Comparing formant extraction methods according to speaking style and added noise in Forensic Voice Comparison

Dávid Sztahó[1], Attila Fejes[2], György Szaszák[1]

[1] *Budapest University of Technology and Economics, Faculty of Electrical Engineering and Informatics, Department of Telecommunication and Media Informatics*
[2] *Hungarian National University of Public Service*

Abstract

In forensic voice comparison, formant measurements are a "traditional" way of comparing speaker identities. Deep learning may offer a new way of estimating formant values; therefore, it is essential to compare its performance in a forensic way of use. In this study, four formant estimation methods are compared: three based on LPC and one on deep learning. Several aspects of formant modeling in forensic voice comparison were investigated: comparisons according to utterance lengths, speaking styles, and samples corrupted with various types of noise: reverberation and white noise. Results are reported according to C_{llr}, AUC, and EER metrics. It was found that the length of recordings used as suspect samples influences performance to a large extent. Additionally, formant tracking based on deep learning lags behind the other methods in all metrics. Same and different speaking styles also have a measurable effect on performance. Samples corrupted with reverberation do not deteriorate results but white noise does. There are no exact results on which method is better and which is to be used in studies and works. C_{llr} values show that the three LPC-based methods perform similarly. They all make large mistakes when samples are corrupted with white noise. Although the deep learning-based formant extractor performs slightly worse than the other approaches used in this study, it seems to have more resilience to white noise.

1. Introduction

The paradigm shift in forensic sciences and practice (; Saks & Koehler, 2005; Morrison, 2009a, 2011b) enabled the automatic and semi-automatic evaluation of evidence using various modalities and kinds of measurements (such as DNA, fingerprint) (Bazen & Veldhuis, 2004; Matz & Nielsen, 2005). This new paradigm, called the likelihood-ratio (LR) framework, is feasible for processing

Email addresses: sztaho.david@vik.bme.hu (Dávid Sztahó), fejes.attila@nbsz.gov.hu (Attila Fejes), szaszak.gyorgy@vik.bme.hu (György Szaszák)

Beszédtudomány – Speech Science 2022. 7–35. DOI: 10.15775/Besztud.2022.7-35

by computers. Such is the case in voice processing, where speaker recognition techniques are adapted to the requirements of the framework, namely to produce a probability ratio of same and different speaker for evidence (Mandasari et al., 2011; Morrison, 2011a; Kelly et al., 2019).

There are numerous 'traditional' acoustic features used in forensic voice comparison systems, such as fundamental frequency, mel-frequency cepstral coefficients, and intonation-based features (Chaudhary et al., 2017).

Formants, local maxima of the speech spectrum caused by resonance frequencies of the vocal tract, are also traditional features of forensic voice comparison (Titze & Martin, 1998). Although speech sounds, especially vowels, are characterized by their first two formants, formants are also speaker-dependent. Due to the possible overlap of the first formant and higher fundamental frequency, second and third formants are more frequently used in research works (Gargouri et al., 2006; Guillemin & Watson, 2008) although higher formants may also have problematic calculations (attenuation, superimposed noise, exposure to GSM coding).

There are numerous formant tracking applications and methods that are feasible for calculating formant trajectories of voice samples for forensic purposes. Commonly used methods (Kameny et al., 1974; Snell & Milinazzo, 1993) are based on LPC coefficients and direct spectrum envelope estimation, but novel techniques are also emerging that use deep learning for the same purpose (Zhang et al., 2013; Dissen et al., 2019). Although, it is an open question whether deep learning techniques have superficial performance in forensic speaker verification and also in general formant extraction compared to traditional LPC-based methods. Due to the data-driven basis of deep learning, it may be more robust against different noises speech signals are corrupted with (such as reverberation or white noise) if the training data is prepared to contain these corruptions for each target class. Otherwise, it may be fooled by background noise (Ribeiro et al., 2016). *Deepformants* applies a standard feedforward network architecture to learn LPC parameters, thus, in theory, removing potentially highly inaccurate and redundant estimates (Dissen & Keshet, 2021). The goal of the present

paper is to examine the features derived from these formant extraction methods in realistic forensic scenarios by applying a dataset that is created by forensic protocol requirements. Therefore, the used dataset must meet certain criteria.

From an engineering point of view, various modeling techniques can be adapted to the LR framework. State-of-the-art speaker verification techniques (i-vector, x-vector) (Dehak et al., 2010; Mandasari et al., 2011; Snyder et al., 2018; Kelly et al., 2019) produce speaker embeddings based on the spectrum of the voice and require large amounts of training samples. On the other hand, datasets developed directly for forensic purposes commonly contain limited samples. These corpora must meet certain special requirements (Morrison et al., 2012), such as multiple recordings of individuals separated in time, and multiple speaking styles from each speaker. Due to these highly controlled scenarios, a limited number of speakers can be recorded. Therefore, more simple modeling techniques are commonly used, among these are the multivariate kernel density estimation (MVKD) and Gaussian mixture models (GMM) that are favorably utilized in forensic voice comparison (Becker et al., 2008; Rose & Winter, 2010; Morrison, 2011a; Wang & Zhang, 2015; Hughes, 2017; Tsuge & Ishihara, 2018), These probability-based methods are easily fitted into the LR framework. Following the common practice of forensic voice comparison, we use the MVKD and GMM modeling techniques to compare formants estimated by the investigated formant trackers.

In this paper, we aim to investigate multiple matters on a Hungarian corpus specially created for forensic voice comparison. First, we introduce the Hungarian Database for Forensic Voice Comparison and evaluate the performance of MVKD and GMM on the corpus using formants as a first result obtained on the dataset. We would like to extend the existing works measuring the performance of these methods by applying them to a corpus created for forensic purposes. Coy and his colleagues (2021) compared *deepformants* to formants calculated by *Snack* and found controversial results. Formants extracted by *deepformants* were not found to be superior in all investigated scenarios. However, the ap-

plied corpus did not fulfill the requirements of forensic practices: recordings from multiple sessions from each speaker, and varying speaking styles.

Speaking styles in speaker verification were examined in some studies before, but they used artificial speaking style modifications for their experiments (Afshan et al., 2020) or the speakers were forced to change vocal efforts and speaking styles (Shriberg et al., 2008). Also, deep learning-based formant tracking was tested against noises added to the speech signal (Gowda et al., 2021). Our goal in this paper is not to examine these effects separately, but to measure the formant trackers' performance in a realistic forensic setup that our applied corpus enables. Utterance length is also an essential factor in forensic speaker comparison. There are studies that deal with these phenomena (such as Honglin & Jiangping, 2012), but deep learning-based formant trackers were not yet investigated.

Also, there are gaps in such works in Hungarian corpora, the filling of which would prove very useful. We investigate the effect of recording length on performance. Also, we would like to measure the performance of such methods when different speaking styles are available for offender (sample of evidence from crime scene) and suspect (test sample with speaker identity in question) voice samples. Finally, we examine the effect of noise on the applied formant calculation methods by adding reverberation effects and white noises of two signal-to-noise levels to the clean samples.

