SPEAKER DISCRIMINATION USING FORMANT TRAJECTORIES FROM CASEWORK RECORDINGS: CAN LDA DO IT?

Radek Skarnitzl, Jitka Vaňková & Tomáš Nechanský

Institute of Phonetics, Faculty of Arts, Charles University in Prague
radek.skarnitzl@ff.cuni.cz, jitka.vanka@gmail.com, tomas.nechansky@seznam.cz

ABSTRACT

Formant trajectories have been shown to convey a great deal of speaker-specific information and their speaker-discriminatory potential has been quantified using Linear Discriminant Analysis on laboratory material [16]. This study tests the applicability of LDA on three sets of real-case forensic recordings. Given the limitations of LDA, we used the actual formant trajectory values (F1–F3) and coefficients of the quadratic and cubic fit. As for classification rate, our results indicate that LDA performs comparably to the studio condition, with quadratic fit being the most convenient way of parametrizing the trajectory. However, LDA performed well above chance when discriminating between recordings of the same speaker; it is especially this inability to "identify" the same speaker which makes the use of LDA in forensic practice not recommendable.

Keywords: speaker discrimination, vowel formants, forensic phonetics, LDA, Czech.

1. INTRODUCTION

Forensic phoneticians have been trying for decades to identify acoustic properties of speech which manifest a satisfactory ratio of between-speaker to within-speaker variability. Vowel formants have always been in the centre of these endeavours: they have been traditionally ranked among the most useful parameters in speaker identification [10], [21], because they reflect, to a certain extent, both the physiological properties of the speaker (i.e., vocal tract size) and various sociolinguistically determined factors (e.g., the speaker's regional background, age or socioeconomic class), as well as the differences in phonetic implementation [19], [20] (i.e., the individual realization of vowels mapped on to the given vocal tract). Other characteristics of vowel formants which favour their widespread use include their relative robustness in noisy conditions and also the fact that they can be easily extracted from the signal (but see [3], [9] or [24] for methodological aspects).

For forensic purposes, vowel formants have been used in three ways. First, it is possible to extract and compare the mean formant values from the phonetic target [3], [21]. The second method consists in the analysis of long-term formant (LTF) distributions, as proposed by Nolan and Grigoras [21]; this method has recently been gaining in popularity in the forensic phonetic community [12], [13], [18], because it captures a speaker's global speaking habits such as a tendency to palatalize or labialize. The time-free LTF analysis is nicely complemented by the last method, the comparison of formant trajectories which, conversely, takes the temporal dimension into account. Although first speaker identification studies exploiting formant trajectories date decades back [6], [7], it is the belief of Nolan and colleagues [20], [22] that dynamic properties of speech provide most cues to speaker identity which has stimulated most research, especially by McDougall [15], [16], [17]. While formant values in the vocalic targets, mentioned above, do exhibit some speaker-specific potential, they are largely constrained by the speaker's linguistic system. It appears to be the transitions *between* individual phonetic targets where idiosyncratic implementation – reflecting both the speaker's physiology and acquired speaking habits – is free to manifest itself.

So far, most studies investigating the speaker-specificity of formant trajectories have focused on English, whose vowel system is quite specific: McDougall analyzed the formant contours in words ending in [a͡ɪk] (*hike*, *bike*) or in sequences like [əˈɹV] (*peruse*, *charade*). As not all languages have such an abundance of suitable sounds, Fejlová et al. [5] examined the speaker-discriminating potential of formant trajectories in Czech, a language where diphthongs and long vowels are comparatively rare. In addition, to increase the ecological validity of the findings, the study analyzed vowels occurring in more segmental contexts (ones which were both symmetric and asymmetric from the perspective of articulation place, for instance [dVs] and [pVk], respectively), as well as in different prosodic contexts. Their results indicate that formant contours of even short vowels contain an reasonable degree of speaker-specific information and that the second-degree polynomial (i.e., quadratic) function fit to the formant trajectory captures the formant dynamics adequately, while considerably reducing the number of predictors necessary for the Linear Discriminant Analysis (LDA); *cf.* the results reported by [16] and see [28: 276] for LDA limitations.

The objective of this study is to continue in this line of research and to enhance the ecological validity still more, by analyzing real forensic material provided by the Czech Police. To our knowledge, such research has not been conducted before.

