The Variability of Vowels' Formants in Forensic Speech

Sonia Cenceschi, Chiara Meluzzi, and Alessandro Trivilini

Speech analysis plays a pivotal role in the exploitation of forensic recordings in resolving a wide range of questions. Although this topic may include a large set of methodologies based on varied digital features (MFCC, Centroid, Harmonicity, VOT, etc., cf. [1]), this work approaches the theme from a phonetic perspective, taking into consideration the vowel formants, and focusing on their variability and difficulty of measurement. Formants correspond to the resonant frequencies of the vocal tract and are, therefore, sensible to specific speaker-related variations such as age and sex. In this respect, formants' variation contributes to characterizing the subjective timbre of the person. For this reason, formants' values are largely used in forensics [2], with all the practical problems that come with them, and in particular when dealing with speaker recognition or discrimination [3], [4]. Indeed, formants' values correspond to specific frequencies of the sound signal and are usually reported in Hertz. They are, however, affected by numerous internal and external variables, so that although on average they are characteristic of the individual speaker, they always vary within frequency bands that cannot be defined in absolute terms [5]. How, then, is it possible to provide reliable answers to forensic questions? In essence, it is up to the specialist to determine if there are the conditions to carry out an analysis, and to understand, for example, whether the differences between formants' values could be ascribed to two different speakers or the difference is too subtle to justify this claim. What should be measurable today, and must be defined at the level of jurisprudence, is therefore the professionalism of the expert. However, this is still heterogeneous in different countries, although several codes of practice such as the International Association for Forensic Phonetics or engineering society ones have been validated.

With these premises, the focus of this work is to provide an overview of the instability of vocal parameters and a methodological proposal for formants analysis to be applied in forensics, through the open source software Visible Vowels [6]. The paper is organized as follows: in the first section, we describe the importance of formants' values in performing forensic reports; in the second section, we deal with speech variability from both an acoustic engineering and a sociophonetic point of view, by emphasizing intra- and extra-speaker variability; in the third section, we present a possible use of the free online tool *Visible Vowels* for forensic purposes; and finally, we offer some (still preliminary) conclusions in the fourth section.

Formants and Forensic Research

The formants are one of the milestones in the forensic analysis of spoken language, and they are mainly used in the speaker's identification and recognition [7], for describing the accent variation across dialects, and sociolinguistic variability. In particular for speaker recognition, the methods used in practice are very heterogeneous [8]. For this reason, and because of the extreme variability of the sources for quality and context of the recordings, the analysis is done, when possible, by applying different methodologies in parallel, including: semi-automatic analysis (with software such as Nuance solution or *iVocalise* based on spectral and phonetic features), semi-automatic phonetic analysis (mainly based on formants extraction with script), and perceptive (through tests aimed at a large set of carefully chosen listeners). Formants are the most used phonetic parameters for the first two sets of methods, but, to date, the main issues in forensic speech comparison remain the measurement of their variability and validation of the results' reliability. The expert's intellectual processes are always mandatory in the evaluation of the availability of the data (e.g., in relation to noise, quantity of material, modality of speech), the best model, and/or the application and choice of statistical methods to make the results usable and useful in forensics.

Speech Variability as Reflected through Vowels' Formants

Human speech is extremely variable at the suprasegmental (prosodic) level, and this variability is gradually reflected in the segmental behavior of its components. Starting from a general perspective, the variables that influence the sentence's prosody of a single speaker are splitted into two main groups [9]: one composed of fully defined characteristics directly related to the dialogic dimension (e.g., motivational state, rhetorical form,

1094-6969/21/$25.00©2021IEEE

emotions, etc.), and another that takes into account dimensions that exist regardless of the presence of interaction (e.g., language or dialect, social context, etc.). In this regard, it is possible to understand how speech analyses are (still) extremely difficult to generalize, if not at the cost of delimiting the observed sample. This is not possible in forensics, where the recordings are provided by third parties, and cannot be defined *a priori*. Moreover, the prosodic behavior is reflected on the phonetic level, and precisely on formants and other acoustic features used in investigations, so that introducing complementary and statistical analysis therefore becomes necessary.

The issue of variability at the phonetic level has been recently fully investigated by a new field of linguistic research known as sociophonetics. By combining the methodological and theoretical considerations of previous sociolinguistics, sociophonetics has focused on the variation of acoustic features in the production of speech as a means to signal not only the speaker's identity but also his/her orientation towards the hearer or the emotions attached to a certain topic. Thus, when explored from a sociophonetic perspective, the variability is not intended as a simple noise in the data, but as a resource used by speakers in everyday interactions, most of the time without being aware of it. In this respect, phonetic cues provide social meaning.

