INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

A comparison of the accuracy of Dissen and Keshet's (2016) DeepFormants and traditional LPC methods for semi-automatic speaker recognition

Thomas Coy, Vincent Hughes, Philip Harrison, Amelia Gully

Department of Language and Linguistic Science, University of York. UK
tomalexcoy@gmail.com, {vincent.hughes|philip.harrison|amelia.gully}@york.ac.uk

Abstract

There is a growing trend in the field of forensic speech science towards integrating the vanguard of speech technology with traditional linguistic methods in pursuit of both scalable (i.e. automatable) and accurate evidential methods. To this end, this paper investigates DeepFormants, a DNN formant estimator which its creators, Dissen and Keshet [1], claim constitutes an accurate tool ready for use by linguists. In the present paper, DeepFormants is integrated into semi-automatic speaker recognition systems using long-term formant distributions and compared against systems using traditional linear predictive coding. The readiness of the tool is assessed on overall speaker recognition performance, measured using equal error rates (EER) and the log LR cost functions (C_{llr}). In high-quality conditions, DeepFormants outperforms the best performing LPC systems. Much poorer overall performance is found in channel mismatch conditions for DeepFormants, suggesting it is not adaptable to conditions it was not originally trained on. However, this is also true of LPC methods, raising questions over the validity of using formant analysis at all in such cases. A major benefit of DeepFormants over LPC is that the analyst does not need to specify settings. We discuss the implications of this with regard to results for individual speakers.

Index Terms: Formant Analysis; DNN, Semi-Automatic Speaker Recognition; Long-Term Formant Distribution; Forensic Voice Comparison

1. Introduction

1.1. Interpretability in Forensics

Forensic Voice Comparison (FVC) makes up the largest proportion of work done by forensic speech scientists. FVC consists of analytical comparison between a known speaker's reference sample (e.g. a police interview) and an unknown, commonly criminal, voice (e.g. a threatening phone call). The speaker-specificity of vowel formants is well documented (e.g. [2,3,4]), and as such formant analysis has been a de facto norm for FVC in the UK since the 2002 appeal case R-v-O'Doherty [5] as well as in other jurisdictions around the world. In the courtroom, the expert witness must make their findings understandable to the triers-of-fact who are not trained in linguistics. This makes the interpretability of data particularly important should they come under scrutiny. For this reason, amongst others, automatic approaches to FVC have not been well received in England and Wales. In R-v-Slade & Ors [6], the appeal court declined to make a decision on the admissibility of ASR (automatic speaker recognition) evidence, in part, due to the complex and non-linear nature of their input features (typically cepstral coefficients) which "left a lot of questions unanswered". Formants are more pervasive throughout FVC [7] as they can be more readily explained to a lay audience as directly measurable frequencies that map onto articulation in a way that cepstral coefficients cannot. The trend in recent years to try and combine linguistic and automatic systems, ideally integrating the best from both, presents a potential solution: semi-automatic speaker recognition (SASR) systems which combine linguistically interpretable input features with automatic modelling, scoring and evaluation to vastly increase workload potential.

1.2. Long-Term Formant Distributions and LPC

Perhaps the most common form of SASR takes long-term formant distributions (LTFD) as input, as they are well suited to automatic scoring and evaluation. LFTDs were first surmised by [8] and developed further in [9] with the inclusion of bandwidth information. [9] and [10] showed that adding deltas generally improves performance and the study in [11] supports their use in forensics. However, when using the current state-of-the-art for formant estimation, linear predictive coding (LPC), accurate measurement is highly contingent on LPC order. LPC estimates for a given speaker can vary drastically whether as the result of applying fixed settings [12] or inter-analyst inconsistency [13]. Consequently, the results that an SASR system produces for individual comparisons can be considerably unstable and unpredictable across systems as a result of differing LPC order settings, as [14] revealed.