The usage of real casework material has several implications which deserve attention. First of all, the recordings were obtained by means of police wiretaps on mobile telephones. It is well known that mobile telephone transmission affects the formant frequencies of vowel sounds; fortunately, the effect seems smaller for male voices and for lower formants [4], [8], [26], [27]. In addition, since speakers often move freely when using mobile phones, the resulting recordings often contain differing degrees of environmental noise, from wind or traffic sounds to the sound of competing talkers. McDougall [16] used Linear Discriminant Analysis on a controlled material and she states that her positive results do not automatically mean that LDA will be applicable in actual forensic phonetic casework – naturally, all the above-mentioned factors are expected to impair the performance of LDA. Our aim is therefore to find out the nature of this effect and, specifically, to see whether LDA may be of any assistance to forensic practitioners.

2. METHOD

The data for this study consist of three sets of recordings, each related to a single casework solved by the Czech Police. The individual sets contained recordings of mostly unknown (suspect) speakers, typically originating from mobile phone intercepts. Altogether we worked with 20 minutes of recordings corresponding to as many as 23 male voices (the actual number of speakers was smaller, as more recordings came from the same speaker). As we will present classification results for each set separately, focusing on different aspects of the evaluations, the details regarding the speakers and recordings will be introduced more thoroughly in the results section.

The transcribed recordings were automatically aligned using Prague Labeller [23], and the boundaries of all vowel segments were manually corrected in line with the recommendations by [14]. Subsequently, we extracted from each vowel eleven equidistant values for the first three formants using the *Formant Tracker* implemented in Praat [1]; default settings – corresponding to an average male vocal tract size, with 500 Hz, 1500 Hz and 2500 Hz for F1, F2 and F3 respectively – were used for formant tracking.

Since vowel formant extraction is not always reliable [24], especially in lower-quality recordings such as those used here, the formant values were also inspected visually. The vowels in which the automatically extracted formant values appeared unlikely due to jumps in the contour or due to one formant being mistaken for another, or those where the formants differed widely from those expected for Czech male speakers [25] were removed. In the end, we worked with the total of 3,659 vowels.

The resulting formant contours were then fitted with second and third degree polynomial functions (quadratic and cubic fits) using least-squares linear regression in Matlab. The extracted formant values (in SET 1) and coefficients of polynomial regression (in SETS 1–3) served as input for LDA.

3. RESULTS AND DISCUSSION

3.1. Parametrizing the formant contours

As indicated in the previous section, the three sets of recordings were analyzed separately, attention always being paid to a different aspect of analysis. The presentation and discussion of results will therefore focus on one set at a time.

In SET 1, our aim was to compare the performance of LDA based on different predictors. Overall, 1,490 vowels from four unknown speakers were analyzed. We used the extracted formant values and the coefficient values of quadratic and cubic polynomial regression; only the 1st, 2nd, 6th, 10th, and 11th value was taken from the formant contour, due to the above-mentioned limitations of LDA.

First, the most general results were compared with those obtained by [5] on a controlled material. Figure 1 shows classification rates of the formant values and of the quadratic fit, for all vowels combined, in the two studies (results for cubic fit are not reported but were quite comparable to those obtained for quadratic fit). The comparison is rather favourable: the classification rate given by LDA is well above chance not only for controlled recordings ([5], in the upper part of the figure), but also for recordings from real forensic cases (the lower part). The exact location of the chance level in the two studies differs due to a different number of speakers, which was 12 in [5], corresponding to the chance level of 8.3%, and 4 in this study, hence 25%).

Furthermore, the performance of formant values (marked by a circle) and of the quadratic fit (marked by a square) is largely comparable in both studies, with only a small tendency for formant values to score higher. Although the quadratic fit performs slightly worse, its significant advantage is that the number of predictors is 9, while in the latter case it is 15. Given the limitations posed by LDA on the number of predictors [28], a quadratic fit appears as a reasonable compromise: despite reducing the number of predictors, the performance is not significantly impaired. The Wilks' lambda value for the whole discrimination model equals 0.67 with the actual formant values and 0.69 with quadratic fit coefficients, which also supports our conclusion that the polynomial fitting does not reduce the overall effectiveness to a larger degree. Only quadratic fits are thus used in the subsequent analyses.

