Dieter Maurer

Acoustics of the Vowel

Preliminaries

Peter Lang

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 rea-

sons 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.