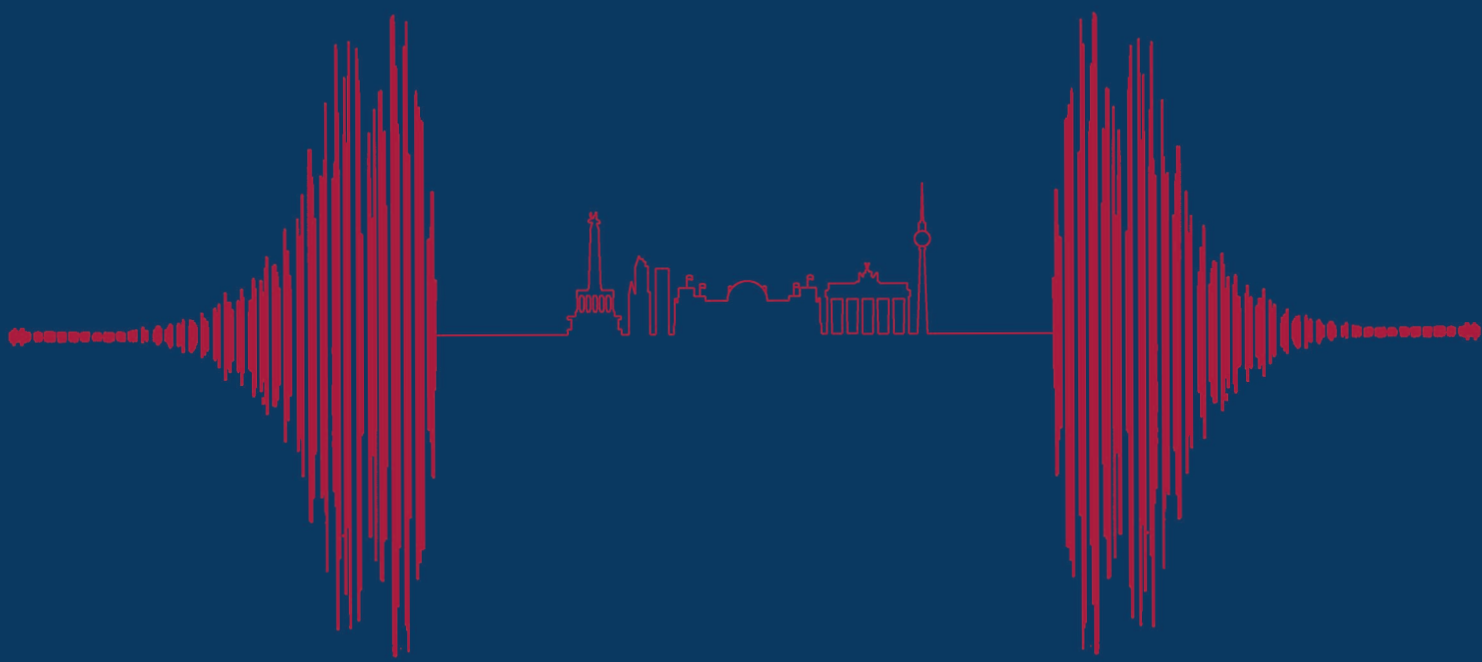
P&P 13

Proceedings of the Conference on Phonetics & Phonology in German-speaking countries



Leibniz-Zentrum
Allgemeine Sprachwissenschaft

HUMBOLDT-UNIVERSITÄT ZU BERLIN

Why a phenomenology of vowel sounds is needed

Dieter Maurer

Institute for the Performing Arts and Film, Zurich University of the Arts

dieter.maurer@zhdk.ch

Abstract

In the literature, there is an extensive and often controversial debate on the primary acoustic and perceptual cues of vowel quality, resulting in two main viewpoints that these cues are contained in either the formants, or, alternatively, in the spectral shape. However, in our understanding, one aspect is highly underestimated: the fact that any spectral representation of vowel-quality is directly or indirectly pitch-related. Hence, a given formant pattern as well as a given spectral envelope is in many – if not all – cases ambiguous in terms of representing sounds of different vowel qualities, if speakers equal in size and gender produce the sounds on very different pitches. Neither of the two above-mentioned viewpoints can account for this issue, however. – The present paper (i) summarises the ongoing debate, (ii) describes the empirical evidence for pitch-related spectral representation of vowel quality, (iii) concludes that existing approaches to determine the acoustic cues for vowel quality do not account for all recognisable vowel sounds, both conceptually and methodologically, and (iv) argues for the need of a phenomenology of the acoustics of vowels in terms of building up large-scale, language-specific sound descriptions, addressing all variations of production parameters and their possible extension relevant for perceived vowel quality.