Figure 1: Comparison of classification rates reported by [5] for twelve speakers recorded in studio conditions (**a.**) and classification rates for four speakers in SET 1 (**b.**).

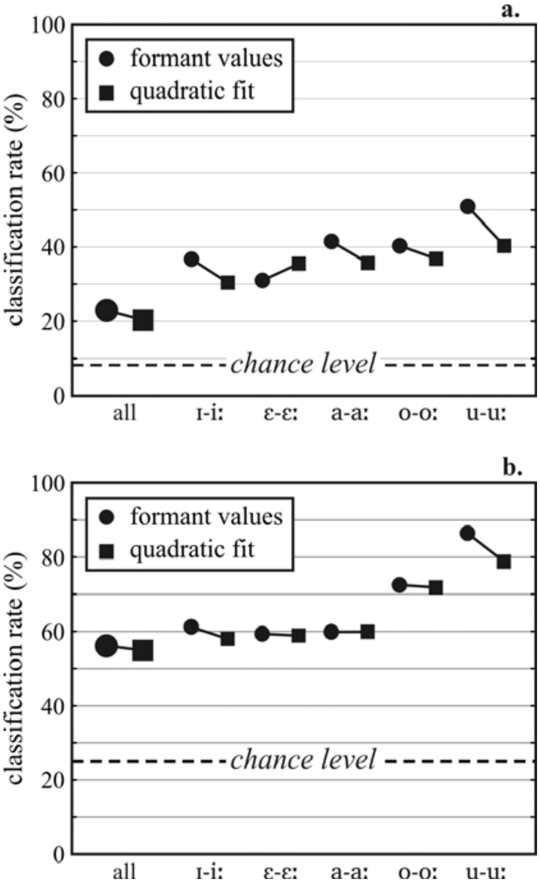

Figure 1 also reveals that higher classification rates can be observed when individual vowels are analyzed. This is not surprising, as it is likely that only some vowel qualities show idiosyncrasy for a given speaker – either due to its truly idiosyncratic realization or, possibly, as a result of a linguistic change. For English, higher classification rate (and hence higher speaker-specificity) has been identified for vowels undergoing diachronic change [2]. Our results indicate that this could be the case for Czech, too: the best performing vowels were the close back vowels [u uː] which also seem to be undergoing a quality shift [25] (it should be noted, however that [u uː] were also the least numerous vowels in our dataset, as there are only 52 such items in SET 1 for all the four speakers). On the other hand, the vowels achieving the second highest classification rate [o oː] do not seem to be undergoing any such shift and appear to be quite stable.

This first experiment thus shows that LDA performs in a comparable way as on laboratory speech. Obviously, the classification rates are low for applicability in forensic casework; however, it would be unrealistic to expect better based on formant analysis alone. Classification may turn out to be considerably higher if combined with other acoustic parameters.

3.2. Within-speaker variability

Apart from between-speaker differences, also the variability within one speaker is crucial for forensic purposes, as discussed in the introduction, and needs to be addressed. As a next step, therefore, our goal was to find out whether LDA is an adequate tool from the perspective of intra-speaker variability, too.

First, we were interested in finding out the potential effect of stylistic and contextual variability. In SET 2, there were six recordings, all obtained from police wiretaps. Three recordings came from an unknown speaker who was talking with an accomplice (these will be referred to as type A); in the other three recordings the speaker tries to arrange an illegal deal with a potential business partner, who was actually a police informer (these are referred to as type B recordings). Since we are talking about actual forensic cases, one should not make absolute claims about the identity of the speakers; however, based on thorough auditory analysis (which is in line with the required combination of acoustic and auditory approaches to forensic speaker identification [10], [11], [20]) we can state with very high probability that the unknown and known recordings come from the same speaker. Most importantly for the present analysis, the recordings differ markedly in speech style: while the speaker is clearly at ease with the accomplice, he uses colloquial language and engages in friendly banter, the other three recordings are characterized by a much more distant, formal speaking style. In total, we worked with 1,181 vowels in SET 2.

What we were interested in is whether the different contexts have some impact on formants; since all the recordings come from the same speaker, no such effect should be observed. We would expect to find similar classification rates when comparing same-style recordings (a type A recording with another A recording, or two B recordings) as when comparing across styles; in Table 1, then, the sum in all quadrants should be similar. That is, however, not what we can observe in the table. Instead, there are 346 cases where an A type recording is classified as A and 478 items when B type recordings are classified as B, but only 196 cases where A is identified as B, and 161 cases of B identified as A. Although all the recordings are of similar technical quality (mobile phone recordings with little background noise and essentially no waveform clipping), the LDA matches same-style recordings more that different-style recordings.