In the following chapters, we introduce the corpus, the formant tracker methods, modeling techniques, and evaluation metrics applied. Next, the experimental setup is described, followed by the achieved results. In the end, a detailed discussion is given with the conclusion of the work.

2. Methods

2.1. Hungarian Database for Forensic Voice Comparison

We introduce for the first time the FORVOICE audio database for Forensic Voice Comparison in the Hungarian language. To this day, this database con-

tains samples of 80 speakers, 39 females and 41 males (between the ages of 18 and 35). Net total speaking time (without silences and pauses) is 24.74 hours. Four speaking styles are recorded: free dialogue, controlled information exchange, monologue, and prescripted answers to questions (simulation of refused answers at interrogation). Two recordings were made per speaker, separated by a period of at least two weeks. These are marked as sessions 1 and 2 in the rest of the paper. The average duration of speaking durations per speaker and session is 266.66, 153.71, 124.39, and 12 seconds for each speaking style in respective order as mentioned before. Prior to recording, subjects expressed written consent to record their voice for the given research purposes. Head-mounted microphones were used to ensure the best possible recording quality (format: PCM, 44kHz, 16-bit). Manual transcriptions and phoneme and word level segmentations are available for all recordings.

2.2. Formant trackers and features

Four formant tracking methods were compared, three LPC based and one deep learning based:

- *Praat* (Boersma & Weenink, 2021) with 'Burg' method,
- *Snack Sound Toolkit* (Kåre, 2021),
- *Voicebox* (Brookes, 2021) (a *Matlab* based toolbox), and
- *Deepformants* (Dissen & Keshet, 2021 trained on the training set of VTR-TIMIT (Dissen et al., 2019). Based on the formant space examination in Van Heuven, 2016, it may not be a problem that the model is trained on a different language than it is used on.

25 ms as the window size and 10 ms as the time step were applied. Phoneme level segmentation (generated by Hidden Markov Model Toolkit (HTK) [Young & Young, 1993] using forced alignment based on the manual transcriptions) was available for the whole corpus. Formants (F) were measured for vowels /i/, /u/,

and /a:/, the ones with the farthest in the F1 and F2 space (nodes of a triangle) in Hungarian. For each vowel, the mean, the standard deviation, and the first three discrete cosine transform (DCT) coefficients (the 0^{th} coefficient was skipped) of the first three formants were calculated, resulting in a 15-dimension feature vector for each vowel.

Figure 1 shows the formant space (in the space of F1 and F2 derived from /i/, /u/, and /a:/ vowels) measured in samples without noise abruption (*clean* samples). The formant space was modeled by GMM probability density functions. The figure is not derived from the extracted features but it is only a sample of the deviation between the formant tracking methods. As the figure clearly shows, neither calculation framework gives the same (or completely accurate) formant space even in the case of clean samples. With added noise, even larger differences are found. Figures 2 and 3 show the formant space of *clean* and *white noise with 10 dB SNR* in the case of *Praat* and *deepformants*. The method *deepformants* seems to have calculated formants more robustly, but even in this case, a large deviation can be observed between the two acoustic scenarios. This may affect speaker verification results.

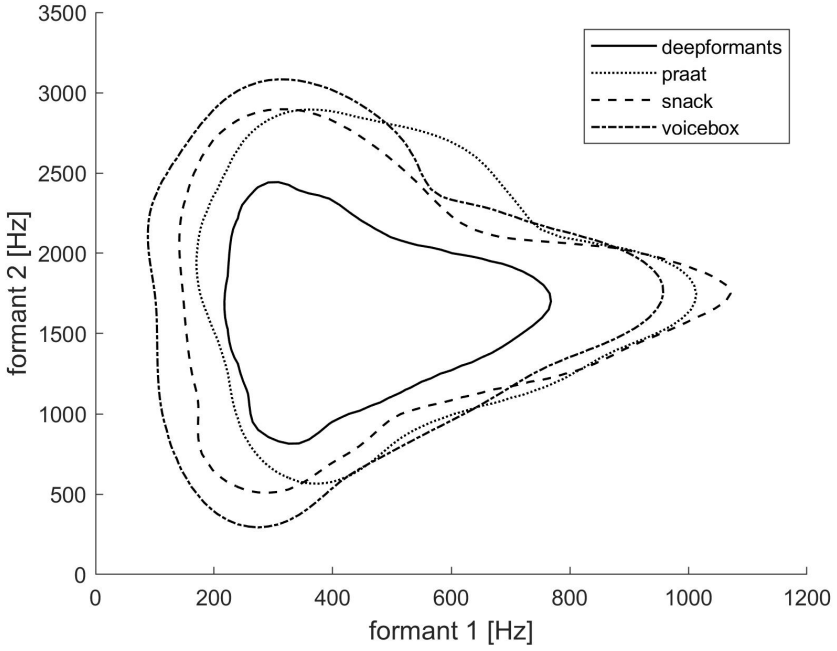

Figure 1: Formant space (F1-F2 triangle derived from /i/, /u/, and /aː/ vowels) measured at *clean* samples.

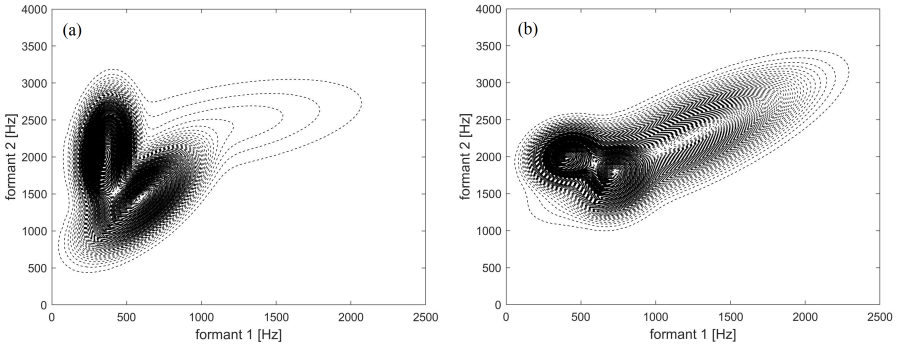

Figure 2: Formant space measured at *clean* (top) and *white noise with 10 dB SNR* (bottom) samples with *Praat*. The GMM-modeled PDFs are shown.

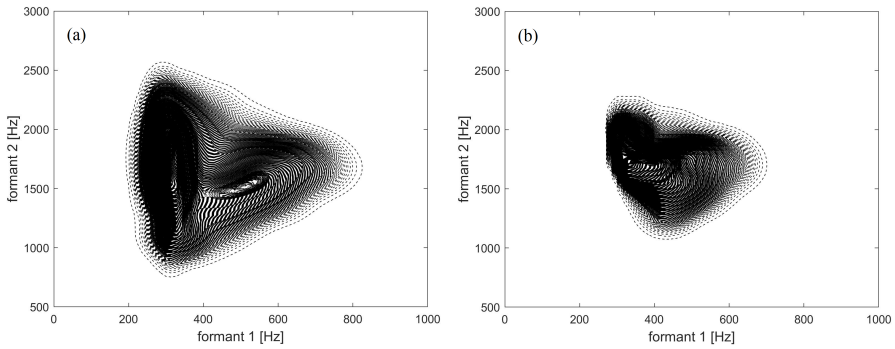

Figure 3: Formant space measured at *clean* (top) and *white noise with 10 dB SNR* (bottom) samples with *deepformants*. The GMM-modeled PDFs are shown.

