Dieter Maurer

Acoustics of the Vowel

Preliminaries

Peter Lang

Acoustics of the Vowel

Preliminaries

It seems as if the fundamentals of how we produce vowels and how they are acoustically represented have been clarified: we phonate and articulate. Using our vocal chords, we produce a vocal sound or noise which is then shaped into a specific vowel sound by the resonances of the pharyngeal, oral, and nasal cavities, that is, the vocal tract. Accordingly, the acoustic description of vowels relates to vowel-specific patterns of relative energy maxima in the sound spectra, known as patterns of formants.

The intellectual and empirical reasoning presented in this treatise, however, gives rise to scepticism with respect to this understanding of the sound of the vowel. The reflections and materials presented provide reason to argue that, up to now, a comprehensible theory of the acoustics of the voice and of voiced speech sounds is lacking, and consequently, no satisfying understanding of vowels as an achievement and particular formal accomplishment of the voice exists. Thus, the question of the acoustics of the vowel—and with it the question of the acoustics of the voice itself—proves to be an unresolved fundamental problem.

www.peterlang.com

Acoustics of the Vowel

Dieter Maurer

Acoustics of the Vowel

Preliminaries

PETER LANG
Bern · Berlin · Bruxelles · Frankfurt am Main · New York · Oxford · Wien

Bibliographic information published by die Deutsche Nationalbibliothek
Die Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data is available on the Internet at ‹http://dnb.d-nb.de›.

British Library Cataloguing-in-Publication Data: A catalogue record for this book is available from The British Library, Great Britain

Library of Congress Control Number: 2015959255

Published with the support of the Swiss National Science Foundation within the scope of the pilot project OAPEN-CH.

Published as volume 12 of the series *subTexte*, edited by Anton Rey, Institute for the Performing Arts and Film, Zurich University of the Arts. www.zhdk.ch/index.php?id=subtexte

Layout and cover design: Jacques Borel, Zurich

ISBN 978-3-0343-2031-3 pb.
ISBN 978-3-0351-0912-2 eBook
ISBN 978-3-0351-9782-2 ePub
ISBN 978-3-0343-2391-8 mobi

© Peter Lang AG, International Academic Publishers, Bern 2016
Hochfeldstrasse 32, CH-3012 Bern, Switzerland
info@peterlang.com, www.peterlang.com

Acknowledgements

We thank the many children, women and men—untrained speakers and professional singers, actresses and actors—who participated in our studies and who lent us their voices for an understanding of what we are questioning.

We thank Anton Rey, Head of the Institute for the Performing Arts and Film, Zurich University of the Arts, Switzerland, for his unswerving support of our research, and we are very happy to have this text published within the publication series *subTexte* of the Institute.

We thank Volker Dellwo, Head of the Phonetics Laboratory at the Department of Comparative Linguistics, University of Zurich, Switzerland, and Daniel Friedrichs, participating in some of the ongoing studies, for all the long discussions of many of the aspects considered in this treatise. These discussions were a strong help with regard to the development of an appropriate concept for the line of argument and the form of presentation.

We owe Heidy Suter, both a linguist and a professional singer, much for here exceptional ability to intellectually re-enact matters of our research and to relate them to voice production, both when speaking and singing herself as a subject of research as well as when advising professional and non-professional singers during recording sessions as a research associate. Moreover, we thank her for her extraordinary effort in editing and proofreading the text.

The strongest influence on the present text exerted Christian d’Heureuse. More than two decades ago, when we first discussed the present matter, he immediately and fully understood the core problem described here, his criticism was always persistent, precise and challenging, and he may become one of the scholars which will provide promising new approaches. Additionally, his conception and implementation of the database software “Media Archive Tool” was and is irreplaceable for the investigation of our large sound corpus.

We thank David Michael for his thoroughly elaborated proofreading and his prudent advices for the improvement of the text and its structure.

We thank Jacques Borel for his talent, taste and expertise in giving the text, tables and figures a fluid, readable and elegant look. We are aware of the many details of the layout structure and the typography he had to consider and of the very time consuming work he was confronted with during the realisation of the book.

We thank the publisher Peter Lang Publishing Group in general, and Adrian Stähli in particular, for accepting to publish this treatise and for the very attentive and proficient support during the editing and production processes.

Funding by the Swiss National Science Foundation SNSF

This publication relates to an ongoing research on voice and vowel qualities by comparing trained and untrained speakers of the three speaker groups of children, women and men (Maurer, Suter, Friedrichs, & Dellwo, 2015, Maurer, n.d.), supported by two grants of the Swiss National Science Foundation SNSF (grant no. 100016_143943 and no. 100016_159350). Within their efforts to fund open-access publications (pilot project OAPEN-CH), the Swiss National Science Foundation has selected this treatise and has also covered the entire financial needs for this publication (grant no. B-OA10_163510.)

The *subTexte* series

As mentioned, this book is published as volume 12 of the series *subTexte*, edited by Anton Rey, Institute for the Performing Arts and Film, Zurich University of the Arts. The *subTexte* series is dedicated to presenting original research within two fields of inquiry: Performative Practice and Film. The series offers a platform for the publication of texts, images, or digital media emerging from research on, for, or through the performative arts or film. The series contributes to promoting art based research beyond the ephemeral event and the isolated monograph, to reporting intermediate research findings, and to opening up comparative perspectives. From conference proceedings to collections of materials, *subTexte* gathers a diverse and manifold reflections on, and approaches to, the performative arts and film.—For further information and a list of all volumes, please refer to:
https://www.zhdk.ch/index.php?id=subtexte

Contents

Part III Experiences and Observations

Part IV Falsification

Part V Commentary

Materials

Materials Part I

Materials Part II

Materials Part III

Experiments

Introduction

Topic and Aims

The vocal cords—when oscillating and modulating air expelled from the lungs—produce a sound (a source sound), which is transformed by the resonances of the pharyngeal, oral and nasal cavities: depending on the position of the larynx, velum, tongue, lips and jaw, different shapes of these cavities are formed thus creating different resonance characteristics, allowing different vocal sounds (phones) to be produced and perceived accordingly. If a vocal sound is perceived to belong to a particular linguistic unit (more precisely, a basic linguistic unit, a phoneme), and if the cavity formed by the pharynx and the mouth remains open, then the sound produced is referred to as a vowel sound and its linguistic identity as a vowel quality or simply as a vowel.

The prevailing theory of vowel acoustics begins with such formulations, or similar ones. According to this theory, with respect to human utterances, the vocal cords produce a general sound, which is transformed into a specific vowel sound by the resonances of the (supralaryngeal) vocal tract: as human beings, we phonate and articulate.

Because of this, vowel sounds, as sounds, are expected to exhibit relative spectral energy maxima in those frequency ranges that correspond to the resonances of the vocal tract during speech production. These spectral energy maxima are known as formants.

Such a perspective gives rise to the prevailing psychophysical principle of the vowel: vowel sounds that are perceived as having the same vowel quality have similar formant patterns, that is, similarly patterned relative spectral energy maxima. By contrast, vowel sounds that are perceived as different vowel qualities have dissimilar formant patterns.

At first glance, such a conception of vowel production and of the subsequent physical representation of vowels seems plausible or even self-evident. Our vocal cords do vibrate when we speak, we do move our mouths (more precisely, our articulators) to form different vocal sounds, and we are indeed often able to “lip read” the words uttered from such movements, an ability highly developed by deaf people.

Moreover, the vast majority of statistical investigations seem to confirm the correlation between vowels and vowel-specific formant patterns.

Vowel synthesis, transforming artificial source sounds by filters, have also proven to be very capable of producing recognisable vowel sounds.

From such a perspective, existing problems in analysing and determining the physical characteristics of vowel sounds according to the perceived vowel quality are not considered with regard to the principle of prevailing theory, but they are related to the dynamics and complexity of the production and perception of speech. Furthermore, isolated vowel sounds, for which a simple and statistical correspondence between the perceived vowel quality and its specific formant pattern is to be expected, are often considered as playing only a marginal role in everyday speech. In speech, vowel sounds and perceived vowel qualities are generally embedded in syntactic and semantic contexts, in contexts of other vocal sounds and of meaning. Such embedded vowel sounds exhibit distinct dynamic processes and above all transitions from one sound to another. Thus, vowel sounds may be perceived in speech even if distinct, static sound elements are absent, and a vowel sound isolated from speech as a sound fragment may be perceived as a different vowel quality than the same sound in connected speech. This explains, for example, why speech can remain intelligible even when substantial interferences or transformations affect its transmission. And so on.

Consequently, the current scientific discussions mainly focus on specific matters such as different types of phonation and articulation when producing vowel sounds, sound variations and dynamic processes related to the respective syntactic and semantic context, sounds produced by speakers of different age and gender and corresponding normalisation attempts, attempts to improve formant pattern estimation and attempts to relate acoustic findings and processes of auditory perception. And so on.

Having said that, notwithstanding, the present consideration returns to the basic assertion of the current acoustic theory of the vowel cited at the beginning of this introduction. It presents a critical reading, indeed a falsification, of this assertion. Further, it seeks to demonstrate that whereas prevailing theory indicates (is an index of) the actual physical characteristics of vowels, it fails to designate these characteristics adequately. As such, this work highlights an unresolved fundamental problem of the voiced speech sound, and thus of the voice as such, and raises this problem once again for discussion.

The form of this treatise is, in part, unusual in a scientific context. However, with the exception of the four aspects discussed below, this introduction dispenses with lengthy prefatory explanations. In its course, the argument and its form of presentation should become self-evident. Besides, additional comments in the afterword further expand on, and hopefully clarify, matters.

As mentioned, however, four introductory aspects are to be explained at this juncture. They concern linguistic expression and style, referencing, the significance of argumentation and the perspective adopted here.

Many parts of the main body of the text are “abstract” in their presentation, which is to say, they are “technical”. This might complicate the reading. Moreover, with the exception of Sections 1.10, 2.1 and 2.2, the text is not accompanied by illustrated examples or tables listing statistical data. Further, from Part III onwards, the text requires the reader to reflect thoroughly on the prevailing theory of the vowel as presented in Part I. The text also calls upon the reader to approach the related terms and concepts and the statistical values for formant patterns with a certain amount of self-assurance. However, such a procedure is necessary: the text insists on the discussion of a few fundamental reflections and general facts, and their interrelations, in the attempt, as mentioned, to highlight a fundamental problem.

Most of the issues considered here have already been discussed in the literature, and most of the corresponding publications were presented by other authors. However, they have often been interpreted in a way that differs from the point of view taken here. Yet, aside from the illustrations and tables mentioned, the text largely dispenses with explicit references to previous studies, including our own, so as to pursue its main argument without any detailed discussion and referencing of individual aspects. The Materials section (for the structure of this text, see below), however, includes a considerable number of citations, together with references to existent publications. Moreover, as mentioned above, my colleagues and I have discussed most of the aspects addressed here elsewhere. The present text is new in its course of argument, as is the arrangement and presentation of citations, comments, illustrated examples and outlines of experiments in the Materials and Experiments sections. However, new content but concerns aspects discussed in Part V and in the afterword, some presentations in the Materials section (see Sections M8.2, M10-A) and some examples in the Experiments section.

The empirical basis of this treatise, to which many of the statements made here refer, above all in Part III and IV, consists of recordings from various areas of everyday life, the entertainment sector and art, that is, stage voices in music and straight theatre. Whereas one part of these recordings forms the basis of single, published investigations undertaken in the past, another part is unpublished and the corresponding recordings have not been subject to any further identification tests, apart from the identification by the author. Thus, the reflections in Part

III and IV lay no claim to consistent verification in terms of the existing scientific standards. Instead, they are formulated as hypotheses in view of general findings that are conceivable or even predictable. In line with this, illustrated examples are given in the Materials section.

Accordingly, this treatise is limited to presenting and interrelating those reflections, experiences and observations anew that tend to refute the assertion that vowel qualities are physically represented by formant patterns. If this undertaking proves successful, then—to repeat and insist—this once again raises the question of the voiced speech sound as a fundamental problem.

The argument focuses on and is limited to the relationship between individual vowel sounds, perceived vowel qualities, corresponding sound spectra and formant patterns in the sense of patterns of formant frequencies. Formant bandwidths and amplitudes, to mention two aspects of possible importance, are not discussed in detail.

This treatise adopts a decidedly psychophysical perspective. Only general reference is made to the production and perception of sounds: sound production is referred to because the concept of formants itself refers to vocal tract resonances and also because this relationship needs to be emphasised repeatedly in the course of the argument. Sound perception is referred to because the reflections presuppose that the vowel sounds discussed can be attributed to (perceptually identified as) the specific vowel qualities in question. Beyond these general references, however, production and perception are not further discussed.

By no means does excluding a consideration of further details of sound production and perception from the present discussion suggest that these aspects are unimportant for the physical description of vowels. Doing so merely serves to focus on the psychophysical question of the vowel: given that an utterance—or its reproduction, manipulated or not, or a synthesis for that matter—is perceived as a specific vowel quality, which describable physical characteristic or which ensemble of physical characteristics may be said to represent that quality?

In line with this, the argument focuses on voiced oral vowel sounds produced either in isolation or isolated (extracted) from syntactic and semantic contexts. Thus, nasalisation and the syntactic and semantic context are as such also excluded from discussion. With regard to the different types of phonation, only whispered vowels are considered here, and are mentioned only briefly. Again, this is intended to enable the straightforward discussion of the psychophysical question of the vowel.

In no way does limiting the consideration to voiced vowel sounds isolated from syntactic and semantic contexts and exhibiting quasi-static spectral characteristics suggest that such static spectral characteristics are absolutely necessary for vowel recognition. Thus, the limitation adopted here does not run counter to the phenomena described in the literature concerning the possibility of vowel recognition in the case of sounds exhibiting predominantly dynamic spectral characteristics. This study does, however, refute the conclusion partly drawn in the literature that isolated vowel sounds or sound fragments with quasi-static spectral characteristics are essentially less easily recognisable than vowel sounds occurring in a syntactic context and associated with distinctively dynamic spectral characteristics and transitions, or that the former are even insufficiently recognisable. The afterword will return to this aspect.

As this treatise reveals, there is good reason to understand and pursue the psychophysics of voiced speech sounds as a phenomenology: that is, for research not to start from a model and to conduct single experiments based on it, but instead from an open-ended and continually expanding collection and compilation of vocal utterances, together with a simultaneously evolving description of their physical characteristics related to perceived vowel qualities.

With the adoption of such a perspective, it may become understandable why the present treatise, despite its narrow focus on phonetics, is not published by a correspondingly specialised university institute, but rather by an institute affiliated with an arts university. In contrast to many approaches, here there is no assumption of a “normal case” of speaking, based on which “other kinds” of utterances are treated as “special cases”, such as emotionally tinged utterances with corresponding variations of fundamental frequency and vocal effort, or utterances produced with a “head voice”, or shouting, or singing, or acting, and so on. Such a view is not borne out either by everyday experience or by creative expression.

In the first instance, vocal utterances and thus speech sounds do not obey narrowly restricted norms of production, and the only reliable representation of the human voice and speech that critical reflection and the development of an empirical approach can refer to, is the artistic or interpretative utterance. Only art is able to represent the “artificiality”—that is, the reduction, standardisation and coding—of any specific utterance whilst, at the same time, overcoming it, albeit only to some extent. Referring to the fact that any utterance is a token, not a type, only art involves the quasi-systematic variation of vocal utterances, without which any investigation and consideration of the relationship between the sounds produced and the qualities perceived run the risk of interpreting findings about concrete and specific utterances as findings about general characteristics and principles. The afterword will return to this point, too.

Vowel sounds, perceived as isolated single sounds, can be intelligible. This fact is central to human voice and speech: vowel sounds must be intelligible as such because elementarisation—manifest in the aptitude of speech for a phonetic system of writing—is at the core of speech and language. Such an assumption underlies the reflections advanced here. Consequently, vowel qualities—or rather the differences between the vowel qualities of any given language—are considered to be represented physically. As this treatise aims to show, it is likely that such a representation cannot be derived from a physical model but, instead, needs to be described as an achievement of the human voice itself.

Structure

This treatise is divided into a main body and the two sections Materials and Experiments.

The main body is divided into five parts, followed by an afterword:

- Part I reviews the prevailing theory of the physical characteristics involved in vowel representation.
- Part II presents reflections that, according to the author's reading of the literature, oppose the understanding of the theory, that is, its intellectual re-enactment and validation.
- Part III formulates several hypotheses about the actual relationship between vowel sounds, sound spectra and formant patterns. These hypotheses refer to the recordings mentioned in the introduction and to related analyses and observations.
- Part IV explains why the reflections, experiences and observations compiled here falsify prevailing theory.
- Part V discusses the resulting state of affairs and points to the need to devise a phenomenology and to develop a new theory. This part also includes an excursus on the harmonic spectrum as being vowel specific.
- The afterword presents various additional comments.

The Materials section contains selected excerpts from the literature, commented on in part, and presents exemplary series of vowel sounds and related acoustic analysis. An extended version of the materials is also presented in digital form online; please refer to:
http://www.phones-and-phonemes.org/vowels/acoustics/preliminaries

The treatise concludes with a list of possible experiments that allow for empirical exploration of the problems discussed here under laboratory conditions.

The main body of this text—excluding Section 13.3 which was added to this edition separately—is a revised and translated version of an earlier publication in German titled *Akustik des Vokals – Präliminarien* (Maurer, 2013). The Materials section is an entirely revised and substantially enlarged version of the digitally published sound archive of the German version. The Experiments section is new.

Tables and figures are numbered separately for each chapter. In the Materials section, the figure legends are positioned at the top.

The citations in the Materials section are given in their original version, including the corresponding writing style and format.

If included in the citations of the Materials section, figures referred to are not given in this treatise and publications referred to are not listed in the References section. For corresponding details, please consult the publications in question.

Terms and Notation

To facilitate reading, the key terms, notation style and abbreviations adopted in the text are explained below.

Vocal tract. The term “vocal tract” is used as a short form referring to the supralaryngeal (or supraglottal) vocal tract in terms of the pharyngeal, oral and nasal cavities.

Sound, vocal sound, speech sound. The distinction between “sound” (*Klang,* a quasi-periodic sound with a pitch and a harmonic spectrum) and “noise” (*Geräusch,* a non-periodic sound with no pitch) is made in the English version of this treatise only when it matters for the argument. In all other cases, the term sound is used as a generic term.

The distinction between “vocal sound” (*Laut,* voiced or unvoiced) and “speech sound” (*Sprachlaut*) is made here to refer to the fact that not every vocal utterance is linguistic in a narrow sense, that is, not every vocal utterance can be attributed to a phoneme.

Vowel sound, vowel quality. The term “vowel sound” refers to a single concrete vocal sound possessing linguistic value, that is, a phone. It is termed a vowel sound—in distinction from other phones—because it is perceived to have vowel quality (see below). According to the literature, vowel sounds are quoted in square brackets, for instance [a]. In part, additional suprasegmental characteristics are also given, for instance, in the distinction between [a:] in the German word *Kahn* and [a] as in *Kamm* (long and short vowel sound).

The term “vowel quality” denotes a class of vowel sounds of an individual language, that is, a phoneme. Thus, concrete single vowel sounds as phones are attributed to abstract classes of vowel qualities as phonemes. In the literature, vowel qualities are quoted between two slashes, such as /a/.

Vowel qualities are quoted according to the symbols of the International Phonetic Alphabet (revised to 2005).

Whenever context allows, the terminological distinction between vowel sounds and vowel qualities is shortened to the distinction between vowel sounds and vowels, or sounds and vowels.

In general, the reflections, experiences and observations presented in Part II refer to the long vowels of Standard German /i, y, e, ø, ɛ, a, o, u/. Included here is the vowel /ɑ/, which is encountered in the Swiss pronunciation of Standard German. Therefore, the corresponding vowel area is assigned as /a–ɑ/, including all allophones of /a/ or /ɑ/. In the Materials section, some sounds of the vowel /ɔ/ are also included in order to discuss the spectral phenomena occurring between /a–ɑ/ and /o/.

In the text, these vowels are often subsumed under three groups: as front vowels /i, y, e, ø, ɛ/, as vowel area /a–ɑ/ and as back vowels /ɔ, o, u/. The terms “front vowels” and “back vowels” are adopted from the literature, but they have no further significance here. In particular, their attributed relationship with the tongue position in sound production plays no part.

Note that, depending on the subject of discussion or demonstration, the vowel order sometimes deviates from a consistent front–back direction.

The discussion focuses on German vowels because most of the author’s experiences and observations to date concern the sounds of the German language. However, the corresponding general statements also apply to other individual languages.

Fundamental frequency. The term “fundamental frequency” refers to the measured fundamental frequency of the sound. However, no distinction is made in the text between fundamental frequency and pitch, because such a differentiation is insignificant to the discussion. Thus, both terms are used synonymously.

Here, F0 is used as an abbreviation for fundamental frequency. Thereby, depending on the context, the abbreviation refers to fundamental frequency in general terms or to a specific level (or range) of fundamental frequency in Hz.

Spectrum, harmonic spectrum. The term “spectrum” refers to the sound spectrum of a vowel sound, generally resulting from a of Fourier analysis. In certain cases, the term can refer to a spectrogram because, in many empirical studies, formant values are appraised or verified on the basis of this type of spectrum. Important differences exist between these two types of spectral representation. However, because the present consideration concerns only general aspects, with a few exceptions, these differences are negligible here. In the exceptional cases referred to, corresponding differentiations will be made.

The term “harmonic spectrum” refers to a series of harmonics in the sound spectrum, a series of partials (sinusoidal components of a complex tone) whose frequencies are an integral multiple of the fundamental frequency. However, even if this terminology is common, it is not unquestionable. Above all, vowel spectra may not always exhibit the first (or the first few lower) harmonics (consider, for example, high-pass filtering), and the perceived pitch may not always correspond to the acoustically measured fundamental frequency. The emerging terminological question is left open here.

Relative spectral energy maximum, spectral envelope peaks. The term “relative spectral energy maximum” refers to a narrowly delimited frequency range of a spectrum that exhibits significantly increased energy compared to the frequency ranges immediately preceding and immediately following such spectral enhancement. In the literature, such relative maxima are in general determined on the basis of evaluating a spectral envelope (in the sense of an imaginary smooth line drawn to enclose an amplitude spectrum, see Chapter M6) and are termed “spectral envelope peaks”.

Formant, formant pattern, formant statistics. The term “formant” is used in different ways in the literature. In particular, it can refer either to a resonance as a physical property of the vocal tract, to a spectral envelope peak as a physical characteristic of a vowel sound, or to a filter as a part of a series of filters related to an analytical method of speech processing. The term can also denote two or even all three of these aspects at the same time.

Here, a basic distinction is made between the resonances of the vocal tract and the formants of the vowel sound produced. Such a distinction corresponds to the perspective adopted, namely, not to discuss the production of a vowel sound but, instead, the vowel sound itself, including the related perception of the corresponding vowel quality.

At the beginning of the present contribution, the term “formant” refers to spectral envelope peaks as well as to filters used in speech analyses, because in the literature, when formulating vowel-specific physical characteristics is at issue, both characteristics are generally assumed to correspond. In the course of argument, when considering current empirical studies and corresponding formant values, it will become clear that, today, the concept of vowel-specific formants is generally limited to the filters used in speech analyses.

In the literature, formant abbreviations are often used to distinguish between formant frequencies, bandwidths and amplitudes or levels. Such a distinction is dispensed with here. Instead, single formants are referred to as F1, F2, F3, . . . F(i) and configurations as F1–F2 or F1–F2–F3, termed as “formant patterns”. Depending on the context, as is the case for F0, these abbreviations refer to formants in general terms or to specific levels (or ranges) of formant frequencies in Hz. Formant bandwidths and amplitudes play no substantial role in the discussions.

Accordingly, formants and formant frequencies of vowel synthesis are abbreviated as F1’, F2’, F3’, … F(i)’ and vocal tract resonances are abbreviated as R1, R2, R3 … R(i).

Note that abbreviations of fundamental, formant and resonance frequencies with lower case numbers—F_0, F_1, F_2, F_3 . . . —are used only in tables showing formant statistics and in citations.

If references are made to formant values as given in formant statistics for voiced vowel sounds, corresponding investigations generally concern formant measurements for sounds produced in citation-form words with medium or spontaneous vocal effort at related fundamental frequencies, in a quiet room in front of a microphone. These values are often assumed to be representative of so-called “normal speech”, and the limitation of measurement in terms of not considering vowel sounds produced by single speakers at very different fundamental frequencies is often ignored and remains unmentioned. (Please note that, for reasons explained in the text and on the basis of observations documented in the Materials section, we do not consider the expression "normal speech" appropriate and, with regard to both fundamental frequency and formant patterns, we question the representative character of sounds produced in citation-form words for the utterances in everyday life. However, the analysis of sounds produced in citation-form words may be comparable to the analyses of relaxed speech.)

For the ongoing debate on terminology and abbreviations, please refer to Section M6.

LPC. The abbreviation "LPC" stands for Linear Predictive Coding, which is a method used to analyse the acoustic characteristics of speech sounds.

Indications of frequency ranges and frequency limits. Frequency ranges and frequency limits for observed aspects of vowel spectra and formant patterns and for methodological considerations are given as rough approximations. (Note that the vowel-specific frequency range for sounds of back vowels and of /a–ɑ/ is given as ≤ 1.5 kHz. However, for some sounds of /a/, the upper limit of this frequency range may exceed 1.5 kHz; see Section 2.1, for example.)

Speaker group. The term "speaker group" is used as a short form for age- and gender-specific groups of speakers, that is, children, women and men, as they are referred to in the literature. (Note that some scholars term these groups age- and size-specific speaker groups; others differentiate further in terms of age, gender and size.) As explained in the text, the differentiation of these three speaker groups is motivated by three different average vocal-tract sizes.

In the literature, age- and gender-specific speaker groups are generally given in the order "men, women, children". However, a systematic adherence to this order carries with it an age and gender bias and poses a corresponding problem. Moreover, it mirrors a tradition in phonetics to favour the analysis of men's voices (see also Chapter M6). If, in this text, other studies are referred to, the order of listing accords to the cited study. Apart from those cases, the order is inverted. This makes for a formal inconsistency of the text. For future investigations in the field of phonetics, the standard for the listing order should be discussed and an adequate linguistic form should be established.

Part I Prevailing Theory and Empirical References

The first part of the main text reviews the prevailing theory of the physical characteristics involved in vowel representation.

1 Prevailing Theory

1.1 General Acoustic Characteristics of Vowel Sounds

With respect to human utterances, the following is said to apply: The vocal cords—when oscillating and modulating air expelled from the lungs—produce a sound (a source sound), which is transformed by the resonances of the pharyngeal, oral and nasal cavities: depending on the position of the larynx, velum, tongue, lips and jaw, different shapes of these cavities are formed thus creating different resonance characteristics, allowing different vocal sounds (phones) to be produced and perceived accordingly. If a vocal sound is perceived to belong to a particular linguistic unit (more precisely, a basic linguistic unit, a phoneme), and if the cavity formed by the pharynx and the mouth remains open, then the sound produced is referred to as a vowel sound and its linguistic identity as a vowel quality or simply as a vowel (see the introduction).

According to this approach, the production of a vowel sound involves two quasi-independent processes: the production of sound and its transformation by resonance, termed phonation and articulation. Sound production or phonation is not vowel specific. By contrast, the respective resonance effect or articulation is vowel specific. The two-part model arising from such an understanding of speech production is known as the source-filter model of speech production.

Physiologically, the perceived linguistic identity of a vowel sound corresponds to a vowel-specific articulation in terms of an ensemble of possible positions of the vocal tract, which produce quasi-identical (that is, very similar) patterns of resonances.

Acoustically, the perceived linguistic identity of a vowel sound corresponds to vowel-specific spectral energy maxima, which are quasi-identical to the vowel sounds of the same vowel quality. In acoustic analysis, these spectral energy maxima appear as spectral envelope peaks, generally known as formants.

In cases of whispered vowels, phonation does not involve periodic sound, but noise.

1.2 Language-Specific Acoustic Characteristics of Vowel Sounds

In general, not all formants of a vowel but only the first two (lowest in their frequencies) correspond to a perceived vowel quality. The higher formants refer to other qualities of vocal expression.

In certain languages, exceptions to this rule concern sounds of high front vowels and of r-coloured front vowels. In such cases, the frequencies of the first two formants of sounds of two vowels are quasi-identical, and only the difference within the respective frequency of the third formant corresponds to the difference in the perceived vowel quality.

1.3 Speaker Group-Specific Acoustic Characteristics of Vowel Sounds

In general, children have a considerably smaller vocal tract than adults, just as women have a smaller tract than men. Because of this, the acoustic correspondence between vowel qualities and formant patterns, formulated above in general terms, are related to the different speaker groups of children, women and men in terms of age and gender: thus, for each group and the respective average vocal-tract length, the sounds of a given vowel correspond physiologically to a specific articulation involving a specific resonance pattern, and acoustically to a specific formant pattern.

1.4 Phonation Type-Specific Acoustic Characteristics of Vowel Sounds and Limitation to Voiced Oral Sounds

The geometry, and thus the resonances, of the glottal area of the vocal tract vary for different types of phonation. Therefore, for example, the formant patterns of voiced and whispered vowel sounds of one perceived vowel quality differ substantially. Consequently, the acoustic correspondence between vowels and formant patterns must also be related to the various types of phonation: thus, for each single speaker group too, depending on the respective average vocal-tract length and type of phonation, the sounds of a given vowel correspond physiologically to a specific articulation involving a specific resonance pattern, and acoustically to a specific formant pattern.

Existing empirical reference values for formant patterns—formant statistics—predominantly concern voiced vowel sounds produced in citation-form words, comparable to relaxed speech with limited variation of fundamental frequency. Statistical reference values for vowel sounds involving other phonation types are rare. Further, the various kinds of phonation are related to different methodological problems of formant pattern estimation. The following discussion therefore concentrates on voiced vowel sounds. Only passing reference is made to vowel sounds involving other types of phonation.

Nasal vowel sounds are also related to specific methodological problems of formant pattern estimation and are therefore not considered here either. Hence, the following discussion is restricted to voiced oral vowel sounds.

1.5 Limitation to Isolated Vowel Sounds

The perception of vowel sounds can depend on the semantic context: in some cases, a vowel sound embedded in a syllable or a word may be perceived as a certain vowel quality, which, if extracted from the context and presented as an isolated sound fragment, may be perceived to have a different quality.

Whether or not the perception of vowel sounds can also depend directly on their syntactic context, for example when produced in nonsense syllables or non-words, is left open here.

Consequently, the discussion of the acoustic correspondence between vowels and formant patterns is further restricted to vowel sounds produced in isolation or extracted from a concrete syntactic or semantic context.

1.6 Limitation to Vowel Sounds as Monophthongs with Quasi-Constant Sound Characteristics

In general, single voiced oral vowel sounds that feature a perceivably constant vowel quality, a quasi-constant fundamental frequency and a quasi-constant loudness throughout their entire duration, exhibit the characteristics of a quasi-periodic sound wave. With regard to the physical representation of the vowel quality, the corresponding spectral characteristics of such vowel sounds can be described in terms of the average harmonic spectrum of a sound, including the respective spectral envelope and, if occurring, its peaks, and with the latter the corresponding formant patterns.

This does not apply to vowel sounds whose perceived vowel quality, fundamental frequency, or loudness are subject to substantial variation. So as to exclude the ensuing questions about a possible influence of such variations on the perception of vowel qualities and their spectral representation, the following discussion focuses on vowel sounds as monophthongs that possess quasi-constant sound characteristics. Vowel sounds lacking such sound characteristics are again discussed only in passing and by way of incidental comments.

1.7 Speech Community-Specific Acoustic Characteristics of Vowel Sounds

In the first instance, the acoustic correspondence between vowels and formant patterns only applies to speakers and listeners belonging to the same speech community: quasi-constant vowel production and perception exist among the members of such a community, who accordingly attribute sound variations either to one and the same vowel quality or to different vowel qualities.

However, the methodological question of how to determine empirically the consistency of such an attribution is not discussed further here. The present discussion generally assumes that the vowel sounds considered, when subjected to a concrete identification test involving listeners of one speech community, specially trained for such a perception test, will exhibit a consistent attribution substantially above a 50% level for any given vowel quality.

Yet to be discussed elsewhere are correspondences that reach beyond one particular speech community as well as one particular linguistic community.

1.8 The Prevailing Theory of Physical Vowel Representation

Given that

- vowel sounds are produced by individuals belonging to one of the three speaker groups of children, women, or men of a given speech community;
- vowel sounds are either produced as isolated voiced oral sounds or as voiced oral sound fragments extracted from their concrete syntactic and semantic context of production, with neither transitions at the beginning nor the end;
- vowel sounds are produced with a quasi-constant fundamental frequency and loudness and exhibit the characteristics of a quasi-periodic sound wave;
- vowel sounds are perceived as belonging to one vowel quality by other individuals of the same speech community;

then the following applies to the individual vowel sound:

- physiologically, its perceived linguistic identity as a specific vowel quality corresponds to a specific position of the vocal tract which, by means of (according to their frequency position) the first two (in some cases of high front vowels and r-coloured front

vowels of certain languages the first three) resonances of the tract, transforms the source sound of the vocal cords to that sound;
– acoustically, its perceived vowel quality hence corresponds to the first two (or the first three) lower formants of the sound spectrum.

Given the same assumptions, for two vowel sounds perceived as two different vowel qualities, this implies that:

- physiologically, the difference in vowel perception corresponds to two different positions of the vocal tract, each with a different pattern of the lower two (or three) resonances;
- acoustically, the difference in vowel perception corresponds to two different patterns of the first two (or first three) lower formants of their respective spectra.

For the sounds of a particular vowel, albeit produced by speakers of different speaker groups, this implies that:

- physiologically, their perceived linguistic identity as the same vowel quality corresponds to different patterns of the first two (or first three) lower resonances of the vocal tract, related to the difference in average vocal tract length of the speaker groups compared;
- acoustically, their perceived linguistic identity as the same vowel quality hence corresponds to different speaker group-specific patterns of the first two (or first three) lower formants of the respective spectra.

These formulations are central to the prevailing theory of the physical representation of the vowel.

1.9 Formalising Prevailing Theory

For isolated, voiced oral vowel sounds that possess quasi-constant sound characteristics and are produced by individuals belonging to a given speech community and a given speaker group of children, women, or men, the following applies:

- vowel sounds perceived as one vowel quality correspond to quasi -identical (that is, similar) R1–R2 (R1–R2–R3 in some cases of high front vowels and r-coloured front vowels in certain languages) and, at the same time, quasi-identical F1–F2 (or F1–F2–F3, respectively);

- vowel sounds perceived as different vowel qualities correspond to dissimilar R1–R2 (R1–R2–R3, respectively) and, at the same, dissimilar F1–F2 (F1–F2–F3, respectively).

1.10 Illustration

Figure 1 is an illustration of this prevailing understanding of vowel production and perception, typical of many publications in the field. (The illustration is simplified in that it lacks any differentiation of the actual characteristics of the source spectrum on the one hand, and of the radiation impedance occurring when a sound is emitted into space on the other. This differentiation is not discussed further here because it is irrelevant to the present argument.)

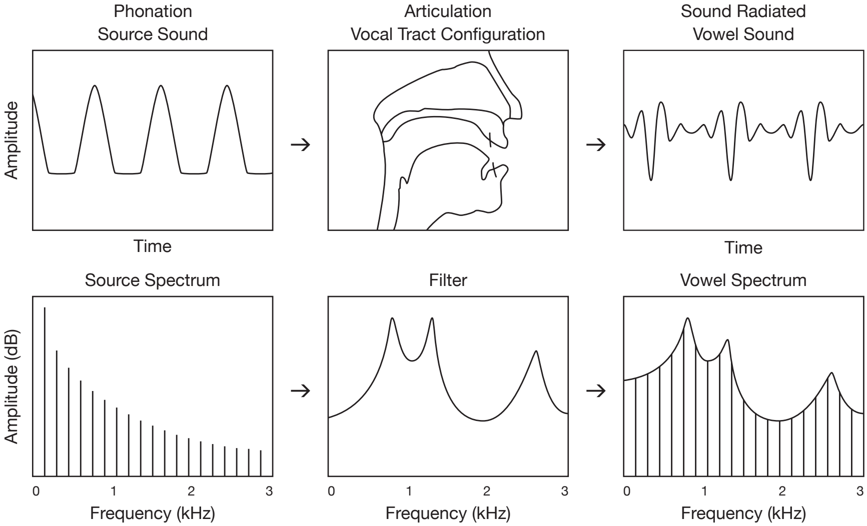


Figure 1. Illustration of prevailing theory.

Figure 2 shows examples of spectra, filter curves (LPC curves) and formant patterns (maxima of filter curves) of specially selected sounds of different vowels. This kind of illustration, which is limited to the acoustic perspective, is also widespread in the literature.

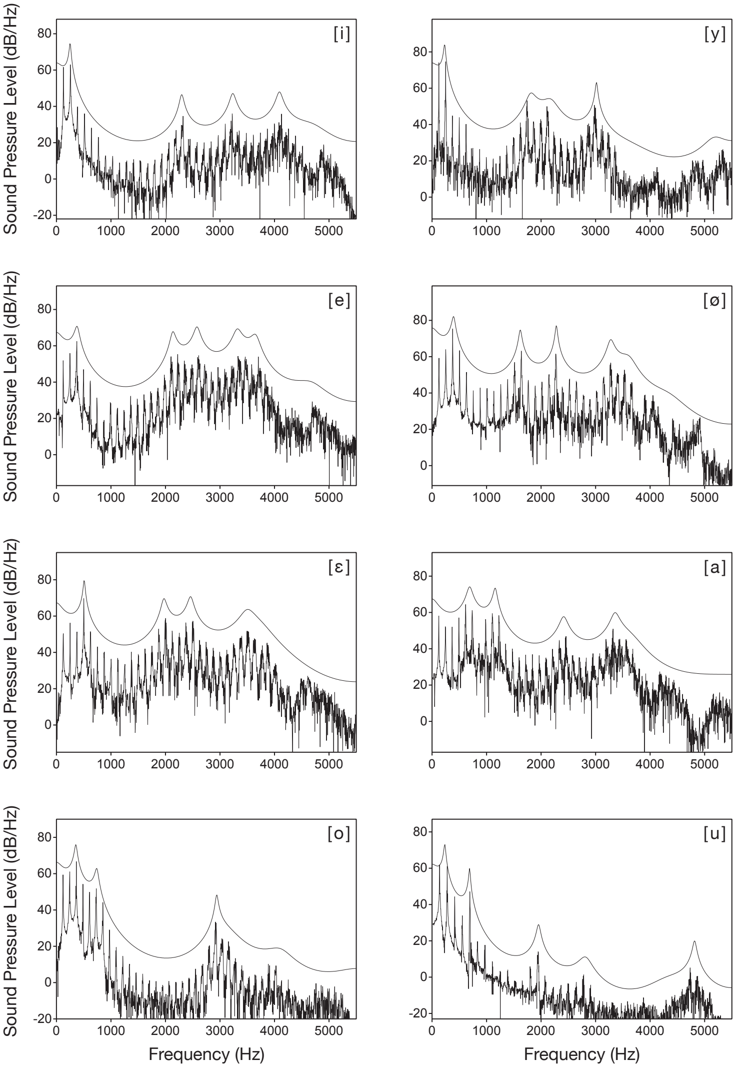


Figure 2. Examples of sounds of different vowels produced in isolation by adult male speakers at fundamental frequencies of 120–140 Hz. Corresponding spectra and filter curves (LPC curves) are shown. The examples are specially selected in order to illustrate prevailing theory.

2 Prevailing Empirical References

2.1 General References

The first extensive statistical study of the correspondence between vowels and formant patterns with reference to the three speaker groups, children, women and men was conducted by Peterson and Barney (1952, see Table 1, and Figure 1). Their study focused on American English and later became one of the dominant references in the literature.

Hillenbrand, Getty, Clark, and Wheeler (1995) used new recording and measurement methods (digitisation, LPC analysis) as well as an extended set of 12 vowels to replicate the classic study of Peterson and Barney (see Table 2).

Parallel to Peterson and Barney, Fant (1959) published a statistical study of Swedish vowels. However, Fant's study was limited to the two speaker groups of men and women (see Table 3).

Presumably, the vowel-specific formant patterns as given by Peterson and Barney (1952) and Hillenbrand et al. (1995) are the most widely cited references in general discussions of the physical characteristics of vowels. The statistics of Fant (1959) also played an important role in the development of the source-filter theory.

Table 1. Formant statistics for American English vowels (Peterson & Barney, 1952). Average values for fundamental frequency, formant frequencies F_1–F_2–F_3 (Hz) and formant amplitude levels L_1–L_2–L_3 (dB) are given for the speaker groups of men (M), women (W) and children (Ch).

		i	ɪ	ɛ	æ	ɑ	ɔ	U	u	ʌ	ɝ
Fundamental frequencies	M	136	135	130	127	124	129	137	141	130	133
(cps)	W	235	232	223	210	212	216	232	231	221	218
	Ch	272	269	260	251	256	263	276	274	261	261
Formant frequencies (cps)	M	270	390	530	660	730	570	440	300	640	490
	W	310	430	610	860	850	590	470	370	760	500
F_1	Ch	370	530	690	1010	1030	680	560	430	850	560
	M	2290	1990	1840	1720	1090	840	1020	870	1190	1350
F_2	W	2790	2480	2330	2050	1220	920	1160	950	1400	1640
	Ch	3200	2730	2610	2320	1370	1060	1410	1170	1590	1820
F_3	M	3010	2550	2480	2410	2440	2410	2240	2240	2390	1690
	W	3310	3070	2990	2850	2810	2710	2680	2670	2780	1960
	Ch	3730	3600	3570	3320	3170	3180	3310	3260	3360	2160
Formant amplitudes	L_1	−4	−3	−2	−1	−1	0	−1	−3	−1	−5
(db)	L_2	−24	−23	−17	−12	−5	−7	−12	−19	−10	−15
	L_3	−28	−27	−24	−22	−28	−34	−34	−43	−27	−20

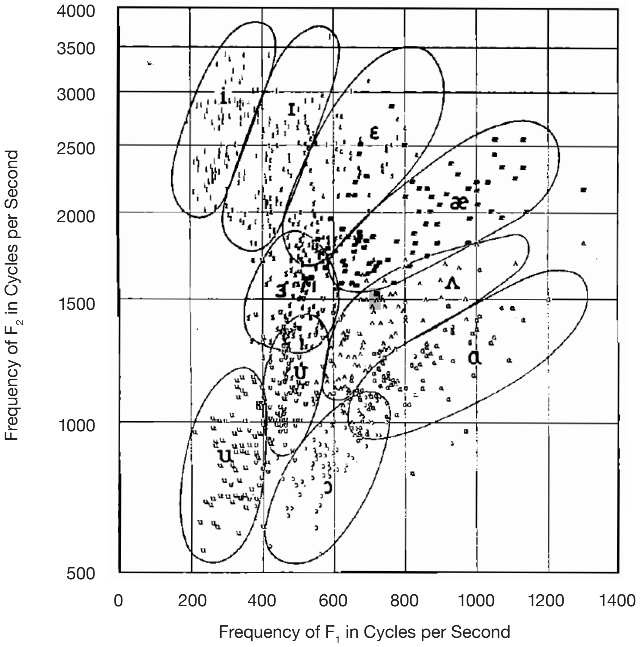


Figure 1. Illustration of the distribution of the first two formants for American English vowels (Peterson & Barney, 1952; data of 76 speakers, 33 men, 28 women, 15 children). x-axis = formant frequencies (Hz) for F1; y-axis = format frequencies (Hz) for F2. (Reproduced with kind permission of Peterson & Barney [1952]. Copyright 1952, Acoustical Society of America.)

Table 2. Formant statistics for American English vowels (Hillenbrand et al., 1995). Average values for fundamental frequency F0 (Hz) and formant frequencies F1–F2–F3–F4 (Hz) are given for men (M), women (W) and children (C).

		/i/	/ɪ/	/e/	/ɛ/	/æ/	/ɑ/	/ɔ/	/o/	/U/	/u/	/ʌ/	/ɜ˞/
F0	M	138	135	129	127	123	123	121	129	133	143	133	130
	W	227	224	219	214	215	215	210	217	230	235	218	217
	C	246	241	237	230	228	229	225	236	243	249	236	237
F1	M	342	427	476	580	588	768	652	497	469	378	623	474
	W	437	483	536	731	669	936	781	555	519	459	753	523
	C	452	511	564	749	717	1002	803	597	568	494	749	586
F2	M	2322	2034	2089	1799	1952	1333	997	910	1122	997	1200	1379
	W	2761	2365	2530	2058	2349	1551	1136	1035	1225	1105	1426	1588
	C	3081	2552	2656	2267	2501	1688	1210	1137	1490	1345	1546	1719
F3	M	3000	2684	2691	2605	2601	2522	2538	2459	2434	2343	2550	1710
	W	3372	3053	3047	2979	2972	2815	2824	2828	2827	2735	2933	1929
	C	3702	3403	3323	3310	3289	2950	2982	2987	3072	2988	3145	2143
F4	M	3657	3618	3649	3677	3624	3687	3486	3384	3400	3357	3557	3334
	W	4352	4334	4319	4294	4290	4299	3923	3927	4052	4115	4092	3914
	C	4572	4575	4422	4671	4409	4307	3919	4167	4328	4276	4320	3788

Table 3 (see pages 25 and 26). Formant statistics for Swedish vowels (Fant, 1959). Values for fundamental frequency F0 and formant frequencies F1–F2–F3–F4 (Hz) and for formant amplitude levels L1–L2–L3–L4 (dB). Gj–n = values for a single adult male speaker; M–r = values for a single adult female speaker; M = average values for men; F = average values for women.

V	S	F_0 Hz	L_0 dB	F_1 Hz	L_1 dB	F_2 Hz	L_2 dB	F_3 Hz	L_3 dB	F_4 Hz	L_4 dB
[u]	Gj–n	125	0.5	325	3	640	2.5	2400	–40	3500	–40
	M–r	256	6	270	6	740	–5	2550	–37	(3200	–35)
	M	127	–0.5	307	4.5	730	–8	2230	–37	3300	–38
	F	222	2	340	5	690	–6	2900	–43	(4000	–45)
[o]	Gj–n	125	2	405	6	700	–1	2450	–32	3200	–29
	M–r	256	5	380	8	850	–10	2800	–38	— —	— —
	M	132	–0.5	402	6	708	–2	2460	–31	3150	–33
	F	223	1	433	7	815	–9	2840	–38	(3600	–37)
[ɔ]	Gj–n	125	1.5	500	7	800	–1	2530	–26	3150	–25
	M–r	257	5	510	8	900	0	2800	–35	(3000	–35)
	M	123	–0.5	487	6	825	1	2560	–26	3250	–28
	F	217	–2	518	5	840	–6	2825	–36	(3500	–40)
[ɑ]	Gj–n	125	1.5	600	7	935	0	2620	–18	3150	–22
	M–r	250	4	650	6	1125	0	2800	–18	— —	— —
	M	126	–1	582	7.5	940	4	2480	–21	3290	–20
	F	225	–1	682	4	1075	4	2930	–22	(3800	–28)
[a]	Gj–n	125	2.5	680	6	1075	–1	2720	–18	3350	–18
	M–r	255	3	770	8	1250	1	2800	–15	— —	— —
	M	124	–1	680	6	1070	1	2520	–10	3345	–20
	F	215	–1	860	4	1195	4	2830	–23	— —	— —
[æ]	Gj–n	125	2.5	560	6	1740	–7	2470	–12	3200	–16
	M–r	257	3	600	6	1740	–7	2900	–20	— —	— —
	M	125	0.5	606	7	1550	–3	2450	–12	3400	–15
	F	213	0	785	5	1820	–6	2950	–18	(3600	–17)
[ɛ]	Gj–n	125	2.5	480	5	1870	–7	2480	–9	3250	–18
	M–r	255	4	535	7	1870	–8	2600	–18	— —	— —
	M	125	0	438	6	1795	–9	2385	–12	3415	–19
	F	214	0	545	5	2140	–11	2860	–20	— —	— —

V	S	F_0	L_0	F_1	L_1	F_2	L_2	F_3	L_3	F_4	L_4
		Hz	dB	Hz	dB	Hz	dB	Hz	dB	Hz	dB
[e]	Gj–n	125	1	325	4	2210	–11	2650	–12	3400	–20
	M–r	256	7	320	7	2200	–13	2700	–12	— —	— —
	M	124	0	334	6	2050	–12	2510	–13	3400	–16
	F	215	1	365	6	2540	–15	2950	–18	— —	— —
[i]	Gj–n	140	2	275	2	2205	–17	3100	–12	3500	–17
	M–r	256	7	270	7	2200	–23	3100	–15	— —	— —
	M	128	0	256	3	2066	–23	2960	–20	3400	–23
	F	218	3	278	5	2520	–24	3450	–24	(3900	–28)
[y]	Gj–n	140	3	275	2	2050	–12	2300	–15	3325	–21
	M–r	257	8	260	8	2070	–22	2820	–17	3300	–24
	M	128	1	257	4.5	1928	–17	2421	–19	3300	–24
	F	215	5	270	6	2480	–21	2920	–23	3575	–26
[ʉ]	Gj–n	125	1	290	3	1690	–12	2170	–15	3300	–22
	M–r	256	6	300	6	1760	–14	2270	–14	3100	–34
	M	126	0	283	5.5	1633	–13	2140	–17	3314	–26
	F	217	4	300	5	1910	–18	2600	–22	3450	–34
[ɵ]	Gj–n	125	2.5	375	4	1070	–12	2500	–20	3500	–27
	M–r	257	9	370	10	1050	–14	2400	–22	— —	— —
	M	125	–0.5	416	6	1070	–7	2315	–24	3300	–29
	F	216	2	410	7	1175	–11	2700	–31	3600	–35
[ø]	Gj–n	125	2.5	345	5	1735	–10	2250	–12	3400	–22
	M–r	257	8	350	8	1800	–2	2250	–8	— —	— —
	M	126	0	363	6.5	1690	–9	2200	–11	3390	–20
	F	215	2	372	5	2000	–14	2610	–18	3650	–28
[œ]	Gj–n	125	2.5	470	5	1195	–7	2550	–16	3300	–24
	M–r	257	4	500	10	1300	–11	2600	–22	3500	–35
	M	124	–0.5	524	6	1103	–4	2430	–22	3250	–19
	F	217	1	565	8	1290	–6	2730	–21	3700	–29

2.2 Empirical Reference for Standard German

Pätzold and Simpson (1997) conducted a statistical study of vowels of Standard German, produced by men and women (see Table 4, limited to monophthongs). These values are given here because, as mentioned in the introduction, most of the author's experiences and observations to date concern the sounds of the German language, and corresponding references are made in the text as from Part II.

2.3 Other Statistical References

References to other formant statistics and additional data of interest to the present discussion can be found in the Materials section. Such information includes formant statistics for vowels of different languages, model-like formant patterns, formant statistics for whispered vowels and indications concerning formant patterns of vowel sounds at different fundamental frequencies.

Table 4. Formant statistics for Standard German vowels (Pätzold & Simpson, 1997). Average values for formant frequencies F1–F2–F3 (Hz) are given, including additional data on the lower and upper quartiles (lq and uq). Part (a) shows the values of women, part (b) the values of men (see next page).

(a)	Vowel	F1	lq	uq	F2	lq	uq	F3	lq	uq	n
	iː	329	292	385	2316	2125	2496	2796	2644	3000	719
	ɪ	391	350	442	2136	1905	2348	2867	2660	3026	1014
	yː	342	312	401	1667	1485	1833	2585	2437	2691	125
	Y	406	369	466	1612	1475	1735	2631	2518	2779	105
	eː	431	382	495	2241	1949	2472	2871	2691	3055	579
	ɛ	592	517	687	1944	1774	2100	2867	2679	2997	607
	øː	434	391	482	1646	1551	1739	2573	2440	2708	108
	œ	509	452	584	1767	1620	1870	2640	2488	2757	48
	aː	779	665	880	1347	1236	1439	2785	2644	2941	452
	a	751	651	838	1460	1346	1583	2841	2679	2983	810
	oː	438	395	487	953	789	1102	2835	2673	2990	269
	ɔ	573	509	660	1174	1055	1279	2825	2668	2965	279
	uː	350	319	405	1048	885	1220	2760	2624	2877	299
	U	450	387	504	1184	1074	1302	2749	2570	2960	434
	ɐ	590	494	685	1608	1430	1754	2829	2679	2968	610
	ə	420	369	482	1746	1554	1948	2811	2649	2968	1338

(b)	Vowel	F1	lq	uq	F2	lq	uq	F3	lq	uq	n
	iː	290	266	337	1986	1813	2106	2493	2328	2668	710
	ɪ	343	303	380	1803	1640	1956	2483	2309	2632	1009
	yː	310	278	349	1505	1362	1624	2205	2117	2321	126
	Y	374	333	401	1431	1345	1529	2284	2131	2445	102
	eː	372	328	436	1879	1700	2006	2486	2324	2614	580
	ɛ	498	443	552	1639	1517	1755	2451	2299	2599	613
	øː	375	333	414	1458	1383	1505	2220	2104	2319	107
	œ	437	407	501	1504	1376	1598	2179	2121	2327	49
	aː	639	570	700	1225	1166	1292	2477	2316	2613	452
	a	608	529	674	1309	1224	1386	2466	2317	2618	831
	oː	380	352	429	907	774	1009	2415	2269	2570	265
	ɔ	506	455	550	1060	992	1127	2415	2295	2546	283
	uː	309	283	343	961	835	1145	2366	2247	2520	291
	U	382	332	439	1058	966	1165	2363	2225	2522	435
	ɐ	503	440	561	1372	1253	1463	2430	2288	2570	610
	ə	370	321	424	1521	1391	1660	2368	2219	2547	1286

Part II Reflections

Part II presents reflections that, according to the author's reading of the literature, oppose the understanding of the theory, that is, its intellectual re-enactment and validation.

3 Vowels and Number of Formants

3.1 Inconstant Number of Vowel-Specific Relative Spectral Energy Maxima in Sounds of Back Vowels and of /a–ɑ/

As reported in the literature, when analysing samples of sounds of back vowels and of /a–ɑ/, some sounds may exhibit only one distinct vowel-specific spectral envelope peak, whereas other sounds of the same vowels exhibit the expected two pronounced peaks.

Empirically, the number of vowel-specific relative spectral energy maxima proves to be inconstant for sounds of single vowels.

3.2 Inconstant Correspondence between Vowel-Specific Relative Spectral Energy Maxima and Calculated Vowel-Specific Formant Patterns

If sounds of back vowels and of /a–ɑ/ exhibit only a single vowel-specific spectral envelope peak, according to the literature, formant analysis (e.g. using LPC analysis) often reveals two close formant frequencies. Such cases are therefore referred to as formant merging. It follows that, for the sounds in question, the spectral envelope peak and the calculated first two formants do not correspond to one another.

Yet, if sounds of back vowels and of /a–ɑ/ exhibit two vowel-specific spectral envelope peaks, such a correspondence is generally found.

Thus, the observation of an inconstant number of vowel-specific spectral envelope peaks of sounds of one and the same vowel calls into question the fundamental relationship between spectral envelopes and calculated formants.

No direct parallelism exists between relative spectral energy maxima and calculated formants.

Consequently, formants prove to be constructs of a specific method of analysis (see Section 6.1).

3.3 Inconstant Number of Vowel-Specific Relative Spectral Energy Maxima and of Calculated Vowel-Specific Formants

As shown in Part I, with regard to high front vowels and r-coloured front vowels of some languages, sounds belonging to these vowels can exhibit, in part, similar first and second lower spectral envelope peaks and formant analysis can reveal similar F1–F2. Thus, the sounds of the corresponding vowels are physically distinct only with regard to the third spectral envelope peak and the third formant, respectively.

For such languages, it follows that back vowels, as well as some of the front vowels, are physically describable in terms of different patterns of F1–F2, whereas the remaining front vowels have to be described only in terms of different patterns of F1–F2–F3.

> Empirically, the number of vowel-specific relative spectral energy maxima and of calculated vowel-specific formants proves to be inconstant among different vowels.

With regard to spectral envelope peaks, then, the quality of some sounds of back vowels is represented by a single peak, the quality of other sounds of back vowels and sounds of some front vowels by two peaks and the quality of some front vowels by three peaks.

3.4 Addition: “Spurious” Formants

In the spectra of the sounds of certain speakers, an additional spectral envelope peak may occur between the expected first and second or second and third formant. According to the prevailing methodological rules for determining formants, this maximum is not interpreted as vowel specific but as a specific characteristic of the speaker’s voice in question. Therefore, it is referred to as a “spurious” formant.

Such “spurious” spectral envelope peaks also need to be considered within the context of the inconstant number of vowel-specific spectral envelope peaks.

3.5 Addition: “Flat” Vowel Spectra

In the literature, some indications for possible vowel perception related to “flat” spectral parts, lacking any clearly distinctive relative energy maxima, are also given.

3.6 Addition: Inconstant Number of Vowel-Specific Formants in Synthesis

Synthetically produced—and easily recognisable—vowel sounds can be generated for most vowel qualities using three- and two-formant synthesis. For certain vowels, in particular for back vowels and /a–ɑ/, this is also possible by way of a one-formant synthesis.

With regard to synthesised sounds perceived as belonging to one vowel quality, a comparison of the sounds with F1'–F2' (two-formant synthesis) and the sounds with F1'–F2'–F3' (three-formant synthesis) reveals differences for F2', in particular for sounds of front vowels. Similarly, a comparison of the sounds with F1' (one-formant synthesis) and the sounds with F1'–F2' (two-formant synthesis) reveals differences for F1'. (However, in the corresponding comparative studies, the fundamental frequency used in synthesis the was not varied systematically.)

Synthesis thus confirms the inconstant number of observable vowel-specific formants. Further, synthesis involving different numbers of formants (different numbers of filters) indicates differences for F1' or F2', respectively, although the sounds in question are perceived as belonging to the same vowel.

4 Vowels and Fundamental Frequency

4.1 Fundamental Frequency, First Formant and “Grade” of Vowels

According to prevailing theory, vowel-specific formant patterns are independent of the fundamental frequency of their respective individual sounds.

In general, the frequencies of the first formant of all vowels, as specified in current formant statistics for sounds produced in citation-form words, comparable to relaxed speech, lie within the range of the possible fundamental frequencies for the speakers of a given speaker group. Concerning long German vowels, the lowest statistical values for F1 are given for /i, y, u/, medium values for /e, ø, o/, followed by values for /ɛ, ɔ/ and the highest values are indicated for /a–ɑ/.

If the fundamental frequency involved in producing vowel sounds exceeds the frequencies of the first formant of /i, y, u/ and approaches the frequencies of the first formant of /e, ø, o/, then it is to be expected that the vowels /i, y, u/ become unintelligible because their first vowel-specific formant is no longer physically representable. Thus, the vowels /i, y, u/ would be of a “lower grade”, that is, more restricted in their production, physical representation and intelligibility than the other vowels. The same would apply to /e, ø, o/ compared to /ɛ, a, ɑ, ɔ/ and to /ɛ, ɔ/ compared to /a–ɑ/.

In line with prevailing theory, the possibility that the fundamental frequency of a vowel sound can exceed the first formant frequency of a vowel quality as given in formant statistics leads to the assumption that the “grade” of vowels differs because of vowel-specific acoustic characteristics.

However, everyday experience refutes such a generalising conclusion. If speakers of a given speaker group produce vowel sounds, and if the fundamental frequency of these sounds exceeds the frequencies of the statistically given first formant of /i, y, u/ and approaches the frequencies of the first formant of /e, ø, o/, then all of the six vowels mentioned can be produced with the same “grade” of vowel perception, given speakers with correspondingly good vocal abilities. There is no general impairment of vowel perception for the sounds of /i, y, u/ if the fundamental frequency exceeds statistical F1.

The same holds true—although it is less obvious in everyday utterances and only for good voices—for the vowels /e, ø, o/ produced at fundamental frequencies higher than the statistical values of their first formant frequencies.

Speakers with excellent vocal abilities can even produce clearly intelligible cardinal vowels up to a fundamental frequency that corresponds to the highest statistical F1 of all vowels of the language they master.

In this context, special attention needs to be given to everyday speaking styles or habits that exhibit a fundamental frequency variation of one octave or more. Such styles and habits plainly reveal the significance of the problem of fundamental frequencies above statistical first-formant frequencies, confronting the prevailing acoustic theory of the vowel.

Special attention also needs to be given to utterances of stage voices (in musical and straight theatre, entertainment, film, television etc.) because extensive fundamental frequency variation is one of the hallmarks of the singing and speaking voice in the context of art and entertainment.

Generally, with regard to a fundamental frequency range up to the maximum frequency of the first formant as given in formant statistics, no principally different “grades” of vowel perception in relation to fundamental and first formant frequency can be experienced.

4.2 Fundamental Frequency, Spectral Envelope, Formant Pattern and “Grade” of Vowels

If the fundamental frequency of a sound increases, so too does the frequency spacing between the harmonics in the spectrum. As a consequence, determining the spectral envelopes and their maxima becomes difficult. The same applies to the calculation of formant frequencies. According to prevailing theory, it is to be expected that the “grade” of vowel perception is in general also dependent on the fundamental frequency of the sounds: with regard to fundamental frequency, the expected tendency for vowel perception is: the lower, the better; the higher, the worse.

Indeed, considering vowel sounds at higher pitches, many scholars interpret these sounds as related to a spectral undersampling of the formants.

However, one does not only have to consider a general interrelation between fundamental frequency, harmonic spectrum, spectral envelope and expected formant frequencies, but also a formant-specific role within this interrelation: depending upon given statistical frequency values of vowel-specific formants, comparisons show that sounds at higher fundamental frequencies may in some cases exhibit frequencies and relative amplitude maxima of harmonics that correspond to the statistical formant frequencies for the vowels in question, whereas the frequencies of the harmonics of sounds at lower fundamental frequencies lie in between these formant frequencies. For the latter, the formants are subsequently expected to appear as envelope peaks either only indistinctly or not at all, and the corresponding vowel perception is expected to be impaired when compared to sounds at higher fundamental frequencies for which the frequencies of the harmonics match statistical vowel-specific formant frequencies.

Such reasoning leads to the assumption that there is not only a general but also a discontinuous relationship between the intelligibility of vowel sounds and their fundamental frequency: accordingly, vowel sounds at lower fundamental frequencies would, as a rule, be more intelligible than vowel sounds at higher frequencies, but vowel intelligibility would also depend upon the respective relationships between fundamental frequency, harmonic spectrum and vowel-specific formant patterns (as given in formant statistics).

In line with prevailing theory, the relationship between fundamental frequency, harmonic spectrum, spectral envelope and expected vowel-specific formant pattern leads to the same assumption that the "grade" of vowels differs in relation to vowel-specific acoustic characteristics.

However, as explained, everyday experience refutes such a generalised conclusion. Thus, a theory of vowels as elements of language that formulates an inherently qualitative and at the same time discontinuous relationship between fundamental frequency and vowel perception stands in contrast with the—possibly "sensational"—characteristic of a voiced element of language being independent of pitch within the range of intelligible speech.

5 Formant Patterns and Speaker Groups

5.1 Fundamental Frequency, Spectral Envelope, Formant Pattern and “Grade” of Vowels Uttered by Children, Women and Men

If one further extends the reasoning developed in the previous chapter, namely that—according to prevailing theory—the intelligibility of a vowel sound is expected to relate to the respective fundamental frequency of the sound and the (statistically given) first formant frequency of the vowel, then, correspondingly, the “grade” of vowel perception should also depend upon the speaker group: vowel intelligibility should prove to be best for men, average for women and worst for children.

According to prevailing theory, the above relationship between fundamental and first formant frequencies, spectral characteristics and expected differences in the “grade” of intelligibility of different vowel qualities leads to the assumption that the “grade” of vowels varies for different speaker groups (children, women, or men).

Everyday experience also refutes this generalisation. Thus, again, a theory of vowels as elements of language that formulates a inherently qualitative relationship between age and gender on the one hand, and vowel perception on the other, stands in contrast with the—possibly (yet again!) “sensational”—characteristic of a voiced element of language being quasi-independent of a speaker’s constitution (if not impaired).

Vowels as such are related neither to age nor to gender. If direct comparisons of utterances of single speakers show that some speakers produce vowel sounds “better” (better in vowel intelligibility) than others, then, this has to do with the vocal abilities of the individual speakers investigated, not with vowels, speaker groups, or vocal-tract sizes (with the exception of very young children acquiring their first language). As a rule, vowels, as speech sounds of a given language, can potentially be produced with equal intelligibility by speakers of all general speaker groups. Vowels are not attributes of an individual, but elements of language. Vowels are “abstracted” from the individual.

5.2 One Vowel, Different Formant Patterns

In the literature, empirical reference values for vowel-specific formant patterns are given separately for each speaker group (children, women, or men), that is, in group-specific terms (see, for example, Chapter 2). In the first instance, these differences in formant patterns are not explained in terms of varying average fundamental frequencies, but in terms of varying average vocal-tract size.

This view leads to the assumption that each vowel is physically represented by three different speaker group-specific formant patterns, not only in terms of the different fundamental frequencies, but also in terms of the same fundamental frequency: in general, women and men are able to produce clearly recognisable vowel sounds at a child's fundamental frequency—for instance, at around 250 Hz (see Section 2.1; note, in this context, that in the statistics of Hillenbrand et al., F0 differences between women and children do not exceed 20 Hz). Given such cases of sounds at similar fundamental frequencies, three sounds of the same vowel, produced by a man, a woman and a child respectively, are expected to exhibit three substantially different formant patterns, despite the similarity in vowel perception.

According to prevailing theory, the relationship between vowel-specific formant patterns and age- and gender-related speaker groups leads to the assumption that the physical representation of a vowel is based upon different formant patterns.

Such reasoning also leads to the assumption that women and men are capable of producing sounds of a given vowel with fundamental frequencies substantially higher than those of children, albeit with substantially lower corresponding formant patterns.

The problem that the particular sound configurations in question pose to the theoretical approach discussed here becomes particularly evident when considering corresponding sounds of the vowels /a, ɑ, ɔ, o, u/, which are low-pass filtered with a cut-off frequency of 2 kHz (note that, for these vowels, statistical values of vowel-specific formant patterns F1–F2 for all three speaker groups discussed here are given as ≤ 2 kHz): then, neither different fundamental frequencies nor different higher spectral energy configurations can play a role in vowel perception and can explain why three different patterns of F1–F2 can be expected to represent the same vowel.

It goes without saying that the above also holds true for the restricted comparison between women and men.

The problem described here becomes particularly acute if, instead of natural vocalisations, corresponding sound configurations are studied by means of vowel synthesis, applying similar fundamental frequencies but different patterns F1'–F2'.

However, in its turn, such a conclusion runs counter the requirement of a psychophysical parallel between perceived vowel quality and physical representation: formant patterns are either vowel specific, which means that clearly distinct formant patterns do not represent the same vowel—regardless of the fundamental frequency—or they are, as such, not directly vowel specific. According to the first stance, the assumption of speaker group-specific formant patterns would have to be questioned. According to the second stance, the assumption of vowel-specific formant patterns in general would have to be questioned.

5.3 Different Vowels, One Formant Pattern

Disregarding the comment in the previous paragraph, the pursuit of the reasoning developed in Section 5.2 leads to the further assumption that a single formant pattern can represent two different vowels: given that the sounds of a vowel produced by a speaker of one speaker group exhibit higher vowel-specific formant frequencies than the sounds of the same vowel produced by a speaker of another speaker group, and that the fundamental frequency plays no substantial role in the physical representation of the vowel in terms of formant patterns, and also given that the vowel-specific formant frequencies of the sounds of the first speaker lie within the frequency range of the possible vowel-specific formant frequencies of the second speaker, then it must be possible to find cases of comparisons of two sounds, each produced by one of these two speakers, that exhibit similar vowel-specific formant patterns, yet are perceived as different vowels.

According to prevailing theory, the relationship between vowel-specific formant patterns and age- and gender-related speaker groups leads to the assumption that a single formant pattern can physically represent two different vowels.

Again, the problem that such sound configurations pose to the theoretical approach discussed here becomes particularly evident when considering corresponding sounds of the vowels /a, ɑ, ɔ, o, u/, because the vowel-specific formant frequencies of the corresponding sounds of all speaker groups are given in formant statistics ≤ 2 kHz, and in such a frequency range, adults can reproduce sounds exhibiting any of the F1–F2 pattern found in sounds of children. The same holds true when comparing the sounds of men and women.

The problem described here becomes particularly acute again if replicated by means of vowel synthesis, above all including extensive variation of the fundamental frequency.

However, in line with the explanation given above, the assumption of a possibility of twofold representation, according to which a single formant pattern can correspond physically to the sounds of two different vowels, runs counter to the requirement of a psychophysical parallel between perceived vowel quality and physical representation. At the same time, indeed, it directly contradicts prevailing theory.

This consideration engenders a decided scepticism about the claim that vowel-specific formant patterns are both fundamentally and continuously dependent upon the speaker group, that is, upon vocal-tract size. A fundamental dependence is already difficult to understand from an intellectual standpoint because, as mentioned, vowels do not "have" an age or gender. Besides, the simple fact that sounds of back vowels can be synthesised at fundamental frequencies, observable in sounds of children as well as in sounds of men, paradigmatically illustrates the problem: if, in synthesis, F1–F2 is changed substantially but the fundamental frequency is held constant, in general, the perceived vowel quality also changes, irrespective of whether the F1–F2 of the synthesis corresponds to a pattern observed for natural sounds of a child or of a man.

At the same time, the above reflection suggests an alternative explanation for the existing empirical findings, which seemingly provide evidence for speaker group-specific formant patterns: vowel-specific spectral energy configuration, and with this this calculated formant patterns, can depend upon fundamental frequency.

It is remarkable that, in general, formant statistics deemed worthy of reference in the literature do not give frequency values of formant patterns of the different speaker groups for systemically varied fundamental frequencies. Thus, currently, there is no empirical evidence in the literature to support the claim that observed, speaker group-specific formant patterns of vowels should in principle not be attributed to the different—and simultaneously observed—fundamental frequencies of the respective sounds but, instead, to different average vocal-tract sizes. With regard to the first formant for all vowels, and probably also to the second formant for back vowels, the present reflection indicates that such evidence cannot be furnished.

5.4 A Gap in the Reasoning

As indicated, existing formant statistics suggest that, irrespective of fundamental frequency and perceived vowel quality, adults are capable of producing sounds for almost all variants of F1–F2 patterns as found in children's vowels. Thus, even though adults have larger vocal tracts than children, for most vowels, they are nevertheless capable of producing sounds that exhibit the same vowel-specific formant patterns, above all F1–F2, as evidenced for the sounds of children.

If it is indeed the case that speakers of all three speaker groups are considered to be capable of producing the same vowel-specific patterns for a substantial part of vowels, then how are the pattern differences discussed above to be understood? (Many scholars assume that the schwa sound defines the midpoint of a speaker's vowel space and plays a central role for the formant pattern differences discussed: because of different average vocal tract lengths and different resonance patterns of related open tubes of speakers of different age and gender, it is deduced that different vowel-related format patterns mirror different midpoint reference patterns. However, in the present context, such an assumption does not dispense from the question posed: sounds of schwa, too, can be produced on different fundamental frequencies, and the independence or dependence of related formant patterns on fundamental frequency for perceptually unaltered schwa quality has not yet been clarified.)

Even though existing statistical values list vowel-specific formant patterns for children exceeding those for adults, and for women exceeding those for men, there are nevertheless exceptions: in some cases, as shown by some statistics, single vowel-specific formant frequencies, or even vowel-specific formant patterns F1–F2 or F1–F2–F3, for sounds produced by men do not differ from those for sounds produced by women; they may even slightly exceed the latter. (Thus, remarkably, the formant patterns given by Fant, 1959, for a single male and a single female speaker do not show a consistent speaker group related difference; see Section 2.1, Table 3. Besides, there are cases in which the statistical F1 of women slightly exceeds the F1 of children, see, for instance, Section 2.1, Table 2, and the corresponding values for the vowel /ʌ/.) This raises the same question as above.

The relationship between vowel-specific formant patterns and age- and gender-related speaker groups described in terms of prevailing theory fails to explain why, despite different vocal-tract sizes, similar vowel-specific formant patterns are basically possible at least for the majority of vowels but are—according to theory—not realised (actually not produced).

In addition, this formulation could also prove to be generally applicable: it could prove to be the case that all vowel-specific formant patterns, F1–F2 and F1–F2–F3 as given in formant statistics for children, can also be produced by women and men. (With regard to this aspect, utterances of voice-over artists are of particular interest.)

Repeating and insisting: given a psychophysical perspective, the correspondence between intelligible vowel sounds and the vowel-related physical characteristics must be formulated as such. The formulation of speaker-independent and, in a strict and direct sense, vowel-specific acoustic features represents the touchstone for any acoustic theory of the vowel.

5.5 Addition: Formant Patterns of Voiced and Whispered Vowel Sounds

Empirical studies comparing voiced and whispered vowel sounds indicate substantial differences in the formant patterns related to the perceived vowel qualities. In particular, the first formant frequency of whispered sounds of a given vowel (and, according to some studies, the second formant frequency, too) are found on significantly higher frequency levels than those of voiced sounds. (As mentioned in Section 1.4, such differences are explained as a consequence of differences in the geometry, and thus the resonances, of the glottal area of the vocal tract for the two different phonation types in question.)

This finding relativises again the attempt to establish a direct correspondence between vowels and formant patterns: the sounds of the same vowel can exhibit different formant patterns, not only because of different average vocal-tract sizes but also because of different kinds of phonation acting upon a configuration of a single vocal tract.