Current industry standard software such as Praat [15] and WaveSurfer [16] use LPC for formant estimation. Empirical testing shows that optimal accuracy is achieved when using settings on a speaker- and vowel-specific basis [17]. This is, however, a key obstacle to automation. SASR systems generally apply fixed settings to all tokens for all speakers, thus providing a potential glut of relatively inaccurate estimations [18]. [19] provide a potential "step toward an unsupervised model" which moves away from "the best estimate" to the "most likely estimate" by running through LPC orders to produce 189 estimates per token. While this addresses the issue of fixed settings, there remains a tremendous number of estimates through which the practitioner must sift in order to obtain accurate results.

1.3. DeepFormants

Given the importance of formant estimation in FVC (and in other disciplines of linguistics; e.g. phonetics and sociophonetics), it is paramount that the best tools are found. Dissen and Keshet's [1] DeepFormants has obvious potential: a standard feedforward DNN in which LPC orders, among other processes, are selected based on what was learned in training. The network was trained and tested using read American English from the VTR corpus [20] which comprises hand-corrected formant measurements of speech originating in TIMIT [21]. VTR contains a set of broadband recordings from

http://dx.doi.org/10.21437/Interspeech.2021-1487

eight American dialect regions, where each of 346 speakers (173 male, 173 female) contribute one "phonetically-compact" and one "phonetically diverse" sentence [20, p.2]. As noted in [18], a genuine ground truth formant value is a somewhat moot concept outside of hyper-controlled conditions, a fact comprehensively explored with the use of synthesized speech in [17]. However, the VTR corpus represents the closest to some kind of ground truth that is available with natural speech. Based on their results, Dissen and Keshet subsequently write that it is "doubtful that higher accuracy can be achieved with automated tools seeing as manual annotation cannot" [1, p.960], thus showing great promise as a scalable formant estimator for real-world practitioners. In theory, DeepFormants automates the choice of token-specific LPC orders, removing both inter-analyst inconsistency and the surfeit of potentially highly inaccurate and redundant estimates.

1.4. This Study

Since DeepFormants was already tested for the accuracy of estimates themselves in [1], rather than analyse raw formant data, the present study evaluates the performance of LTFD-based SASRs that use DeepFormants for feature extraction, for which speaker identity is a verifiable ground truth. To do this, we evaluate not only overall system validity, but also the results for individuals within the system in terms of the strength of the evidence that the systems produce. These results are compared with identical systems using traditional LPC derived formants, extracted via the Snack Sound Toolkit [22]. This study builds directly on work done in [10], [12] and [14], which investigated the stability of speaker scores in channel mismatch and the effects of LPC order on system performance. This study will also assess the performance of DeepFormants in mismatched channel conditions, namely landline and simulated mobile telephone transmission, both of which represent very commonly encountered audio in FVC. Previous work in [12] suggests that system performance will deteriorate with mismatch. This is particularly expected in this study in GSM-emulated conditions as they present DeepFormants with new frequency ranges to which it has not previously been exposed. Furthermore, its adaptability to a new linguistic variety and style will be observed as DeepFormants was trained on read American English but is here tested with spontaneous Southern Standard British English.

By using a state-of-the-art DNN to extract linguistic features this study is another step in ongoing efforts to take automatic methods from the vanguard of speech technology and integrate them with linguistic-based methods.

2. Method

2.1. Materials

The same recordings as in [12] were used, which included a total of 97 speakers of Southern Standard British English from the Dynamic Variability in Speech (DyViS) [23] corpus. The recordings comprised of a high-quality (HQ) mock police interview recording (Task 1) and a spontaneous telephone conversation (Task 2), recorded both near-end via a microphone (HQ) and far-end post-transmission (TEL). These tasks were recorded a few hours apart on the same day, with durations of between 9 and 30 minutes. GSM mobile simulations, using the near-end HQ recordings of Task 2, were also used. These had been produced previously in [14] by bandpass filtering the recordings between 300-3400 Hz before using the AMR Speech Codec Platform [24] to produce high and low bitrate recordings: GSM$_{HQ}$ – 12.2 kbps, GSM$_{LQ}$ – 4.75 kbps.

