Formant Contours in Czech Vowels: Speaker-discriminating Potential

Dita Fejlová, David Lukeš, Radek Skarnitzl

Institute of Phonetics, Faculty of Arts, Charles University in Prague, Czech Republic

dita.fejlova@gmail.com, dafydd.lukes@gmail.com, radek.skarnitzl@ff.cuni.cz

Abstract

The usefulness of dynamic formant properties for speaker discrimination was demonstrated on English, mostly exploiting long vowels or diphthongs, characterized both by actual measures along the formant tracks and coefficients from polynomial regression on these tracks. This study applies this paradigm to Czech, using less tightly controlled (more forensically realistic) material and taking into account the specific properties of the Czech vocalic system in which long vowels are much rarer than short ones. When all vowels are pooled together, the best results are achieved for unstressed vowels in asymmetrical CVC contexts. When individual vowels are considered separately, classification rates are best for long [i:] and [a:], but, most importantly, short vowels also show promising results. The performance of actual formant values and regression coefficients as predictors in discriminant analysis appears comparable.

Index Terms: vowel, formant contour, speaker discrimination, Czech

1. Introduction

In the current best practice of forensic speaker identification, auditory and acoustic analysis should complement each other [1], [2]. While the first type is more focused on linguistic-phonetic information (voice quality, long-term articulatory settings, dialect) [2], it also brings the necessary human judgement into the second type of analysis, which quantifies certain phonetic features, some of which may be less readily accessible to the human auditory system. It appears that at this point, acoustic analyses cannot yet be carried out, at least not in forensic casework, relying fully on automatic computing [2], using methods like GMM-UBM (see [3] for a review).

The chief goal of forensic phonetics is to determine which acoustic properties of speech carry some idiosyncratic information. In the history of speaker identification research, phoneticians have mostly focused on the fundamental frequency, vowel formants, or long-term average spectra (LTAS). Formants, which are the focus of this study, are the main spectral resonances of sonorant speech sounds such as vowels, which can be easily detected both visually in the spectrogram and automatically, using freely available software (e.g., [4]). Vowels as linguistic units are described mainly by their first and second formants (F1 and F2), while higher formants appear to convey more idiosyncratic information.

Vowel formants can be analyzed in numerous ways. History has seen three main types of analysis: In the first, static formant values are extracted from the middle portion of a vowel, where the vowel is considered to have reached its target [1], [5], [6]. These measures, typically taken from the first three formants, can then be tested in discriminant analysis to prove their speaker-distinguishing potential. The second type of analysis consists in calculating long-term formant distributions (LTFs): formant values are extracted, typically every 10 ms, from all sonorant sounds in a recording, and a histogram of each formant is plotted across all target segments [1], [7], [8]. This time-free analysis is instrumental especially in revealing a speaker's global habits, such as a tendency to lip-rounding or palatalization [1]. The third type also involves extracting more than one value from each formant, thus providing information about its dynamic properties as they unfold in time [9], [10], [11], [12]. This follows the reasoning by [13] that dynamic properties of speech may convey more speaker-specific information than static acoustic measures extracted from phonetic targets.

One of the indisputable advantages of analyzing formant dynamics is that it captures the individual specificities not only of reaching a target sound, but also of the transition to neighbouring sounds. The resulting trajectories of speech formants are derived from an interplay between individual vocal anatomies and the speaker-specific articulatory strategies for realizing and integrating the target speech sounds [1]. Therefore, this approach could make vowels an even more valuable source of idiosyncratic speaker information.

McDougall, the greatest proponent of investigating formant dynamics for speaker identification purposes, analyzed several sequences of speech sounds. In one of her first studies, she examined the diphthong [aɪ] followed by [k] [14]. She extracted values of F1–F3 at 10 equidistant intervals throughout the diphthong and the subsequent discriminant analysis yielded classification rates of about 90 %. In another study [12], she focused on [aɪk] and [əˈɹV] sequences and fitted the formant contours with polynomial functions using linear regression. She then compared the effectiveness of using the actual formant values and the polynomial coefficients as predictors for LDA. The related reduction of predictors led only to a minor decrease in classification rate, suggesting that this method is a promising way to capture the dynamic properties of vowel formants. Similar results were obtained by [15] for the vowel [uː] in the word *who'd*.