Moreover, comparisons between published formant frequencies of whispered and voiced vowel sounds indicate that all F1, and the majority of F2 ≤ 1.5 kHz, of whispered sounds produced by men generally exceed the corresponding F1 and F2 of voiced sounds produced by women, given the same perceived respective vowel identities and notwithstanding men's larger vocal tract. The same applies to a comparison between whispered sounds of women and voiced sounds of children. Restricted to F1, this also applies to the comparison between whispered sounds of men and voiced sounds of children.

This observation relativises in turn the assumption of a correspondence between vocal-tract size and vowel-specific formant patterns: based on the values given in the literature, such a correspondence is documented only for sounds of one and the same phonation type, not for a comparison of sounds of different phonation types. Besides, it should be noted that the frequency differences of the lower formants for the sounds of a given vowel, which relate to different types of phonation, e.g. voiced versus whispered sounds, are in general greater than the corresponding formant frequency differences between the different speaker groups.

Thus, most importantly, vowel-related formant patterns produced by one vocal tract can differ more than vowel-related formant patterns produced by different vocal tracts with very different tract sizes.

Moreover, referring to Section 5.3, a single formant pattern seems able to physically represent different vowels not only if the corresponding sounds are produced by speakers belonging to different speaker groups, but also if an individual speaker varies his or her phonation.

Such consideration will be discussed further in Part III: comparisons between the formant patterns of voiced and whispered sounds, as documented in the literature, refer only to the average (lower) fundamental frequency of voiced vowel sounds produced in citation-form words, but not to a comparison including a systematic variation in fundamental frequency of voiced sounds. (Such an experimental arrangement assumes, once again, that formant patterns are independent of fundamental frequency and are, therefore, negligible when comparing voiced and whispered sounds.)

6 Terms of Reference, Methods of Formant Estimation

6.1 Formant and Sound Spectrum

Given that the terms "resonance" and "formant" are distinguished from each other, as a means of distinguishing the characteristics of the vocal tract from those of the sound spectrum, then the psychophysical question of the vowel relates to formants only. According to prevailing theory, it is assumed that, in the first instance, the spectrum of a vowel sound exhibits determinable relative energy maxima, which are related to vowel-specific frequency ranges, and that, as a rule, the frequencies of these relative spectral energy maxima correspond to calculated formant frequencies, for example, applying LPC analysis. (Note that, nowadays, formant frequencies are no longer derived as numerical values from the spectral envelope but, instead, are calculated as filters of an analytical model, although the corresponding numerical results are in many cases crosschecked on the basis of a spectrogram.)

As discussed in Sections 3.1 and 3.2, the sound spectra of back vowels and of /a–ɑ/ can exhibit only one single vowel-specific spectral energy maximum, although formant analysis using an analytical model (e.g. LPC analysis)—under involvement of "phonetic knowledge" and sometimes with interactive manual adjustment of parameter settings—indicates two vowel specific formants, often close in frequency. This contradicts the assumption that the number and frequency of relative spectral energy maxima, that is the envelope peaks, always correspond to analytically determined formants.

As mentioned in Section 4.2, due to the increasing frequency spacing of the harmonics, the higher the fundamental frequency, the more difficult it becomes to determine the spectral envelope and its peaks (for further details, see also Section 6.4). This in turn impedes the formulation of a general correspondence between relative spectral energy maxima and calculated formant frequencies.

Regarding the current procedures used in formant analysis and the corresponding numerical values of formant patterns, it follows that in many cases—and thus in principle—the term formant often does not designate a characteristic of the sound spectrum itself, but instead a construct or even artefact of the respective method of analysis.

In the current literature, the term formant—if distinguished from resonance—generally refers neither to any actual characteristic of the vocal tract nor to any actual characteristic of the sound spectrum. The term generally refers to filters of an analytical model. At the same time, formants are not determined on the basis of spectra but on the basis of such an analytical model.

Thus, the assumption that a direct correspondence exists between resonances as a physical property of the vocal tract, spectral energy maxima as a physical characteristic of the vowel sound produced and filter frequencies derived from methods used in the acoustic analysis of vocal sounds, loses its plausibility.

6.2 Speaker Group and Vocal-Tract Size

As discussed, prevailing theory supposes a relationship between vowel-specific formant patterns and age- and gender-related speaker groups and explains corresponding differences in terms of the respective average vocal-tract sizes.

It can be assumed that some women have larger vocal tracts than some men. Comparing the vowel sounds of these female and male speakers, the following constellation is of particular interest in the present context: the sounds of the female speakers in question exhibit fundamental frequencies corresponding to the average fundamental frequency values for women in general, as given in formant statistics, and the sounds of the male speakers in question exhibit substantially lower fundamental frequencies. Then, according to prevailing theory, the vowel-specific formants of these female voices would have to exhibit lower frequencies—despite comparatively higher fundamental frequencies—than the corresponding formant patterns of these male voices.

Extending such consideration, this comparison raises the question of a systematic investigation of the relationship between vocal-tract size and vowel-specific formant patterns within a single speaker group.

Besides the lack of an empirical basis for the questions raised here, the above reflections again point to the fact that prevailing theory does not claim that vowel-specific formant patterns depend in principle on age and gender, but that different vowel-specific formant patterns exist for different vocal-tract sizes: prevailing theory only refers to speaker group-specific differences in average vocal-tract sizes.)

The term “age- and gender-related speaker group” is related to the term “age- and gender-related average vocal-tract size”.

6.3 Formant Analysis and Objectivisation

Concerning natural vocalisations, current analytical methods for determining formants apply a model-like procedure in order to calculate a specific configuration of source sound and filters which, by means of transformation of source by filters, “reproduces” a sound that best corresponds to the real sound. (The same applies to whispered vowel sounds, in relation to the source as noise.)

Such a procedure must not only assume certain characteristics of the source sound but also a certain number and certain characteristics of the filters involved in the frequency range under investigation. (Note that, according to prevailing theory, different numbers of formants are expected for a given frequency range in relation to different speaker groups because of their different average vocal-tract size. Thus, the number of filters for the analysis of a sound must be set accordingly.) How closely the characteristics of the source sound approach actual phonation remains open. The same applies to the question of whether the number of filters and their characteristics actually correspond to real articulation and its resonance.

Thus, formants cannot be determined reliably on the basis of a vowel sound alone. Analysis requires at least some prior knowledge of whether the sound under investigation has been produced by a man, woman, or child, assuming that this information is sufficient to deduce the number of filters (related to the frequency range of interest) to be used in formant analysis.

Besides, subsequent automatically calculated formant frequency values are often double-checked visually on the basis of the sound spectrogram: if the values calculated in the first step—based on analytical parameters according to existing standards and known speaker group—do not correspond to the relative spectral energy maxima of the analysed sound, then the number of filters is varied and analysis is performed until such a correspondence occurs. As a rule, the characteristic of the source sound is not altered. However, this only applies to cases where such an interactive analysis is able to produce vowel-specific numbers and frequencies of formants that correspond to the number and frequency ranges to be expected according to prevailing theory and established statistical patterns, and which are also clearly indicated in the spectrogram. If an interactive procedure of analysis yields no values with such a correspondence, then the respective vowel sounds are often excluded from further studies, irrespective of vowel perception. Exceptions include so-called “formant merging”, as discussed in Section 3.2.

Thus, current methods of formant analysis presuppose that researchers have the necessary analytical skills, that is, a knowledge of the existing phonetic principles and rules of interpretation as well as extensive first-hand experience of conducting such an analysis. This involves prior training because such an analysis involves contextual knowledge, the ability to visually compare numerical values with a corresponding sound spectrogram, together with the ability to interpret the latter visually, and also the skills to vary filter settings interactively and to perform the repetition of numerical analysis. Consequently, methods of formant analysis are not completely objectifiable. If they were, then researchers would play no part as individuals in such research.

Strictly speaking, methods of formant analysis are not fully objectifiable; accordingly, they cannot be fully automated.

Most importantly, these procedures are also very time consuming. Therefore, investigations based on very extensive samples of sounds are problematic with regard to method. This is the case particularly if the fundamental frequency is varied: then, specific problems of analysis aggravate the costly character of the method as such. Obviously, this holds true for all repetitions and verifications of existing investigations.

6.4 Formant Analysis, Fundamental Frequency and Speaker Group or Vocal-Tract Size

In addition to formant analysis not being fully objective and automated, it also depends on the respective fundamental frequencies of the sounds. To repeat: the higher the fundamental frequency, the more difficult it becomes to determine the spectral envelope peaks expected because the frequency spacing between the harmonics become too large to accurately define the spectral envelope. It also becomes increasingly difficult to determine the formants within any of the existing analytical frameworks.

With regard to critical limits of fundamental frequencies, above which methods of formant analysis become unreliable, two kinds of reference values need to be considered: firstly, half the frequency of the lowest first formant for a speaker group in terms of an average vocal-tract size, and secondly, the frequency of the lowest formant for a speaker group.

For a fundamental frequency above half of the first formant frequency (F0 > ½F1), the frequency spacing between the harmonics is already so extended that defining a spectral envelope and evaluating the calculated numerical formant frequencies becomes problematic. (Note that for such sounds, the formants may not be clearly indicated by at least two harmonics.) According to this first kind of limit, and referring to the standard values established by Hillenbrand et al. (1995) for F1 of /i/ (the lowest average value for F1 in these reference statistics), formant analysis becomes critical for fundamental frequencies higher than:

- 226 Hz for sounds of children (involving short vocal tracts)
- 219 Hz for sounds of women (involving medium-sized vocal tracts)
- 171 Hz for sounds of men (involving long vocal tracts)

For a fundamental frequency above the lowest first (statistically given) formant frequency for a given speaker group, under the assumption of independence of formants from fundamental frequency, it is basically impossible to distinguish all F1 of all vowels produced by speakers of that group, not to mention the aggravated problem of determining the spectral envelope. According to this second kind of limit, and again referring to the above statistics, methods of formant analysis lack a methodological basis for fundamental frequencies higher than:

- 452 Hz for sounds of children (involving short vocal tracts)
- 437 Hz for sounds of women (involving medium-sized vocal tracts)
- 342 Hz for sounds of men (involving long vocal tracts)

Note that referring to the statistics of Pätzold and Simpson (1997) for German vowels, shown in Section 2.2, the limits would have to be set even on lower frequencies: ½F1 of /i/ corresponds to 165 Hz for women (medium-sized vocal tracts) and to 145 Hz for men (long vocal tracts), respectively; F1 of /i/ corresponds to 329 Hz for women and to 290 Hz for men or long vocal tracts, respectively.

In this context, attention should also be given to the fact that, according to several formant statistics, the frequency distance between F1 and F2 for sounds of some back vowels is given < 500 Hz. Thus, the frequency spacing of the first two harmonics in a spectrum of a sound on a fundamental frequency above this frequency limit exceeds the F1–F2 distance mentioned, which renders formant estimation obsolete within the existing theoretical framework.

The first lists of frequency limits given above for $F0 > ½F1$ suggests that methodologically speaking the analysis of vowel sounds of children and women must be considered problematic in general. The critical fundamental frequency value mentioned for children is considerably lower than the empirically determined average fundamental frequency that children exhibit when producing vowels in citation-form words, which can be considered as related to relaxed speech on a comparatively low fundamental frequency (see, for example, the statistics in Section 2.1). Thus, most vowel sounds produced by children in their everyday expression, exhibit substantially higher fundamental frequencies.—According to Hillenbrand et al. (1995), the mentioned critical fundamental frequency value for women corresponds to the average fundamental frequency of women producing vowels in citation-form words. In everyday speech, however, vowel sounds in a fundamental frequency range of up to one octave higher than this value are the norm. Moreover, according to Pätzold and Simpson (1997), the mentioned critical fundamental frequency value for women is again considerably lower than the average fundamental frequency generally given in vowel statistics.—The problem discussed here seems to be less pronounced among men than among women and children, but it nevertheless concerns a substantial part of their utterances.

The second list of frequency limits reveals that, for methodological reasons, any determination of formant patterns of vowel sounds exhibiting fundamental frequencies that exceed low first-formant frequencies does not make sense, since general rules for formant estimation can no longer be formulated. In this regard, particular consideration needs to be given to voices exhibiting extensive prosodic variations in fundamental frequency, which can be experienced in everyday speech and, very pronounced, in the field of art and entertainment. (Noticeable, with regard to everyday speech, the literature does not provide ample documentation of the occurrence and significance of such extensive variation in fundamental frequency, allowing for a validation of the significance of the methodological problem of formant estimation discussed here. However, in the Materials section, examples of corresponding utterances are documented; see Section M8.2.)

Within the prevailing theoretical framework, the reliability of formant analysis depends on fundamental frequency and the age- and gender-related speaker group, that is, vocal-tract size. Depending on the latter, formant frequency estimation becomes critical for fundamental frequencies above c. 175 Hz, and formant frequency estimation can no longer be methodologically substantiated for fundamental frequencies substantially above 350 Hz. Consequently, formant analysis cannot be applied to all cases of clearly intelligible vowel sounds.

A part of the literature tends to equate the methodological problem with a particular characteristic of vowel perception, which leads us back to the two assumptions discussed in Sections 4.1 and 5.1: firstly, that vowels produced by children and women are basically less intelligible than those produced by men; and secondly, that at least some vowels of sounds at a fundamental frequency substantially above 350 Hz can no longer be clearly distinguished. As suggested, however, both assumptions contradict actual vowel perception.

6.5 Addition: Parameter Adjustments in Formant Analysis and Inconsistent References to Vocal-Tract Size

On the one hand, formant parameters in current procedures of formant analysis are defined prior to analysis of the sounds depending on the corresponding speaker group, that is, the assumed average vocal-tract size of the speakers. On the other hand, these parameter settings are sometimes interactively altered during the procedure if the calculated numerical values do not yield the expected number of formants in the expected vowel-specific frequency ranges compared to the respective spectrogram.

Thus, for example, with regard to sounds of a single speaker, LPC analysis involving standard parameters according to the related speaker group (average vocal-tract size) may yield the expected values for only a part of the sounds, whereas the analysis of other sounds may require the parameters to be set to the standard of another speaker group (average vocal-tract size) or to a setting that is entirely different from any speaker-group related standard given in the literature.

This reveals an inconsistency in how parameter settings are established: in the first instance, default settings of analytical parameters are related to specific vocal-tract sizes, whereas any corrections of these settings are related to the respective general (not vocal tract related) degree of “formant resolution” of the analysis.

6.6 Addition: Spectrum, Formant Pattern, Resynthesis

As explained in Section 6.1, current methods of analysis yield no consistent and direct relationship between spectrum, spectral envelope and formant frequencies. Consequently, this raises the question of the existence of a general relationship between a natural vowel sound, the determined formant pattern and resynthesis.

Currently, resynthesis is indeed being used to examine the reliability of calculated formant patterns. However, this kind of verification is unable to substantially relativise the general problems of the existing methods of analysis described above: resynthesis is feasible only if formant analysis is not fundamentally at issue and only with regard to a limited variation of analytical parameters.

Moreover, the question of resynthesis must be discussed against the background of synthesised sounds as discussed in Section 3.1, indicating the possibility of substantial differences in formant patterns of sounds of one vowel: if a certain analytically determined formant pattern used in a resynthesis reveals an “expected” vowel identity in a perceptual test, then this does not mean that another determined formant pattern, based on a different parameter setting, and applied in a second resynthesis, in principle cannot reveal the same vowel identity. Further, the possibility cannot be excluded that there are cases of sounds for which, with regard to the perceived vowel quality, based on “unexpected” formant patterns may produce a better approximation to the quality of the natural sounds in question than based on “expected” formant patterns.

6.7 Addition: Formant Analysis and Objectivity with Regard to Synthesised Vowel Sounds

It is noteworthy that, if a sound is synthesised using a specific pattern of filters and filter bandwidths, the formant pattern of a subsequent analysis may differ from the synthesis filters if the number of filters used is not communicated to the scholar conducting the analysis.

Moreover, the problem of possible differences of filters used in synthesis and formant patterns obtained in analysis will be substantially enhanced if the fundamental frequency is varied independent of the filters.

6.8 Addition: Formant Patterns and Resynthesis outside of the Framework of Prevailing Theory

It is also noteworthy that, if formant patterns are calculated outside the framework of prevailing theory, for example, using LPC analysis as a method to decompose any sound into a source and a set of filters, irrespective of the fundamental frequency and the perceptual quality and not relating the decomposition to existing formant or resonance statistics (and therefore not considering a direct relationship between spectral peaks and resonances of the vocal tract), and if the results of analysis are used in resynthesis, for many examples of natural utterances, resynthesis reproduces similar intelligible vowel qualities, even for very high fundamental frequencies. Obviously, then, formant patterns will sometimes deviate strongly from the statistical patterns given in the literature.

Part III Experiences and Observations

The third part of the main text formulates several hypotheses about the actual relationship between vowel sounds, sound spectra and formant patterns. These hypotheses refer to the recordings mentioned in the first part of the introduction and to related analyses and observations.

7 Unsystematic Correspondence between Vowels, Patterns of Relative Spectral Energy Maxima and Formant Patterns

7.1 Inconstant Number of Vowel-Specific Relative Spectral Energy Maxima and Incongruence of Vowel-Specific Formant Patterns

As discussed in Section 3.1, sounds of back vowels and of /a–ɑ/ can exhibit only one relative spectral energy maximum within their vowel-specific frequency range ≤ 1.5 kHz (≤ 2 kHz for some sounds of /a/), in contrast to other sounds of the same vowels, which have two such maxima. Consequently, the number of vowel-specific energy maxima is inconstant.

The spectral envelopes and formant patterns of such vowel sounds cannot in all cases be interpreted as "formant merging": examples of sound pairs of back vowels can be observed for which both sounds exhibit the lowest spectral envelope peak at a similar frequency level, but only one of them has a pronounced second envelope peak within the frequency range mentioned. Then, the first spectral envelope peak of both sounds corresponds to the vowel quality in question, whereas the second spectral envelope peak may be linked to an additional "colouring" of that sound. However, it plays a marginal role in vowel perception and, in such a case, does not posses vowel-differentiating value.

For both sounds of such sound pairs, formant analyses using current methods may reveal two lower formants. However, calculating F2 for the first sound of the respective sound pair mentioned, exhibiting only one lower spectral envelope peak, may prove highly contingent on the number of filters chosen, above all for sounds of children. In addition, its amplitude can be very low and its bandwidth can be very large, that is, far beyond reference values as given in the literature.

With regard to front vowels, the frequency of observable second envelope peaks, and with them also calculated F2, can vary strongly. Because of this, there are examples of sound pairs of front vowels for which the second envelope peak and calculated F2 of one sound approaches or even exceeds the third envelope peak and calculated F3 of the other sound. (Such observations in general relate to sounds of speakers of different speaker groups, which are produced at similar fundamental frequencies. However, this can also be observed for the sounds of speakers of the same speaker group.)

Thus, it is not possible to designate a standard number of consecutive relative spectral energy maxima related to delimited frequency ranges that represent any given vowel. The same holds true for formants, although it is less obvious. There are also formant patterns of sounds of single vowels whose reciprocal correspondence of single formants is open to discussion.

The number of vowel-specific relative spectral energy maxima is inconstant, and formant patterns are incongruent in some cases.

7.2 Partial Lack of Manifestation of Vowel-Specific Relative Spectral Energy Maxima

In their vowel-specific range of the spectrum ≤ 1.5 kHz, sounds of back vowels and of /a–ɑ/ produced at fundamental frequencies ≤ 350 Hz can exhibit series of harmonics with consistent, quasi-identical amplitudes. These vowel-specific parts of harmonic spectra seem to be “flat”, lacking any clearly distinctive relative energy maxima. Of special interest in this respect are the sounds of /a, ɑ, ɔ, o/ in cases where the amplitudes of the first three to five harmonics are not markedly different.

In their vowel-specific range of the spectrum ≥ 1.5 kHz, sounds of front vowels produced at fundamental frequencies ≤ 350 Hz can also exhibit series of harmonics with consistent, quasi-identical amplitudes. Thus, what applies to back vowels and to /a–ɑ/ for their entire vowel-specific frequency range also applies to front vowels for the higher part of their vowel-specific frequency range.

In addition, cases of such vowel-specific, “flat” spectral portions also exist for sounds produced at fundamental frequencies > 350 kHz, even if, in relation to the large frequency spacing of the harmonics, this generally remains limited to the sounds of the vowels / i, e, ɛ, a, ɑ /. For certain fundamental frequencies of the sounds of /ɔ, o/, the first two harmonics can exhibit equal amplitudes.

Also worth mentioning in this context are the sounds of back vowels and of /a–ɑ/, which exhibit continuously decreasing amplitudes in the vowel-specific lower frequency range. In the spectra of these sounds, the first harmonic generally forms the actual spectral maximum.

Thus, the set of problems concerning a formulation of a general relationship between the perceived vowel quality and its physical representation based on a certain number of relative spectral energy maxima is again extended.

Spectral envelope maxima, as described in the literature, are not a precondition for the physical representation of vowels.

The relationship between “flat”, vowel-specific parts of sound spectra and calculated formant frequencies using current methods of analysis cannot be described in simple and general terms. The same holds true for the relationship between continuously decreasing amplitudes of the harmonics in the vowel-specific lower frequency range and calculated formant patterns. Therefore, the issue is left open to discussion here. However, it has to be considered as an additional methodological problem of formant analysis.

7.3 Addition: Resynthesis and Synthesis

Inconstancy in the number of vowel-specific relative spectral energy maxima, possible incongruence of formant patterns and vowel sounds with “flat” or decreasing vowel-specific spectrum portions can be replicated using resynthesis.

The same also applies to formant patterns or harmonic spectra not derived directly from natural vowel sounds.

8 Lack of Correspondence between Vowels and Patterns of Relative Spectral Energy Maxima or Formant Patterns

8.1 Dependence of Vowel-Specific, Relative Spectral Energy Maxima and Lower Formants ≤ 1.5 kHz on Fundamental Frequency

If investigated empirically and systematically, it becomes evident that the first spectral envelope peak—if it exists—and the first calculated formant of vowel sounds often depend on fundamental frequency.

For a range of fundamental frequencies ≤ 350 Hz for which formant analysis is not critical in principle, this dependence is particularly evident in the sounds of the vowels /e, ø, o/ at fundamental frequencies in the range of 200 Hz to 350 Hz.

For a range of fundamental frequencies > 350 Hz, this dependence is, above all, indicated in sounds of the vowels /i, y, u/, because the first harmonic generally exhibits the highest amplitude; thus, the lowest spectral peak rises with increasing fundamental frequency.

In addition, such a dependence can also be observed for the second formant for cases of sounds of back vowels.

For sounds of /ɛ/ and of /a–ɑ/, however, indications of a dependence of F1 on fundamental frequency may prove to be weak and corresponding observations may require a comparison of sounds with a very extended vocal range.

Moreover, the observation of a dependence of F1 on fundamental frequency is not only related to frequency ranges of the latter and vowel qualities but also to single speakers and their phonation characteristics, including vocal effort. (Note that marked differences in the vocal effort of vowel production have a substantial effect on spectral peaks and calculated formant frequencies, and this effect has to be taken into account when investigating the relationship between F0, spectral peaks and formants.) But although the indications for the dependence discussed here prove to be unsystematic, the findings of intelligible vowel sounds at fundamental frequencies > 500 Hz (see next chapter) and of formant pattern ambiguity (see Chapter 9) force us to relate the lower spectral peaks and the lower formants to fundamental frequency.

The possible relationship between fundamental frequency and higher vowel-specific spectral envelope peaks or formants > 1.5 kHz for sounds of front vowels is left open here for discussion.

These assertions hold true for vowel sounds produced by one and the same speaker. Thus, they apply to vowels and their physical representation.

In this respect, what is of particular importance is the observation that the dependence of lower spectral envelope peaks and lower formants ≤ 1.5 kHz does not represent a phenomenon generally related to "oversinging" the first formant of a vowel: most importantly, the shifts of F1 in the sounds of the vowels /e, ø, o/ can already be observed at fundamental frequencies substantially below the corresponding statistical values for F1 as given in the literature for sounds produced in citation-form words. Moreover, given a range of fundamental frequencies of c. 200–350 Hz, the shifts of F1 for the sounds of the vowels /e, ø, o/ are in many cases much more pronounced than for the sounds of the vowels /i, y, u/, although, for the former, the literature gives significantly higher statistical values for F1 than for the latter.

Also of particular importance—and foreshadowing formant pattern ambiguity of vowel sounds (see Chapter 9)—is the observation that, in many cases of sounds of a vowel produced by a single speaker, the shifts of F1 in relation to fundamental frequency exceed the F1 differences of two neighbouring vowels as given in formant statistics for a corresponding speaker group (for speakers with corresponding vocal-tract size). In line with this, the shifts mentioned also exceed speaker-group differences in F1 for that same vowel as given in the format statistics mentioned.

Vowel-specific relative spectral energy maxima ≤ 1.5 kHz (if determinable) and calculated vowel-specific formant patterns (if methodologically substantiated) are dependent on fundamental frequency.

8.2 Vowel Perception at Fundamental Frequencies above Statistical Values of the First-Formant Frequency

Speakers possessing a large vocal range and good phonation and articulation are able to form the sounds of the vowels /i, y, e, ø, ɛ, a, o, u/ in a recognisable and distinguishable way up to a fundamental frequency of c. 700–800 Hz. Such sounds can be readily experienced up to a fundamental frequency of c. 600 Hz because they occur frequently in everyday speech, in particular among children and women. However, these sounds can also be evidenced for men using “falsetto”.

Speakers possessing excellent vocal abilities are even able to form the sounds of the corner vowels /i, a, u/ in a clearly recognisable and distinguishable way up to a fundamental frequency of c. 800–1000 Hz. (Ongoing research also indicates that other vowels, too, are intelligible in this vocal range.)

Correspondingly, the respective sound spectra exhibit vowel-specific differences, even if these have to be described other than in terms of spectral envelopes and formant patterns, for example in terms of vowel-specific configurations in the levels of the harmonics (see below, Sections 13.2 and 13.3).

Note that a fundamental frequency of 700 Hz lies above the statistical F1 values given for sounds of all long German vowels produced by women or men, except for /a/ of women. A fundamental frequency of 800–1000 Hz even lies above the statistical F1 values for all long German vowels, for both women and men (see Section 2.2).

The vowel quality of sounds produced at fundamental frequencies above statistical values of the vowel-related first-formant frequency is intelligible in principle.

The possibility of such vowel production and perception contradicts the designation of established, statistically determined formant patterns as “vowel-specific” patterns, irrespective of the methodological problems of determining envelope peaks and formant frequencies. At the same time, vowel perception and discrimination at such high fundamental frequencies confirms that lower spectral energy maxima (if determinable) and lower formants (if methodically substantiated) depend on fundamental frequency.

The vowel quality of sounds of back vowels and of /a–ɑ/ produced at fundamental frequencies > 500 Hz can be physically represented solely in terms of the first two or three harmonics and their amplitudes. This accentuates the basic problem of assuming that relative spectral energy maxima, that is, envelope peaks in closely delimited frequency ranges, are a pervasive physical characteristic of the sound of a vowel.

Here, the question of the maximal fundamental frequency up to which all vowels of any given language can in principle be produced in a recognisable way is left open for discussion.

8.3 “Inversions” of Relative Spectral Energy Maxima and Minima and “Inverse” Formant Patterns in Sounds of Individual Vowels

Given that spectral envelope peaks ≤ 1.5 kHz (if determinable) depend on fundamental frequency, pairs of sounds of a back vowel produced at different fundamental frequencies can exhibit “inverse” relative spectral maxima and minima in the form of “inverse” spectral envelope curves ≤ 1.5 kHz without any change in vowel perception: whereas we see a relative minimum in the spectrum for one sound, we may observe a spectral maximum for the other, and vice versa. The same holds true for comparisons between the respective calculated filter curves and formant patterns (if methodologically substantiated): where for one sound, the filter curve exhibits a relative minimum, for another sound, the curve may exhibit a maximum, and vice versa.

In the case of some front vowels, such “inversions” can also be observed for the higher vowel-specific frequency range, even if the question of the relationship between such “inversions” and fundamental frequency variation is left open here.

This observation reaffirms the lack of a general correspondence between vowels, vowel-specific spectral envelope curves and corresponding formant patterns.

> With regard to vowel-specific frequency ranges, the spectral envelope curves of two sounds of the same vowel produced at two different fundamental frequencies can exhibit “inverse” behaviour. The same holds true for formant patterns.

8.4 Addition: Whispered Vowel Sounds, Fundamental-Frequency Dependence of Vowel-Specific Spectral Characteristics and “Inversions”

As discussed in Section 5.5, formant statistics indicate increased vowel-specific formant frequencies F1 and F2 for whispered sounds when compared to voiced sounds. However, according to the corresponding recording procedures of the comparative investigations, this only applies to the lower range of fundamental frequency of the voiced sounds produced in citation-form words, comparable to relaxed speech in an enclosed space.

Given that a whispered sound exhibits higher first and second formants than a voiced sound of the same vowel and given that the latter’s fundamental frequency is gradually increased during its production, then in many cases it is possible to determine a certain fundamental frequency for which F1 and F2 of the whispered and voiced sound correspond with each other.

Whether this represents an actual rule is left open here.

If the fundamental frequency of a voiced sound is increased further, then there will be cases in which F1 or F1–F2 of the whispered sound are lower than F1 or F1–F2 of the voiced sound.

In any event, the general statement that whispered sounds exhibit fundamentally higher vowel-specific formant patterns than voiced sounds does not apply.

Over the course of such experimentation, cases involving comparisons between whispered and voiced sounds exhibiting the described “inversions” may also be found.

8.5 Addition: Resynthesis and Synthesis

All the above aspects of the lack of correspondence between vowels and patterns of relative spectral energy maxima or formant patterns, discussed in relation to natural vowel sounds, can be evaluated and replicated using resynthesis.

The same holds true for resynthesis at fundamental frequencies > 350 Hz related directly to the harmonic spectra of natural vowel sounds.

The same also applies to synthesis involving formant patterns or harmonic spectra not derived directly from natural vowel sounds.

9 Ambiguous Correspondence between Vowels and Patterns of Relative Spectral Energy Maxima or Formant Patterns or Complete Spectral Envelopes

9.1 Ambiguous Patterns of Relative Spectral Energy Maxima and Ambiguous Formant Patterns

All these reflections and observations come down to the conjecture that two sounds of two different vowels, produced at two different fundamental frequencies, can exhibit quasi-identical relative spectral energy maxima and quasi-identical formant patterns within their vowel-specific frequency range. Indeed, many patterns of spectral envelope peaks and formants prove to be ambiguous empirically. As such, they often physically represent two (or even several) different vowels.

> In many cases the patterns of relative spectral energy maxima do not prove to be vowel specific, but ambiguous. The same holds true for formant patterns.

This observation becomes particularly evident when comparing vowel sounds for their entire range of fundamental frequencies for which vowels are recognisable and distinguishable and when evaluating the correspondences between relative spectral energy maxima and minima also in a direct comparison of harmonic spectra, aside from determining spectral envelopes and formant frequencies.

9.2 Ambiguous Spectral Envelopes

In certain cases, this ambiguity also concerns the entire course of the spectral envelope.

> Spectral envelopes can be equally ambiguous.

9.3 Ambiguity and Individual Vowels

For all German vowels discussed here, there are cases of sounds with ambiguous patterns of relative spectral energy maxima or with ambiguous formant patterns within the respective vowel-specific frequency ranges.

To what extent this is also true for complete spectral envelopes is left open for discussion.

If vowel sounds are compared for their entire range of fundamental frequencies for which vowels are recognisable and distinguishable and if a possible correspondence of relative spectral energy maxima and minima is evaluated in a direct comparison of harmonic spectra, then, the above ambiguity can be observed not only for sounds of neighbouring vowel pairs but also for other sound pairs and sometimes for sounds of more than two different vowels. This holds particularly true when comparing sounds produced by all of the three age- and gender-related speaker groups.

The ambiguity described is not limited to only a part of the vowels or to neighbouring vowel pairs, and it can affect more than two vowels simultaneously.

The question of whether there are sounds of certain vowels that exhibit strict vowel-specific patterns of relative spectral energy maxima and strict vowel-specific formant patterns, which cannot be found in sounds of any other vowel—for example for sounds of /a/—is left open for further discussion.

9.4 Addition: Resynthesis and Synthesis

The ambiguity discussed above in relation to natural vowel sounds can be evaluated and replicated using resynthesis.

The same also applies to synthesis involving formant patterns or harmonic spectra not derived directly from natural vowel sounds.

10 Lack of Correspondence between Patterns of Relative Spectral Energy Maxima or Formant Patterns and Speaker Groups or Vocal-Tract Sizes

10.1 Similar Patterns of Relative Spectral Maxima and Similar Formant Patterns ≤ 1.5 kHz for Different Speaker Groups or Different Vocal-Tract Sizes

If sounds of a vowel are produced at equal fundamental frequencies by children, women and men, and if these sounds perceptually correspond with each other not only in terms of their general attribution to a vowel quality but also in terms of the respective "vowel-colour" variant—which makes for the greatest possible correspondence as regards perception—then, empirically, both the relative spectral energy maxima (if determinable) and the formant patterns (if methodically substantiated) often prove to be similar in the lower frequency range ≤ 1.5 kHz, apart from possible differences due to the different parameter settings involved in formant analysis. Expected age- and gender-related spectral differences decrease or disappear if the fundamental frequency of the utterances correspond for children, women and men.

Further, for sounds of back vowels and sounds produced by men at higher fundamental frequencies than women, it follows that the sounds of men (at higher F0) may exhibit higher relative spectral energy maxima (if determinable) and higher F1 or even F1–F2 patterns (if methodically substantiated) than the sounds of women (on lower F0), as holds true for F1 of front vowels. The same may also occur in a corresponding comparison of sounds of adults and children.

No statements are made here on /a–ɑ/ since our observations do net yet allow for general formulations for all sounds of /a–ɑ/ (see Section 8.1).

Thus, the question arises whether the lower range of the vowel spectrum mentioned is indeed dependent on age- and gender-related speaker groups, that is, on vocal-tract size. In the literature, this lower frequency range is referred to as being entirely vowel specific for all back vowels and, concerning F1, vowel specific for all other vowels.

In any event, the general statement that the sounds produced by children exhibit the highest, the sounds of women intermediate and the sounds of men the lowest patterns of vowel-specific relative spectral energy maxima and formant frequencies does not apply.

Within the frequency range of ≤ 1.5 kHz, vowel-specific patterns of relative spectral energy maxima (if determinable) and formant patterns (if methodically substantiated) often prove to be empirically independent of the age- and gender-related speaker group, that is, the vocal-tract size. Given strict perceptual correspondences, then, differences refer directly to the differences in fundamental frequency.

As mentioned, the possible relationship between fundamental frequencies and higher vowel-specific spectral envelope peaks or formants for sounds of front vowels is left open for discussion. In the present context, this also concerns the question of whether or not higher frequency ranges are in principal specific to vocal-tract sizes.

10.2 The Dichotomy of the Vowel Spectrum

As mentioned repeatedly, while the dependence of vowel-specific spectral characteristics and formants on fundamental frequency for the lower frequency range ≤ 1.5 kHz is easily understandable and reproducible empirically, this is not the case for the higher frequency range. At the same time, lower spectral ranges and lower formant frequencies are not generally specific to speaker groups and vocal-tract sizes. Whether this is also the case for higher spectral ranges and formant frequencies is still in question. Thus, the spectrum of a vowel sound needs a twofold rather than a uniform consideration.

The spectrum of a vowel proves to be dichotomous.

In this context, with regard to the sounds of front vowels, it is particularly important to consider that, in certain cases, higher relative spectral energy maxima (if determinable) and higher formants (if methodically substantiated) > 2 kHz may be simultaneously related to vowel identity and perceived speaker group: differences in this higher frequency range can often be observed for sounds of a front vowel produced by children, women and men if the speakers form these sounds at similar fundamental frequencies, even if there is no such difference found in the lower frequency range.

However, it is left open for further investigation whether this is also the case if men imitate so-called “female voices” or if adults imitate “children’s voices”.

10.3 Addition: Whispered Vowel Sounds and Speaker Groups or Vocal-Tract Sizes

No results of comparative studies of formant patterns for whispered vowel sounds of children, women and men have been published to date that have obtained a reference status as is the case for reference statistics of voiced vowel sounds referred to in Part II. However, the studies that compare whispered sounds of different speaker groups (limited in number and generally not including all vowels of a language) refer to corresponding differences between formant patterns.

Notwithstanding the reflections and comments made so far, these differences can be understood as an indication of a general relationship between patterns of relative spectral energy maxima and formant patterns on the one hand, and speaker groups, that is, average vocal-tract sizes on the other, including the lower frequency ranges.

This aspect and its significance regarding the relationship between vowels and related spectral characteristics is left open to discussion here and needs to be clarified and discussed elsewhere.

10.4 Addition: Vowel Imitations by Birds

Sounds of animals imitating utterances of humans are also of primary importance in the discussion of vowel sounds, related spectral characteristics, formant patterns, perceived speaker groups and vocal-tract sizes.

Fundamental in this respect is the question of how birds are able to imitate human sounds despite lacking the means of phonation and articulation—in particular, a corresponding vocal-tract resonance.

According to our own preliminary examination of vowel imitation by common hill myna birds who excel at such mimicry (results unpublished, although some clear examples are given in the Materials section), we conclude the following: if these birds imitate words, and if individual imitated vowel sounds are isolated as sound fragments in a way that they possess a quasi-static character in terms of quasi-static spectral characteristics (above all, that transitions are excluded), then vowel perception and a distinction of such sounds by humans is possible. For part of these sound fragments, complete F1–F2–F3 formant patterns comparable to patterns given for human sounds can be interpreted. For the remaining fragments, only a partial correspondence in formant patterns can be observed. (However, this statement must be relativised: strictly speaking, any calculation of vowel-related formant patterns of bird sounds is methodically unsubstantiated; see below.)

The fact that birds are able to imitate human vowel sounds with vowel-specific spectral characteristics and formant patterns comparable to those of humans contradicts, in its turn, a strict correspondence between the spectral characteristics of the produced sound and vocal-tract resonance. The same holds true for a strict correspondence between spectral characteristics of the produced sound and vocal-tract size. Consequently, any critical investigation and discussion of vowels must focus on the possibility that the same sound characteristics can be produced under substantially different physical and physiological conditions.

Besides, if birds are able to mimic human utterances, they must be able to perceptually differentiate different vocal sounds. However, their perception cannot rely on any sensomotoric and conceptual experience of vowel production comparable to the experience of humans. Thus, it can be speculated that their perception relies on a more "abstract" acoustic "form" of the vowel sound. (Such speculation would meet the claim that a phenomenological approach to the physical representation of vowels is needed; see Part V.)

10.5 Addition: Resynthesis and Synthesis

Again, the lack of a general correspondence between patterns of relative spectral energy maxima or formant patterns and speaker groups or vocal-tract sizes can be evaluated and replicated using resynthesis and synthesis.

11 Lack of Correlation between Methodological Limitations of Formant Determination and Limitations of Vowel Perception

11.1 Vowel Perception at Fundamental Frequencies > 350 Hz

As discussed in Section 8.2, recognisable vowels can be produced at fundamental frequencies substantially exceeding the critical limit above which formants can no longer be reliably determined for methodological reasons.

> Vowel perception is maintained for sounds at fundamental frequencies > 350 Hz. Yet, for these middle and higher fundamental frequency ranges, formant pattern estimation is questionable for methodological reasons. Thus, the methodological limitation of determining formant patterns of vowel sounds at fundamental frequencies > 350 Hz does not coincide with impaired vowel intelligibility.

Consequently, formulating a general theory of the physical representation of vowels based on formant patterns proves to be critical due to the related methodological limitations.

11.2 Lack of Correspondence between Methodological Problems of Formant Pattern Estimation at Fundamental Frequencies ≤ 350 Hz and Impaired Vowel Perception

Vowel sounds produced at fundamental frequencies ≤ 350 Hz, for which the estimation of formant patterns proves questionable for reasons other than fundamental frequency—for instance, if expected relative spectral energy maxima are “missing” or if vowel-related parts of a spectrum are “flat”—are not less recognisable than vowel sounds for which formant pattern estimation may be said to be unproblematic.

> Methodological problems regarding the determination of formant patterns of vowel sounds at fundamental frequencies ≤ 350 Hz do not necessarily coincide with impaired vowel intelligibility.

11.3 Addition: Lack of Methodological Basis of Determining Formant Patterns for Vowel Mimicry by Birds

Given the prevailing methodological standards, strictly speaking, the imitation of human vowel sounds by birds cannot be studied in terms of formant patterns. As explained in Section 6.3, formant calculation requires parameter settings for the frequency range and the maximum number of filters used in the analysis in relation to a specific vocal-tract size. Birds, however, have no vocal tract comparable to that of humans. Hence, it is impossible to determine how many filters should be used in analysing a vowel-like sound produced by a bird to determine vowel-specific formants.

Thus, in a first step, comparisons between the utterances of humans and birds must be based on a direct comparison of the respective spectra and must relate to the interpretation of observable relative spectral energy maxima. However, in a subsequent step, formant analysis double-checked by resynthesis may be applied even if methodically unsubstantiated, in order to foster the discussion.

Again, this methodological limitation of mimicry analysis does not coincide with a principal difficulty to identify the imitated vowel sounds involved.

Part IV Falsification

The fourth part of the main text explains why the reflections, experiences and observations compiled here falsify prevailing theory.

12 Empirical Falsification despite Methodological Limitations of Determining Patterns of Relative Spectral Envelope Maxima or Formant Patterns

12.1 Lack of Methodological Basis for Verifying Prevailing Theory

Concerning isolated vowel sounds exhibiting quasi-static spectral characteristics and allowing for clear perceptual vowel recognition and distinction, it is not possible, in a particular language, to formulate general rules for determining patterns of relative spectral energy maxima or of formant patterns which consistently correspond to the perceived vowel quality of the sounds.

Consequently, it is not possible to gather general statistical data on vowel-specific formant frequencies of recognisable vowel sounds referring to the entire realm of utterances.

Prevailing theory cannot be verified for methodological reasons.

From a methodological perspective, prevailing theory, thus, is not endowed with adequate analytical instruments for capturing and describing the phenomenon of the vowel.

Existing references regarding formant statistics do not disclose this problem for two reasons: firstly, the investigated speakers are generally not subject to a qualitative selection regarding their vocal abilities; secondly, such statistics generally exclude any systematic and extensive variation of fundamental frequency. Both factors, however, are essential prerequisites for studying the possible fundamental frequency ranges of intelligible vowel sounds and for examining the appropriateness of the methods of acoustic analysis with regard to the entire realm of utterances. (Moreover, qualitative speaker selection also allows for the study of other important aspects of vowel-sound variation, above all variation of vocal effort, register and phonation type.)

12.2 Systematic Divergence of Empirical Findings from Predictions of Prevailing Theory

If lower relative spectral energy maxima can be determined and if correspondent formant frequency calculation can be methodically substantiated, in most cases, the corresponding patterns ≤ 1.5 kHz, that is the lower frequency range of the spectra, prove to be dependent on the fundamental frequency of the sounds relative to the recognised vowel.

Speakers of a given speech community, despite having different vocal-tract sizes and thus belonging to different speaker groups, are nevertheless able to produce the sounds of one and the same vowel at quasi-identical fundamental frequencies and with quasi-identical lower formant frequencies ≤ 1.5 kHz. Moreover, speakers with comparatively larger vocal-tract sizes can produce sounds of some vowels at higher fundamental frequencies and with higher F1 or even higher F1–F2 values than speakers with comparably smaller vocal-tract sizes.

These empirical findings are reciprocally related. They diverge systematically from both the predicted independence of vowel-specific formant patterns on fundamental frequency and the predicted pervasive dependence of vowel-specific formant patterns on speaker-group or vocal-tract size, respectively.

Empirical findings diverge systematically from the predictions of prevailing theory.

From an empirical perspective, prevailing theory thus proves to be inadequate.

12.3 Empirical Findings Directly Contradicting Prevailing Theory

A single speaker may not only occasionally produce different isolated sounds of different vowels exhibiting the same formant patterns F1–F2 or F1–F2–F3 but, for some vowel qualities, this formant pattern ambiguity of vowel sounds in relation to the perceived vowel quality is systematic if the entire range of fundamental frequency of intelligible vowel sounds is investigated. In these cases of ambiguity, speakers cannot substantially vary fundamental frequency, maintain vowel quality and also maintain formant patterns: if the speaker maintains the vowel quality, the formant pattern will alter, or if the formant pattern is kept constant, the vowel quality will change.

This observation also holds true for patterns of spectral energy maxima. Moreover, in some cases, as mentioned, even the entire interpretable spectral envelope proves to be ambiguous.

Empirical findings can directly contradict the predictions of prevailing theory.

Consequently, prevailing theory is falsified because, for a substantial portion of vowel sounds, the opposite of what the theory claims to be true actually applies: in many cases, given a variation of fundamental frequency, vowel sounds with very different formant patterns allow for a perception of the same vowel quality, while vowel sounds with similar formant patterns allow for a perception of different vowel qualities.

Part V Commentary

The fifth part of the main text discusses the resulting state of affairs and points to the need to devise a phenomenology and to develop a new theory.

13 Preliminaries

13.1 Impediments to Adjusting Prevailing Theory

In response to the principal difficulties in intellectually re-enacting the prevailing theory of the acoustics of the vowel and in response to the empirical observations discussed in the previous chapters, there are several arguments against adjusting or modifying prevailing theory and the corresponding methods of acoustic analysis.

According to prevailing theory, formant patterns are deduced from patterns of vocal-tract resonances. The formulation of a substantial interrelation between these resonances and fundamental frequency in the production of vowel sounds would directly contradict the two-part model of source and filter and the corresponding understanding of phonation and articulation, namely, the production of a general source sound and its transformation by vocal-tract resonances. Fundamental frequency is a primary characteristic of the source, and resonances are a primary characteristic of the vocal tract. These resonances are independent of the sounds or noises affecting them. (Interactions of source and filter, as described in the literature, do not relate to the aspects discussed here.) This amounts to a fundamental conceptual obstacle when it comes to differentiating or modifying prevailing theory.

Current methods of formant analysis neglect fundamental frequency as a source characteristic in the calculation of filters. There is little scope for changing this approach within the existing procedural framework.

Besides, even if formants are not considered to be directly linked to vocal tract resonances, interpreting them solely as results of an analytical decomposition of a sound in a source and a set of filters, it proves difficult to imagine a corresponding method of acoustic analysis applicable to all recognisable sounds and all of the aspects discussed in Part III. This lack of projection itself impedes the modification of prevailing methodology.

The observable behaviour of vowel-specific patterns of relative spectral energy maxima (if determinable) and of formants (if methodically substantiated) cannot be formulated in terms of a general rule, such as relating these characteristics to fundamental frequency as a simple ratio, whether or not such a ratio is based on an auditory scale. Empirically, these characteristics prove to be unsystematic: in general, the shifts in the spectral envelope peaks and the formants discussed are distinctly evident only at fundamental frequencies above c. 200 Hz; the shifts affect the lower spectral frequency ranges and the higher ranges differently; thus, the shifts affect the entire vowel-specific frequency range of back vowels in a direct way but only affect the vowel-specific frequency range of front vowels partly; the shifts relate to vowel quality, yet in parallel, they also relate to the frequency levels of the spectral envelope peaks or formants in question; in addition, a strong variation in vocal effort also affects the frequency location of the spectral envelope peaks and the calculated formants.

Because of this lack of systematic empirical evidence and because there is no uniform method for analysing vowel-specific acoustic characteristics, including all utterances allowing for vowel perception, no robust basis exists for a further differentiation of the description of the vowel-specific spectral characteristics within the prevailing approach to relate to patterns of spectral peaks or patterns of formants.

These reflections, experiences and observations constitute the scepticism expressed in this treatise about attempting to adjust or modify prevailing theory and related methods of further analysis.

13.2 Prevailing Theory as an Index

Given that a voiced vowel sound is produced in isolation and that it exhibits a quasi-constant periodic spectral characteristic, and given its unambiguous perception as belonging to a specific vowel quality (related to a particular language), then, its average harmonic spectrum, measured for the entire duration of the respective sound, is said to be vowel specific: for a frequency range concerning the physical representation of all vowels of the corresponding language, a series of harmonics quasi-identical in number, frequencies and levels, can only be found for other sounds of the same vowel but not for other sounds of any other vowel. Such a statement is formulated in terms of a hypothesis here.

The same holds true for corresponding sounds that are isolated from a particular syntactic and semantic context and that are analysed accordingly as sound fragments.

Obviously, a direct comparison of harmonic spectra always relates to sounds at quasi-identical fundamental frequencies.

Harmonic spectra, as claimed here, are vowel-specific and, further, may also prove to be orthogonal in vowel representation: on their basis, the respective sounds are expected to be reproducible without any change in the perceived vowel quality.

Hence, the fundamental aspects of the problems discussed in the previous parts of this treatise can neither be attributed to dynamic processes occurring within a sound nor to the particular characteristics of the syntactic and semantic context, nor indeed to special perceptual processes. Nor can these aspects be relativised accordingly. On the contrary, they constitute an ensemble of individual problems that first needs to be explained, just as the physical representation of the vowel itself, as a phenomenon, needs to be clarified.

Given that voiced vowel sounds are compared at similar fundamental frequencies, and given that the spectral envelope is determined by the amplitude values of the harmonics, obviously, such an envelope is also vowel specific. However, concerning spectral envelope peaks, no simple statement can be derived if all fundamental frequencies of intelligible vowel sounds are considered.

Given that voiced vowel sounds are compared at similar fundamental frequencies, and given a methodological substantiation, it can be expected that calculated formant patterns (including formant bandwidths) may, in most cases, also prove to be vowel specific and that, on their basis and not altering fundamental frequency, the respective sounds can be reproduced without substantial change in the perceived vowel quality.

Thus, prevailing theory “hints” or “points” at the basic characteristic of the physical representation of vowel quality in an indexical manner. Prevailing theory proves to be an index of this representation.

13.3 Excursus: Vowel Quality and Harmonic Spectrum

To repeat: given that a voiced vowel sound is produced in isolation and that it exhibits a quasi-constant periodic spectral characteristic, and given its unambiguous perception as belonging to a specific vowel quality, then its average harmonic spectrum, measured for the entire duration of the respective sound, is said to be vowel specific. For a frequency range concerning the physical representation of all vowels of a language, a series of harmonics quasi-identical in number, frequencies and levels can only be found for other sounds of the same vowel but not for other sounds of any other vowel.

At first glance, such a statement seems trivial. But it is not.

To say that a harmonic spectrum of a vowel sound is specific for the perceived vowel quality—given the above conditions for the sounds under investigation—is not to say that all sounds of a vowel have very similar spectra of this kind. As shown, large spectral variations can be found for the sounds of one vowel, particularly if vocal effort is varied during the sound production, if sounds of different speaker groups are compared and if different speaking and singing modes and styles, including stage voices, are also considered.

Therefore, an attempt to directly assess the spectral difference related to a perceptual difference of two vowels simply by calculating an average harmonic spectrum for all sounds of one vowel at a given fundamental frequency and comparing it with the similarly averaged harmonic spectrum of the other vowel may, in many cases, not result in a clear spectral difference, that is, in a frequency limit from which the two averaged spectra begin to diverge with no overlap. Exceptions may occur at high fundamental frequencies because the perceived vowel quality is represented by a greatly reduced number of harmonics.

Considering both the direct relation between harmonic spectrum and perceived vowel quality on the one hand and the observably large variation of harmonic spectra for sounds of single vowels on the other, and speculating that instead of looking at a static spectral configuration we should consider looking at a kind of spectral foreground-background relation, another attempt may provide more evidence.

If the harmonic spectrum of a reference sound of a vowel is compared with both the spectra of other sounds of the same vowel and the spectra of sounds of a second vowel, then there will be a frequency limit above which the spectrum of the reference sound diverges from any spectrum of the sounds of the second vowel, but not from any spectrum of the sounds of the same vowel.

More precisely, any single sound of a vowel compared with sounds of another vowel (given similar fundamental frequencies of the sounds) is assumed to be describable in terms of a relation of maximal spectral similarity and subsequent—related—spectral difference: for a (lower) frequency range, the harmonic spectrum of the single sound of the first vowel of comparison can resemble some other harmonic spectra of the second vowel, but if the maximum of this frequency range of possible resemblance is reached, its spectrum differs from all the spectra of the second vowel sharing the maximal similarity, while still resembling some other spectra of the first vowel.

This principle is taken here as the most conservative but also the most promising approach and basis for future research on the acoustics of the vowel: it is testable and falsifiable in a fully objective manner for all levels of fundamental frequency of comparison, it does not need further differentiations related to speaker groups or vocal effort or speaking or singing styles or modes and, therefore, its testing does not require any integration of further phonetic knowledge. Moreover, it also applies to synthesised sounds which are produced using a harmonic synthesiser. Thus, it “hints” or “points” at the basic characteristic of the physical representation of vowel quality in a much stronger manner than prevailing theory, i.e. it is a stronger index of this representation.

Moreover, if developed in more detail, it leads to an entire system comprising various possible relations of spectral similarities and related spectral differences of sounds of all vowels for a given language.

Although formulated on the basis of a very extended knowledge of vowel spectra, obviously, these short reflections are but general assumptions open to further clarification and empirical verification or falsification, and even if they can be empirically demonstrated as valid, they would still remain a fragmentary and temporary basis in the course of reformulating the acoustics of the vowel. Therefore, the drawbacks of investigating the harmonic spectrum—above all, the impossibility of comparing spectra related to very different fundamental frequencies directly, and the impossibility of including the analysis of vowel sounds not exhibiting quasi-static periodic characteristics of the sound wave—are not further discussed. The same applies to the limitation of the principle formulated, namely, that it only relates to vowel-specific spectral differences but not to a full determination of vowel-related acoustic characteristics.

However, for further advances in the investigation of the acoustics of the vowel, an assessment of the reliability of every given statement is needed, and the possibility of a falsification plays a crucial role in this assessment: it is the falsification of a generalised assumption of vowel-related formant patterns that called for this treatise.

Up to now, concerning the acoustics of the vowel, there are only two statements that apply to all vowel sounds: vowel sounds, perceived as isolated single sounds, are intelligible and therefore, the vowel quality must be physically represented in the corresponding sound wave and its characteristics. According to this view, an investigation of the harmonic spectra is one of the most promising approaches, even if it is limited to quasi-constant voiced vowel sounds.

Such a step-by-step procedure will be needed as long as there is no objective and orthogonal method to describe the acoustic characteristics that physically represent the perceived vowel quality, including all types of vowel sounds (see also below). However, during this procedure, a kind of rule-based knowledge will emerge and provide a basis for the development of an objective and comprehensive method.

13.4 "Forefield"

All of the above leads to the conclusion that, at present, no theory of the acoustic representation of the vowel exists. However, empirical evidence exists that indicates the possibility of such a theory and that will contribute to its development. Thus, it is currently in its preliminary stages.

13.5 Two Approaches

Prevailing theory is characterised by its explanation and description of vowel sounds within a physical model unspecific to speech: all kinds of sounds and noises are transformed by filters in the same way, irrespective of whether or not they concern utterances (speech events).

One possible way to respond to the difficulties of understanding prevailing theory in terms of its intellectual re-enactment and to the fact that empirical findings can contradict its predictions might be to supplement the existing source-filter model or to replace that model by another physical model external to language and speech.

Another approach might be to assume that the production and formation of vocal sounds is speech specific and, based on such a premise, to develop a method for describing vowel sounds in form-related terms. This second approach assumes that the vowel sound and its manifestations elude description within a purely physical model.

Whether this covers all of the possible approaches is left open for discussion here.

As explained in Section 13.1, there are substantial reasons for scepticism about the possibility of adjusting prevailing theory and the related methods of acoustic analysis. One further and important aspect, in addition to the arguments already mentioned, concerns the following consideration: it would be possible for humans to produce a vowel-unspecific source sound and transform that sound using vocal-tract resonances, thereby producing the respective vowel-specific physical characteristics according to which listeners perceive vowels, both unambiguously and independently of fundamental frequency. But it belongs to the actual acoustic phenomenon of the vowel sound to systematically deviate from this. The empirical evidence for vowel sounds suggests that humans do not produce such sounds as systematically as physics and physiology seem to predetermine. This contrast might prove fundamental for future theory building.

Elsewhere, the author has formulated the state of affairs as follows: (i) either resonances as such, and thus the corresponding pharyngeal, oral and nasal resonance patterns of the vocal tract, fail to represent in full the physical quantity to which language and speech directly refer, but another physical quantity can be found instead; if this is the case, then it is simply a matter of replacing the existing (physical) model with another rather than adopting a fundamentally different perspective; (ii) or, aside of the human voice, no construction, no instrument and no process can be found to exist in physics that would explain and allow for the production of vowel sounds including basic variations of sound characteristics, for example, fundamental frequency and phonation type; then, the physical representation of human voice cannot be related to a simple voice-independent physical quantity, but instead, the voice would produce a "substance" or "quantity".

Based on all the reflections, experiences and observations presented, this treatise belongs to the second kind of undertaking. This calls for a corresponding phenomenology and for theory building.

13.6 Phenomenology

On the one hand, the existing documentation of vowel sounds hitherto published is no more than fragmentary and on the other, the methods for describing their acoustic characteristics have substantial shortcomings and limitations. Thus, as argued above, a phenomenology is needed, that is, a step by step build-up of systematic compilations of vowel sounds related to individual languages, including the variation of all relevant production parameters. In its course, attempts for describing acoustic characteristics related to vowel qualities in terms of knowledge-based rules will become possible (see above).

In the first instance, such a phenomenology refers to the vowel sounds of a particular language, produced in isolation or detached from sound context, exhibiting quasi-constant spectral characteristics and allowing for high scores of vowel identification in listening tests, involving listeners of the speech community of that particular language.

13.7 Theory Building

As said, vowel sounds perceived as isolated single sounds can be intelligible. This fact is central to human voice and speech. With regard to such sounds, the psychophysical question rises as to which describable physical characteristic or which ensemble of physical characteristics may be said to represent the perceived vowel qualities.

Theory building thus faces a threefold challenge. Firstly, it must produce a uniform, systematic and orthogonal method to describe vowel-specific acoustic characteristics. Only such a descriptive method enables a systematic synthetic reproduction of vowel sounds, based on empirically determined characteristics of natural vocalisations, and thus the verification of the significance of corresponding analyses. Secondly, in relation to the phenomenology discussed, theory building must deduce hypotheses that predict the physical representation of vowel quality irrespective of the individual cases of the vowel sounds, thus extrapolating the phenomenological description. These hypotheses must satisfy the requirements of verification and falsification on the one hand, and be transferable to different languages on the other. Thirdly, theory building must seek to explain empirical findings and the hypotheses deduced from such findings.

Afterword

This treatise is bound to raise many questions, which are not discussed in detail here. Moreover, according to previous experiences regarding academic discussions, some of the considerations and arguments presented here are likely to be refuted on principle.

Some major issues discussed in the literature have not been considered in depth in this text, so that its main argument could be presented in straightforward, general and clear terms. Moreover, in-depth consideration of the issues mentioned has also been dispensed with because they appear in a different light from the perspective adopted here and, thus, they need to be discussed in another context than is usually the case. Within the following exemplary comments, however, some indications are given.

Against the background of the present reflections, experiences and observations, we conclude that explaining the lacking distinctiveness of expected spectral energy maxima in terms of the characteristics of auditory perception as formant merging without taking into account the entire systematics of empirically observable, vowel-specific spectral characteristics—in particular their dependence on fundamental frequency and the possible ambiguity of spectral envelope peaks and of formant patterns—is questionable.

The same holds true for normalisation attempts with regard to the presumed general differences between the vowel-specific formant patterns among children, women and men: such normalisation attempts would have to be approached quite differently if the comparisons of the formant patterns of the three speaker groups did not only include different but also similar fundamental frequencies of the sounds of all groups.

The same also applies when attempting to generally relate formant shifts, which occur when the fundamental frequency for sounds of one vowel is raised, to paralinguistic characteristics, in particular vocal effort: low- and high-pitched sounds can both be formed loudly and softly, and the calculated formant patterns of vowel sounds do not only depend on the vowel quality but also on the fundamental frequency in principle. Hence, one has to expect the occurrence of ambiguous formant patterns for sounds produced with equal vocal effort. Thus, we conclude that the shifts of the lower formants with raising fundamental frequency and formant pattern ambiguity as such are not necessarily related to paralinguistic aspects.

To mention one last example, the same also applies when attempting to relate, in general terms, formant shifts that occur when raising fundamental frequency in singing to formant tuning: evidence given in the studies published on this matter does not allow for a conclusion of whether the documented observations refer to the idiosyncrasies of individual singers, to the stylistic characteristics of a particular singing technique or style with changes in vowel quality (possibly caused by so-called vowel modification), to vocal effort, or whether the observations indeed refer to the fundamental characteristics of vowel sounds.

However, the tendency in the literature to consider vowel sounds at lower fundamental frequencies—with limited frequency variation—to be characteristic of speech, and vowel sounds at middle and higher fundamental frequencies—with extensive frequency variation—to be characteristic of singing, and the tendency to conduct investigations based on these assumptions, needs to be refuted in its turn: neither in everyday life, nor in the entertainment sector, nor indeed in musical art and vocal interpretation is there any such thing as "normal speech" for which, in contrast to singing, a single "average" fundamental frequency could be statistically determined.

The corresponding indications in existing formant statistics are not representative of experienceable speech and observable acoustic characteristics of vowel sounds: they are only representative of sounds uttered into a microphone in a small room in a relaxed and quasi-monotonous manner. (Such a restricted formulation still lacks contextual relativisation in terms of a particular language and "culture".) There is no essential difference between the fundamental frequency ranges for speaking on the one hand, and for singing on the other, no matter how these categories are determined and distinguished in a scientifically reasonable way. If one attentively listens to everyday utterances and to utterances in theatre and film—nowadays easily accessible due to television—, the corresponding experiences make this plain, and both fields of experience need to be integrated into a phenomenology of vowel acoustics (see the Materials section for corresponding examples).

With this consideration in mind, as mentioned, some of the major aspects discussed and interpreted in the literature appear in a different context than is often reflected upon.

As indicated at the beginning of this afterword, the present critical take on the prevailing theory of vowel acoustics must, in turn, prompt scepticism, as has already become evident in many scholarly debates, together with the respective counterarguments. This text has attempted to take into account these arguments. Additional comments follow below.

Whatever the extraordinary and often surprising role of perception in the recognition of speech sounds, this role neither relativises the fact that isolated vowel sounds with a quasi-static sound characteristic can be intelligible beyond a concrete syntactic and semantic context nor that their harmonic spectra are vowel specific. Thus, a psychophysical approach to such vowel sounds, that is, a theory of the relationship between perceived vowel quality and physical characteristic or ensemble of characteristics, must not only be deemed possible but also necessary. The psychophysics of vowel sounds constitutes the basis for an investigation of human voice and speech.

In particular, however, there has been a lack of a robust empirical, extensive, systematic and representative documentation of the aspects discussed in this treatise, and this reason is considered paramount here.

We adopt the viewpoint—and, therefore, have written this text—that any attempt to formulate such a theory in terms of formant patterns cannot be successful. Consequently, a different approach needs to be formulated.

What kind of explanation could be provided to explain the fact that most previous studies of vowel sounds, and thus of voice and speech, have not integrated such a line of argument? There seem to be several reasons for this shortcoming. In particular, however—and this reason is considered paramount here—, there has been a lack of a robust empirical, extensive, systematic and representative documentation of the aspects discussed in this treatise. Thus, as a consequence of the absence of such reference documentation, the discussion lacks a binding empirical basis any interpretation must account for. At the same time, the basis of a formulation of an alternative theory is lacking, too.

Thus, whereas existing individual values obtained in studies of vowel sounds apply to the specific conditions under which these data were gathered, the values are often interpreted in terms of a general physical representation of the vowel, which is empirically contradicted. Generalisation is the critical issue at stake.

To repeat: whereas average formant patterns (as determined statistically and separately for each of the three age- and gender-related speaker groups and related to average fundamental frequencies of relaxed and quasi-monotonous speaking into a microphone in a small, enclosed space) are in general vowel specific, the same does not hold true for substantial fundamental frequency variations evident as prosodic characteristics already in everyday language. Whereas for vowel sounds produced by men and involving a fundamental frequency variation of one octave but not exceeding 200 Hz by much, formant patterns (if methodically substantiated) of vowel sounds in most cases appear independently of fundamental frequency, the same does not hold true for the vowel sounds produced by the majority of women and by almost all children also involving a fundamental frequency variation of one octave. Whereas, in sound synthesis, a specific set of filters related to a specific fundamental frequency makes it possible to perceive a certain vowel quality, in many cases it does not hold true that the same vowel quality is perceived if the filter pattern remains constant but the fundamental frequency is significantly altered. And so on.

Because there is a lack of reliable, extensive, systematic and representative empirical references, including the documentation of variation of all basic production parameters needed in order to evaluate which physical characteristic is related to a single production parameter and which is in general related to vowel quality, and because, in many handbooks of phonetics, the acoustic characteristics of vowels are often treated briefly, in generalised and summary accounts yet without relativisation and problematisation, the reflections, experiences and observations reported in this treatise are partly unfamiliar, are rarely reconsidered and are in general not integrated when interpreting individual findings of other studies. In the first instance, this complicates the discussion within phonetics and psychophysics. Beyond this, however, attention has to be given to the significance of this lack of relativisation and problematisation for other areas of science—not only for fields such as speech recognition, speech pathology, audiology, or neuropsychology, but also for the investigation of voice as such, including philosophy and art, and for voice and speech education and training. How are these fields meant to relate to reliable basic knowledge and understanding of voice and speech production, and how are these fields meant to design reliable experiments if the unresolved problem of generalising individual measurements is not placed at the centre of understanding and investigation?

Moreover, some scholars are fundamentally critical of basing the psychophysics of the vowel on isolated vowel sounds and they question the recognisability and the linguistic function of such sounds. This critical position generally relates to a linguistic definition of the vowel as a vocoid and as syllabic. However, the previous reflections have shown that this treatise does not concur with the resulting notion of a fundamental opposition between isolated versus context-bound sounds, static versus dynamic spectral processes and “functionless” utterances versus those with a linguistic function. Much could be said in response both to such an opposition and a critical take on the psychophysics of isolated vowel sounds. Within the limited scope of the present study, however, only a few aspects can be mentioned (the problem as such is a matter for future debate and research):

- As said in the introduction, we take the stand that the recognisability of vowels (monophthongs) as single speech sounds perceived in isolation by listeners of a given speech community belongs to the elementarisation as a basic characteristic of vocal expression and speech and thus to the aptitude of the latter for a phonetic system of writing. Thus, structurally, isolated vowel sounds must be intelligible as such.
- Refuting the fundamental recognisability and the function of isolated sounds—their function in its broad sense, emotional and aesthetic qualities included—is borne out neither by any experience of art, vocal interpretation and entertainment, nor by everyday experience. (This order of denomination, artistic utterances first, everyday utterances last, is chosen to indicate that all phenomena discussed here may first be experienced in a direct way in the arts; then, when familiar with the correspondingly various types of possible utterances and expressions, they will continuously also get one’s attention in everyday utterances; see also the corresponding consideration in the introduction.)
- In this respect, it is worth pointing out the central role that is played by sustaining vowel sounds, sometimes for as long as possible, in musical composition and vocal interpretation—either in isolation or in a sound context and with or without fundamental frequency variation as a melody. The same holds true for basic and advanced voice training in the field of interpretation and performance.
- In this respect, it is also worth noting the occurrence of vowel sounds produced in isolation in vocal expressions such as exclamations or affirmations. (The German exclamations “Ahhh”, “Ohhh”, “Uhhh” and “Ihhh”, to give a paradigmatic example, have a different meaning depending on the context of expression and the vowels must be understood as such.)
- Dynamic processes are often represented and considered as formant transitions. Yet the lack of a general correspondence between vowel qualities and related formant patterns and the

> limited methodological reliability of formant determination—discussed in this study in relation to quasi-static sounds—must also be linked to dynamic descriptions. Thus, for instance, it is not evident how formant transitions for a sound produced at a fundamental frequency of approximately 200 Hz are supposed to be compared with those of another sound of the same vowel but at a fundamental frequency of 500 Hz.

Furthermore, whereas other scholars have recognised some or all of the problems discussed here, they often reproach the present kind of fundamental deliberation for not formulating a new theory. Such argumentation, however, does not correspond to the views and the stance of this treatise. If reasoned, well founded and applicable, any criticism of prevailing acoustic theory has its own intrinsic value, utterly irrespective of whatever it is that is offered or proposed beyond that criticism. Above all, it allows for an identification and formulation of challenges and, spurred by the need to resolve them, it drives the search for a new approach.

Pursuing a phenomenology and building a new theory requires a considerable effort along with the appropriate resources. Doing so presupposes that the scholarly community acknowledges the importance of such a venture. Any such acknowledgment, however, requires a comprehensible critique of prevailing theory to be advanced, together with a reinterpretation of previous empirical findings.

The author has also written this text because he does not know how far-reaching his contribution and that of his research colleagues is to the phenomenology and a new theoretical framework. However, two attempts are in progress. With regard to phenomenology, a research team is currently creating a large corpus of vowel sounds for Standard German, produced by children, women and men, including extensive variation of basic production parameters and including both untrained and trained speakers and singers (see Maurer, n.d.). In this way, we attempt to contribute to the creation of a systematic reference basis for vowels of single languages. With regard to theory, in a subsequent treatise, we will investigate in detail the thesis of vowel-specific harmonic spectra.

To conclude, the general significance of acoustic characteristics of vowel sounds should, as indicated, not be regarded as solely the subject of phonetics. Above all, it concerns the understanding of the voice as such.

The voice is currently attracting particular attention in the humanities. Deliberations in these fields are directly related to the knowledge and experience gained in artistic creation and interpretation, and there is a strong emphasis for the need for an interdisciplinary approach. In line with such a claim, the research culture in the aesthetics of the voice ought to adopt a particular stance toward the acoustics of the voice, too: namely, not to only cite phonetics with regard to existing descriptions of vocal utterances, but to critically discuss these descriptions and link them to considerations and experiences of art, interpretation and entertainment. In this context, a call should emerge not to take the "Western" perspectives and production styles as the starting point of investigation for the acoustics of the voice, but initially to consider any vocal expression, habit and style of any cultural context as equivalent. In doing so and in facing the diversity of possible vocal expressions, at least in the first instance, no classification of „normal" and „differing" phenomena and no hierarchical order should be imposed, but a decided descriptive perspective should be adopted. As said, there should be no underestimation or misunderstanding of the fact that raising questions regarding voiced speech sounds raises questions regarding the voice itself.

Our vocal cords produce sound. The resonances of the pharyngeal, oral and nasal cavities could form its characteristics into a formant pattern that always and uniquely represents a vowel physically, and thus allows the listener to perceive it accordingly. Empirical investigation reveals, however, that the spectral characteristics of vowel sounds systematically deviate from such an option. This observation leads to the conclusion that, at present, we are but in the preliminary stages of understanding the physical representation of the vowel and, thus, its materialised form.

Materials

The Materials section contains selected excerpts from the literature and presents exemplary series of vowel sounds and related acoustic analyses. An extended version of the Materials is also presented in digital form online; please refer to:
http://www.phones-and-phonemes.org/vowels/acoustics/preliminaries

Materials Part I

The first part of the Materials section contains selected excerpts from the literature that are related to the first part of the main text.

M1 Prevailing Theory

Vowels

“Vowel […]. 1. (also vocoid) In phonetics, a segment whose articulation involves no significant obstruction of the airstream, such as [a], [i] or [u]. Strictly speaking, a glide such as [j] of [w] may also be regarded as a (brief) vowel in this sense. 2. In phonology, a segment which forms the nucleus of a syllable. 3. Any letter of the alphabet which, generally or in a particular case, represents a vowel in sense 2.” (Trask, 1996, p. 382)

“Vocoid […]. 1. A synonym for vowel in the phonetic sense of that term (sense 1), introduced in an effort to remove the ambiguity between the phonetic and the phonological sense of ‘vowel’. While possibly useful, the term has never become established. Pike (1943). 2. More narrowly, a vocoid in sense 1 which is also syllabic: a true vowel, as opposed to a glide or approximant. Sense 2: Laver (1994).” (Trask, 1996, p. 378)

“Vowels and Consonants. Phonetics has traditionally classified the segments of speech into two basic varieties which are called vowels and consonants. Once again, there has never been a straightforward definition of these terms. Early linguists in India also grappled with the concepts of vowel, consonant, and syllable around 800 BC, and they recognized that the three notions are hopelessly intertwined […]. The definitions used here will be similar to those of the ancient Sanskrit scholars, and in fact, the development of modern phonetics in the West owes much to the transmission of knowledge in translation from the Sanskrit sources.

A vowel is defined as a ‘vowel-like segment’ (what Pike […] termed a vocoid) that occupies the nucleus of a syllable. A segment is considered to be a vocoid when its articulation permits the relatively free passage of air through the center of the mouth. This definition is also rather loose, but in roughly familiar terms, most segments that are at least as open as an English *w* or *y*-sound (the latter is transcribed [j] in IPA) are vocoids, all others being non-vocoids. A consonant is then defined simply as a non-vocoid, no matter what syllable position it occupies. This imperfect dichotomy leaves room for a middle category, that of the semivowel, which is defined as a vocoid located outside the nucleus of a syllable. Semivowels, in spite of being vocoids, are usually regarded as a special sort of consonant (often called a ‘glide’) in the interests of preserving the consonant-vowel dichotomy. The interplay of consonants, vowels, and syllables in the speech stream is given a slightly different (more acoustic) view by Orlikoff and Kahane: 'Consonants differ from vowels primarily by the amount of vocal tract constriction employed in their production […] Speech can be considered to be an overlay of consonants on the vocal signal. The dispersion of consonants results in an amplitude modulation of the acoustic energy that, for the most part, gives rise to our perception of syllables.'" (Fulop, 2011, pp. 8–9)

Speech production: source and filter

"The speech wave is the response of the vocal tract filter systems to one or more sound sources. This simple rule, expressed in the terminology of acoustic and electrical engineering, implies that the speech wave may be uniquely specified in terms of *source* and *filter* characteristics. In spite of the technical phrasing it is apparent that this statement also covers essentials of the phonetician's concept of speech production." (Fant, 1960, p. 15)

See also Chapter M4.