2.2. Feature extraction and pre-processing

The process of feature extraction is described in more detail in [10], [12] and [14]. The *vadsohn* function in the VOICEBOX toolbox [25] performed voice activity detection, for which silence was defined as at least 100ms of adjacent non-speech frames. StkCV [26] was used to segment speech into vowels and consonants and extract timestamps for their boundaries. A subset of the segmentations produced by StkCV were manually inspected to assess accuracy. It was found that the approach was generally over conservative, removing vocalic material often at the start and end of segments. However, reassuringly this bias meant that the script rarely (if ever) classified consonants as vowels. Only the first 60 seconds of vowel material from each recording was used. This removed three speakers with insufficient vowel material, providing the final total of 97 speakers. At the boundary of each segment, three frames were deleted to ensure delta values do not erroneously provide information about the change in frequency between originally nonadjacent segments [27]. In Snack, F1-F4 estimates were extracted using 20 ms window-lengths with 10 ms shift (50% overlap). Formant estimates using combinations of fixed LPC orders (12-16) and maximum numbers of formants (3-4) were already available from [14], generating 25 sets of estimates per channel condition. Deltas were also appended in MATLAB.

F1-F4 estimates were also extracted using DeepFormants from precisely the same 20ms frames across the 60s of vowels per speaker. After resampling to 16 kHz, for each estimate 30 LPCCs are generated by looping through LPC orders 8 to 17 and 50 discrete cosine transform (DCT) coefficients are derived from the spectrum. These 350 coefficients are then fed into the prediction model: a multilayer-perceptron which was trained on the VTR corpus [20]. Estimates were exported to MATLAB wherein deltas were calculated and frames timestamped. For further details on the DeepFormants architecture, see [20].

2.3. Scoring and Conditions

Task 1 was always used as the nominal suspect recording, and the various Task 2 recordings as the nominal offender sample. Likelihood ratio (LR)-like scores were obtained using the GMM-UBM approach [28], implemented with the Microsoft Speaker Recognition Identity Toolbox [29]. GMM-UBM was selected over state-of-the-art i-vectors or x-vectors because 1) the relationship between input and output is more transparently intelligible due to the relatively small amount of data manipulation and 2) GMM-UBM requires considerably less data processing and is therefore much more time-efficient. The speakers were split into three groups of 32 reference, 33 training and 32 test speakers, providing 32 same speaker scores and 496 different speaker scores per evaluation. Scores from the test data were calibrated using a logistic-regression model trained on scores from the training data to produce calibrated $\log_{10}$ LRs (LLRs). System performance based on the calibrated LRs was assessed using equal error rate (EER) and log likelihood ratio cost (C_{llr}). For both metrics better performance corresponds to numbers closer to zero.

3. Results

3.1. Overall Results

Figure 1 summarises and compares DeepFormants overall system performance with the best and worst performing Snack systems, for which LPC orders are listed in Table 1. The best and worst snack systems were determined based on both the EER and C_{llr} of the 25 systems (per condition) previously produced in [12].

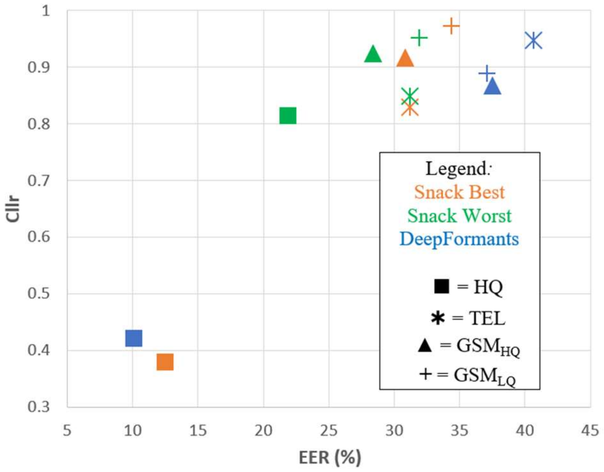

Figure 1: *Scatterplot comparing DeepFormants and Snack results across channel conditions.*