The aim of the present study is to assess the extent to which similar approaches to the analysis of formant trajectories, as opposed to static mean values of formants, are useful for discriminating between speakers of the Czech language. Our motivation consists in the fact that the Czech vowel system differs markedly from that of English, not only in its inventory but also in aspects which may have crucial implications for the methodological background for speaker identification in Czech forensic settings.

The distribution of Czech monophthongs is largely symmetrical: the system consists of five short and five long monophthongs, /ɪ iː ɛ ɛː a aː o oː u uː/, and the quality of each short/long pair is very similar with the exception of /ɪ/ and /iː/ (and possibly /u/ and /uː/); see [16] for a more detailed discussion. More importantly, the long vowels are considerably less frequent than the short vowels, as are the diphthongs [17]. Therefore, while McDougall's studies in English were based on analyses of diphthongs and inherently dynamic sounds like [ɹ], any analysis which is to be ultimately applicable in the Czech forensic context must rely predominantly on short vowels. It is to be expected that the dynamic properties of formants extracted from short monophthongs will convey a lower but acceptable degree of speaker-specific information. In addition, we will not restrain ourselves to uniform consonantal contexts in which the target vowels are located, but include vowels embedded in different segmental and prosodic contexts (*cf.* the discussion in [18]), so as to approach forensically realistic speech material.

2. Method

2.1. Material & subjects

The material for this study was recorded in a sound-treated booth at 32-kHz sampling frequency with 16-bit quantization, using an AKG C4500 condenser microphone. The subjects were 12 adult male native speakers of Czech, aged 22–28, who were asked to read 35 sentences in a natural way after sufficient preparation. Since some long vowels of Czech are quite infrequent (see above), the following monophthongs were analyzed in this study: /iː, ɪ, ɛ, aː, a, o, u/. We examined a total of 1,123 target vowels, each of them embedded in a CVC sequence.

It is obvious that formant trajectories will differ in their initial and final portions depending on the consonants flanking the vowel. When the consonantal context is symmetrical, we may expect less leeway in the individual realization; conversely, the individual strategies associated with the transitions from one sound to another may be more exposed in asymmetrical contexts. That is why the vowels were divided into two groups, those appearing in symmetrical contexts like /pVp/, /dVs/, /gVk/ (n = 483), and those in asymmetrical contexts where the flanking consonants differed in the place of articulation (n = 640). In terms of the prosodic context, 584 vowels constituted the nucleus of stressed syllables, while 539 appeared in unstressed ones (see Table 1 for a summary).

	stressed	unstressed
symmetrical	260	223
asymmetrical	324	316

Table 1. *Overview of counts of vowel tokens according to segmental and prosodic context.*

The recordings were automatically segmented using the Prague Labeller [19] and then adjusted by hand in Praat [4] following segmentation guidelines by [20], who consider full formant structure onset and offset as the primary criteria for the placement of vowel boundaries.

2.2. Formant extraction & analysis

F1, F2 and F3 values were extracted using a script in Praat at ten equidistant points within the target vowels. The Burg algorithm was applied with three formants detected in the frequency range 0–3.3 kHz. Mean formant values were also obtained from the central third of each vowel. The automatically extracted values were checked and those which appeared unlikely (such as those involving jumps in the contour or mistaking one formant for another) were manually corrected by inspection of the relevant portions of the spectrogram.

The resulting formant contours were then fitted with first, second and third degree polynomial functions using least-squares linear regression in Matlab. This step reduced the number of values characterizing each contour from 10 actual measures to 2, 3, or 4 coefficients, depending on the degree of the polynomial (see Figure 1 for an example). Both the actual measures and the coefficients were in turn used, along with their respective formant means, as predictors for Linear Discriminant Analysis (LDA), performing a closed-set attribution of the vowel tokens to the speakers.