Background

Phonetic summaries generally state that vowel sounds exhibit spectral peaks (termed formants) as the primary acoustic and perceptual cue for the perceived vowel quality, and that these peaks are the consequence of vowel-specific resonance characteristics of the vocal tract. However, different conceptual understandings of formants exist side by side, and there is an extensive and often controversial debate in the literature addressing topics that are considered either as aspects of methodology, or as additional cues, or as aspects that are difficult to understand in the framework of a formant concept. (For excellent overviews, see e.g. Harrington, 2012, Kiefte et al., 2013; concerning overviews and exemplary discussions of single aspects, a few references are given below, and an extended list is given online, see Maurer, n.d.).

Formant concept: "Unfortunately, the common definition between a formant and a resonance is yet to be established." (Titze et al., 2015) Above all, formants are understood in terms of either resonances of the vocal tract, or peaks of the spectral envelope, or filters resulting from an acoustic analysis and related to a corresponding algorithm.

Formant estimation: Up to now, no objective method of formant estimation exists, regardless of the algorithm applied: formant patterns are generally estimated by means of an interactive measurement procedure involving general phonetic knowledge and analytical skill of the examiner, context information (size and gender of the speaker), visual crosschecks of calculated values on the basis of the sound spectrum and spectrogram, sometimes related to changes of parameter settings and recalculation of the patterns, and manual interpolations of calculated formants. Further, formant estimation loses methodological substantiation with increasing fundamental frequency (f_o). Some scholars consider the critical f_o frequency level as corresponding to half of the first formant frequency (F_1) of a sound, others assume an f_o level in the F_1 region of the closed vowels, i.e. an f_o level of c. 350 Hz as representing that limit.

Formants and additional cues: The debate on additional cues that potentially affect the acoustics of vowel sounds and the perception of vowel quality, concerns different types of phonation, speaker characteristics (above all size and gender differences) and f_o, duration, vowel-inherent spectral change, context and transitions, formant amplitude, spectral contrast and spectral tilt, and auditory spectral averaging process.

Aspects difficult to understand in the framework of formants: Besides the lack of an objective method for formant estimation, the debate on aspects that are difficult to understand in the framework of a formant concept concerns, above all, the lack of evidence that the data reduction process, implied by this concept, corresponds to the auditory processing of speech sounds, as well as observed nonlinearities in the relation between shifts of formant frequencies and shifts in the perceived vowel quality, and the lack of evidence for a peak picking mechanism of perception as indicated by recognisable vowel sounds with suppressed single formants or flat spectra.

Formants versus spectral shape: Referring to Swanepoel et al. (2012), we conclude that the entire debate on the multitude of aspects mentioned and their often controversial appraisal still have left us with only two main viewpoints, that the major acoustic and perceptual cues are contained in either formant frequency patterns or, alternatively, in the spectral shape, all other aspects of minor or additional effect. Thereby, spectral shape is commonly understood as the envelope of the spectrum derived from some kind of smoothing operation.

Methodological limitations of spectral envelope estimation: With rising f_o, as is true for formant estimation, spectral smoothing becomes also problematic because of spectral undersampling and interrelated distortions. The problem is severe for $f_o \geq 300$ Hz.

Core problem: vowel quality-specific spectral representation is pitch-related

However, in our understanding of the matter, two aspects are highly underestimated: firstly, the fact that vowel quality-specific spectral representation is directly or indirectly pitch-related, and secondly that, as a consequence, a given formant pattern as well as a given spectral envelope is in many – if not all – cases ambiguous in terms of acoustically and perceptually representing sounds of different perceived vowel qualities, if the sounds are produced with equal vocal effort by speakers equal in size and gender on very different pitches. *Thus, pitch-related spectral representation of vowel quality as such cannot primarily be tied to speaker differences in size and gender or to paralinguistic variation.* – Details are given in the following paragraphs.

Formants and f_o: Most scholars conclude for a marginal or very limited effect of f_o on the vowel quality of sounds of speakers equal in size and gender (see Cheveigné & Kawahara, 1999, Barreda & Nearey, 2012). However, most of the studies related to this conclusion reported values for f_o variation below 300 Hz. Yet, the few studies which included higher f_o levels, concluded for a substantial effect of f_o on vowel recognition (Maurer & Landis, 2000). This finding was either interpreted as calling for some kind of intrinsic normalisation of f_o and formants, possibly also related to paralinguistic variation of vocal effort, or as an indication of pitch-related spectral representation of vowel quality, a perspective adopted here.