2.3. LR framework

In a speaker verification method, we have to investigate two hypotheses: (1) "What is the possibility that the sample in question originates from the suspect?" and (2) "What is the possibility that the sample in question originates from a randomly selected speaker of a background population?" The ratio of these expressions expresses the strength of the evidence (Eq. 1). LR is the likelihood ratio, E is the evidence, H_{so} is the hypothesis of same-origin speakers, and H_{do} is the hypothesis of different-origin speakers.

$$LR = \frac{p(E|H_{so})}{p(E|H_{so})} \tag{1}$$

There exist several procedures to calculate LR value from univariate and multivariate continuous data. In this study, we use multivariate kernel density (MVKD) (Aitken & Lucy, 2004) and Gaussian mixture models (GMM) to estimate the density functions of features described in Section 2.2.

2.3.1. Multivariate Kernel Density

Kernel density estimation is a nonparametric technique for density estimation. It can be viewed as a generalization of histogram density estimation with improved statistical properties. In MVKD, the density function of the samples is estimated by using kernels centered at each data point. The estimated density

function is obtained by summing these kernels. This eliminates the problem of choosing the correct bins for histograms.

MVKD uses the summation of a set of equally-weighted kernels with one kernel per group centered on the mean vector of the measurements from that group for modeling between-group distribution. In forensic voice comparison, the speaker is such a group. The Gaussian kernel is used with a scaled covariance matrix of the pooled within-group covariance matrix. The degree of kernel smoothing (scaling) is determined by a function of the number of background database groups (Morrison, 2011a). The procedure is described in detail in Aitken & Lucy, 2004 and Morrison, 2011a. For detailed information and equations about calculating the LR score of MVKD, see Morrison, 2011a.

2.3.2. GMM-UBM

Gaussian mixture models (Reynolds & Rose, 1995; Hansen & Hasan, 2015) is a combination of Gaussian probability density functions (PDFs) that are commonly used to model multivariate data. It does not only cluster data in an unsupervised way, but also gives its PDF. Applying GMM to speaker modeling provides the speaker-specific PDF, from which a probability score can be obtained. Thus, by testing a sample with an unknown label, based on the probability scores of the speaker GMMs, a decision can be made.

A GMM is a mixture of Gaussian PDFs parameterized by a number of mean vectors, covariance matrices, and weights (Eq. 2). π_g, μ_g, and Σ_g indicate the weight, mean vector, and covariance matrix of the g^{th} mixture component. For a sequence of acoustic features ($X = x_n | n \in 1 \cdots T$), the probability of observing these features is computed as Eq. 3.

$$f(x_n|\lambda) = \sum_{g=1}^{M} \pi_g N(x_n|\mu_g, \Sigma_g) \tag{2}$$

$$p(X|\lambda) = \prod_{n=1}^{T} p(x_n|\lambda) \tag{3}$$

For the speaker verification scheme, a slightly different approach was developed in (Hansen & Hasan, 2015). Besides the claimed speaker's model, an alternate

model is necessary, which represents an 'opposing' model. This alternate model is called the universal background model (GMM-UBM). The GMM-UBM represents a background speaker population and it is trained on a large number of speaker samples. This is used to answer the question if the given test sample is more likely to be sampled from a target speaker or not and helps to calculate the LR score. It was first used in Reynolds et al. (2000). Later, UBM was used as an initial model for the speaker models: rather than training GMMs on speaker data directly, the specific speaker models were created by adapting a prior UBM (Gauvain & Lee, 1994). In the GMM-UBM scheme, H_{so} and H_{do} are represented by speaker-dependent GMM and the GMM-UBM, respectively.

2.3.3. Fusing calibration

Because formants were measured at multiple vowels, it is necessary to use the information obtained from them jointly. The evaluated final results are generated by fusing scores of individual vowel LR scores. Logistic regression (Hastie et al., 2009) is a probabilistic classification method, offering a common score-to-likelihood-ratio transformation, and it is feasible to calibrate a single set of scores and fuse multiple sets of scores (Brummer et al., 2007). It takes the same and different speaker labels as target labels and the extracted features as input, then it fits a logistic curve to the data, which can be interpreted as the probability of each class (same and different speaker identities). The logistic regression-based calibration and fusion need multiple scores of comparisons from same and different speaker sample pairs combined into training data. Tokens of the three applied vowels are used as acoustic-phonetic comparisons as parallel comparisons on the same speech sample required by the algorithm. Calibration is an affine transformation to a set of scores optimizing a cost function. In forensic voice comparison, this cost function to be minimized is the log-likelihood-ratio cost (C_{llr}) (Van Leeuwen & Br"ummer, 2007) (Eq. 4),

$$C_{llr} = \frac{1}{2}\left(\frac{1}{N_{so}}\sum_{i=1}^{N_{so}}\log_2\left(1+\frac{1}{LR_{so_i}}\right)+\frac{1}{N_{do}}\sum_{j=1}^{N_{do}}\log_2\left(1+LR_{do_j}\right)\right), \qquad (4)$$

where N_{so} and N_{do} are the number of same-origin and different-origin comparisons and LR_{so} and LR_{do} are the likelihood ratios derived from same-origin and different-origin comparisons. C_{llr} is a function measuring the balance of LR scores of same and different speaker comparisons. Ideal same-origin and different-origin comparison have $\log(LR_{so}) > 0$ and $\log(LR_{do}) < 0$, respectively. Incorrect (not as ideal as the mentioned inequalities) produce a higher C_{llr}. The better the performance of a forensic comparison system, the more correct LR values are produced, the lower C_{llr} is achieved, supplying the evidence magnitude. Thus, calibrated C_{llr} on known same-origin and different-origin sample pairs provides a metric of system validity. Calibration in this study followed Morrison (2011a) and was based on the UBM set. Same and different speaker pairs were built from all speaker pairs, skipping mirrored pairs due to symmetrical LR values (this is not exactly precise, but they are similar enough to be skipped). Throughout the paper, fused C_{llr} is the base of all evaluation metrics. C_{llr} is commonly drawn as tippet plots, showing LR components in a clean way (Morrison, 2009b).

2.4. The experimental setup

Two speaking styles were used in the study:

- **free dialogue (fd):** completely free dialogue with a talking partner without restrictions ($\sim$ 10 minutes). Due to individual head-mounted microphones, crosstalk volume was minimal between speakers;
- **monologue (m):** speakers should tell the events of their previous day objectively ($\sim$ 3 minutes per speaker).