Table 1: Classification matrix showing the discrimination of three type A and three type B recordings from the same speaker (SET 2); see text.

	A1	A2	A3	B1	B2	B3
A1	181	9	11	1	84	5
A2	40	8	15	0	22	7
A3	55	5	22	0	69	8
B1	52	1	4	1	144	0
B2	63	3	7	1	265	0
B3	13	11	7	0	54	13

Let us turn to the analysis of SET 3, which involves lower complexity of the examined material; that allowed us to dig deeper and try to identify other possible sources of variability in the data.

SET 3 included three suspect recordings known beyond any doubt to originate from one and the same speaker. All these recordings involved mobile phone calls to a police station, and they were of a largely comparable quality. Altogether, these three recordings yielded 353 vowels, with the number of vowels per recording ranging between 64 and 147.

If LDA is a convenient tool for tackling forensic recordings, it should not discriminate between recordings coming from one speaker; in other words, classification rates should not exceed the chance level too much, because it simply should not be possible to distinguish one recording from another recording of the same speaker.

Table 2: Classification matrix showing the discrimination of three recordings from the same speaker (SET 3).

	A1	A2	A3	class. rate
A1	0	27	37	0.0%
A2	1	99	42	69.7%
A3	1	44	102	69.4%
total:	2	170	181	**56.9%**

The classification matrix shown in Table 2 reveals, however, that the discrimination does not yield values near the chance level – though the Wilks' λ of 0.77 is relatively higher than in SET 1 (i.e., it is more difficult to discriminate between the three voices), the overall classification rate of 56.9% is well above the chance level of 33.3% (since we had three voices which we were trying to discriminate between).

It is especially the classification of vowels coming from recording A1 which is curious: none of the vowels was classified by LDA as coming from that recording. Since recording A1 also yielded the fewest vowel tokens, this led us to formulate a research question concerning the number of vowels available for LDA: the 0% classification rate may possibly stem from a lower number of vowels in this recording.

To answer this question, we randomly selected 64 vowels from recordings A2 and A3 so as to equalize their number across the three recordings, and ran LDA again. The data presented in Table 3 show that the classification rates for the individual recordings are more levelled and the overall classification rate dropped considerably when compared with Table 2, but still lies above the 33.3% chance level. Even with the number of vowel items equalized, then, LDA finds information in the vowel formants which allow it to distinguish between the recordings with a higher than chance classification rate. Clearly, however, an effect of the data size on classification rate could be observed.

Table 3: Classification matrix showing the discrimination of three recordings from the same speaker (SET 3), after the numbers of vowel tokens have been equalized across recordings.

	A1	A2	A3	class. rate
A1	20	22	22	31.3%
A2	16	37	11	57.8%
A3	17	16	31	48.4%
total:	53	75	64	**45.8%**

4. CONCLUSION

As Linear Discriminant Analysis (LDA) has been successfully used in research on laboratory speech [5], [15], [16], [17], the aim of this study was to verify the applicability of LDA on forensic casework material. McDougall herself [16] states that high classification rates do not automatically make the method suitable for forensic casework. In our first analysis, we showed that classification rates obtained from casework material are comparable with those obtained for laboratory material. However, based on the fact that LDA appears to be sensitive to aspects like speech style and especially its tendency to discriminate above chance level between same-style recordings of one speaker (in other words, its inability to identify recordings from the same speaker), we have to conclude that LDA is not suitable as a tool in forensic phonetic casework.

ACKNOWLEDGEMENT

This research was supported by the Czech Science Foundation (GAČR 406/12/0298) and by Specific Charles University Research Programme granted to the second (VG184) and third (VG118) author.

5. REFERENCES

[1] Boersma, P., Weenink, D. 2014. *Praat: doing phonetics by computer (Version 5.4)*. Retrieved from http://www.praat.org on October 14, 2014.

[2] De Jong, G., McDougall, K., Nolan, F. 2007. Sound change and speaker identity: An acoustic study. In: Müller, C. (ed), *Speaker Classification II*. Berlin; New York: Springer, 130–141.