Formants

"The spectral peaks of the sound spectrum $|P(f)|$ are called *formants.* Referring to *Fig. 1.1-2,* it may be seen that one such resonance has its counterpart in a frequency region of relatively effective transmission through the vocal tract. This selective property of $|T(f)|$ is independent of the source. The frequency location of a maximum in $|T(f)|$, i.e. the *resonance frequency,* is very close to the corresponding maximum in spectrum $P(f)$ of the complete sound. Conceptually these should be held apart, but in most instances resonance frequency and formant frequency may be used synonymously. Thus, for technical applications dealing with voiced sounds it is profitable to define formant frequency as a property of $T(f)$.

The basic principle of the theory of voiced sounds is that, to a first order of approximation, the filter function is independent of the source. The formant peak will thus only accidentally coincide with the frequency of a harmonic. The formant frequencies can change only as a result of an articulatory change affecting the dimensions of the various parts of the vocal tract cavity system and thus the filter function. Conversely, but with the limitations implied by the concept of compensatory forms of articulation, the formant frequencies provide information about the position of the speaker's articulatory organs. If these formant frequencies are held constant and the fundamental frequency is raised one octave, the result is ideally that twice as many pulses per second are emitted from the voice organs. The distance between adjacent harmonics in the spectrum will be doubled, and the number of harmonics up to a certain fixed frequency limit will thus be halved. If a specific formant, for instance the first, comes close to the *6th* harmonic at the lower pitch, it will be the *3rd* harmonic that comes closest to the same formant in the case of the higher pitch. The concepts of formant frequency and harmonic number should not be confused." (Fant, 1960, p. 20)

See also Chapters M4 and M6.

Vowel-specific formants

"Usually vowels can be quite well characterized in terms of the frequencies of just the first and second formants, but the third formant should also be measured for high front vowels and for r-colored vowels." (Ladefoged, 2003, p. 105)

Age- and gender-specific formants

"The length of the pharyngeal-oral tract depends on the physical size of the speaker. The length affects the frequency locations of all of the vowel formants; this fact helps us to predict where the formant peaks in the spectrum will appear for men, women, and children. A very simple rule relates the frequencies of the formants to the overall length of the tract from glottis through lips. The rule for this relation is:

Length Rule. The average frequencies of the vowel formants are inversely proportional to the length of the pharyngeal-oral tract. In other words, the longer the tract, the lower are its average formant frequencies.

The neutral vowel formants for the average man, with an oral tract 17.5 cm in length, are at 500, 1500, 2500 Hz, and so on, with the lowest formant at 500 Hz and frequency spacing of 1000 Hz between all formants.

An easy way to remember the neutral formant frequencies is to think of the odd numbers 1, 3, 5, 7, 9, and so on, because the formant frequencies of a uniform tube that is closed at one end and open at the other, like the pharyngeal-oral tract, are always odd multiples of the frequency of the lowest formant. For example, begin with the basic formant frequency, 500 Hz, as the unit or 1; then the formant frequencies above that are 500 × 3 = 1500 Hz, 500 × 5 = 2500 Hz, and so on. This method, calculating the formants above F1 as multiples of F1, applies only as a model of a neutral tract shape.

The pharyngeal-oral tract length of an infant is approximately half the length of that of a man. Therefore, following our Length Rule about formant frequency locations, the formants of a neutral-shaped infant tract in relation to a man's would be at frequency locations that are a factor of the reciprocal of ½, or twice those of the man. On this basis the infant formant locations for a neutral vowel would be as follows: F1 is 500 × 2 = 1000 Hz, F2 is 1500 × 2 = 3000 Hz, F3 is 2500 × 2 = 5000 Hz, and so on.

Following the same procedure, a woman's vocal tract, on the average, is about 15% shorter than that of a man. The ratio corresponding to this amount of shortening is approximately 5/6. The reciprocal of 5/6 is 6/5, which is equal to a factor of 1.20, which, when multiplied by the man's neutral formant frequencies, gives the woman's values of 20% higher: F1 is 500 × 1.2 = 600 Hz, F2 is 1500 × 1.2 = 1800 Hz, F3 is 2500 × 1.2 = 3000 Hz, and so on. […]

The Length Rule tells us approximately where we may find the formants for the very young as well as for older, larger persons. However, the neutral locations of F1 and F2 for an individual are also affected by the length proportions of the vocal tract between the oral and pharyngeal cavities (Fant, 1973, Chapter 4). In general, the location and spacing of formants F3 and above are more closely correlated with length of vocal tract than for F1 and F2. The average locations of F1 and F2 for an individual are also affected somewhat by language environment and training." (Pickett, 1999, pp. 38–40)

See also Chapter M5.

M2 Prevailing Empirical References

Illustration: including radiation factor/radiation impedance

For a more differentiated graphic illustration, showing a 12db/octave slope of the source and a 6dB/octave intensity increase because of the radiation impedance, see Ladefoged (1996, p. 104), Figure 7.7 and the related comment: “Figure 7.7 shows a source-filter view of the production of a vowel. The spectrum of the glottal pulse is shown on the left of the figure. In this case we have taken the vocal folds to be vibrating at 100 Hz, so the components are at 100 Hz intervals. To the right of the spectrum is the set of curves specifying the vocal tract response. The output of the vocal tract can be regarded as the input to another box entitled ’radiation factor,’ which we must now take into account. […] these vibrations […] inside the mouth […] are not themselves the variations in air pressure that we hear. The air in the vocal tract vibrates so that the air particles at the open end between the lips move backward and forward. It is these movements that start the air outside the lips vibrating. The air between the lips acts like a piston, a source of sound producing variations in air pressure that radiate out from the lips just as the variations in air pressure radiate out from a source of sound such as a tuning fork. The movements of this piston of air are more effective in causing variations in pressure in the surrounding air at some frequencies than others. The higher the frequency, the greater the response of the surrounding air to the action of the air vibrating in the vocal tract. This effect, which we have termed the ‘radiation factor’ (‘radiation impedance’ is the term used in more technical books), can be regarded as a kind of filter that boosts the higher frequencies by 6 dB per octave. The curve representing the radiation factor is shown above the third box in figure 7.7.

The output produced at the lips depends on the vocal cord source, the filtering action of the vocal tract, and the further modifications produced by the radiation factor. Normally the vocal cord source is the same for each vowel, apart from variations of pitch. The vocal folds may be vibrating at 100 Hz, or at 200 Hz, as in the examples we have been considering, or at any other frequency in the range of the human voice. But irrespective of the fundamental frequency, the spectral slope of the cord pulse will usually be approximately –12 dB per octave. The filtering action of the vocal tract will be different for each position of the vocal organs, thus producing formants (peaks in the resonance curve) at different frequencies. The spectrum of the waveform beyond the lips (shown on the right of figure 7.7) will have peaks in regions which depend on the filter characteristics of the vocal tract. The general slope of the output spectrum will be influenced by the slope of the spectrum of the glottal pulse (–12 dB/octave) and the radiation factor (+6 dB/octave). Taken together these two slope factors account for a –6 dB/octave slope in the output spectrum. The major characteristics of the output spectrum – the formant peaks – are superimposed on this general slope. They are primarily dependent on the filtering characteristics of the vocal tract." (Ladefoged, 1996, pp. 104–105)

Formant statistics by Fant et al.

With regard to the study of Fant (1959; see Section 2.1, Table 3), see also the later study of Fant, Henningsson, and Stalhammar (1969) concerning statistical formant patterns for long Swedish vowels produced by men.

Formant statistics for Standard German

Older studies concerning formant patterns of German vowels were published by Jørgensen (1969), Iivonen (1970, 1986), Rausch (1972), Wängler (1981), and Ramers (1988). For further indications of formant statistics for Standard German, see the online digital version of the materials.

Formant statistics for other languages

For further indications of formant statistics of other languages, see also the online digital version of the materials.

Materials Part II

The second part of the Materials section contains selected excerpts from the literature as well further indications and discussions relating to the second part of the main text.

M3 Vowels and Number of Formants

Formant merging

"If you know you are analyzing a low back vowel, don't be surprised to find one thick bar on the spectrogram that really corresponds to two formants close together below 1'000 Hz." (Ladefoged, 2003, p. 114)

Referring to vocalisations of /ɔ/ as in caught: "When the formants are close together […] neither the wide- nor the narrowband spectrum gives a good indication of the formant frequencies. […] The first two formants appear as a single peak below 1'000 Hz. Their frequencies cannot be determined from these spectra." (Ladefoged, 2003, pp. 119–120)

Spurious formant

"Sometimes it is not immediately obvious whether a particularly wide band represents one formant or two. Figure 5.8 is a spectrogram of the word *bud,* spoken by a female speaker of Californian English. There is a wide band below 1,000 Hz, but is this one formant or two formants close together as in Figure 5.7? Noting that there is a clear formant at about 1,500 Hz in Figure 5.8, and additional formants higher, we must take it that there is only a single formant below 1,000 Hz. It seems that there is some kind of extra formant near the first formant, making this dark bar wider. From the evidence of this one vowel it is impossible to say whether the additional energy is above or below the first formant. Further analysis of this speaker's voice showed that there was often energy around the 1,000 Hz region, irrespective of the vowel. This spurious formant is not connected with the vowel quality, but is simply a characteristic of the particular speaker's voice. This is a good example of the necessity of looking at a representative sample of a speaker's voice before making any measurements of the formants." (Ladefoged, 2003, pp. 114–115)

"Flat" vowel spectra

"Flat-spectrum stimuli, consisting of many equal-amplitude harmonics, produce timbre sensations that can depend strongly on the phase angles of the individual harmonics. For fundamental frequencies in the human pitch range, many realizable timbres have vowel-like perceptual qualities. This observation suggests the possibility of constructing intelligible voiced speech signals that have flat-amplitude spectra." (Schroeder & Strube, 1986)

M4 Vowels and Fundamental Frequency

Independence of formants and fundamental frequency

“Obviously, formant frequency is independent from the fundamental frequency […] Changes in formant frequency are due to changes in the shape of the vocal tract cavity or cavities; changes in pitch frequency to stretching of the vocal cords. If the two physiological events are independent, so are the acoustic results of each event […].” (Delattre, 1958/1980)

“[…] when a complex wave consists of a damped waveform repeated at regular intervals, the component frequencies will always have the same relative amplitudes as the corresponding components in the continuous spectrum representing the isolated occurrence of the damped wave. Consequently, altering the rate at which the vocal folds produce pulses will affect the fundamental frequency of the complex wave; but it will not alter the formants (the peaks in the spectrum), which correspond to the basic frequencies of the damped vibrations of the air in the vocal tract. It is in this sense that we may say that the formants of a sound are properties of the corresponding mouth shape. […] the formants which characterize a given vowel irrespective of the rate at which pulses are produced by the vocal cords […]

We saw in Chapter 6 that the pitch of a sound depends mainly on the fundamental frequency. Accordingly, when there is a variation in the rate at which pulses are produced by the vocal cords, there will be a change in the pitch of the sound (although there will be no change in the formants, and hence no change in the characteristic vowel quality). It is usually possible to alter the pitch of a vowel sound without altering its characteristic quality, because each of these factors is controlled by a separate physiological mechanism. As we have seen, the pitch depends on the action of the vocal cords, and the characteristic quality depends largely on the formants, which have certain fixed values for each particular shape of the vocal tract.” (Ladefoged, 1996, pp. 98–99)

See also the citation of Hillenbrand (n.d.) in Chapter M6.

“Undersampling” the formants I: formants at middle and high fundamental frequencies

“According to the undersampling account of the effects of f0 on vowel identifiability, the sparser distribution of harmonics at high f0s yields poorer definition of the peaks and valleys in the spectral envelope, creating a more ambiguous stimulus.” (Diehl, Lindblom, Hoemeke, & Fahey, 1996)

“However, in this range of frequency (500 to 1000 Hertz), you could not tell apart different vowels anyway, because the harmonics of the voice are so far apart that they are not ‘sampling’ the locations of the formants enough for you to tell where the formants lie. Therefore operatic writers only put words intended to be intelligible in the lower part of a soprano’s range.” (Moore, 2006, p. 11)

“Oversinging” the first formant

“For the U it is also by no means easy to find the pitch of the resonance by a fork, as the smallness of the opening makes the resonance weak. Another phenomenon has guided me in this case. If I sing the scale from *c* upwards, uttering the vowel U for each note, and taking care to keep the quality of the vowel correct, and not allowing it to pass into O, I feel the agitation of the air in the mouth, and even on the drums of both ears, where it excites a tickling sensation, most powerfully when the voice reaches *f.* As soon as *f* is passed the quality changes, the strong agitation of the air in the mouth and the tickling in the ear cease. […] The resonance of the mouth for U is thus fixed at *f* with more certainty than by means of tuning forks. But we often meet with a U of higher resonance, more resembling O, which I will represent by the French Ou. Its proper tone may rise as high as *f’*.” (von Helmholtz, 1885/1954, p. 110; *c* = 131 Hz, *f* = 175 Hz, *f’* = 349 Hz)

“Above *f’,* the characterization of U becomes imperfect even if it is closely assimilated to O. But so long as it remains the only vowel of indeterminate sound, and the remainder allow of sensible reinforcement of their upper partials in certain regions, this negative character will distinguish U. On the other hand a soprano voice in the neighbourhood of *f”* should not be able to clearly distinguish U, O, A; and this agrees with my own experience.” (von Helmholtz, 1885/1954, p. 114; *f”* = 699 Hz)

“It is reasonable to assume […] that it is impossible to produce recognizable vowels at musical pitches very much higher than their first formants. […]

The following table is offered as a practical guide: Vowels start seriously losing intelligibility when the fundamental reaches these frequencies:

(i u y)	350 cps (roughly middle F)
(e o ø)	450 cps (roughly middle A)
(ɛ ɔ œ)	600 cps (roughly high D)
(æ ɑ a)	750 cps (roughly high G)”

(Howie & Delattre, 1962)

"[…] only very few correct identifications of isolated vowels can be expected when fundamental frequency reaches or exceeds the usual first formant of a vowel." (Hollien, Mendes-Schwartz, & Nielsen, 2000)

"[…] vowel identifiability is inevitably compromised once f_0 exceeds R_1 […]" (Joliveau, Smith, & Wolfe, 2004)

"We have seen that female singers gain considerably in sound level by abandoning the formant frequencies typical of normal speech when they sing at high pitches. At the same time, F1 and F2 are decisive to vowel quality. This leads to the question of how it is possible to understand the lyrics of a song when it is performed with the 'wrong' F1 and F2 values. Both vowel intelligibility and syllable/text intelligibility can be expected to be disturbed. This aspect of singing has been studied in several investigations.

As a thought-provoking reminder of the difficulties in arranging well-controlled experimental conditions in the past, an experiment carried out by the German phonetician Carl Stumpf (1926) may be mentioned. He used three singer subjects: a professional opera singer and two amateur singers. Each singer sang various vowels at different pitches, with their backs turned away from a group of listeners who tried to identify the vowels. The vowels that were sung by the professional singer were easier to identify. Also, overall, the percentages of correct identifications dropped as low as 50% for several vowels sung at the pitch of G5 (784 Hz).

Since then, many investigations have been devoted to intelligibility of sung vowels and syllables (see, e.g. Benolken & Swanson, 1990; Gregg & Scherer, 2006; Morozov, 1965). Figure 12 gives an overview of the results in terms of the highest percentage of correct identifications observed in various investigations for the indicated vowels at the indicated pitches. The graph shows that vowel intelligibility is reasonably accurate up to about C5 and then quickly drops with pitch to about 15% correct identification at the pitch of F5. The only vowel that has been observed to be correctly identified more frequently above this pitch is /a/. Apart from pitch and register, larynx position also seems to affect vowel intelligibility (Gottfried and Chew, 1986; Scotto di Carlo and Germain, 1985).

Smith and Scott (1980) strikingly demonstrated the significance of consonants preceding and following a vowel. This is illustrated in the same graph. Above the pitch of F5, syllable intelligibility is clearly better than vowel intelligibility. Thus, vowels are easier to identify when the acoustic signal contains some transitions (Andreas, 2006). Incidentally, this seems to be a perceptual universal: changing stimuli are easier to process than are quasi-stationary stimuli.

The difficulties in identifying vowels and syllables sung at high pitches would result both from singers' deviations from the formant frequency patterns of normal speech and from the fact that high-pitched vowels contain few partials that are widely distributed over the frequency scale, producing a lack of spectral information.

In addition, a third effect may contribute. Depending on phonation type, the F0 varies in amplitude. At a high pitch, F1 may lie between the first and the second partial. Sundberg and Gauffin (1982) presented synthesized, sustained vowel sounds in the soprano range and asked subjects to identify the vowel. The results showed that an increased amplitude of the F0 was generally interpreted as a drop in F1." (Sundberg, 2013, pp. 86–88)

"Grade" of vowels

As discussed in Sections 4.1 and 4.2, prevailing theory gives reason to assume that a general but also discontinuous relationship exists between the intelligibility of vowel sounds and their fundamental frequency: accordingly, vowel sounds at lower fundamental frequencies would, as a rule, be more intelligible than vowel sounds at higher frequencies, but vowel intelligibility would also depend upon the respective relationships between fundamental frequency, harmonic spectrum and the vowel-specific formant pattern (as given in formant statistics).

Concerning the former, consider the following model cases:

- Comparison of two sounds of /ɛ/ produced by a woman at F0 of 200 and 400 Hz, related to a common formant pattern F1–F2 = 600–2000 Hz (compare Section 2.2, the formant statistics for Standard German); F1 will be "undersampled" for the sound at higher F0, i.e. F1 lying in between the first and the second harmonics, whereas for the first sound, the third harmonic matches with F1 indicating a "sampled" formant pattern F1–F2 as a better condition for vowel perception.
- Comparison of two sounds of /ɔ/ produced by a woman at F0 of 285 and 340 Hz, related to a common formant pattern F1–F2 = 570–1140 Hz (compare Section 2.2, the formant statistics for Standard German); F1–F2 will be "undersampled" for the sound at higher F0, i.e. F1 lying in between the first and the second, and F2 lying in between the third and the fourth harmonics, while for the first sound, the second and the fourth harmonics match with F1 and F2.
- And so on.

Concerning the latter, consider the following model cases:

- Comparison of two sounds of /i/ produced by a woman at F0 of 200 and 300 Hz, related to a common formant pattern F1–F2 = 300–2700 Hz (compare Section 2.1, the formant statistics of Peterson and Barney, 1952); F1 and F2 will be "undersampled" for the sound at lower F0, with F1 lying in between the first and the second, and F2 lying in between the twelfth and the thirteenth harmonics, while for the second sound, the first and the ninth harmonics match with F1 and F2 indicating a "sampled" formant pattern F1–F2 as a better condition for vowel perception.
- Comparison of two sounds of /ɑ/ produced by a woman at F0 of 270 and 330 Hz, related to a common formant pattern F1–F2 = 660–990 Hz (compare Section 2.1, the formant statistics of Fant, 1959); F1 and F2 will be "undersampled" for the sound at lower F0, i.e. F1 lying in between the second and the third, and F2 lying in between the third and the fourth harmonics, while for the second sound, the second and the third harmonics match with F1 and F2.
- Comparison of two sounds of /u/ produced by a woman at F0 of 200 and 300 Hz, related to a common formant pattern F1–F2 = 300–900 Hz; F1 and F2 will be "undersampled" for the sound at lower F0, i.e. F1 lying in between the first and the second, and F2 lying in between the fourth and the fifth harmonics, while for the second sound, the first and the third harmonics match with F1 and F2.
- And so on.

"Undersampling" the formants II: resonances and formants

If a basic distinction is made between the resonances of the vocal tract and the formants of the vowel sound produced, strictly speaking, only resonances can be undersampled in the sense of a large frequency distance between harmonics and no harmonic matching an existing resonance frequency. Formants in their turn are always a result of a method of measurement.

M5 Formant Patterns and Speaker Groups

Thesis of age- and gender-related differences in vowel-specific format patterns

"Because of shorter cavity lengths females […] have larger average formant spacings and higher average formant frequencies than males. Similar relations hold for children compared with adults […]." (Fant, 1960, p. 21)

"Men, women, and children generally differ with respect to average vocal tract length, which is significant for the formant frequencies, as we know. For this reason, the same vowel is usually represented by different formant frequencies in men, women, and children.

[…] average formant frequency differences between male and female adults are expressed as the percentages by which the three lowest formant frequencies of a given vowel in female adults exceed those in male adults (Fant, 1975). […] they vary considerably between vowels, particularly for the lowest two formants. […] these percentage differences occur similarly in various languages. The first formant frequency shows a maximum percentage difference in the open /a:/ vowel of the Italian word caro. The second formant frequency shows high values for all front vowels. The difference, averaged over the entire set of vowels, amounts to 12%, 17%, and 18% for the three lowest formants. Children's average formant frequencies are about 20% higher than those for female adults, or 32%, 37%, and 38% higher than those of male adults. Probably most of these differences are due to inequalities in the vocal tract dimensions between the various groups of speakers. Thus, younger children tend to have higher formant frequencies than older children because of their shorter vocal tracts.

If the proportions of the average female and male vocal tracts are compared, one finds that the female vocal tract is not merely a small-scale version of the male vocal tract. According to Nordstrom (1977), the average mouth length of a female adult is about 85% of that of the average male adult, while the female pharynx length is only 77% of the corresponding male value. In other words, the average female pharynx is much shorter than the average male pharynx, while the average difference is smaller with regard to the mouth.

If one computes the formant frequency differences that would result from these dissimilarities in the mouth and pharynx proportions between adult males and females, one finds a discrepancy between prediction and reality; the differences that have been found in the dimensions do not explain the actual formant frequency differences, according to

Nordstrom (1977). The reason for this is not well understood. The existence of sex dialects, or ‘sexolects’, cannot be excluded; it is possible that females and males use a slightly different articulation of some vowels. The reason may be hidden in the largely unknown processes used by our sense of hearing and our brain in order to identify vowels.

We correctly infer that the actual reasons for the formant frequency differences between children and adult males and females are not understood in every detail. However, it is also interesting to see to what extent the voice timbre differences between these groups of speakers can be accounted for by the formant frequency differences. Colem (1976) has published an interesting investigation on this topic. In an experiment in which subjects tried to identify the sex of speakers by listening to the voice quality, he found that phonation frequency was a much more important factor than formant frequencies as illustrated in Figure 5.10; the average of the three lowest formant frequencies showed little or no correlation with maleness and femaleness in voice timbre. The faint trace of a correlation that appears to exist between the average of the three lowest formant frequencies and the perceived maleness or femaleness was due to an equally low correlation between phonation frequency and this formant frequency average.

It may be important to these results that the three lowest formant frequencies were not separated but were converted into an average in this investigation. It is not clear whether such an average catches all of the timbral voice differences between the sexes, and it is also possible that the results would have come out differently if the fourth formant had been included in the average; the higher the formant frequency, the more its frequency depends on nonarticulatory factors such as vocal tract length.

It seems clear that the perceptually most important difference in voice quality between the two sexes depends on phonation frequency rather than formant frequencies. The mean phonation frequency difference is almost one octave, which is much greater than the formant frequency difference. We realize that our brain is quite smart: it is more impressed by the great phonation frequency difference than by the small formant frequency difference when guessing the sex of a speaker.” (Sundberg, 1978)

Concerning indications of similar formant patterns for sounds of different vowels produced by speakers of different speaker groups, see, for example, the vowel synthesis experiment in Potter and Steinberg (1950), and the [e]–[ø] ambiguity reported by Fant, Carlson, and Granström (1974). See also the indications of similar F1–F2 for /U/ and /u/, and for /ʌ/ and /o/ in the statistics of Hillenbrand et al. (1995), comparing the patterns of women and men, and of children and men, respectively.

Questioning this thesis: von Helmholtz (1885), Potter and Steinberg (1950)

“ […] the proper tones of the cavity of the mouth are nearly independent of age and sex. I have in general found the same resonances in men, women, and children. The want of space in the oral cavity of women and children can be easily replaced by a great closure of the opening, which will make the resonance as deep as in the larger oral cavities of men.” (von Helmholtz, 1885/1954, p. 105)

Note that this statement by von Helmholtz stands in contradiction to his self-experiment, on the basis of which he concluded a vowel-specific resonance for U at 175 Hz (see Chapter M2): particularly for the speech of children, the fundamental frequency is substantially above 175 Hz, not allowing for a production of U, if vowel-specific resonances are independent of age and gender.

“Audible Form and Vowel Identification: Form or pattern of the formant positions appears to be important in discriminating between sounds. One of the first results found was that, for a given vowel sound, the actual formant frequency positions for a man’s voice differ markedly from those for a woman’s or a child’s voice. To illustrate this difference the frequencies of the formants in the vowel sound [æ] as spoken by a man, a woman and a child are shown on the left hand side of Fig. 5 by short horizontal lines designated F1, F2, F3. […] Listening tests indicate that these three sounds are identified as the same vowel. Yet the values of the formant frequencies are quite different. Certainly we cannot regard a vowel as completely specified by fixed regions of energy concentration. […]

If we view the formant positions in relation to positions of fundamental frequency, they fall into better alignment. This suggests that the fundamental frequency of the voiced sounds might offer a means for normalizing the formant positions. However, this seems a dubious possibility because the formant positions for a given vowel are probably directly related to the dimensions of the vocal cavities and only incidentally related to fundamental frequency. For example, whispered vowels can be identified readily. Also there may well be cases of high fundamental frequency with large vocal cavities, and vice versa, that would need to be considered.

To obtain preliminary information on the question of how pitch affects vowel identification we have synthesized sounds having the same formant outlines but different fundamental frequencies. One such case is illustrated in Fig. 6. The two upper charts show the spectra for the [æ] (had) sounds of Fig. 5, for the adult male and child's voices. The fundamental frequencies are 109 and 264 cycles respectively. The lower chart shows an unnatural spectrum, namely, the adult male's formant outline with a fundamental frequency of 256 cycles, approximating that of the child's voice. This frequency was chosen so that the peaks of the formants would not be shifted markedly in position. Sounds corresponding to the three spectra were synthesized by means of a spectrum generator […].

The first two synthesized sounds were readily identified by ear as [ae] sounds. The third sound, however, was neither the man's nor the child's [ae]. It seemed to be somewhere between the child's [ae] and [ɛ]. This phonetic shift may indicate an association between fundamental frequency and formant position. But the shift could also arise if the ear assigns different pitch centers or positions to the energy concentrations representing the formants in the upper and lower cases.

The effects become more pronounced when the back vowels are used in such a comparison. Figure 7 shows spectra similar to the ones in Fig. 6, except that they are for the [ɑ] (father) sound.

In this case, the first two sounds were clear [ɑ's]. The third sound was more like a child's [ɔ] (awl) than the [ɑ] (father). Here there is also a question of association or actual shift in the ear's assignment of formant position. Still if one considers the bar positions of these sounds as illustrated in Fig. 8, there is some support for an association of fundamental frequency and formant position. […] We have seen that an increase in fundamental frequency seems to require that both bars be raised in frequency position to maintain the identification of a given vowel (Fig. 5). Hence, in the case of the [ɑ] sound, the combination of adult formants with the child's fundamental frequency shifts the sound toward the [ɔ]. It must be admitted, though, that the association of adult formants and child's fundamental frequency is an unnatural one giving sounds that do not correspond to any of the natural sounds." (Potter & Steinberg, 1950)

Exceptions in existing formant statistics

Although in formant statistics, the highest frequency values of vowel-specific formants are generally given for children, middle values for women and the lowest values for men, exceptions can be found. Some examples of such exceptions are listed below, ordered according to vowel quality. Abbreviations used are: “*” = values for the comparison of voiced vowel sounds, “**” values for the comparison of whispered vowel sounds; “SinSp” = values for the comparison of the sounds of a single male and a single female speaker as given in Fant (1959); “Av” = average values for a speaker group in the statistics of Fant (1959). Examples of single formants or formant patterns for which higher frequency values are given for men than for women:

/i/ F1*-F2*-F3* (Fant, 1959, SinSp); F1* (Fant, 1959, A), F1* (compare Pols, Tromp, & Plomp, 1973, van Nierop, Pols, & Plomp, 1973)

/y/ F1* (Fant, 1959, SinSp; marginal difference for F2*), F1* (compare Pols et al., 1973, Van Nierop et al., 1973

/e/ F1*-F2* (Fant, 1959, SinSp)

/ɵ/ F1*-F2*-F3* (Fant, 1959, SinSp); F1* (Fant, 1959, A)

/ɛ/ F2* (Fant, 1959, SinSp)

/æ/ F2* (Fant, 1959, A); F2** (Sharifzadeh, McLoughlin, & Russell, 2012)

/ɔ/ F1** (Sharifzadeh et al., 2012; marginal difference F2**; marginal differences also for F1*-F2*)

/o/ F1* (Fant, 1959, SinSp)

/ʊ/ F1*-F2* (Fant, 1959, SinSp); F1* (Fant, 1959, A); F1*, F1**-F2** (Sharifzadeh et al., 2012)

/u/ F1* (Fant, 1959, SinSp); F2* (Fant, 1959, A); F1* (compare Pols et al., 1973, Van Nierop et al., 1973); F1* (Zee, 2003); F1** (Sharifzadeh et al., 2012)

See also Hillenbrand et al. (1995) for slightly higher F1 values of /ʌ/ for women than for children.

“We have argued […] that for the vowels /u/, /i/ and /y/ as well, F1 can be chosen so that its average value is higher for female speakers than for male speakers. However, F1 then becomes about equal to 2xF0 (490 Hz) which is much too high. The data on the vowels /u/, /i/ and /y/ do not confirm the usual upward shift of formant frequencies for female speakers. We do not suggest that the anomaly for these three vowels reflects the actual resonance frequencies of the vocal tract.” (van Nierop et al., 1973)

Zee (2003) found lower F1 for women than for men for the vowel /u/ when investigating formant frequencies of Cantonese vowels and comments his finding as follows: “In any case, it is not clear as to why the F1 value for [u] does not follow the general pattern.”

“In looking at the ranges for each vowel formant frequency for the male and female groups, the overlap between genders was considerable. In all cases, the highest formant value for the male group was markedly above the lowest formant value for the female group for each formant of both vowels. This would suggest that in some individual cases, the formants of a male speaker might be the same as, or even higher than, the formants of a female speaker.” (Gelfer & Bennett, 2013)

M6 Terms of Reference, Methods of Formant Estimation

Terms of reference

“Formant […]. A concentration of acoustic energy within a particular frequency band, especially in speech. Any given configuration of the vocal tract produces resonance, and hence formants, in certain frequency ranges. During the articulation of a vowel, these formants show up prominently in a sound spectrogram as thick dark bars; the three lowest of these, known as first, second and third formants (F1, F2 and F3) are highly diagnostic, and vowels are distinguished acoustically by the positions of these formants.” (Trask, 1996, p. 148)

“Some refer to a formant as a peak in the acoustic spectrum. In this usage, a formant is an acoustic feature that may or may not be evidence of a vocal tract resonance. Others use the term formant to designate a resonance, whether or not actual empirical evidence is found for it.” (Kent & Read, 2002, p. 24)

“Resonances, formants and spectral peaks: Unfortunately, the meaning of the word ‘formant’ has expanded to describe two or three different things. Fant (1960) gives this definition: ‘The spectral peaks of the sound spectrum $|P(f)|$ are called formants.’ Resonance frequencies are then defined in terms of the gain function $T(f)$ of the tract by ‘The frequency location of a maximum in $|T(f)|$, i.e. the resonance frequency, is very close to the corresponding maximum in spectrum $|P(f)|$ of the complete sound.’ Fant then writes: ‘Conceptually these should be held apart but in most instances resonance frequency and formant frequency may be used synonymously.’ Benade (1976) uses a similar definition of formant: ‘The peaks that are observed in the spectrum envelope are called formants.’ More recently, the acoustical properties of the vocal tract are often modelled using an all-pole autoregressive filter (Atal and Hanauer, 1971). For many voice researchers, formants now refer to the poles of this filter model. To others, formant means the resonance frequency of the tract. Finally, many researchers, particularly in the broader field of acoustics, retain the original meaning: a broad peak in the spectral envelope of a sound (of a voice, musical instrument, room etc.). The original meaning of formant is also retained, almost universally, when discussing the singers formant and actors formant: these terms refer to a peak in the spectral envelope around 3 kHz (discussed below). As Fant observes, while these uses are often closely related, they are conceptually quite distinct. Further, the resonant frequency, the pole of the fitted filter function and the peak spectral maximum need not coincide. Moreover, it is now possible to measure resonances of the vocal tract quite independently of the voice. Consequently, it is sometimes essential to make a clear distinction among a resonance frequency (a physical property of the tract), a filter pole (a value derived from data processing) and a spectral peak (a property of the sound)." (Wolfe, Garnier, & Smith, 2009)

"Formant is used by James Jeans (1938) to mean the collection of harmonics of a note that are augmented by a resonance.
Formant was defined by Gunnar Fant (1960): **'The spectral peaks of the sound spectrum |*P(f)*| are called *formants'.***
Benade (1976) writes: **'The peaks that are observed in the spectrum envelope are called formants'.**

In its standards for acoustical terminology, the Acoustical Society of America (1994) defines formant thus: **"Of a complex sound, a range of frequencies in which there is an absolute or relative maximum in the sound spectrum.** Unit, hertz (HZ). NOTE-The frequency at the maximum is the formant frequency." (Wolfe, n.d.)

"Does it matter? For the voice, a resonance at a frequency R(i) gives rise to a spectral maximum at frequency F(i) which may produce in a filter model a pole at frequency P(i). Usually, the three frequencies have similar values. However, as Fant observed, they are conceptually distinct. Let's take some examples:

– Consider a vocal tract with a resonance at 500 Hz, which is being excited by the larynx producing a fundamental frequency of 1 kHz (near C6, the high C for sopranos). There is no spectral maximum at 500 Hz. In this case there is a resonance R1 but no corresponding spectral peak F1. Here of course the difference does matter.
– Consider the singers formant or singing formant, a broad band of enhanced power noticed in the spectral envelope of classically trained male singers (and possible others) in a range. Sundberg (1974) attributes this formant to a clustering of the third, fourth and fifth resonances of the vocal tract. Here, where three resonances are thought to give rise to one formant, the distinction between formant and resonance is important.
– Consider a glottal source with a negative spectral slope, input to a vocal tract that (including radiation impedance) has a resonance at R1. The peak in the spectral envelope of the radiated sound in this case has a frequency less than R1. In this case, if one is estimating the spectral peak from the harmonic spectrum

of the output voice, the difference between the two is less than the precision of the estimation, so the distinction is usually not important.

- Consider a musical wind instrument, whose bore radiates weakly below some frequency f, and which is excited by a reed or lip valve whose spectral envelope falls with frequency. Here the output sound has a spectral envelope peak that has nothing at all to do with the resonances of the bore.
- Consider this quote, from Stevens and House (1961): 'When resonant frequencies are sufficiently close, however, they are not necessarily identical with the frequencies of the peaks in the spectrum. For example, when two resonances with bandwidths of about 100 cps are about 100 cps apart, the spectrum envelope may show only one prominence: the frequency of the peak will be somewhere between the two resonant frequencies. In the discussion that follows, the levels of the resonances will be defined to be the levels of the spectral envelope at the frequencies of the resonances (rather than at the spectral peaks).'

In our laboratory, the distinction is important. We routinely measure the resonances independently of the voice (Epps et al, 1997; Dowd et al, 1997; Joliveau et al, 2004a, b). We are often interested in comparing formants and resonances.

What to do? Our preference would be to retain the original meaning for the word formant. We prefer to say 'A resonance at frequency Ri gives rise to a formant at frequency Fi. This may be modelled by a filter with a pole at frequency Pi'. While acousticians will broadly agree with this use, some members of the speech research and modelling community may not. We therefore suggest that, when discussing the voice, the word formant should be defined, to make it clear which meaning is intended. In principle, one could consider abandoning the word. However 'broad peak in the spectral envelope' is a long phrase, so it is useful to retain formant for that reason.

[…]

Whatever your choice of definition, you should make it clear. And, in literature and in discussions, prepare for some confusion. For instance, some researchers who use formant to mean resonance will also talk about 'formant level'. When such people then talk of 'formant level', or say that the second formant is 10 dB lower than the first, I suspect that they refer to the amplitude of a peak in the sound spectrum. In a scientific talk, I have heard the sentence: 'Trained sopranos tune the first formant near the note sung, but they usually don't have a strong singer's formant'. When that speaker said 'first formant' he presumably meant ‘first resonance’ and when he said ‘singer’s formant’ he meant a spectral peak probably due to two or more resonances. So we have the same person using the word in two of its three different meanings in the one sentence.” (Wolfe, n.d.)

“With regard to airway resonances, historical precedence and current usage of terminology are also slightly at odds. Joe Wolfe and colleagues suggest that the symbol R be used to stand separate from the symbol F for formant (Wolfe, 2014). The distinction is being made because a formant was originally defined as a peak in the output spectrum envelope radiated from the mouth (Hermann, 1894, 1895; Russell, 1929; Fant, 1960, p. 20). A similar definition appears in the current ASA standard of acoustic terminology (Acoustical Society of America, 2004), namely, that a formant is ‘a range of frequencies in which there is absolute or relative maximum in the sound spectrum. The frequency at the maximum is the formant frequency.’ As such, a formant involves both the source and the filter. However, as speech analysis and synthesis have progressed in a half century, the definition has not been universally maintained. Fant (1960, pp. 20, 53) defined formants as the poles of the transfer function of the supraglottal vocal tract, and labeled the pole frequencies F1, …, Fn and their bandwidths B1, …, Bn. He was followed in this path by many authors, such as Titze (1994, p. 156) or Stevens (1998, p.131). It is noteworthy that Flanagan (1965, p. 57) was aware of the dual definition (and possible evolution) by using the term ‘formant resonance.’ While Benade (1976) maintained the definition of ‘peaks in the spectral envelope of the radiated sound,’ Badin and Fant (1984) computed formant frequencies and bandwidths on the basis of x-ray area function resonances of the supraglottal vocal tract, not peaks in the output spectrum envelope. Story et al. (1996) did similar calculations based on magnetic resonance imaging (MRI). Differentiation between the formant frequencies and resonance frequencies of the vocal tract can be found in some papers comparing measurements from phonation (formants) to those derived from vocal tract impedance measurements or from calculations based on MRI or computer tomography (CT) data (resonance frequencies) (e.g., Stoffers et al., 2006; Vampola et al., 2013).

What is relevant here for nomenclature and symbolic notation is that the letter R is easily distinguishable from the letter F or f, both in speaking and writing. Hence, it is useful as a subscript to separate source and filter symbols. Discussion can continue on whether or not a formant is a meaningful representation of any particular resonance. Some authors describe resonances pertaining to the supraglottal airway only (assuming no coupling to the glottal or subglottal system), while others describe the net effect of complex interactions of multiple resonators above, below, and within the larynx. […]

Unfortunately, the common definition between a formant and a resonance is yet to be established." (Titze et al., 2015)

Note that Titze et al. (2015) propose a new and consistent terminology for the frequencies, magnitudes and bandwidths of harmonics, resonances and formants.

"**Spectrum Envelope:** The term **spectrum envelope** refers to an imaginary smooth line drawn to enclose an amplitude spectrum. Figure 3-17 shows several examples. This is a rather simple concept that will play a very important role in understanding certain aspects of auditory perception. For example, we will see that our perception of a perceptual attribute called **timbre** (also called **sound quality**) is controlled primarily by the shape of the spectrum envelope, and not by the fine details of the amplitude spectrum. The examples in Figure 3-17 show how differences in spectrum envelope play a role in signaling differences in one specific example of timbre called **vowel quality** (i.e., whether a vowel sounds like /i/ vs. /a/ vs. /u/, etc.). For example, panels *a* and *b* in Figure 3-17 show the vowel /ɑ/ produced at two different fundamental frequencies. (We know that the fundamental frequencies are different because one spectrum shows wide harmonic spacing and the other shows narrow harmonic spacing.) The fact that the two vowels are heard as /a/ despite the difference in fundamental frequency can be a ttributed to the fact that these two signals have similar spectrum envelopes. Panels *c* and *d* in Figure 3-17 show the spectra of two signals with different spectrum envelopes but the same fundamental frequency (i.e., with the same harmonic spacing). As we will see in the chapter on auditory perception, differences in fundamental frequency are perceived as differences in pitch. So, for signals (a) and (b) in Figure 3-17, the listener will hear the same vowel produced at two different pitches. Conversely, for signals (c) and (d) in Figure 3-17, the listener will hear two different vowels produced at the same pitch." (Hillenbrand, n.d., pp. 16–17)

Methods of formant estimation I: general aspects

"The difficulties involved in measuring formant frequencies have been well known since the early days of the spectrograph, and involve errors related to (i) the ambiguous definition of the object to be measured, (ii) spectral features of the speech wave, (iii) intermodulation distortion, (iv) the spectrographic record, and (v) the measuring procedure:

- A formant is seen both as a spectral prominence in the speech wave and as a filter property of the vocal tract; a definition comprising both components contradicts itself; a definition embracing just the first component presupposes that the relevant information for speech perception is immediately available in the speech wave; a definition based on the second part alone is production oriented and sees the true formant value as a vocal tract pole frequency that is being measured from its (sometimes poor) reflection in the speech wave.
- The resolution of the spectral envelope depends on the interval between the partials, which is equal to the fundamental frequency; a spectral peak may be asymmetrical within the formant band; individual spectral peaks become less well defined as they approach each other or as their bandwidths increase. […]

Lindblom's advice is thus still valid today. It is still necessary to apply one's knowledge and experience of speech production and expected envelope shapes to the problem of how to select samples to measure and where to look for spectral peaks." (Wood, 1989, referring to Lindblom, 1962)

"[…] At this point we should remember that an LPC filter lumps together several aspects of speech production […]. An LPC spectrum represents not only the formant frequencies due to the resonances of the vocal tract but also the effects of the lip radiation and the spectrum of the pulse from the vocal folds. Nevertheless, the peaks in the LPC spectrum are usually good indicators of the formant frequencies. Problems may arise when two formants are close together, in which case the spectrum may appear to have only a single peak corresponding to both of them, or when one formant has a lower amplitude, so that it appears as only a kink in the curve representing another formant. These problems lead us to another way of considering LPC analysis.

It is also possible to analyze an LPC expression so as to determine the exact frequencies corresponding to the poles (which, however, may not be exactly those of the formants in the vocal tract transfer function). For every pair of LPC terms we get a pair of numbers corresponding to the frequency and the bandwidth of a pole in the filter. We know […] that there will be a formant at 500Hz, 1,500 Hz, 2,500 Hz, and so on in a neutral vowel for a speaker with a vocal tract of 17.5 cm. In general, for such a speaker there will be one formant for every 1,000 Hz interval. So with a 10,000 Hz sample rate and an upper frequency limit of 5,000 Hz, we can expect to find five formants. This will require ten LPC terms. If we want to allow two further terms to account for higher formants that may be influencing the spectrum or a pole due to the glottal pulse shape, then we should make a twelve-point LPC analysis. If the speaker might have a shorter vocal tract so that we could only expect four formants below 10,000 Hz, then we could use a ten point LPC.

Choosing the right number of coefficients for an LPC analysis is somewhat of an art. If one chooses too many, the analysis will produce poles corresponding to spurious formants; if one chooses too few, formants may be lumped together because the higher formants or the glottal pulse may require more complex specification. The problem is compounded by the fact that an LPC analysis is equivalent to trying to model the spectrum using only poles, and there may be zeros (antiresonances) in the vocal tract transfer function. There certainly will be antiresonances in any vocal tract shape that contains the equivalent of a side tube, such as the oral cavity in the case of a nasal sound. LPC analysis is not reliable for nasalized vowels. A general rule of thumb for the number of coefficients is the sample rate in kHz plus 2, e.g. 10,000 Hz = 10 kHz plus 2 equals 12. But a better rule is to use several different analyses with different numbers of coefficients and see which gives the most interpretable results." (Ladefoged, 1996, pp. 210–212)

"Good spectrograms are a great help in determining where the formants are. This is often not as easy one might imagine. You have to know where to look for formants before you can find them. The best practical technique is to look for one formant for every 1,000 Hz. The vowel ə, for example, has formants at about 500, 1,500 and 2,500 Hz for a male speaker (all slightly higher for a female speaker). Other vowels will have formants up or down from this mid range. But there are exceptions to this general rule of one formant per 1,000 Hz. It would be more true to say that there is, on average, one formant for every 1,000 Hz. Low back vowels may have two formants below 1,000 Hz, but nothing between 1,000 and 2,000 Hz, and then the third formant somewhere between 2,000 and 3,000 Hz." (Ladefoged, 2003, pp. 113–114)

Methods of formant estimation II: methodological limits related to F0

"[…] in the case of female speech, formant analysis is extremely difficult. The fundamental frequency is so high that formants are often poorly defined. […] We had difficulties in determining the position of a formant in about 40% of the 300 vowel segments, if no a priori knowledge was used." (Van Nierop et al., 1973)

"[…] because formant frequencies are hard to determine when fundamental frequency is higher than about half of the frequency of the first formant." (Sundberg, 1987, pp. 124–125)

"Accurate measurement of formant frequencies is important in many studies of speech perception and production. Errors in formant frequency estimation by eye, using a spectrogram, or automatically, using linear prediction, have been reported to be as high as 60 Hz at $F0 < 300$ Hz. This exceeds the typical auditory difference limens (DLs) for formant frequencies and is also greater than some of the variation that one would like to study, e.g. the acoustic effects of varying vocal effort. The problem becomes substantially worse when F0 is as high as 500 to 600 Hz, which is not uncommon in the speech of women and children at high vocal efforts." (Traunmüller & Eriksson, 1997)

"Measurements of the frequency position of the formants, considered as the resonances of the vocal tract, are affected by substantial errors when F0 is as high as it is when people communicate over large distances. This holds for LPC-based methods as well as when using visual inspection of spectrograms." (Traunmüller & Erikkson, 2000)

"The problem is that it is difficult to determine reliably the resonance frequencies of the tract from the sound alone, using either spectral analysis or linear prediction, once F0 exceeds 350 Hz (Monson and Engebretson, 1983), and essentially impossible once F0 exceeds 500 Hz." (Joliveau et al., 2004)

"[…] it is difficult to determine unambiguously the frequencies of the resonances with a resolution much finer than f0/2." (Swerdlin, Smith, & Wolfe, 2010)

Methods of formant estimation III: "One wonders, for example, if the source-filter theory of speech production would have taken the same course of development if female voices had been the primary model early on."

"To a large extent, the early work in acoustic phonetics focused on the adult male speaker. There were a number of reasons for this focus, including social and technical factors. Only rather recently has the study of acoustic phonetics been broadened to encompass significant research on populations other than men. This is not to say that children and women were neglected altogether in the early history of acoustic speech research. Peterson and Barney's (1952) classic study included acoustic data on vowels for men, women and children, making it clear that acoustic values vary markedly with age and gender characteristics of speakers […].

The problem is that the research effort given to the speech of women and children has been on a smaller scale than that given to the speech of men. Consequently, there is a continuing need to gather acoustic data for diverse populations. The concentration on male speakers had several consequences, not all of which facilitated research on the speech of women and children. One consequence was the choice of an analyzing bandwidth (300 Hz for the 'wide-band' analysis) on early spectrographs that worked well enough for most adult male voices but was deficient for many women and children. The unsuitability of the analyzing bandwidth probably discouraged acoustic analyses of women's and children's speech.

The implications of the male emphasis may have reached even to theory; Titze (1989, p. 1699) commented, 'One wonders, for example, if the source-filter theory of speech production would have taken the same course of development if female voices had been the primary model early on.' Klatt and Klatt (1990, p. 820) remarked on the same point: 'informal observations hint at the possibility that vowel spectra obtained from women's voices do not conform as well to an all-pole [i.e. all formant] model, due perhaps to tracheal coupling and source/ tract interactions.' The acoustic theory for vowels […] assumed that the vocal tract transfer function is satisfactorily represented by formants (poles) and that antiformants (zeros) are required only for modifications such as nasalization. It is advisable to bear in mind that this theory is predicated largely on the characteristics of adult male speech and that it may have to be altered to account for the characteristics of both children and women." (Kent & Read, 2002, pp. 189–190)

Materials Part III

The third part of the Materials section presents exemplary series of vowel sounds and related acoustic analyses linked to the third part of the main text, including further indications on previously published data.

Note on the Method

Empirical basis

As mentioned in the introduction, the empirical basis of this treatise—and the basis of the series of vowel sounds selected for presentation here—consists of recordings from various areas of everyday life, the entertainment sector and art, that is, stage voices in music and straight theatre. (For an additional investigation of sounds of birds imitating human utterances, see Section M10.A.)

The recordings were collected over a time period of more than 20 years with different techniques related to different sound qualities, and they represent utterances of speakers different in age and gender, producing vowel sounds in different contexts, with different durations and different vocal efforts. However, such variation is not a shortcoming but an intention here, since this treatise focuses on the psychophysical question of the vowel (see the introduction and Section 13.7): given that different vowel sounds are perceived as being related to a single vowel quality—in contrast to the variation of other vocal sound characteristics—, which describable physical characteristic or which ensemble of physical characteristics may be said to represent that quality?

Concerning the acoustic characteristics of vowel sounds, the sound examples presented here were produced in isolation or in word context by native German or Swiss-German speakers, with a few exceptions, and the vowel qualities correspond to Standard German. Because of the psychophysical perspective adopted here, and because of the large fundamental frequency range considered—including many high-pitched vowel sounds produced in isolation or in the context of high-pitched speech by untrained children, women and men as well as by professional actresses and actors—, no principal difference is made between speaking and singing for isolated vowel sounds or extracted vowel nuclei and no corresponding indication is given in the figures which would relate to a classificatory system of modes of vowel production.—Acoustic analysis as well as perceptual identification relates to sounds produced in isolation or extracted as vowel nuclei from words.

Concerning the acoustic characteristics of pitch contours, the examples presented here (see Section 8.2) only concern contours of speech. Thereby, they relate to utterances of speakers of different languages (see the corresponding figure legends).

Whereas one part of these recordings forms the basis of single, published investigations undertaken in the past, which included listening tests, another part is unpublished and the corresponding recordings have not been subject to any further identification tests, apart from the identification by the author: in the course of creating this publication, for each of the sound series of a single figure presented in the Materials section, the author has evaluated the perceptual vowel quality of each sound separately. Moreover, only sounds are presented for which the intended and the perceived vowel quality correspond.

Acoustic analysis

With regard to the acoustic analysis of the sounds in general and to the calculation of fundamental and formant frequencies in particular, automatically calculated values using routines from the PRAAT Software (Boersma & Weenink, 2015) related to corresponding standard parameters are given in the figures of Chapters 7 to 10.

Acoustic analysis was conducted on isolated vowel sounds or on extracted vowel nuclei and concerned F0, spectrum, formant frequencies and LPC curve. (Note that the digital version of the Materials further includes pitch contour, spectrogram, formant tracks and comparison of three formant patterns and three LPC curves related to the three standard parameter settings for children, women and men.)

For longer vowel sounds, a middle sound fragment of 0.3 s, and for shorter sounds, a middle vowel nucleus excluding onset and offset was analysed.

The fundamental frequency of a sound fragment was calculated as average value using the Praat command To Pitch. Calculated values were perceptually crosschecked. If calculation errors occurred, the parameters “pitch floor” and “pitch ceiling” were adjusted.

The spectrum of a sound fragment was calculated as average spectrum for 0–5.5 kHz.

The formant frequencies of a sound fragment were automatically calculated as average values of LPC analysis using the Praat command To Formant (robust), with standard parameters according to the age and/or gender of the speaker and for a frequency range of 0–5.5 kHz. For illustration purposes, an LPC curve was calculated related to the analysis window in the middle of the sound fragment analysed.

Please note:

- Spectrum and numerical formant frequencies are calculated as averaged for the entire sound fragment analysed, but the LPC curve is related to a single window in the middle of the fragment. As a consequence, for a few sounds, the LPC filter curve does not correspond to the vowel spectrum and the numerical formant pattern.
- Because of automatic calculation and averaged values, calculated F1 for sounds of /i, y, u/ at middle and high fundamental frequencies is sometimes given as slightly below F0. In these cases, F1 can be estimated as roughly matching F0.
- A few of the calculated frequencies of the formants considered deviate so strongly from the sound spectrum and its amplitude minima and maxima that they are set in parenthesis or have been replaced by a rough estimation related to the spectrum. Exceptions are the sounds produced by birds for which the automatically calculated formant frequencies are given without consideration of their validity.

For longer recordings of speech (see Section M8.2), only the pitch contour was analysed and perceptually crosschecked. If major calculation errors occurred, the parameters "pitch floor" and "pitch ceiling" were again adjusted.

Illustrations

Each figure includes a series of vowel sounds (represented as vowel spectra) or examples of speech (represented as pitch contours). The subject matter of illustration is explained in the text and indicated in short form in the figure legend.

A vowel spectrum is given as the sound pressure level (SPL) in dB/Hz (y-coordinate) for a frequency range of 0–5500 Hz (x-coordinate). If, in the text, a vowel spectrum is considered in relation to calculated formants and/or to an LPC curve, this curve is also shown; if not, only the spectrum is presented. Below a spectrum, the following indications are given in the first line: figure number and number of the spectrum in the figure, vowel quality, fundamental frequency (F0), identification number of the speaker, gender of the speaker (w=woman/female, m=man/male), age group of the speaker (C=children, A=adults; note B=birds) and record number (R) of the recording in the database. For some figures, depending on the context of consideration, selected formant frequencies are indicated in addition in the second line.

Since the single vowel spectra relate to single vowel sounds, the vowel quality is given in square brackets. Note that in the figures, the vowel quality of /a–ɑ/ is represented by the character "a" with no further differenciation.

Pitch contours of speech are given as the pitch frequency in Hz (y-coordinate) over a time range in s (x-coordinate). Below a pitch contour, the following indications are given in the first line: figure number and number of the contour in the figure, [speech] as the mode of vocal expression and the content of recording, identification number of the speaker, gender and age group of the speaker and record number (R) of the recording in the database. In the second line, the overall F0 range for all contours of a speaker presented in a figure is given.

Note that the order of sound presentation in relation to vowel qualities and to F0 is not uniform throughout the entire Materials section; for each single section, this order accords to the subject matter illustrated and to the choice of the author.

Digital version of the Materials

More details on the method and, as mentioned, an extended documentation of the results of acoustic analysis are provided in the digital version of the Materials at:
http://www.phones-and-phonemes.org/vowels/acoustics/preliminaries

M7 Unsystematic Correspondence between Vowels, Patterns of Relative Spectral Energy Maxima and Formant Patterns

M7.1 Inconstant Number of Vowel-Specific Relative Spectral Energy Maxima and Incongruence of Vowel-Specific Formant Patterns

Figures 1 to 3 show examples of sounds of the back vowels /u, o/ and of /a–ɑ/ exhibiting only one relative spectral energy maximum within their vowel-specific frequency range ≤ c. 1.5 kHz. Each series corresponds to sounds produced by speakers of one speaker group (children, women, men). Note that for the sounds of /a–ɑ/, a dominant first harmonic is ignored here when interpreting relative spectral energy maxima. Note also that the examples 1, 3 and 4 in Figure 1 perceptually represent /ɔ/ rather than /a–ɑ/.

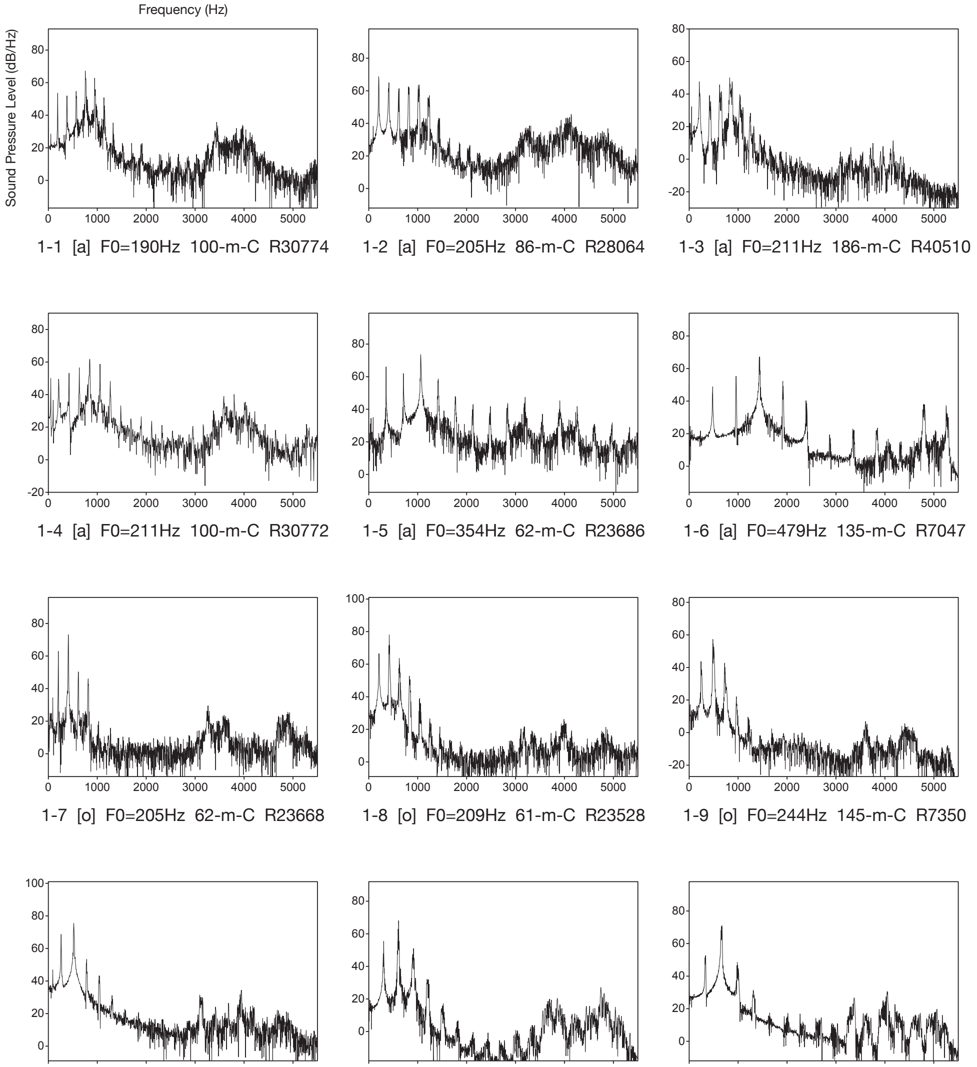
For each of the speaker groups and each of the three vowels in question, Figures 4 to 6 show three examples exhibiting two relative spectral energy maxima within their vowel-specific frequency range ≤ c. 1.5 kHz, as is usually assumed to be the "normal" case for sounds of these vowels.

Note that the spectra of the sounds of /u, o/ shown in Figures 1 to 3 cannot be interpreted as a general manifestation of "formant merging": if these spectra are compared with the spectra of the corresponding vowel sounds shown in Figures 4 to 6, the lowest spectral envelope peaks occur at similar frequency levels, given similar F0. Thus, the first spectral envelope peak of all sounds corresponds to the vowel quality in question, whereas the second spectral envelope peak for the sounds shown in Figures 4 to 6 may be related to an additional sound "colouring" that, however, does not possess vowel-differentiating value. Figure 7 illustrates this phenomenon by direct comparison of selected sounds of /u, o/ in Figures 1 to 3 with selected sounds of /u, o/ in Figures 4 to 6.

Figures 8 and 9 show examples of sound pairs of the vowels /i/ and /e/, each pair produced by speakers of one speaker group, for which differences in F0 and F1 are small but differences in the higher vowel-related spectral parts are substantial, up to F2 of the second sound matching or exceeding F3 of the first. Figure 10 shows more sound pairs of this kind but, in this case, comparing sounds of children and men, in order to document the phenomenon in its very extreme.

For earlier accounts, see Maurer, Landis, and d'Heureuse (1991), Maurer and Landis (1995).

Figure 1. Sounds of /a–ɑ, o, u/, produced by children, which exhibit only one relative spectral energy maximum within their vowel-specific frequency range ≤ c. 1.5 kHz.



1-1 [a] F0=190Hz 100-m-C R30774

1-2 [a] F0=205Hz 86-m-C R28064

1-3 [a] F0=211Hz 186-m-C R40510

1-4 [a] F0=211Hz 100-m-C R30772

1-5 [a] F0=354Hz 62-m-C R23686

1-6 [a] F0=479Hz 135-m-C R7047

1-7 [o] F0=205Hz 62-m-C R23668

1-8 [o] F0=209Hz 61-m-C R23528

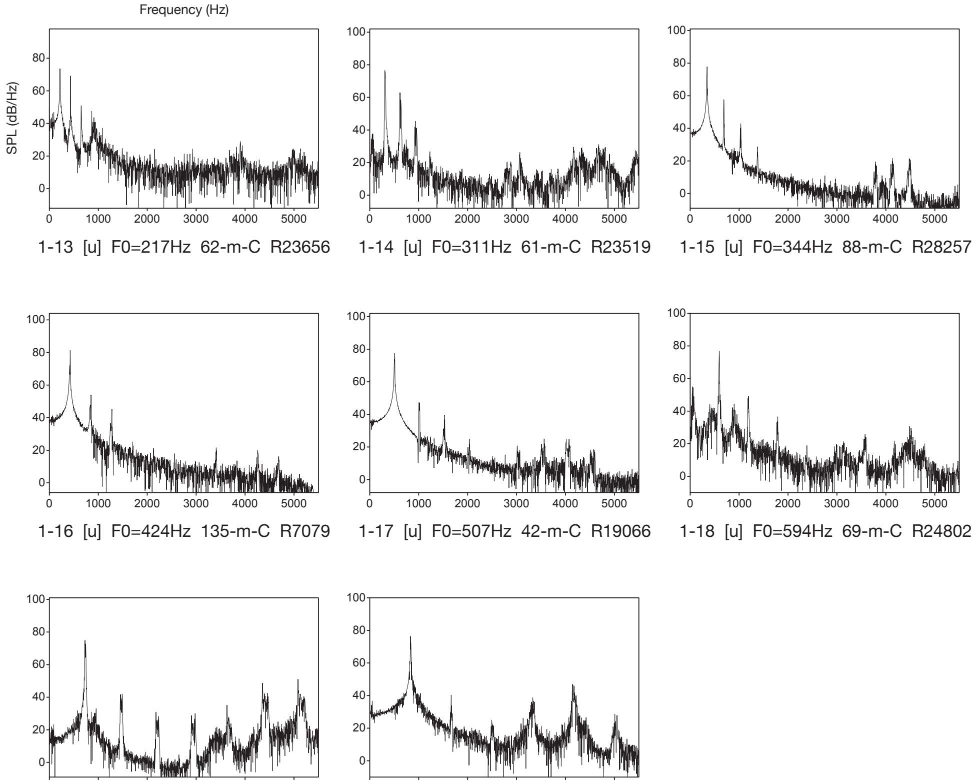
1-9 [o] F0=244Hz 145-m-C R7350

1-10 [o] F0=260Hz 86-m-C R28057

1-11 [o] F0=303Hz 145-m-C R7338

1-12 [o] F0=329Hz 186-m-C R40430

(Figure 1, continuation)



1-13 [u] F0=217Hz 62-m-C R23656

1-14 [u] F0=311Hz 61-m-C R23519

1-15 [u] F0=344Hz 88-m-C R28257

1-16 [u] F0=424Hz 135-m-C R7079

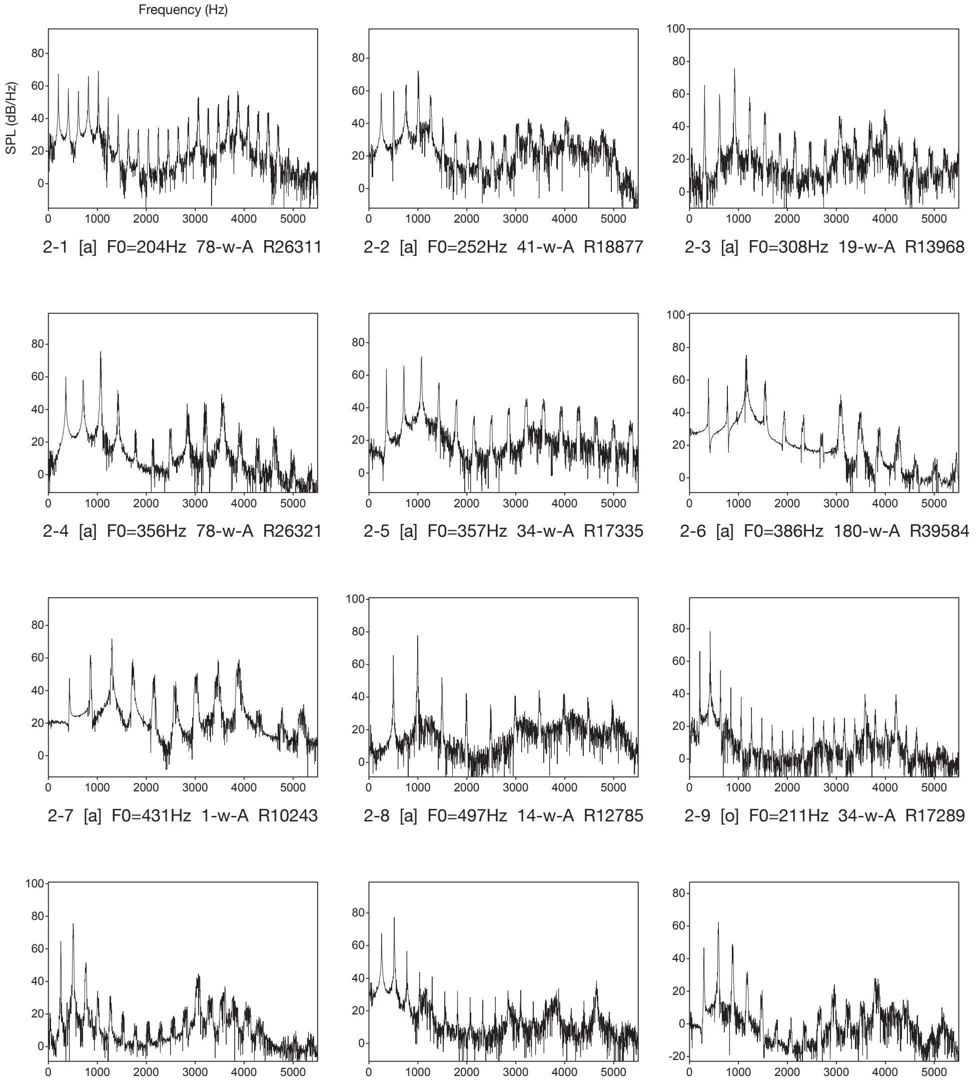
1-17 [u] F0=507Hz 42-m-C R19066

1-18 [u] F0=594Hz 69-m-C R24802

1-19 [u] F0=736Hz 61-m-C R23622

1-20 [u] F0=834Hz 38-w-C R18452

Figure 2. Sounds of /a–ɑ, o, u/, produced by women, which exhibit only one relative spectral energy maximum within their vowel-specific frequency range ≤ c. 1.5 kHz.



2-1 [a] F0=204Hz 78-w-A R26311

2-2 [a] F0=252Hz 41-w-A R18877

2-3 [a] F0=308Hz 19-w-A R13968

2-4 [a] F0=356Hz 78-w-A R26321

2-5 [a] F0=357Hz 34-w-A R17335

2-6 [a] F0=386Hz 180-w-A R39584

2-7 [a] F0=431Hz 1-w-A R10243

2-8 [a] F0=497Hz 14-w-A R12785

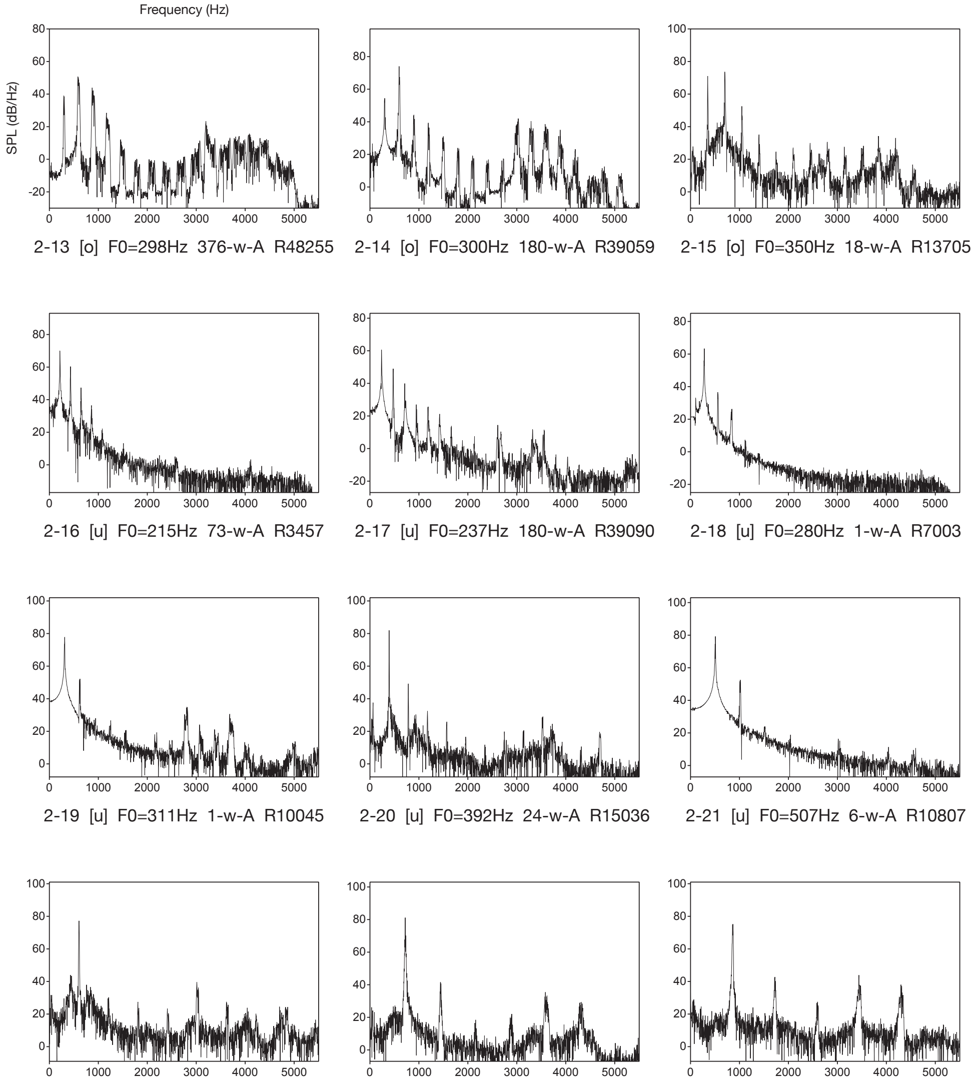
2-9 [o] F0=211Hz 34-w-A R17289

2-10 [o] F0=254Hz 6-w-A R10811

2-11 [o] F0=258Hz 22-w-A R14596

2-12 [o] F0=294Hz 377-w-A R48254

(Figure 2, continuation)



2-13 [o] F0=298Hz 376-w-A R48255

2-14 [o] F0=300Hz 180-w-A R39059

2-15 [o] F0=350Hz 18-w-A R13705

2-16 [u] F0=215Hz 73-w-A R3457

2-17 [u] F0=237Hz 180-w-A R39090

2-18 [u] F0=280Hz 1-w-A R7003

2-19 [u] F0=311Hz 1-w-A R10045

2-20 [u] F0=392Hz 24-w-A R15036

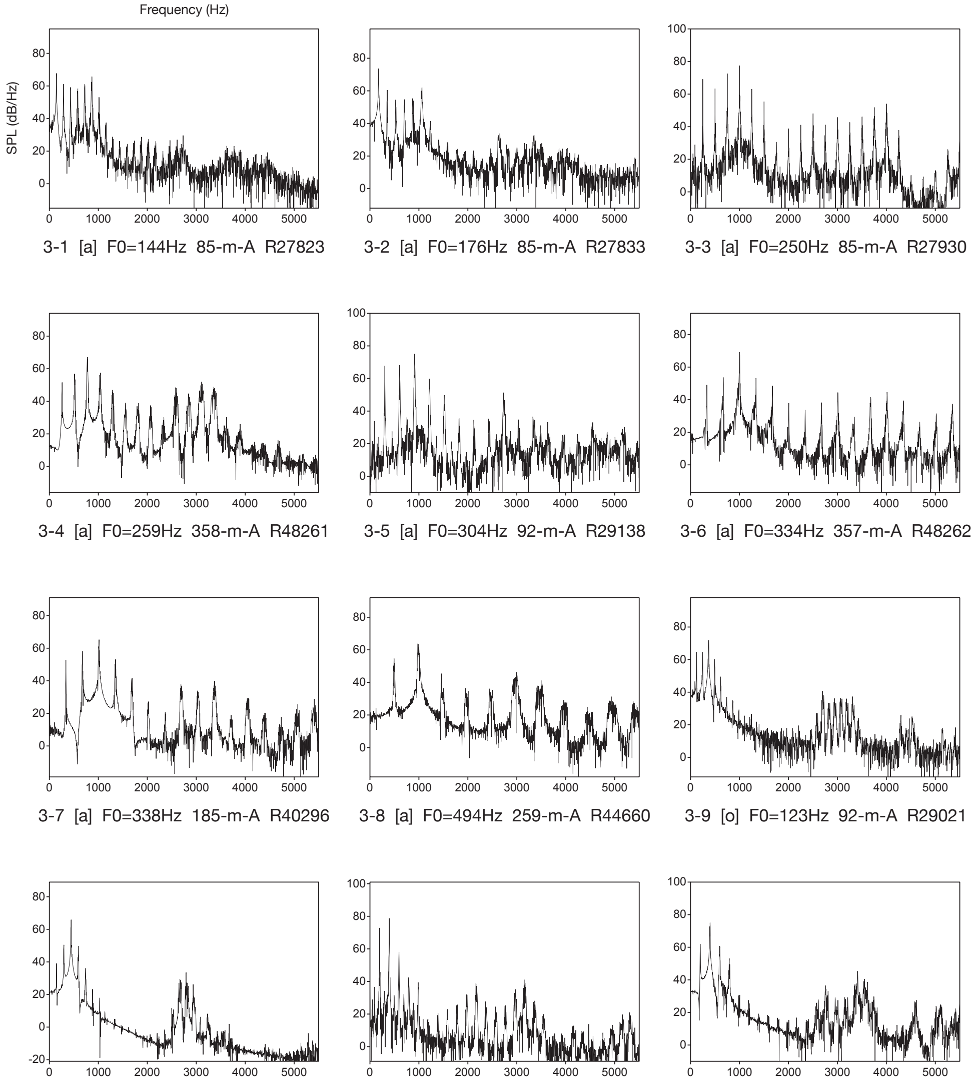
2-21 [u] F0=507Hz 6-w-A R10807

2-22 [u] F0=606Hz 33-w-A R17229

2-23 [u] F0=721Hz 14-w-A R12887

2-24 [u] F0=863Hz 53-w-A R21540

Figure 3. Sounds of /a–ɑ, o, u/, produced by men, which exhibit only one relative spectral energy maximum within their vowel-specific frequency range ≤ c. 1.5 kHz.



