ENHANCING THE OBJECTIVITY OF INTERACTIVE FORMANT ESTIMATION: INTRODUCING EUCLIDEAN DISTANCE MEASURE AND NUMERICAL CONDITIONS FOR NUMBERS AND FREQUENCY RANGES OF FORMANTS

Thayabaran Kathiresan[1], Dieter Maurer[2], Heidy Suter[2], Volker Dellwo[1]

[1] Phonetics Laboratory, University of Zurich, Switzerland
[2] Institute for the Performing Arts and Film, Zurich University of the Arts, Switzerland
thayabaran.kathiresan@uzh.ch

Abstract: Current formant measurement studies of vowel sounds generally use a Linear Predictive Coding (LPC) algorithm and rely on an interactive method of formant estimation which includes a comparison of measured formant tracks and characteristics of the spectrogram. Thereby, the selection of LPC parameters is based on the assumption that the number of poles for the analysis of a given frequency range is age- and gender-specific. However, when crosschecking measured formant tracks with the spectrogram, mismatches occur in a significant number of cases. In these cases, the investigators try to minimize these mismatches by modifying the number of poles of LPC. Such an interaction is based on phonetic knowledge, analytical experience and related expectations. Several authors have pointed towards the lack of objectivity and the inherent circularity as well as the fact that similar formant estimations performed by different researchers may yield different results. As of yet, the issue of an improvement and objectification of formant estimation procedure is still a matter of debate. The present paper describes such a corresponding approach: basing the LPC pole-number selection on objective criteria by introducing Euclidean distance measure and formant frequency conditions as references for interactive formant frequency estimation. The paper further presents and discusses the results of a pilot evaluation using the method proposed on 224 long Standard German vowel sounds /i-y-e-ø-ɛ-a-o-u/ produced by eight children, ten women and ten men on fundamental frequencies of 262 Hz (children), 220 Hz (women) and 131 Hz (men), respectively.

1 Introduction

In terms of the acoustic analysis of speech sounds, formants *F*(i) are understood as frequency ranges in which there are absolute or relative maxima in the sound spectrum, with formant frequencies $F_{(i)}$ as the frequencies of the maxima (see the current ASA standard of acoustic terminology [1]; for a discussion of the definition of the formant as well as for the form of abbreviation, see [2]). There has been a long history of methods for measuring formant frequencies from the speech signal. However, as of yet, formant patterns are generally estimated by means of an interactive measurement procedure (for a reference study, see [3]): Firstly, based on general phonetic knowledge and expectation, a number of poles corresponding to age and gender of the speaker are set for the LPC analysis. Secondly, the formant tracks resulting from the LPC analysis are visually crosschecked on the basis of the spectrogram (and sometimes also the spectrum) and are interpolated manually (by doing this, "gaps" in the tracks are also filled). Thereby, again on the basis of general phonetic knowledge, expectation and practice of acoustic analysis, the conditions for the number and the frequency ranges of the formants are also considered. However, in many cases – e.g. related to "formant merging", "spurious formants", weak or broad spectral peaks, low vocal effort, middle or high fundamental frequencies (f_o) – the automatically calculated formant tracks may not match the spectrogram. In these cases, an investigator manually changes the number of poles for LPC analy-

sis [4], crosschecks and interpolates the newly calculated formant tracks and may also introduce further manual formant frequency estimations that relate to the spectrogram or spectrum directly. In some studies, certain sounds are even excluded from analysis because no LPC pole number setting can be found which allows for a suitable calculation of formant tracks that can be approved by the investigators via the crosscheck mentioned. Many authors have pointed towards the lack of objectivity and the inherent circularity in this method (see e.g. [3, 5]). Moreover, a recent study showed substantial differences when comparing the results of formant analysis of four different investigators [6]. Thus, the issue of how the procedure of formant estimation can be improved and objectivised is still a matter of debate. Against this background, the present paper describes an approach and its first empirical evaluation which bases the selection of the LPC pole number on objective criteria by introducing Euclidean distance measure and formant frequency conditions as references for interactive formant frequency estimation. The empirical evaluation – formant analysis performed with the new method – consisted of an examination of 224 sounds of the long Standard German /i-y-e-ø-ɛ-a-o-u/ produced by eight children, ten women and ten men on fundamental frequencies of 262 Hz (children), 220 Hz (women) and 131 Hz (men), respectively. The significance of the described procedure and the first evaluation results as well as improvements to be addressed in future studies are discussed.