"Oversinging" F_1 as generally given in formant statistics: Pätzold and Simpson (1997) reported statistical F_1 for six of the eight long Standard German vowels /i-y-e-ø-o-u/ < 400 Hz for men, and < 450 Hz for women. Summarising studies on vowel recognition in Western classical singing, Sundberg (2013, pp. 86–88) concluded that recognition can be maintained for all vowels up to C5 (523 Hz). Studies on vowel sounds produced in other artistic styles or involving untrained speakers, however, showed even higher f_o limits for general vowel recognition up to f_o in the range of 660–1046 Hz (dependent on the conditions of vowel production and of the listening tests), and the corner vowels were found to be recognisable up to 1046 Hz (Friedrichs et al., 2017). Thus, at least for a substantial part of vowels of a language, they can be produced and recognised on f_o above statistical F_1 obtained for relaxed speech.

"Oversinging" the f_o frequency limit for formant and spectral envelope estimation: The finding that vowel sounds can generally be recognised at f_o of c. 600 Hz and even above indicates a discrepancy between perception and methods of acoustic analysis: vowels can be recognised at pitches for which no formant frequency and no spectral envelope estimation is methodologically substantiated; further, the assumption of a direct relation between "spectral undersampling" and degradation of vowel quality is also contradicted.

Significance of extensive f_o variation in vowel production and perception: There is a strong tendency in the phonetic literature to describe the acoustic characteristics of vowel sounds on f_o levels related to citation-form words and to relaxed speech, and to consider extensive f_o variation as a phenomenon of either size and age-differences of the speakers, or specific (strong) emotions, or shouting, or to singing. However, we assume that the

significance of f_o variation should be reflected on differently: (i) f_o ranges of speakers different in size and gender substantially overlap. (ii) There is no principally pitch-related difference of spoken and sung vowels and, in art, the transition between speaking and singing can be fluid. (iii) Western classical opera style cannot be regarded as providing a general reference for vowel production and recognition, because the style-specific need for vocal power and instrumental sound timbre is often superordinated to vowel differentiation. (iv) Roughly spoken, according to Hollien (1972) and his terminology, vocal expressions can be experienced up to f_o = c. 500 Hz for men and c. 800 Hz for women in modal register, and up to f_o = c. 800 Hz for men and even substantially above 1 kHz for women in loft or falsetto register. (v) Noteworthy, everyday speech with register changes and/or with strong emotional variations, strong vocal efforts (including shouting), as well as specific speaking styles (ethnolects, infant directed speech, speech in a large audience, artistic speaking and singing styles etc.) include extensive f_o variation.

Spectral representation of vowel quality is f_o or pitch-related: This all comes down to the conclusion that, concerning the acoustics and perception of (radiated) voiced sounds, spectral representation of vowel quality is directly f_o or pitch-related (for the difference, see below). For unvoiced sounds (whisper phonation), this relation is indirect in the sense, that their estimated formant patterns and spectral envelopes correspond to patterns and envelopes of only a part of voiced sounds of the same vowel quality, within a limited f_o range of the latter.

Ambiguity of formant patterns and spectral envelopes: If formant patterns and spectral envelopes for sounds with different f_o differ, then what is to be expected are sounds with quasi-identical formant patterns or even quasi-identical spectral envelopes which, however, represent different vowel qualities, the main acoustic difference being their level of f_o. This kind of ambiguity is indicated in several studies of vowel synthesis, from the very early studies onwards, and it is also demonstrated for the neural open tube resonance patterns. However, and most importantly, the ambiguity was also demonstrated for natural vocalisations, including sounds produced by speakers equal in size and gender or even by single speakers (Maurer & Landis, 2000).

f_o versus pitch: Because the two phenomena discussed here can also be observed in cases of a "missing fundamental", strictly speaking, we consider the phenomena as related to pitch perception. In most cases, however, both f_o and pitch are concerned.

Non-systematic relation between f_o/pitch and vowel quality-related sound spectrum: As discussed earlier (Maurer, 2016, p. 59 and pp. 158–169), the relation between f_o/pitch, spectral peaks and envelope of the sound (if methodologically substantiated), and vowel quality is not systematic. It varies according to f_o/pitch range and course of the spectral envelope in general, and according to frequencies, levels and harmonic resolution of the spectral peaks in particular, the peaks represented, e.g., in calculated values of formant frequencies, bandwidths and levels. However, roughly speaking, ambiguous spectral peaks and envelopes occur if f_o/pitch is varied substantially above c. 200 Hz, and they primarily concern sounds of closed and mid-open vowel qualities.

A lack of conception and methodology

The concept of formants as being the major acoustic and perceptual cue for vowel quality does not account for the phenomena related to pitch-dependent spectral representation of vowel quality: (i) According to this concept, f_o is considered as an aspect of phonation, i.e. an aspect of the source, and formants are considered as an aspect of articulation and vowel differentiation, i.e. an aspect of the filter. These two parameters are assumed as quasi-independent, and the finding of nonlinear dynamics in the source-filter relationship does not principally contradict this assumption. (ii) Concerning the method of formant estimation, as said, no methodological substantiation exists to estimate formant patterns for the entire f_o/pitch range of recognisable vowel sounds.