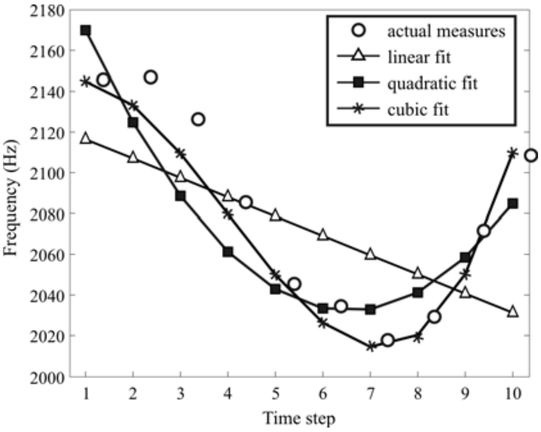

Figure 1. *Plot of F2 contour in one token of /ɪ/ from speaker FLIG: actual measures, linear ($y = Ax + B$), quadratic ($y = Ax^2 + Bx + C$) and cubic ($y = Ax^3 + Bx^2 + Cx + D$) fits.*

3. Results and Discussion

3.1. General results

Our first objective was to determine how the different combinations of formant metrics would contribute to the discrimination of the speakers. Since LDA puts a limit to the number of predictors with respect to the number of tokens, we could only use a subset of the ten equidistant values. In addition to the formant means, we thus included four equidistant measures of the formant trajectory (i.e., values 1, 4, 7, and 10; henceforth referred to as condition A), as well as only the edge measures (values 1, 2, 9, and 10; henceforth referred to as condition B).

As we can see in Table 2, the overall classification rates are comparatively low when all vowels are considered: based only on the mean values of F1–F3, the classification rate reached 15.5 %, with the chance level being 8.3 %. Upon the inclusion of dynamic formant measures, the best classification rate (23.0 %) was obtained with the edge measures (condition B), which supports the assumption that it is the transitions of formant trajectories to the neighbouring speech sounds that allow for the greatest expression of idiosyncratic behaviour. Condition B also yielded the lowest value of Wilks' lambda.

	classification rate (%)	Wilks' lambda
means (F1, F2, F3)	15.5	0.85
means + condition A	22.8	0.68
means + condition B	23.0	0.65

Table 2. *Classification rates and Wilks' lambda values based on mean formant values and two combinations of dynamic formant measurements.*

Table 3 shows discrimination results based on the mean values and the coefficients of polynomial regression. Although we are dealing with monophthongs and we would therefore not expect great movements of formants within the vowels, the cubic approximations still yielded a comparatively higher classification rate, as well as the lowest value of Wilks' lambda.

	classification rate (%)	Wilks' lambda
means + linear	20.4	0.72
means + quadratic	22.5	0.68
means + cubic	24.0	0.66

Table 3. *Classification rates and Wilks' lambda values based on mean formant values and coefficients of polynomial regression.*

In the following analysis, we were interested in the role lexical stress may play in the discrimination of speakers. In English, the absence of stress is often associated with a radical reduction in vowel quality (e.g., *contract* as [ˈkɒntrækt] or [kənˈtrækt]). That is why, in the context of English, only stressed vowels have been analyzed as to their speaker specificity. In Czech, however, vowel quality is mostly maintained regardless of word stress and there is no phonological reduction; in fact, informal observations suggest that mild centralization is more frequent in some stressed than in unstressed syllables. That is why one of our objectives was to find out whether vowels in stressed and unstressed syllables will perform differently in discriminant analysis. According to the classification rates given in Figure 2, vowels in unstressed syllables performed better than in stressed positions, with the exception of condition B. This finding would lend indirect support to the above-mentioned observation that stressed syllables may undergo partial centralization.

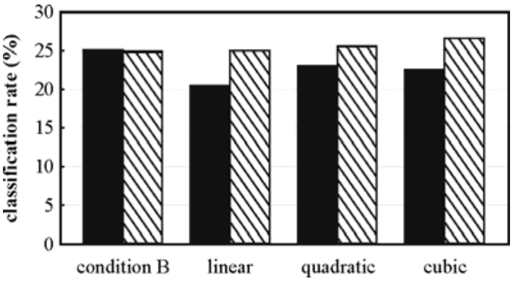

Figure 2. *Classification rates of vowel formants in stressed (black) and unstressed (stripes) positions, ordered by type of predictors used in analysis.*

As we mentioned in section 2.1, we were also interested in determining whether classification rate would be better in asymmetrical consonantal contexts than when the vowel is flanked by consonants articulated at the same place. As can be seen in Figure 3, the asymmetrical contexts reach higher classification rates in all types of analyses, although the difference is negligible in condition B.