3-1 [a] F0=144Hz 85-m-A R27823

3-2 [a] F0=176Hz 85-m-A R27833

3-3 [a] F0=250Hz 85-m-A R27930

3-4 [a] F0=259Hz 358-m-A R48261

3-5 [a] F0=304Hz 92-m-A R29138

3-6 [a] F0=334Hz 357-m-A R48262

3-7 [a] F0=338Hz 185-m-A R40296

3-8 [a] F0=494Hz 259-m-A R44660

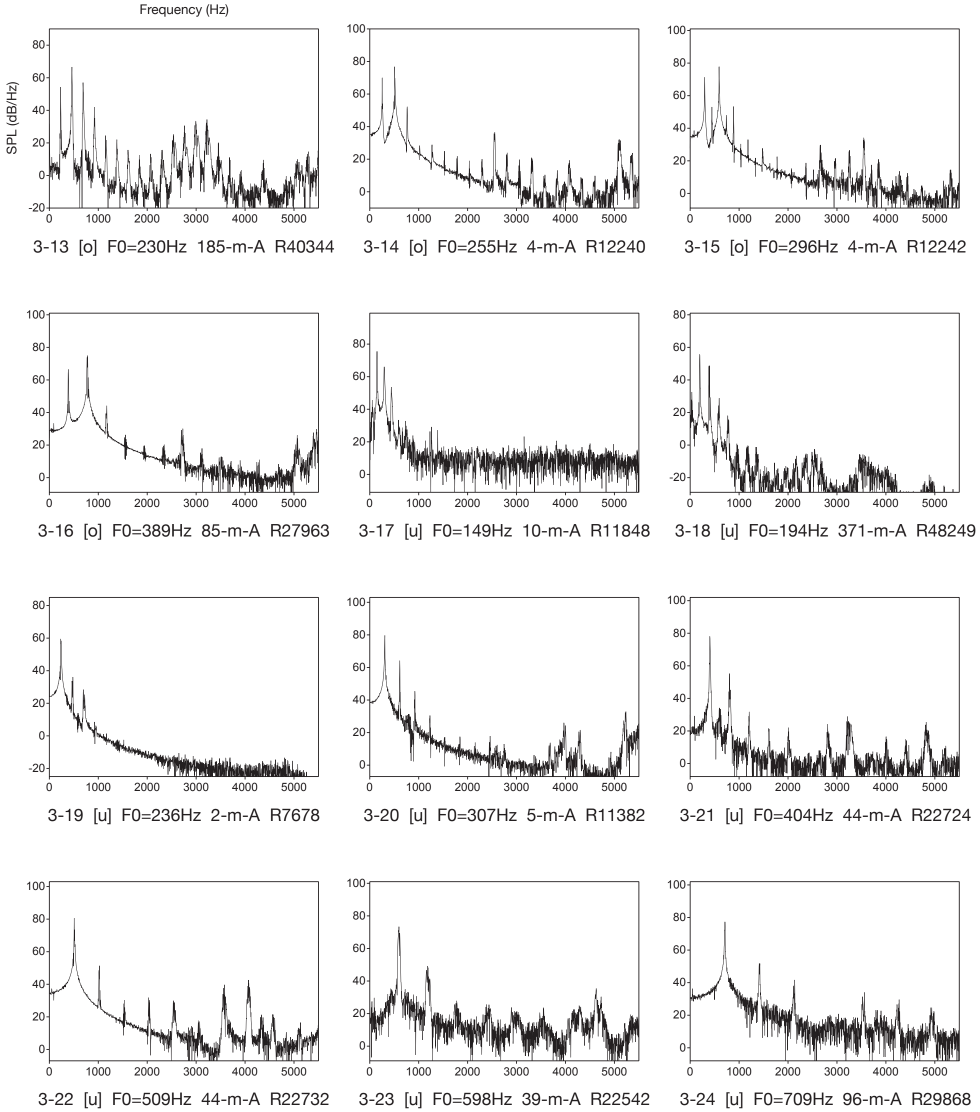
3-9 [o] F0=123Hz 92-m-A R29021

3-10 [o] F0=148Hz 2-m-A R48257

3-11 [o] F0=198Hz 98-m-A R30167

3-12 [o] F0=199Hz 59-m-A R23151

(Figure 3, continuation)



3-13 [o] F0=230Hz 185-m-A R40344

3-14 [o] F0=255Hz 4-m-A R12240

3-15 [o] F0=296Hz 4-m-A R12242

3-16 [o] F0=389Hz 85-m-A R27963

3-17 [u] F0=149Hz 10-m-A R11848

3-18 [u] F0=194Hz 371-m-A R48249

3-19 [u] F0=236Hz 2-m-A R7678

3-20 [u] F0=307Hz 5-m-A R11382

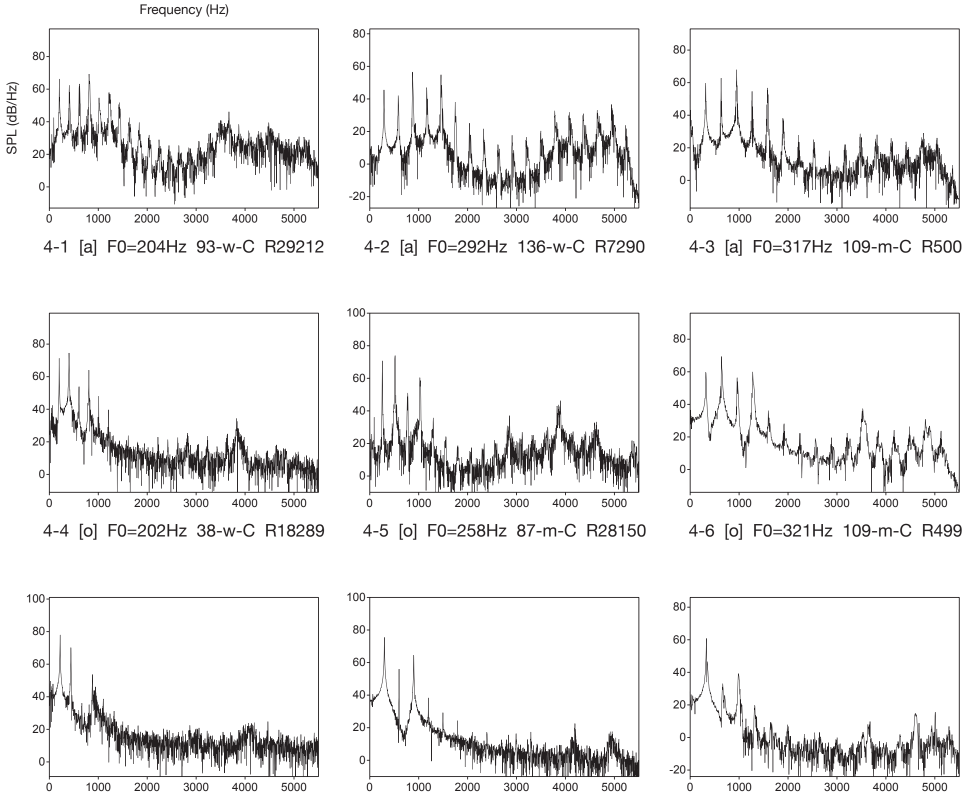
3-21 [u] F0=404Hz 44-m-A R22724

3-22 [u] F0=509Hz 44-m-A R22732

3-23 [u] F0=598Hz 39-m-A R22542

3-24 [u] F0=709Hz 96-m-A R29868

Figure 4. Sounds of /a–ɑ, o, u/, produced by children, which exhibit two relative spectral energy maxima within their vowel-specific frequency range ≤ c. 1.5 kHz.



4-1 [a] F0=204Hz 93-w-C R29212

4-2 [a] F0=292Hz 136-w-C R7290

4-3 [a] F0=317Hz 109-m-C R500

4-4 [o] F0=202Hz 38-w-C R18289

4-5 [o] F0=258Hz 87-m-C R28150

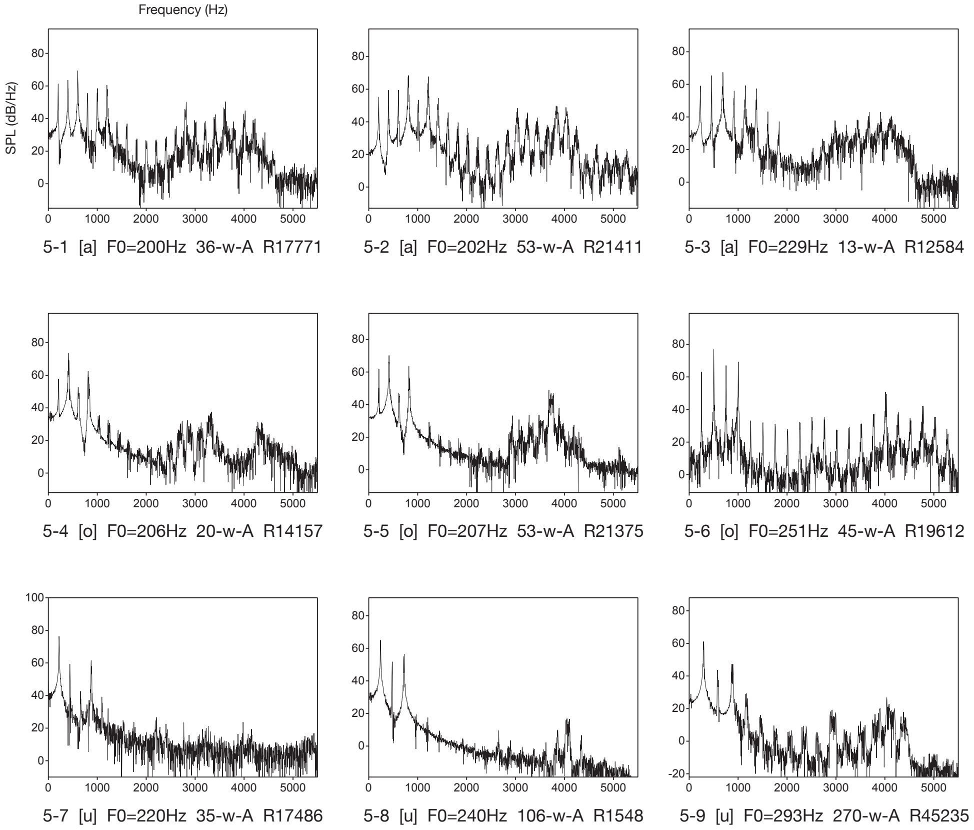
4-6 [o] F0=321Hz 109-m-C R499

4-7 [u] F0=221Hz 66-w-C R24152

4-8 [u] F0=299Hz 54-w-C R21593

4-9 [u] F0=336Hz 186-m-C R40392

Figure 5. Sounds of /a–ɑ, o, u/, produced by women, which exhibit two relative spectral energy maxima within their vowel-specific frequency range ≤ c. 1.5 kHz.



5-1 [a] F0=200Hz 36-w-A R17771

5-2 [a] F0=202Hz 53-w-A R21411

5-3 [a] F0=229Hz 13-w-A R12584

5-4 [o] F0=206Hz 20-w-A R14157

5-5 [o] F0=207Hz 53-w-A R21375

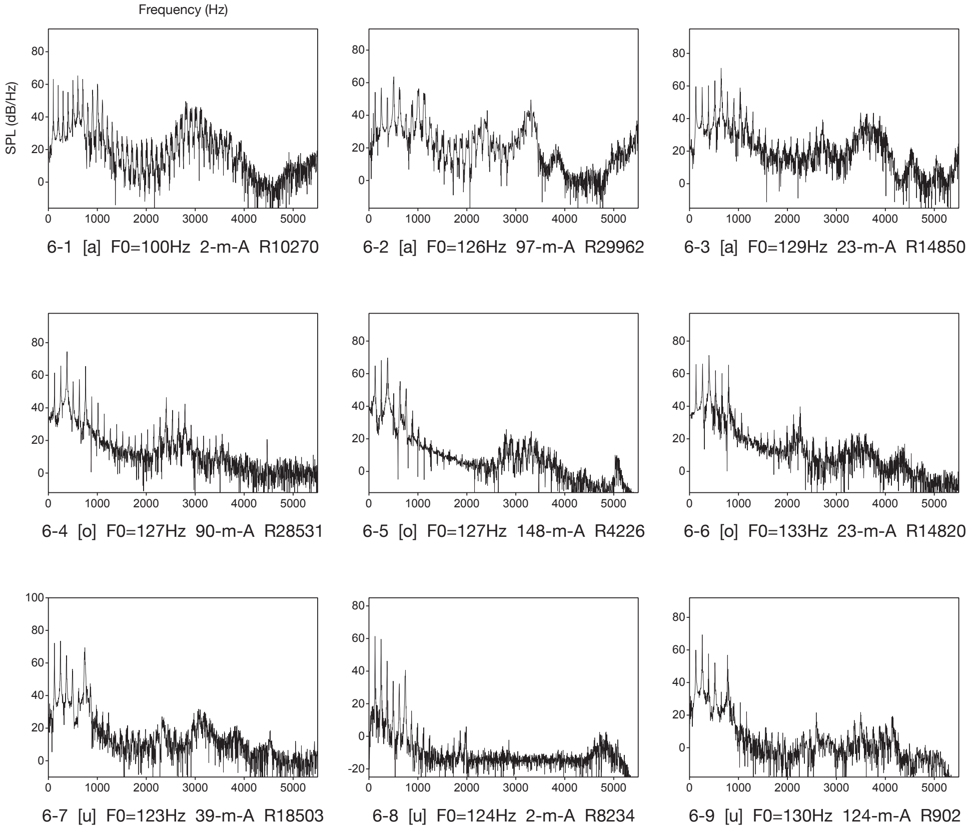
5-6 [o] F0=251Hz 45-w-A R19612

5-7 [u] F0=220Hz 35-w-A R17486

5-8 [u] F0=240Hz 106-w-A R1548

5-9 [u] F0=293Hz 270-w-A R45235

Figure 6. Sounds of /a–ɑ, o, u/, produced by men, which exhibit two relative spectral energy maxima within their vowel-specific frequency range ≤ c. 1.5 kHz.



6-1 [a] F0=100Hz 2-m-A R10270

6-2 [a] F0=126Hz 97-m-A R29962

6-3 [a] F0=129Hz 23-m-A R14850

6-4 [o] F0=127Hz 90-m-A R28531

6-5 [o] F0=127Hz 148-m-A R4226

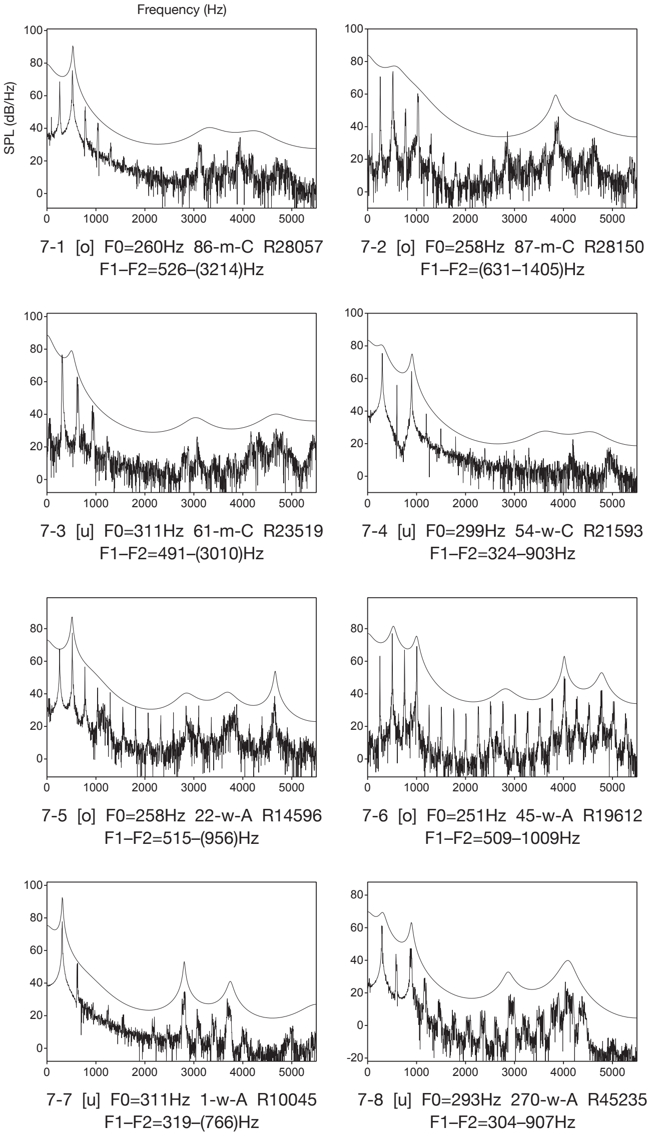
6-6 [o] F0=133Hz 23-m-A R14820

6-7 [u] F0=123Hz 39-m-A R18503

6-8 [u] F0=124Hz 2-m-A R8234

6-9 [u] F0=130Hz 124-m-A R902

Figure 7. Direct comparisons of sounds of back vowels with one or two relative spectral energy maxima ≤ c. 1.5 kHz. (Sounds of children are selected from Figures 1 and 4, those for women from Figures 2 and 5 and those for men from Figures 3 and 6.)



7-1 [o] F0=260Hz 86-m-C R28057
F1–F2=526–(3214)Hz

7-2 [o] F0=258Hz 87-m-C R28150
F1–F2=(631–1405)Hz

7-3 [u] F0=311Hz 61-m-C R23519
F1–F2=491–(3010)Hz

7-4 [u] F0=299Hz 54-w-C R21593
F1–F2=324–903Hz

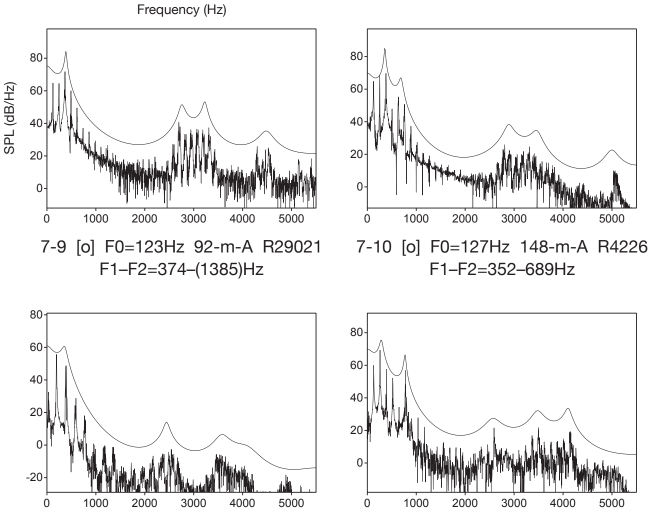
7-5 [o] F0=258Hz 22-w-A R14596
F1–F2=515–(956)Hz

7-6 [o] F0=251Hz 45-w-A R19612
F1–F2=509–1009Hz

7-7 [u] F0=311Hz 1-w-A R10045
F1–F2=319–(766)Hz

7-8 [u] F0=293Hz 270-w-A R45235
F1–F2=304–907Hz

(Figure 7, continuation)



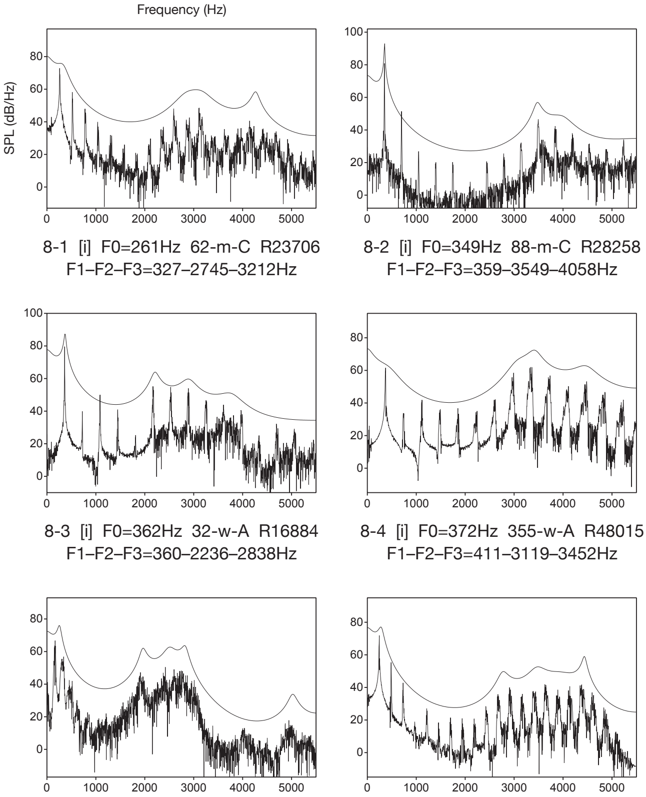
7-9 [o] F0=123Hz 92-m-A R29021
F1–F2=374–(1385)Hz

7-10 [o] F0=127Hz 148-m-A R4226
F1–F2=352–689Hz

7-11 [u] F0=194Hz 371-m-A R48249
F1–F2=301–(1006)Hz

7-12 [u] F0=130Hz 124-m-A R902
F1–F2=291–781Hz

Figure 8. Sound pairs of /i/, each pair produced by speakers of one and the same age and gender-related speaker group, with small differences in F0 and F1 but substantial differences in the higher vowel-related spectral range.



8-1 [i] F0=261Hz 62-m-C R23706
F1–F2–F3=327–2745–3212Hz

8-2 [i] F0=349Hz 88-m-C R28258
F1–F2–F3=359–3549–4058Hz

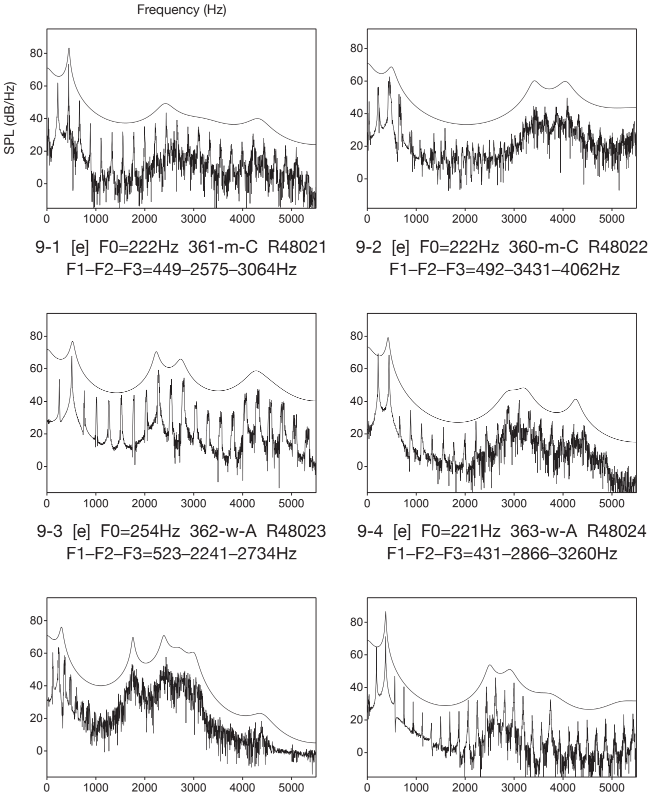
8-3 [i] F0=362Hz 32-w-A R16884
F1–F2–F3=360–2236–2838Hz

8-4 [i] F0=372Hz 355-w-A R48015
F1–F2–F3=411–3119–3452Hz

8-5 [i] F0=162Hz 356-m-A R48016
F1–F2–F3=258–1918–2500Hz

8-6 [i] F0=243Hz 129-m-A R5149
F1–F2–F3=312–2761–3519Hz

Figure 9. Sound pairs of /e/, each pair produced by speakers of one and the same age and gender-related speaker group, with small differences in F0 and F1 but substantial differences in the higher vowel-related spectral range.



9-1 [e] F0=222Hz 361-m-C R48021
F1–F2–F3=449–2575–3064Hz

9-2 [e] F0=222Hz 360-m-C R48022
F1–F2–F3=492–3431–4062Hz

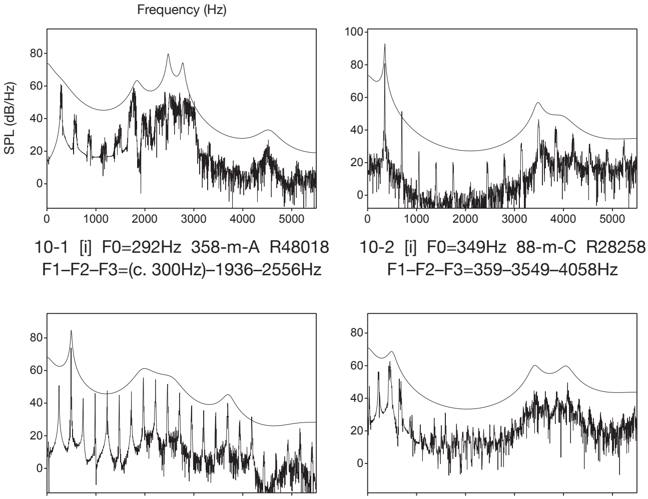
9-3 [e] F0=254Hz 362-w-A R48023
F1–F2–F3=523–2241–2734Hz

9-4 [e] F0=221Hz 363-w-A R48024
F1–F2–F3=431–2866–3260Hz

9-5 [e] F0=121Hz 358-m-A R48025
F1–F2–F3=280–1760–2425Hz

9-6 [e] F0=187Hz 357-m-A R48026
F1–F2–F3=368–2507–2957Hz

Figure 10. A sound pair of /i/ and a corresponding pair of /e/, each pair comparing productions of a man and a child, with small differences in F0 and F1 but very pronounced differences in the higher vowel-related spectral ranges.



10-1 [i] F0=292Hz 358-m-A R48018
F1–F2–F3=(c. 300Hz)–1936–2556Hz

10-2 [i] F0=349Hz 88-m-C R28258
F1–F2–F3=359–3549–4058Hz

10-3 [e] F0=246Hz 359-m-A R48019
F1–F2–F3=500–1923–2303Hz

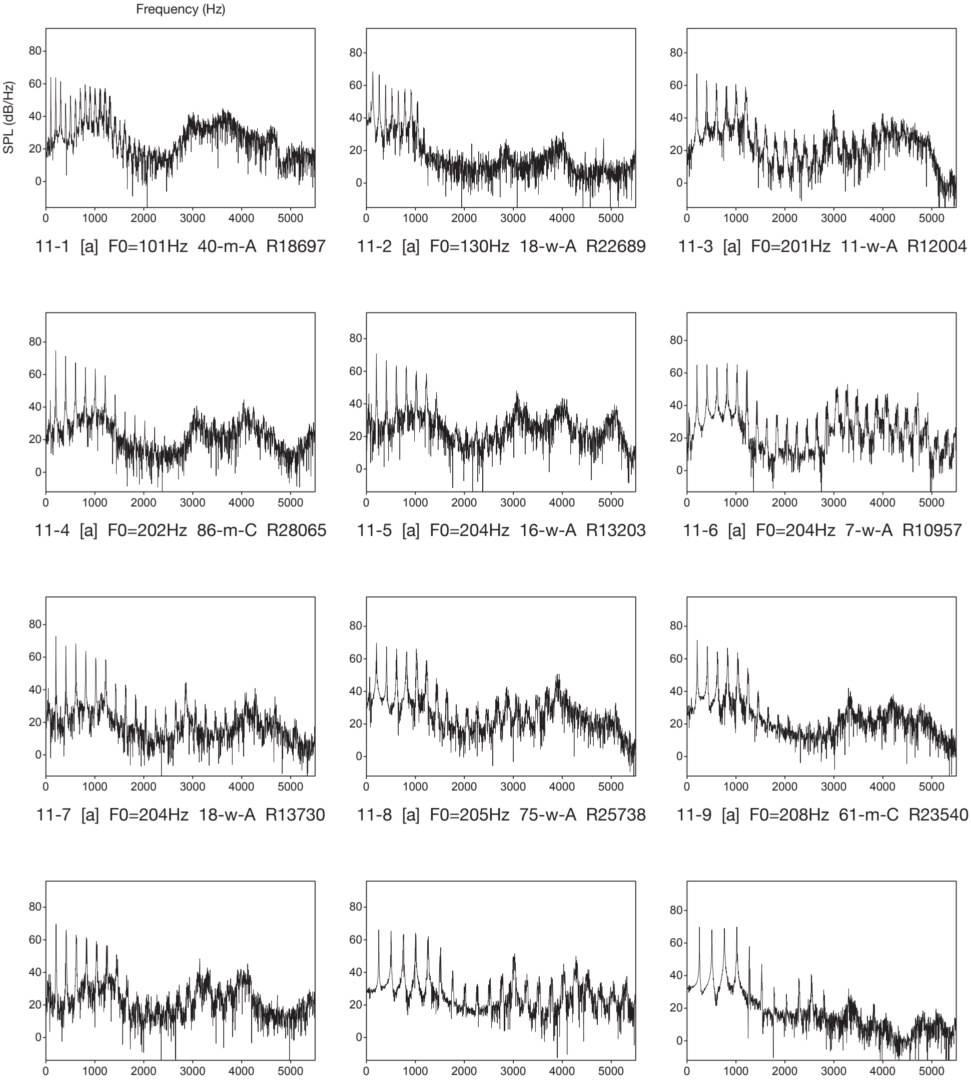
10-4 [e] F0=222Hz 360-m-C R48020
F1–F2–F3=492–3431–4062Hz

M7.2 Partial Lack of Manifestation of Vowel-Specific Relative Spectral Energy Maxima

Figures 11 and 12 show examples of sounds of the vowels /a–ɑ/ and of /o/ with “flat” or “sloping” spectral portions in their vowel-specific frequency range < c. 1.5 kHz which are lacking a clearly determinable peak. Note that the perceived vowel quality of some sounds intentionally produced as /a–ɑ/ lies in between /ɑ/ and /ɔ/, and of some sounds intentionally produced as /o/ in between /o/ and /ɔ/. Note also that for the sounds of /a–ɑ/, a dominant first harmonic is again ignored here when interpreting relative spectral energy maxima. (For cases of “sloping” lower spectral portions in sounds of /u/, see Section M7.1, Figures 1 to 3.)

Figures 13 and 14 show corresponding observations for sounds of front the vowels /i, e/ with “flat” higher spectral portions in their upper vowel-specific frequency range of 1.5–5 kHz which are lacking a clearly determinable pattern of vowel-related peaks.

Figure 11. Sounds of /a–ɑ/, produced by children, women and men, which exhibit “flat” or “sloping” lower spectral portions < c. 1.5 kHz lacking a clearly determinable vowel-related peak.



11-1 [a] F0=101Hz 40-m-A R18697

11-2 [a] F0=130Hz 18-w-A R22689

11-3 [a] F0=201Hz 11-w-A R12004

11-4 [a] F0=202Hz 86-m-C R28065

11-5 [a] F0=204Hz 16-w-A R13203

11-6 [a] F0=204Hz 7-w-A R10957

11-7 [a] F0=204Hz 18-w-A R13730

11-8 [a] F0=205Hz 75-w-A R25738

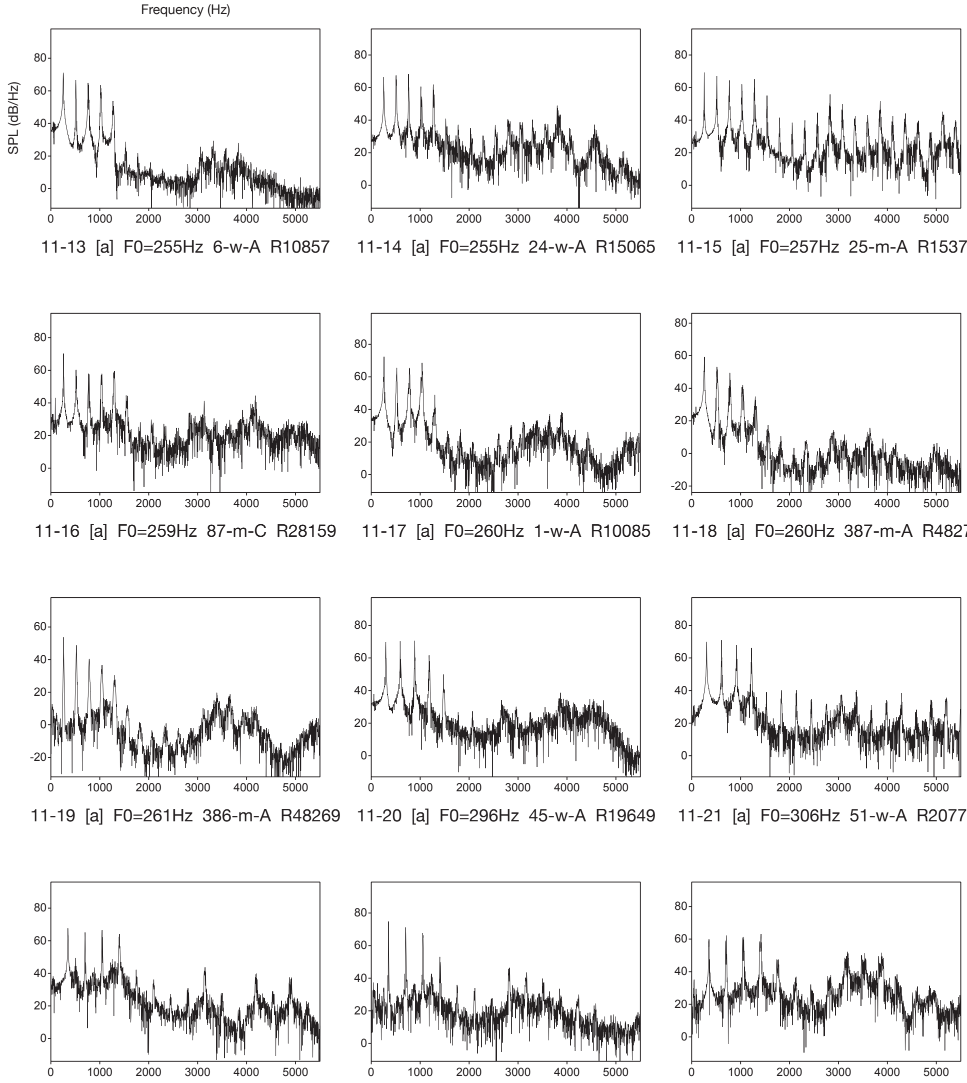
11-9 [a] F0=208Hz 61-m-C R23540

11-10 [a] F0=208Hz 72-w-C R25142

11-11 [a] F0=252Hz 47-w-A R19901

11-12 [a] F0=254Hz 27-m-A R15837

(Figure 11, continuation)



11-13 [a] F0=255Hz 6-w-A R10857

11-14 [a] F0=255Hz 24-w-A R15065

11-15 [a] F0=257Hz 25-m-A R15371

11-16 [a] F0=259Hz 87-m-C R28159

11-17 [a] F0=260Hz 1-w-A R10085

11-18 [a] F0=260Hz 387-m-A R48270

11-19 [a] F0=261Hz 386-m-A R48269

11-20 [a] F0=296Hz 45-w-A R19649

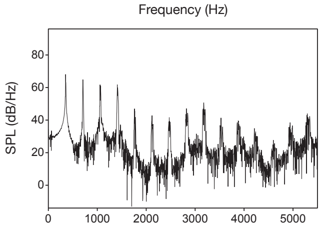
11-21 [a] F0=306Hz 51-w-A R20778

11-22 [a] F0=350Hz 42-m-C R19091

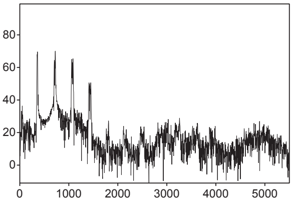
11-23 [a] F0=352Hz 21-w-A R14410

11-24 [a] F0=352Hz 50-w-A R20551

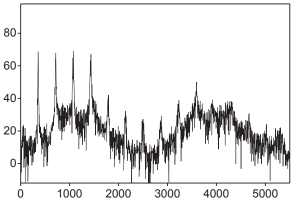
(Figure 11, continuation)



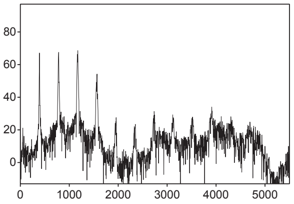
11-25 [a] F0=353Hz 25-m-A R15373



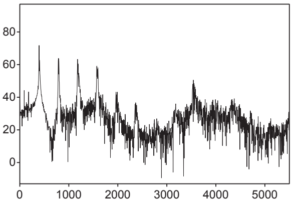
11-26 [a] F0=357Hz 51-w-A R20782



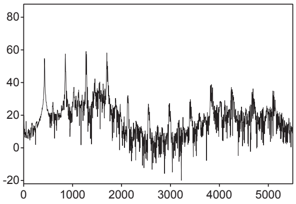
11-27 [a] F0=357Hz 69-m-C R24720



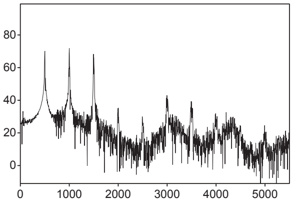
11-28 [a] F0=390Hz 45-w-A R19657



11-29 [a] F0=395Hz 93-w-C R29243

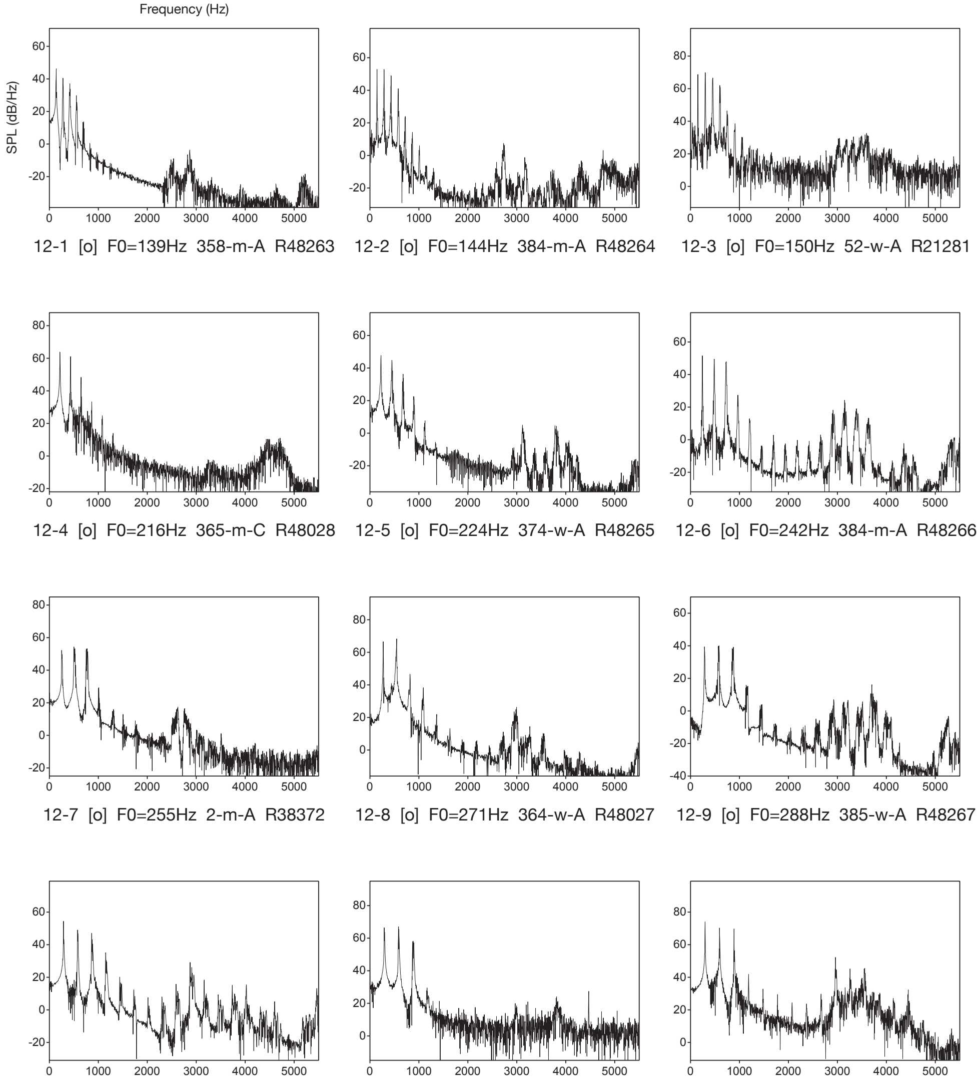


11-30 [a] F0=426Hz 184-m-C R40061



11-31 [a] F0=500Hz 54-w-C R21655

Figure 12. Sounds of /o/, produced by children, women and men, which exhibit “flat” or “sloping” lower spectral portions < c. 1.5 kHz lacking a clearly determinable vowel-related peak.



12-1 [o] F0=139Hz 358-m-A R48263

12-2 [o] F0=144Hz 384-m-A R48264

12-3 [o] F0=150Hz 52-w-A R21281

12-4 [o] F0=216Hz 365-m-C R48028

12-5 [o] F0=224Hz 374-w-A R48265

12-6 [o] F0=242Hz 384-m-A R48266

12-7 [o] F0=255Hz 2-m-A R38372

12-8 [o] F0=271Hz 364-w-A R48027

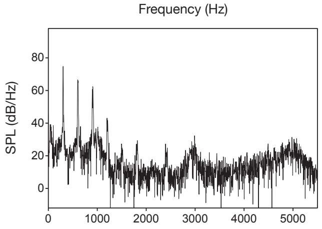
12-9 [o] F0=288Hz 385-w-A R48267

12-10 [o] F0=289Hz 378-w-A R48268

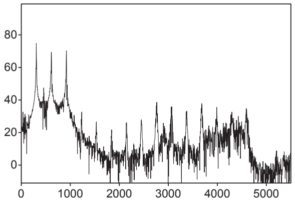
12-11 [o] F0=293Hz 71-w-A R24881

12-12 [o] F0=296Hz 24-w-A R15050

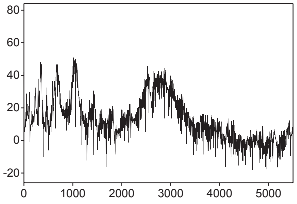
(Figure 12, continuation)



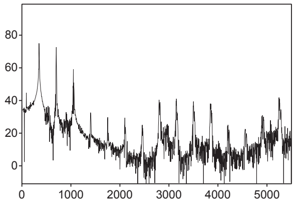
12-13 [o] F0=301Hz 16-w-A R13173



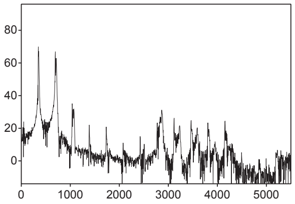
12-14 [o] F0=307Hz 11-w-A R11966



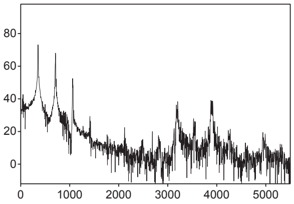
12-15 [o] F0=347Hz 156-m-A R36721



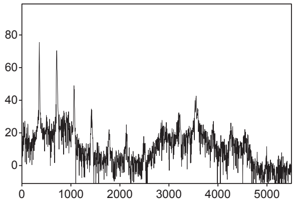
12-16 [o] F0=350Hz 97-m-A R30062



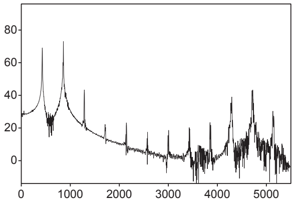
12-17 [o] F0=350Hz 4-m-A R12244



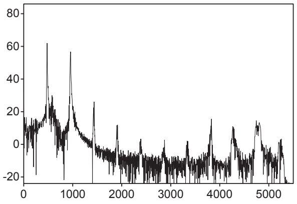
12-18 [o] F0=354Hz 14-w-A R12741



12-19 [o] F0=355Hz 32-w-A R16793

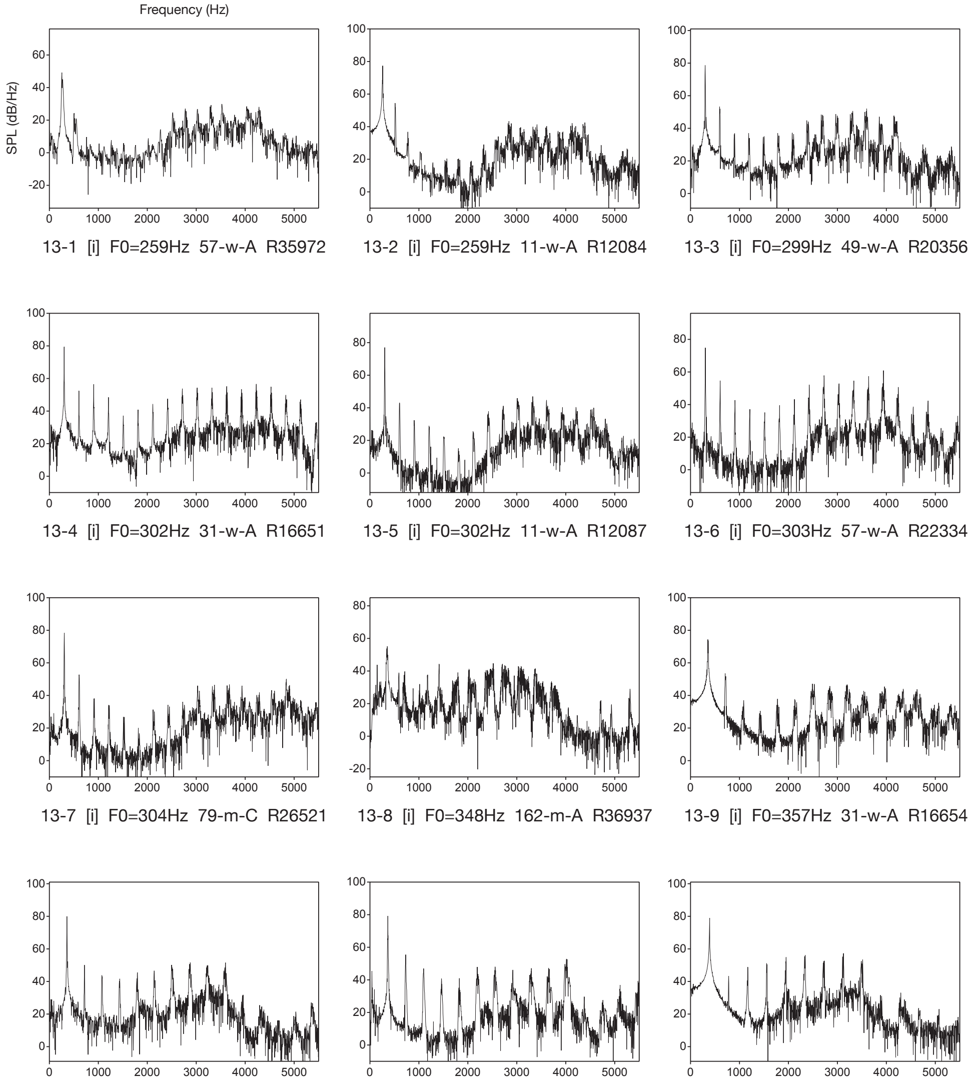


12-20 [o] F0=429Hz 135-m-C R7078



12-21 [o] F0=477Hz 135-m-C R7080

Figure 13. Sounds of /i/, produced by children, women and men, which exhibit “flat” higher spectral portions in the frequency range of 1.5–5 kHz lacking a clearly determinable pattern of vowel-related peaks.



13-1 [i] F0=259Hz 57-w-A R35972

13-2 [i] F0=259Hz 11-w-A R12084

13-3 [i] F0=299Hz 49-w-A R20356

13-4 [i] F0=302Hz 31-w-A R16651

13-5 [i] F0=302Hz 11-w-A R12087

13-6 [i] F0=303Hz 57-w-A R22334

13-7 [i] F0=304Hz 79-m-C R26521

13-8 [i] F0=348Hz 162-m-A R36937

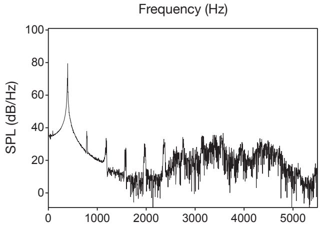
13-9 [i] F0=357Hz 31-w-A R16654

13-10 [i] F0=358Hz 82-m-A R27223

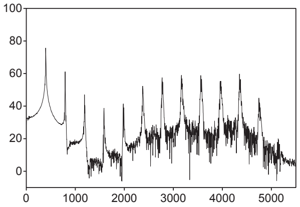
13-11 [i] F0=365Hz 39-m-A R22506

13-12 [i] F0=390Hz 82-m-A R27224

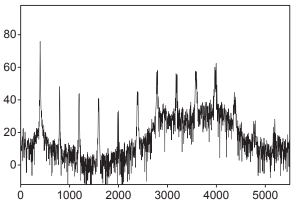
(Figure 13, continuation)



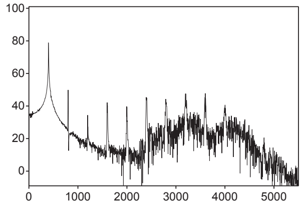
13-13 [i] F0=395Hz 86-m-C R28093



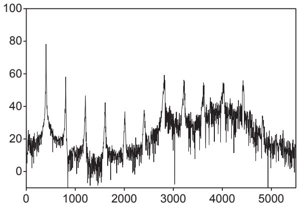
13-14 [i] F0=396Hz 31-w-A R16658



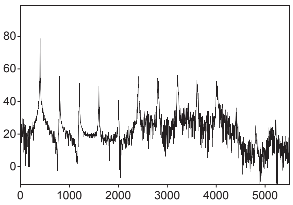
13-15 [i] F0=398Hz 34-w-A R17391



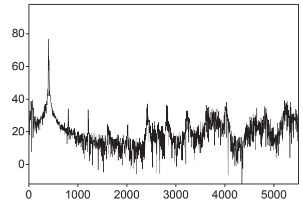
13-16 [i] F0=400Hz 49-w-A R20361



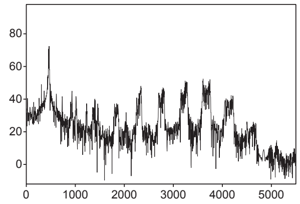
13-17 [i] F0=402Hz 34-w-A R17390



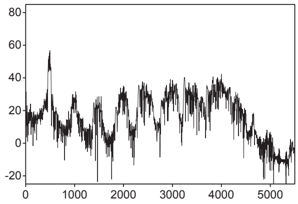
13-18 [i] F0=402Hz 57-w-A R22338



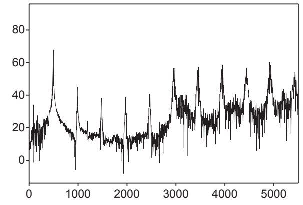
13-19 [i] F0=403Hz 1-w-A R10135



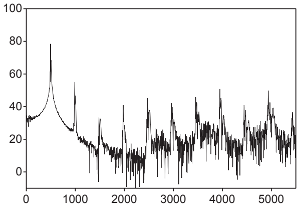
13-20 [i] F0=461Hz 266-m-C R44766



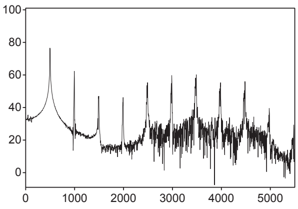
13-21 [i] F0=487Hz 266-m-C R44804



13-22 [i] F0=492Hz 94-m-C R29325

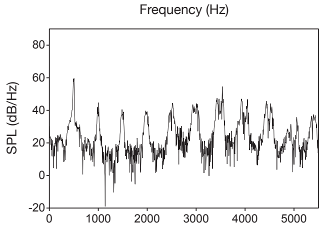


13-23 [i] F0=496Hz 64-w-C R23980

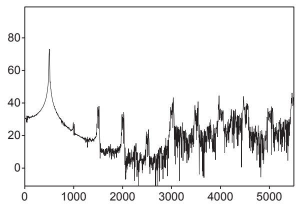


13-24 [i] F0=497Hz 31-w-A R16659

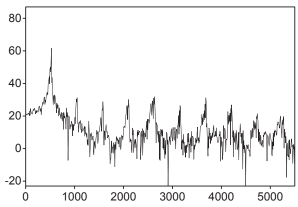
(Figure 13, continuation)



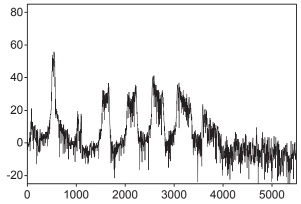
13-25 [i] F0=500Hz 163-m-A R37221



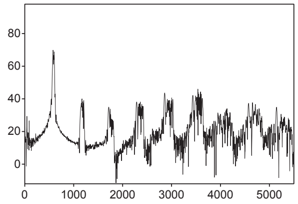
13-26 [i] F0=501Hz 42-m-C R19129



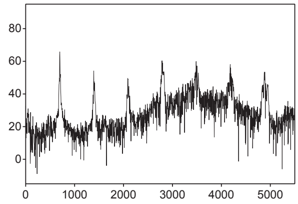
13-27 [i] F0=524Hz 194-m-A R43715



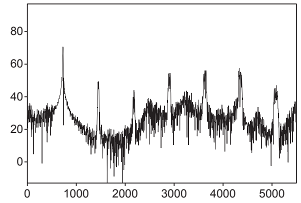
13-28 [i] F0=534Hz 273-m-A R45616



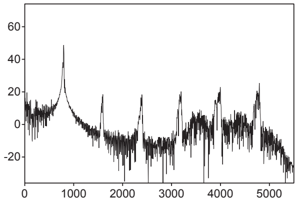
13-29 [i] F0=593Hz 66-w-C R24266



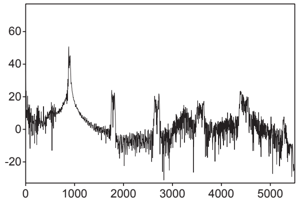
13-30 [i] F0=697Hz 21-w-A R14543



13-31 [i] F0=726Hz 12-w-A R12511

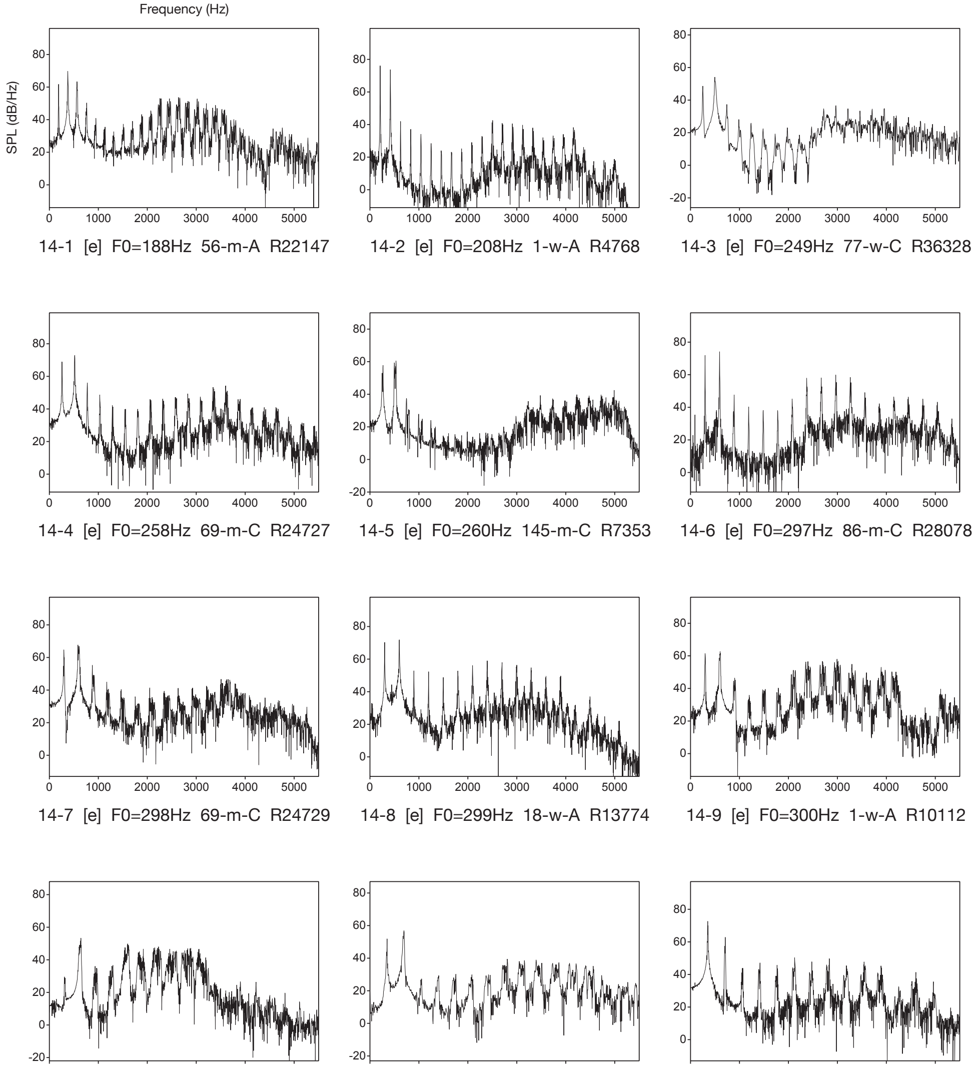


13-32 [i] F0=795Hz 135-m-C R7043



13-33 [i] F0=888Hz 135-m-C R7064

Figure 14. Sounds of /e/, produced by children, women and men, which exhibit "flat" higher spectral portions in the frequency range of 1.5–5 kHz lacking a clearly determinable pattern of vowel-related peaks.



14-1 [e] F0=188Hz 56-m-A R22147

14-2 [e] F0=208Hz 1-w-A R4768

14-3 [e] F0=249Hz 77-w-C R36328

14-4 [e] F0=258Hz 69-m-C R24727

14-5 [e] F0=260Hz 145-m-C R7353

14-6 [e] F0=297Hz 86-m-C R28078

14-7 [e] F0=298Hz 69-m-C R24729

14-8 [e] F0=299Hz 18-w-A R13774

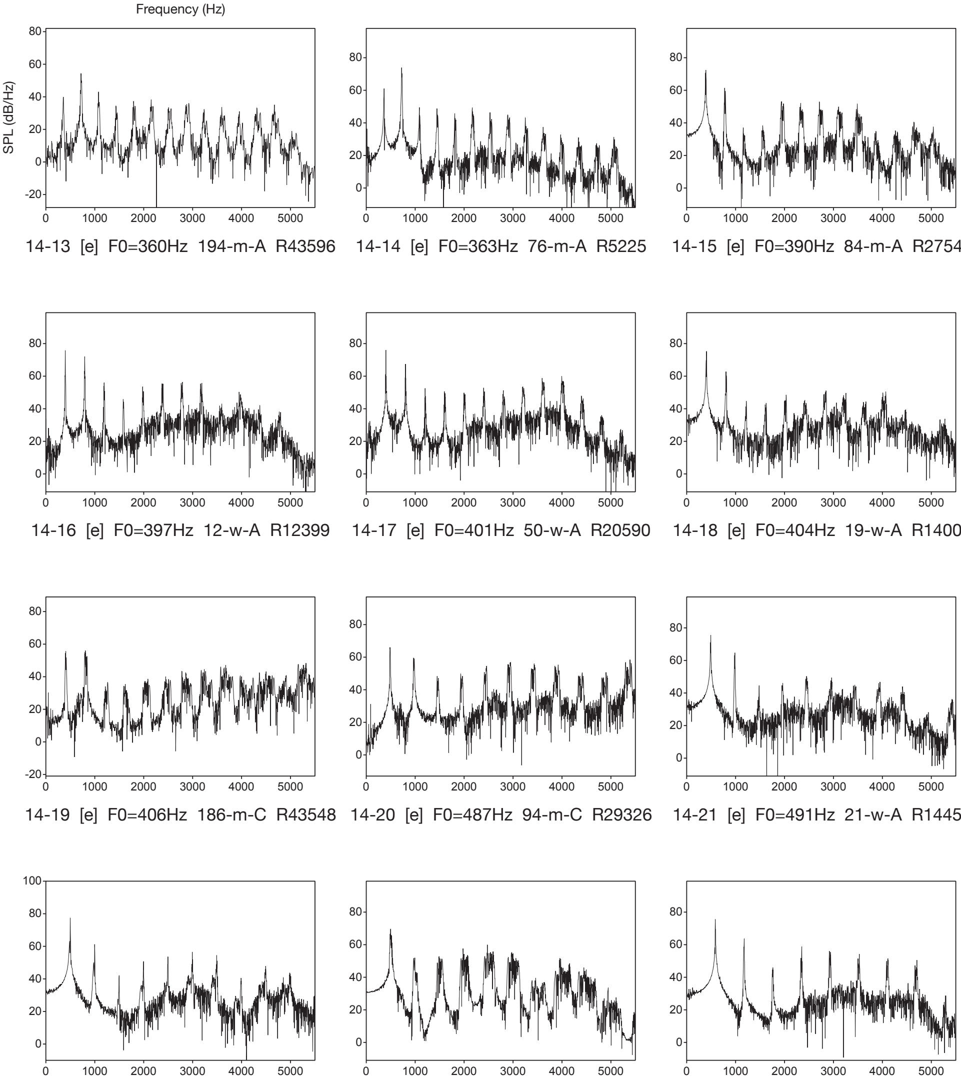
14-9 [e] F0=300Hz 1-w-A R10112

14-10 [e] F0=314Hz 273-m-A R45635

14-11 [e] F0=348Hz 42-m-C R35719

14-12 [e] F0=353Hz 84-m-A R27541

(Figure 14, continuation)



14-13 [e] F0=360Hz 194-m-A R43596
14-14 [e] F0=363Hz 76-m-A R5225
14-15 [e] F0=390Hz 84-m-A R27549

14-16 [e] F0=397Hz 12-w-A R12399
14-17 [e] F0=401Hz 50-w-A R20590
14-18 [e] F0=404Hz 19-w-A R14008

14-19 [e] F0=406Hz 186-m-C R43548
14-20 [e] F0=487Hz 94-m-C R29326
14-21 [e] F0=491Hz 21-w-A R14452

14-22 [e] F0=498Hz 1-w-A R10117
14-23 [e] F0=499Hz 309-w-A R47336
14-24 [e] F0=586Hz 31-w-A R16732

M8 Lack of Correspondence between Vowels and Patterns of Relative Spectral Energy Maxima or Formant Patterns

M8.1 Dependence of Vowel-Specific, Relative Spectral Energy Maxima and Lower Formants ≤ 1.5 kHz on Fundamental Frequency

Figure 1 shows examples of sounds of the vowels /o, ø, e/ produced at different F0 by a woman (/o/), a man (/ø/) and a child (/e/; age 8). In the frequency range of F0 of c. 200–400 Hz, the second partial is generally dominant thus indicating a shift of the lowest spectral peak with rising F0, which is also indicated by the corresponding calculated F1. In more detail: For the sound series of the vowel /o/, the shift in F0 is 170–400 Hz, the frequency shift of the dominant second harmonic is 340–800 Hz and the shift of calculated F1 is c. 380–800 Hz. (Note that for the sound at F0 = 400 Hz, the first calculated formant value at 560 Hz is ignored here because it is associated with a bandwidth of 928 Hz and, as a consequence, the LPC filter curve does not show a corresponding peak.)—For the sound series of the vowel /ø/, the shift in F0 is c. 110–360 Hz, the frequency shift of the dominant harmonic (third harmonic up to F0 = 167 Hz, then second harmonic) is c. 330–720 Hz and the shift of calculated F1 is c. 350–710 Hz.—For the sound series of the vowel /e/, the shift in F0 is c. 210–360 Hz, the frequency shift of the dominant second harmonic is c. 420–720 Hz (dominance is weak but constant) and the shift of calculated F1 is c. 420–720 Hz.

Figure 2 shows examples of sounds of the vowels /u, y, i/ produced at different F0 by a woman (/u/), a child (/y/; age 13, transition to adolescence) and a woman (/i/). For all sounds, the first partial is generally dominant thus indicating a shift of the lowest spectral peak with rising F0, which is also indicated by the corresponding calculated F1. (Note that for higher levels of F0, the calculation of F1 is methodically unsubstantiated; however, the calculated values correspond to the dominant first harmonics.) In more detail: For the sound series of the vowel /u/, the shift in F0 is c. 220–870 Hz, as is true for the frequency shift of the first dominant harmonic and the shift of calculated F1 is c. 230–870 Hz.—For the sound series of the vowel /y/, the shift in F0 is c. 210–710 Hz, as is true for the frequency shift of the first dominant harmonic, and the shift of calculated F1 is c. 380–740 Hz. (Note the problem of automatic calculation of F1 for the example in Figure 2-14.)—For the sound series of the vowel /i/, the shift in F0 is c. 210–830 Hz, as is true for the frequency shift of the first dominant harmonic and the shift of calculated F1 is c. 240–900 Hz.

Note the very pronounced spectral differences for the three sounds of /i, y, u/ in the frequency range of F0 of 700–800 Hz which reinforces the thesis of a parallelism between differences in perceived vowel quality and related acoustic differences, that is, the thesis of vowel-specific harmonic spectra of high-pitched sounds.

However, as mentioned in Section 8.1, indications for an F0-dependence of the lower spectral peaks and lower formants ≤ 1.5 kHz are not systematic: above all, the indications in question relate to frequency ranges of F0, to vowel qualities and to single speakers and their phonation characteristics, including vocal effort.

Concerning the F0 ranges, the indications for the F0-dependence in question are generally weak or absent for F0 < c. 200 Hz for the sounds of all vowels (see, for example, Figure 1 in this chapter, the corresponding sounds of /ø/).

Concerning vowel quality, the indications of the F0-dependence in question are particularly evident in the sounds of /i, y, e, ø, o, u/ but often unsystematic, weak or even absent for the sounds of /ɛ/ and of /a–ɑ/. In terms of an illustration, Figure 3 shows examples of sounds of /a–ɑ/ produced by a child (age 13, transition to adolescence) on different F0. The harmonic spectrum strongly varies and peak and formant estimation is difficult to conduct. However, no clear indication of a relation between F0 and the lower spectral envelope is evident.

Concerning single speakers and their phonation characteristics, including vocal effort, Figure 4 shows examples of sounds of /o/ produced at different F0 by a woman; in contrast to the corresponding sound series in Figure 1, only a very weak indication of a relation between F0 and the lower spectrum is evident.

But, as mentioned in Section 8.1, although the indications for the dependence discussed here prove to be unsystematic, the findings of intelligible vowel sounds at fundamental frequencies > 500 Hz (see next chapter) and of formant pattern ambiguity (see Chapter M9) force us to relate the lower spectral peaks and the lower formants to fundamental frequency.

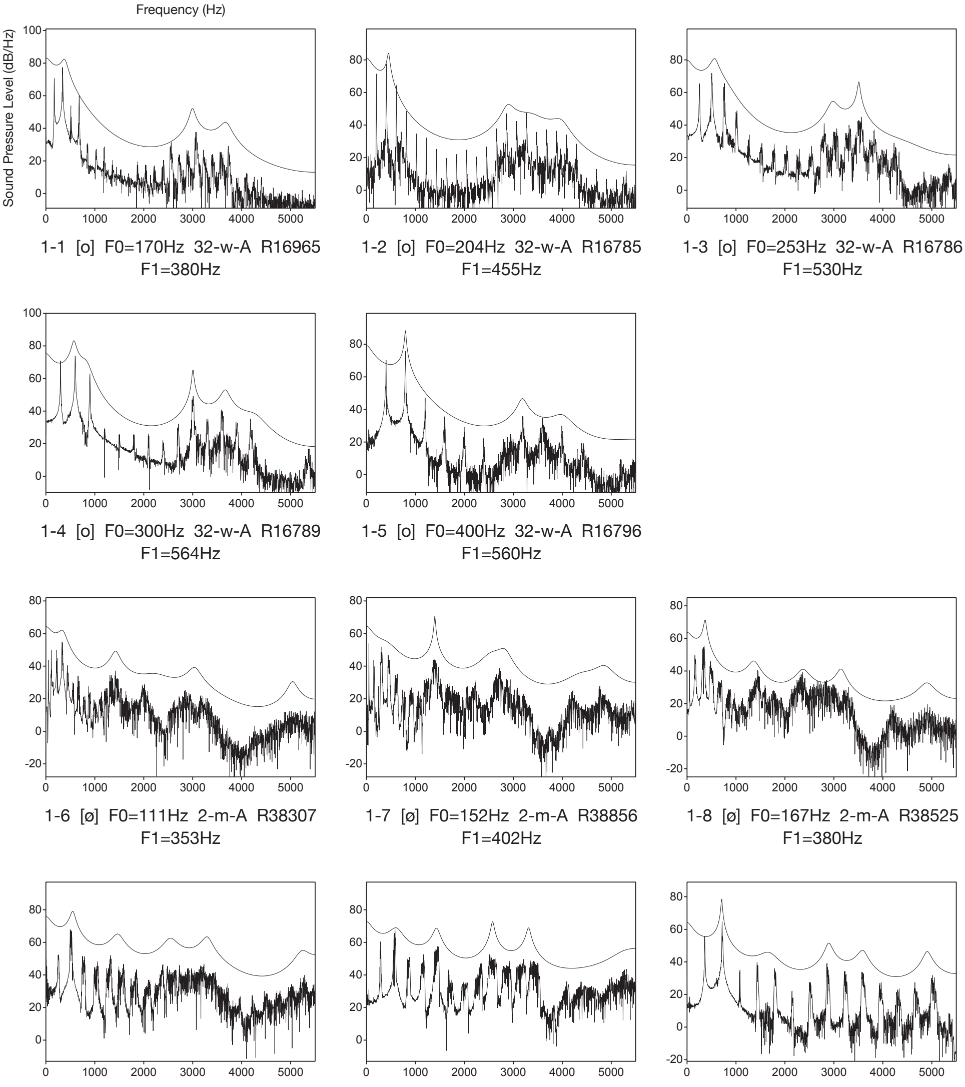
In addition, such a dependence can also be observed for the second formant for cases of sounds of back vowels (see, for example, Section 10.1, Figure 1).

In the context of such F1 shifts with rising F0, “inverted” frequency levels of the lowest spectral peak and of calculated F1 can be observed for two sounds of two different vowels: where statistical values give lower formant frequencies for F1 for one vowel quality than for the other, higher values can be found for sounds of the former than for sounds of the latter if F0 variations are included into the investigation. Figures 5 shows examples of such cases in terms of sound pairs of /o, u/ and /e, i/. (The sound pairs produced by children, women and men are presented separately.) The lowest spectral peaks < 1.5 kHz for the sounds of /u/ are above those of the sounds of /o/, as is the case for the sounds of /i/ compared with the sounds of /e/. Moreover, no clear indication of a second peak < 1.5 kHz and a corresponding marked F2 is manifest for the sounds of /o, u/, and the calculated F2 for the sound pairs of /e, i/ are also “inverted”, i.e. F2 for the sounds of /i/ is found below F2 for the sounds of /e/.

This observation foreshadows formant pattern ambiguity of vowel sounds, as documented in detail in Chapter M9.

For earlier accounts, see Maurer, Landis, and d’Heureuse (1991), Maurer and Landis (1995, 1996, 2000); see also Traunmüller (n.d.) for synthesised examples.