The formant concept represents a powerful model for the description and prediction of vowel quality-related acoustic characteristics of a part of vowel sounds, namely, sounds produced within limited ranges of certain production parameters (above all within certain f_o/pitch limits, but also with regard to other parameters such as phonation type and vocal effort), but it cannot account for vowel sounds independently of these parameters and the limits set by methods of formant estimation (Maurer, 2016).

The same holds true for a corresponding concept of spectral shape as being the major acoustic and perceptual cue for vowel quality:

the acoustic representation of vowel-quality as spectral envelope also comes with a pitch-constraint and its determination also loses methodological substantiation with rising f_o.

A phenomenology is needed

Against this background, we argue that there is no robust approach to determine and predict acoustic sound characteristics directly related to perceived vowel quality for all recognisable vowel sounds, neither conceptually nor methodologically. We further conclude that a phenomenological approach to the acoustics of vowel sounds is needed, with three major aims: (i) reaffirming a pitch-related spectral representation of vowel quality and providing corresponding evidence for different fields of the scientific community, (ii) demonstrating the extension of spectral variation of recognisable vowel sounds, and (iii) building up empirical references for future competing approaches to assess the major acoustic cues of vowel quality, generally valid for all recognisable vowel sounds.

General structure and form: A phenomenology of the acoustics of vowels is considered here in terms of large-scale sound descriptions related to a given language, which include and interrelate sounds of all size and gender-related speaker groups, and address all variations of production parameters and their possible relevance for vowel quality.

Vowels: Principally, all vowels of a language are subject of investigation. However, at first, long vowels may be brought into focus because of their duration and their often quasi-constant (steady-state) sound nucleus.

Production parameters to vary: The primary variable parameters required in building up a within-speaker subsample of sounds are phonation types, f_o including register change, vocal effort, and phoneme context (isolated sounds, CVCV or CVC, minimal pairs, read speech).

Artistic speech and singing styles are of high interest because of the vocal abilities of the artists and the expressed variation of production parameters, including style-specific aspects. Therefore, sounds of non-professionals and professionals, and untrained and trained speakers and singers have to be included in a phenomenology, and production style as well as the differentiation of speaking and singing (and corresponding subtypes of vocal expression) have to be added as parameters to vary.

Acknowledgement

This work was supported by the Swiss National Science Foundation SNSF Grant No. 100016_159350.

References

Barreda, S., & Nearey, T. M. (2012). The direct and indirect roles of fundamental frequency in vowel perception. *The Journal of the Acoustical Society of America*, *131*(1), 466–477.

Friedrichs, D., Maurer, D., Rosen, S., & Dellwo, V. (2017): Vowel recognition at fundamental frequencies up to 1kHz reveals point vowels as acoustic landmarks. *Journal of the Acoustical Society of America, 142*(2), 1025–1033.

Harrington, J. (2012). Acoustic phonetics. In W. J. Hardcastle, J. Laver, & F. E. Gibbon (Eds.), *The handbook of phonetic sciences* (pp. 81–129). (2nd ed.). Malden MA: Wiley-Blackwell

Hollien, H. (1974). On vocal registers. Communication Sciences Laboratory Quarterly Report, 10(1), 1–33.

Kiefte, M., Nearey, T. M., & Assmann, P. F. (2013). Vowel perception in normal speakers. *Handbook of vowels and vowel disorders* (pp. 160–185). New York NY: Psychology Press.

Maurer, D. (2016). *Acoustics of the Vowel – Preliminaries*. Bern: Peter Lang.

Maurer, D. (n.d.). References for the paper "Why a phenomenology of vowel sounds is needed". http://phones-and-phonemes.org/PuP/13.html. Accessed October 01, 2017.

Maurer, D., d'Heureuse, C., & Landis, T. (2000). Formant pattern ambiguity of vowel sounds. *International Journal of Neuroscience, 100* (1–4), 39–76.

Pätzold, M., & Simpson, A. P. (1997). Acoustic analysis of German vowels in the Kiel Corpus of Read Speech. *Arbeitsberichte des Instituts für Phonetik und digitale Sprachverarbeitung Universität Kiel*, *32*, 215–247.

Sundberg, J. (2013). Perception of singing. In D. Deutsch (Ed.), *The Psychology of music*. (3rd ed.). London: Academic Press.

Swanepoel, R., Oosthuizen, D. J., & Hanekom, J. J. (2012). The relative importance of spectral cues for vowel recognition in severe noise. *The Journal of the Acoustical Society of America, 132*(4), 2652–2662.

Titze, I. R., Baken, R. J., Bozeman, K. W., Granqvist, S., Henrich, N., Herbst, C. T., ... & Kreiman, 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.