The original clean recordings were also augmented with two types of distortions: reverberation and white noise. Reverb effects were applied by *Matlab* (Dattorro, 1997) with the following *Matlab* function arguments: density of reverb tail – 0.5 (scale: 0–1), decay factor of reverb tail – 0.5 (scale: 0–1), ratio of reverberated to original signal – 0.3 (scale: 0–1). The values were chosen by *Matlab* recommendations. White noise was added to clean recordings to achieve

two mean signal-to-noise levels: 10 dB and 15 dB. SNR was calculated as the sound intensity level ratio measured between the original clean recording and the added noise. This augmentation resulted in four final sets: clean, reverb, white noise with 10 dB SNR (*'wnoise10'*) and white noise with 15 dB SNR (*'wnoise15'*).

Multiple scenarios were considered to be evaluated: same speaking style and cross-speaking style experiments were carried out. In the same speaking style setup, suspect and offender (evidence from crime scene) samples are both taken from the *free dialogue* and *monologue* tasks and are evaluated. In the cross-speaking style setup, suspect and offender samples were taken from different tasks. In the last case, all speaking styles were used. From the 80 speakers, 60 were used to train the universal background model, the remaining 20 speakers were selected for suspect and offender sets. Sex distribution of the UBM matches the suspect-offender speakers. From the 20 speakers' samples, session 1 was used as the offender, while session 2 was applied as the suspect. Background models were trained on both speaking styles and sessions of the selected 60 speakers. The resulting scenarios for evaluation are summarized in Table 1. Formant measurements for UBM and offender models were always taken from the total sample duration. As suspect data, various durations were considered and evaluated. Sample chunks with lengths of $20, 40, 60, \cdots, 300$ seconds were applied (15 possible chunk durations). Figure 4 shows the number of vowel tokens in the function of sample chunk durations. It shows how the median and standard deviation of token numbers of speakers change as the function of sample duration increases. We would expect that the more tokens we have to extract features from, the more robust the speaker verification is. For the sake of data uniformity, if the total duration of a given sample was exceeded by the chunk length, the vowel tokens were selected from the total sample leaving the original chunk length as notation. For example, for a 120 seconds long speech sample, the 160 seconds long chunk length used the total speech sample but was still marked as '160'. Through initial experiments, the mixture number of the

Table 1: Summary of evaluation scenarios according to speaking styles

test case	UBM	#speakers	session	speaking style
	offender	60	1 and 2	free dialogue and monologue
1.	suspect	20	1	free dialogue
	suspect	20	2	free dialogue
	UBM	60	1 and 2	free dialogue and monologue
2.	offender	20	1	monologue
	suspect	20	2	monologue
	UBM	60	1 and 2	free dialogue and monologue
3.	offender	20	1	free dialogue
	suspect	20	2	monologue
	UBM	60	1 and 2	free dialogue and monologue
4.	offender	20	1	monologue
	suspect	20	2	free dialogue

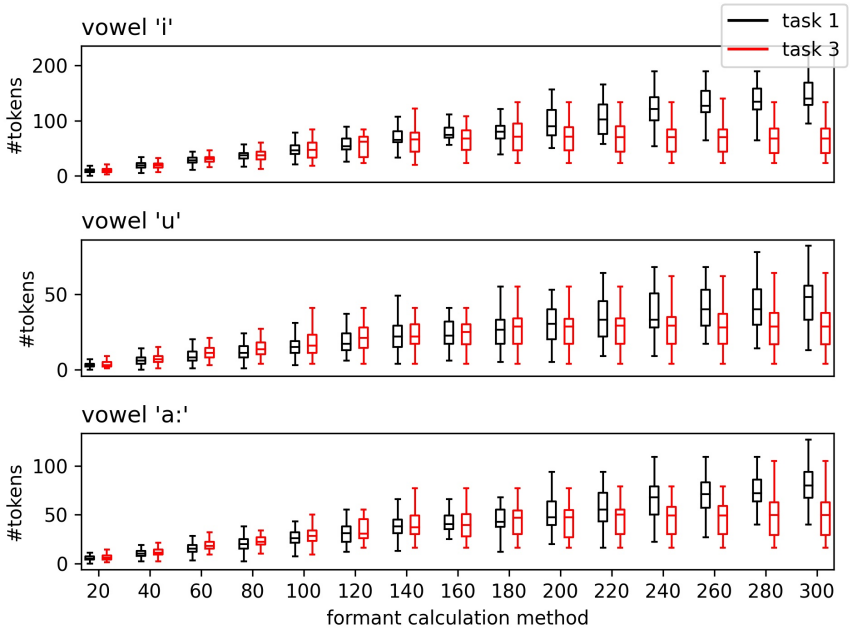

Figure 4: Number of vowel tokens according to chunk durations.

GMM model was selected as 11 and all features were selected to be included in the final feature vector.

2.5. Evaluation metrics

Multiple evaluation metrics are used to assess the performance of the test scenarios:

- Post-calibration fused C_{llr} value as described earlier (lower score means better performance).
- Equal error rate (EER). EER is the level where the false acceptance rate and the false rejection rate are equal, commonly used in biometric security systems. Lower EER means better performance.
- Receiver operating characteristics (ROC) curve. ROC can measure the performance of a binary classifier system as its discrimination threshold is varied. In our case, by discrimination threshold, we mean the level of LR score of accepting a speaker pair as same. The area under the ROC curve (AUC) value, when using normalized units, is equal to the probability that a classifier will rank a randomly chosen same speaker pair higher than a randomly chosen different origin speaker pair. A higher AUC value means better performance.

One-way ANOVA tests were used for checking mean value equivalence for investigated groups. To measure speaking style difference significance, generalized linear mixed models were applied (Hedeker, 2005).

3. Results

3.1. Chunk lengths

The effect of chunk lengths on performance was evaluated. Line plots are used for visualization to emphasize the possible trends of performance values. Figure 5 shows the C_{llr} values measured at *clean* samples as a function of chunk lengths. The mean performance of each formant calculation method is shown

separately. Due to the four experimental setups, a standard deviation range can also be depicted. Starting from 20 s, a continuously decreasing trend can be observed when applying both GMM and MVKD modeling techniques. Although MVKD resulted in a lower C_{llr} value at the beginning, GMM seems to outperform when enough tokens are available from all vowels. The EER and AUC values develop in the same way (Figures 6 and 7) as one would expect looking at the C_{llr} values. EERs fall (and AUC values rise) sharply until the 100 s chunk length and continue to decrease (AUC: increase) in a slighter way afterwards. Considering the various formant calculation methods, *praat* and *snack* had the lowest C_{llr}. The single method based on deep learning seems to lag behind the others, having an increased C_{llr} throughout the chunk lengths. ANOVA tests show no significant differences (p-values are 0.622 and 0.113 for GMM and MVKD, respectively) at 20 s chunk lengths between formant measurement methods. In the case of 300 s, there is a significant difference in mean values of C_{llr} in the case of GMM (p-values of ANOVA tests are 0.026 and 0.066 for GMM and MVKD, respectively).