[3] Duckworth, M., McDougall, K., de Jong, G., Shockey, L. 2011. Improving the consistency of formant measurement. *International Journal of Speech, Language and the Law* 18, 35–51.

[4] Enzinger, E. 2010. Measuring the Effects of Adaptive Multi-Rate (AMR) codecs on formant tracker performance. *Proc 2nd Pan-American/Iberian Meeting on Acoustics*.

[5] Fejlová, D., Lukeš, D., Skarnitzl, R. 2013. Formant Contours in Czech Vowels: Speaker-discriminating Potential. In: *Proc Interspeech* Lyon, 3182–3186.

[6] Goldstein, U. G. 1976. Speaker-identifying features based on formant tracks. *Journal of the Acoustical Society of America* 59, 176–182.

[7] Greisbach, R., Esser, E., Weinstock, C. 1995. Speaker identification by formant contours. *BEIPHOL 64, Studies in Forensic Phonetics*, 49–55.

[8] Guillemin, B. J., Watson, C. 2008. Impact of the GSM mobile phone network on the speech signal: some preliminary findings. *International Journal of Speech, Language and the Law* 15, 193–218.

[9] Harrison, P., Clermont, F. 2012. The influence of LPC order on the accuracy of formant measurements across speakers. *Proc. IAFPA* Santander.

[10] Hollien, H. 2002. *Forensic voice identification*. San Diego, Calif: Academic Press.

[11] Jessen, M. 2012. *Phonetische und linguistische Prinzipien des forensischen Stimmenvergleichs.* München: Lincom.

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

[13] Jessen, M., Enzinger, E., Jessen, M. 2013. Experiments on Long-Term Formant Analysis with Gaussian Mixture Modeling using VOCALISE. *Proc. IAFPA* Tampa.

[14] Machač, P., Skarnitzl, R. 2009. *Principles of Phonetic Segmentation*. Praha: Epocha.

[15] McDougall, K. 2004. Speaker-specific formant dynamics: An experiment on Australian English /aɪ/. *International Journal of Speech, Language and the Law* 11, 103–130.

[16] McDougall, K. 2006. Dynamic Features of Speech and the Characterisation of Speakers: Towards a New Approach Using Formant Frequencies. *International Journal of Speech, Language and the Law* 13, 89–126.

[17] McDougall, K., Nolan, F. 2007. Discrimination of speakers using the formant dynamics of /u:/ in British English. *Proc ICPhS* Saarbrücken, 1825–1828.

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

[19] Nolan, F. 1983. *The phonetic bases of speaker recognition*. Cambridge: Cambridge University Press.

[20] Nolan, F. 1999. Speaker Recognition and Forensic Phonetics. In: Hardcastle, W., Laver, J. (eds), *The Handbook of Phonetic Sciences*. Oxford: Blackwell, 744–767.

[21] Nolan, F., Grigoras, C. 2005. A case for formant analysis in forensic speaker identification. *International Journal of Speech, Language and the Law* 12, 143–173.

[22] Nolan, F., McDougall, K., de Jong, G., Hudson, T. 2009. The DyViS database: style-controlled recordings of 100 homogeneous speakers for forensic phonetic research. *International Journal of Speech, Language and the Law* 16, 31–57.

[23] Pollák, P., Volín, J., Skarnitzl, R. 2007. HMM-Based Phonetic Segmentation in Praat Environment. *Proc SPECOM* Moscow, 537–541.

[24] Skarnitzl, R., Vaňková, J., Bořil, T. 2015 (in print). Optimizing formant extraction. In: Niebuhr, O., Skarnitzl, R. (eds), *Tackling the Complexity in Speech*. Praha: Faculty of Arts.

[25] Skarnitzl, R., Volín, J. 2012. Reference values of vowel formants for young adult speakers of Standard Czech. *Akustické listy* 18, 7–11. [in Czech]

[26] Vaňková, J., Bořil, T. 2014. Telefonní přenos. In: Skarnitzl, R. (ed), *Fonetická identifikace mluvčího*. Praha: Faculty of Arts, 104–115.

[27] Vaňková, J., Bořil, T. submitted. Impact of the GSM AMR codec on automatic vowel formant measurement in Praat and VoiceSauce (Snack). *International Journal of Speech Technology*.

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