In the HQ condition, DeepFormants was comparable with the best performing Snack system. Indeed only DeepFormants HQ (EER = 10.13%, C_{llr} = 0.4210) and the best HQ Snack system (EER = 12.50%, C_{llr} = 0.3790) performed well at all, with DeepFormants producing the overall lowest EER and only a marginally higher C_{llr} than the best Snack system. Importantly, however, it must be highlighted that each of the worst Snack systems performed very poorly, even in the HQ condition (EER = 21.93%, C_{llr} = 0.8137). In fact, only the best Snack HQ system produced a C_{llr} of less than 0.8. This highlights the potential for vast variability and serious degradation in system performance as a function of the LPC settings chosen. For both Snack and DeepFormants, performance was considerably worse in the channel degraded and mismatch conditions. As reported in [12], the ordering of performance by channel is predictable, with TEL generally performing better than GSM_{HQ}, followed by GSM_{LQ}.

Table 1: *LPC order for each Snack condition.*

Channel Condition	Previous Performance	Offender LPC	Suspect LPC
HQ	Best	14	14
	Worst	12	16
TEL	Best	16	16
	Worst	15	12
GSM_{HQ}	Best	12	14
	Worst	16	12
GSM_{LQ}	Best	13	16
	Worst	16	12

In these three conditions, DeepFormants performed similarly poorly (TEL = 40.68%, GSM_{HQ} = 37.50%, GSM_{LQ} = 37.10%) and consistently worse than even the worst performing Snack systems in terms of EER (TEL = 31.25%, GSM_{HQ} = 30.85%, GSM_{LQ} = 34.43%). However, all of the C_{llr} values were close to 1, indicating that none of the systems were able to capture much speaker-specific information and so would, in reality, be of little use in a forensic case.

3.2. Individual Analysis

As well as considering overall performance, it is important to assess the effects of different formant measurement methods on LLRs for individual speakers. Given that performance was so poor in all non-HQ conditions, only LLRs generated in the HQ condition are discussed here. Individuals' LLRs varied considerably across systems and the nature of that variability was itself inconsistent and unpredictable. Figure 2 shows three speakers' mean same speaker (SS) and different speaker (DS) LLRs for each of the 25 HQ Snack systems relative to DeepFormants. These speakers were chosen specifically to exemplify the different patterns found across the dataset. Plots

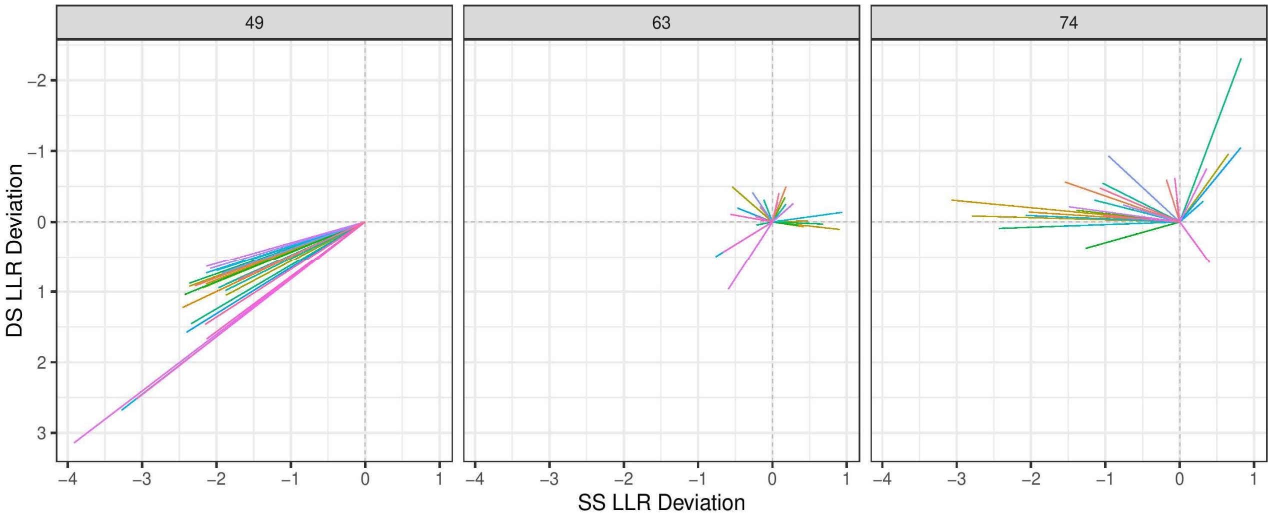

Figure 2: *Plots comparing mean same speaker and DS LLRs for speakers 49, 63 and 74 across all 25 HQ systems. DeepFormants scores are the origin and all lines represent the relative differences in Snack systems*