3.2. Same and different speaking styles

Speaking style can have a significant effect on voice comparison by an LR framework (Drygajlo et al., 2015). Among the four experimental setups, two used the same speaking style for suspect and offender data (#1 and #2), and likewise, two used different ones (#3 and #4). For the experiments, the total lengths of the samples are used. All results obtained in these experimental setups are included in Table 2. By depicting the single C_{llr} values of the same and different speaking styles as a scatter plot (Figure 8), it is clear that by applying the same speaking style as enrollment and target samples, lower C_{llr} can be achieved. Same and different speaking style C_{llr} measurements are noted by '+' and 'x', respectively. The C_{llr} values depicted are single values for a single experiment scenario. The same effect of noise corruption can also be observed: better and indistinguishable results of clean and reverberated samples, and almost identically worse results for white noise added. Generalized linear mixed

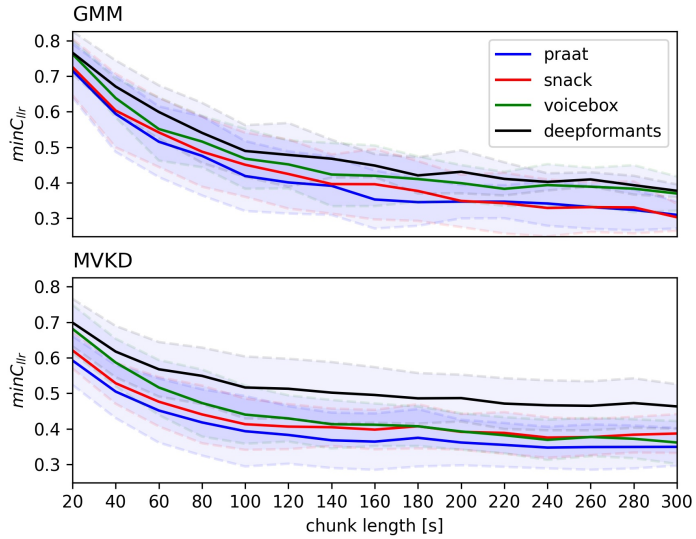

Figure 5: C_{llr} values according to split lengths using *clean* samples (top: GMM, bottom: MVKD).

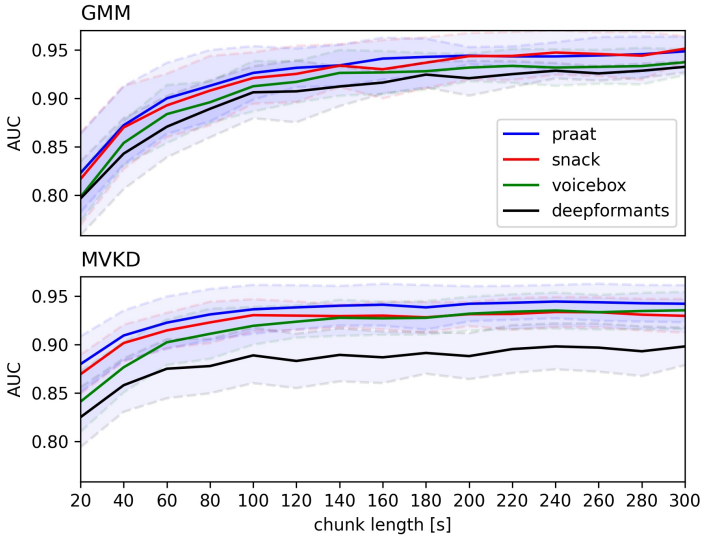

Figure 6: AUC values according to split lengths using *clean* samples (top: GMM, bottom: MVKD).

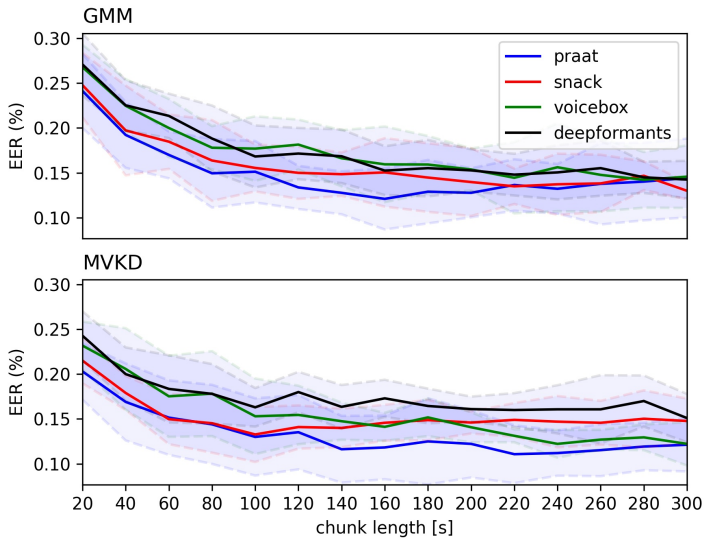

Figure 7: Equal error rates according to split lengths using *clean* samples (top: GMM, bottom: MVKD).

models were fitted to the data with C_{llr} as the target variable and speaking style (same or different), noise, and method as fixed variables. In the case of both GMM and MVKD, speaking style and noise were significant factors ($p < 0.05$) and the method was not ($p > 0.05$).

By calculating the difference of the mean of the C_{llr} values for same and different speaker styles (Figure 9), only positive values are present, showing that different speaking styles always deteriorate performance.

3.3. Noise types

Noises applied to the clean samples have different effects. Reverberation did not decrease the performance of the LR framework. On the contrary, there are multiple cases when the reverb effect did lower the C_{llr} value. Figure 10 shows mean and standard deviations (as error bars) of C_{llr} for every formant calculation method and sample quality. Only the results of 300 s long chunks are depicted because all chunk durations showed the same behavior in evaluation.

Table 2: Results aggregated according to different or same speaking styles in suspect and offender samples.

method	noise	speaking style	AUC		minC_{llr}		eer	
			GMM	MVKD	GMM	MVKD	GMM	MVKD
deepformants	clean	different	0.930	0.885	0.385	0.505	0.153	0.164
		same	0.940	0.911	0.371	0.422	0.132	0.139
	reverb	different	0.888	0.899	0.475	0.492	0.188	0.179
		same	0.949	0.926	0.310	0.378	0.125	0.140
	wnoise10	different	0.871	0.871	0.550	0.544	0.236	0.224
		same	0.908	0.950	0.449	0.316	0.157	0.108
	wnoise15	different	0.911	0.885	0.459	0.523	0.205	0.214
		same	0.909	0.951	0.458	0.324	0.165	0.141
praat	clean	different	0.938	0.929	0.332	0.384	0.153	0.132
		same	0.959	0.956	0.288	0.316	0.136	0.111
	reverb	different	0.929	0.948	0.361	0.341	0.159	0.132
		same	0.959	0.964	0.273	0.282	0.139	0.135
	wnoise10	different	0.835	0.820	0.622	0.647	0.238	0.237
		same	0.859	0.911	0.550	0.447	0.191	0.164
	wnoise15	different	0.844	0.849	0.604	0.590	0.252	0.209
		same	0.909	0.916	0.451	0.441	0.154	0.179
snack	clean	different	0.944	0.915	0.325	0.434	0.135	0.164
		same	0.959	0.945	0.282	0.340	0.125	0.132
	reverb	different	0.924	0.930	0.400	0.397	0.135	0.180
		same	0.955	0.950	0.284	0.328	0.108	0.109
	wnoise10	different	0.846	0.873	0.583	0.540	0.208	0.203
		same	0.883	0.937	0.527	0.378	0.211	0.160
	wnoise15	different	0.850	0.888	0.595	0.500	0.213	0.193
		same	0.914	0.940	0.418	0.368	0.188	0.138
voicebox	clean	different	0.933	0.925	0.394	0.394	0.153	0.135
		same	0.942	0.946	0.346	0.330	0.139	0.110
	reverb	different	0.938	0.934	0.357	0.377	0.147	0.142
		same	0.946	0.957	0.342	0.296	0.112	0.110
	wnoise10	different	0.833	0.826	0.610	0.657	0.231	0.249
		same	0.850	0.899	0.602	0.464	0.224	0.179
	wnoise15	different	0.830	0.811	0.604	0.679	0.256	0.274
		same	0.900	0.886	0.492	0.515	0.181	0.210

