Reviewing the performance of formants for Forensic Voice Comparison: a meta-analysis of forensic speech science research

Lois Fairclough[1], Georgina Brown[1,2] and Christin Kirchhübel[2]

[1]Department of Linguistics and English Language, Lancaster University, UK
[2]Soundscape Voice Evidence, Lancaster, UK
{l.fairclough | g.brown5}@lancaster.ac.uk, ck@soundscapevoice.com

ABSTRACT

It is widely accepted in forensic speech science that formants have speaker discriminatory power, and therefore formant measurements commonly feature in forensic speech casework. Courts in Northern Ireland even go as far as to insist that formant analysis has to be carried out as part of forensic voice comparison analysis. However, work in the broader phonetics field has started to challenge the weight that has been traditionally attached to formants in a range of related subdisciplines – this justifies a review of the use and performance of formants in forensic speech science research studies. This paper therefore presents part of a meta-analysis of forensic speech science studies which test formants as a parameter for speaker discrimination. The results are highly variable across the 277 results from the 37 papers included. Some performance trends that might be expected are not evident from this meta-analysis.

Keywords: forensic phonetics, vowel formants, speaker discrimination

1. INTRODUCTION

The job of a forensic speech analyst typically involves comparing questioned and known speech samples in order to address the question of whether it is the same speaker or different speakers featuring in the recordings. The auditory phonetic and acoustic approach is a well-established method applied to the forensic voice comparison (FVC) task, as shown in a survey of forensic speech practitioner practices [1]. Within this approach, analysts consider multiple parameters, one being vowel formants. In fact, in UK practice, there is a general expectation that practitioners will include an analysis of formants in their voice comparison analyses. This is because formants are said to reflect the resonances of the vocal tract, thereby tapping into the physical characteristics of the speaker [2].

Despite the weight that is given to formants on theoretical grounds, there is not a clear account of how well formants perform in practice at speaker discrimination. In recognition of this, the overarching objective of this paper is to carry out a systematic review of the relevant research literature and to offer a description of the performance of formants for the purpose of discriminating speakers in the forensic setting.

Section 2 lays out further background information and motivation for this work. Section 3 explains the methods used in this meta-analysis, while section 4 reveals the initial findings of the analysis. Section 5 discusses the findings and some implications around the use of formants in forensic speech science.

2. BACKGROUND

This section first describes what formants are, and how they can be measured in forensic speech science research and practice. The section then moves on to outline the status given to formants in FVC.

2.1. What are formants?

Formants represent the resonant frequencies of the vocal tract [3], and as such reflect a speaker's physical and articulatory characteristics. While F1 and F2, to a great extent, reflect the phonetic quality of a vowel, the higher formants (F3, F4 and F5) are said to represent more speaker-specific characteristics [3]–[5]. Generally, F4 and F5 are not available for analysis in forensic casework due to the frequency bandwidth reduction typical of the types of recordings involved.

Formants can be analysed in different ways. Three key ways of measuring formants include midpoint measurements, dynamic measurements and long-term formant analysis (LTF). All three of these appear in the forensic speech science literature. Midpoint measurements capture formant frequencies from the centre of the vocalic segment in question [6], while dynamic measurements capture multiple time points across the segment [7]. LTF analysis captures formants across all vocalic portions of speech in a sample [8]. Within each of the three techniques there are different strategies by which the formant analysis can be carried out. For example, the intervals from which LTFs can be calculated might differ – some might take measurements every 5ms, others every 10ms.

Irrespective of the type of formant measurement used, recent discussion in acoustic phonetics has questioned the reliability of formants as a parameter in speech analysis. This emphasises that the methods used to obtain formants need more attention and caution [9], and reinforces the idea that formant measurements are actually only "estimates", not true values [10]. Together, this puts into question the representation of formants in the forensic literature and beyond.

2.2. Formants in FVC casework

Formants are widely used by forensic speech practitioners. In the survey by [1], it was found that of 36 experts, 97% of them carried out some form of formant analysis. Of the practitioners using formants, all measure F2, 87% measure F1 and F3, while 17% measure F4. Importantly, 94% measured the centre frequencies of monophthongs, 71% reported measuring the trajectories of diphthongs, and 45% measured vowel-consonant or consonant-vowel formant transitions [1]. In addition, practitioners also make use of LTF analysis [11, 12]. Detailed results of formant measurement practice in [13] revealed diversity in carrying out formant measurements, highlighting the flexibility in the field when it comes to integrating formants into FVC analyses.