Figure 1. Sounds of /o, ø, e/ produced at different F0 by a woman (/o/), a man (/ø/) and a child (/e/) indicating a shift of the lowest spectral peak as well as of calculated F1 with rising F0.



1-1 [o] F0=170Hz 32-w-A R16965
F1=380Hz

1-2 [o] F0=204Hz 32-w-A R16785
F1=455Hz

1-3 [o] F0=253Hz 32-w-A R16786
F1=530Hz

1-4 [o] F0=300Hz 32-w-A R16789
F1=564Hz

1-5 [o] F0=400Hz 32-w-A R16796
F1=560Hz

1-6 [ø] F0=111Hz 2-m-A R38307
F1=353Hz

1-7 [ø] F0=152Hz 2-m-A R38856
F1=402Hz

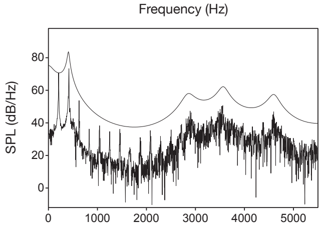
1-8 [ø] F0=167Hz 2-m-A R38525
F1=380Hz

1-9 [ø] F0=256Hz 2-m-A R38433
F1=539Hz

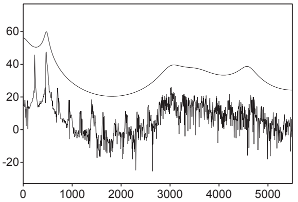
1-10 [ø] F0=287Hz 2-m-A R38435
F1=579Hz

1-11 [ø] F0=359Hz 2-m-A R7597
F1=710Hz

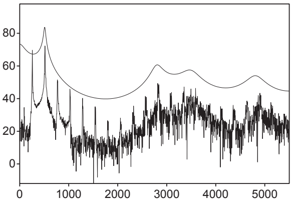
(Figure 1, continuation)



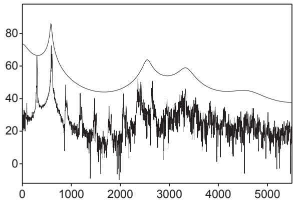
1-12 [e] F0=208Hz 72-w-C R25154
F1=420Hz



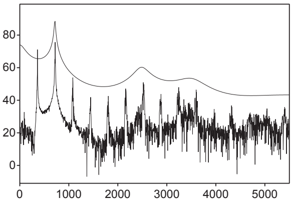
1-13 [e] F0=235Hz 72-w-C R36238
F1=472Hz



1-14 [e] F0=257Hz 72-w-C R25156
F1=534Hz

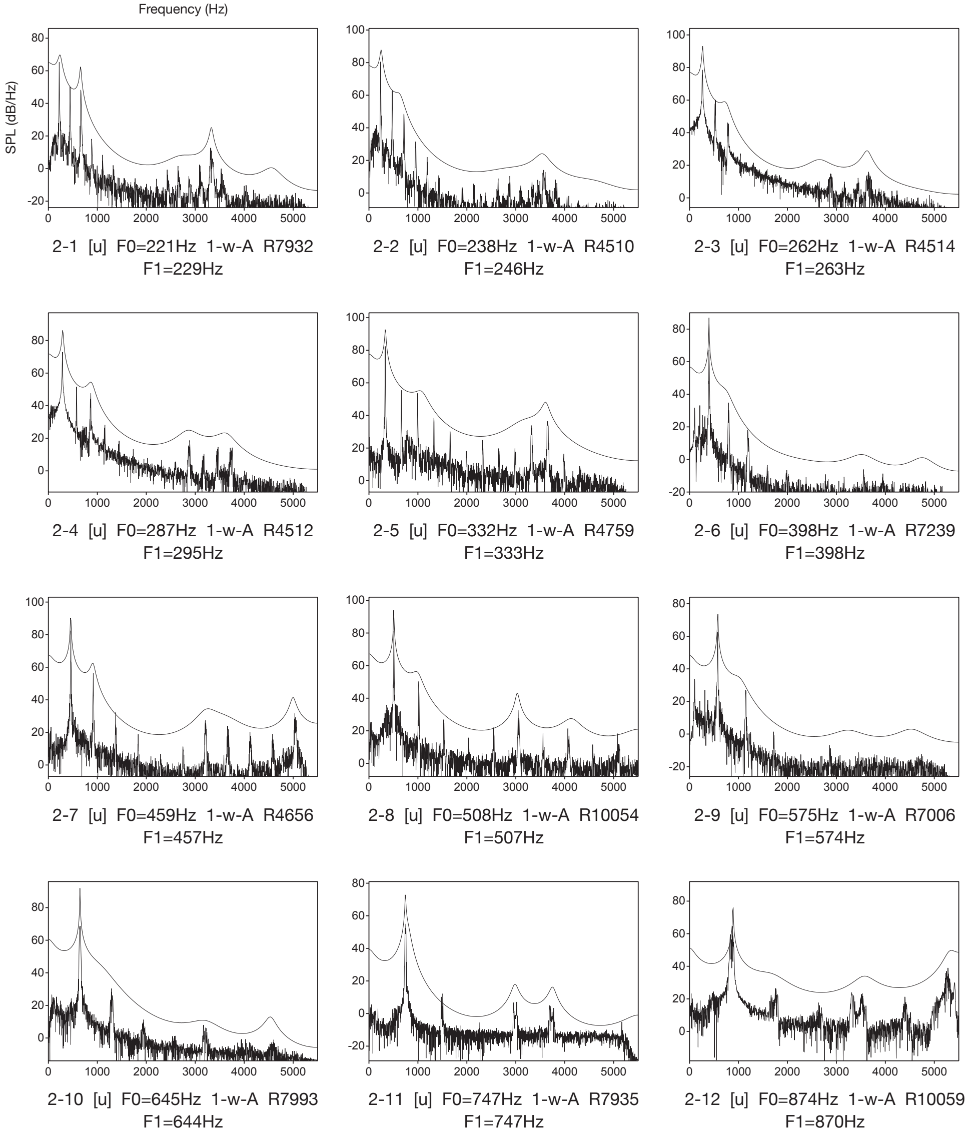


1-15 [e] F0=296Hz 72-w-C R25159
F1=596Hz



1-16 [e] F0=360Hz 72-w-C R25160
F1=718Hz

Figure 2. Sounds of /u, y, i/ produced at different F0 by a woman (/u/), a child (/y/) and another woman (/i/) indicating a shift of the lowest spectral peak as well as of calculated F1 with rising F0.



2-1 [u] F0=221Hz 1-w-A R7932
F1=229Hz

2-2 [u] F0=238Hz 1-w-A R4510
F1=246Hz

2-3 [u] F0=262Hz 1-w-A R4514
F1=263Hz

2-4 [u] F0=287Hz 1-w-A R4512
F1=295Hz

2-5 [u] F0=332Hz 1-w-A R4759
F1=333Hz

2-6 [u] F0=398Hz 1-w-A R7239
F1=398Hz

2-7 [u] F0=459Hz 1-w-A R4656
F1=457Hz

2-8 [u] F0=508Hz 1-w-A R10054
F1=507Hz

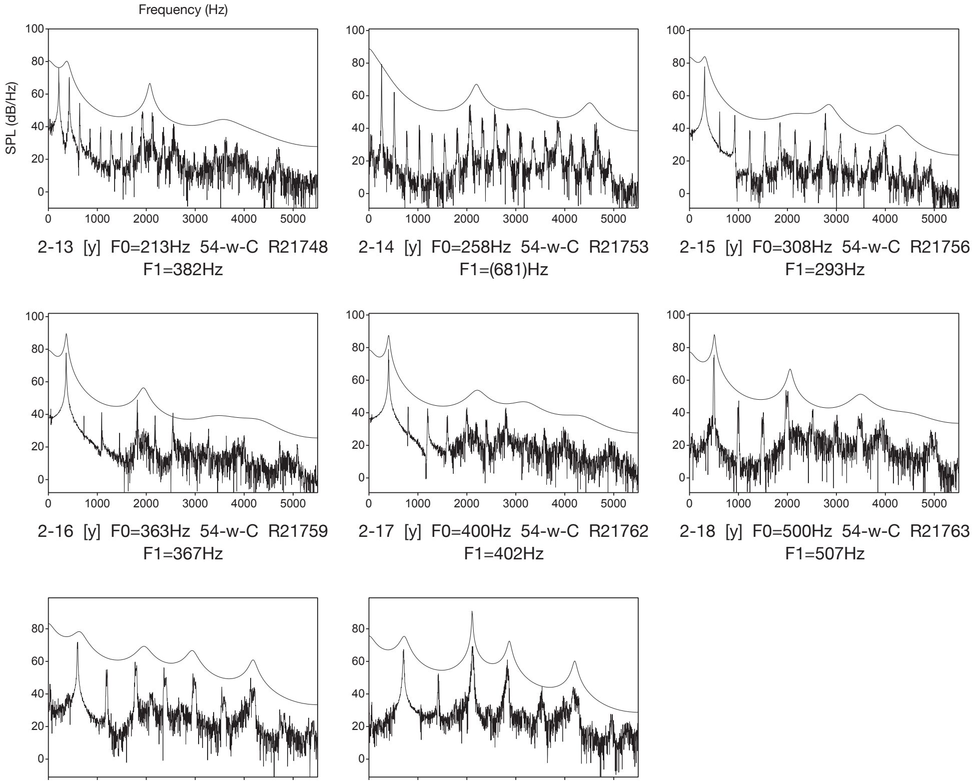
2-9 [u] F0=575Hz 1-w-A R7006
F1=574Hz

2-10 [u] F0=645Hz 1-w-A R7993
F1=644Hz

2-11 [u] F0=747Hz 1-w-A R7935
F1=747Hz

2-12 [u] F0=874Hz 1-w-A R10059
F1=870Hz

(Figure 2, continuation)



2-13 [y] F0=213Hz 54-w-C R21748
F1=382Hz

2-14 [y] F0=258Hz 54-w-C R21753
F1=(681)Hz

2-15 [y] F0=308Hz 54-w-C R21756
F1=293Hz

2-16 [y] F0=363Hz 54-w-C R21759
F1=367Hz

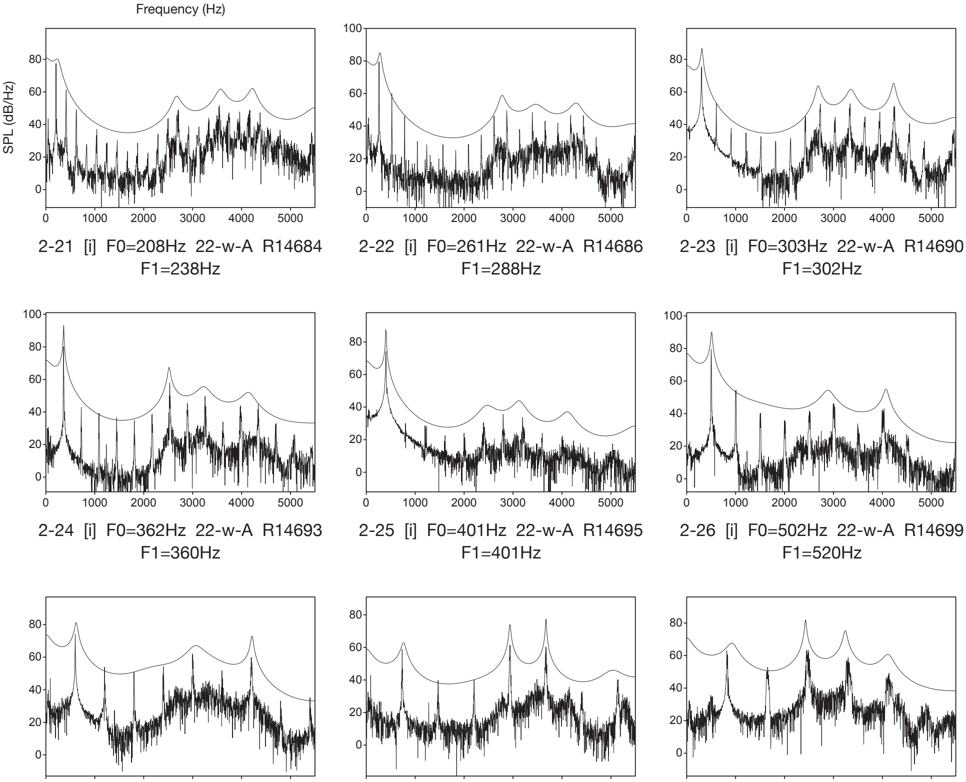
2-17 [y] F0=400Hz 54-w-C R21762
F1=402Hz

2-18 [y] F0=500Hz 54-w-C R21763
F1=507Hz

2-19 [y] F0=596Hz 54-w-C R21787
F1=630Hz

2-20 [y] F0=707Hz 54-w-C R21788
F1=744Hz

(Figure 2, continuation)



2-21 [i] F0=208Hz 22-w-A R14684
F1=238Hz

2-22 [i] F0=261Hz 22-w-A R14686
F1=288Hz

2-23 [i] F0=303Hz 22-w-A R14690
F1=302Hz

2-24 [i] F0=362Hz 22-w-A R14693
F1=360Hz

2-25 [i] F0=401Hz 22-w-A R14695
F1=401Hz

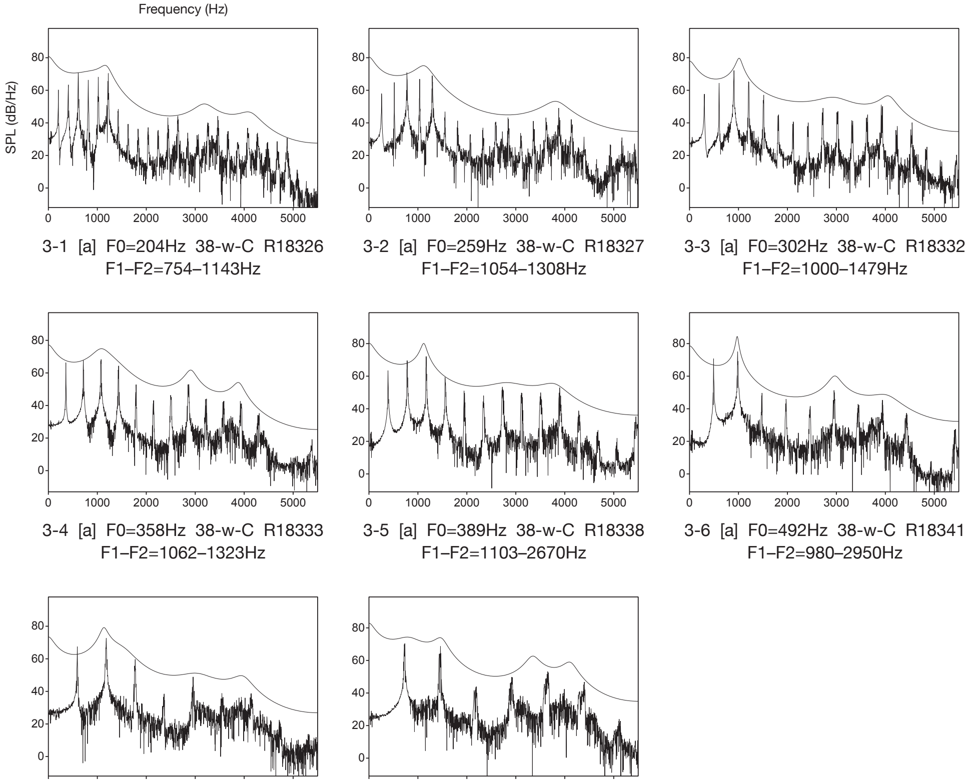
2-26 [i] F0=502Hz 22-w-A R14699
F1=520Hz

2-27 [i] F0=600Hz 22-w-A R22775
F1=632Hz

2-28 [i] F0=734Hz 22-w-A R22779
F1=750Hz

2-29 [i] F0=825Hz 22-w-A R14773
F1=899Hz

Figure 3. Sounds of /a–ɑ/, produced at different F0 by a child, for which there is no clear indication of a relation between F0 and the lower spectral envelope (even if the harmonic spectrum strongly varies).



3-1 [a] F0=204Hz 38-w-C R18326
F1–F2=754–1143Hz

3-2 [a] F0=259Hz 38-w-C R18327
F1–F2=1054–1308Hz

3-3 [a] F0=302Hz 38-w-C R18332
F1–F2=1000–1479Hz

3-4 [a] F0=358Hz 38-w-C R18333
F1–F2=1062–1323Hz

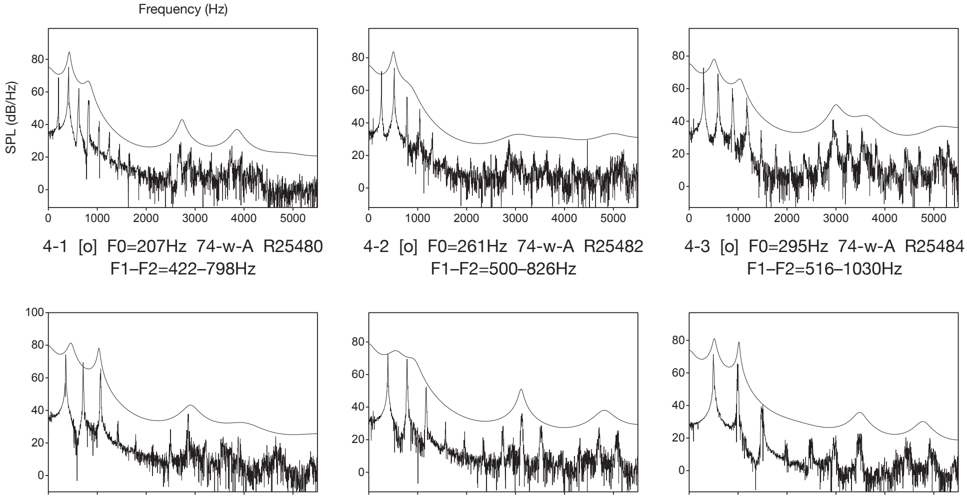
3-5 [a] F0=389Hz 38-w-C R18338
F1–F2=1103–2670Hz

3-6 [a] F0=492Hz 38-w-C R18341
F1–F2=980–2950Hz

3-7 [a] F0=590Hz 38-w-C R18459
F1–F2=1179–1414Hz

3-8 [a] F0=727Hz 38-w-C R18460
F1–F2=868–1356Hz

Figure 4. Sounds of /o/, produced at different F0 by a woman, for which only a very weak indication of a relation between F0 and the lower spectrum is manifest.



4-1 [o] F0=207Hz 74-w-A R25480
F1–F2=422–798Hz

4-2 [o] F0=261Hz 74-w-A R25482
F1–F2=500–826Hz

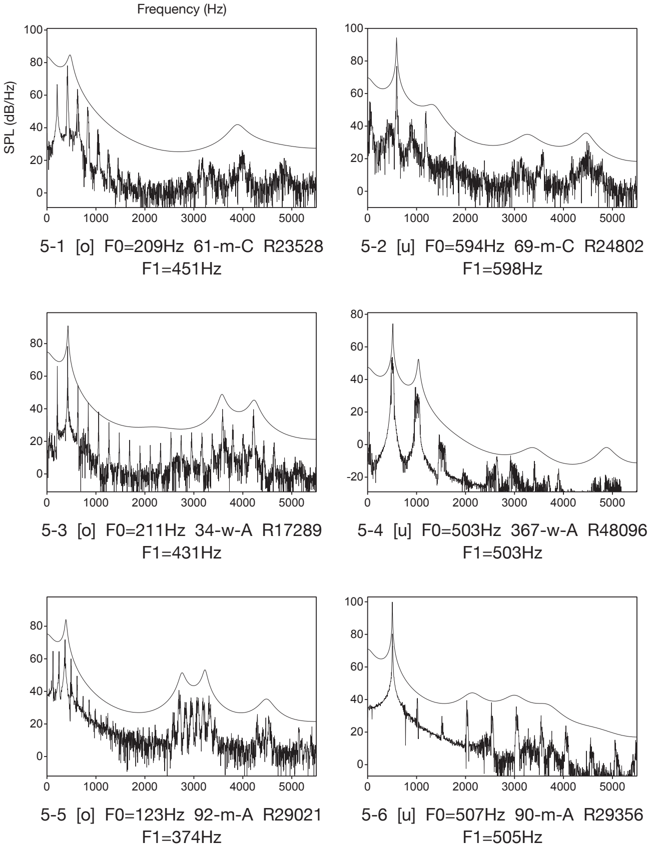
4-3 [o] F0=295Hz 74-w-A R25484
F1–F2=516–1030Hz

4-4 [o] F0=355Hz 74-w-A R25487
F1–F2=463–1019Hz

4-5 [o] F0=391Hz 74-w-A R25491
F1–F2=535–917Hz

4-6 [o] F0=495Hz 74-w-A R25493
F1–F2=490–1014Hz

Figure 5. Three sound pairs of /o, u/ and three sound pairs of /e, i/, produced by children, women and men, exhibiting a higher first spectral peak frequency for /u/ than for /o/, and for /i/ than for /e/, respectively. Note also the absent second spectral peak <1.5 kHz for the sounds of the back vowels and higher calculated F2 for /e/ than for /i/.



5-1 [o] F0=209Hz 61-m-C R23528
F1=451Hz

5-2 [u] F0=594Hz 69-m-C R24802
F1=598Hz

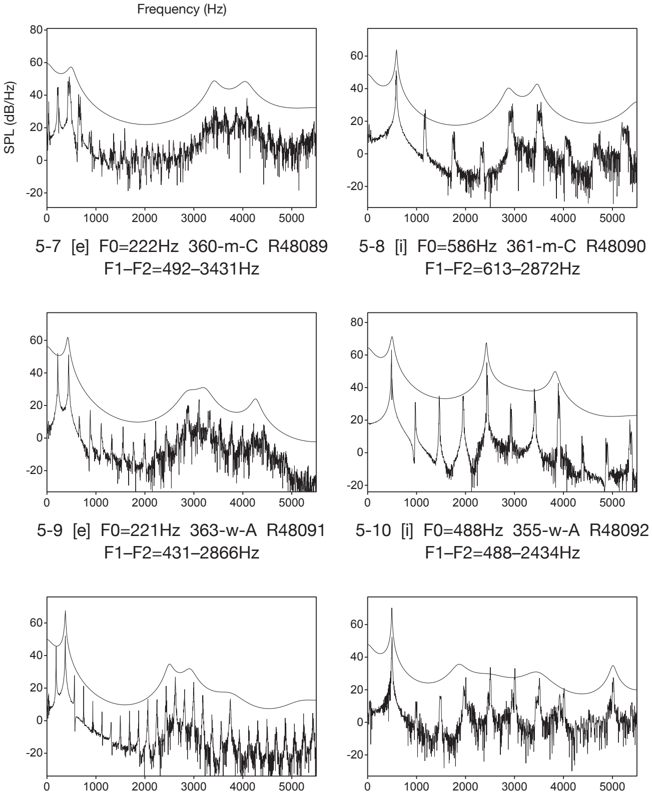
5-3 [o] F0=211Hz 34-w-A R17289
F1=431Hz

5-4 [u] F0=503Hz 367-w-A R48096
F1=503Hz

5-5 [o] F0=123Hz 92-m-A R29021
F1=374Hz

5-6 [u] F0=507Hz 90-m-A R29356
F1=505Hz

(Figure 5, continuation)



5-7 [e] F0=222Hz 360-m-C R48089
F1–F2=492–3431Hz

5-8 [i] F0=586Hz 361-m-C R48090
F1–F2=613–2872Hz

5-9 [e] F0=221Hz 363-w-A R48091
F1–F2=431–2866Hz

5-10 [i] F0=488Hz 355-w-A R48092
F1–F2=488–2434Hz

5-11 [e] F0=187Hz 357-m-A R48093
F1–F2=368–2507Hz

5-12 [i] F0=496Hz 369-m-A R48094
F1–F2=497–2331Hz

M8.2 Vowel Perception at Fundamental Frequencies above Statistical Values of the Respective First Formant Frequency

Figure 6 shows intelligible high-pitched sounds of the vowels /y, e, ø, ɛ, o/ at F0 of c. 750 Hz, and Figure 7 exhibits intelligible high-pitched sounds of the corner vowels /i, a, u/ at F0 of c. 850 Hz. Note again the pronounced spectral differences for these high-pitched sounds of different vowels supporting the thesis of a parallelism between differences in perceived vowel quality and related acoustic differences, that is, the thesis of vowel-specific harmonic spectra.

Figures 8 to 10 show examples of speech extracts of untrained speakers, journalists, TV hosts and actresses and actors, which manifest pitch contours for utterances of single speakers exceeding age- and gender-related statistical F1 of the vowels /i, y, u/ (450 Hz for children, 400 Hz for women and 350 Hz for men). The ranges of F0 indicated—overall ranges for the speech sounds of a single speaker or a group of speakers (see below)—were determined acoustically in terms of approximations by listening to the sounds. (Please ignore some errors in the graphics exceeding the verified ranges given below. These errors are due, for example, to background noise or music, or the sound of an audience or to automatic pitch calculation.) The order of presentation within a figure accords, firstly, to the number of examples per speaker or a group of speakers, and secondly, to the identification number of the speaker.

Figure 8 shows pitch contours of speech extracts produced by untrained speakers, journalists, TV hosts and actresses talking on TV (not acting), to experience in every day life:

- The examples for speaker 172 (see pitch contours 8-1 to 8-3) relates to extracts of a woman selling grilled chicken in a market in Paris. Overall range of F0 = c. 220–700 Hz (excluding very high-pitched exclamations).
- The examples for the two speakers subsumed under the ID number 379 and for the speaker 380 (see pitch contours 8-4 to 8-6) relate to extracts of two American women and one American man demonstrating infant child directed speech. Overall range of F0 = c. 200–800 Hz for the women (except one higher peak at c. 1 kHz) and c. 150–600 Hz for the man.
- The examples for speaker 336 (see pitch contours 8-7 and 8-8, the latter from 0.7 to 2.5 sec.) relate to extracts of a female Indonesian singer talking in a TV show and to an exclamation of her name during the show. Overall range of F0 = c. 350–950 Hz.

- The two examples for the speakers subsumed under the ID number 348 (see pitch contours 8-9 and 8-10) relate to extracts of two female TV hosts announcing the results of a singing contest (announcements in English). Overall range of F0 = c. 200–700 Hz.
- The example for speaker 135 (see pitch contour 8-11) relates to two sentences of a boy (age 6). Range of F0 = c. 220–600 Hz.
- The example for speaker 174 (see pitch contour 8-12) relates to an extract of a female North American journalist speaking on television. Range of F0 = c. 175–600 Hz.
- The example for speaker 217 (see pitch contour 8-13) relates to an extract of a North American woman talking about her child on television. Range of F0 = c. 160–550 Hz.
- The example for speaker 220 (see pitch contour 8-14) relates to an extract of a female French doctor talking on television. Range of F0 = c. 250–520 Hz.
- The example for speaker 238 (see pitch contour 8-15) relates to an extract of a male French TV host. Range of F0 = c. 130–420 Hz (exceeding only gender-related statistical F1 of the vowels /i, y, u/).
- The example for speaker 383 (see pitch contour 8-16) relates to an extract of a French woman talking on television in a TV spot. Range of F0 = c. 220–830 Hz.
- The example for two speakers subsumed under the ID number 379 (see pitch contour 8-17) relates to an extract of a female French journalist (first part) questioning a French woman on the street, and the answer of the latter (second part). Overall range of F0 for the utterances of both women = c. 230–600 Hz.

Figure 9 shows pitch contours of speech extracts of performing actresses (film, comic, voice-over, dubbing):

- The example for speaker 216 (see pitch contours 9-1 and 9-6) relates to extracts of a female Swiss narrator of fairy tales. Overall range of F0 = c. 150–900 Hz.
- The examples for speaker 177 (see pitch contours 9-7 to 9-9) relate to extracts of a French comic actress performing on stage. Overall range of F0 = c. 180–780 Hz.
- The examples for speaker 178 (see pitch contours 9-10 to 9-12) relate to extracts of another French comic actress performing on stage. Overall range of F0 = c. 200–850 Hz.
- The examples for speaker 212 (see pitch contours 9-13 to 9-15) relate to extracts of the speech of a French actress in a cartoon. Overall range of F0 = c. 300–700 Hz.

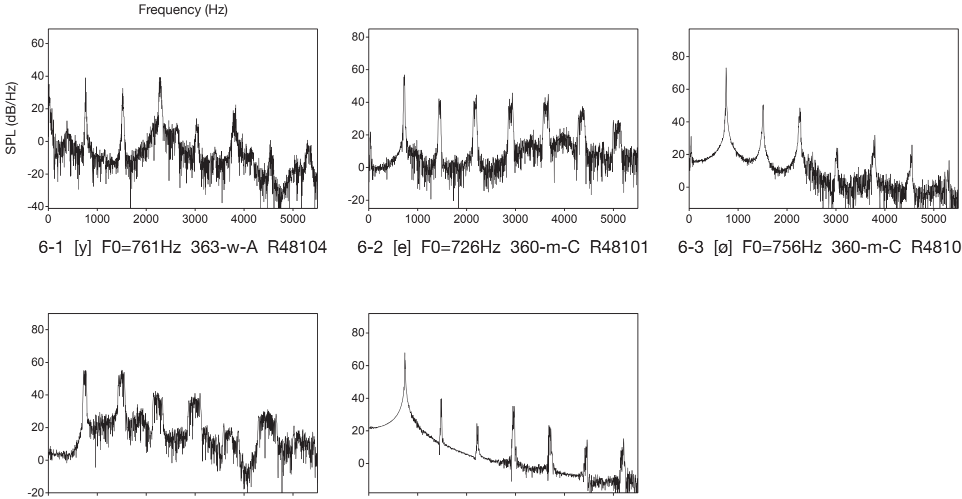
- The examples for speakers 251 (see pitch contours 9-16 to 9-18) relate to extracts of two British actresses performing as the voices of the two main characters in a computer-animated fantasy film. Overall range of F0 = c. 150–800 Hz.
- The examples for speaker 276 (see pitch contours 9-19 to 9-21) relate to extracts of a French comedy actress performing on stage. Overall range of F0 = c. 400–780 Hz.
- The example for speaker 175 (see pitch contour 9-22) relates to an extract of a North American actress performing as a female character in a film. Range of F0 = c. 270–700 Hz (excluding one high-pitched exclamation at F0 of c. 880 Hz).
- The example for speaker 223 (see pitch contour 9-23) relates to an extract of a German actress dubbing a female character in a film. Range of F0 = c. 220–780 Hz (excluding one high-pitched exclamation at the end).
- The example for speaker 234 (see pitch contour 9-24) relates to an extract of a French comic actress performing on stage. Range of F0 = c. 200–850 Hz.
- The example for speaker 258 (see pitch contour 9-25) relates to an extract of a French actress performing as the voice of a female character in an animation film. Range of F0 = c. 220–780 Hz.
- The example for speaker 275 (see pitch contour 9-26) relates to an extract of a German comic actress performing on stage. Range of F0 = c. 180–850 Hz.
- The example for speaker 291 (see pitch contour 9-27) relates to an extract of a British actress performing in a fantasy film. Range of F0 = c. 100–700 Hz.
- The example for speaker 296 (see pitch contour 9-28) relates to an extract of a German comic actress. Range of F0 = c. 150–600 Hz.
- The example for speaker 350 (see pitch contour 9-29) relates to an extract of a North American actress performing as a female character in a film. Range of F0 = c. 160–900 Hz (excluding some very high-pitched exclamations).
- The example for speaker 398 (see pitch contour 9-30) relates to an extract of a North American actress performing as a female character in a TV series. Range of F0 = c. 300–980 Hz.

Figure 10 shows pitch contours of speech extracts of performing actors (film, comic, voice-over, dubbing):

- The examples for speaker 225 (see pitch contours 10-1 to 10-4) relate to speech extracts of a Swiss comic actor performing as a female character. Overall range of F0 = c. 220–780 Hz.
- The examples for speaker 163 (see pitch contours 10-5 to 10-7) relate to extracts of an Indonesian comic actor performing on stage in a Drama Gong. Overall range of F0 = c. 300–600 Hz.
- The examples for speaker 169 (see pitch contours 10-8 and 10-10) relate to extracts of a German actor dubbing a male character in a film. Overall range of F0 = c. 100–700 Hz.
- The examples for speaker 214 (see pitch contours 10-11 to 10-13) relate to extracts of a Japanese Kabuki actor. Overall range of F0 = c. 250–700 Hz.
- The examples for speaker 297 (see pitch contours 10-14 to 10-16) relate to extracts of speech of another Swiss comic actor performing in a TV show. Overall range of F0 = c. 130–620 Hz.
- The examples for speaker 194 (see pitch contours 10-17 and 10-18) relate to extracts of a French comic actor performing on stage. Overall range of F0 = c. 130–700 Hz.
- The example for speaker 394 (see pitch contours 10-19 and 10-20) relates to extracts of two French actors performing as the voices of male characters in an animation film. Overall range of F0 = c. 310–650 Hz.
- The example for speaker 171 (see pitch contour 10-21) relates to extracts of speech of a German actor dubbing the voice of a male character. Range of F0 = c. 180–550 Hz.
- The example for speaker 274 (see pitch contour 10-22) relates to extracts of speech of a Swiss actor performing as ventriloquist. Range of F0 = c. 120–600 Hz.
- The example for speaker 294 (see pitch contour 10-23) relates to an extract of speech of a North American actor performing as the voice of a female character in a comedy-variety film. Range of F0 = c. 200–800 Hz.
- The example for speaker 351 (see pitch contour 10-24) relates to an extract of speech of a German comic actor performing in a TV show. Range of F0 = c. 150–580 Hz (excluding one high-pitched exclamation at F0 of c. 780 Hz).

For earlier accounts, see Maurer and Landis (1996, 2000), Maurer, Mok, Friedrichs, and Dellwo (2014), Friedrichs, Maurer, and Dellwo (2015), Friedrichs, Maurer, Suter, and Dellwo (2015).

Figure 6. Five intelligible sounds of /y, e, ø, ɛ, o/ produced by children and women at F0 in the range of 700–800 Hz.



6-1 [y] F0=761Hz 363-w-A R48104

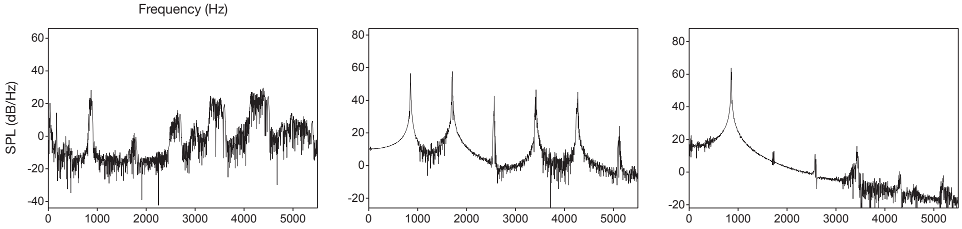
6-2 [e] F0=726Hz 360-m-C R48101

6-3 [ø] F0=756Hz 360-m-C R48100

6-4 [ɛ] F0=746Hz 355-w-A R48103

6-5 [o] F0=739Hz 391-w-A R48358

Figure 7. Three intelligible sounds of the corner vowels /i, a, u/ produced by women at F0 of c. 850 Hz.

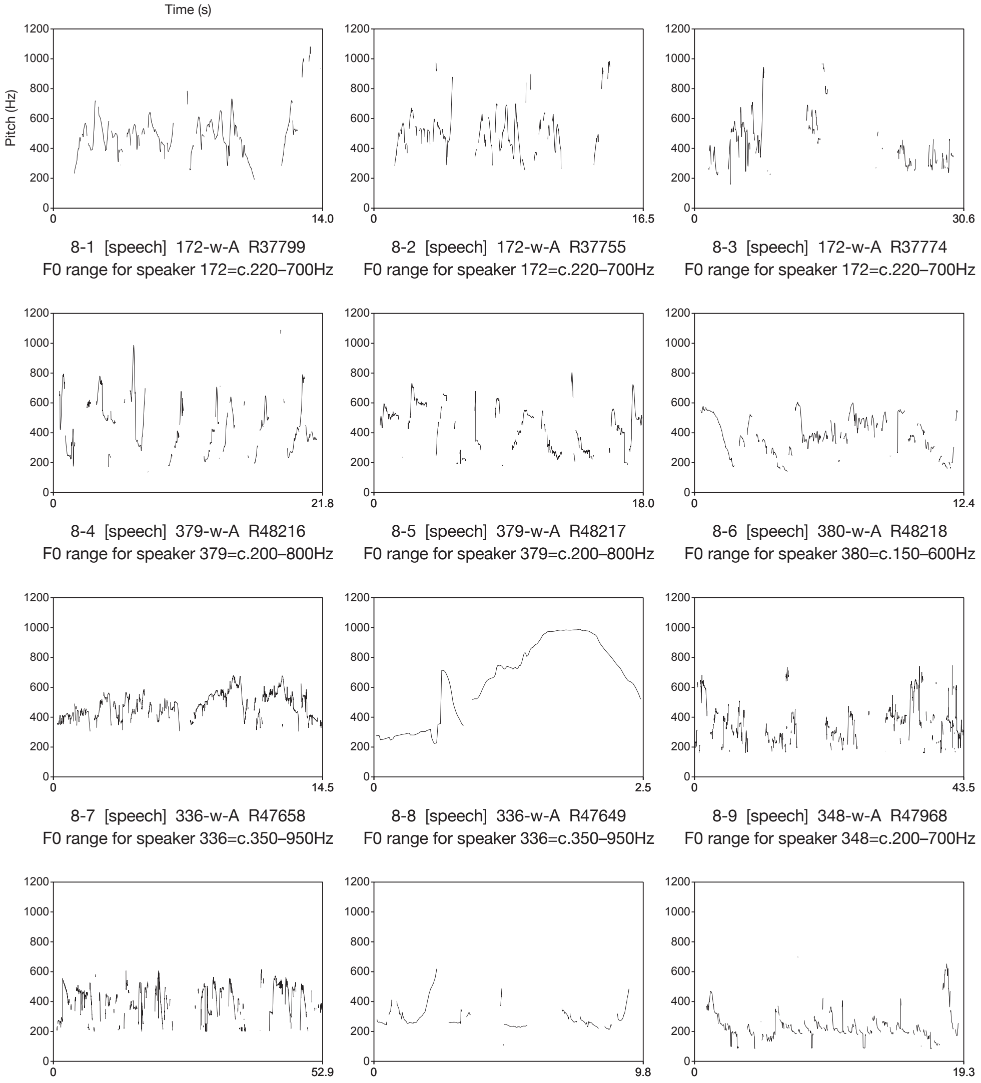


7-1 [i] F0=876Hz 376-w-A R48099

7-2 [a] F0=853Hz 363-w-A R48098

7-3 [u] F0=859Hz 363-w-A R48097

Figure 8. Pitch contours of speech extracts produced by untrained speakers, journalists, TV hosts and actresses talking on TV (not acting), to experience in every day life.



8-1 [speech] 172-w-A R37799
F0 range for speaker 172=c.220–700Hz

8-2 [speech] 172-w-A R37755
F0 range for speaker 172=c.220–700Hz

8-3 [speech] 172-w-A R37774
F0 range for speaker 172=c.220–700Hz

8-4 [speech] 379-w-A R48216
F0 range for speaker 379=c.200–800Hz

8-5 [speech] 379-w-A R48217
F0 range for speaker 379=c.200–800Hz

8-6 [speech] 380-w-A R48218
F0 range for speaker 380=c.150–600Hz

8-7 [speech] 336-w-A R47658
F0 range for speaker 336=c.350–950Hz

8-8 [speech] 336-w-A R47649
F0 range for speaker 336=c.350–950Hz

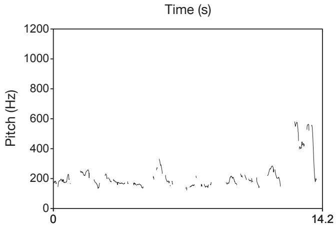
8-9 [speech] 348-w-A R47968
F0 range for speaker 348=c.200–700Hz

8-10 [speech] 348-w-A R47972
F0 range for speaker 348=c.200–700Hz

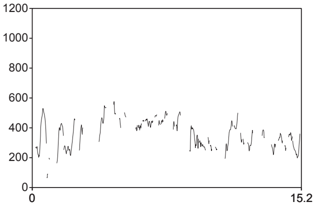
8-11 [speech] 135-m-C R43741
F0 range for speaker 135=c.220–600Hz

8-12 [speech] 174-w-A R44696
F0 range for speaker 174=c.175–600Hz

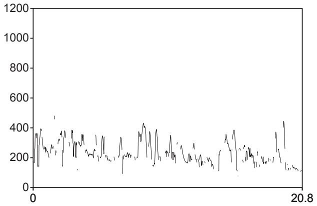
(Figure 8, continuation)



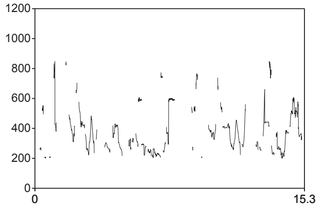
8-13 [speech] 217-w-A R43252
F0 range for speaker 217=c.160–550Hz



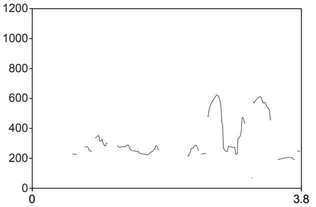
8-14 [speech] 220-w-A R43743
F0 range for speaker 220=c.250–520Hz



8-15 [speech] 238-m-A R43988
F0 range for speaker 238=c.130–420Hz

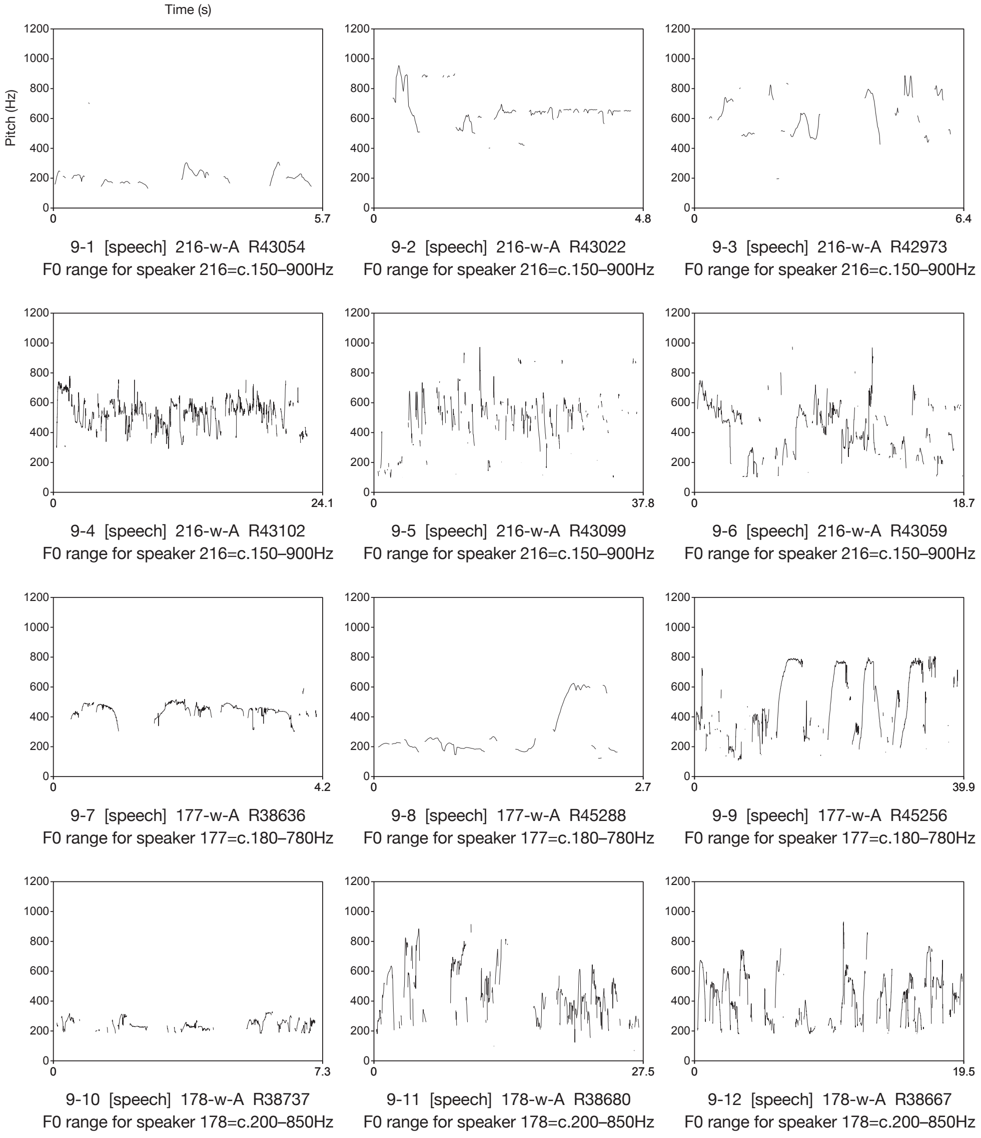


8-16 [speech] 383-w-A R48251
F0 range for speaker 383=c.220–830Hz



8-17 [speech] 379-w-A R48219
F0 range for speaker 379=c.230–600Hz

Figure 9. Pitch contours of extracts of speech produced by actresses while performing (film, comic, voice-over, dubbing).



9-1 [speech] 216-w-A R43054
F0 range for speaker 216=c.150–900Hz

9-2 [speech] 216-w-A R43022
F0 range for speaker 216=c.150–900Hz

9-3 [speech] 216-w-A R42973
F0 range for speaker 216=c.150–900Hz

9-4 [speech] 216-w-A R43102
F0 range for speaker 216=c.150–900Hz

9-5 [speech] 216-w-A R43099
F0 range for speaker 216=c.150–900Hz

9-6 [speech] 216-w-A R43059
F0 range for speaker 216=c.150–900Hz

9-7 [speech] 177-w-A R38636
F0 range for speaker 177=c.180–780Hz

9-8 [speech] 177-w-A R45288
F0 range for speaker 177=c.180–780Hz

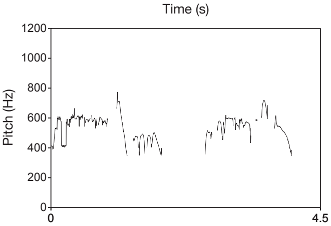
9-9 [speech] 177-w-A R45256
F0 range for speaker 177=c.180–780Hz

9-10 [speech] 178-w-A R38737
F0 range for speaker 178=c.200–850Hz

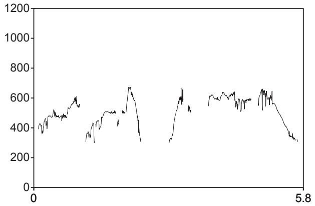
9-11 [speech] 178-w-A R38680
F0 range for speaker 178=c.200–850Hz

9-12 [speech] 178-w-A R38667
F0 range for speaker 178=c.200–850Hz

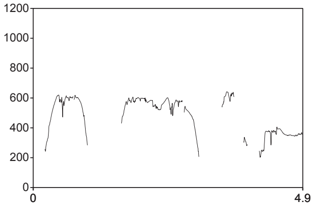
(Figure 9, continuation)



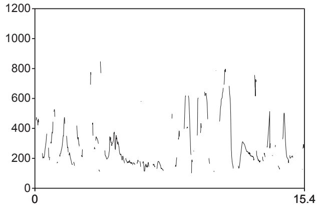
9-13 [speech] 212-w-A R42755
F0 range for speaker 212=c.300–700Hz



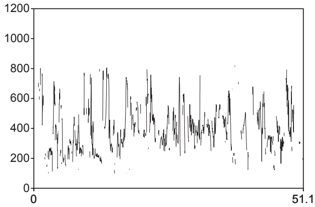
9-14 [speech] 212-w-A R42737
F0 range for speaker 212=c.300–700Hz



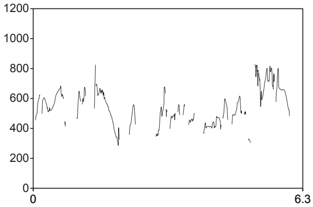
9-15 [speech] 212-w-A R42747
F0 range for speaker 212=c.300–700Hz



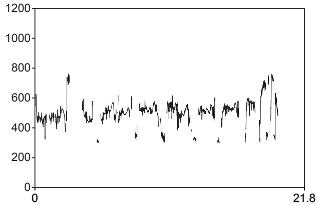
9-16 [speech] 251-w-A R44572
F0 range for speaker 251=c.150–800Hz



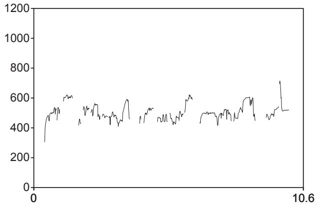
9-17 [speech] 251-w-A R44577
F0 range for speaker 251=c.150–800Hz



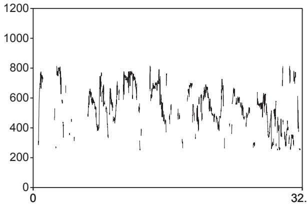
9-18 [speech] 251-w-A R44569
F0 range for speaker 251=c.150–800Hz



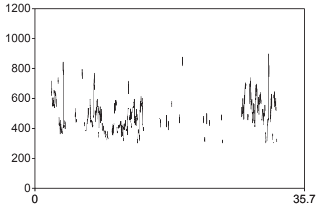
9-19 [speech] 276-w-A R45758
F0 range for speaker 276=c.400–780Hz



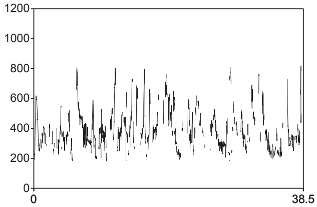
9-20 [speech] 276-w-A R45757
F0 range for speaker 276=c.400–780Hz



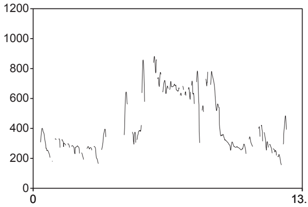
9-21 [speech] 276-w-A R45762
F0 range for speaker 276=c.400–780Hz



9-22 [speech] 175-w-A R46869
F0 range for speaker 175=c.270–700Hz

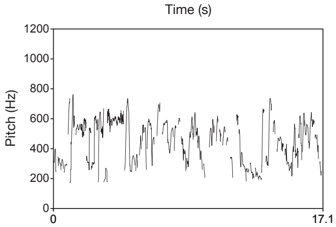


9-23 [speech] 223-w-A R43748
F0 range for speaker 223=c.220–780Hz

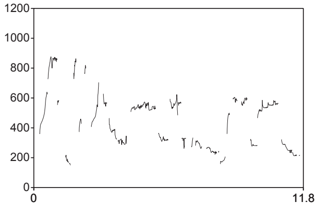


9-24 [speech] 234-w-A R43907
F0 range for speaker 234=c.200–850Hz

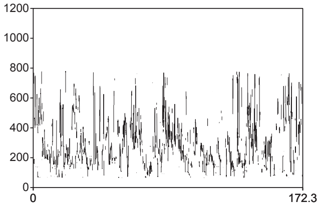
(Figure 9, continuation)



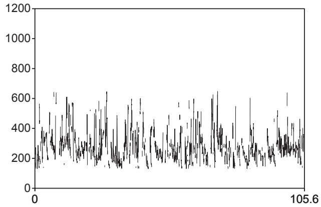
9-25 [speech] 258-w-A R46233
F0 range for speaker 258=c.220–780Hz



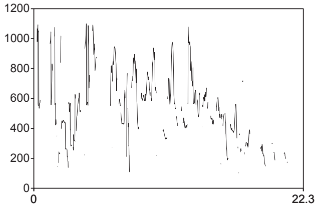
9-26 [speech] 275-w-A R45728
F0 range for speaker 275=c.180–850Hz



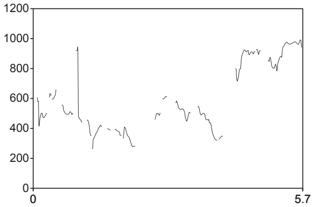
9-27 [speech] 291-w-A R46797
F0 range for speaker 291=c.100–700Hz



9-28 [speech] 296-w-A R46819
F0 range for speaker 296=c.150–600Hz

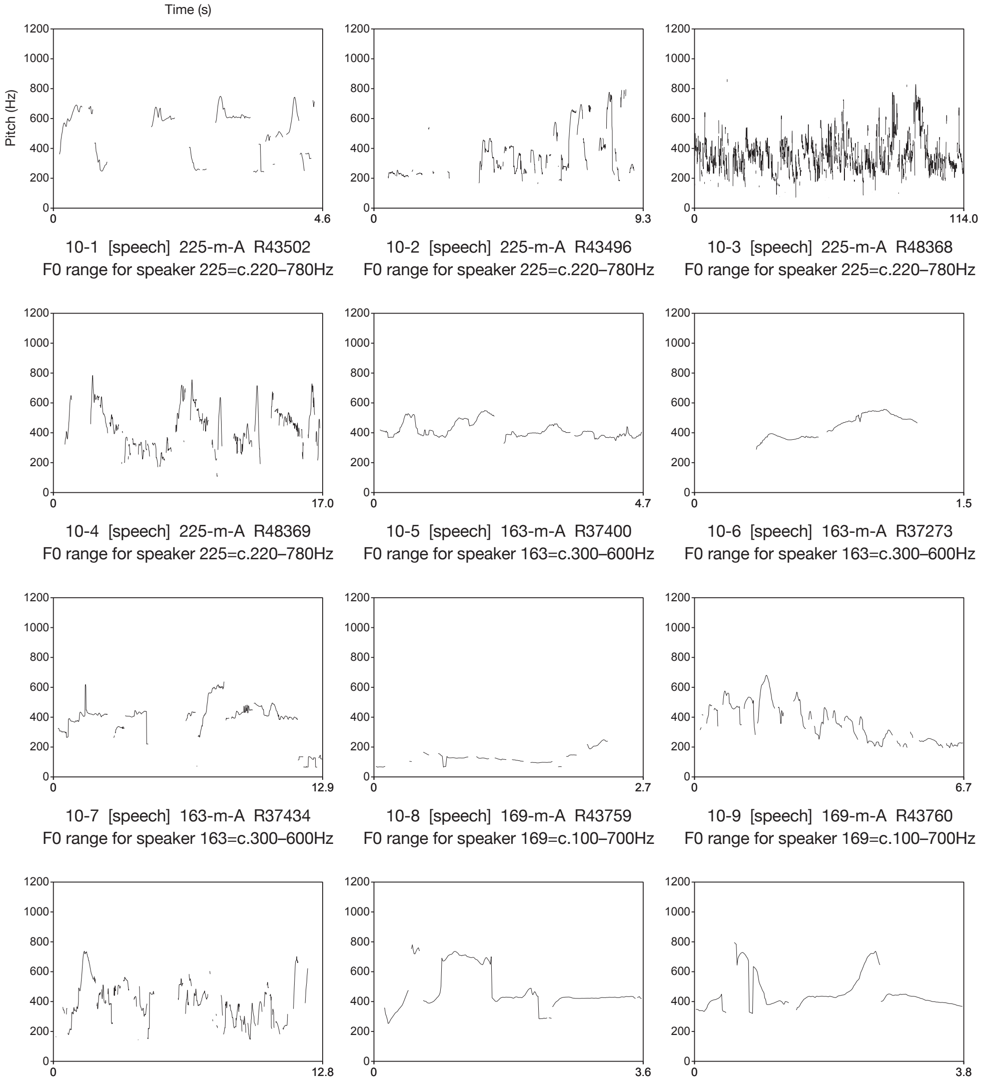


9-29 [speech] 350-w-A R47975
F0 range for speaker 350=c.160–900Hz



9-30 [speech] 398-w-A R48431
F0 range for speaker 398=c.300–980Hz

Figure 10. Pitch contours of extracts of speech produced by actors while performing (film, comic, voice-over, dubbing).



10-1 [speech] 225-m-A R43502
F0 range for speaker 225=c.220–780Hz

10-2 [speech] 225-m-A R43496
F0 range for speaker 225=c.220–780Hz

10-3 [speech] 225-m-A R48368
F0 range for speaker 225=c.220–780Hz

10-4 [speech] 225-m-A R48369
F0 range for speaker 225=c.220–780Hz

10-5 [speech] 163-m-A R37400
F0 range for speaker 163=c.300–600Hz

10-6 [speech] 163-m-A R37273
F0 range for speaker 163=c.300–600Hz

10-7 [speech] 163-m-A R37434
F0 range for speaker 163=c.300–600Hz

10-8 [speech] 169-m-A R43759
F0 range for speaker 169=c.100–700Hz

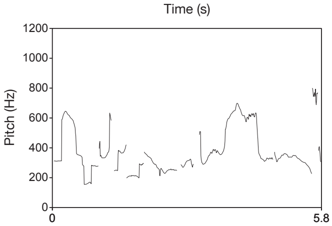
10-9 [speech] 169-m-A R43760
F0 range for speaker 169=c.100–700Hz

10-10 [speech] 169-m-A R43753
F0 range for speaker 169=c.100–700Hz

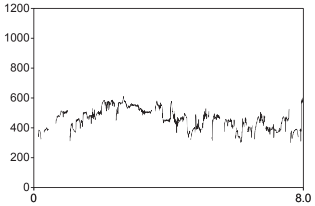
10-11 [speech] 214-m-A R42792
F0 range for speaker 214=c.250–700Hz

10-12 [speech] 214-m-A R42791
F0 range for speaker 214=c.250–700Hz

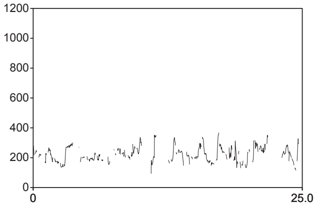
(Figure 10, continuation)



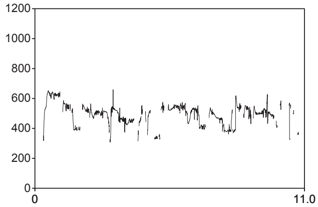
10-13 [speech] 214-m-A R42790
F0 range for speaker 214=c.250–700Hz



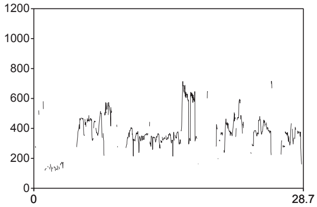
10-14 [speech] 297-m-A R46877
F0 range for speaker 297=c.130–620Hz



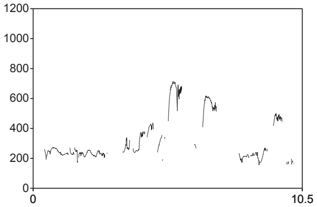
10-15 [speech] 297-m-A R46871
F0 range for speaker 297=c.130–620Hz



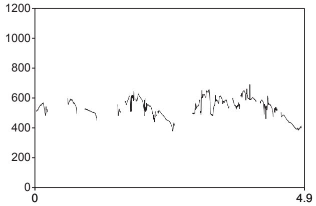
10-16 [speech] 297-m-A R46873
F0 range for speaker 297=c.130–620Hz



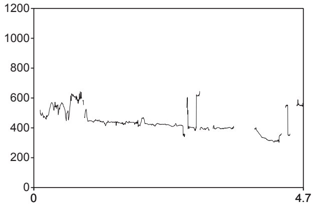
10-17 [speech] 194-m-A R43592
F0 range for speaker 194=c.130–700Hz



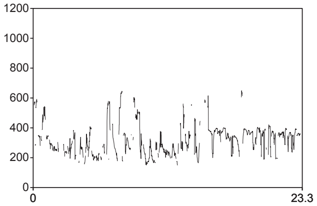
10-18 [speech] 194-m-A R43667
F0 range for speaker 194=c.130–700Hz



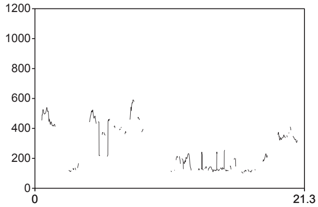
10-19 [speech] 394-m-A R44518
F0 range for speaker 394=c.310–650Hz



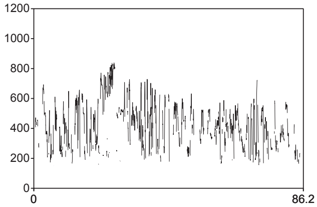
10-20 [speech] 394-m-A R44517
F0 range for speaker 394=c.310–650Hz



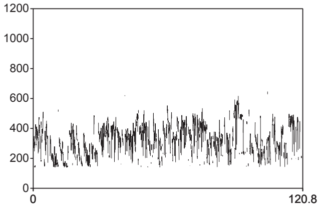
10-21 [speech] 171-m-A R43746
F0 range for speaker 171=c.180–550Hz



10-22 [speech] 274-m-A R45648
F0 range for speaker 274=c.120–600Hz



10-23 [speech] 294-m-A R46801
F0 range for speaker 294=c.200–800Hz

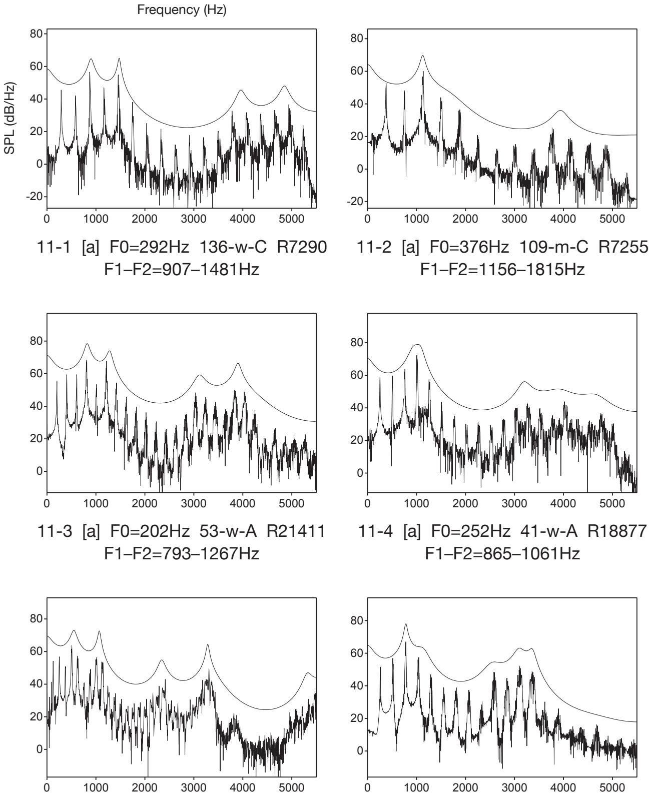


10-24 [speech] 351-m-A R47989
F0 range for speaker 351=c.150–580Hz

M8.3 “Inversions” of Relative Spectral Energy Maxima and Minima and “Inverse” Formant Patterns in Sounds of Individual Vowels

For each of the vowels /a–ɑ, o, u/ and for each speaker group, Figures 11 to 13 show pairs of sounds produced at different fundamental frequencies exhibiting “inverse” relative spectral maxima and minima in terms of “inverse” spectral envelope curves ≤ 1.5: whereas a relative minimum in the spectral envelope occurs for one sound of a pair, a peak for the other sound is manifest, and vice versa; however, the perceived vowel quality is maintained. The same holds true for comparisons of the respective calculated filter curves and, for most cases, for comparisons of patterns of manifest formants.

Figure 11. Sounds of /a–ɑ/, produced at different F0 by children, women and men, which exhibit “inverse” relative spectral maxima and minima in terms of “inverse” spectral envelope curves ≤ 1.5 kHz.



11-1 [a] F0=292Hz 136-w-C R7290
F1–F2=907–1481Hz

11-2 [a] F0=376Hz 109-m-C R7255
F1–F2=1156–1815Hz

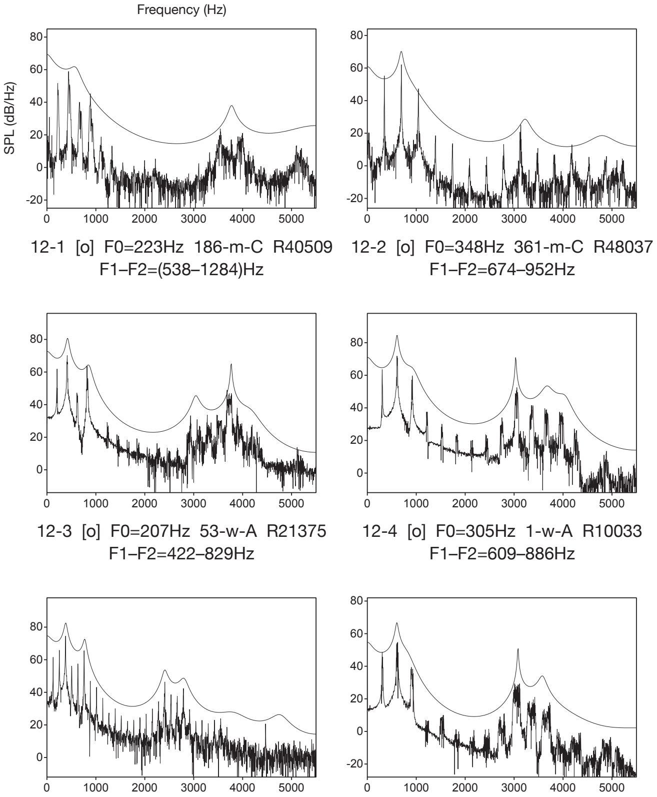
11-3 [a] F0=202Hz 53-w-A R21411
F1–F2=793–1267Hz

11-4 [a] F0=252Hz 41-w-A R18877
F1–F2=865–1061Hz

11-5 [a] F0=126Hz 97-m-A R29962
F1–F2=553–1071Hz

11-6 [a] F0=259Hz 358-m-A R48261
F1–F2=776–1120Hz

Figure 12. Sounds of /o/, produced at different F0 by children, women and men, which exhibit "inverse" relative spectral maxima and minima in terms of "inverse" spectral envelope curves ≤ 1.5 kHz.



12-1 [o] F0=223Hz 186-m-C R40509
F1–F2=(538–1284)Hz

12-2 [o] F0=348Hz 361-m-C R48037
F1–F2=674–952Hz

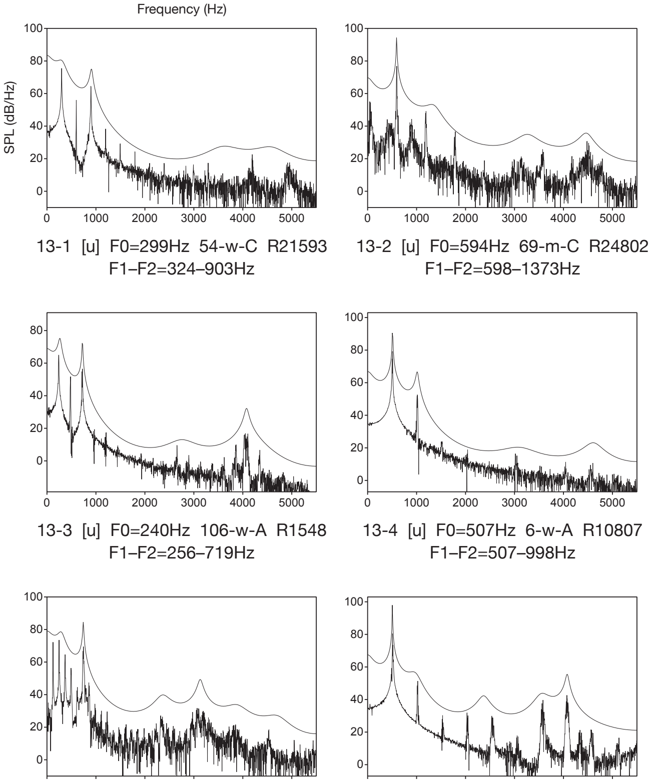
12-3 [o] F0=207Hz 53-w-A R21375
F1–F2=422–829Hz

12-4 [o] F0=305Hz 1-w-A R10033
F1–F2=609–886Hz

12-5 [o] F0=127Hz 90-m-A R28531
F1–F2=387–764Hz

12-6 [o] F0=306Hz 2-m-A R7801
F1–F2=599–771Hz

Figure 13. Sounds of /u/, produced at different F0 by children, women and men, which exhibit “inverse” relative spectral maxima and minima in terms of “inverse” spectral envelope curves ≤ 1.5 kHz.



13-1 [u] F0=299Hz 54-w-C R21593
F1–F2=324–903Hz

13-2 [u] F0=594Hz 69-m-C R24802
F1–F2=598–1373Hz

13-3 [u] F0=240Hz 106-w-A R1548
F1–F2=256–719Hz

13-4 [u] F0=507Hz 6-w-A R10807
F1–F2=507–998Hz

13-5 [u] F0=123Hz 39-m-A R18503
F1–F2=296–749Hz

13-6 [u] F0=509Hz 44-m-A R22732
F1–F2=507–937Hz

M9 Ambiguous Correspondence between Vowels and Patterns of Relative Spectral Energy Maxima or Formant Patterns or Complete Spectral Envelopes

M9.1 Ambiguous Patterns of Relative Spectral Energy Maxima and Ambiguous Formant Patterns

Figures 1 to 21 show series of sounds of different vowels produced at different F0 but exhibiting similar patterns of relative spectral energy maxima and/or similar patterns of calculated formant frequencies within their supposed vowel-specific frequency range related to statistical F1 and F2. In all cases, the actual differences of the patterns for the sounds of different vowels presented in a single series are far smaller than the observable differences (variations) of corresponding patterns for sounds of a single vowel.—In some series that include sounds at high fundamental frequencies, the overall spectral envelopes and the harmonic spectra are considered for the comparison in question.

For each series, roughly estimated average frequencies of the two lower relative spectral energy maxima and/or of the calculated frequencies F1–F2 are given below in terms of model patterns for the sounds compared. Exceptions concern a few comparisons of sounds of back vowels, for which only a single spectral peak is manifest in the sound spectra (for these comparisons, the corresponding peak frequency is given), and an additional exception concerns a comparison of sounds /a–ɑ, u/, for which only the spectrum as such > 1.5 kHz is considered.

The first sound series shown include sounds of the vowels /a–ɑ, o, u/, divided into two groups, one presenting sounds of different speakers, the other presenting sounds of single speakers. The second series shown include sounds of front vowels, again divided into the two groups mentioned. (Figures 9 and 11 include exceptions that illustrate the ambiguity discussed for sounds of different and of single speakers.) Within a series, the sounds are organised according to fundamental frequency.

Comparisons of sounds of back vowels and of /a–ɑ/ produced by different speakers:

Figure 1 Sounds of /a–ɑ, o, u/; model pattern of spectral peaks and/or of calculated formant frequencies = 600–1200 Hz

Figure 2 Sounds of /a–ɑ, o, u/; model pattern of spectral peaks and/or of calculated formant frequencies = 600–1050 Hz

Figure 3 Sounds of /a–ɑ, o/; model pattern of spectral peaks and/or of calculated formant frequencies = 660–1320 Hz

Sounds of /u/ are included in the first three series because the first harmonic corresponds to F1 of the model pattern in question; however, no clear spectral indication can be found for F2 even if LPC analysis gives a (weak) second formant at a frequency level which corresponds to the model pattern of a series.

Comparisons of sounds of back vowels and of /a–ɑ/ produced by single speakers:

Figure 4 Three comparisons of sounds of /a–ɑ, o, u/ produced by a man and two women; model pattern of spectral peaks and/or of calculated formant frequencies = 600–1200 Hz

Figure 5 Two comparisons of sounds of /a–ɑ, o/ produced by a man (sounds sung by a tenor); model pattern of spectral peaks and/or of calculated formant frequencies = 600–1200 Hz for the first comparison, similar spectral peaks and spectral envelopes for the second comparison

Figure 6 Sounds of /a–ɑ/ and of /u/ produced by a woman which exhibit comparable spectral envelopes < 1.5 kHz

Figure 7 Sounds of /ɔ, o, u/ produced by a woman; model pattern of spectral peaks and/or of calculated formant frequencies = one clear peak at c. 550 Hz (exceptionally, sounds of the vowel /ɔ/ are included in order to show a possible shift in perceived vowel quality from /ɔ/ to /o/ related to two levels of F0 of c. 175 Hz and c. 260 Hz)

Figure 8 Two comparisons of sounds of /o, u/ produced by two children (age 12 and 6); model patterns of spectral peaks and/or of calculated formant frequencies = one clear peak at c. 400 Hz (first sound pair) and at c. 520 Hz (second sound pair), respectively.

Comparisons of sounds of front vowels produced by different speakers:

In contrast to many other comparisons presented in this chapter, the ambiguity illustrated in Figures 9 to 11 does not always relate to substantial differences in F0 but also to the configuration of the levels of the harmonics, to the spectrum above F2 and to the levels of calculated formants including F3. This is the case particularly for direct comparisons of sounds of /e/ and of /ø/, and of /i/ and of /y/, respectively. Moreover, a sound produced with creak phonation is exceptionally included into the comparison (see the first vowel spectrum of Figure 9).