2. Methods

2.1 Speakers, recordings, listening test

The present study analyses recordings of twenty-eight German native speakers, twenty adults (gender balanced, age 18 to 52 years) and eight children (gender balanced, age 7 to 10) with no record of hearing or speaking disorders. The speakers were asked to produce isolated sounds of the vowels /i-y-e-ø-ɛ-a-o-u/ on f_o corresponding to age- and gender-related levels reported for utterances in citation-form words (see. e.g. [3, 7]) that is, f_o= 131 Hz for men, f_o= 220 Hz for women and f_o = 262 Hz for children (adapted to fit the C-major scale). As a reference, before each recording, the pitch was played back on a digital piano via loudspeaker, and the speaker was asked to produce the sounds of the eight vowels mentioned on that particular pitch in isolation, with medium vocal effort and with a duration of 1–2 sec. Concerning the pronunciation of /a/, the vowel quality of the isolated sounds produced by the speakers corresponds to a range in between /ɑ/ and /a/. The speakers were recorded in standing position in a noise-controlled room with a speaker–microphone distance of 30 cm. The sounds were digitally recorded on a PC (cardioid condenser microphone Sennheiser MKH 40 P48, pop shield, audio interface Fireface UCX). The sampling frequency of the recordings was 44.1 kHz. Five phonetic expert listeners (professionally trained singers and actors with no record of hearing disorder) identified all sounds investigated in a multiple-choice identification tasks. A correspondence of the vowel intention of the speaker's production and the recognition of a majority of the listeners was found for all sounds, with recognition rates = 5/5 (5 of 5 listeners) for 210 sounds, 4/5 for 12 sounds and 3/5 for 2 sounds.

2.2 Acoustic analysis

The acoustic analysis was performed on the middle 0.3 sec sound nucleus of each recorded vowel sound. As shown in Figure 1 (top), for this sound nucleus and a frequency range of 0–5.5 kHz, three formant frequency patterns F_1–F_2–F_3 were calculated in parallel using the LPC Burg algorithm in Praat, with pole numbers 10, 12 and 14. These three pole numbers are generally considered appropriate for formant estimation of vowel sounds produced by children, women and men, respectively, i.e. they represent standard age- and gender-related LPC settings. For each of the three parallel measures, formant tracks and average formant frequency values were calculated and processed by the algorithm described in the next section.