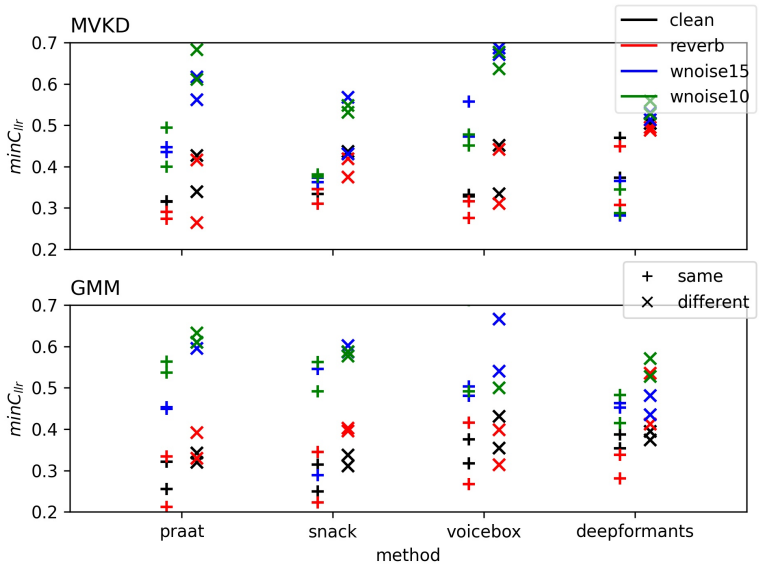

Figure 8: Scatter plot of C_{llr} values according to speaking styles (top: MVKD, bottom: GMM).

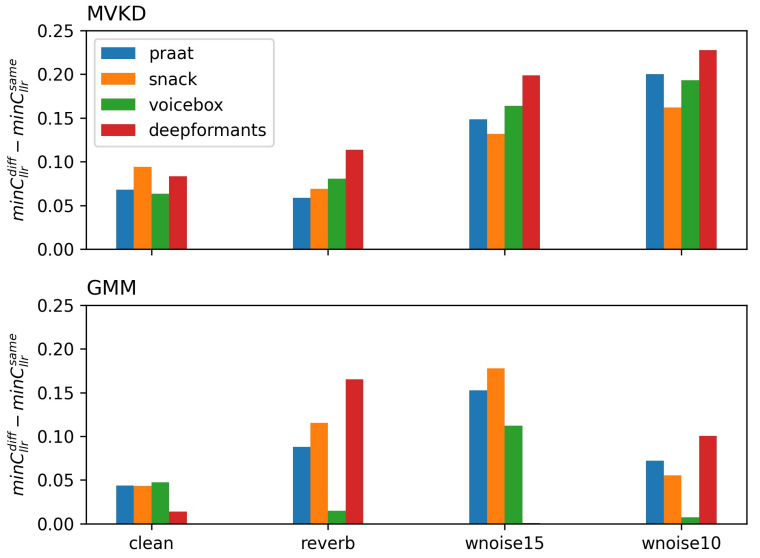

Figure 9: Difference of C_{llr} values according to speaking styles ($C_{llr}{}^{diff} - C_{llr}{}^{diff}$). Positive values mean lower C_{llr} scores in same speaking style cases (top: MVKD, bottom: GMM).

However, various observations can be made by observing the measured values. As expected, white noise added to the samples has a higher impact on performance. LPC-based calculation methods exhibit a high degree of deterioration in all evaluation metrics. The method *deepformants*, however, seems to have a more robust resistance to white noise. In the case of MVKD, the C_{llr} did not decrease at all, and in the case of GMM, *deepformants* had the lowest mean values across white noise corrupted samples. There are no differences between the 10 dB and 15 dB SNR samples. The p-values of one-way ANOVA tests are shown in Table 3 marking if there are any significant differences between sample qualities and formant calculation methods. Besides *deepformants*, all other methods show significant differences across noise types. On the other hand, if we consider noise types, neither case reaches a significance level to show that the methods would differ for the given noise type.

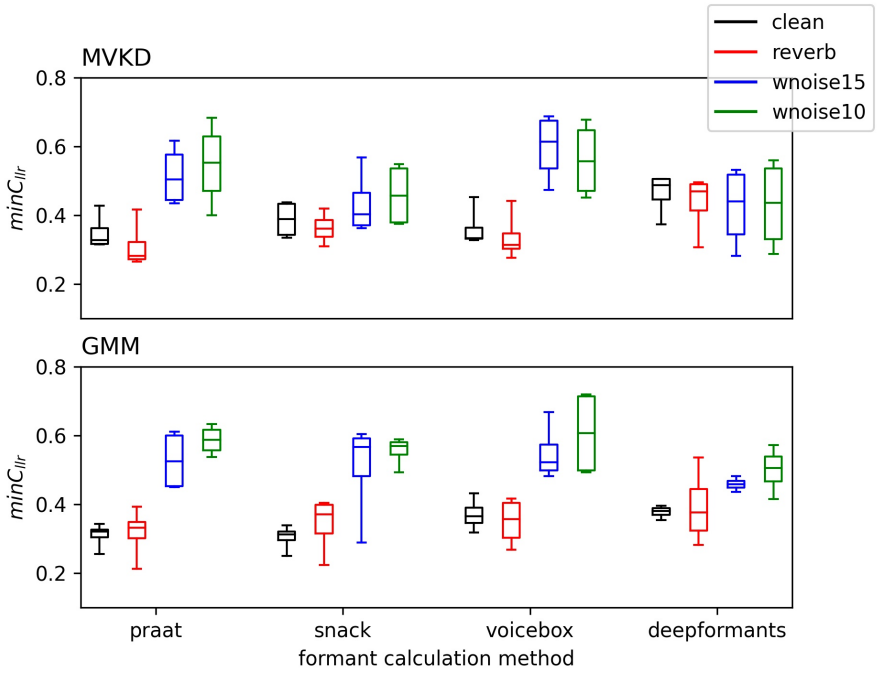

Figure 10: C_{llr} values according to formant calculation methods and noises (chunk size: 300s). Top: MVKD, bottom: GMM.

Table 3: The p-values of one-way ANOVA tests in case of 300 ms chunk lengths for noises across methods and methods across noises. Values < 0.05 are marked with '*'.