for all 32 test speakers can also be found at: https://vincehughes.files.wordpress.com/2021/03/deepformantsalldeviations-1.pdf. The speaker's performance in the DeepFormants system is the origin (0,0) and the relative performance of all the Snack systems are represented by the coloured lines as deviations from the origin. The closer the grouping of the lines around (0,0) the less variability there is across the different Snack systems relative to DeepFormants.

Speaker #49 (for reference, speaker numbers here refer to the identifying numbers given to speakers within the DyViS database) performs unambiguously better in the DeepFormants system, with stronger same speaker and different speaker LLRs compared with any Snack system. This is indicated by the fact that all lines move towards the bottom left with higher different speaker LLRs and lower same speaker LLRs. For some Snack systems, mean LLRs are as much as 4 orders of magnitude stronger using DeepFormants. Speaker #74, conversely, shows improvement on the DeepFormants system in a few Snack systems, however in many more there are significantly weaker same speaker LLRs. Lastly, speaker #63 remains, relatively speaking, fairly consistent across all conditions. None of the 32 test speakers performed markedly and consistently better in Snack systems than they did in the DeepFormants system.

4. Discussion

It was predicted that DeepFormants would produce inaccuracies when attempting to measure formants from materials that did not match its original training data. Positively, DeepFormants did not struggle with the newly encountered British English dialect and spontaneous speech style, as evidenced by it producing the best EER and a low C_{llr} in the HQ condition; at least this is the case based on SASR system performance rather than analysis of the raw formant data itself. That this is only true in HQ shows that the DeepFormants systems remain highly sensitive to channel mismatch, however this was also the case for all Snack systems which also performed poorly in non-HQ conditions. It seems that all estimation methods fail to produce formant values that are useful when input into an LTFD-based SASR which compares audio across channel conditions. This suggests that it is inappropriate to use exclusively formant-based systems, i.e. based on raw formant data alone, in forensic cases involving channel mismatch and degradation. Since this study focused only on the performance of non-HQ systems in mismatched conditions, the results do not readily clarify whether such poor performance is simply a lack of robustness to channel mismatch or rather a lack of robustness to non-HQ channels more generally.

Nevertheless, the results presented here for the different Snack systems cast into sharp relief the significant impact of LPC order settings on individual LLRs as well as overall system performance. It is also important to highlight that the variability in performance, either overall system performance or individual LLRs, as a function of LPC settings was in no way predictable. Some combinations of LPC order happened to produce better or worse overall performance. This is the major advantage of a DeepFormants system over, for example, a Snack system: one need not select an LPC order since the multi-layer perceptron combines information from all orders based on its training experience. Unlike [12], wherein all possible LPC order combinations were run and the best performing system taken as the most appropriate based on prior knowledge of the ground truth, employing DeepFormants entirely removes this otherwise highly arduous and imprecise process by automating it and producing a single result. DeepFormants is therefore the logical next step beyond what Kendall and Vaughn [19] achieved with their Praat simulation. As they put it 'although certain combinations of settings are clearly "wrong", there is no such thing as a priori "correct" or "best" settings' [19, p. 7] and DeepFormants appears to successfully learn which are 'clearly wrong' in training.

Individual LLRs reveal that no speakers performed substantially worse in the DeepFormants system than in Snack systems, and importantly that LPC order settings have no uniform, predictable effect across individuals. For example, while speaker #74 showed the potential for a small improvement in a few Snack systems, most LPC settings provide much worse same speaker LLRs (i.e. producing weak or contrary-to-fact evidence). It therefore seems preferable to accept the risk of slightly worse performance for a small minority of speakers using DeepFormants (and, as an analyst, not have to select settings) rather than the risk of poor LPC order selection which could end in the extremely poor performance (and weak or contrary-to-fact LLRs) seen in a great number of Snack systems. As such, DeepFormants provides a much more reliable and convenient front-end input: it provides the strongest discriminatory power as evinced by it producing the lowest EER; it considerably reduces manual workload through automation, making it eminently scalable; and it eliminates the task of selecting LPC orders and thus the potential for serious human error, which in turn avoids the poor performances seen in the worst Snack systems.