In the legal context, there is a court ruling in the jurisdiction of Northern Ireland that places an expectation on practitioners to include formants in their FVC analyses. It states that "*no prosecution should be brought in Northern Ireland in which one of the planks is voice identification given by an expert which is solely confined to auditory analysis… there should also be expert evidence of acoustic analysis… which includes formant analysis*" [14]. This reinforces the significance placed on using formant analysis for FVC.

Given the theory that underpins formants, and the significance that has been placed on formants within and outside the forensic speech science community, it seems fitting to carry out a review of the forensic speech science research literature to interrogate their speaker discrimination performance. This can be achieved through a meta-analysis of the existing research literature.

3. METHODOLOGY

In selecting papers for inclusion in the meta-analysis, an initial screening of each paper was undertaken to ensure that formants were used as a parameter for investigation in isolation of other parameters. A basic sampling approach was adopted by collecting papers from Google Scholar, university library online resources, and journals which have particularly relevant contributions to forensic speech science. From this initial set of papers, the references were examined for more papers, increasing the sample size.

The above exercise resulted in over 100 forensic speech science studies being considered for inclusion in the meta-analysis. Reasons for exclusion comprised, among other things, the fact that some studies were only represented by a conference abstract, or they combined formant measurements with other parameters (e.g., MFCCs) to generate results. Of the papers that were included, two key types of study motivation emerged: 1) analysis of formants for the purpose of speaker discrimination and 2) analysis of formants to test their 'robustness' (i.e., within-speaker variation owing to speaker-internal or speaker-external factors). This paper focuses solely on the analysis of speaker discrimination-based results which led to the inclusion of 277 results in this meta-analysis.

Each paper was interrogated according to 23 different factors which cover the nature of the dataset that was used, the formant measurement technique, the analysis, and the results. The present paper pays particular attention to the following factors: measurement type (e.g., midpoint, dynamic, LTF), the number of formants (e.g., F1, F2, F3), and the performance result.

Papers generally reported speaker discrimination performance through Equal Error Rates (EERs) or Classification Rates (CRs). However, there were instances of alternative quantitative measures and qualitative comments which did not fall into the EER or CR categories. EERs reflect the threshold for a system's false acceptance and false rejection rates. The lower the EER, the higher the performance of the measurement parameters for speaker discrimination [15]. CRs most often present the results of a discriminant analysis, whereby linear combinations of features are identified to characterise and group speakers. This multivariate approach can be used to determine whether a set of predictors can be combined to predict group membership, or in this case speaker membership [16]. Importantly, EERs are produced as a result of studying speaker discrimination as an "open-set" type of problem (i.e., does the analysis support the same-speaker or different-speaker view?), whereas the studies that use CRs to reflect speaker discrimination performance assume a "closed-set" type of problem (i.e., which speaker is it out of the "closed-set" of X speakers?). As such, the CR results are heavily dependent on the number of speakers that were included in the "closed set" of speakers.

This paper focuses on EERs and CRs as they account for the majority of results in the speaker discrimination papers.

4. FINDINGS

The meta-analysis results are derived from 37 different papers from 25 different first authors. Some authors feature frequently, with the most frequent appearing as first author on 6 publications and many more as a co-author. The papers included are dated from 1996-2021, therefore spanning 25 years. The number of speakers in each study range from 5 to 171. The findings cover 8 languages including Cantonese, Czech, Dutch, English, German, Mandarin, Shanghainese, and Swedish. The research predominantly included English (21 of the 37 papers). Within English, three varieties were investigated: Australian English, North American English and Standard Southern British English (SSBE). Of the 37 papers, 12 analyse spontaneous speech, 18 read speech, and 7 semi-spontaneous. The majority of the papers are based on laboratory recorded speech (both direct and using telephone transmission), and only two papers are based on casework recordings.

Two mixed-effects linear regression analyses were run for each type of result (i.e., EER and CR). For these analyses, only factors that featured in all of the studies were included. Number of speakers, linguistic variety, and measurement type were included as fixed effects, while the paper ID was included as a random effect. The analysis of EER results brought about one significant effect which corresponded to the number of speakers ($p = 0.04$).