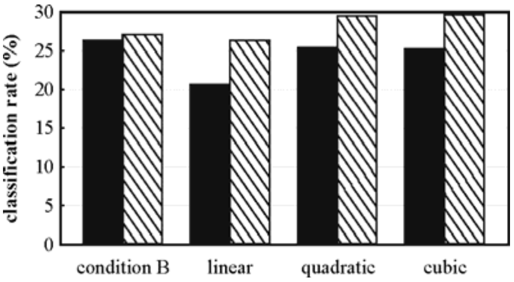

Figure 3. *Classification rates of vowel formants in symmetrical (black) and asymmetrical (stripes) consonantal contexts, ordered by type of predictors used in analysis.*

To summarize this section, vowels in unstressed positions and asymmetrical consonantal contexts seem to hold more speaker-specific information in Czech. The classification rate obtained for vowels in asymmetrical contexts using the cubic coefficients (29.84 %) is nearly 6 % higher than the highest rate from the overall analysis (24 %; see Table 3).

3.2. Vowel-specific results

In order to be able to compare with studies cited in the Introduction, it is necessary to consider each vowel separately: as mentioned above, the analyzed speech material was much more constrained in all these studies, limited to a single vowel or to a controlled consonantal context.

When analyzing the individual vowels, we included condition B (i.e., mean values and measures taken in steps 1, 2, 9, 10), as well as the three sets of polynomial regression coefficients. Figure 4 illustrates how well the vowels performed depending on the predictors used; Wilks' lambdas for condition B are listed below the symbol for each vowel. The decrease in their value, compared with Tables 2 and 3, points to a notable improvement in condition B's reliability when separate vowels are taken into account.

We can see that discrimination was most successful with the phonologically long vowels – /iː/ and /aː/ (60 % and 58.3 %, respectively, in condition B) – followed by the short /u/. It would thus appear that vowels representing the extremes of a speaker's vocalic space provide for more variety in individual realization. It is noteworthy that the classification rates of /iː/ were the most "compact" across the different sets of predictors. Classification rate of /o/ was also tightly clustered, despite the fact that it was the worst-performing vowel (40.1 % in condition B). On the other hand, the classification rates of /a/ differ considerably in each analysis, the greatest step between two percentage values being 17.5 %.

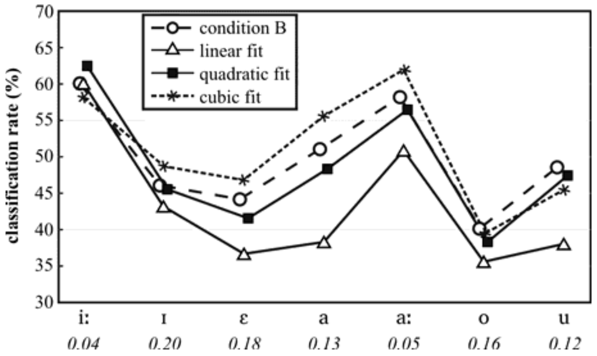

Figure 4. *Classification rates for the individual vowels in the four types of analysis. Values of Wilks' lambda for condition B are given below each vowel.*

Of the polynomial coefficients, the cubic ones ensured best performance in the discriminant analysis. However, the quadratic ones lagged behind only by an average 2.4 %. In the case of /iː/ and /u/, quadratic coefficients even made for better success rates than the cubic ones. The fact that quadratic coefficients capture the formant's properties nearly as aptly as cubic ones could have a positive consequence for discriminant analyses of vowels in the future. There are three instead of four coefficients per formant, and so the number of predictors can be substantially lower (also discussed by [12], [21: 276].