		GMM	**MVKD**
method	praat	0.0001*	0.0054*
	snack	0.0042*	0.3079
	voicebox	0.0025*	0.0021*
	deepformants	0.0720	0.9501
noise	clean	0.0257	0.0656
	reverb	0.6741	0.1329
	wnoise15	0.5985	0.1053
	wnoise10	0.2848	0.3498

4. Discussion

In this study, four formant trackers were systematically evaluated for forensic voice comparison. We analyzed the performance of various formant trackers in a forensic voice comparison setup depending on speaking style, length of suspended recording, and noises. Based on our results, the deep learning approach (*deepformants*) showed significant differences. However, the other methods, although all are based on LPC, also did not produce the same results (see Figure 1 for initial formant measurement example). On the contrary, significant differences could be observed. The method *deepformants* shows a narrower F1–F2 space than the other methods, explained by the real 'tracking' nature of its algorithm. While the LPC-based ones are estimating formants in each frame independently, *deepformants* takes neighboring frames into consideration when estimating formants in a given frame. This results in a narrower F1–F2 space but its advantage is its more robust nature. As Figure 8 shows, there are fewer differences between samples with different noise corruption in its case (although it performs basically worse also in the case of *clean* samples).

Our results confirm that the longer speech segments are used (more formant tokens are involved), the better the performance can be measured. Below $120-140$ seconds, there is a rapid decline in C_{llr}. If we look at EERs, a value of

around 0.15 can be achieved from $60 - 80$ seconds long chunks. This implies that formant estimates are rather unreliable if they are based on short segments.

Speaking style had a significant effect on performance measurements. When the offender and suspect speech samples originated from the same speaking style (free dialog and monologue in this study), lower C_{llr}, EER and higher AUC could be calculated compared to cases when the offender and suspect samples contained different speaking styles. Figure 10 shows that when the mean of the C_{llr} of different speaking style test trials was subtracted from same speaking style mean values, no negative values could be calculated. This implies that when comparing formants in a forensic voice comparison situation, speaking style matters a lot. It is always advised to compare the speech material of suspects with evidence of the same style.

Reverberation didn't seem to influence formant estimation in a bad way. In some cases, even lower C_{llr} and EER could be calculated compared to *clean* samples. White noise, however, mostly deteriorated the performance to a large extent. Besides *deepformants*, which seemed to be robust against white noise, but performed worse even with *clean* samples, only *snack* was robust against *white noise* with 15 dB SNR. *White noise* with 10 dB SNR always deteriorated performance, only *deepformants* was resistant to it using MVKD modeling.

There are no exact results on which method is better and which is to be used in studies and works. C_{llr} values of Figure 8 shows that the three LPC based methods perform similarly. They all make large mistakes when samples are corrupted with white noise. Although the method *deepformants* performs slightly worse than the others used in this study, it seems to have more resilience to white noise. This may be due to its real formant tracking approach, as mentioned before. This implies that it cannot be stated that the DNN approach outperforms the LPC-based ones.

5. Conclusion

In this study, several aspects of formant modeling in forensic voice comparison were investigated. A corpus containing 80 speakers with multiple speaking styles is used to estimate first and second formant values by four different formant estimation methods (three based on LPC, one on deep learning). Formants were modeled by multivariate kernel density estimation and Gaussian mixture models. It was found that the length of recording used as suspect samples influences performance to a large extent. Additionally, formant tracking based on deep learning lags behind the other methods in all metrics. Same and different speaking styles also have a measurable effect on performance. Samples corrupted with reverberation do not deteriorate results but white noise does. The continuation of the work will be to investigate these effects also by fully automatic voice comparison systems, such as x-vector and i-vectors.

Acknowledgements

The work was funded by project no. FK128615, which has been implemented with the support provided from the National Research, Development, and Innovation Fund of Hungary, financed under the FK_18 funding scheme.

References

Afshan, A., Guo, J., Park, S. J., Ravi, V., McCree, A., & Alwan, A. (2020). *Variable frame rate-based data augmentation to handle speaking-style variability for automatic speaker verification*. arXiv (Cs, Eess). doi:https://doi.org/10.48550/arXiv.2008.03616. arXiv:2008.03616.

Aitken, C. G., & Lucy, D. (2004). Evaluation of trace evidence in the form of multivariate data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *53*, 109–122. doi:https://doi.org/10.1046/j.0035-9254.2003.05271.x.

Bazen, A. M., & Veldhuis, R. N. (2004). Likelihood-ratio-based biometric verification. *IEEE Transactions on Circuits and Systems for Video Technology*, *14*, 86–94. doi:https://doi.org/10.1109/TCSVT.2003.818356.

Becker, T., Jessen, M., & Grigoras, C. (2008). Forensic speaker verification using formant features and Gaussian mixture models. In *Ninth Annual Conference of the International Speech Communication Association*.

Boersma, P., & Weenink, D. (2021). *Praat: Doing phonetics by computer*. [Computer program] (Version 6.1.42). Retrieved 15 April 2021. URL: http://www.praat.org/.

Brookes, M. (2021). *Voicebox*. URL: http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html.

Brummer, N., Burget, L., Cernocky, J., Glembek, O., Grezl, F., Karafiat, M., Van Leeuwen, D. A., Matejka, P., Schwarz, P., & Strasheim, A. (2007). Fusion of heterogeneous speaker recognition systems in the stbu submission for the nist speaker recognition evaluation 2006. *IEEE Transactions on Audio, Speech, and Language Processing*, *15*, 2072–2084. doi:https://doi.org/10.1109/TASL.2007.902870.

Chaudhary, G., Srivastava, S., & Bhardwaj, S. (2017). Feature extraction methods for speaker recognition: A review. *International Journal of Pattern Recognition and Artificial Intelligence*, *31*, 1750041. doi:https://doi.org/10.1142/S0218001417500410.

Coy, T., Hughes, V., Harrison, P., & Gully, A. J. (2021). A comparison of the accuracy of dissen and keshet's (2016) deepformants and traditional lpc methods for semi-automatic speaker recognition. *Interspeech 2021*, (pp. 406–410). doi:https://doi.org/10.21437/Interspeech.2021-1487.

Dattorro, J. (1997). Effect design, part 1: Reverberator and other filters. *Journal of the Audio Engineering Society*, *45*, 660–684.

Dehak, N., Kenny, P. J., Dehak, R., Dumouchel, P., & Ouellet, P. (2010). Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, *19*, 788–798.

Dissen, S., & Keshet, J. (2021). *DeepFormants*. URL: `https://github.com/MLSpeech/DeepFormants`.

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

Drygajlo, A., Jessen, M., Gfroerer, S., Wagner, I., Vermeulen, J., & Niemi, T. (2015). *Methodological guidelines for best practice in forensic semiautomatic and automatic speaker recognition*. Verlag f"ur Polizeiwissenschaft.

Gargouri, D., Kammoun, M. A., & Hamida, A. B. (2006). A comparative study of formant frequencies estimation techniques. In *Proceedings of the 5th WSEAS International Conference on Signal Processing, Istanbul, Turkey* (pp. 15–19).