2.3 Processing the acoustic measures

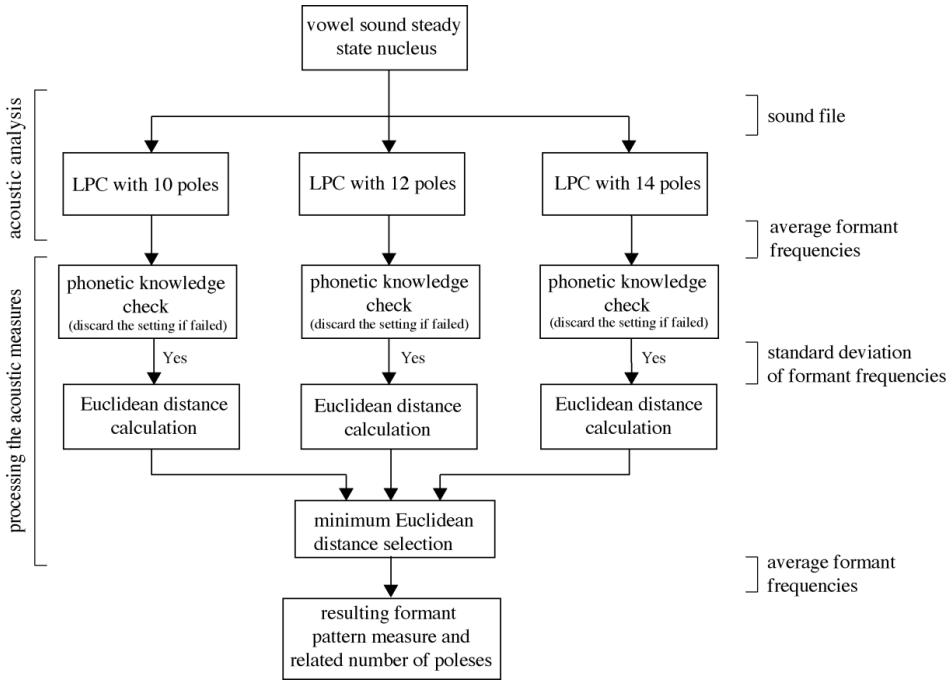

Figure 1 – Schematic flow diagram of automatic formant estimation

In the subsequent processing of the three parallel formant measures, firstly, the corresponding three average formant frequency patterns were passed on to the respective "Phonetic Knowledge" checking block. If, for a single measure, the formant frequencies were found within the age- and gender-related frequency ranges as shown in Table 1, the measure was processed further. The frequency ranges in the "Phonetic Knowledge" table were set according to the average formant frequencies for Standard German vowel sounds produced by women and men as given by [8]. These values were further adapted and generalised for all three speaker groups on the basis of the vowel spectra investigated.

Table 1 – "Phonetic Knowledge" table (all the values kHz)

Vowels	Children			Women			Men		
	F_1	F_2	F_3	F_1	F_2	F_3	F_1	F_2	F_3
i	< 1	2.5 – 3.5	3 – 5	< 1	2 – 3	2.5 – 4	< 1	1.9 – 3	2.4 – 4
y	< 1	1.7 – 2.4	2.2 – 4	< 1	1.6 – 2.3	2 – 3	< 1	1.5 – 2.3	2 – 2.8
e	< 1	2 – 3.5	3 – 5	< 1	2 – 3	2.5 – 4	< 1	1.7 – 3	2.4 – 4
ø	< 1	1.7 – 2.4	2.2 – 4	< 1	1.4 – 2.3	2 – 3	< 1	1.3 – 2.3	2 – 2.8
ɛ	< 1	2 – 3.5	2.5 – 5	< 1	1.6 – 2.6	2.5 – 3.5	< 1	1.5 – 2.6	2 – 3.5
a	< 1.6	< 2	2.5 – 4	< 1.5	< 2	2 – 3.5	< 1.3	< 2	2 – 3.5
o	< 1	<1.5	2.5 – 4	< 1	< 1.5	2 – 3.5	< 1	< 1.5	2 – 3.5
u	< 1	<1.5	2 – 4	< 1	< 1.5	2 – 3.5	< 1	< 1.5	1.9 – 3.5

Secondly, for the formant measures that concur with these age- and gender-related frequency conditions, the standard deviation (σ) from the formant tracks are used to calculate the respective Euclidean Distance (ED) according to the formula (1) (for illustration, see Figure 2):

$$\mathrm{ED} = \sqrt{[\sigma(F_1)]^2 + [\sigma(F_2)]^2 + [\sigma(F_3)]^2} \quad (1)$$

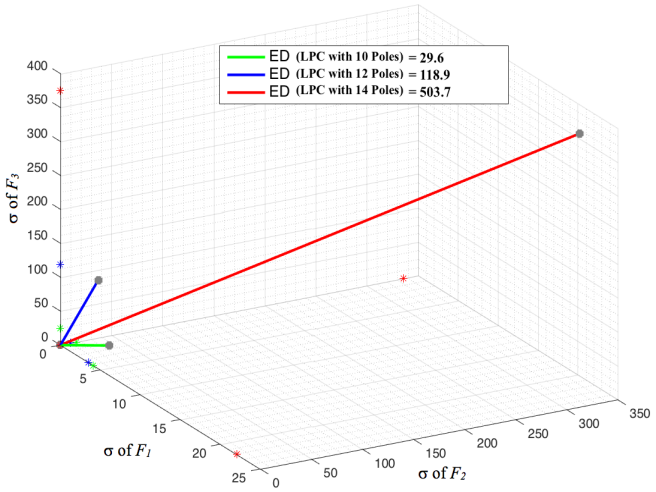

Figure 2 - ED projection of the first three formants in 3-dimensional space of the vowel /ɛ/ produced by a man