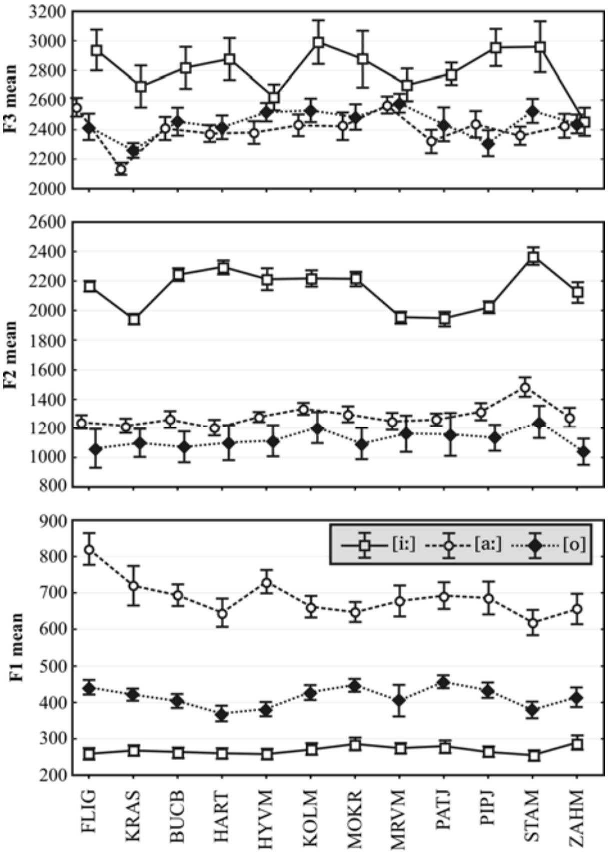

Figure 5. *Mean values of F1–F3 along with 95% confidence intervals for the vowels* [iː aː o] *for the twelve speakers (see text for more detail).*

Figure 5 compares the classification performance of the two most successful vowels, /iː/ and /aː/, with the least successful one, /o/. It shows the distribution of each of the three formants' means for the twelve speakers. As we can see, /iː/ displays greatest variability in F2 and F3 and almost none in F1. The vowel /aː/ appears to be most variable in its F1 and F3, but it is the F2 that shows the lowest degree of within-speaker variability, as suggested by the 95% confidence intervals. Finally, the formant means of /o/ manifest the least between-speaker variability. All of these tendencies are supported by the *F*-values of the F1–F3 means illustrated in Figure 6: it is F2 that contributes the most to the discrimination between speakers through /iː/ and /aː/.

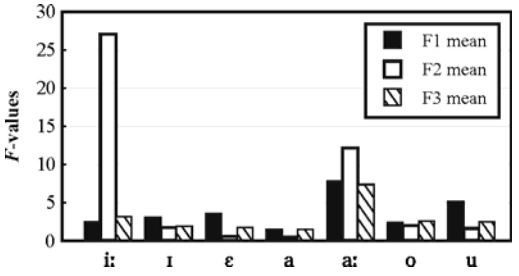

Figure 6. F-*values for F1–F3 in the individual vowels.*

4. General discussion and Conclusions

The aim of this study was to examine whether the dynamic properties of formant trajectories are applicable for speaker discrimination in Czech, whose vocalic system differs importantly from the one in English. Our results, obtained from a comparatively natural speech material, suggest that even the short vowels of Czech allow for idiosyncratic solutions, although the long monophthongs /iː/ and /aː/ performed better. It would be interesting to find out, in a more controlled study, whether the high classification rate of /iː/ and /aː/ is related to their phonological length, but our present material did not provide for such an analysis.

Of the different parameterizations of formant contours, the edge values seem to contain most idiosyncratic information, provided that the formant means are also used as predictors. The use of polynomial regression coefficients also yields good results. Similarly to the findings presented in [12], quadratic coefficients describe the dynamic properties of vowel formants nearly as well as cubic ones. Contrary to our expectations and to other studies, it was F2 and not F3 which proved to contribute the most to the discrimination of our speakers.

The results are encouraging for the Czech forensic context in that they show that a greater portion of material can be used for reliable analyses. First, we do not have to restrain our examinations to stressed syllables (in fact, it was unstressed syllables which performed better in LDA). Second, the variability of consonantal contexts in actual casework does not seem to deteriorate speaker classification.