Gauvain, J.-L., & Lee, C.-H. (1994). Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains. *IEEE Transactions on Speech and Audio Processing*, *2*, 291–298. doi:`https://doi.org/10.1109/89.279278`.

Gowda, D. N., Bollepalli, B., Kadiri, S. R., & Alku, P. (2021). Formant tracking using quasi-closed phase forward-backward linear prediction analysis and deep neural networks. *IEEE Access*, *9*, 151631–151640. doi:`https://doi.org/10.1109/ACCESS.2021.3126280`.

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 & the Law*, *15*. doi:`https://doi.org/10.1558/IJSLL.V15I2.193`.

Hansen, J. H., & Hasan, T. (2015). Speaker recognition by machines and humans: A tutorial review. *IEEE Signal Processing Magazine*, *32*, 74–99. doi:https://doi.org/10.1109/MSP.2015.2462851.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. Springer Science & Business Media.

Hedeker, D. (2005). *Generalized linear mixed models*. Encyclopedia of Statistics in Behavioral Science.

Honglin, C., & Jiangping, K. (2012). Speech length threshold in forensic speaker comparison by using long-term cumulative formant (ltcf) analysis. In *2012 Second International Conference on Instrumentation, Measurement, Computer, Communication and Control* (pp. 418–421). doi:https://doi.org/10.1109/IMCCC.2012.103.

Hughes, V. (2017). Sample size and the multivariate kernel density likelihood ratio: How many speakers are enough? *Speech Communication*, *94*, 15–29. doi:https://doi.org/10.1016/j.specom.2017.08.005.

Kameny, I., Brackenridge, W. A., & Gillmann, R. (1974). Automatic formant tracking. *The Journal of the Acoustical Society of America*, *56*, S28–S28. doi:https://doi.org/10.1121/1.1914097.

Kåre, S. (2021). *Snack Sound Toolkit*. URL: http://www.speech.kth.se/snack.

Kelly, F., Forth, O., Kent, S., Gerlach, L., & Alexander, A. (2019). Deep neural network based forensic automatic speaker recognition in vocalise using x-vectors. In *Audio Engineering Society Conference: 2019 AES International Conference on Audio Forensics*.

Mandasari, M. I., McLaren, M. L., & van Leeuwen, D. A. (2011). *Evaluation of i-vector speaker recognition systems for forensic application*. Florence, Italy.

Matz, M. V., & Nielsen, R. (2005). A likelihood ratio test for species membership based on dna sequence data. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *360*, 1969–1974. doi:https://doi.org/10.1098/rstb.2005.1728.

Morrison, G. S. (2009a). Forensic voice comparison and the paradigm shift. *Science & Justice*, *49*, 298–308. doi:https://doi.org/10.1016/j.scijus.2009.09.002.

Morrison, G. S. (2009b). Likelihood-ratio forensic voice comparison using parametric representations of the formant trajectories of diphthongs. *The Journal of the Acoustical Society of America*, *125*, 2387–2397. doi:https://doi.org/doi.org/10.1121/1.3081384.

Morrison, G. S. (2011a). A comparison of procedures for the calculation of forensic likelihood ratios from acoustic–phonetic data: Multivariate kernel density (mvkd) versus gaussian mixture model–universal background model (gmm–ubm). *Speech Communication*, *53*, 242–256. doi:https://doi.org/10.1016/j.specom.2010.09.005.

Morrison, G. S. (2011b). Measuring the validity and reliability of forensic likelihood-ratio systems. *Science & Justice*, *51*, 91–98. doi:https://doi.org/10.1016/j.scijus.2011.03.002.

Morrison, G. S., Rose, P., & Zhang, C. (2012). Protocol for the collection of databases of recordings for forensic-voice-comparison research and practice. *Australian Journal of Forensic Sciences*, *44*, 155–167. doi:https://doi.org/10.1080/00450618.2011.630412.

Reynolds, D. A., Quatieri, T. F., & Dunn, R. B. (2000). Speaker verification using adapted Gaussian mixture models. *Digital Signal Processing*, *10*, 19–41. doi:https://doi.org/10.1006/dspr.1999.0361.

Reynolds, D. A., & Rose, R. C. (1995). Robust text-independent speaker identification using Gaussian mixture speaker models. *IEEE Transactions on Speech and Audio Processing*, *3*, 72–83. doi:https://doi.org/10.1109/89.365379.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). *"Why Should I Trust You?": Explaining the Predictions of Any Classifier*. arXiv. URL: http://arxiv.org/abs/1602.04938. arXiv:1602.04938.

Rose, P., & Winter, E. (2010). Traditional forensic voice comparison with female formants: Gaussian mixture model and multivariate likelihood ratio analyses. *SST*, (pp. 42–45).

Saks, M. J., & Koehler, J. J. (2005). The coming paradigm shift in forensic identification science. *Science*, *309*, 892–895. doi:https://doi.org/10.1126/science.1111565.

Shriberg, E., Graciarena, M., Bratt, H., Kathol, A., Kajarekar, S. S., Jameel, H., Richey, C., & Goodman, F. (2008). Effects of vocal effort and speaking style on text-independent speaker verification. In *Ninth Annual Conference of the International Speech Communication Association*. 22–26 September 2008, Brisbane, Australia.

Snell, R. C., & Milinazzo, F. (1993). Formant location from lpc analysis data. *IEEE Transactions on Speech and Audio Processing*, *1*, 129–134. doi:https://doi.org/10.1109/89.222882.

Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., & Khudanpur, S. (2018). X-vectors: Robust dnn embeddings for speaker recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5329–5333). 15–20 April 2018, Calgary, Alberta, Canada.

Titze, I. R., & Martin, D. W. (1998). *Principles of voice production*. Acoustical Society of America.

Tsuge, S., & Ishihara, S. (2018). Text-dependent forensic voice comparison: Likelihood ratio estimation with the hidden markov model (hmm) and gaus-

sian mixture model. In *Proceedings of the Australasian Language Technology Association Workshop 2018* (pp. 17–25).

Van Heuven, V. J. (2016). An acoustic characterisation of English vowels produced by american, dutch, chinese and hungarian speakers. *Hungarian Journal of Applied Linguistics*, *16*, 1–20.

Van Leeuwen, D. A., & Br"ummer, N. (2007). An introduction to application-independent evaluation of speaker recognition systems. In *Speaker classification I* (pp. 330–353). Springer. doi:https://doi.org/10.1007/978-3-540-74200-5_19.

Wang, H., & Zhang, C. (2015). Forensic automatic speaker recognition based on likelihood ratio using acoustic-phonetic features measured automatically. *Journal of Forensic Science and Medicine*, *1*, 119. doi:https://doi.org/10.4103/2349-5014.169617.

Young, S. J., & Young, S. (1993). *The HTK hidden Markov model toolkit: Design and philosophy*. Cambridge University Engineering Department.

Zhang, C., Morrison, G. S., Enzinger, E., & Ochoa, F. (2013). Effects of telephone transmission on the performance of formant-trajectory-based forensic voice comparison–female voices. *Speech Communication*, *55*, 796–813. doi:https://doi.org/10.1016/j.specom.2013.01.011.