One of the principal phonetic variables used in sociophonetic research, as well as in previous sociolinguistic and dialectological inquiries, are vowels. In particular, measures of vowel quality (including the vowel's height, advancement and lip rounding) and duration have been approached differently and with various tools. Vowels' formants remain the main clues for analyzing vowels' variation. The first formant (F1) refers to the tongue's height along a high-mid-low continuum, and presents in an inversely proportional way: that is, vowel [a] is pronounced with the lowest tongue height and shows the highest F1 values. The second formant (F2) refers to the advancement of the tongue, by distinguishing between front and back vowels (e.g., [i] *vs*. [u]) with the more front vowels having the highest F2 values. Lip rounding is usually associated with a variation in both F2 and F3 (i.e., the third formant) values, and this is particularly helpful for the study of those languages (e.g., English) in which the opposition between rounded and unrounded vowels is also phonological (i.e., it is a means for distinguishing between different words). As a side note, it is possible to highlight that phoneticians used to transcribe between square brackets speech sounds as they were actually pronounced by the speakers. Phonetic notation could be very precise and refers to a set of characters and symbols collected in the International Phonetic Alphabet (IPA, cf. www.internationalphoneticalphabet.org). It should be noted, however, that a very accurate phonetic notation is not easily readable by non-experts, therefore being quite unusable outside its specific field of research (e.g., for forensic purposes).

Intra- and Inter-speaker Variability

Since formants' values correspond to specific frequencies in the sound signal, they are highly variable according to speakers' specific physical attributes: for instance, women and children notoriously have higher formants' values with respect to men. This means that speaker's specific characteristics, including both biological and social features, play a role in shaping formants' variability in speech, also across different speaking situations (for an overview, cf. [10]). A way to solve this inter-speaker problem is through formants' normalization with various formulas (e.g., by transforming hertz values in bark). However, the technician must be warned against the use of normalization procedures, especially in small speech samples, because they tend to squeeze variability, thus making it unusable for sociophonetic analysis. A recent work has also demonstrated how bark normalization drastically reduces within-speaker's semi-automatic recognition in a possible forensic setting [11].

Furthermore, vowels' formants could be measured with different techniques and at different points in time. A main difference is between static and dynamic approaches to formants' measurements: the first centers on formants' values as extracted at vowels' midpoints, whereas the latter focuses on the variation of F1 and F2 values through 5 to 7 timepoints during vowels' durations [10].

Another perspective considers not only the single formant's variation, but the whole vowel space as created by all the vowels pronounced by a speaker (or, at least, the cardinal ones). Vowel space could be measured through Euclidean distances but also by using F1-F2 co-variation as cardinal points on a Cartesian plane, mapping the movements of vowels in a bi-dimensional space. As we will see, there are tools that help in the visual representation of this variation, thus being a resource also for forensic purposes.

Signal Variability

The variability of the signal (and consequently of the phonetic parameters) is also influenced by the quality and context of the recording, which in the forensic field can change a lot according to the environment of the recording that can range from a crowded restaurant to a silent street. Moreover, the analysis of vowels' formants variation across recording modalities shows that formants are also modified by the compression format [11].

Comparing the voices of possibly the same speaker from different recording devices for forensic purposes is possible [12], but a qualitative analysis must be combined with a quantitative one.

Qualitative and Quantitative Analysis

In forensic settings, reliability of the results is particularly relevant because of the possible juridical consequences of the reports provided by different experts. In voice comparison, the most accredited method for statistical elaboration of formants' values is based on a Bayes factor, obtained as a likelihood ratio of the marginal likelihood of two competing hypotheses (usually a null and an alternative) in order to obtain a discrimination score. However, it requires extrapolation of a high number of vowels according to their type, and this is often not possible for forensic recordings. Indeed, forensic reports are often based on small speech samples, up to 10 seconds or less, to be compared across different modalities and usually with lots of background noise. Due to these conditions, a proper phonetic investigation through formants' automatic extraction and further analysis are impossible for different reasons. Firstly, samples could be recorded in low quality format (e.g., mp3), thus compressing frequency ranges and making a comparison between low-quality and high-quality recordings impossible to perform. Secondly, due to the small dimension of speech samples it could be the case that very few tokens are available for analyzing vowels, and as a result, it may be impossible to confront the same vowels across two samples (e.g., for identifying speakers).