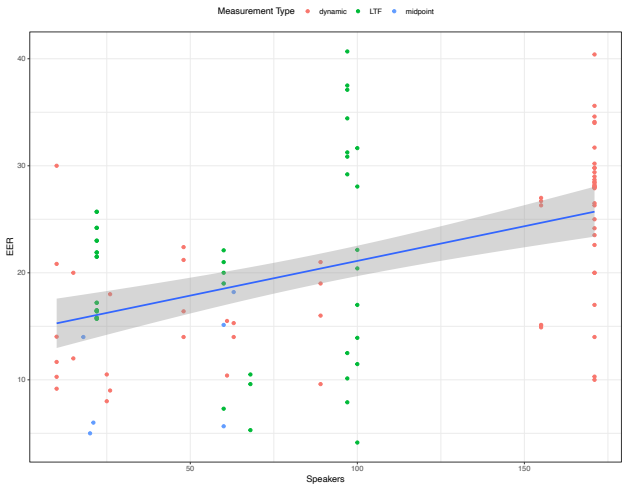

Figure 1: Number of speakers by EER results

Looking more closely, there is a weak tendency for studies which include a larger number of speakers to display higher EERs. This could be connected to the idea that a larger pool of speakers creates more room for different speakers to sound more similar to one another – therefore there is opportunity for more discrimination errors. As we would expect, the analysis of the CR results uncovered a significant effect for number of speakers ($p = 0.01$). In addition, however, linguistic variety also resulted in a significant effect ($p = 0.01$), and measurement type resulted in a near-significant effect ($p = 0.07$).

Figure 2 shows the results of the EER dataset. Studies which have measured the midpoint have the lowest median EERs and they all fall below 20% EER. Studies which have measured dynamic formant trajectories and those which have included LTF measurements display similar median EERs and similarly wide ranges (i.e., EERs ranging from below 10% to above 30%). Importantly, there is only a very low number of studies which have analysed midpoints and reported EERs. This might be the reason for the lower degree of variation in midpoint results compared to dynamic and LTF results.

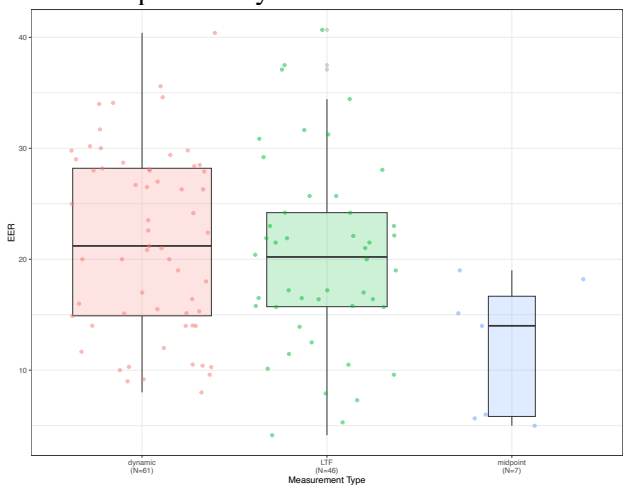

Figure 2: Measurement type by EER results

To further investigate these patterns, the results were visualised according to the specific formant frequencies analysed. This was in order to discover whether certain formants have greater speaker distinguishing power than others (see Figure 3).

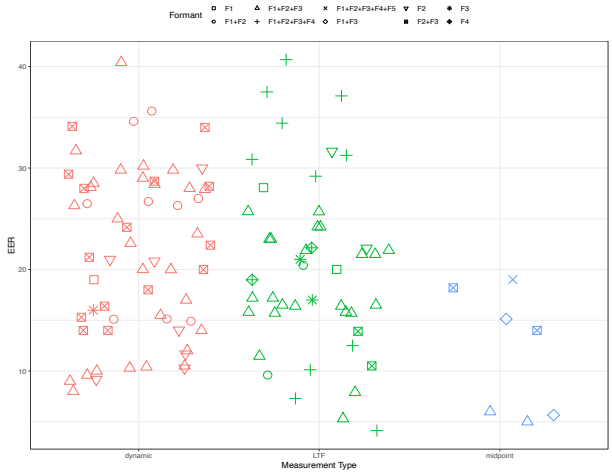

Figure 3: Formant frequencies and measurement type by EERs

There is no clear pattern in terms of specific formant frequencies and low EERs. However, it appears that adding more formants into the analysis does not show