Foulkes et al [18, p.2] are correct in observing, in specific reference to DeepFormants, 'that until such techniques are widely available to analysts, LPC remains the state-of-the-art'. However, it seems we are not too far off the emergence of a reliable, deep-learning formant estimator, perhaps in the form of the latest version of DeepFormants [30], the code for which remains unavailable at the time of writing. As and when the training code becomes available for DeepFormants or similar DNNs, it would be pertinent to investigate the impact on system performance with either a specialised model trained for telephony, or a generalised model which incorporates telephony among a broad array of channel conditions. The final step would then be to develop a user-friendly embedding for industry standard software.

5. Conclusions

This study has compared the performance of long-term formant distribution-based SASRs which use traditional LPC methods for formant extraction with systems employing the DeepFormants DNN formant estimator. Findings show that DeepFormants SASRs perform comparably with the better LPC systems in HQ conditions with no discernible consistency in the direction of speaker-specific performance across systems. Such systems thus present a preferable option in skirting the efforts of manually setting LPC orders or producing a surfeit of substandard data. Conversely, poor performance in non-HQ conditions show that manual LPC methods remain the preferred option in those conditions. Since DeepFormants has successfully, and seemingly quite accurately, automated the task of formant estimation, there appears to be great potential for pre-trained DNNs becoming the new state-of-the-art for semi-automatic speaker recognition systems.

6. References

[1] Dissen, Y., & Keshet, J. (2016). Formant estimation and tracking using deep learning. *INTERSPEECH 2016,* 958–962.

[2] Peterson, G. E., & Barney, H. L. (1952). Control Methods Used in a Study of the Vowels. *The Journal of the Acoustical Society of America. 24*(2), 175–184.

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

[4] de Jong, G., McDougall, K., Hudson, T., & Nolan, F. (2007). The speaker discriminating power of sounds undergoing historical change: A formant-based study. *ICPhS XVI*, 1813–1816.

[5] Regina v Anthony O'Doherty (2002) ref: NICB3173 19th April 2002.

[6] Regina v Slade & Ors (2015) EWCA Crim 71 10th February 2015.

[7] Gold, E., & French, P. (2011). International Practices in Forensic Speaker Comparison. *International Journal of Speech Language and the Law, 18*(2), 293–307. https://doi.org/10.1558/ijsll.v18i2.293

[8] Nolan, F., & Grigoras, C. (2005). A case for formant analysis in forensic speaker identification. *International Journal of Speech, Language and the Law, 12*(2), 143–173. https://doi.org/10.1558/sll.2005.12.2.143

[9] Becker, T., Jessen, M., & Grigoras, C. (2008). Forensic Speaker Verification Using Formant Features and Gaussian Mixture Models. *INTERSPEECH 2008*.

[10] Hughes, V., Harrison, P., Foulkes, P., French, P., Kavanagh, C., & San Segundo, E. (2017). Mapping Across Feature Spaces in Forensic Voice Comparison: The Contribution of Auditory-Based Voice Quality to (Semi-)Automatic System Testing. *INTERSPEECH 2017,* 3892–3896. https://doi.org/10.21437/Interspeech.2017-1508.

[11] Gold, E., French, P., & Harrison, P. (2013). Examining long-term formant distributions as a discriminant in forensic speaker comparisons under a likelihood ratio framework. *The Journal of the Acoustical Society of America, 133*(5), 3294–3294. https://doi.org/10.1121/1.4805427.

[12] Hughes, V., Harrison, P. T., Foulkes, P., French, J. P., & Gully, A. J. (2019). Effects of formant settings and channel mismatch on semi-automatic systems in forensic voice comparison. *Proceedings of the International Congress of Phonetic Sciences, August 4-10, Melbourne, Australia, 2019*, 3080–3084.