Then thirdly, comparing the ED values of the parallel measures, the measure with the minimum ED was chosen and the corresponding formant tracks and average formant values together with the related number of poles were allocated as the values to crosscheck on the basis of the spectrogram and the spectrum of the vowel sound in question (see next section). The same procedure was repeated for F_1 and F_2 values only, with ED calculated for the two lower formants.

2.4 Crosscheck

The results produced by the algorithm were crosschecked by the second author by comparing formant tracks with spectrogram, and LPC curve of the middle of the sound nucleus with average spectrum of the vowel sound in question (see Figure 3 for illustration). Thereby, the crosscheck adhered to rules proposed by expert phoneticians (see e.g. [3, 9, 10]), with the exception that no further interpolation of the formant tracks was performed. The crosscheck also related to extended experience of formant pattern estimations undertaken in previous studies (see [11]). If formant tracks and LPC curve corresponded with spectrogram and spectrum (with an estimated frequency range of ± 100 Hz for formant frequencies ≤ 1.5 kHz and of ± 150 Hz for formant frequencies > 1.5 kHz as a range of match between the numerical formant frequency value and the frequency of the related absolute or relative spectral maxima), the results of the algorithm were considered approved; else they were rejected. The crosscheck was performed in two separate runs, on the full F_1–F_2–F_3 pattern and on the reduced F_1–F_2 pattern.

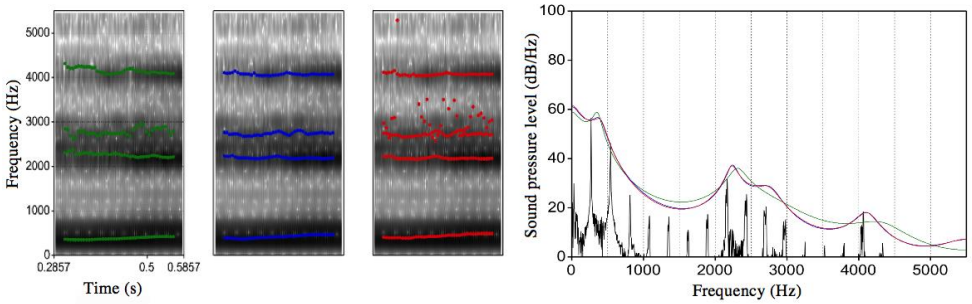

Figure 3 - Graphics presented to a phonetician to crosscheck the algorithm's formant selection: formant tracks and spectrogram (left), LPC curve (middle analysis window of the sounds nucleus) and spectrum (right). Results of LPC are shown for three different pole settings, i.e. 10 (green), 12 (blue) and 14 (red). In the example, the algorithm selected the formant pattern related to a pole-number setting = 12.

3. Results

As shown in Table 2, from a total of 224 sounds investigated, the measures selected by the algorithm were approved in the crosscheck for 185 sounds (83%) concerning F_1–F_2–F_3 and for 208 sounds (93%) concerning F_1–F_2. The investigator rejected the measures provided by the algorithm for 14 sounds (6%) because of lacking correspondence with spectrogram and spectrum, and for 2 sounds (1%), the values failed to fulfil the processing conditions entirely.

Table 2 - General performance of the algorithm after crosschecking the vowel spectrogram and spectrum

Sounds investiagted	224	100%
Approved for F_1–F_2–F_3	185	83%
Approved for F_1–F_2	208	93%
Disapproved (Investigator)	14	6%
Data processing failed	2	1%

Out of all the algorithm measures approved in the crosscheck, 52%–63% corresponded to the LPC pole settings generally assumed to be age- and gender-related (see Figure 4).