Since this is quite often the working scenario in forensic settings, it appears that it is rather impossible to provide quantitatively-based reports. Although the use of inferential statistics generally provides reliability, in these cases we lack the fundamental premises to perform a statistical analysis on our data. Furthermore, when comparing audio samples across recording modalities, semi-automatic analysis could lead to misinterpretation of data.

Qualitative analysis also must be performed with caution and by testing some basic prerequisites (e.g., the audio quality), but they could be preferred in order to avoid sampling issues. We intend to conduct qualitative analysis comparisons between audio samples also based on the extraction and the evaluation of vowels' formants, if possible, but integrated with visual inspection of the spectrograms. Furthermore, in qualitative paradigms the selection of the examples must be very well-defined. For example, in the case of vowels' formants, it would be better to select cardinal vowels, or conversely, those vowels that are known from the literature to have a high degree of geographic variability (e.g., middle-vowels in Italian, or the GOOSE vowel in many English dialects). By accurately selecting the target vowels, even with few tokens it will be possible to provide a more reliable report with respect to both within-speaker identification or sociophonetic profiling.

Some Tools to Visually Represent Variability

After providing reliable analysis on our (often small) audio samples, another issue arises, that is the visual representation of those data. This is also the case in proper phonetic analysis, but it assumes a different importance in forensic settings: indeed, it is important that the expert's report is clear and understandable by the judge and the attorney. In this respect, the use of a specific vocabulary could lead to possible misunderstanding and making the lawyers underestimate or overestimate the results of a phonetic comparison based on such a small and variable element like vowels' formants. Therefore, it is important for the expert to offer not only reliable results, both qualitatively and quantitatively, but also to present them in a visually clear way. One possible tool for visually inspecting vowels' formants' variation is *Visible Vowels*. The tool has been developed as part of a huge sociophonetic project on Dutch variation in Frisia. It is free and allows standardized formants' values to be computed through different formulas. It also offers the possibility to visualize both the single formant's variation in both static and dynamic analyses and also to investigate vowel space. To our knowledge, the application of this tool in forensic reports has been firstly proposed by the second author of this paper (Cenceschi et al., to appear), and implemented also with the other authors in different work cases reports. Fig. 1 offers an example of vowels' comparison made through the online tool Visible Vowels.

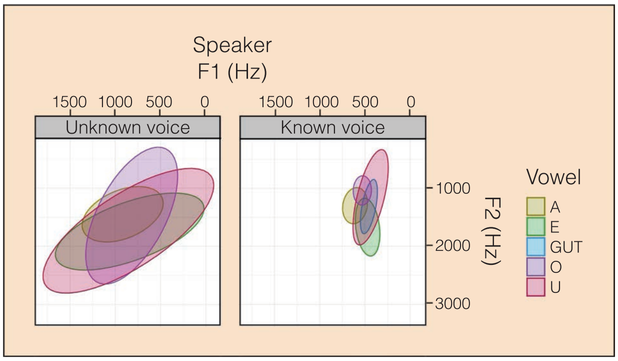

Fig. 1. An example of visualization on the webtool Visible Vowels for comparing a known and unknown voice.

In this case, the experts have to provide evidence that will differentiate between a known voice (KV) and an unknown one (UV). To put it simply, the expert has to provide the lawyers the proof of whether UV and KV could be the same speaker. Alongside with other measurements, one piece of evidence was constituted by the different formants' values between KV' and UV' vowels. The use of a visual representation through Visible Vowels helped the experts to provide evidence for their claim that the two voices belong to different speakers.

However, at the present state of the research, there is not a set of reference values of formants' within-speaker variation, not to mention the fundamental between-speaker variation. Different scholars have provided reference values for the phonological vowels of their language, but these values are based on a prototypical adult male voice. Thus, when performing a forensic comparison on vowels' formants, we are in definitional uncertainty, and only experience in the field, statistical analysis (whether applicable), and an in-depth knowledge of the (socio)phonetics literature on that particular language could help to avoid terrible misinterpretation of the data.

Conclusions: Good Practice for Forensic Analysis

Among the different measurements available for forensic sound comparison, vowels' formants are extremely reliable in providing a good measure of both intra- and inter-speaker variability. Although different techniques have been developed in sociophonetics (e.g., static *vs*. dynamic analysis, vowel space, etc.), in forensic analysis the speech sample is not wide enough to allow for a quantitative analysis which must be associated with a qualitative one. Furthermore, the lack of reference values for formants' variability makes it extremely difficult to address inter-speaker comparison, especially if the audio samples to be compared are extremely compromised, as it usually happens in forensic settings. For these reasons, substantial precaution in linguistic and phonetic analysis is essential.