Figure 9 Sounds of /ø, e, y, i/; model pattern of spectral peaks and/or of calculated formant frequencies = 330–2000 Hz; note the ambiguity for sounds of /ø, i/ for the single speaker 391 and the ambiguity for the sounds of /ø, y/ for the single speaker 376

Figure 10 Sounds of /ø, e, y, i/; model pattern of spectral peaks and/or of calculated formant frequencies = 350–2150 Hz

Figure 11 Sounds of /ø, e, y, i/; model pattern of spectral peaks and/or of calculated formant frequencies = 420–2150 Hz; note the ambiguity for sounds of /ø, y/ for the single speaker 402

Figure 12 Sounds of /ɛ, e, i/; model pattern of spectral peaks and/or of calculated formant frequencies = 500–2250 Hz

Figure 13 Sounds of /ɛ, e, i/; model pattern of spectral peaks and/or of calculated formant frequencies = 600–2450 Hz

Figure 14 Sounds of /e, i/; model pattern of spectral peaks and/or of calculated formant frequencies = 400–2600 Hz

Figure 15 Sounds of /ɛ, e, y/; model pattern of spectral peaks and/or of calculated formant frequencies = 500–2000 Hz

Figure 16 Sounds of /ɛ, ø, y/; model pattern of spectral peaks and/or of calculated formant frequencies = 430–2000 Hz

Figure 17 Sounds of /ɛ, ø, y/; model pattern of spectral peaks and/or of calculated formant frequencies = 475–1900 Hz

Figure 18 Sounds of /ɛ, y/; model pattern of spectral peaks and/or of calculated formant frequencies = 650–1950 Hz

Comparisons of sounds of front vowels, produced by single speakers:

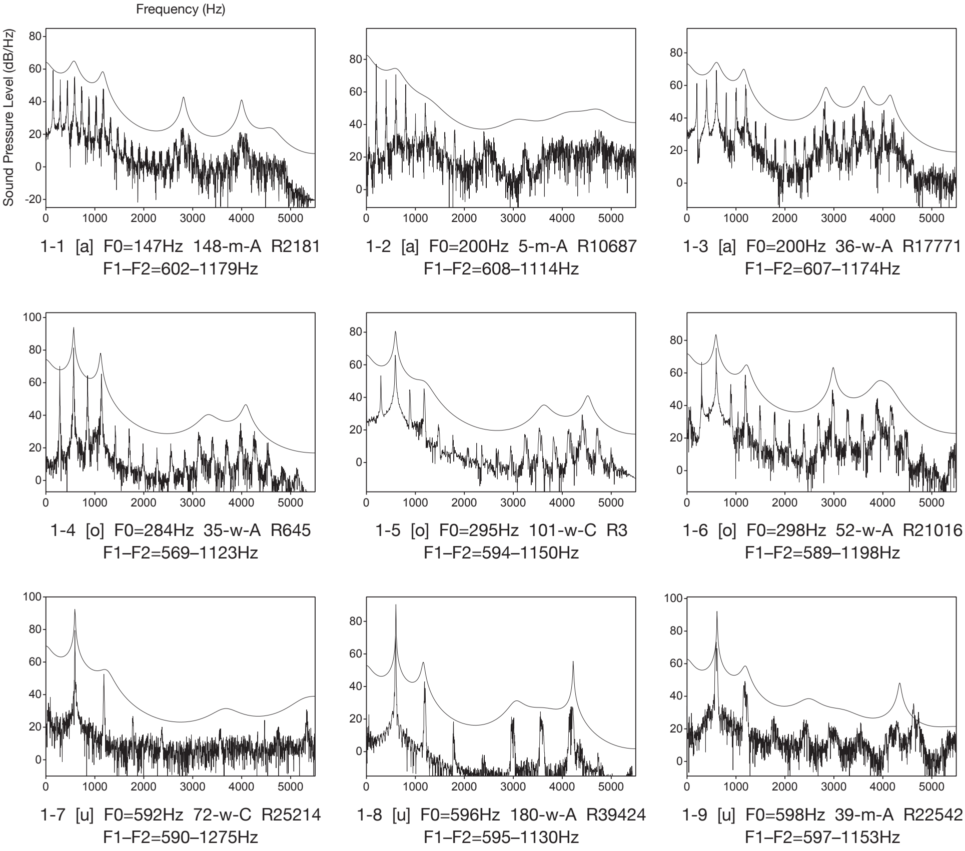
Figure 19 Two comparisons of sounds of /ɛ, e, i/ produced by two women; model patterns of spectral peaks and/or of calculated formant frequencies = 510–2550 Hz and 600–2400 Hz, respectively

Figure 20 Three comparisons of sounds of /e, i/ produced by three children (age 7 to 9); model patterns of spectral peaks and/or of calculated formant frequencies = 450–3000 Hz and 400–3000 Hz, respectively

Figure 21 Three comparisons of sounds of /ø, y/ produced by a man, a woman and a child (age 12); model patterns of spectral peaks and/or of calculated formant frequencies = 320–1600 Hz, 320–2000 Hz and 400–2000 Hz, respectively

For earlier accounts, see Maurer and Landis (2000).

Figure 1. Sounds of /a–ɑ, o, u/ produced by different speakers; related model pattern of spectral peaks and/or of calculated formant frequencies = 600–1200 Hz.



1-1 [a] F0=147Hz 148-m-A R2181
F1–F2=602–1179Hz

1-2 [a] F0=200Hz 5-m-A R10687
F1–F2=608–1114Hz

1-3 [a] F0=200Hz 36-w-A R17771
F1–F2=607–1174Hz

1-4 [o] F0=284Hz 35-w-A R645
F1–F2=569–1123Hz

1-5 [o] F0=295Hz 101-w-C R3
F1–F2=594–1150Hz

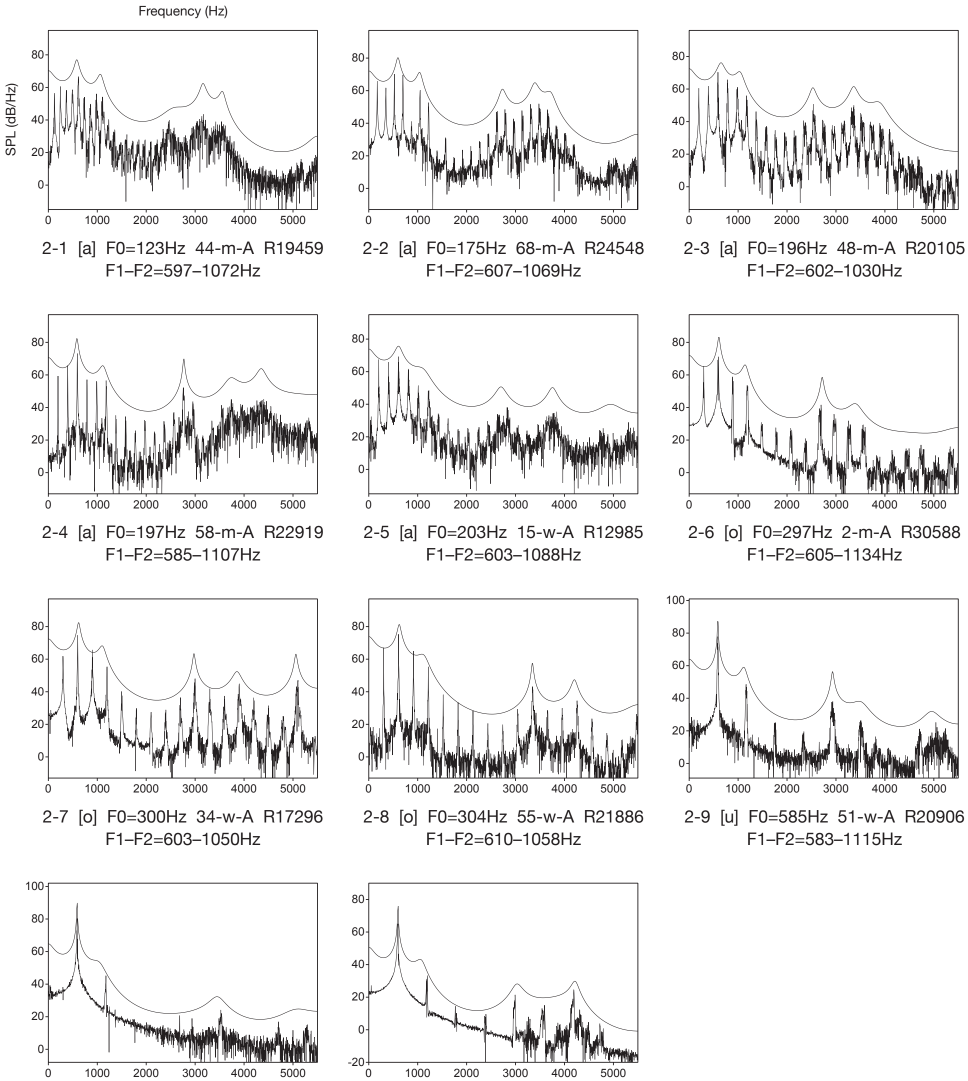
1-6 [o] F0=298Hz 52-w-A R21016
F1–F2=589–1198Hz

1-7 [u] F0=592Hz 72-w-C R25214
F1–F2=590–1275Hz

1-8 [u] F0=596Hz 180-w-A R39424
F1–F2=595–1130Hz

1-9 [u] F0=598Hz 39-m-A R22542
F1–F2=597–1153Hz

Figure 2. Sounds of /a–ɑ, o, u/ produced by different speakers; related model pattern of spectral peaks and/or of calculated formant frequencies = 600–1050 Hz.



2-1 [a] F0=123Hz 44-m-A R19459
F1–F2=597–1072Hz

2-2 [a] F0=175Hz 68-m-A R24548
F1–F2=607–1069Hz

2-3 [a] F0=196Hz 48-m-A R20105
F1–F2=602–1030Hz

2-4 [a] F0=197Hz 58-m-A R22919
F1–F2=585–1107Hz

2-5 [a] F0=203Hz 15-w-A R12985
F1–F2=603–1088Hz

2-6 [o] F0=297Hz 2-m-A R30588
F1–F2=605–1134Hz

2-7 [o] F0=300Hz 34-w-A R17296
F1–F2=603–1050Hz

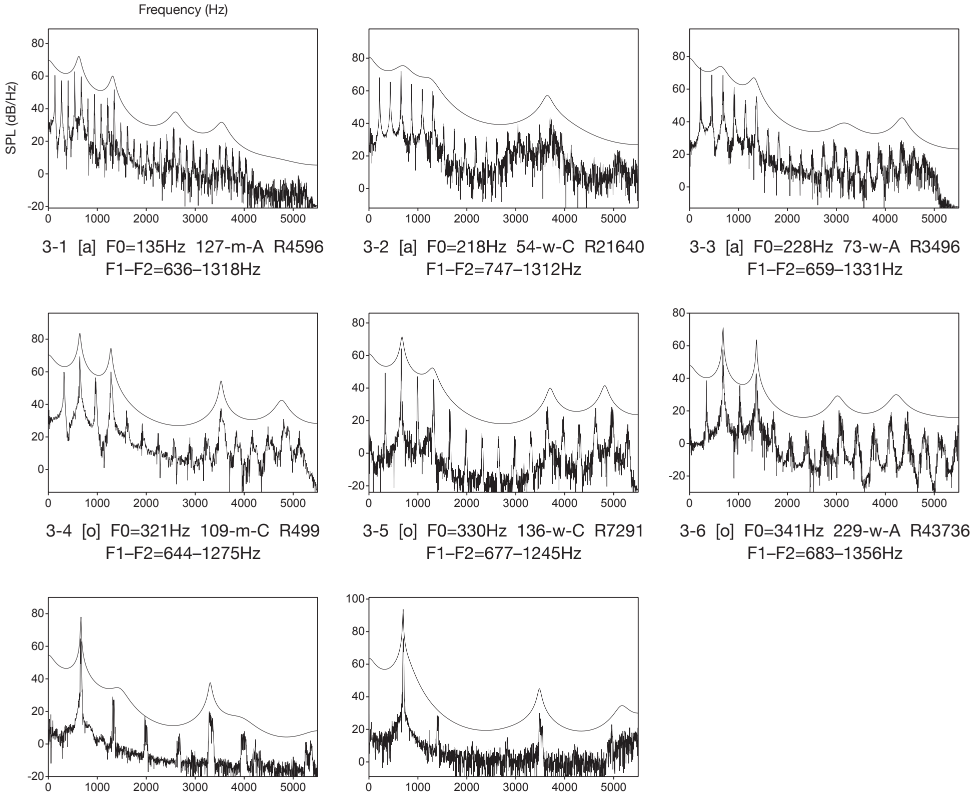
2-8 [o] F0=304Hz 55-w-A R21886
F1–F2=610–1058Hz

2-9 [u] F0=585Hz 51-w-A R20906
F1–F2=583–1115Hz

2-10 [u] F0=587Hz 7-w-A R10886
F1–F2=588–975Hz

2-11 [u] F0=595Hz 180-w-A R39659
F1–F2=596–1024Hz

Figure 3. Sounds of /a–ɑ, o, u/ produced by different speakers; related model pattern of spectral peaks and/or of calculated formant frequencies = 660–1320 Hz.



3-1 [a] F0=135Hz 127-m-A R4596
F1–F2=636–1318Hz

3-2 [a] F0=218Hz 54-w-C R21640
F1–F2=747–1312Hz

3-3 [a] F0=228Hz 73-w-A R3496
F1–F2=659–1331Hz

3-4 [o] F0=321Hz 109-m-C R499
F1–F2=644–1275Hz

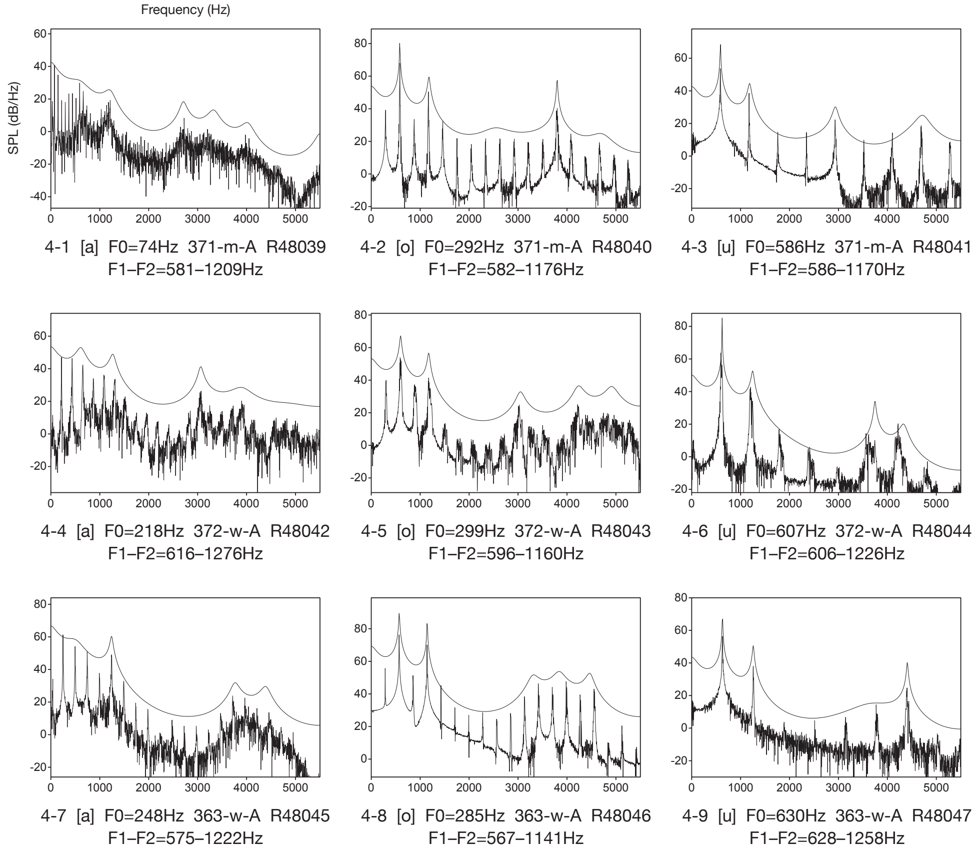
3-5 [o] F0=330Hz 136-w-C R7291
F1–F2=677–1245Hz

3-6 [o] F0=341Hz 229-w-A R43736
F1–F2=683–1356Hz

3-7 [u] F0=663Hz 180-w-A R39603
F1–F2=664–1425Hz

3-8 [u] F0=702Hz 1-w-A R10058
F1–F2=703–(803)Hz

Figure 4. Three comparisons of sounds of /a–ɑ, o, u/ produced by a man and two women; related model pattern of spectral peaks and/or of calculated formant frequencies = 600–1200 Hz.



4-1 [a] F0=74Hz 371-m-A R48039
F1–F2=581–1209Hz

4-2 [o] F0=292Hz 371-m-A R48040
F1–F2=582–1176Hz

4-3 [u] F0=586Hz 371-m-A R48041
F1–F2=586–1170Hz

4-4 [a] F0=218Hz 372-w-A R48042
F1–F2=616–1276Hz

4-5 [o] F0=299Hz 372-w-A R48043
F1–F2=596–1160Hz

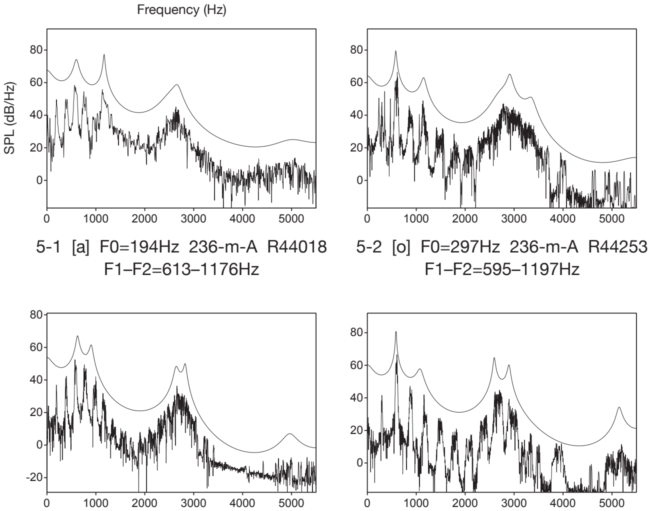
4-6 [u] F0=607Hz 372-w-A R48044
F1–F2=606–1226Hz

4-7 [a] F0=248Hz 363-w-A R48045
F1–F2=575–1222Hz

4-8 [o] F0=285Hz 363-w-A R48046
F1–F2=567–1141Hz

4-9 [u] F0=630Hz 363-w-A R48047
F1–F2=628–1258Hz

Figure 5. Two comparisons of sounds of /a–ɑ, o/ produced by a man (sounds sung by a tenor); related model pattern of spectral peaks and/or of calculated formant frequencies = 600–1200 Hz for the first comparison; similar spectral peaks and spectral envelopes for the second comparison.



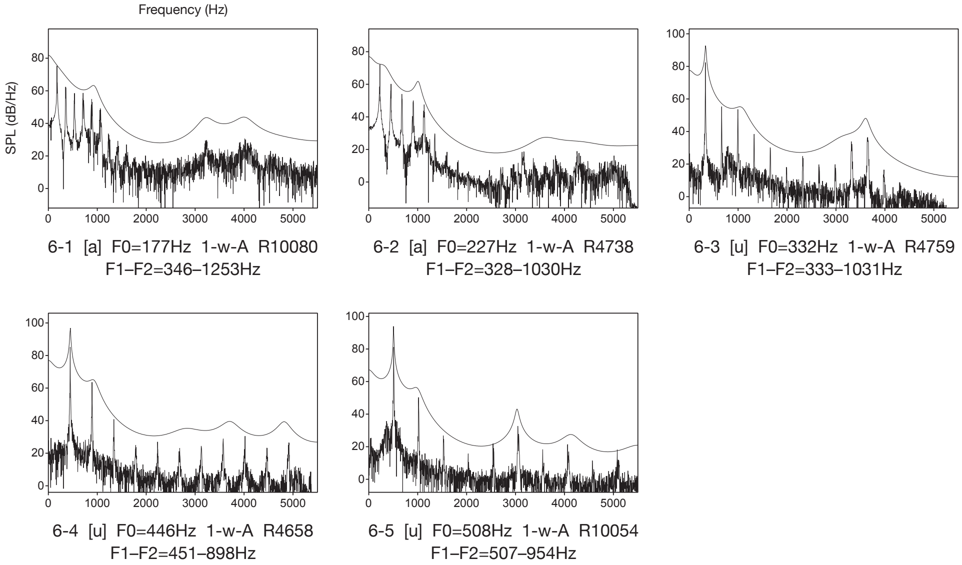
5-1 [a] F0=194Hz 236-m-A R44018
F1–F2=613–1176Hz

5-2 [o] F0=297Hz 236-m-A R44253
F1–F2=595–1197Hz

5-3 [a] F0=194Hz 236-m-A R44235
F1–F2=618–917Hz

5-4 [o] F0=298Hz 236-m-A R44306
F1–F2=604–1093Hz

Figure 6. Sounds of /a–ɑ/ and of /u/, produced by a woman, which exhibit comparable spectral envelopes < 1.5 kHz.



6-1 [a] F0=177Hz 1-w-A R10080
F1–F2=346–1253Hz

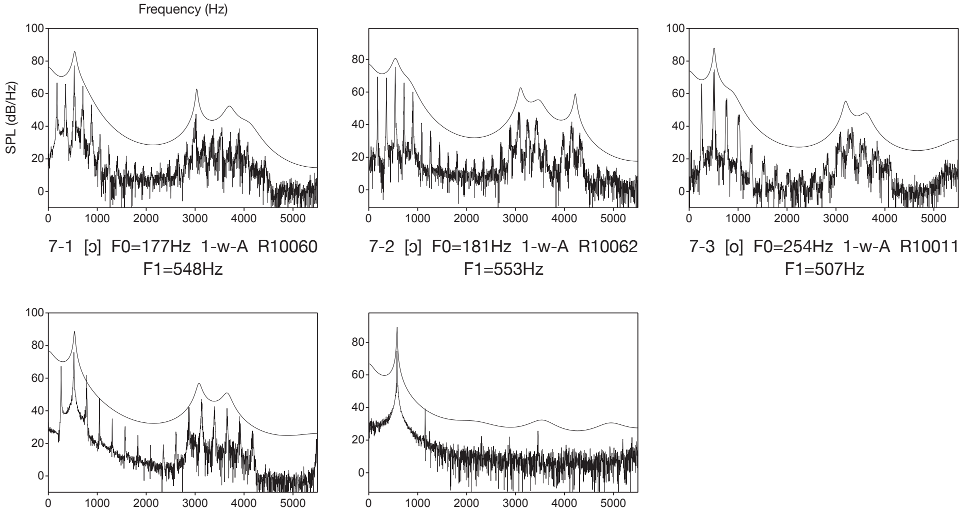
6-2 [a] F0=227Hz 1-w-A R4738
F1–F2=328–1030Hz

6-3 [u] F0=332Hz 1-w-A R4759
F1–F2=333–1031Hz

6-4 [u] F0=446Hz 1-w-A R4658
F1–F2=451–898Hz

6-5 [u] F0=508Hz 1-w-A R10054
F1–F2=507–954Hz

Figure 7. Sounds of /ɔ, o, u/ produced by a woman; related model pattern of spectral peaks and/or of calculated formant frequencies = one clear peak at c. 550 Hz.



7-1 [ɔ] F0=177Hz 1-w-A R10060
F1=548Hz

7-2 [ɔ] F0=181Hz 1-w-A R10062
F1=553Hz

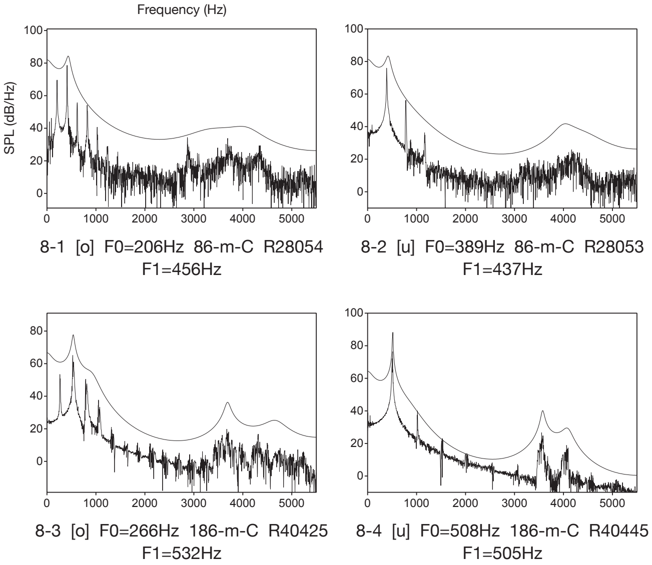
7-3 [o] F0=254Hz 1-w-A R10011
F1=507Hz

7-4 [o] F0=261Hz 1-w-A R10032
F1=519Hz

7-5 [u] F0=576Hz 1-w-A R10057
F1=569Hz

Figure 8. Two comparisons of sounds of /o, u/ produced by two children (age 12 and 6); related model patterns of spectral peaks and/or of calculated formant frequencies = one clear peak at c. 400 Hz (first sound pair) and at c. 520 Hz (second sound pair), respectively.

.



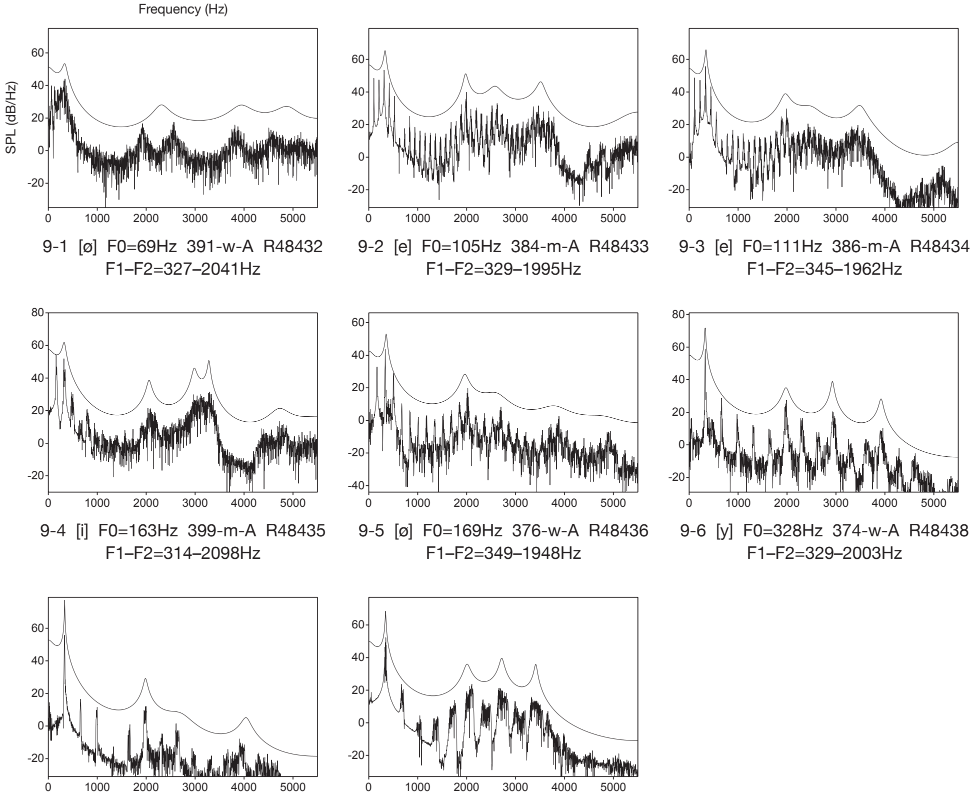
8-1 [o] F0=206Hz 86-m-C R28054
F1=456Hz

8-2 [u] F0=389Hz 86-m-C R28053
F1=437Hz

8-3 [o] F0=266Hz 186-m-C R40425
F1=532Hz

8-4 [u] F0=508Hz 186-m-C R40445
F1=505Hz

Figure 9. Sounds of /ø, e, y, i/ produced by different speakers; model pattern of spectral peaks and/or of calculated formant frequencies = 330–2000 Hz.



9-1 [ø] F0=69Hz 391-w-A R48432
F1–F2=327–2041Hz

9-2 [e] F0=105Hz 384-m-A R48433
F1–F2=329–1995Hz

9-3 [e] F0=111Hz 386-m-A R48434
F1–F2=345–1962Hz

9-4 [i] F0=163Hz 399-m-A R48435
F1–F2=314–2098Hz

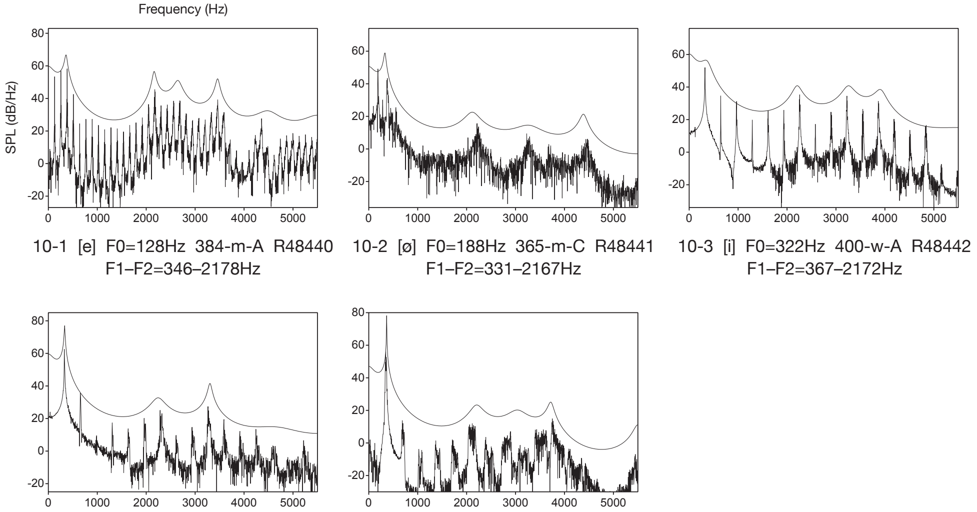
9-5 [ø] F0=169Hz 376-w-A R48436
F1–F2=349–1948Hz

9-6 [y] F0=328Hz 374-w-A R48438
F1–F2=329–2003Hz

9-7 [y] F0=329Hz 376-w-A R48437
F1–F2=329–2001Hz

9-8 [i] F0=345Hz 391-w-A R48439
F1–F2=344–2066Hz

Figure 10. Sounds of /ø, e, y, i/ produced by different speakers; model pattern of spectral peaks and/or of calculated formant frequencies = 350–2150 Hz.



10-1 [e] F0=128Hz 384-m-A R48440
F1–F2=346–2178Hz

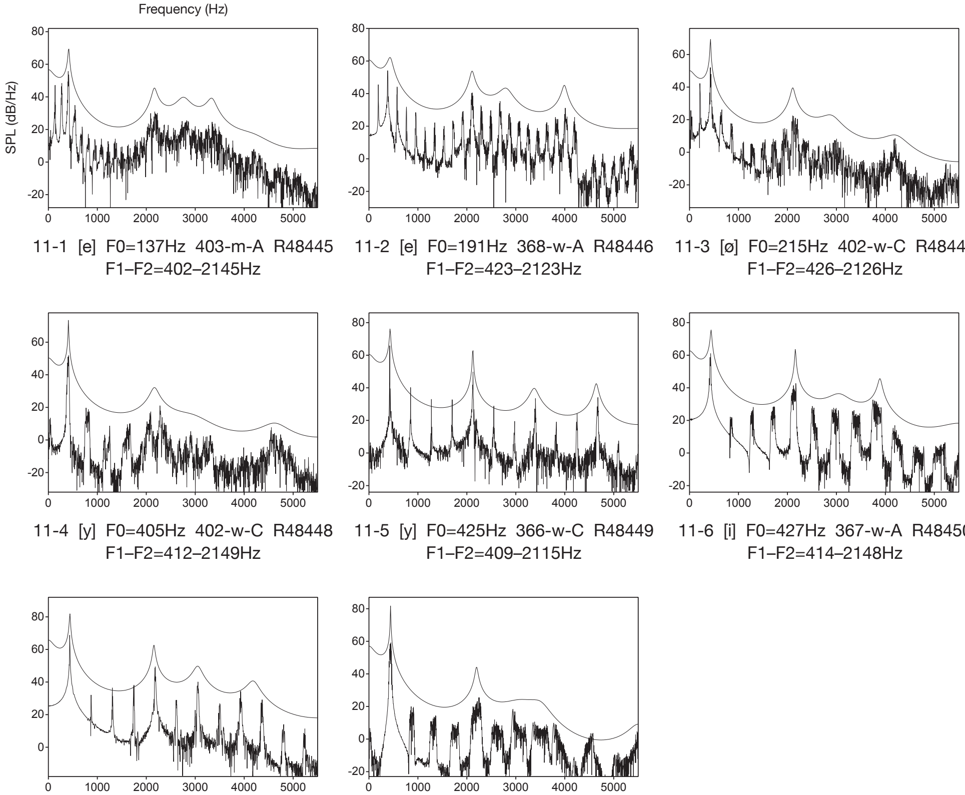
10-2 [ø] F0=188Hz 365-m-C R48441
F1–F2=331–2167Hz

10-3 [i] F0=322Hz 400-w-A R48442
F1–F2=367–2172Hz

10-4 [y] F0=328Hz 375-m-C R48443
F1–F2=345–2184Hz

10-5 [i] F0=351Hz 367-w-A R48444
F1–F2=353–2180Hz

Figure 11. Sounds of /ø, e, y, i/ produced by different speakers; model pattern of spectral peaks and/or of calculated formant frequencies = 420–2150 Hz.



11-1 [e] F0=137Hz 403-m-A R48445
F1–F2=402–2145Hz

11-2 [e] F0=191Hz 368-w-A R48446
F1–F2=423–2123Hz

11-3 [ø] F0=215Hz 402-w-C R48447
F1–F2=426–2126Hz

11-4 [y] F0=405Hz 402-w-C R48448
F1–F2=412–2149Hz

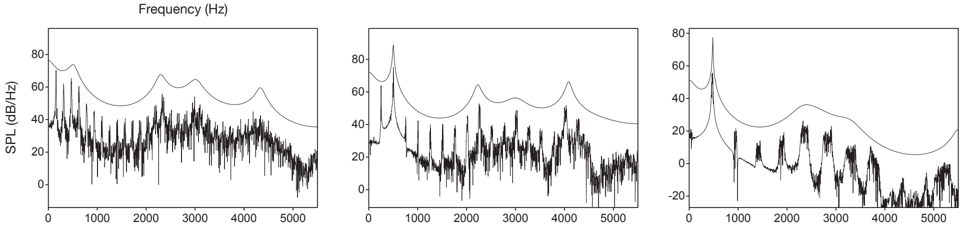
11-5 [y] F0=425Hz 366-w-C R48449
F1–F2=409–2115Hz

11-6 [i] F0=427Hz 367-w-A R48450
F1–F2=414–2148Hz

11-7 [i] F0=436Hz 401-w-A R48452
F1–F2=435–2149Hz

11-8 [i] F0=439Hz 367-w-A R48451
F1–F2=440–2203Hz

Figure 12. Sounds of /ε, e, i/ produced by different speakers; related model pattern of spectral peaks and/or of calculated formant frequencies = 500–2250 Hz.

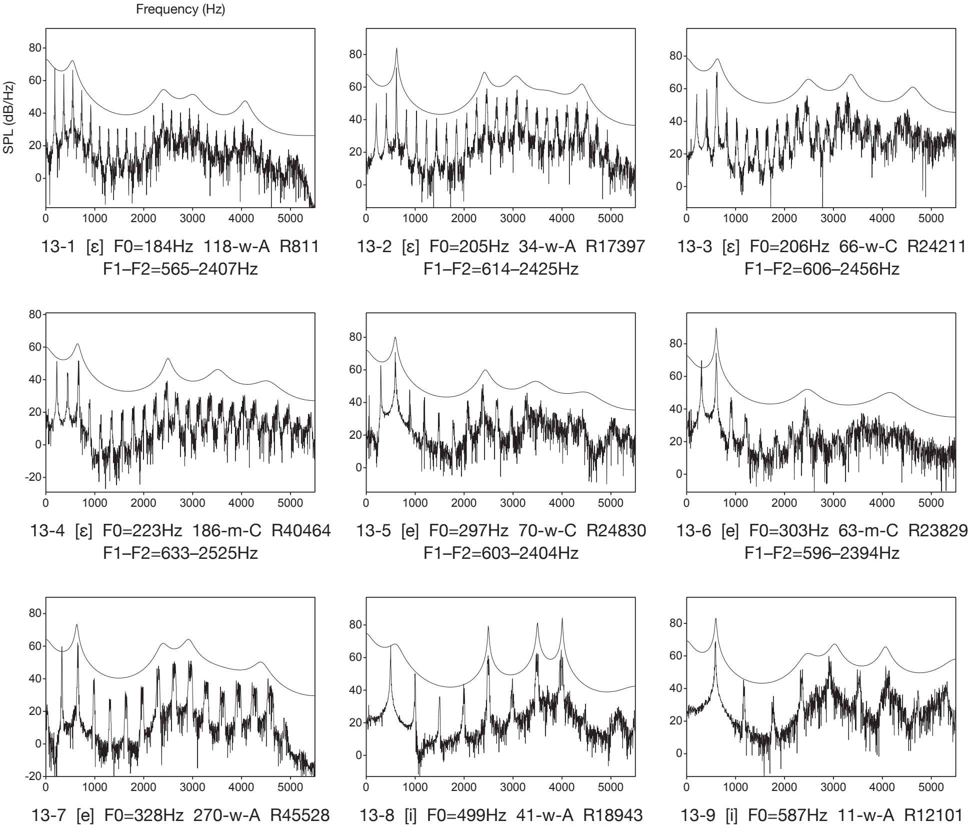


12-1 [ε] F0=156Hz 21-w-A R22616
F1–F2=494–2301Hz

12-2 [e] F0=252Hz 6-w-A R13359
F1–F2=495–2230Hz

12-3 [i] F0=475Hz 367-w-A R48048
F1–F2=472–2273Hz

Figure 13. Sounds of /ɛ, e, i/ produced by different speakers; related model pattern of spectral peaks and/or of calculated formant frequencies = 600–2450 Hz.



13-1 [ɛ] F0=184Hz 118-w-A R811
F1–F2=565–2407Hz

13-2 [ɛ] F0=205Hz 34-w-A R17397
F1–F2=614–2425Hz

13-3 [ɛ] F0=206Hz 66-w-C R24211
F1–F2=606–2456Hz

13-4 [ɛ] F0=223Hz 186-m-C R40464
F1–F2=633–2525Hz

13-5 [e] F0=297Hz 70-w-C R24830
F1–F2=603–2404Hz

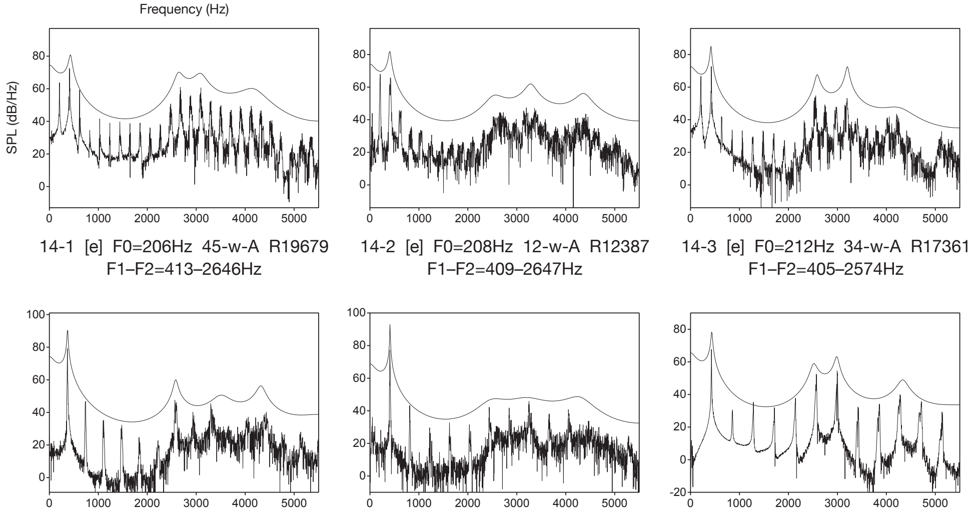
13-6 [e] F0=303Hz 63-m-C R23829
F1–F2=596–2394Hz

13-7 [e] F0=328Hz 270-w-A R45528
F1–F2=611–2445Hz

13-8 [i] F0=499Hz 41-w-A R18943
F1–F2=615–2488Hz

13-9 [i] F0=587Hz 11-w-A R12101
F1–F2=576–2440Hz

Figure 14. Sounds of /e, i/ produced by different speakers; related model pattern of spectral peaks and/or of calculated formant frequencies = 400–2600 Hz.



14-1 [e] F0=206Hz 45-w-A R19679
F1–F2=413–2646Hz

14-2 [e] F0=208Hz 12-w-A R12387
F1–F2=409–2647Hz

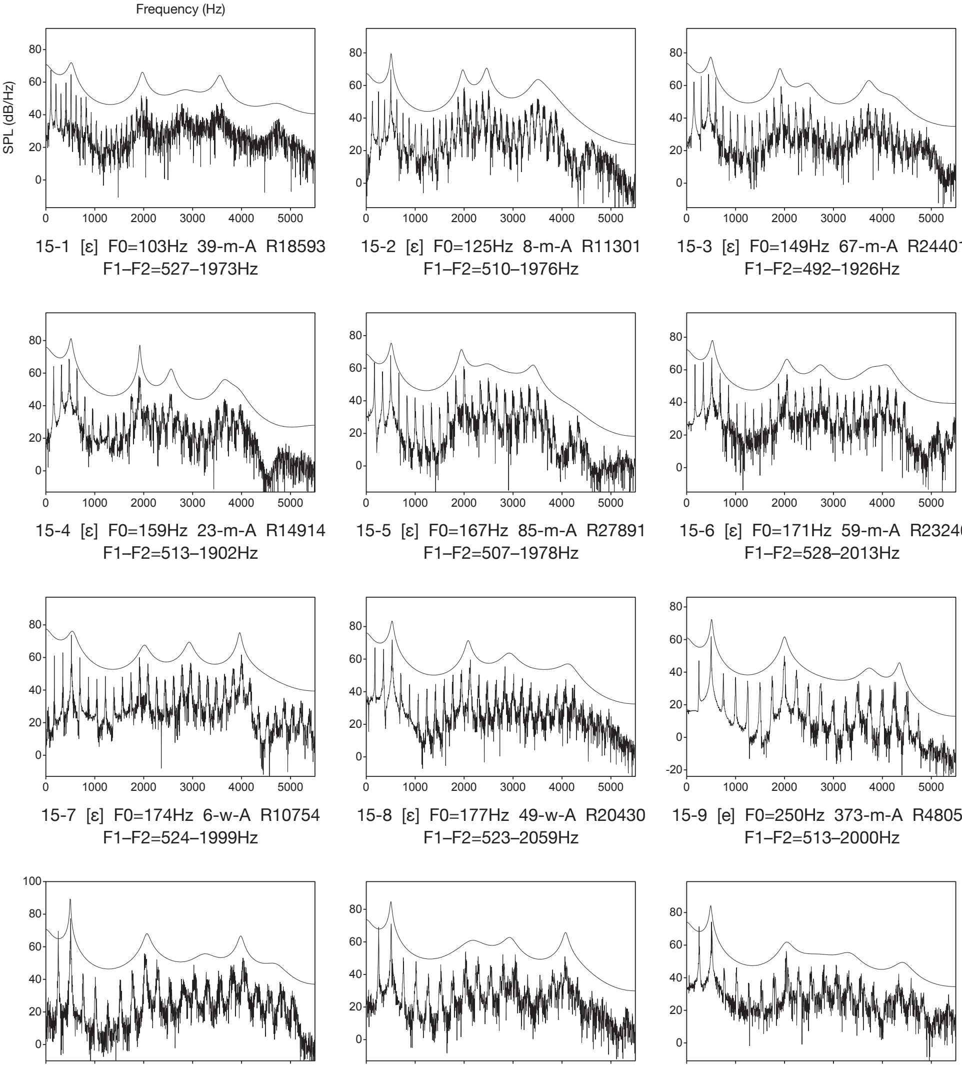
14-3 [e] F0=212Hz 34-w-A R17361
F1–F2=405–2574Hz

14-4 [i] F0=368Hz 14-w-A R12815
F1–F2=372–2593Hz

14-5 [i] F0=408Hz 19-w-A R14027
F1–F2=409–2555Hz

14-6 [i] F0=428Hz 362-w-A R48049
F1–F2=425–2575Hz

Figure 15. Sounds of /ɛ, e, y/ produced by different speakers; related model pattern of spectral peaks and/or of calculated formant frequencies = 500–2000 Hz.



15-1 [ɛ] F0=103Hz 39-m-A R18593
F1–F2=527–1973Hz

15-2 [ɛ] F0=125Hz 8-m-A R11301
F1–F2=510–1976Hz

15-3 [ɛ] F0=149Hz 67-m-A R24401
F1–F2=492–1926Hz

15-4 [ɛ] F0=159Hz 23-m-A R14914
F1–F2=513–1902Hz

15-5 [ɛ] F0=167Hz 85-m-A R27891
F1–F2=507–1978Hz

15-6 [ɛ] F0=171Hz 59-m-A R23240
F1–F2=528–2013Hz

15-7 [ɛ] F0=174Hz 6-w-A R10754
F1–F2=524–1999Hz

15-8 [ɛ] F0=177Hz 49-w-A R20430
F1–F2=523–2059Hz

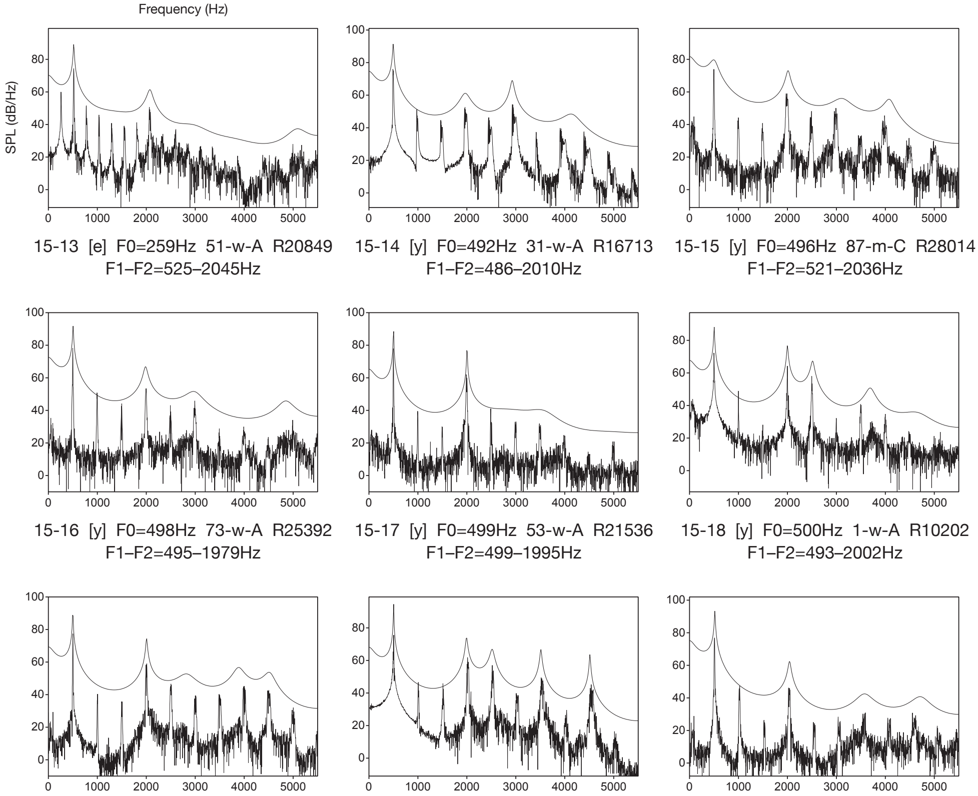
15-9 [e] F0=250Hz 373-m-A R48050
F1–F2=513–2000Hz

15-10 [e] F0=253Hz 36-w-A R17809
F1–F2=488–2089Hz

15-11 [e] F0=254Hz 20-w-A R14229
F1–F2=518–2074Hz

15-12 [e] F0=255Hz 96-m-A R29842
F1–F2=515–2058Hz

(Figure 15, continuation)



15-13 [e] F0=259Hz 51-w-A R20849
F1–F2=525–2045Hz

15-14 [y] F0=492Hz 31-w-A R16713
F1–F2=486–2010Hz

15-15 [y] F0=496Hz 87-m-C R28014
F1–F2=521–2036Hz

15-16 [y] F0=498Hz 73-w-A R25392
F1–F2=495–1979Hz

15-17 [y] F0=499Hz 53-w-A R21536
F1–F2=499–1995Hz

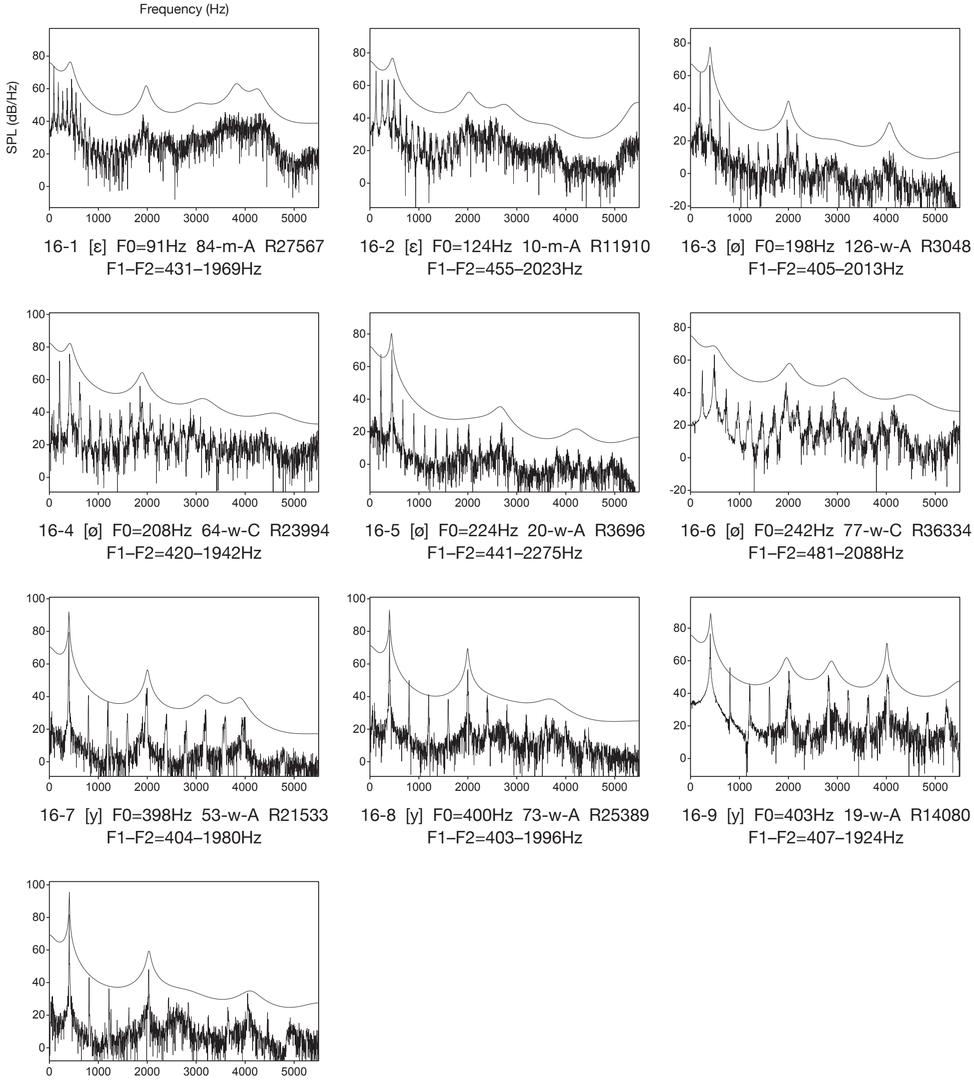
15-18 [y] F0=500Hz 1-w-A R10202
F1–F2=493–2002Hz

15-19 [y] F0=500Hz 24-w-A R15166
F1–F2=506–2020Hz

15-20 [y] F0=505Hz 35-w-A R17665
F1–F2=506–2005Hz

15-21 [y] F0=509Hz 63-m-C R23882
F1–F2=513–1984Hz

Figure 16. Sounds of /ɛ, ø, y/ produced by different speakers; related model pattern of spectral peaks and/or of calculated formant frequencies = 430–2000 Hz.



16-1 [ɛ] F0=91Hz 84-m-A R27567
F1–F2=431–1969Hz

16-2 [ɛ] F0=124Hz 10-m-A R11910
F1–F2=455–2023Hz

16-3 [ø] F0=198Hz 126-w-A R3048
F1–F2=405–2013Hz

16-4 [ø] F0=208Hz 64-w-C R23994
F1–F2=420–1942Hz

16-5 [ø] F0=224Hz 20-w-A R3696
F1–F2=441–2275Hz

16-6 [ø] F0=242Hz 77-w-C R36334
F1–F2=481–2088Hz

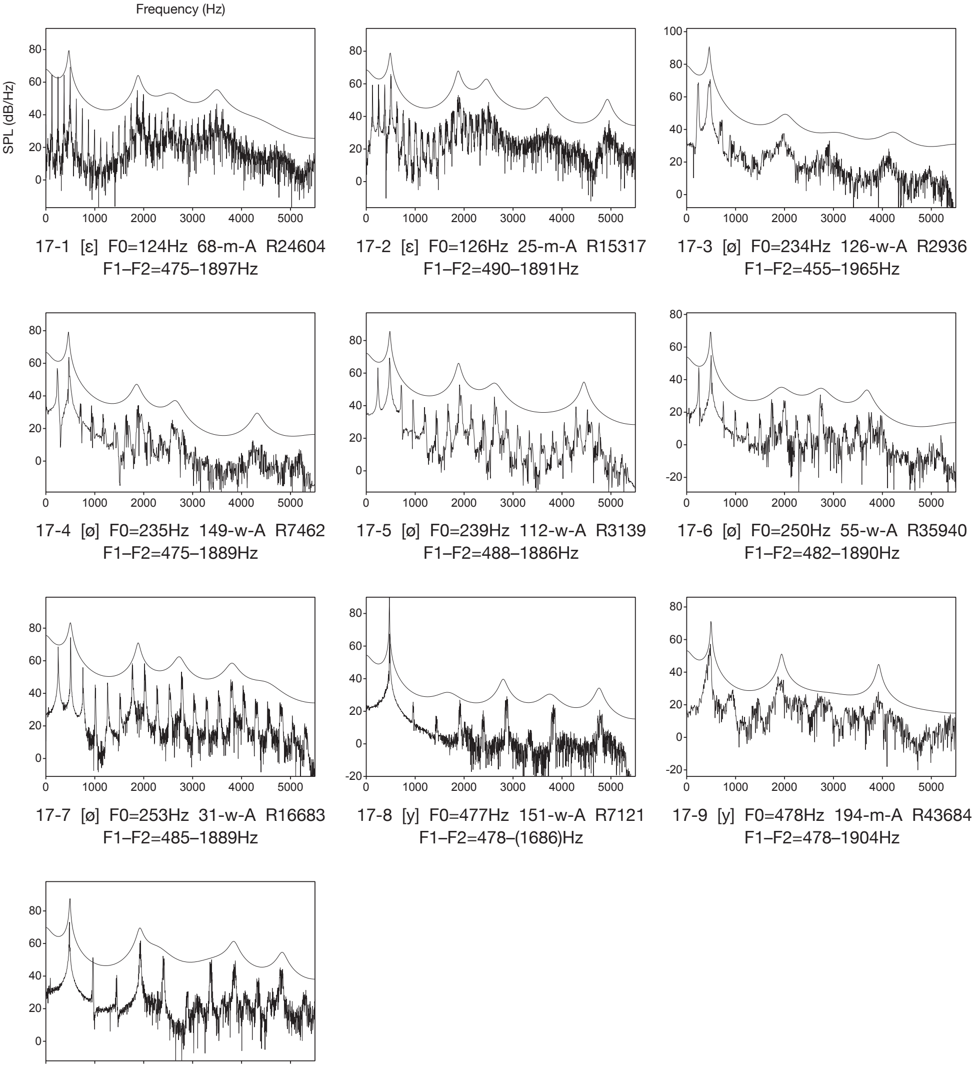
16-7 [y] F0=398Hz 53-w-A R21533
F1–F2=404–1980Hz

16-8 [y] F0=400Hz 73-w-A R25389
F1–F2=403–1996Hz

16-9 [y] F0=403Hz 19-w-A R14080
F1–F2=407–1924Hz

16-10 [y] F0=405Hz 16-w-A R13322
F1–F2=403–2031Hz

Figure 17. Sounds of /ɛ, ø, y/ produced by different speakers; related model pattern of spectral peaks and/or of calculated formant frequencies = 475–1900 Hz.



17-1 [ɛ] F0=124Hz 68-m-A R24604
F1–F2=475–1897Hz

17-2 [ɛ] F0=126Hz 25-m-A R15317
F1–F2=490–1891Hz

17-3 [ø] F0=234Hz 126-w-A R2936
F1–F2=455–1965Hz

17-4 [ø] F0=235Hz 149-w-A R7462
F1–F2=475–1889Hz

17-5 [ø] F0=239Hz 112-w-A R3139
F1–F2=488–1886Hz

17-6 [ø] F0=250Hz 55-w-A R35940
F1–F2=482–1890Hz

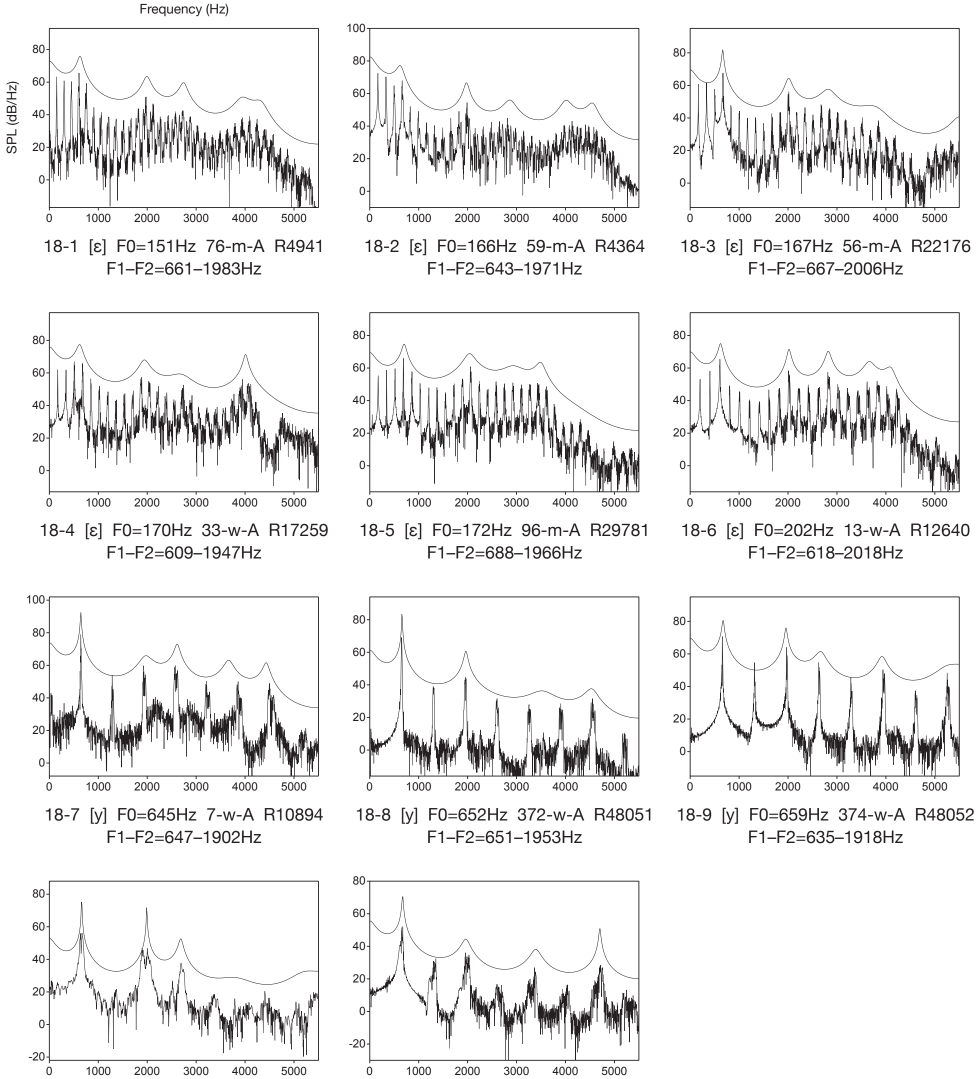
17-7 [ø] F0=253Hz 31-w-A R16683
F1–F2=485–1889Hz

17-8 [y] F0=477Hz 151-w-A R7121
F1–F2=478–(1686)Hz

17-9 [y] F0=478Hz 194-m-A R43684
F1–F2=478–1904Hz

17-10 [y] F0=481Hz 58-m-A R26586
F1–F2=481–1906Hz

Figure 18. Sounds of /ɛ, y/ produced by different speakers; related model pattern of spectral peaks and/or of calculated formant frequencies = 650–1950 Hz.



18-1 [ɛ] F0=151Hz 76-m-A R4941
F1–F2=661–1983Hz

18-2 [ɛ] F0=166Hz 59-m-A R4364
F1–F2=643–1971Hz

18-3 [ɛ] F0=167Hz 56-m-A R22176
F1–F2=667–2006Hz

18-4 [ɛ] F0=170Hz 33-w-A R17259
F1–F2=609–1947Hz

18-5 [ɛ] F0=172Hz 96-m-A R29781
F1–F2=688–1966Hz

18-6 [ɛ] F0=202Hz 13-w-A R12640
F1–F2=618–2018Hz

18-7 [y] F0=645Hz 7-w-A R10894
F1–F2=647–1902Hz

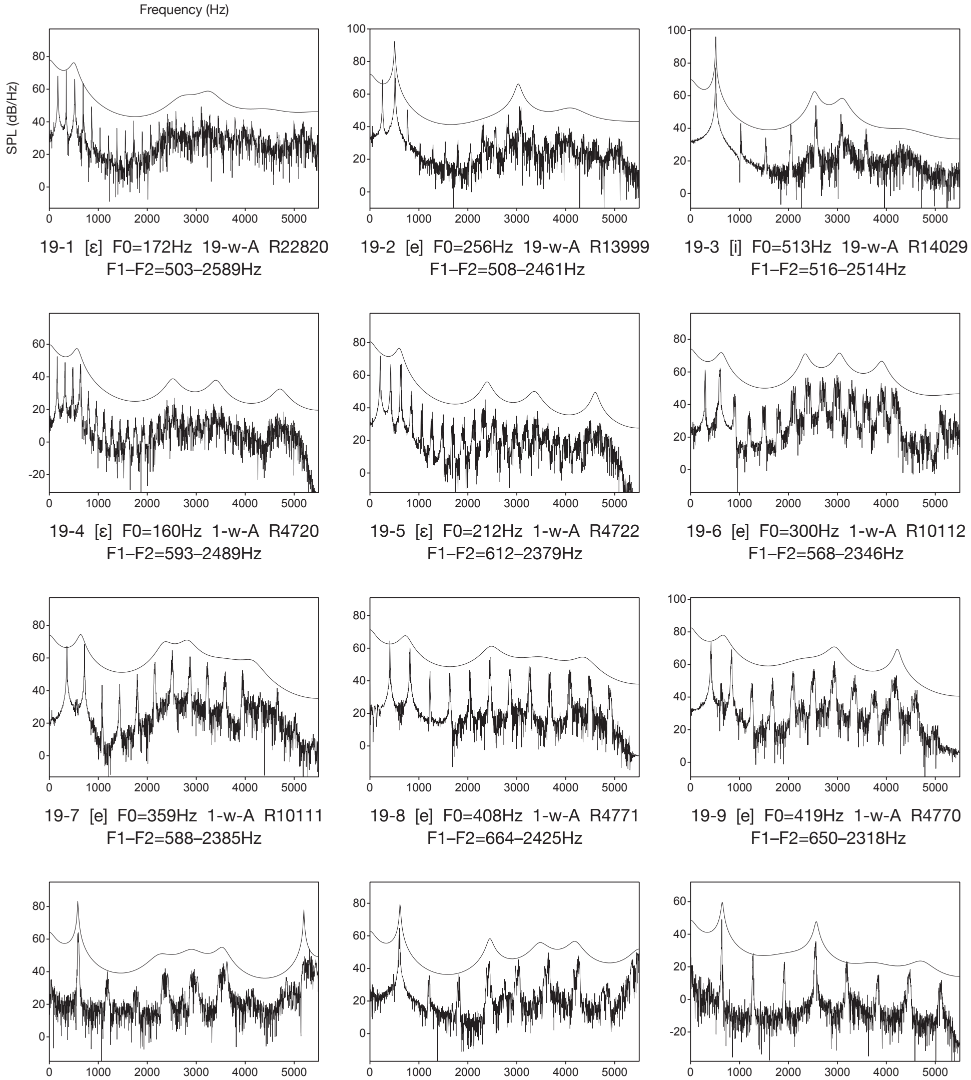
18-8 [y] F0=652Hz 372-w-A R48051
F1–F2=651–1953Hz

18-9 [y] F0=659Hz 374-w-A R48052
F1–F2=635–1918Hz

18-10 [y] F0=660Hz 177-w-A R45252
F1–F2=661–1978Hz

18-11 [y] F0=661Hz 375-m-C R48053
F1–F2=668–1960Hz

Figure 19. Two comparisons of sounds of /ɛ, e, i/ produced by two women; related model patterns of spectral peaks and/or of calculated formant frequencies = 510–2550 Hz and 600–2400 Hz, respectively.



19-1 [ɛ] F0=172Hz 19-w-A R22820
F1–F2=503–2589Hz

19-2 [e] F0=256Hz 19-w-A R13999
F1–F2=508–2461Hz

19-3 [i] F0=513Hz 19-w-A R14029
F1–F2=516–2514Hz

19-4 [ɛ] F0=160Hz 1-w-A R4720
F1–F2=593–2489Hz

19-5 [ɛ] F0=212Hz 1-w-A R4722
F1–F2=612–2379Hz

19-6 [e] F0=300Hz 1-w-A R10112
F1–F2=568–2346Hz

19-7 [e] F0=359Hz 1-w-A R10111
F1–F2=588–2385Hz

19-8 [e] F0=408Hz 1-w-A R4771
F1–F2=664–2425Hz

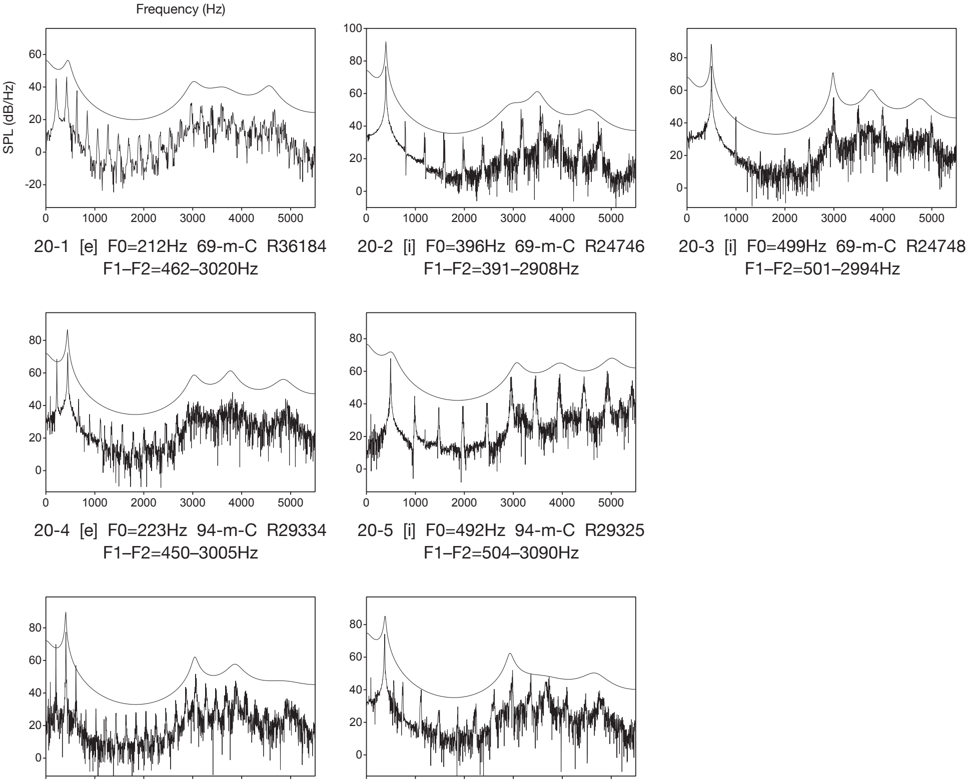
19-9 [e] F0=419Hz 1-w-A R4770
F1–F2=650–2318Hz

19-10 [i] F0=589Hz 1-w-A R10139
F1–F2=590–2370Hz

19-11 [i] F0=605Hz 1-w-A R10138
F1–F2=614–2558Hz

19-12 [i] F0=638Hz 1-w-A R4691
F1–F2=616–2329Hz

Figure 20. Three comparisons of sounds of /e, i/ produced by three children (age range 7 to 9); related model patterns of spectral peaks and/or of calculated formant frequencies = 450–3000 Hz and 400–3000 Hz, respectively.



20-1 [e] F0=212Hz 69-m-C R36184
F1–F2=462–3020Hz

20-2 [i] F0=396Hz 69-m-C R24746
F1–F2=391–2908Hz

20-3 [i] F0=499Hz 69-m-C R24748
F1–F2=501–2994Hz

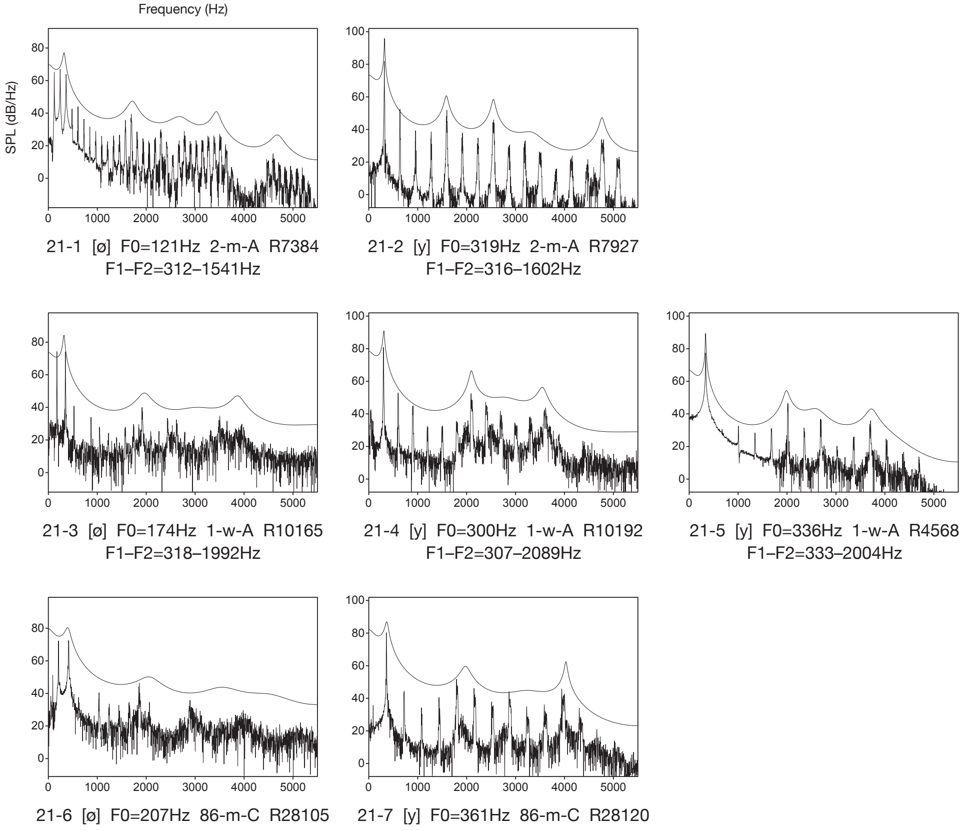
20-4 [e] F0=223Hz 94-m-C R29334
F1–F2=450–3005Hz

20-5 [i] F0=492Hz 94-m-C R29325
F1–F2=504–3090Hz

20-6 [e] F0=204Hz 93-w-C R29213
F1–F2=405–3015Hz

20-7 [i] F0=185Hz 93-w-C R29256
F1–F2=389–2925Hz

Figure 21. Three comparisons of sounds of /ø, y/ produced by a man, a woman and a child (age 12); related model patterns of spectral peaks and/or of calculated formant frequencies = 320–1600 Hz, 320–2000 Hz, and 400–2000 Hz, respectively.



21-1 [ø] F0=121Hz 2-m-A R7384
F1–F2=312–1541Hz

21-2 [y] F0=319Hz 2-m-A R7927
F1–F2=316–1602Hz

21-3 [ø] F0=174Hz 1-w-A R10165
F1–F2=318–1992Hz

21-4 [y] F0=300Hz 1-w-A R10192
F1–F2=307–2089Hz

21-5 [y] F0=336Hz 1-w-A R4568
F1–F2=333–2004Hz

21-6 [ø] F0=207Hz 86-m-C R28105
F1–F2=405–2007Hz

21-7 [y] F0=361Hz 86-m-C R28120
F1–F2=359–1970Hz

M9.2 Ambiguous Spectral Envelopes

For the frequency range relevant for the perceived vowel qualities in question, many of the sound series presented in the previous chapter do not only show similar patterns of vowel-related spectral peaks and similar patterns of calculated F1–F2 but also similar vowel-related spectral envelope shapes for sounds of different vowels, including similar patterns of calculated F1–F2–F3 for sounds of front vowels (for all calculated formant frequencies refer to the online digital version of the Materials).