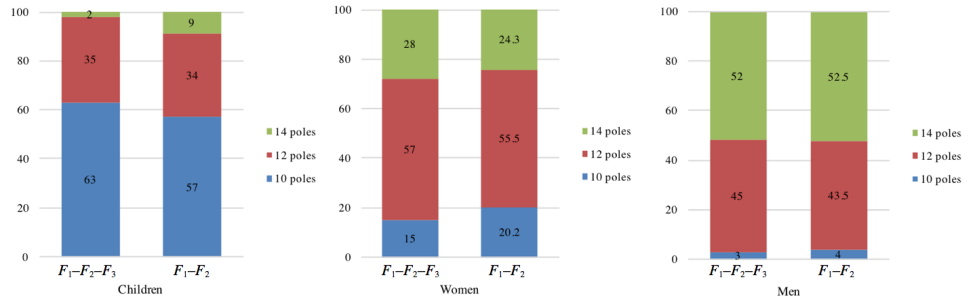

Figure 4 - Distribution of pole-number settings for the approved F_1–F_2–F_3 and F_1–F_2 patterns (all the values in % of the approved formant measures)

Thereby, some differences in the distribution of the selected number of poles are found to relate to speaker groups, front and back vowels, and higher/lower pole numbers compared with the expected standard setting in question (see Table 3). Figure 4 illustrates the resulting distribution of F_1–F_2 values for each of the three speaker groups.

Table 3 - Results of the crosscheck of the formant measures selected by the algorithm, given separately for front vowels, and back vowels and /a/

		front vowel formants		back vowel formants and /a/		total	
speaker group	number of poles	F_1–F_2–F_3	F_1–F_2	F_1–F_2–F_3	F_1–F_2	F_1–F_2–F_3	F_1–F_2
children	10	19	–	13	1	32	1
	12	14	–	4	2	18	2
	14	1	–	0	4	1	4
women	10	9	3	1	2	10	5
	12	21	1	16	3	37	4
	14	8	–	8	–	16	–
men	10	2	1	0	–	2	1
	12	20	2	12	–	32	2
	14	23	1	14	3	37	4
total		**117**	**8**	**68**	**15**	**185**	**23**

4. Discussion

The present approach does not aim at resolving the methodological problem of formant pattern estimation for all vowel sounds. It only represents an attempt to objectivise formant estimation for a frequency range of f_o for which formant estimation is not in general critical, that is for $f_o \leq 350$ Hz. Thereby, the main intention is to obtain an improved methodological basis for a re-examination of existing concepts of the relationship between perceived vowel quality and vowel-related formant patterns, and related empirical findings, as reported in the literature. Accordingly, the objective is to limit the visual and manual interaction of the investigators to only approving or rejecting calculated and processed formant measures but not allowing for any modification of numerical values. Such limitations will render it possible to directly compare the results of formant pattern estimations of different investigators and to differentiate between sounds with unanimously approved formant patterns when crosschecking the vowel spectrogram and spectrum, sounds with divided approval and sounds with disapproval. Given such a perspective, in the present approach, a rule-based processing of numerical formant measures of LPC is proposed that suggests that the LPC pole-number setting should be based on objective criteria by introducing Euclidean distance measure and formulating "phonetic knowledge expectations" in terms of numerical age- and gender-related formant frequency conditions.

The first test of the presented algorithm on the vowel sample described above provides promising results, since vowel related formant patterns F_1–F_2–F_3 could be assessed for 83% sounds of the sample, and F_1–F_2 patterns for 93% sounds, without modifying numerical formant values (no interpolation), without manual changes of LPC pole-number settings, and according to numerical formant frequency conditions.