a clear improvement. To illustrate, taking the largest number of formants in a measurement combination for the dynamic category, i.e., F1+F2+F3, this shows values ranging from the lowest EERs to the highest. Likewise, when focusing on the F1+F2+F3+F4 combination for the LTF category, we can see a similarly broad range of performance results. The midpoint category contained the largest measurement combination that occurred across all the studies (F1+F2+F3+F4+F5) - this resulted in the highest EER for the midpoint category. Overall, relatively little research has focused on the higher formants, i.e., F3, F4 and F5, in isolation. It is anticipated that some of the variability within the dynamic and midpoint measurement types may be connected to the specific vocalic segments that were in focus. In addition, differences in recording context are also likely to have contributed to the variability in performance across all studies. Further work would be needed to unpack these interactions.

Turning to the CR dataset, Figure 4 presents a plot of the number of speakers relative to the CR achieved. Obviously, the more speakers included in the discrimination task, the worse the CR result. More interesting, though, is the incidental relationship between measurement type and the number of speakers included in the "closed-set". This may explain the overall better classification rates of dynamic over midpoint measurements (and the near-significance of measurement type). To illustrate, a cluster of midpoint values at the 50-speaker mark has poor classification rates compared to the high CR rates of between 50 and 90% using dynamic measurements with <10 speakers.

Figure 4 also shows that even when the number of speakers remains stable – for instance all the midpoint results – there are high levels of variation in classification rates. This highlights the importance of understanding the influences on CRs other than the number of speakers, such as the specific formant frequencies measured, the vowel segments included and the recording contexts.

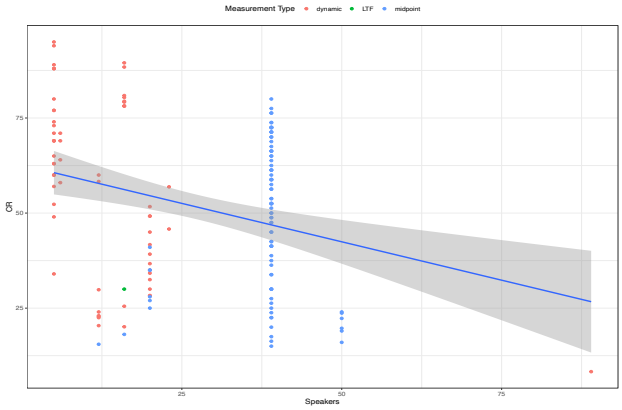

Figure 4: Number of speakers by classification rates

Having further investigated the significant effect of linguistic variety, it transpires that SSBE shows overall poorer classification rates compared to other varieties. However, this finding should not be overinterpreted in view of the low number of studies.

5. DISCUSSION

It might come as a surprise to learn that relatively few studies use midpoint formant measurements in the speaker discrimination research literature. Midpoint measurements seem to exhibit better performance overall than dynamic and LTF measurements, potentially as a result of the small number of studies that use midpoint measurements and report EERs. This overall low usage of midpoint measurements in the forensic speech science research literature seems to be at odds with the 97% of practitioners who reported to use midpoint measurements in forensic casework [1]. This discrepancy between research and practice further extends to the data itself, with the majority of the forensic speech science research using lab-based speech which is not representative of casework material. This meta-analysis reveals other trends that go against some of our expectations including the notion that adding to the number of formants is necessarily better for speaker discrimination, dynamic measurements are more fruitful than midpoint measurements, and that some formant analysis processes might be "better" than others. Additionally, there is little research to support the notion that higher formants are necessarily better speaker discriminators.