M9.3 Ambiguity and Individual Vowels

The series in Section M9.1 present ambiguities as discussed here for all combinations of the long German back vowels and /a–ɑ/ and for all combinations of the long German front vowels. Thus, the ambiguities are not a phenomenon of overlapping F1–F2 spaces of neighbouring vowel qualities but, in most cases, a consequence of the dependence of vowel-specific, relative spectral energy maxima and lower formants ≤ 1.5 kHz on fundamental frequency, interrelated with an observable variation of higher vowel-related spectral parts for sounds of front vowels.

However, two restrictions apply.

Concerning the sounds of back vowels and of /a–ɑ/ investigated, the demonstration of a possible ambiguity of the lower spectral envelope and of F1–F2 is unquestionable for comparisons of sounds of /u/ and of /o/, and of /o/ and of /a–ɑ/. For the comparison of sounds of /u/ and of /a–ɑ/, however, the demonstration of a possible ambiguity is limited to similar calculated F1–F2, but because of high F0 of the sounds of /u/, this calculation is methodically unsubstantiated. Further direct comparison of the spectral envelope and the configuration of the levels of the harmonics generally provides no clear indication. Notwithstanding, it is important to consider the fact that sounds of /u/ can be produced at a level of F0 that can corresponds to F1 of sounds of /a/ and that, in such cases, exhibit a dominant first harmonic.

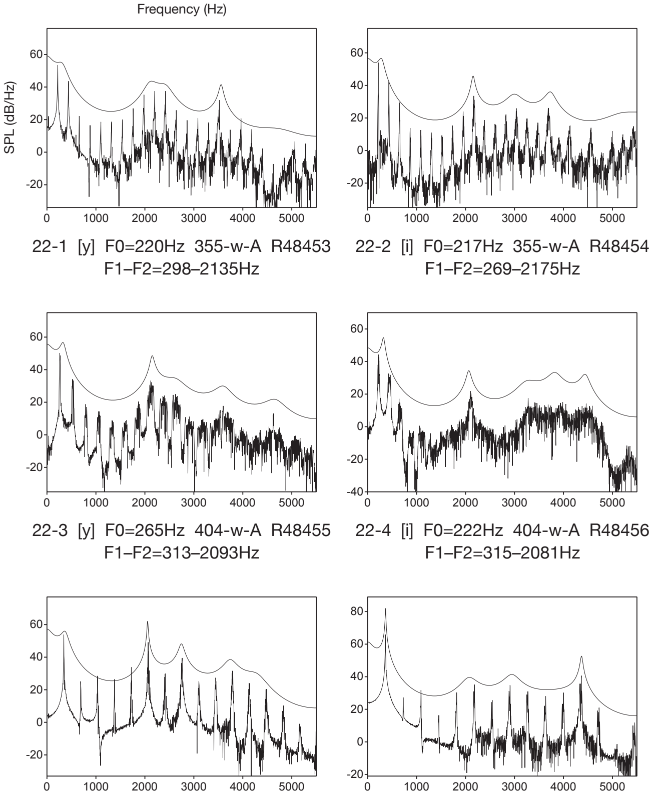
Concerning the sounds of front vowels investigated, the demonstration of a possible ambiguity, which is related to differences in F0 of the sounds compared, does not concern the direct comparisons of sounds of /e, ø/, and of /i, y/. As mentioned, in such cases, the ambiguity relates to the configuration of the levels of the harmonics, to the spectrum above F2 and to the levels of calculated formants including F3. This phenomenon is again illustrated in the following three figures.

Figure 22 Three sound pairs of /y, i/, each pair produced by single female speakers; model patterns of spectral peaks and/or of calculated formant frequencies = 290–2150 Hz, 315–2100 Hz and 350–2100 Hz, respectively

Figure 23 Sounds of /y, i/ produced by different male speakers; model pattern of spectral peaks and/or of calculated formant frequencies = 230–2050 Hz

Figure 24 A sound pair of /ø, e/ produced by a single male speaker; model pattern of spectral peaks and/or of calculated formant frequencies = 350–1700 Hz

Figure 22. Three sound pairs of /y, i/, each pair produced by single female speakers; model patterns of spectral peaks and/or of calculated formant frequencies = 290–2150 Hz, 315–2100 Hz and 350–2100 Hz, respectively.



22-1 [y] F0=220Hz 355-w-A R48453
F1–F2=298–2135Hz

22-2 [i] F0=217Hz 355-w-A R48454
F1–F2=269–2175Hz

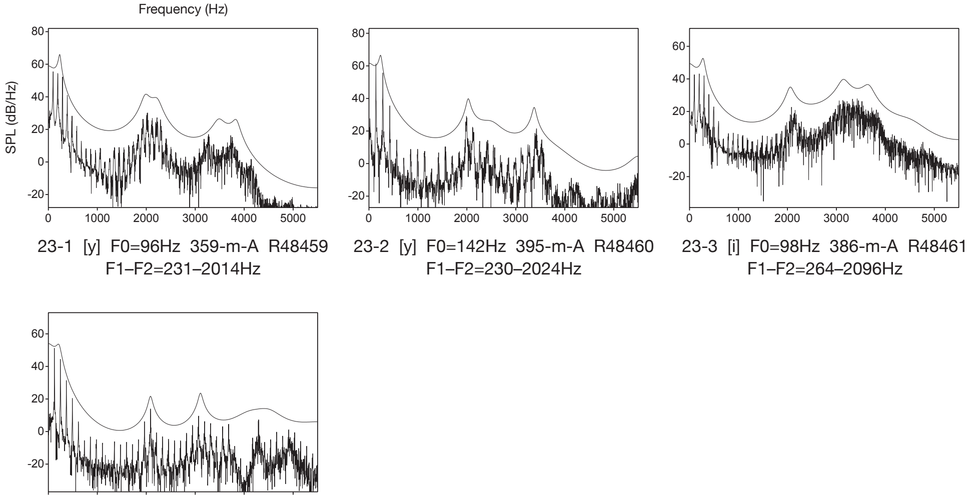
22-3 [y] F0=265Hz 404-w-A R48455
F1–F2=313–2093Hz

22-4 [i] F0=222Hz 404-w-A R48456
F1–F2=315–2081Hz

22-5 [y] F0=345Hz 401-w-A R48457
F1–F2=349–2073Hz

22-6 [i] F0=363Hz 401-w-A R48458
F1–F2=346–2116Hz

Figure 23. Sounds of /y, i/ produced by different male speakers; model pattern of spectral peaks and/or of calculated formant frequencies = 230–2050 Hz.



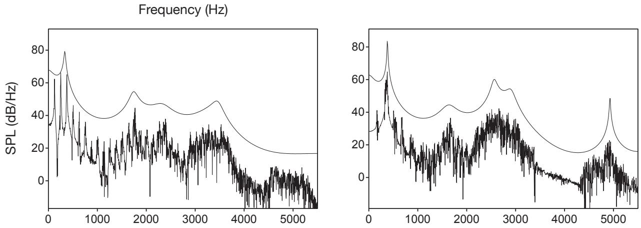
23-1 [y] F0=96Hz 359-m-A R48459
F1–F2=231–2014Hz

23-2 [y] F0=142Hz 395-m-A R48460
F1–F2=230–2024Hz

23-3 [i] F0=98Hz 386-m-A R48461
F1–F2=264–2096Hz

23-4 [i] F0=122Hz 384-m-A R48462
F1–F2=217–2100Hz

Figure 24. A sound pair of /ø, e/ produced by a single male speaker; model pattern of spectral peaks and/or of calculated formant frequencies = 350–1700 Hz.



24-1 [ø] F0=126Hz 2-m-A R7431
F1–F2=328–1741Hz

24-2 [e] F0=178Hz 2-m-A R48425
F1–F2=368–1692Hz

M10 Lack of Correspondence between Patterns of Relative Spectral Energy Maxima or Formant Patterns and Age- and Gender-Related Speaker Groups or Vocal-Tract Sizes

M10.1 Similar Patterns of Relative Spectral Maxima and Similar Formant Patterns ≤ 1.5 kHz for Different Age- and Gender-Related Speaker Groups or Vocal-Tract Sizes

Figure 1 shows sounds of the vowel /o/ produced by a child (age 8), a woman and a man. Each speaker produced sounds at different F0 in a way that allowed for a comparison of the sounds of the three speakers (representing the three main speaker groups according to age and gender) at different and similar F0. The comparison shows that age- and gender-related differences ≤ 1.5 kHz as given in formant statistics for citation-form words can decrease or even disappear if F0 of the vocalisations correspond for children, women and men. In this regard, comparisons of vocalisations of /o/ are of special interest (and shown first) because an F0-dependence of the lower spectral frequency range can be observed for F0 clearly below statistical F1, and because the frequency range ≤ 1.5 kHz covers the entire range related to the vowel identity in question. — Data for speakers, ranges of F0 and calculated F1 and F2:

Spectra 1-1 to 1-6	Child; F0 = 196–322 Hz, F1 = 424–624 Hz, F2 = 777–1092 Hz
Spectra 1-7 to 1-13	Woman; F0 = 162–320 Hz, F1 = 363–576 Hz, F2 = 804–1141 Hz
Spectra 1-14 to 21	Man; F0 = 129–326 Hz, F1 = 343–577 Hz, F2 = 672–1143 Hz

Figure 2 demonstrates this phenomenon for sounds of the vowel /e/ produced by a child (age 10), a woman and a man, concerning the lowest spectral peak and F1.—Data for speakers, ranges of F0 and calculated F1:

Spectra 2-1 to 2-6	Child; F0 = 180–330 Hz, F1 = 395–563 Hz
Spectra 2-7 to 2-13	Woman; F0 = 160–325 Hz, F1 = 389–622 Hz
Spectra 2-14 to 2-21	Man; F0 = 122–336 Hz, F1 = 370–566 Hz (excluding the last sound for which automatic calculation of F1 does not provide a reliable result)

Similar indications as shown for sounds of /e/ can be found for sounds of /ø/.

Figure 3 demonstrates this phenomenon for sounds of the vowel /u/ produced by a child (age 8), a woman and a man. However, only the first lower peak and calculated F1 are discussed because, for several sounds, an interpretation of F2 lacks methodological substantiation.—Data for speakers, ranges of F0 and calculated F1:

Spectra 3-1 to 3-6	Child; F0 = 237–492 Hz, F1 = 273–492 Hz
Spectra 3-7 to 3-13	Woman; F0 = 177–498 Hz, F1 = 300–502 Hz
Spectra 3-14 to 3-21	Man; F0 = 138–519 Hz, F1 = 303–519 Hz

Figure 4 demonstrates this phenomenon for sounds of the vowel /i/ produced by a child (age 8), a woman and a man, concerning the lower spectral peak and calculated F1.—Data for speakers, ranges of F0 and calculated F1:

Spectra 4-1 to 4-6	Child; F0 = 247–533 Hz, F1 = 267–534 Hz
Spectra 4-7 to 4-13	Woman; F0 = 177–518 Hz, F1 = 279–525 Hz
Spectra 4-14 to 4-21	Man; F0 = 134–534 Hz, F1 = 216–550 Hz

Similar indications as shown for sounds of /i/ can be found for sounds of /y/.

With regard to sounds of /a–ɑ/, a compilation of corresponding sound series similar to those presented for the other vowels often encounters some difficulties for two main reasons: spectral peaks and formant patterns often do not shift markedly with rising F0, and children often produce a very open /a/, while many adults produce an intermediate sound of /a–ɑ/ or even a sound of /ɑ/, although all speakers speak the same language and live in a geographically limited area. However, Figure 5 demonstrates a case of comparable vowel spectra and comparable formant patterns for sounds of /a/ produced by a child (age 10), a woman and a man.—Data for speakers, ranges of F0 and calculated F1:

Spectra 5-1 to 5-6	Child; F0 = 196–329 Hz, F1 = 759–1055 Hz, F2 = 1341–1555 Hz
Spectra 5-7 to 5-13	Woman; F0 = 160–329 Hz, F1 = 706–1007 Hz, F2 = 1265–1503 Hz
Spectra 5-14 to 5-21	Man; F0 = 126–324 Hz, F1 = 758–898 Hz, F2 = 1232–1431 Hz

The sounds presented in the previous figures may lead to the question whether, with rising F0 and related shifts of the lower spectral peaks and of the calculated lower formants, the perception of age and gender of the speaker alters, i.e. whether the sounds of adults are perceived as produced by children at F0 > c. 260 Hz, and whether sounds of men are perceived as produced by women > c. 200 Hz. This may indeed be the case for the comparison of the sounds of some speakers, while it does not hold true for others. To demonstrate the latter, Figure 6 shows similar vowel spectra and similar formant patterns for sounds of the vowel /o/ produced by a child (age 10), a woman (untrained speaker) and a man (classical opera singer, baritone). For these sounds, the perceived vowel quality corresponds very well. However, the baritone is always perceived as such at all F0 of his singing, which is represented in his vowel spectra by a so-called "singer's formant cluster". (Again, only the first lower peak and calculated F1 are discussed since most sounds exhibit only one spectral peak; for these sounds, the calculated F2 is weak and its role for vowel perception is questionable; see Section M7.1.) — Data for speakers, ranges of F0 and calculated F1:

Spectra 6-1 to 6-5	Child; F0 = 181–348 Hz, F1 = 377–674 Hz
Spectra 6-6 to 6-11	Woman; F0 = 168–332 Hz, F1 = 344–593 Hz
Spectra 6-12 to 6-17	Man; F0 = 127–325 Hz, F1 = 386–680 Hz

As a direct consequence of the documented observations, it follows that, for back vowels, the sounds of men (at higher F0) may exhibit higher vowel-related spectral peaks and higher calculated F1 or F1–F2 patterns than the sounds of women (at lower F0). The same holds true for the lowest spectral peak and calculated F1 of front vowels and may also occur when comparing sounds of adults and children.

Figure 7 shows such an “inversion” of expected age- and gender-related differences comparing sounds of the vowel /o/ produced by a child and a man, selected from the sound series of the previous Figure 6. If the F0 of the sounds of the man substantially exceeds the F0 of a sound of the child, the first spectral peak and calculated F1 of the sounds of the man are also above the corresponding peak and F1 of the sound of the child (compare Spectra 7-1 to 7-3). The same holds true for calculated F2, but as mentioned, the measurement and perceptual role of F2 are in question. However, if the comparison relates to the sounds of the man at F0 corresponding to statistical values (given for citation-form words), the first spectral peak and calculated F1 (and F2) are found as lower for the man than for the child, as this is generally expected (see Spectra 7-4 and 7-5).—Data for speakers, ranges of F0 and calculated F1 (and F2), in the order of F0:

	“Inverted” age- or size-related difference
Spectra 7-1	Child; F0 = 223 Hz, F1 = 440 Hz (F2 = 764 Hz)
Spectra 7-2, 7-3	Man; F0 = 261–325 Hz, F1 = 511–680 Hz (F2 = 884–950 Hz)
	“Expected” age- or size-related difference
Spectra 7-4	Man; F0 = 127 Hz, F1 = 430 Hz (F2 = 535 Hz)
Spectra 7-5	Child; F0 = 264 Hz, F1 = 538 Hz (F2 = 1069 Hz)

Figure 8 demonstrates this phenomenon < 1.5 kHz by comparing selected sounds of the vowel /e/ shown in Figure 2.—Data for speakers and ranges of F0 and calculated F1:

	“Inverted” age- or size-related difference
Spectra 8-1	Child; F0 = 222 Hz, F1 = 449 Hz
Spectra 8-2, 8-3	Man; F0 = 260–293 Hz, F1 = 506–566 Hz
	“Expected” age- or size-related difference
Spectra 8-4	Man; F0 = 122 Hz, F1 = 370 Hz
Spectra 8-5	Child; F0 = 265 Hz, F1 = 518 Hz

Figure 9 demonstrates this phenomenon < 1.5 kHz by comparing selected sounds of the vowel /u/ shown in Figure 3.—Data for speakers and ranges of F0 and calculated F1:

	“Inverted” age- or size-related difference
Spectra 9-1	Child; F0 = 237 Hz, F1 = 273 Hz
Spectra 9-2, 9-3	Man; F0 = 410–519 Hz, F1 = 412–519 Hz
	“Expected” age- or size-related difference
Spectra 9-4	Man; F0 = 138 Hz, F1 = 303 Hz
Spectra 9-5	Child; F0 = 257 Hz, F1 = 346 Hz

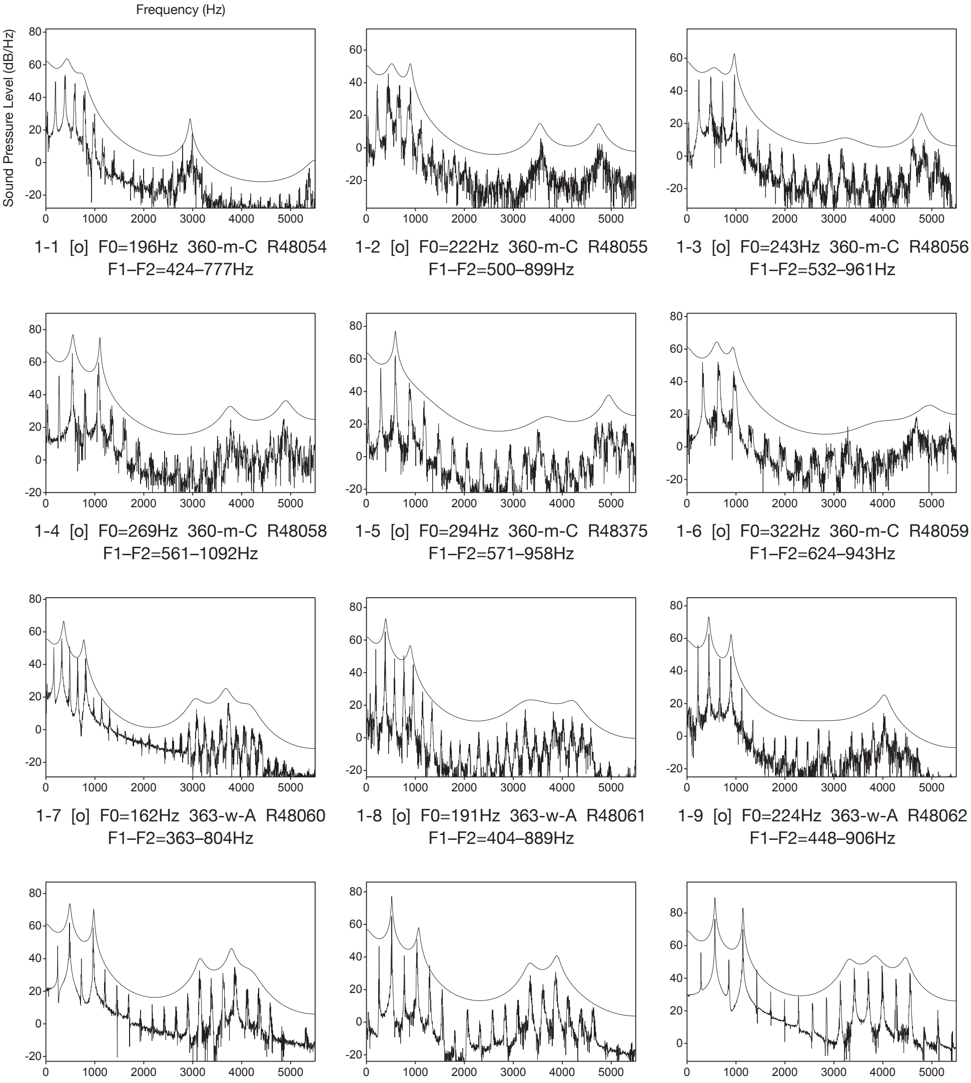
Figure 10 demonstrates this phenomenon < 1.5 kHz by comparing selected sounds of the vowel /i/ shown in Figure 4.—Data for speakers and ranges of F0 and calculated F1:

	“Inverted” age- or size-related difference
Spectra 10-1	Child; F0 = 247 Hz, F1 = 267 Hz
Spectra 10-2, 10-3	Man; F0 = 441–534 Hz, F1 = 444–550 Hz
	“Expected” age- or size-related difference
Spectra 10-4	Man; F0 = 134 Hz, F1 = 269 Hz
Spectra 10-5	Child; F0 = 263 Hz, F1 = 301 Hz

Comparisons are limited to children and men because the corresponding differences in the vocal-tract sizes are assumed to be highest.

For earlier accounts, see Maurer, Cook, Landis, and d’Heureuse (1992), Maurer, Suter, Friedrichs, and Dellwo (2015b); note also some related reflections in Potter and Steinberg (1950).

Figure 1. Sounds of /o/ produced by a child, a woman and a man at comparable levels of F0. Fig. 1-1 to 1-6 = sounds of the child; Fig. 1-7 to 1-13 = sounds of the woman; Fig. 1-14 to 1-21 = sounds of the man.



1-1 [o] F0=196Hz 360-m-C R48054
F1–F2=424–777Hz

1-2 [o] F0=222Hz 360-m-C R48055
F1–F2=500–899Hz

1-3 [o] F0=243Hz 360-m-C R48056
F1–F2=532–961Hz

1-4 [o] F0=269Hz 360-m-C R48058
F1–F2=561–1092Hz

1-5 [o] F0=294Hz 360-m-C R48375
F1–F2=571–958Hz

1-6 [o] F0=322Hz 360-m-C R48059
F1–F2=624–943Hz

1-7 [o] F0=162Hz 363-w-A R48060
F1–F2=363–804Hz

1-8 [o] F0=191Hz 363-w-A R48061
F1–F2=404–889Hz

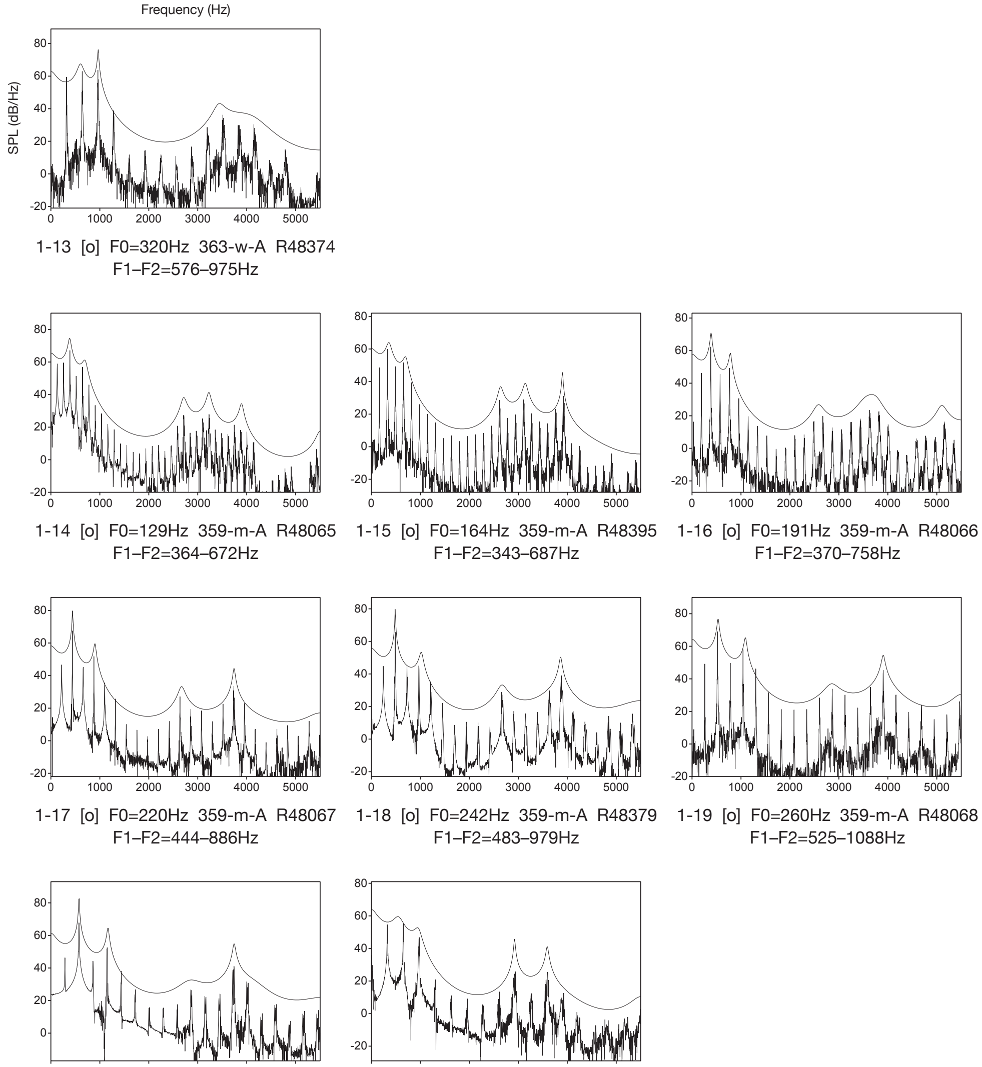
1-9 [o] F0=224Hz 363-w-A R48062
F1–F2=448–906Hz

1-10 [o] F0=242Hz 363-w-A R48376
F1–F2=478–981Hz

1-11 [o] F0=258Hz 363-w-A R48377
F1–F2=512–1061Hz

1-12 [o] F0=285Hz 363-w-A R48378
F1–F2=567–1141Hz

(Figure 1, continuation)



1-13 [o] F0=320Hz 363-w-A R48374
F1–F2=576–975Hz

1-14 [o] F0=129Hz 359-m-A R48065
F1–F2=364–672Hz

1-15 [o] F0=164Hz 359-m-A R48395
F1–F2=343–687Hz

1-16 [o] F0=191Hz 359-m-A R48066
F1–F2=370–758Hz

1-17 [o] F0=220Hz 359-m-A R48067
F1–F2=444–886Hz

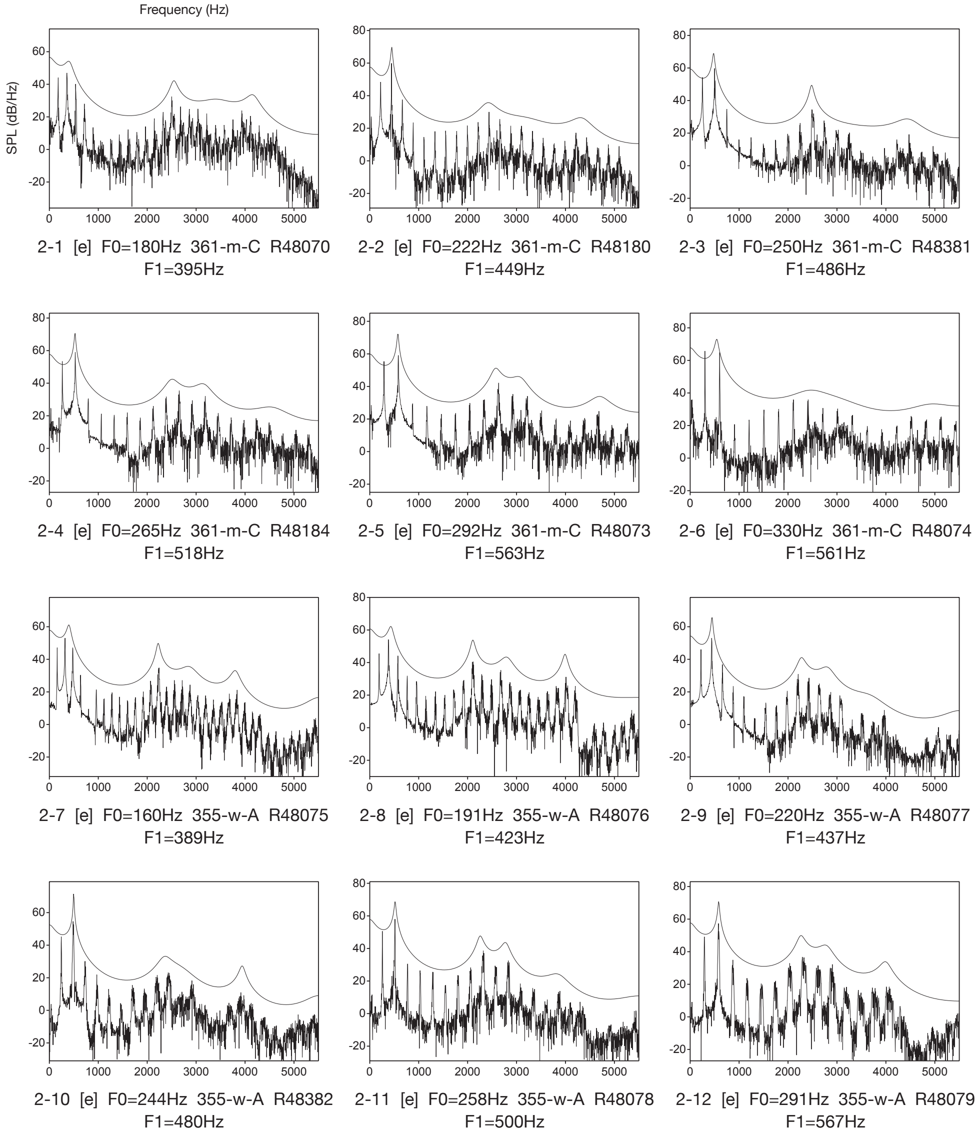
1-18 [o] F0=242Hz 359-m-A R48379
F1–F2=483–979Hz

1-19 [o] F0=260Hz 359-m-A R48068
F1–F2=525–1088Hz

1-20 [o] F0=287Hz 359-m-A R48380
F1–F2=576–1143Hz

1-21 [o] F0=326Hz 359-m-A R48069
F1–F2=577–941Hz

Figure 2. Sounds of /e/ produced by a child, a woman and a man at comparable levels of F0. Fig. 2-1 to 1-6 = sounds of the child; Fig. 2-7 to 2-13 = sounds of the woman; Fig. 2-14 to 2-21 = sounds of the man.



2-1 [e] F0=180Hz 361-m-C R48070
F1=395Hz

2-2 [e] F0=222Hz 361-m-C R48180
F1=449Hz

2-3 [e] F0=250Hz 361-m-C R48381
F1=486Hz

2-4 [e] F0=265Hz 361-m-C R48184
F1=518Hz

2-5 [e] F0=292Hz 361-m-C R48073
F1=563Hz

2-6 [e] F0=330Hz 361-m-C R48074
F1=561Hz

2-7 [e] F0=160Hz 355-w-A R48075
F1=389Hz

2-8 [e] F0=191Hz 355-w-A R48076
F1=423Hz

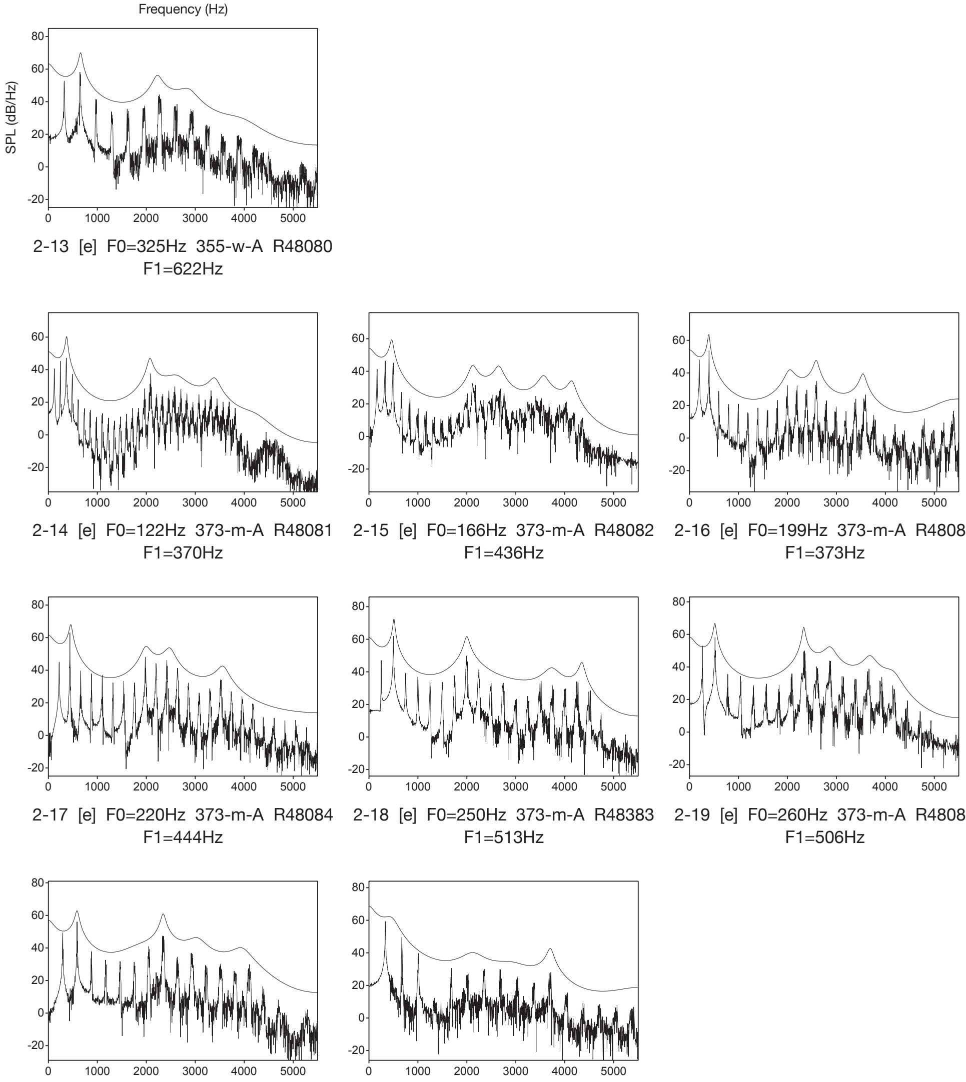
2-9 [e] F0=220Hz 355-w-A R48077
F1=437Hz

2-10 [e] F0=244Hz 355-w-A R48382
F1=480Hz

2-11 [e] F0=258Hz 355-w-A R48078
F1=500Hz

2-12 [e] F0=291Hz 355-w-A R48079
F1=567Hz

(Figure 2, continuation)



2-13 [e] F0=325Hz 355-w-A R48080
F1=622Hz

2-14 [e] F0=122Hz 373-m-A R48081
F1=370Hz

2-15 [e] F0=166Hz 373-m-A R48082
F1=436Hz

2-16 [e] F0=199Hz 373-m-A R48083
F1=373Hz

2-17 [e] F0=220Hz 373-m-A R48084
F1=444Hz

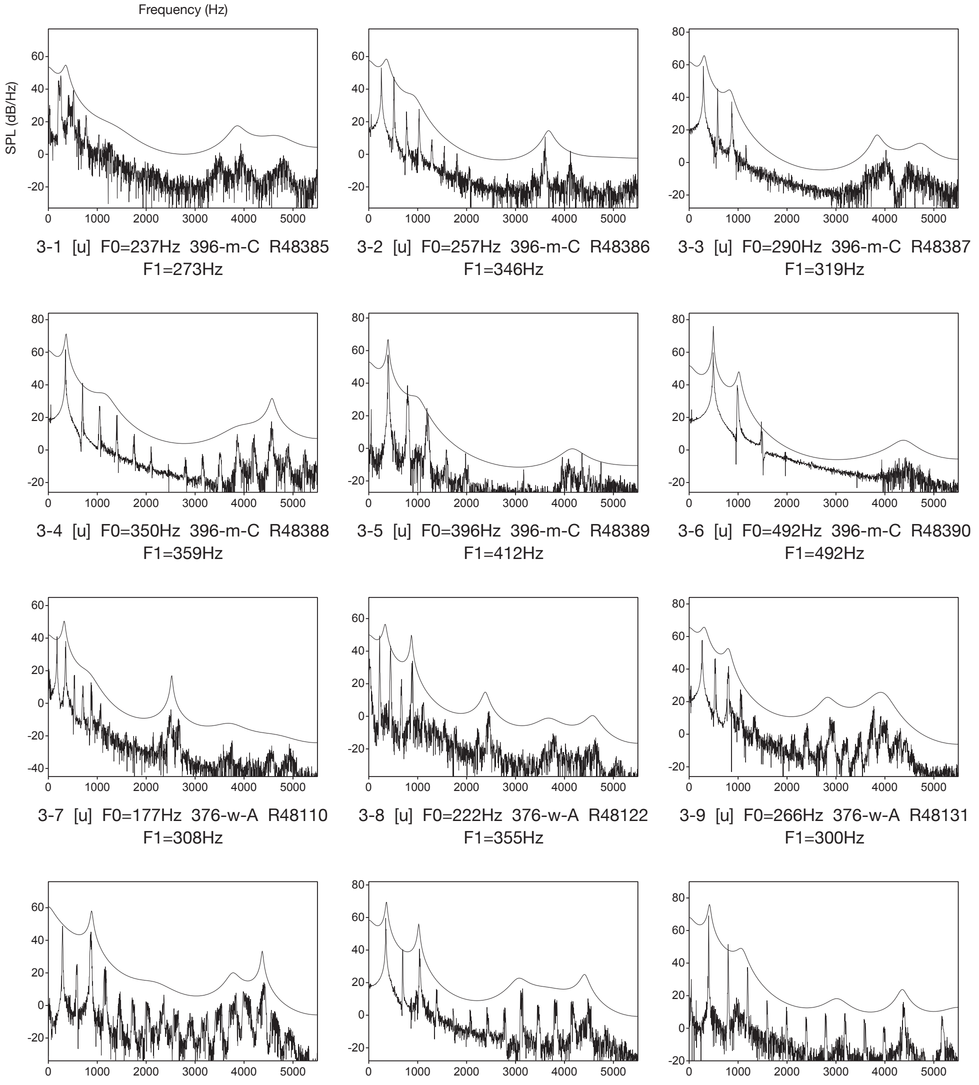
2-18 [e] F0=250Hz 373-m-A R48383
F1=513Hz

2-19 [e] F0=260Hz 373-m-A R48085
F1=506Hz

2-20 [e] F0=293Hz 373-m-A R48086
F1=566Hz

2-21 [e] F0=336Hz 373-m-A R48384
F1=(375)Hz

Figure 3. Sounds of /u/ produced by a child, a woman and a man at corresponding F0. Fig. 1-1 to 1-6 = sounds of the child; Fig. 3-7 to 3-13 = sounds of the woman; Fig. 3-14 to 3-21 = sounds of the man.



3-1 [u] F0=237Hz 396-m-C R48385
F1=273Hz

3-2 [u] F0=257Hz 396-m-C R48386
F1=346Hz

3-3 [u] F0=290Hz 396-m-C R48387
F1=319Hz

3-4 [u] F0=350Hz 396-m-C R48388
F1=359Hz

3-5 [u] F0=396Hz 396-m-C R48389
F1=412Hz

3-6 [u] F0=492Hz 396-m-C R48390
F1=492Hz

3-7 [u] F0=177Hz 376-w-A R48110
F1=308Hz

3-8 [u] F0=222Hz 376-w-A R48122
F1=355Hz

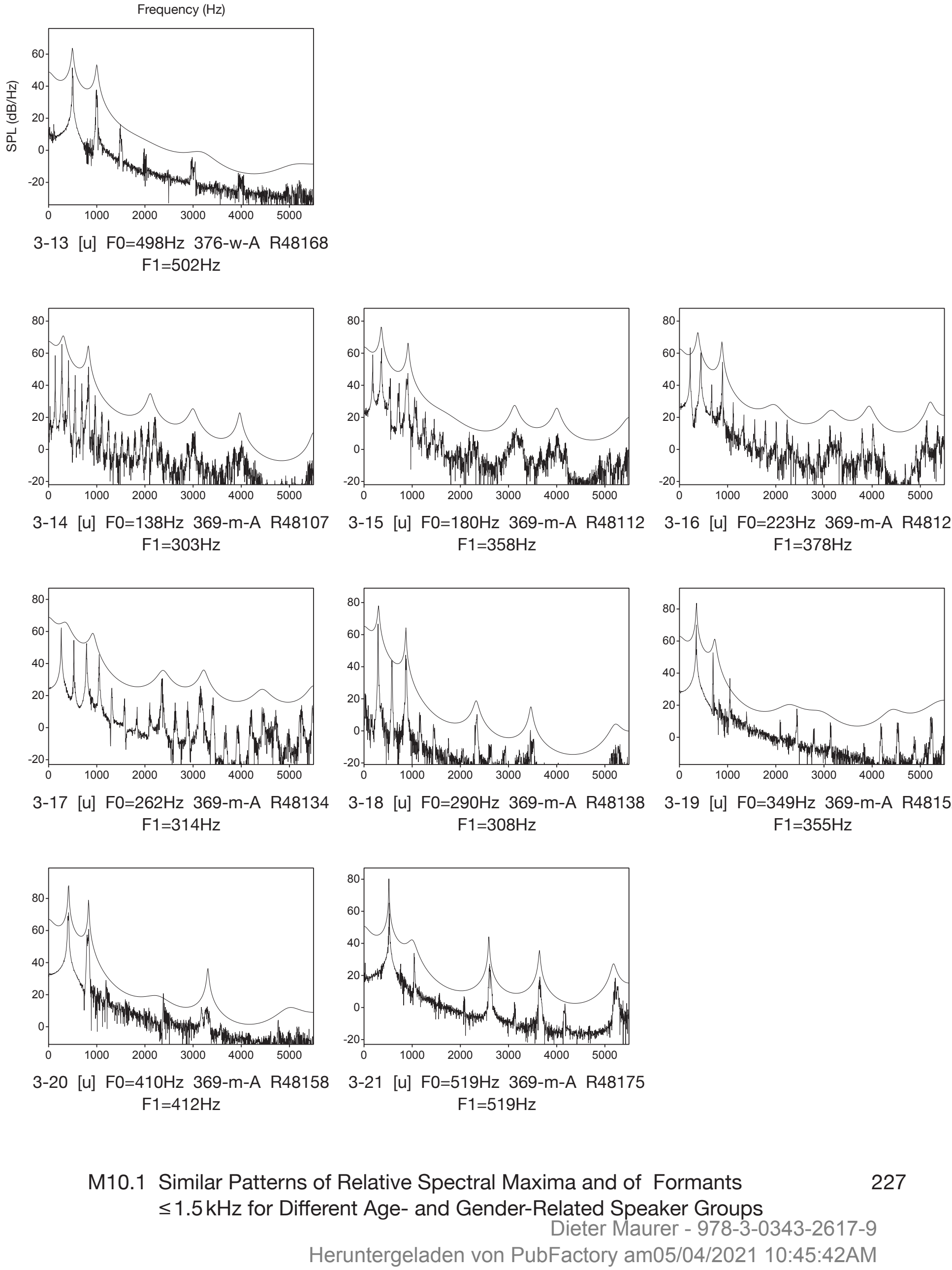
3-9 [u] F0=266Hz 376-w-A R48131
F1=300Hz

3-10 [u] F0=290Hz 376-w-A R48141
F1=(c. 300)Hz

3-11 [u] F0=347Hz 376-w-A R48151
F1=335Hz

3-12 [u] F0=398Hz 376-w-A R48159
F1=404Hz

(Figure 3, continuation)



3-13 [u] F0=498Hz 376-w-A R48168
F1=502Hz

3-14 [u] F0=138Hz 369-m-A R48107
F1=303Hz

3-15 [u] F0=180Hz 369-m-A R48112
F1=358Hz

3-16 [u] F0=223Hz 369-m-A R48120
F1=378Hz

3-17 [u] F0=262Hz 369-m-A R48134
F1=314Hz

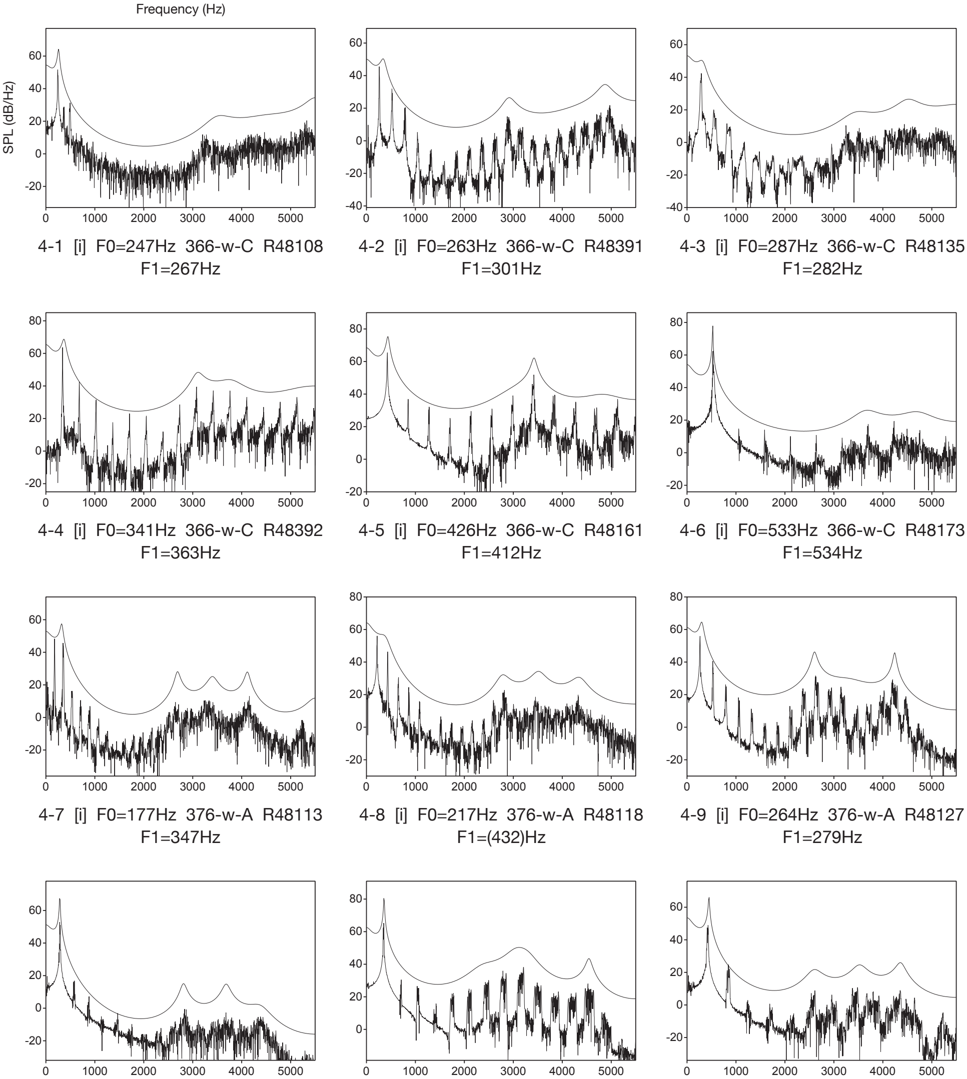
3-18 [u] F0=290Hz 369-m-A R48138
F1=308Hz

3-19 [u] F0=349Hz 369-m-A R48152
F1=355Hz

3-20 [u] F0=410Hz 369-m-A R48158
F1=412Hz

3-21 [u] F0=519Hz 369-m-A R48175
F1=519Hz

Figure 4. Sounds of /i/ produced by a child, a woman and a man at comparable levels of F0. Fig. 4-1 to 4-6 = sounds of the child; Fig. 4-7 to 4-13 = sounds of the woman; Fig. 4-14 to 4-21 = sounds of the man.



4-1 [i] F0=247Hz 366-w-C R48108
F1=267Hz

4-2 [i] F0=263Hz 366-w-C R48391
F1=301Hz

4-3 [i] F0=287Hz 366-w-C R48135
F1=282Hz

4-4 [i] F0=341Hz 366-w-C R48392
F1=363Hz

4-5 [i] F0=426Hz 366-w-C R48161
F1=412Hz

4-6 [i] F0=533Hz 366-w-C R48173
F1=534Hz

4-7 [i] F0=177Hz 376-w-A R48113
F1=347Hz

4-8 [i] F0=217Hz 376-w-A R48118
F1=(432)Hz

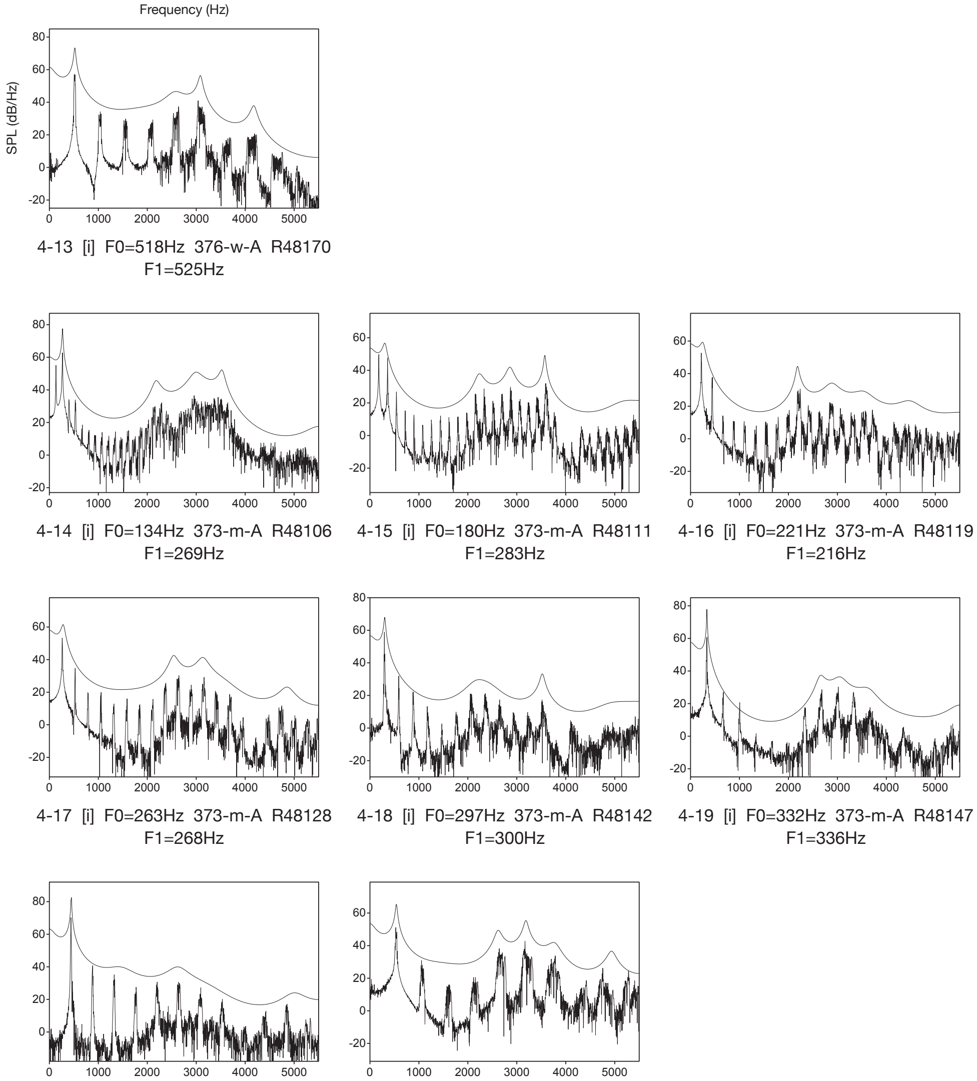
4-9 [i] F0=264Hz 376-w-A R48127
F1=279Hz

4-10 [i] F0=287Hz 376-w-A R48136
F1=293Hz

4-11 [i] F0=351Hz 376-w-A R48393
F1=363Hz

4-12 [i] F0=424Hz 376-w-A R48162
F1=437Hz

(Figure 4, continuation)



4-13 [i] F0=518Hz 376-w-A R48170
F1=525Hz

4-14 [i] F0=134Hz 373-m-A R48106
F1=269Hz

4-15 [i] F0=180Hz 373-m-A R48111
F1=283Hz

4-16 [i] F0=221Hz 373-m-A R48119
F1=216Hz

4-17 [i] F0=263Hz 373-m-A R48128
F1=268Hz

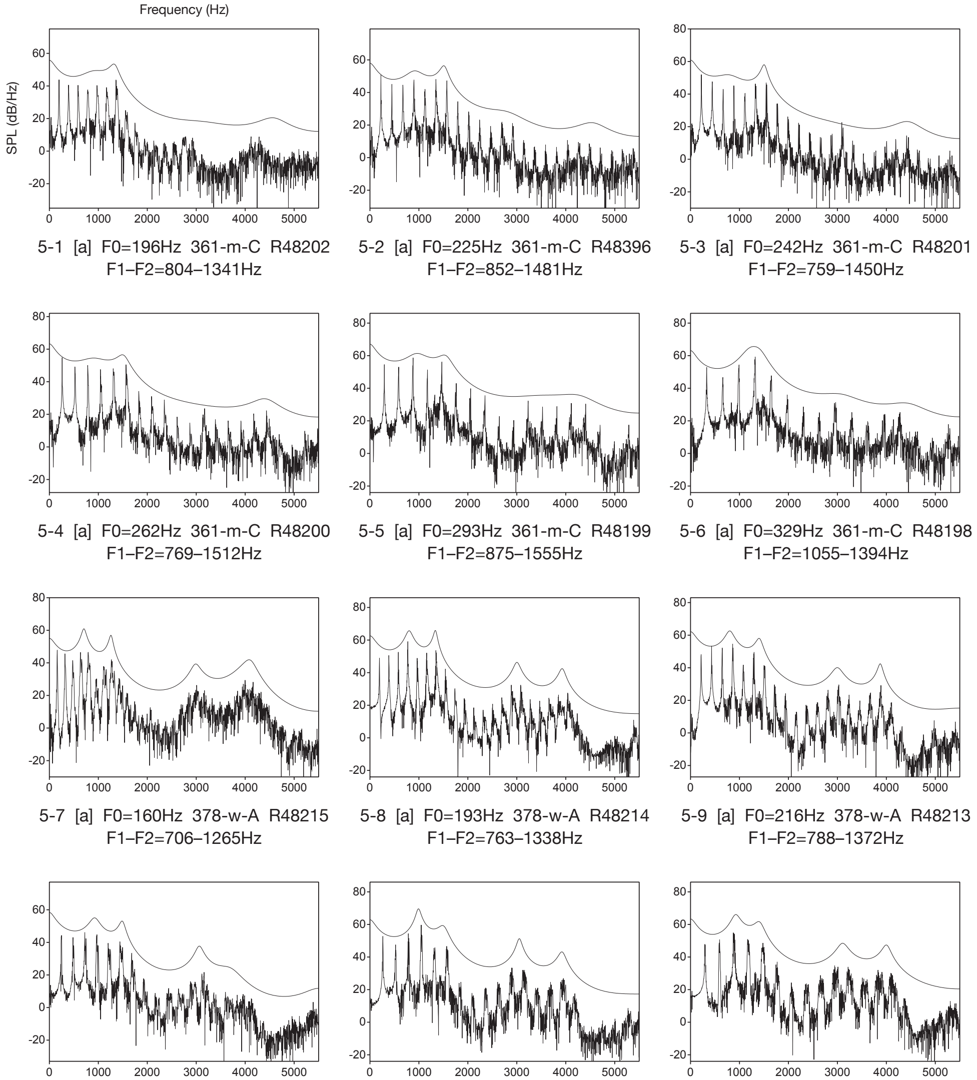
4-18 [i] F0=297Hz 373-m-A R48142
F1=300Hz

4-19 [i] F0=332Hz 373-m-A R48147
F1=336Hz

4-20 [i] F0=441Hz 373-m-A R48163
F1=444Hz

4-21 [i] F0=534Hz 373-m-A R48394
F1=550Hz

Figure 5. Sounds of /a/ produced by a child, a woman and a man at comparable levels of F0. Fig. 5-1 to 5-6 = sounds of the child; Fig. 5-7 to 5-13 = sounds of the woman; Fig. 5-14 to 5-21 = sounds of the man.



5-1 [a] F0=196Hz 361-m-C R48202
F1–F2=804–1341Hz

5-2 [a] F0=225Hz 361-m-C R48396
F1–F2=852–1481Hz

5-3 [a] F0=242Hz 361-m-C R48201
F1–F2=759–1450Hz

5-4 [a] F0=262Hz 361-m-C R48200
F1–F2=769–1512Hz

5-5 [a] F0=293Hz 361-m-C R48199
F1–F2=875–1555Hz

5-6 [a] F0=329Hz 361-m-C R48198
F1–F2=1055–1394Hz

5-7 [a] F0=160Hz 378-w-A R48215
F1–F2=706–1265Hz

5-8 [a] F0=193Hz 378-w-A R48214
F1–F2=763–1338Hz

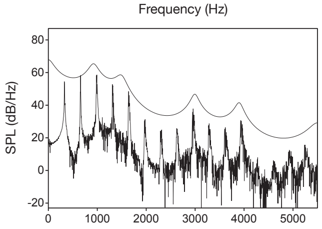
5-9 [a] F0=216Hz 378-w-A R48213
F1–F2=788–1372Hz

5-10 [a] F0=244Hz 378-w-A R48398
F1–F2=956–1502Hz

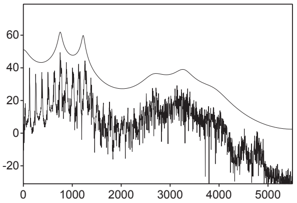
5-11 [a] F0=263Hz 378-w-A R48212
F1–F2=1007–1503Hz

5-12 [a] F0=295Hz 378-w-A R48211
F1–F2=875–1469Hz

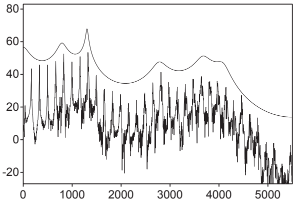
(Figure 5, continuation)



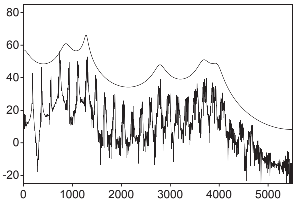
5-13 [a] F0=329Hz 378-w-A R48210
F1–F2=1005–1500Hz



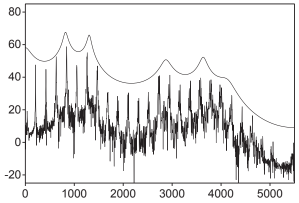
5-14 [a] F0=126Hz 371-m-A R48208
F1–F2=765–1232Hz



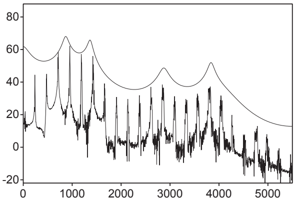
5-15 [a] F0=166Hz 371-m-A R48207
F1–F2=758–1308Hz



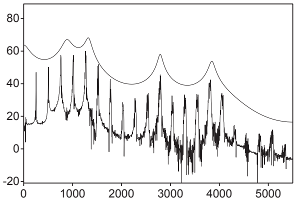
5-16 [a] F0=185Hz 371-m-A R48206
F1–F2=802–1286Hz



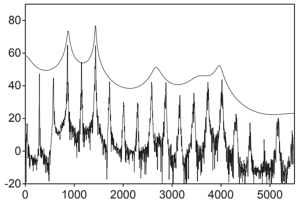
5-17 [a] F0=210Hz 371-m-A R48205
F1–F2=797–1316Hz



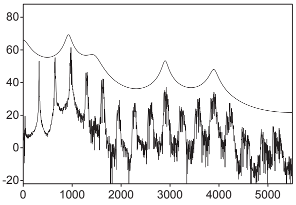
5-18 [a] F0=238Hz 371-m-A R48397
F1–F2=868–1316Hz



5-19 [a] F0=254Hz 371-m-A R48204
F1–F2=882–1363Hz

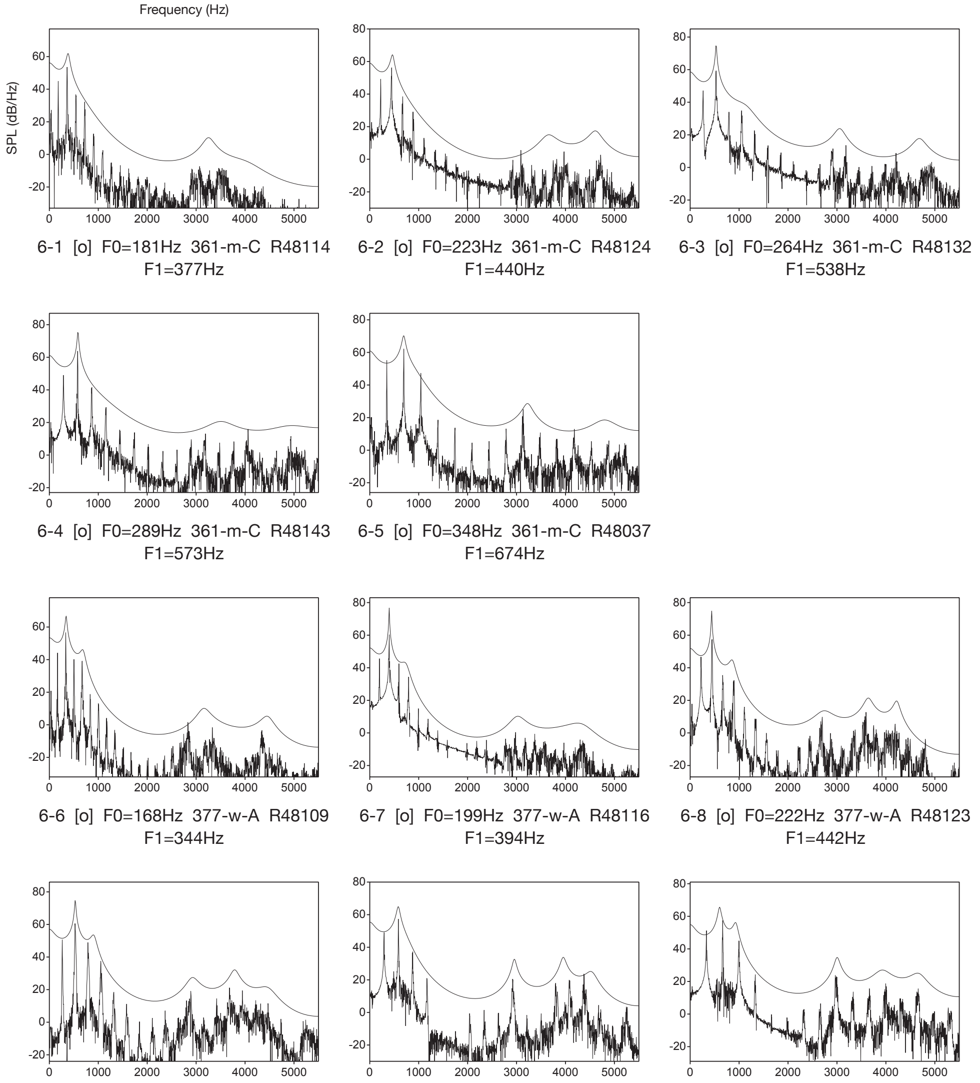


5-20 [a] F0=287Hz 371-m-A R48203
F1–F2=877–1431Hz



5-21 [a] F0=324Hz 371-m-A R48209
F1–F2=898–1417Hz

Figure 6. Sounds of /o/ produced by a child, a woman (untrained speaker) and a man (professional opera singer, baritone) at comparable levels of F0. Fig. 6-1 to 6-5 = sounds of the child; Fig. 6-6 to 6-11 = sounds of the woman; Fig. 6-12 to 6-17 = sounds of the man.



6-1 [o] F0=181Hz 361-m-C R48114
F1=377Hz

6-2 [o] F0=223Hz 361-m-C R48124
F1=440Hz

6-3 [o] F0=264Hz 361-m-C R48132
F1=538Hz

6-4 [o] F0=289Hz 361-m-C R48143
F1=573Hz

6-5 [o] F0=348Hz 361-m-C R48037
F1=674Hz

6-6 [o] F0=168Hz 377-w-A R48109
F1=344Hz

6-7 [o] F0=199Hz 377-w-A R48116
F1=394Hz

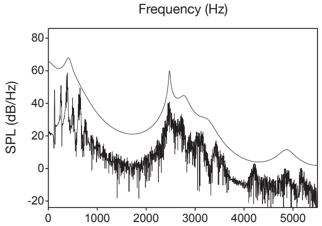
6-8 [o] F0=222Hz 377-w-A R48123
F1=442Hz

6-9 [o] F0=264Hz 377-w-A R48130
F1=531Hz

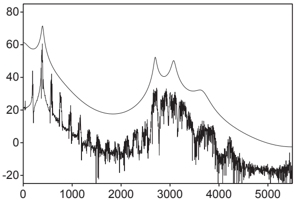
6-10 [o] F0=294Hz 377-w-A R48139
F1=593Hz

6-11 [o] F0=332Hz 377-w-A R48149
F1=584Hz

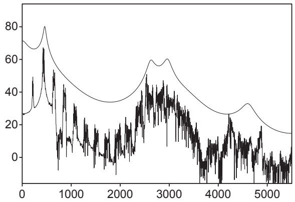
(Figure 6, continuation)



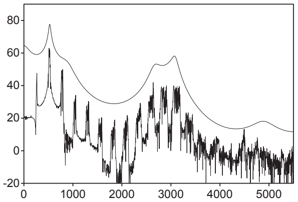
6-12 [o] F0=127Hz 358-m-A R48105
F1=430Hz



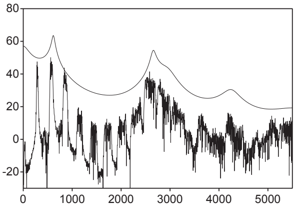
6-13 [o] F0=192Hz 358-m-A R48115
F1=386Hz



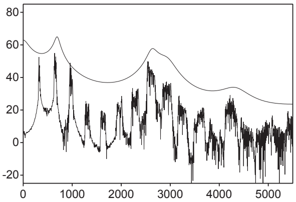
6-14 [o] F0=216Hz 358-m-A R48121
F1=461Hz



6-15 [o] F0=261Hz 358-m-A R48129
F1=511Hz

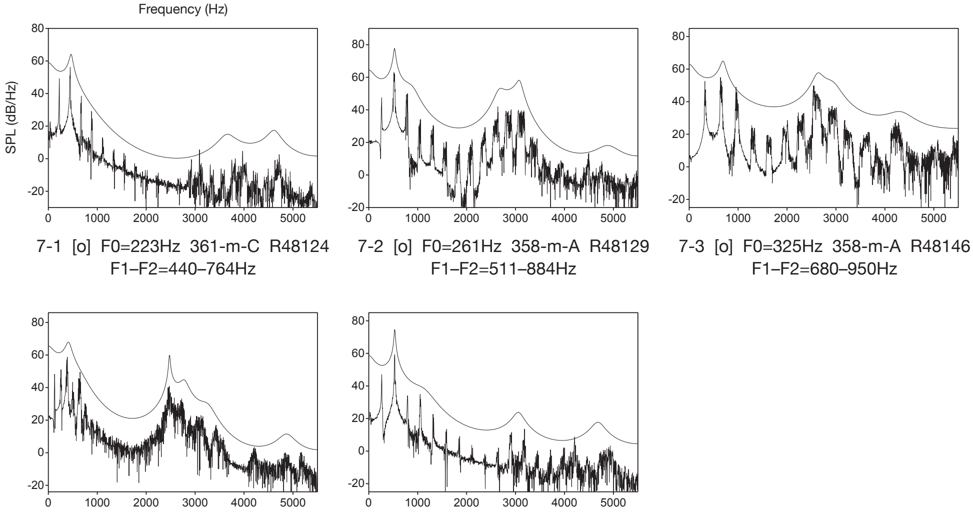


6-16 [o] F0=284Hz 358-m-A R48137
F1=619Hz



6-17 [o] F0=325Hz 358-m-A R48146
F1=680Hz

Figure 7. "Inverted" age- or size-related differences in the vowel-related lower spectral peak(s) and calculated F1 (and F2) for sounds of /o/ produced by a child and a man (see Fig. 7-1 to 7-3), and "expected" age- or size-related differences (see Fig. 7-4 and 7-5). Comparison of selected sounds of Figure 6.



7-1 [o] F0=223Hz 361-m-C R48124
F1–F2=440–764Hz

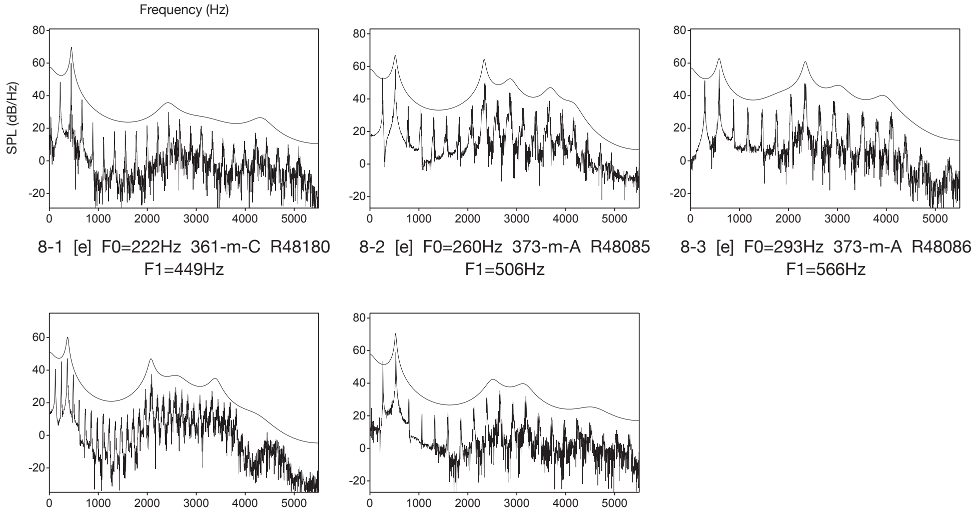
7-2 [o] F0=261Hz 358-m-A R48129
F1–F2=511–884Hz

7-3 [o] F0=325Hz 358-m-A R48146
F1–F2=680–950Hz

7-4 [o] F0=127Hz 358-m-A R48105
F1–F2=430–535Hz

7-5 [o] F0=264Hz 361-m-C R48132
F1–F2=538–1069Hz

Figure 8. “Inverted” age- or size-related differences in the vowel-related lower spectral peak and calculated F1 for sounds of /e/ produced by a child and a man (see Fig. 8-1 to 8-3), and “expected” age- or size-related differences (see Fig. 8-4 and 8-5). Comparison of selected sounds of Figure 2.



8-1 [e] F0=222Hz 361-m-C R48180
F1=449Hz

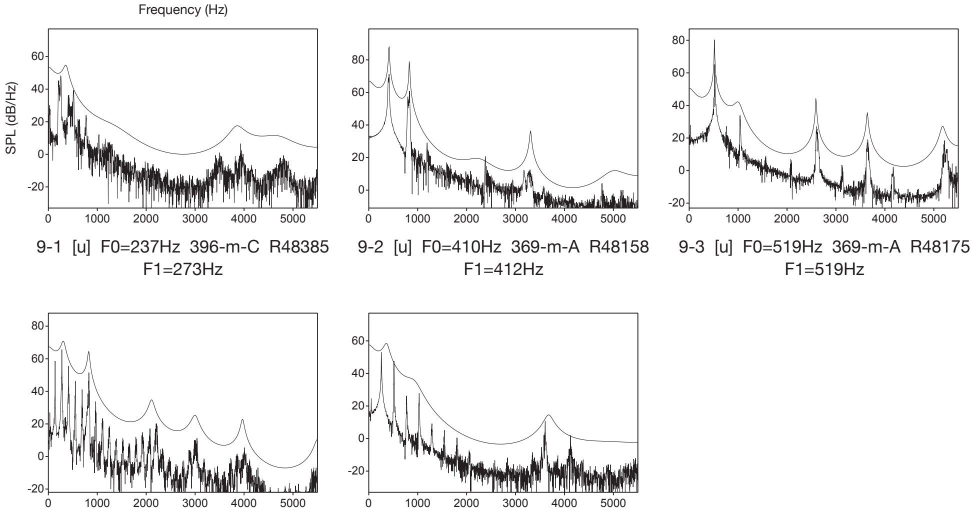
8-2 [e] F0=260Hz 373-m-A R48085
F1=506Hz

8-3 [e] F0=293Hz 373-m-A R48086
F1=566Hz

8-4 [e] F0=122Hz 373-m-A R48081
F1=370Hz

8-5 [e] F0=265Hz 361-m-C R48184
F1=518Hz

Figure 9. “Inverted” age- or size-related differences in the vowel-related lower spectral peak(s) and calculated F1 (and F2) for sounds of /u/ produced by a child and a man (see Fig. 9-1 to 9-3), and “expected” age- or size-related differences (see Fig. 9-4 and 9-5). Comparison of selected sounds of Figure 3.