At the same time, based on the results of this test, some important further conclusions can be made. Firstly, although assumed age- and gender-related standard LPC pole-number settings were found for the majority of the sounds of a given speaker group, deviations from the standards are found for more than 30% of the sounds. By far no marginal phenomenon, the present algorithm offers a methodological framework to account for these deviations and, at

the same time, to objectivise the pole-number selection. Secondly, estimation of formant frequencies for the entire formant pattern proved to be critical for 7% of the sound sample investigated, and for F_3 for 10% of the sounds (critical F_3 measures were observed above all for sounds of back vowels). Thus, formant statistics cannot be based on an entire sample of recordings, and the algorithm presented formulates corresponding rules for the reduction of a sample. Thirdly, in the crosscheck, we found that there were a non-negligible number of sounds for which some investigators may approve the measures provided by the algorithm but others may not. With rare exceptions, this concerned only the correspondence of F_3 with peaks in the spectrogram or spectrum. (Concerning the present sample, we estimate the corresponding number of sounds as c. 10% of the sample.) Accordingly, F_1–F_2 estimations are indicated to be more reliable than F_1–F_2–F_3 estimations, that is, investigator's estimation consensus will be higher for F_1–F_2 than for F_1–F_2–F_3. Since F_3 is often discussed as either being an acoustic cue for the distinction of front vowels or a cue for individual speakers or speaker groups, the lowered reliability of F_3 estimation has to be taken into account when addressing these two issues. Finally, the entire formant estimation routine presented here is by far less time consuming than routines including the manual modifications mentioned above. This is an important advantage for acoustic analysis of a large sample of vowel sounds as well as for the comparison of formant pattern estimation of different investigators.

In future studies, the robustness of the present approach should be tested on vowel sounds produced in syllable or word context (short vowel nucleus), and on vowel sounds produced with very low or very high vocal effort (effect of vocal effort on spectral peaks). Also, further improvements in processing the acoustic formant measures may be achieved by disregarding formants with large formant bandwidths and by excluding single analysis windows showing isolated gaps in the formant tracks. Finally, the rules how to compare numerical formant values and vowel spectrum and spectrogram may also be formulated more explicitly.

5. Acknowledgement

This study was supported by the Swiss National Science Foundation (SNSF), Grant No. 100016_159350/1.

6. References

[1] ANSI (2004). ANSI S1.1-1994, American National Standard Acoustical Terminology (*J. Acoust. Soc. Am*, Melville, NY

[2] Titze, I. R., R. J. Baken, K. W. Bozeman, et al, "Toward a consensus on symbolic notation of harmonics, resonances, and formants in vocalization," *J. Acoust. Soc. Am.* 137, 3005–3007. 2015

[3] HILLENBRAND, J., L. A. GETTY, M. J. CLARK, AND K. WHEELER, "Acoustic characteristics of American English vowels," *J. Acoust. Soc. Am.*, vol. 97, no. 5, pp. 3099–3111, 1995.

[4] VALLABHA, G. K., AND T. BETTY, *"*Systematic errors in the formant analysis of steady-state vowels*"*, *Speech Communication*, Volume 38, Issues 1–2, September 2002, Pages 141-160, ISSN 0167-6393

[5] LADEFOGED. P.: *Three Areas of experimental Phonetics* (Oxford U.P., London). 1967.

[6] SHADLE, C. H., H. NAM, AND D. H. WHALEN, "Comparing Measurement Errors for Formants in Synthetic and Natural Vowels," *J. Acoust. Soc. Am.*, vol. 9, no. August 2014, pp. 713–727, 2015.

[7] PETERSON, G. E., AND H. L. BARNEY, "Control methods used in a study of the vowels," *J. Acoust. Soc. Am.* 24(2) 1952
[8] PÄTZOLD, M., AND A. SIMPSON, "Acoustic analysis of German vowels in the Kiel Corpus of Read Speech," *Arbeitsberichte des Instituts für Phonetik und Digit. Sprachverarbeitung Univ. Kiel*, vol. 32, no. 1978, pp. 215–247, 1997
[9] LADEFOGED, P.: *Elements of acoustic phonetics* (2nd ed.). Chicago: *The University of Chicago Press*, 1996.
[10] LADEFOGED, P.: *Phonetic data analysis: An introduction to fieldwork and instrumental techniques*. Malden, MA: *Wiley-Blackwell*, 2003.
[11] MAURER, D.: "*Acoustics of the Vowel – Preliminaries*," *Peter Lang Verlag*, Bern/Frankfurt a.m., 2016.