5. Acknowledgements

This study is a project output of the Internal grants 2012 VG130 solved at the Faculty of Arts in Prague, and of the project GAČR 406/12/0298 granted to the third author.

6. References

[1] Nolan, F. and Grigoras, C., "A case for formant analysis in forensic speaker identification", Int J Speech Lang Law, 12(2): 143-173, 2005.

[2] Jessen, M., "Phonetische und linguistische Prinzipien des forensischen Stimmenvergleichs", Munich: Lincom, 2012.

[3] Kinnunen, T. and Li, H., "An overview of text-independent speaker recognition: From features to supervectors", SpeCom, 5: 12-40, 2010.

[4] Boersma, P. and Weenink, D. (2012). Praat - Doing phonetics by computer (Version 5.3.30.). Online: http://www.praat.org, accessed on 8 October 2012.

[5] de Jong, G., McDougall, K. and Nolan, F., "Sound Change and Speaker Identity: An Acoustic Study", in C. Müller [Ed], Speaker Classification II, LNAI 4441, 130-141, Springer, 2007.

[6] Duckworth, M., McDougall, K., de Jong, G. and Shockey, L., "Improving the consistency of formant measurement", Int J Speech Lang Law, 18: 35-51, 2011.

[7] Moos, A., "Long-term formant distribution as a measure of speaker characteristics in read and spontaneous speech", The Phonetician, 101/102: 7-25, 2012.

[8] Jessen, M. and Becker, T., "Long-term Formant Distribution as forensic-phonetic feature", ASA 2nd Pan-American/Iberian Meeting on Acoustics, Cancún, México, 2010.

[9] Goldstein, U., "Speaker-identifying features based on formant tracks", J Acoust Soc Am, 1: 176-182, 1976.

[10] Greisbach, R., Esser, O. and Weinstock, C., "Speaker identification by formant contours", In A. Braun and J. Köster [Eds], Studies in Forensic Phonetics: Beiträge zur Phonetik und Linguistik. Trier: Wissenschaftlicher Verlag Trier, 49-55, 1995.

[11] Ingram, J., Prandolini, R. and Ong, S., "Formant trajectories as indices of phonetic variation for speaker identification", Forensic Linguistics, 3(1): 129-145, 1996.

[12] McDougall, K., "Dynamic Features of Speech and the Characterisation of Speakers: Towards a New Approach Using Formant Frequencies", Int J Speech Lang Law, 13(1): 89-126, 2006.

[13] Nolan, F., McDougall, K., de Jong, G. and Hudson, T., "A forensic phonetic study of 'dynamic' sources of variability in speech: The DyViS Project", P. Warren and C. I. Watson [Eds], 11th Australian International Conference on Speech Science & Technology, 13-18, 2006.

[14] McDougall, K., "Speaker-specific formant dynamics: An experiment on Australian English /aɪ/", Int J Speech Lang Law, 11(1): 103-130, 2004.

[15] McDougall, K. and Nolan, F., "Discrimination of speakers using the formant dynamics of /u:/ in British English", In J. Trouvain and W. Barry [Eds], Proc 16th ICPhS, Saarbrücken, 1825-1828, 2007.

[16] Skarnitzl, R. and Volín, J., "Reference values of vowel formants in young adult speakers of standard Czech", Akustické listy, 18: 7-11, 2012. [in Czech]

[17] Bartoň, T., Cvrček, V., Čermák, F., Jelínek, T. and Petkevič, V., "Statistics of Czech", Praha: Lidové noviny/ÚČNK, 2009. [in Czech]

[18] Enzinger, E., "Characterising Formant Tracks in Viennese Diphthongs for Forensic Speaker Comparison", In Proc AES 39th International Conference – Audio Forensics, 47-52, Hillerød, Denmark, 2010.

[19] Pollák P., Volín J. and Skarnitzl R., "HMM-Based Phonetic Segmentation in Praat Environment", Proc SPECOM 2007, 537-541, 2007.

[20] Machač, P. and Skarnitzl, R., "Principles of Phonetic Segmentation", Praha: Epocha, 2009.

[21] Volín, J., "Statistical Methods in Phonetic Research", Praha: Epocha, 2007. [in Czech]