[13] Zhang, C., Morrison, G. S., Ochoa, F., & Enzinger, E. (2013). Reliability of human-supervised formant-trajectory measurement for forensic voice comparison. *The Journal of the Acoustical Society of America*, *133*, EL54–EL60. https://doi.org/10.1121/1.4773223.

[14] Hughes, V., Harrison, P., Foulkes, P., French, P., Kavanagh, C., & San Segundo, E. (2018). The Individual and the System: Assessing the Stability of the Output of a Semi-automatic Forensic Voice Comparison System. *INTERSPEECH 2018*, 227–231. https://doi.org/10.21437/Interspeech.2018-1649.

[15] Boersma, P., & Weenink, D. (2018). Praat: Doing phonetics by computer [computer program] (6.1.05).

[16] Beskow, J., & Sjölander, K. (2000). Wavesurfer—An open source speech tool. *Proceedings of ICSLP*, 4.

[17] Harrison, P. (2013). *Making accurate formant measurements: An empirical investigation of the influence of the measurement tool, analysis settings and speaker on formant measurement.* University of York PhD dissertation.

[18] Foulkes, P., Docherty, G., Hufnagel, S. S., & Hughes, V. (2018). Three steps forward for predictability. Consideration of methodological robustness, indexical and prosodic factors, and replication in the laboratory. *Linguistics Vanguard 42*(s2). https://doi.org/10.1515/lingvan-2017-0032

[19] Kendall, T., & Vaughn, C. (2019). Exploring vowel formant estimation through simulation-based techniques. *Linguistics Vanguard, 6*(s1). https://doi.org/10.1515/lingvan-2018-0060

[20] Li Deng, Xiaodong Cui, Pruvenok, R., Huang, J., Momen, S., Yanyi Chen, & Alwan, A. (2006). A Database of Vocal Tract Resonance Trajectories for Research in Speech Processing. *IEEE International Conference on Acoustics Speed and Signal Processing Proceedings*, 1, I-369-I–372. https://doi.org/10.1109/ICAsame speakerP.2006.1660034

[21] Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., & Dahlgren, N. L. (1993). DARPA TIMIT: Acoustic-phonetic continuous speech corpus CD-ROM, NIST speech disc 1-1.1 (NIST IR 4930; p. NIST IR 4930). *National Institute of Standards and Technology*. https://doi.org/10.6028/NIST.IR.4930

[22] Sjölander, K. (2006). Snack Sound Toolkit. http://www.speech.kth.se/snack/

[23] 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*(1), 31–57. https://doi.org/10.1558/ijsll.v16i1.31

[24] Alzqhoul, E. A. S., Nair, B. B. T., & Guillemin, B. J. (2014). An alternate approach for investigating the impact of mobile phone technology on speech. Proc. World Congress on Engineering and Computer Science, 1.

[25] Brookes, M. (1997). VOICEBOX: Speech Processing Toolbox for MATLAB. http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html.

[26] Andre-Obrecht, R. (1988). A new statistical approach for automatic speech segmentation. *IEEE Transactions on Asame speakerP 36*, 29–40.

[27] Enzinger, E., Morrison, G. S., & Ochoa, F. (2016). A demonstration of the application of the new paradigm for the evaluation of forensic evidence under conditions reflecting those of a real forensic-voice-comparison case. *Science & Justice, 56*(1), 42–57. https://doi.org/10.1016/j.scijus.2015.06.005

[28] Reynolds, D. A., Quatieri, T. F., & Dunn, R. B. (2000). Speaker Verification Using Adapted Gaussian Mixture Models. *Digital Signal Processing, 10*(1–3), 19–41. https://doi.org/10.1006/dspr.1999.0361

[29] Sadjadi, S. O., Slaney, M., & Heck, L. (2013). *MSR Identity Toolbox v1.0: A MATLAB Toolbox for Speaker Recognition Research.*

[30] Dissen, Y., Goldberger, J., & Keshet, J. (2019). Formant estimation and tracking: A deep learning approach. *The Journal of the Acoustical Society of America*, *145*(2), 642–653. https://doi.org/10.1121/1.5088048