Forensic issues must be addressed by acknowledging that many measurements we provide assume a theoretical approach at the interface between applied linguistics, phonetics, and sound engineering. In this paper, we limited our examples to phonetics, a field in which it is quite easy to need collaborations with psychiatrists, computer engineers, etc. The single expert must surely insist on an active collaboration with the police and law enforcement, without closing in on its technological domain but rather working on the scientific communication of methodologies and results.

References

[1] H. G. Kim, E. Berdahl, N. Moreau, and T. Sikora, "Speaker recognition using MPEG-7 descriptors," in *Proc. 8th European Conf. Speech Commun. Technol. (Eurospeech),* 2003.

[2] M. Jessen, "Forensic phonetics," *Language and Linguistics Compass*, vol. 2, no. 4, pp. 671-711, 2008.

[3] H. Hollien, R. H. Bahr, and J. D. Harnsberger, "Issues in forensic voice," *J. Voice*, vol. 28, no. 2, pp. 170-184, 2014.

[4] E. Gold and P. French, "International practices in forensic speaker comparisons: second survey," *Int. J. Speech Language and the Law*, vol. 26, no. 1, pp. 1-20, 2019.

[5] P. Harrison, "Variability of formant measurements," submitted in partial fulfilment of the degree of M.A. with the Department of Language and Linguistic Science, University of York, UK, 2004.

[6] W. Heeringa and H. Van de Velde, "Visible vowels: a tool for the visualization of vowel variation," in *Proc. Common Language Resources and Technology Infrastructure (CLARIN) Ann. Conf.*, pp. 8-10, 2018.

[7] P. Rose, *Forensic Speaker Identification*. London, UK: Taylor and Francis, 2002.

[8] H. J. Künzel, "Current approaches to forensic speaker recognition," in *Proc. ESCA Workshop on Automatic Speaker Recognition, Identification and Verification*, pp. 135-142, 1994.

[9] S. Cenceschi, "Speech analysis for automatic prosody recognition," thesis submitted for the Ph.D. degree with the Department of Computer Engineering at the Politecnico di Milano, Italy, 2019.

[10] E. Thomas, *Sociophonetics: An Introduction*. Hampshire, UK: Macmillan International Higher Education, 2010.

[11] S. Cenceschi, C. Meluzzi, and N. Nese, (to appear), "Speaker's identification across recording modalities: a preliminary phonetic experiment," AISV 7 studies, Milano: Officina Ventuno.

[12] J. P. Campbell, W. Shen, W. M. Campbell, R. Schwartz, J. Bonastre, and D. Matrouf, "Forensic speaker recognition," *IEEE Signal Proc. Mag.*, vol. 26, no. 2, pp. 95-103, 2009.

Sonia Cenceschi (sonia.cenceschi@supsi.ch) is a Researcher at the Digital Forensic Service of the Department of Innovative Technologies of the University of Applied Sciences of Southern Switzerland (SUPSI). She has a Ph.D. degree in computer engineering from Politecnico di Milano, Italy and has been working for more than ten years as audio expert for the criminal field. Her specific research interests include audio forensics, digital speech analysis, and speech prosody.

Chiara Meluzzi (chiara.meluzzi@unipv.it) is a Postdoctoral Fellow in Sociophonetics at University of Pavia, Italy, where she also teaches sociolinguistics and phonetics. She obtained her Ph.D. degree in 2014 with an investigation on the variability in the Italian of Bolzano, and ever since, she has worked on speech variability from a phonetic perspective. She has studied forensic comparisons of voices and in the semi-automatic creation and annotation of speech corpora. Chiara's current research areas include applied linguistics, sociophonetics, and forensic linguistics.

Alessandro Trivilini (alessandro.trivilini@supsi.ch) is Head of the Digital Forensics Lab of the University of Applied Sciences and Arts of Southern Switzerland where he has been a Lecturer and Researcher in software engineering, digital forensics, and cybersecurity since 2003. The lab was nominated by the Swiss Federal Court as official digital forensics partner in January 2019. Alessandro's current research interests are digital security, data protection, and forensic digital investigations. He obtained the Ph.D. degree in computer science (AI/NLP) at the Politecnico di Milano, Italy.