9-1 [u] F0=237Hz 396-m-C R48385
F1=273Hz

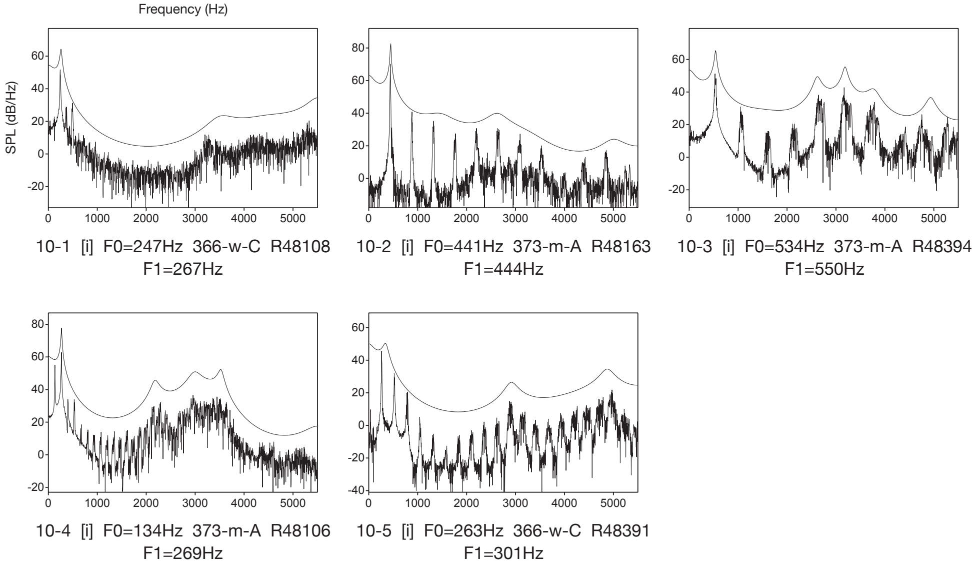
9-2 [u] F0=410Hz 369-m-A R48158
F1=412Hz

9-3 [u] F0=519Hz 369-m-A R48175
F1=519Hz

9-4 [u] F0=138Hz 369-m-A R48107
F1=303Hz

9-5 [u] F0=257Hz 396-m-C R48386
F1=346Hz

Figure 10. “Inverted” age- or size-related differences in the vowel-related lower spectral peak and calculated F1 for sounds of / i / produced by a child and a man (see Fig. 10-1 to 10-3) and “expected” age- or size-related differences (see Fig. 10-4 and 10-5). Comparison of selected sounds of Figure 4.



10-1 [i] F0=247Hz 366-w-C R48108
F1=267Hz

10-2 [i] F0=441Hz 373-m-A R48163
F1=444Hz

10-3 [i] F0=534Hz 373-m-A R48394
F1=550Hz

10-4 [i] F0=134Hz 373-m-A R48106
F1=269Hz

10-5 [i] F0=263Hz 366-w-C R48391
F1=301Hz

M10.2 The Dichotomy of the Vowel Spectrum

In Chapter 10.1, we have argued that the spectrum of a vowel sound needs a twofold rather than a uniform consideration, because only the vowel-related spectrum ≤ 1.5 kHz clearly depends on F0 and, therefore, is not generally specific to speaker groups and vocal-tract sizes. Figures 7 to 10 in the previous chapter illustrate this dichotomy of the vowel spectrum.

M10.A Addition: Vowel Imitations by Birds

The following series show examples of vowel sounds of common hill mynah birds *(Gracula religiosa)* imitating vocal expressions and words of humans. The examples are selected on the basis of extensive recordings of 21 birds, most of them living in Indonesia. (However, they imitated words of different languages.) The spectra presented relate to vowel nuclei extracted from the expressions or words. Both the entire imitated expressions or words as well as the extracted sound fragments are perceptually recognisable.

In each of the series, the sound spectra are given in the order of the birds and of F0. (Note that in several cases, different sound spectra for the same vowel are shown for a bird, in order to document variations in F0 and the sound spectra.)—Acoustic analysis corresponds to the analysis as described in the Note on the Method section. LPC filter curves relate to a parameter setting of the LPC analysis according to the PRAAT standard for women. However, as mentioned in the text, the LPC analysis is not methodically substantiated.

Figure 11 Examples of sounds of imitated / i / in word context produced by five birds, with F0 ranging from c. 110–380 Hz; perceptual vowel quality is / i /, including intermediate qualities / i–j /, / i–y/ and / i–e/

Figure 12 Examples of sounds of imitated /e/ in word context produced by five birds, with F0 ranging from c. 160–330 Hz; perceptual vowel quality is /e/, including intermediate qualities /e–i / and /e–ø/

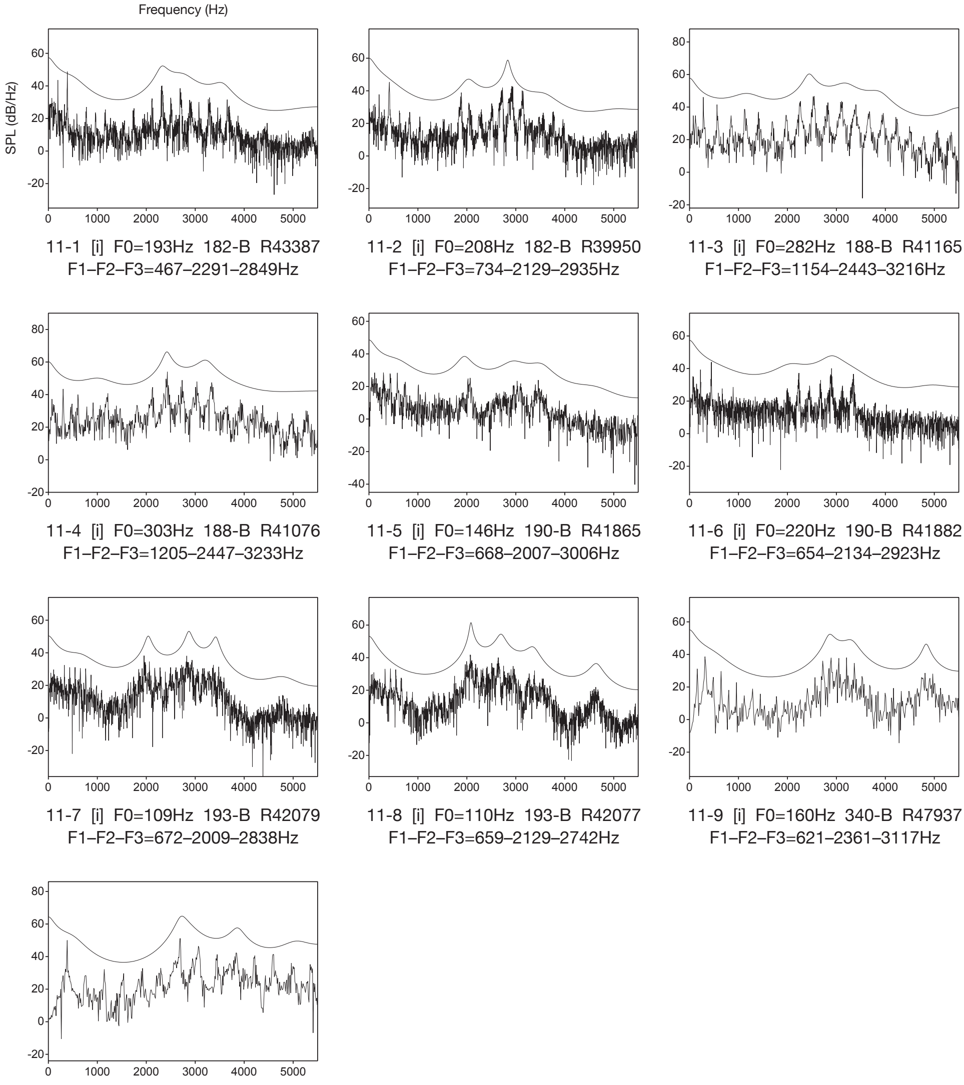
Figure 13 Examples of sounds of imitated /a / in word context produced by twelve birds, with F0 ranging from c. 110–490 Hz; perceptual vowel quality is /a–ɑ/, including intermediate quality /ɑ– ɔ/

Figure 14 Examples of sounds of imitated /o/ in word context produced by eleven birds, with F0 ranging from c. 80–410 Hz; perceptual vowel quality is /o/, including intermediate qualitiy /o–ɔ/

Figure 15 Examples of sounds of imitated /u/ in word context produced by seven birds, with F0 ranging from c. 110–660 Hz; perceptual vowel quality is /u/, including intermediate quality /u–o/

Note that many of the sound spectra of these birds are similar to the vowel spectra of humans presented in the previous sections. However, for some examples of imitations of front vowels, the lower part of the spectral configuration < 1 kHz is "unexpected".

Figure 11. Sounds of /i/ in word context imitated by mynah birds.



11-1 [i] F0=193Hz 182-B R43387
F1–F2–F3=467–2291–2849Hz

11-2 [i] F0=208Hz 182-B R39950
F1–F2–F3=734–2129–2935Hz

11-3 [i] F0=282Hz 188-B R41165
F1–F2–F3=1154–2443–3216Hz

11-4 [i] F0=303Hz 188-B R41076
F1–F2–F3=1205–2447–3233Hz

11-5 [i] F0=146Hz 190-B R41865
F1–F2–F3=668–2007–3006Hz

11-6 [i] F0=220Hz 190-B R41882
F1–F2–F3=654–2134–2923Hz

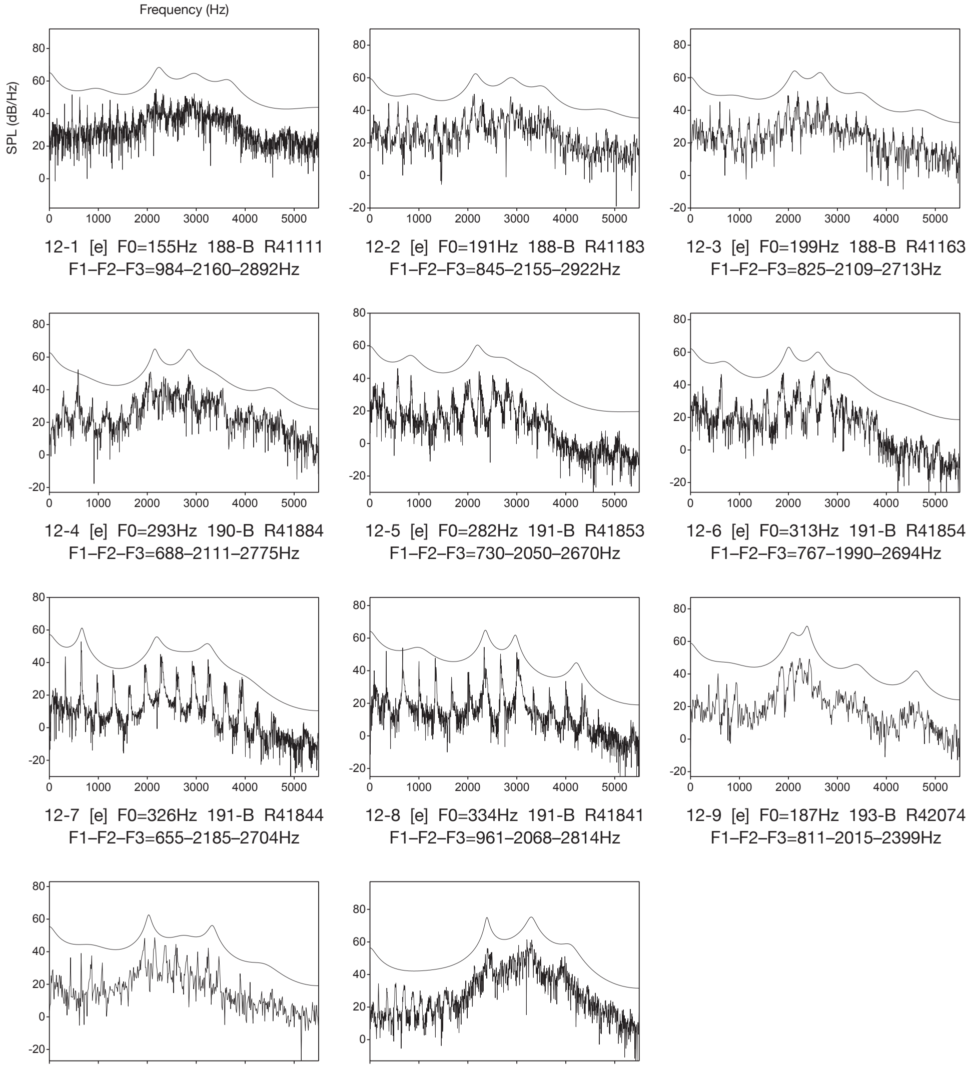
11-7 [i] F0=109Hz 193-B R42079
F1–F2–F3=672–2009–2838Hz

11-8 [i] F0=110Hz 193-B R42077
F1–F2–F3=659–2129–2742Hz

11-9 [i] F0=160Hz 340-B R47937
F1–F2–F3=621–2361–3117Hz

11-10 [i] F0=382Hz 340-B R47949
F1–F2–F3=602–2645–2898Hz

Figure 12. Sounds of /e/ in word context imitated by mynah birds.



12-1 [e] F0=155Hz 188-B R41111
F1–F2–F3=984–2160–2892Hz

12-2 [e] F0=191Hz 188-B R41183
F1–F2–F3=845–2155–2922Hz

12-3 [e] F0=199Hz 188-B R41163
F1–F2–F3=825–2109–2713Hz

12-4 [e] F0=293Hz 190-B R41884
F1–F2–F3=688–2111–2775Hz

12-5 [e] F0=282Hz 191-B R41853
F1–F2–F3=730–2050–2670Hz

12-6 [e] F0=313Hz 191-B R41854
F1–F2–F3=767–1990–2694Hz

12-7 [e] F0=326Hz 191-B R41844
F1–F2–F3=655–2185–2704Hz

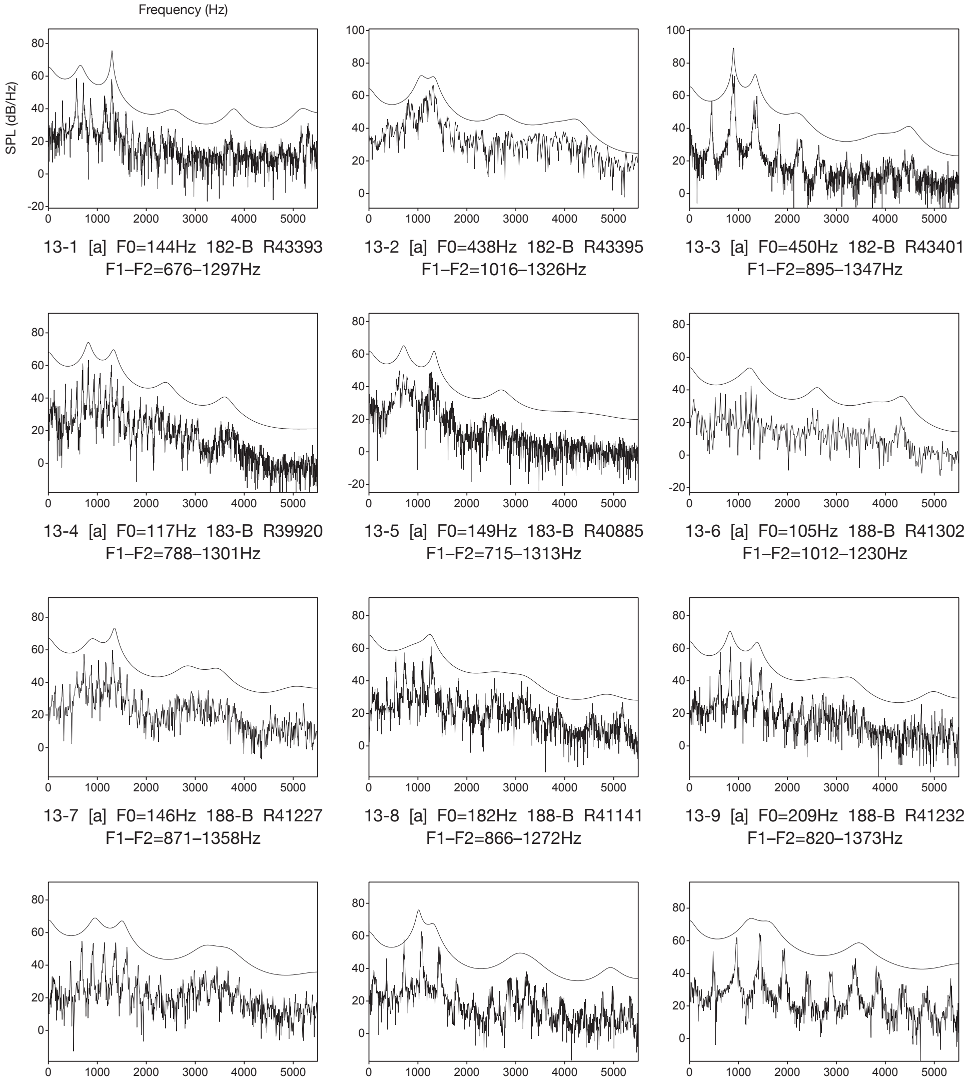
12-8 [e] F0=334Hz 191-B R41841
F1–F2–F3=961–2068–2814Hz

12-9 [e] F0=187Hz 193-B R42074
F1–F2–F3=811–2015–2399Hz

12-10 [e] F0=216Hz 193-B R42067
F1–F2–F3=833–2016–2637Hz

12-11 [e] F0=177Hz 195-B R42183
F1–F2–F3=1230–2412–3206Hz

Figure 13. Sounds of /a–ɑ/ in word context imitated by mynah birds.



13-1 [a] F0=144Hz 182-B R43393
F1–F2=676–1297Hz

13-2 [a] F0=438Hz 182-B R43395
F1–F2=1016–1326Hz

13-3 [a] F0=450Hz 182-B R43401
F1–F2=895–1347Hz

13-4 [a] F0=117Hz 183-B R39920
F1–F2=788–1301Hz

13-5 [a] F0=149Hz 183-B R40885
F1–F2=715–1313Hz

13-6 [a] F0=105Hz 188-B R41302
F1–F2=1012–1230Hz

13-7 [a] F0=146Hz 188-B R41227
F1–F2=871–1358Hz

13-8 [a] F0=182Hz 188-B R41141
F1–F2=866–1272Hz

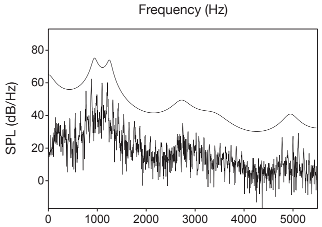
13-9 [a] F0=209Hz 188-B R41232
F1–F2=820–1373Hz

13-10 [a] F0=228Hz 188-B R41160
F1–F2=877–1487Hz

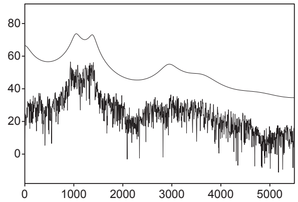
13-11 [a] F0=359Hz 188-B R41280
F1–F2=999–1363Hz

13-12 [a] F0=481Hz 188-B R41273
F1–F2=1223–1644Hz

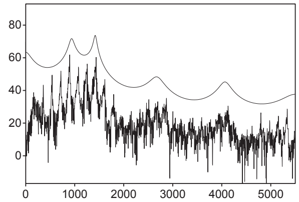
(Figure 13, continuation)



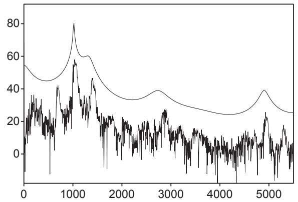
13-13 [a] F0=110Hz 189-B R41508
F1–F2=938–1247Hz



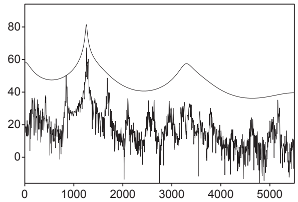
13-14 [a] F0=161Hz 189-B R41721
F1–F2=1020–1381Hz



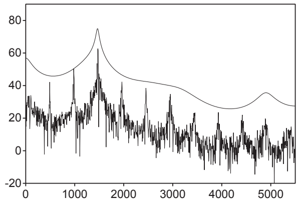
13-15 [a] F0=178Hz 189-B R41655
F1–F2=922–1406Hz



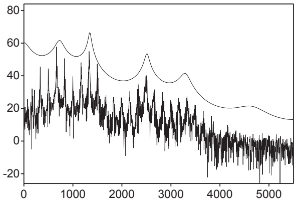
13-16 [a] F0=343Hz 189-B R41625
F1–F2=1016–1285Hz



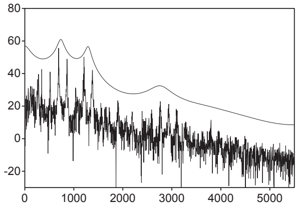
13-17 [a] F0=425Hz 189-B R41607
F1–F2=1260–1421Hz



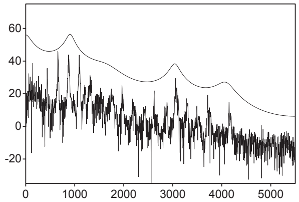
13-18 [a] F0=491Hz 189-B R41695
F1–F2=1287–1492Hz



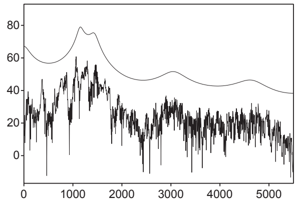
13-19 [a] F0=166Hz 190-B R41871
F1–F2=729–1355Hz



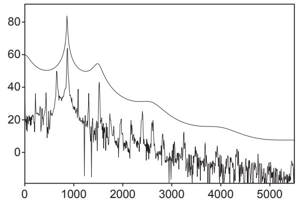
13-20 [a] F0=172Hz 191-B R41848
F1–F2=735–1290Hz



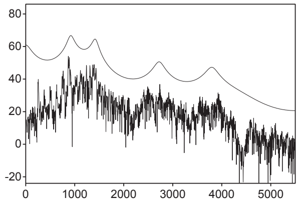
13-21 [a] F0=219Hz 191-B R41849
F1–F2=874–1614Hz



13-22 [a] F0=400Hz 191-B R41858
F1–F2=1046–1476Hz

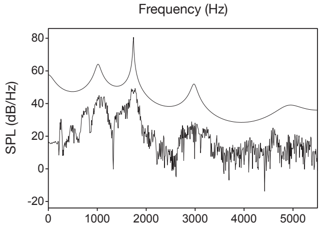


13-23 [a] F0=217Hz 221-B R43451
F1–F2=862–1490Hz

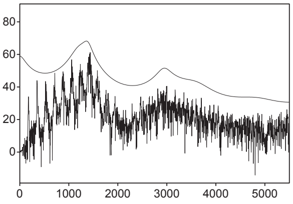


13-24 [a] F0=127Hz 337-B R47708
F1–F2=938–1427Hz

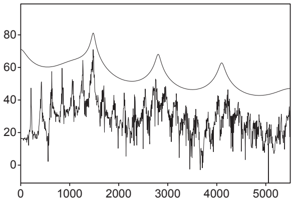
(Figure 13, continuation)



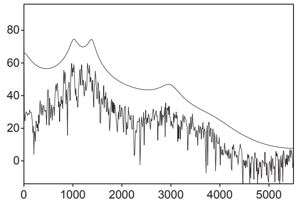
13-25 [a] F0=248Hz 339-B R47952
F1–F2=1044–1726Hz



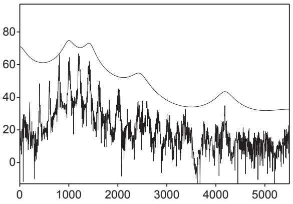
13-26 [a] F0=177Hz 340-B R47947
F1–F2=1068–1435Hz



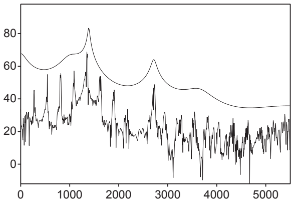
13-27 [a] F0=211Hz 340-B R47948
F1–F2=1110–1470Hz



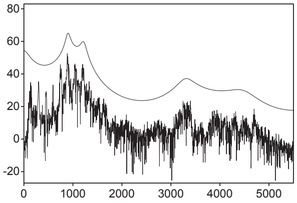
13-28 [a] F0=165Hz 345-B R47913
F1–F2=993–1374Hz



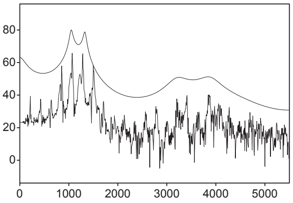
13-29 [a] F0=202Hz 345-B R47912
F1–F2=988–1361Hz



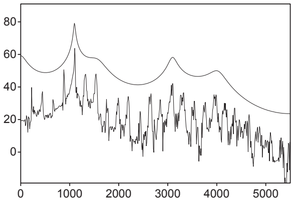
13-30 [a] F0=272Hz 345-B R47907
F1–F2=1118–1361Hz



13-31 [a] F0=149Hz 346-B R47903
F1–F2=890–1217Hz

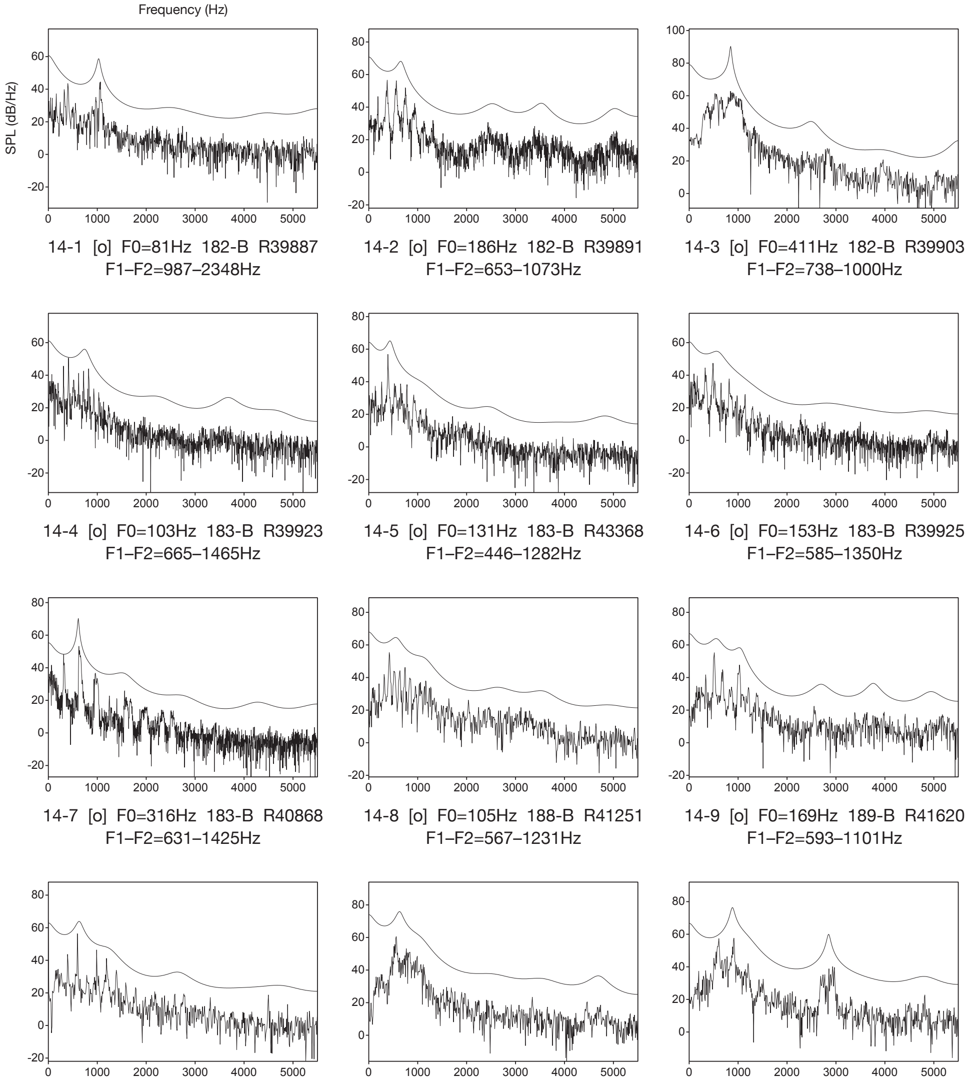


13-32 [a] F0=213Hz 346-B R47899
F1–F2=1055–1347Hz



13-33 [a] F0=220Hz 346-B R47893
F1–F2=1097–1556Hz

Figure 14. Sounds of /o/ in word context imitated by mynah birds.



14-1 [o] F0=81Hz 182-B R39887
F1–F2=987–2348Hz

14-2 [o] F0=186Hz 182-B R39891
F1–F2=653–1073Hz

14-3 [o] F0=411Hz 182-B R39903
F1–F2=738–1000Hz

14-4 [o] F0=103Hz 183-B R39923
F1–F2=665–1465Hz

14-5 [o] F0=131Hz 183-B R43368
F1–F2=446–1282Hz

14-6 [o] F0=153Hz 183-B R39925
F1–F2=585–1350Hz

14-7 [o] F0=316Hz 183-B R40868
F1–F2=631–1425Hz

14-8 [o] F0=105Hz 188-B R41251
F1–F2=567–1231Hz

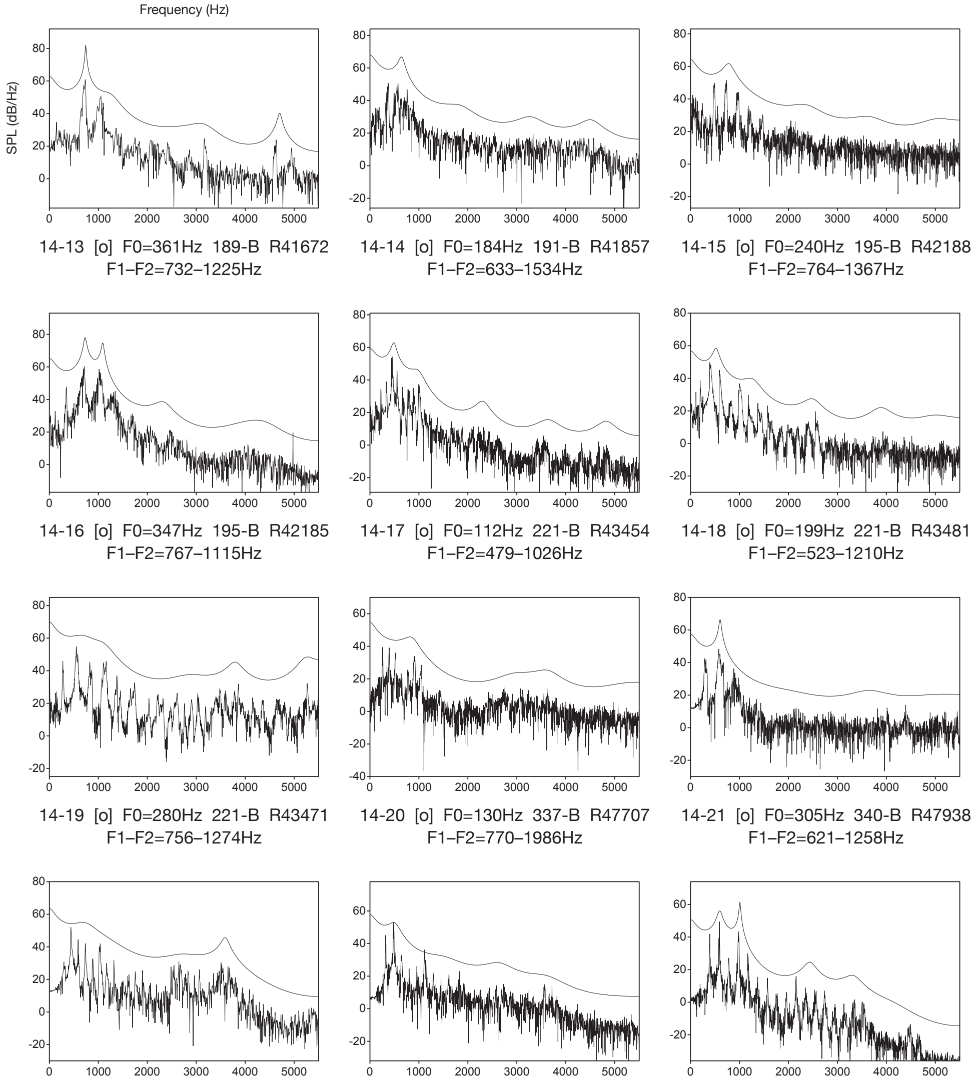
14-9 [o] F0=169Hz 189-B R41620
F1–F2=593–1101Hz

14-10 [o] F0=198Hz 189-B R41652
F1–F2=638–1226Hz

14-11 [o] F0=269Hz 189-B R41681
F1–F2=659–1163Hz

14-12 [o] F0=296Hz 189-B R41678
F1–F2=801–1059Hz

(Figure 14, continuation)



14-13 [o] F0=361Hz 189-B R41672
F1–F2=732–1225Hz

14-14 [o] F0=184Hz 191-B R41857
F1–F2=633–1534Hz

14-15 [o] F0=240Hz 195-B R42188
F1–F2=764–1367Hz

14-16 [o] F0=347Hz 195-B R42185
F1–F2=767–1115Hz

14-17 [o] F0=112Hz 221-B R43454
F1–F2=479–1026Hz

14-18 [o] F0=199Hz 221-B R43481
F1–F2=523–1210Hz

14-19 [o] F0=280Hz 221-B R43471
F1–F2=756–1274Hz

14-20 [o] F0=130Hz 337-B R47707
F1–F2=770–1986Hz

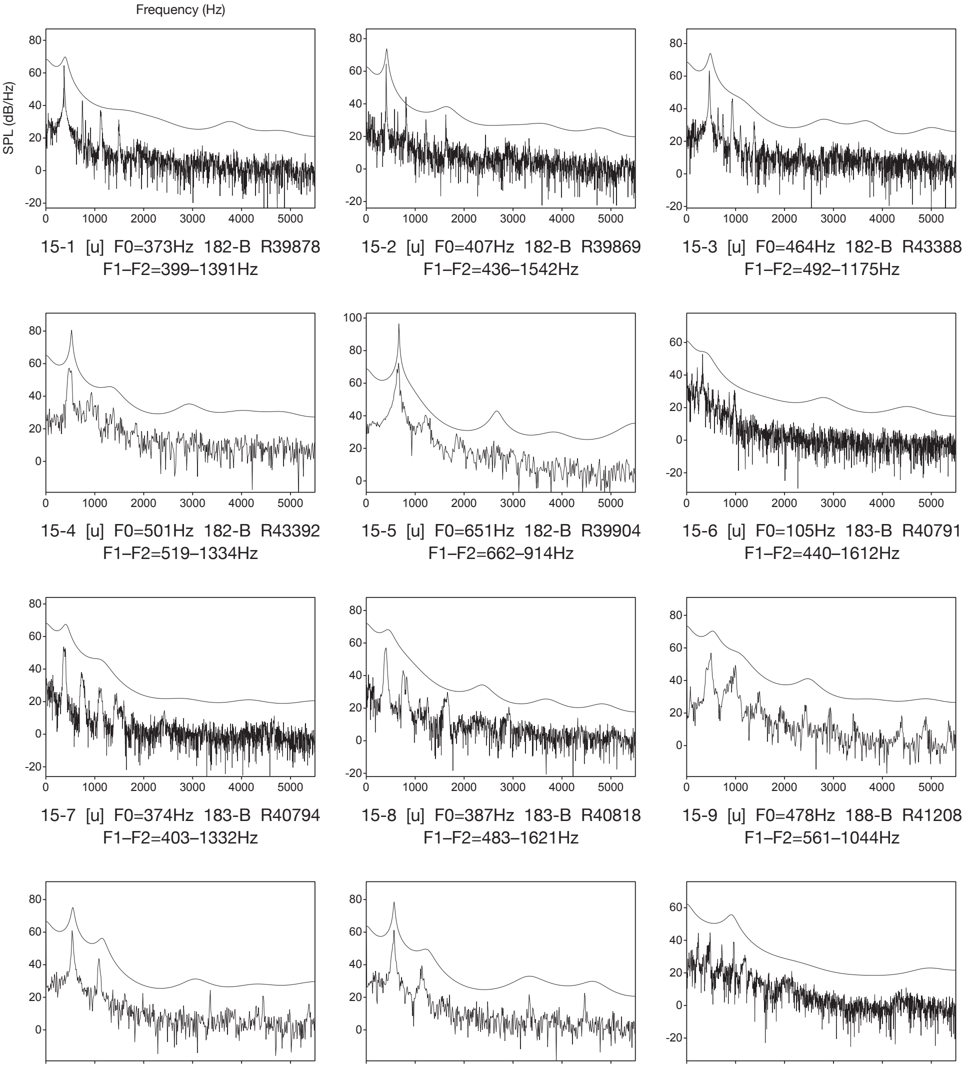
14-21 [o] F0=305Hz 340-B R47938
F1–F2=621–1258Hz

14-22 [o] F0=147Hz 342-B R47934
F1–F2=630–1182Hz

14-23 [o] F0=161Hz 342-B R47923
F1–F2=494–1495Hz

14-24 [o] F0=196Hz 343-B R47932
F1–F2=580–1009Hz

Figure 15. Sounds of /u/ in word context imitated by mynah birds.



15-1 [u] F0=373Hz 182-B R39878
F1–F2=399–1391Hz

15-2 [u] F0=407Hz 182-B R39869
F1–F2=436–1542Hz

15-3 [u] F0=464Hz 182-B R43388
F1–F2=492–1175Hz

15-4 [u] F0=501Hz 182-B R43392
F1–F2=519–1334Hz

15-5 [u] F0=651Hz 182-B R39904
F1–F2=662–914Hz

15-6 [u] F0=105Hz 183-B R40791
F1–F2=440–1612Hz

15-7 [u] F0=374Hz 183-B R40794
F1–F2=403–1332Hz

15-8 [u] F0=387Hz 183-B R40818
F1–F2=483–1621Hz

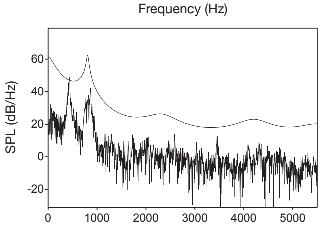
15-9 [u] F0=478Hz 188-B R41208
F1–F2=561–1044Hz

15-10 [u] F0=540Hz 189-B R41572
F1–F2=551–1209Hz

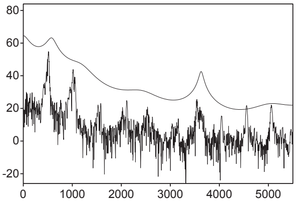
15-11 [u] F0=561Hz 189-B R41631
F1–F2=567–1253Hz

15-12 [u] F0=230Hz 190-B R41894
F1–F2=773–1587Hz

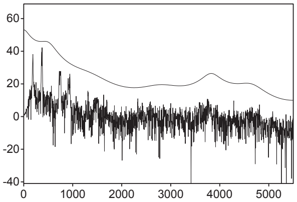
(Figure 15, continuation)



15-13 [u] F0=412Hz 193-B R42086
F1–F2=622–1562Hz



15-14 [u] F0=506Hz 193-B R42071
F1–F2=541–1189Hz



15-15 [u] F0=186Hz 346-B R47896
F1–F2=568–1476Hz

M11 Lack of Correlation between Methodological Limitations of Formant Determination and Limitations of Vowel Perception

M11.1 Vowel Perception at Fundamental Frequencies > 350 Hz

The sound series presented in Sections M8.1 and M8.2 demonstrate that recognisable vowels can be produced at fundamental frequencies substantially exceeding the critical limit above which formants can no longer be reliably determined for methodological reasons.

M11.2 Lack of Correspondence between Methodological Problems of Formant Pattern Estimation at Fundamental Frequencies ≤ 350 Hz and Impaired Vowel Perception

The sound series presented in the Sections M7.1 and M7.2 demonstrate that vowel sounds produced at fundamental frequencies ≤ 350 Hz, for which the estimation of formant patterns proves questionable for reasons other than fundamental frequency—for instance, if expected relative spectral energy maxima are "missing" or if vowel-related parts of a spectrum spectra are "flat"—are not less recognisable than vowel sounds for which formant pattern estimation may be said to be unproblematic.

Experiments

The treatise concludes with a list of possible experiments that allow for empirical exploration of the problems discussed here under laboratory conditions.

E1 Number of Relative Spectral Energy Maxima and Number of Formants

E1.1 Sounds of Back Vowels Showing only One Lower Spectral Peak ≤ 1.5 kHz

To do: (i) Find examples of sounds of back vowels, produced as voiced sounds in isolation, which show only one spectral peak ≤ 1.5 kHz. (ii) Perform a listening test.

Note: For most of the corresponding examples, LPC analysis yields two formants ≤ 1.5 kHz; however, you will find that the second formant is often weak (large second formant bandwidth, low second formant level). You also will find examples for which LPC analysis yields only one lower formant frequency. (Long vowels produced in some languages, such as Standard German, are particularly suited for such an experiment.)

Option: You may also perform resynthesis and perform a related second listening test.

Thesis: You will find many examples for which the vowel identification score is high.

Examples: See Section M7.1, Figures 1 to 3.

E1.2 Sounds of Back Vowels Showing only One Pronounced Lower Formant ≤ 1.5 kHz

To do: (i) From the sample investigated in the previous experiment, select examples of sounds of back vowels for which LPC analysis gives a weak second formant (high bandwidth, low level). (ii) Manipulate these sounds in terms of shaping the spectrum using bandpass filtering including filter slope variation, until LPC analysis gives only one formant ≤ 1.5 kHz. (iii) Perform a listening test.

Thesis: You will find examples for which the perceived vowel quality proves to be maintained for the manipulated sounds.

E1.3 Sounds of Single Front Vowels Showing Non-Corresponding F2 and F3

To do: (i) Find examples of sound pairs of the same intended front vowel, produced as voiced sounds in isolation at similar F0, for which F2 of the first sound is near or above F3 of the second sound. (ii) Perform a listening test.

Option: You may compare sounds produced by speakers of the same age and gender group as well as of different groups. You may also perform resynthesis, and perform a related second listening test. You may also investigate the roles of the higher formants in bandpass filtering single formants.

Thesis: You will find such examples of sound pairs equal in perceived vowel quality.

Examples: See Section M7.1, Figures 8 to 10.

E1.4 Sounds of Back Vowels Showing No Pronounced Spectral Peak ≤ 1.5 kHz

To do: (i) Find examples of sounds of back vowels, produced as voiced sounds in isolation, which show no pronounced spectral peak ≤ 1.5 kHz apart from the fundamental ("flat" spectra, or spectra exhibiting continuously decreasing amplitudes of the harmonics). (ii) Perform a listening test.

Thesis: You will find examples for which the score of vowel identification is high. Further, you may experience examples for which the calculation of F1–F2 depends on rather small amplitude variations of the first harmonics.

Examples: See Section M7.2, Figures 11 and 12.

E1.5 Sounds of Front Vowels Showing No Pronounced Spectral Peak > 2 kHz

To do: (i) Find examples of sounds of front vowels, produced as voiced sounds in isolation, which show no pronounced spectral peak > 2 kHz. (ii) Perform a listening test.

Thesis: You will find examples for which the vowel identification score is high.

Examples: See Section M7.2, Figures 13 and 14.

E2 Patterns of Relative Spectral Energy Maxima, Formant Patterns and Fundamental Frequency

E2.1 Sounds of Single Vowels Produced at Different F0 Exhibiting Different Spectral Peaks and Different Calculated Formant Patterns: Part 1, Dependence of Formant Patterns on F0

To do: (i) Select speakers with excellent vocal abilities. (ii) Investigate all long vowels of the language in question. (iii) Let the speakers produce single words (including word pairs forming minimal pairs), single syllables (including logatomes) and isolated vowel sounds for their entire range of F0 of possible vowel production. (iv) Perform a listening test. (v) Only select sounds with a high identification score. (vi) Perform spectral analysis and LPC analysis.

Options: You may need to train the speakers so as they indeed maintain the perceived vowel while altering F0. You may select professional singers, actresses and actors. You may give special attention to the entertainment sector, including voice-over. You may vary vocal effort. You may include resynthesis. You may also extract words or syllables or vowel nuclei from existing recordings.

Thesis: (i) You will obtain unsystematic results, above all depending on single speakers, F0 levels and vocal effort, frequency ranges of spectral peaks and formants, vowel qualities and additional spectral characteristics of the original sounds. (ii) However, for F0 > 200, the spectral peaks and the calculated lower formants will shift with raising F0 for a substantial part of your sample even if the perceived vowel quality remains the same. (iii) Whether or not you experience a systematic (and not speaker-related) impact of the syntactic or semantic context of the vowel sounds is left open here.

Examples: See Section M8.1, Figures 1 to 5.

E2.2 Sounds of Single Vowels Produced at Different F0 Exhibiting Different Spectral Peaks and Different Calculated Formant Patterns: Part 2, Vowel Intelligibility for Sounds at F0 > 500 Hz

To do: (i) Refer to the sounds of the previous experiment. (ii) Select the sounds at F0 > 500 Hz.

Thesis: (i) You will obtain different results related to the abilities and production styles or habits of the speakers. (ii) However, you will observe possible vowel perception up to F0 corresponding to the upper frequency limit of F1 for men and women as given in formant statistics.

Examples: See Section M8.1, Figures 2 and 3, and Section M8.2, Figures 6 and 7; see also the pitch contours in Section M8.2, Figures 8 to 10.

E2.3 Sounds of Single Vowels Produced at Different F0 Exhibiting Different Spectral Peaks and Different Calculated Formant Patterns: Part 3, Resynthesising a Formant Pattern at Different F0

To do: (i) Refer to the sounds experiment E2.1. (ii) Select two sounds of one vowel exhibiting very different F0 and different spectral peaks or (lower) formants, respectively. (iii) Concatenate these two sounds and insert a pause between them. Eventually, equalise loudness. (iv) Perform resynthesis of the concatenated sound, applying three conditions for F0. Firstly, use F0 of the original sounds; secondly, fix F0 to the value of the original sound at lower F0; thirdly, fix F0 to the value of the original sound at higher F0. (v) Perform a listening test including all sound pairs.

Options: Instead of concatenating two sounds, a singer or speaker with high vocal ability may perform a glissando, and resynthesis is performed at original (altering) F0, fixed F0 corresponding to the lowest, and fixed F0 corresponding to the highest F0 values of the original sound. However, during the production of the glissando, the vowel quality must be strictly maintained.

Thesis: (i) You will obtain unsystematic results (see above). (ii) However, you will find many cases for which the original sounds of a pair as well as the resynthesised sounds, for which the first condition mentioned applies, are perceived as the same vowel, but the resynthesis applying the second and third condition produces a change in vowel perception between the two sounds of a pair.

E2.4 Sounds of Single Back Vowels Produced at Different F0 Exhibiting Inverse Spectral Peaks

To do: Refer to experiment E2.1 but, in particular, consider sound pairs of a back vowel which differ in F0 and exhibit an "inversion" of spectral peaks, that is, the first relative spectral energy maximum (corresponding to its F1) for the sound at higher F0 is found at a frequency level of a relative spectral minimum for the sound at lower F0, in between the first and second spectral peak (in between the F1 and F2) of the latter. Consider also resynthesis and identification scores.

Thesis: You will find many cases for which the sounds of such pairs are perceived as the same vowel.

Examples: See Section M8.3, Figures 11 to 13.

E2.5 Special Note Concerning Inconstant Numerical Relationship between Calculated F0 and Formant Patterns

To do: (i) Refer to sounds at very different F0, above all to sounds of the vowel /e, o/. Include sounds produced with different vocal effort. (ii) Perform a listening test. (iii) Select only sounds with a high identification score. (iv) Calculate formant patterns for these sounds. (v) Perform resynthesis. (vi) Perform a listening test with the resynthesised sounds.

Thesis: (i) You may observe sound pairs for which F1 or F1–F2 of the sound at lower F0 is higher than F1 or F1–F2 at higher F0, thus seemingly indicating an "inverse" dependence of lower formants and F0. (ii) You may also note that resynthesis seems to confirm this observation. (iii) However, you will have to relate such observations to a limited frequency range of F0, differences in vocal effort may have a strong influence on formant estimation and you will have to consider methodological aspects of LPC analysis.

Examples: See Section M8.1, Figure 4 for an indication.

E3 Formant Pattern Ambiguity

E3.1 Formant Pattern Ambiguity in Natural Vocalisations

To do: (i) Select speakers from all three speaker groups with excellent vocal abilities. (ii) Let them produce isolated sounds of long vowels at very different F0. Vary vocal effort, for example medium, low and high vocal effort. Investigate a frequency range of F0 of 220 to 700 Hz for children, 175 to 880 Hz for women and 110 to 523 Hz for men. Investigate different F0 step by step (you may refer to a musical scale). (iii) Perform spectral analysis and formant pattern analysis. With regard to the latter, you may perform the analysis also for F0 > 350 Hz even if there is a lack of methodological substantiation. (iv) Perform a listening test.

Thesis: (i) You will find unsystematic results (see above). (ii) However, comparing the vowel-related patterns of spectral peaks and of formants of the sounds of a single speaker, you will find many examples of similar patterns for sounds at different F0 and two different perceived vowel qualities. You may even encounter examples of such patterns for three vowels. (iii) The same holds true in an extended way for a corresponding comparison of the sounds of different speakers. (iv) You will not be able, in general terms, to directly relate such a pattern ambiguity to differences in speaker group or vocal effort.

Examples: See Section M9.1, Figures 1 to 21.

E3.2 Formant Pattern Ambiguity in Model Synthesis

To do: (i) Refer to the sounds in experiment E3.1. (ii) Select sounds of different vowels for which—apart from differences in F0 and the frequency distance of the harmonics—a direct comparison of the vowel-related spectral region as well as the corresponding spectral peaks and formant patterns can be considered similar, according to prevailing consideration in phonetics. (iii) Use the related formant patterns (including formant bandwidths) as models for vowel synthesis. (iv) Perform vowel synthesis for the entire range of the F0 you have investigated in the previous experiment. (v) Perform a listening test.

Thesis: You will observe that, for selected formant patterns of natural vocalisations that prove to be ambiguous in vowel representation, the alteration of F0 in such a model synthesis generally produces a clear and sometimes very pronounced change in perceived vowel quality.

E4 Patterns of Relative Spectral Energy Maxima, Formant Patterns and Age- and Gender-Related Vocal-Tract Sizes

E4.1 Comparison of Vowel-Specific Spectral Characteristics of Children, Women and Men Related to Different and Similar F0 of Vocalisations: Part 1, Natural Vocalisations

To do: (i) Select a child, a woman and a man with excellent vocal abilities. (ii) Let them produce isolated sounds of long vowels at different F0 according to the C-major scale, for example starting from 220 Hz for the child, from 175 Hz for the woman and from 131 Hz for the man. Investigate a range of F0 up to 523 Hz. Ensure that the sounds correspond with each other perceptually, not only in vowel quality but also in "vowel-colour" variant, which makes for the greatest possible correspondence as regards perception (exclusion of age- and gender-related "dialects"). (iii) Perform a listening test. (iv) Perform spectral analysis and compare the spectra and the spectral peaks of the sounds of a singe vowel. (iv) In parallel, perform formant analysis and compare the formant patterns of the sounds of a single vowel.

Option: You may proceed in a similar way with several speakers from the three speaker groups as to re-examine formant statistics. However, you will not be able to control the correspondences of the vowel qualities as precisely as in an investigation of the utterances of three single speakers.

Thesis: (i) With regard to the spectral characteristics in general and the spectral peaks in particular, you will find the expected differences which are in line with the numbers given in formant statistics for citation-form words, if the F0 of the sounds also concurs with the F0 of the statistics in question, that is, c. 262 Hz for the child, c. 220 Hz for the woman and c. 131 Hz for the man (levels given according to the C-major scale). (ii) However, you will observe that spectral differences ≤ 1.5 kHz decrease or disappear if the speakers vocalise at a similar F0. (iii) You will even observe cases of "inversions" of expected age- and gender-related spectral differences in terms of higher spectral peaks ≤ 1.5 kHz for the sounds of the two adults than for the sounds of the child, if the F0 of the former are also higher than of the latter. The same will hold true for the comparison of sounds of the man with sounds of the woman at correspondingly different F0. (iv) With regard to calculated formant patterns, you will observe similar behaviour. However, methodological problems of analysis will interfere. (v) With regard to formant statistics, you will not be able to resolve the methodological problem of formant pattern analysis at F0 > 350 Hz. Moreover, you will have to consider possible age- and gender-related vowel colouring (age- and gender-related "dialects"). However, for sounds > 220 Hz, you will no longer find a clear indication of generalised age- and gender-related formant patterns < 1.5 kHz, if the F0 of the sounds correspond.

Examples: See Section M10.1, Figures 1 to 10.

E4.2 Comparison of Vowel-Specific Spectral Characteristics of Children, Women and Men Related to Different and Similar F0 of Vocalisations: Part 2, Resynthesis

To do: (i) Select the sounds of the three single speakers of the previous experiment. (ii) Resynthesise them on the basis of formant analysis but, for each single formant pattern of a single vocalisation, perform resynthesis for all F0 levels on which the speaker produced vowel sounds. (iii) Perform a listening test.

Thesis: (i) If resynthesis is performed applying F0 and formant patterns of the original sounds, in general, the perceived vowel quality will not change. (ii) If only the formant patterns correspond to the original sounds but F0 is varied according to the F0-range of the natural sounds, you will obtain unsystematic results (see above). However, for some of the vowels investigated and for F0 of the sounds > 200 Hz, for all three speakers, you will find many examples of sounds for which the perceived vowel quality changes with changing F0.

E5 Patterns of Relative Spectral Energy Maxima, Formant Patterns and Phonation Types

E5.1 Whispered Sounds Compared with Voiced Sounds at Different F0 in Utterances of a Single Speaker

To do: (i) Select a speaker with good vocal abilities. (ii) Let the speaker produce isolated whispered sounds of the long vowels of his language. (iii) Then, let the speaker produce voiced sounds of these vowels at different levels of F0. (You may refer to a musical scale). Investigate a range of F0 up to 523 Hz in minimum. (You may refer to utterances of a woman.) Pay attention to the close correspondence of the produced vowel qualities and vowel colours. (iv) Perform spectral analysis and formant analysis. (v) Perform resynthesis, according to the following conditions: for a given formant pattern of a single sound produced, as source characteristic, apply all F0 investigated as well as noise. (vi) Perform a listening test.

Thesis: (i) You will find unsystematic results (see above). (ii) When comparing whispered sounds with voiced sounds at lower F0, in many cases, you will find indications of higher spectral peaks ≤ 1.5 kHz and higher frequencies of calculated F1 and F2 for the former than the latter, as is indicated in formant statistics for citation-form words. (iii) However, you will also find many cases in which such differences decrease or even disappear if the F0 of the voiced vowel sound is raised. (iv) In parallel, often, no change in vowel perception will be found for a resynthesis using formant patterns of whispered sounds but higher F0 of voiced sounds. (v) In parallel, as metioned above, a change in vowel perception will often be found for resynthesising formant patterns of voiced sounds with regard to all F0 investigated.

E5.2 Whispered Sounds Compared with Voiced Sounds at Different F0 in Utterances of Speakers of Different Speaker Groups

To do: Redo the previous experiment for three speakers, a child, a woman and a man. (You may refer to the three speakers and the sounds of experiment E4.1.)

Thesis: In addition to the results predicted for experiment E4.1, you can question the so-called speaker group differences. Above all, for a given vowel, you may find correspondences of formant patterns of a voiced sound of a child when compared to a whispered sound of an adult, and vice versa.

E5.3 Sounds of Back Vowels Showing Three Spectral Peaks ≤ 1.5 kHz

To do: (i) Search for examples of sounds of back vowels, produced as whispered sounds in isolation, which show three spectral peaks ≤ 1.5 kHz. Also search for correspondingly produced examples that only show two peaks ≤ 1.5 kHz. (ii) Perform a listening test.

Thesis: You will find examples of the first kind for which the identification score is as high as for the examples of the second kind.

E5.4 Sounds of Front Vowels Showing Two Spectral Peaks ≤ 1.5 kHz

To do: (i) Search for examples of sounds of front vowels, produced as whispered sounds in isolation, which show two spectral peaks ≤ 1.5 kHz. Also search for correspondingly produced examples that show only one peak ≤ 1.5 kHz. (ii) Perform a listening test.

Thesis: You will find examples of the first kind for which the identification score is as high as for the examples of the second kind.

E6 Patterns of Relative Spectral Energy Maxima, Formant Patterns and Vowel Imitation by Birds

E6.1 Direct Comparisons of Selected Sounds of Humans and Birds

To do: (i) Create a sample of imitated words produced by birds, for example common hill myna birds. (ii) Select the best examples with regard to intelligibility. (iii) Isolate the sound nuclei corresponding to a vowel. (iv) Perform a listening test. (v) Select the sounds with a high score of consistent vowel perception. (vi) Let a woman and a man in turn imitate the "words" of the birds at the corresponding F0, and isolate the vowel sound nuclei. (vii) Perform a second listening test for all the sounds compared. (viii) Perform spectral analysis and formant pattern analysis. Concerning the sounds of the birds, even if methodologically unsubstantiated, you may apply both standard parameter settings for females and for males.

Thesis: (i) You will be able to observe examples in which a bird can produce a sound with a formant pattern F1–F2–F3 that corresponds to the formant pattern of a woman or a man. (ii) You will also be able to observe examples for which the sound of a bird does exhibit only a part of the formant patterns produced by the woman or man, yet vowel perception is not impaired.

E6.2 Resynthesis Relating to "Anomalous" Formant Patterns of Sounds of Birds

To do: (i) Select the sounds of the birds of the previous experiment with intelligible vowel quality but only partial correspondence of the formant patterns compared with the sounds of the man or the woman. (ii) Perform resynthesis. (iii) Perform a listening test.

Thesis: You will be able to resynthesise these sounds related to "anomalous" formant patterns with no substantial change in perceived vowel quality compared with the natural sounds.

E7 Anomalous Vowel Spectra

E7.1 Spectra with Increasing Number of Harmonics Equal in Amplitude ("Flat" Vowel Spectra)

To do: (i) Perform vowel synthesis using a harmonic synthesiser, that is, create harmonic spectra, perform inverse Fourier analysis and repeat the periods obtained over time for a certain duration, for example 1 s. (ii) Investigate sounds at F0 of 110 Hz and 220 Hz separately. (iii) Start a synthesis with only the first harmonic or fundamental at a given F0. Then continue to add, step by step, harmonics 2, 3, 4, etc. equal in amplitude to the fundamental. (iv) Perform a listening test.

Option: You may also investigate F0 other than the two frequency levels mentioned.

Thesis: You will find some sounds in the sound series created for which the listening test gives a vowel identification of one of the vowels /u/, /o/, /ɔ/ and /a/. Eventually, /ɛ/ is also perceived.

Extension: You may extend the investigation to front vowels concerning "flat" spectral parts > c. 2 kHz. (Try also "flat" spectral parts > c. 1.5 kHz.) You may then start with a series of lower harmonics as found in natural vocalisations and add, step by step, harmonics equal in amplitude from c. 2 kHz (or from c. 1.5 kHz) upwards.

E7.2 Spectra with Increasing Number of Harmonic Pairs Showing Equal Amplitude Differences ("Ridged" Parts of Vowel Spectra)

To do: Apply the same procedure as described in the previous experiment but add, step-by-step, harmonics with periodic increasing and decreasing amplitudes; for example L2 (level of second amplitude) < L1 (level of first amplitude), L3 = L1, L4 = L2, and so on; or vice versa.

Thesis: (i) You will obtain results depending on the extent of the difference in the harmonic level you have set. (ii) However, within a limited range of such a difference, the listening test will provide similar results to those predicted for the previous experiment.

Extension: You may again extend the investigation to front vowels concerning "ridged" spectral parts > 2 kHz.

E8 Aspects of Method

E8.1 Formant Pattern Estimation Related to Non-Standard Parameters

To do: (i) Refer to a large sample of isolated vowel sounds. (ii) Perform LPC analysis applying standard parameters. (iii) Select the sounds for which the calculated formant patterns clearly do not correspond to what is expected if referred to formant statistics. However, do not try to include very high F0. (iv) Perform a listening test. (v) Select only sounds with a high score of identification. (vi) Perform LPC analysis again but alter the parameters.

Option: You may also perform resynthesis.

Thesis: (i) You will find various examples for which LPC analysis based on non-standard parameters as given in the literature—above all based on a non-standard maximum number of formants for a given frequency range, which is usually related to age and gender of the speaker—provides “better” (that is, more “expected”) results than LPC analysis based on standard parameters. (ii) However, you will not be able to relate this finding to a general production characteristic for all vowel sounds produced by a single speaker.

E8.2 Formant Pattern Estimation at F0 > 350 Hz

To do: (i) Select isolated vowel sounds produced at F0 > 350 Hz. (ii) Perform a listening test. (iii) Select sounds with a high identification score. (iv) Perform LPC analysis using standard parameters. (v) On the basis of the corresponding results, perform resynthesis. (vi) Perform a listening test related to the resynthesised sounds.

Thesis: (i) You will find variable results. (ii) However, you will find many examples for which the natural and the resynthesised sound is perceived as the same vowel, although the LPC analysis is not methodologically substantiated and the calculated formant pattern may differ strongly from values given in formant statistics.

E8.3 Resynthesis of Sounds at Varying F0 and Subsequent Formant Pattern Estimation

To do: (i) Select isolated natural vowel sounds. (ii) Perform a listening test. (iii) Select only sounds with a high identification score. (iv) Perform LPC analysis using standard parameters. (v) On the basis of the corresponding results, perform resynthesis for two conditions; first, use F0 of the natural vocalisation; second, use a very different F0 level. (vi) Perform a listening test again and select again only sounds with a high identification score. (7) Perform LPC analysis for both types of the resynthesised sounds.

Thesis: (i) You will find variable results. (ii) However, you will find many examples for which the calculated formant pattern of the resynthesised sounds differs substantially from the original formant pattern, if the F0 of the resynthesised and the natural sounds also differs substantially.

List of Figures
List of Tables
References

Because of the high number of figures shown in the Materials section, the lists of figures and tables are presented at the end of the text, followed by the References section.

List of Figures

M8.2 Vowel Perception at Fundamental Frequencies above Statistical Values of the Respective First Formant Frequency

M8.3 “Inversions” of Relative Spectral Energy Maxima and Minima and “Inverse” Formant Patterns in Sounds of Individual Vowels

M9 Ambiguous Correspondence between Vowels and Patterns of Relative Spectral Energy Maxima or Formant Patterns or Complete Spectral Envelopes

M9.1 Ambiguous Patterns of Relative Spectral Energy Maxima and Ambiguous Formant Patterns

Figures 1 to 3. Comparisons of sounds of back vowels and of /a–ɑ/ produced by different speakers, and related model patterns of spectral peaks and/or of calculated formant frequencies.

Figures 4 to 8. Comparisons of sounds of back vowels and of /a–ɑ/ produced by single speakers, and related model patterns of spectral peaks and/or of calculated formant frequencies.

Figures 9 to 18. Comparisons of sounds of front vowels produced by different speakers, and related model patterns of spectral peaks and/or of calculated formant frequencies.

M9.3 Ambiguity and Individual Vowels

M10 Lack of Correspondence between Patterns of Relative Spectral Energy Maxima or Formant Patterns and Age- and Gender-Related Speaker Groups or Vocal-Tract Sizes

M10.1 Similar Patterns of Relative Spectral Maxima and Similar Formant Patterns ≤ 1.5 kHz for Different Age- and Gender-Related Speaker Groups or Vocal-Tract Sizes

M10.A Addition: Vowel Imitations by Birds

List of Tables

2 Prevailing Empirical References

2.1 General References

2.2 Empirical Reference for Standard German

References

Boersma, P., & Weenink, D. (2015). *Praat: doing phonetics by computer* [Computer program]. Version 5.4.08. Retrieved March 30, 2015, from http://www.praat.org.

Delattre, P. (1980). Vowel color and voice quality. In J. Large (Ed.)*, Contributions of Voice Research to Singing* (pp. 373 –384). Houston, TX: College Hill Press. (Reprinted from *The Bulletin of the National Association of Teachers of Singing,1958, XV*, 4–7.)

Diehl, R. L., Lindblom, B., Hoemeke, K. A., & Fahey, R. P. (1996). On explaining certain male-female differences in the phonetic realization of vowel categories. *Journal of Phonetics, 24*(2), 187–208.

Fant, G. (1959). Acoustic analysis and synthesis of speech with applications to Swedish. *Ericsson Technics, 1*, 1–106.

Fant, G. (1960). *Acoustic theory of speech production.* The Hague: Mouton.

Fant, G., Carlson, R., & Granström, B. (1974). The [e]-[ø] ambiguity. In *Speech Communication Seminar* (pp. 117-121).

Fant, G., Henningsson, G., & Stalhammar, U. (1969). Formant frequencies of Swedish vowels. *Speech Transmission Laboratory Quarterly Progress and Status Report, 10*(4), 26–31.

Friedrichs, D., Maurer, D., & Dellwo, V. (2015). The phonological function of vowels is maintained at fundamental frequencies up to 880 Hz. *The Journal of the Acoustical Society of America, 138*(1), EL36–EL42.

Friedrichs, D., Maurer, D., Suter, H., & Dellwo, V. (2015). Vowel identification at high fundamental frequencies in minimal pairs. In *Proceedings of the 18th International Congress of Phonetic Sciences* (no. 0434, pp. 1–4).

Fulop, S. A. (2011). *Speech spectrum analysis*. Berlin: Springer Science & Business Media.

Gelfer, M. P., & Bennett, Q. E. (2013). Speaking fundamental frequency and vowel formant frequencies: Effects on perception of gender. *Journal of Voice, 27*(5), 556–566.

von Helmholtz, H. L. F. (1954). *On the sensations of tone.* New York, NY: Dover. (Republication of the 2nd edition of the Ellis translation of *Die Lehre von den Tonempfindungen,* Longman & Co., 1885.)

Hillenbrand, J. (n.d.). *The physics of sound.* Retrieved October 1, 2015, from http://homepages.wmich.edu/~hillenbr/206/ac.pdf

Hillenbrand, J., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *The Journal of the Acoustical Society of America, 97*(5), 3099–3111.

Hollien, H., Mendes-Schwartz, A. P., & Nielsen, K. (2000). Perceptual confusions of high-pitched sung vowels. *Journal of Voice, 14*(2), 287–298.

Howie, J., & Delattre, P. (1962). An experimental study of the effect of pitch on the intelligibility of vowels. *The National Association of Teachers of Singing Bulletin, 18*(4), 6–9.

Iivonen, A. (1986). A set of German stressed monophthongs analyzed by RTA, FFT, and LPC. In R. Channon & L. Shockey (Eds.), *In honour of Ilse Lehiste* (pp. 125–138). Dordrecht: Foris.

Iivonen, A. (1970). *Experimente zur Erklärung der spektralen Variation deutscher Phonemrealisationen* (Commentationes Humanarum Litterarum, vol. 45). Helsinki: Societas Scientiarum Fennica.

Joliveau, E., Smith, J., & Wolfe, J. (2004). Vocal tract resonances in singing: The soprano voice. *The Journal of the Acoustical Society of America, 116*(4), 2434–2439.

Jørgensen, H. P. (1969). Die gespannten und ungespannten Vokale in der norddeutschen Hochsprache mit einer spezifischen Untersuchung der Struktur ihrer Formantfrequenzen. *Phonetica, 19*, 217–245.

Kent, R. D., & Read, C. (2002). *The acoustic analysis of speech* (2nd ed.). Clifton Park, NY: Delmar, Cengage Learning.

Ladefoged, P. (1996). *Elements of acoustic phonetics* (2nd ed.). Chicago: The University of Chicago Press.

Ladefoged, P. (2003). *Phonetic data analysis: An introduction to fieldwork and instrumental techniques.* Malden, MA: Wiley-Blackwell.

Maurer, D. (2013). *Akustik des Vokals – Präliminarien*. subTexte 08, A. Rey (Ed.). Zurich: Institute for the Performing Arts and Film, Zurich University of the Arts.

Maurer, D. (n.d.). *Acoustic characteristics of voice in music and straight theatre – towards a systematic empirical foundation. Project description.* Retrieved October 1, 2015, from http://www.phones-and-phonemes.org/project-1.html.

Maurer, D., Cook, N., Landis, T., & d'Heureuse, C. (1991). Are measured differences between the formants of men, women and children due to F0 differences? *Journal of the International Phonetic Association, 21*(2), 66–79.

Maurer, D., & Landis, T. (1995). F0-dependence, number alteration, and non-systematic behaviour of the formants in German vowels. *International Journal of Neuroscience, 83*(1–2), 25–44.

Maurer, D., & Landis, T. (1996). Intelligibility and spectral differences in high-pitched vowels. *Folia Phoniatrica et Logopaedica, 48*(1), 1–10.

Maurer, D., & Landis, T. (2000). Formant pattern ambiguity of vowel sounds. *International Journal of Neuroscience, 100*(1–4), 39–76.

Maurer, D., Landis, T., & d'Heureuse, C. (1991). Formant movement and formant number alteration with rising F0 in real vocalisations of the German vowels [u:], [o:] and [a:]. *International Journal of Neuroscience, 57*(1–2), 25–38.

Maurer, D., Mok, P., Friedrichs, D., & Dellwo, V. (2014). Intelligibility of high-pitched vowel sounds in the singing and speaking of a female Cantonese Opera singer. In *Proceedings of the 15th Conference of the International Speech Communication Association, Interspeech 2014* (pp. 2132–2133). (For an extended version including additional material, see the related internet presentation online at http://is2014.phones-and-phonemes.org, retrieved October 1, 2015.)

Maurer, D., Suter, H., Friedrichs, D., & Dellwo, V. (2015). Acoustic characteristics of voice in music and straight theatre: topics, conceptions, questions. In A. Leemann, M-J. Kolly, S. Schmid, & V. Dellwo (Eds.), *Trends in Phonetics and Phonology. Studies from German-speaking Europe* (pp. 256–265). Bern/Frankfurt: Peter Lang.

Moore, G. D. (2006). *The physics and psychophysics of music* (course page for Physics 224, lecture 28, p. 11). Retrieved November 1, 2015, from http://www.physics.mcgill.ca/~guymoore/ph224/notes/lecture28.pdf.

van Nierop, D. J. P. J., Pols, L. C. W., & Plomp, R. (1973). Frequency analysis of Dutch vowels from 25 female speakers. *Acta Acustica united with Acustica, 29*(2), 110–118.

Pätzold, M., & Simpson, A. (1997). Acoustic analysis of German vowels in the Kiel Corpus of Read Speech. *Arbeitsberichte des Instituts für Phonetik und Digitale Sprachverarbeitung der Universität Kiel (AIPUK), 32*, 215–247.

Peterson, G. E., & Barney, H. L. (1952). Control methods used in a study of the vowels. *The Journal of the Acoustical Society of America, 24*(2), 175–184.

Pickett, J. M. (1999). *The acoustics of speech communication: fundamentals, speech perception theory, and technology.* Boston, MA: Allyn & Bacon.

Pols, L. C. W., Tromp, H. R. C., & Plomp, R. (1973). Frequency analysis of Dutch vowels from 50 male speakers. *The Journal of the Acoustical Society of America, 53*(4), 1093–1101.

Potter, R. K., & Steinberg, J. C. (1950). Toward the specification of speech. *The Journal of the Acoustical Society of America, 22*(6), 807–820.

Ramers, K. H. (1988). *Vokalquantität und -qualität im Deutschen*. Linguistische Arbeiten 213. Tübingen: Niemeyer.

Rausch, A. (1972). Untersuchungen zur Vokalartikulation im Deutschen. In H. Kelz & A. Rausch (Eds.), *Beiträge zur Phonetik* (IPK-Forschungsberichte, vol. 30, pp. 35–82). Hamburg: Buske.

Schroeder, M. R., & Strube, H. W. (1986). Flat-spectrum speech. *The Journal of the Acoustical Society of America, 79*(5), 1580–1583.

Sharifzadeh, H. R., McLoughlin, I. V., & Russell, M. J. (2012). A comprehensive vowel space for whispered speech. *Journal of Voice*, 26(2), e49–56.

Sundberg, J. (1978). Synthesis of singing. *Swedish Journal of Musicology, 60*(1), 107–112.

Sundberg, J. (1987). *The Science of the Singing Voice*. DeKalb, Ill.: Northern Illinois University Press.

Sundberg, J. (2013). Perception of singing. In D. Deutsch (Ed.), *The psychology of music* (3rd ed., pp. 69–105). San Diego, CA: Elsevier.

Swerdlin, Y., Smith, J., & Wolfe, J. (2010). The effect of whisper and creak vocal mechanisms on vocal tract resonances. *The Journal of the Acoustical Society of America, 127*(4), 2590–2598.

Trask, R. L. (1996). *A dictionary of phonetics and phonology*. New York, NY: Routledge.

Titze, I. R., Baken, R. J., Bozeman, K. W., Granqvist, S., Henrich, N., Herbst, C. T., … & Wolfe, J. (2015). Toward a consensus on symbolic notation of harmonics, resonances, and formants in vocalization. *The Journal of the Acoustical Society of America*, *137*(5), 3005–3007.

Traunmüller, H. (n.d.). *The role of F_0 in vowel perception*. Retrieved November 1, 2015, from http://www2.ling.su.se/staff/hartmut/i.htm.

Traunmüller, H., & Eriksson, A. (1997). A method of measuring formant frequencies at high fundamental frequencies. In *Proceedings of Eurospeech* (Vol. 97, No. 1, pp. 477-480).

Traunmüller, H., & Eriksson, A. (2000). Acoustic effects of variation in vocal effort by men, women, and children. *The Journal of the Acoustical Society of America*, *107*(6), 3438–3451.

Wängler, H.-H. (1981). *Atlas deutscher Sprachlaute*. Berlin: Akademie-Verlag.

Wolfe, J. (n.d.). *Formant: what is a formant?* Retrieved November 1, 2015, from http://www.phys.unsw.edu.au/jw/formant.html.

Wolfe, J., Garnier, M., & Smith, J. (2009). Vocal tract resonances in speech, singing, and playing musical instruments. *Human Frontier Science Program Journal, 3*(1), 6–23.

Wood, S. (1989). The precision of formant frequency measurement from spectrograms and by linear prediction. *Speech Transmission Laboratory Quarterly Progress and Status Report, 30*(1), 91–93.

Zee, E. (2003). Frequency analysis of the vowels in Cantonese from 50 male and 50 female speakers. In *Proceedings of the 15th International Congress of Phonetic Sciences (pp. 1117–1120).*