Overall, it seems to be difficult to characterise the performance of formants in speaker discrimination tasks, given the wide range of results presented in the forensic speech science research literature. It is acknowledged that as this is a meta-analysis, all of the studies had varying factors, some of which have been explored and commented on in this paper. Others that have not been commented on in this initial phase of the meta-analysis include the role played by specific vocalic segments. This naturally hinders the ability to directly compare results from all of the included papers. Even so, in view of the wide-ranging results in this meta-analysis and ongoing discussion regarding the obscurities around formants in phonetic research [9], it may be appropriate to review the importance placed on formants in FVC casework. This is not to say that formants should not be analysed in FVC casework at all. Rather, a measured approach to their inclusion should be taken.

6. ACKNOWLEDGEMENTS

This work was supported by the Economic and Social Research Council (ESRC) [grant number 2385933].

7. REFERENCES

[1] E. Gold and P. French, 'International Practices in Forensic Speaker Comparison', *International Journal of Speech, Language and the Law*, vol. 18, no. 2, pp. 293–307, Nov. 2011, doi: 10.1558/ijsll.v18i2.293.

[2] F. Nolan, 'The "telephone effect" on formants: a response', *International Journal of Speech, Language and the Law*, vol. 9, no. 1, pp. 74–82, Mar. 2002, doi: 10.1558/ijsll.v9i1.74.

[3] P. Ladefoged and K. Johnson, *A Course in Phonetics*, 6th ed. Wadsworth, Cengage Learning, 2011.

[4] K. N. Stevens, C. E. Williams, J. R. Carbonell, and B. Woods, 'Speaker Authentication and Identification: A Comparison of Spectrographic and Auditory Presentations of Speech Material', *The Journal of the Acoustical Society of America*, vol. 44, no. 6, pp. 1596–1607, Dec. 1968, doi: 10.1121/1.1911302.

[5] P. Rose, *Forensic Speaker Identification*, 1st ed. Taylor Francis, 2002.

[6] C. Byrne and P. Foulkes, 'The "Mobile Phone Effect" on vowel formants', *International Journal of Speech, Language and the Law*, vol. 11, no. 1, pp. 83–102, Mar. 2004, doi: 10.1558/ijsll.v11i1.83.

[7] K. McDougall, 'Speaker-characterising properties of formant dynamics: A case study', *Proceedings of the 9th Australian International Conference on Speech Science & Technology Melbourne, December 2 to 5, 2002.*, pp. 403–408, 2002.

[8] F. Nolan and C. Grigoras, 'A case for formant analysis in forensic speaker identification', *International Journal of Speech, Language and the Law*, vol. 12, no. 2, pp. 143–173, Aug. 2005, doi: 10.1558/sll.2005.12.2.143.

[9] D. H. Whalen, W.R. Chen, C. H. Shadle, and S. A. Fulop, 'Formants are easy to measure; resonances, not so much: Lessons from Klatt (1986)', *The Journal of the Acoustical Society of America*, vol. 152, no. 2, pp. 933–941, Aug. 2022, doi: 10.1121/10.0013410.

[10] T. Kendall and V. Fridland, *Sociophonetics*. Cambridge, UK: Cambridge University Press, 2021. doi: 10.1017/9781316809709.

[11] M. Jessen, 'Speaker profiling and forensic voice comparison: The auditory-acoustic approach', in *The Routledge Handbook of Forensic Linguistics*, 2nd ed., M. Coulthard, A. May, and R. Sousa-Silva, Eds. London: Routledge, 2020, pp. 382–399.

[12] M. Jessen, *MAP Adaptation Characteristics in Forensic Long-Term Formant Analysis*. 2021, p. 415. doi: 10.21437/Interspeech.2021-1697.

[13] T. Cambier-Langeveld, 'Current methods in forensic speaker identification: Results of a collaborative exercise', *International Journal of Speech, Language and the Law*, vol. 14, no. 2, pp. 223–243, Mar. 2008, doi: 10.1558/ijsll.2007.14.2.223.

[14] R v O'Doherty [2002] NICA 20.

[15] M. Jessen, 'Forensic voice comparison', in *Handbook of Communication in the Legal Sphere*, J. Visconti, Ed. De Gruyter, 2018, pp. 169–200. doi: 10.1515/9781614514664-010.

[16] F. Nolan, K. McDougall, G. De Jong, and T. Hudson, 'A Forensic Phonetic Study of "Dynamic" Sources of Variability in Speech: The DyViS Project', *Proceedings of the 11th Australian International Conference on Speech Science & Technology*, Dec. 2006, [Online].