INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Why Did the x-Vector System Miss a Target Speaker? Impact of Acoustic Mismatch Upon Target Score on VoxCeleb Data

Rosa González Hautamäki and Tomi Kinnunen

Computational Speech Group, School of Computing, University of Eastern Finland, Finland
{rgonza,tkinnu}@cs.uef.fi

Abstract

Modern automatic speaker verification (ASV) relies heavily on machine learning implemented through deep neural networks. It can be difficult to interpret the output of these black boxes. In line with interpretative machine learning, we model the dependency of ASV detection score upon acoustic mismatch of the enrollment and test utterances. We aim to identify mismatch factors that explain target speaker misses (false rejections). We use distance in the first- and second-order statistics of selected acoustic features as the predictors in a linear mixed effects model, while a standard Kaldi x-vector system forms our ASV black-box. Our results on the VoxCeleb data reveal the most prominent mismatch factor to be in F0 mean, followed by mismatches associated with formant frequencies. Our findings indicate that x-vector systems lack robustness to intra-speaker variations.

Index Terms: automatic speaker verification, VoxCeleb, target speaker errors, acoustic mismatch

1. Introduction

Automatic speaker verification (ASV) [1] systems take a pair of utterances (enrolment and test) to predict if the speakers in them are same or different. When the former is actually true, such pairwise comparison is known as a *target* trial, otherwise as a *nontarget* trial. The prediction can be hard binary decision or a real-valued speaker similarity score. Current state-of-the-art relies largely on deep neural networks, such as the *x-vector* [2] architecture, to extract speaker embeddings from each utterance. Speaker similarity score is then formed by comparing the enrolment and test embeddings using a back-end classifier [3].

ASV systems are typically optimized to make accurate predictions for given data, *on average*; not all the speakers or trials are necessarily equally treated. ASV systems are typically required to operate in an *open-world* setting where the number of target speakers (classes) is allowed to increase dynamically. Additionally, ASV is used across varied operating conditions including unseen microphones, environments, and speaking styles. Thus, despite the effort that one spends on optimization, ASV systems are bound to face the unknown. Moreover, reliance on machine learning may yield decisions that humans have difficulty to interpret. The importance of explaining the decisions of machine learning systems is acknowledged and ASV is no exception. Forensic voice comparison is a canonical example of a high-stakes application where importance of explainable decisions is evident. Nonetheless, explaining the decisions of ASV systems is important for researchers, too, as it may reveal system loopholes.

We model the dependency of ASV score upon acoustic mismatch in enrollment and test data. The ASV system is treated as a black-box given *as-is*: we may run it on new speech data to obtain speaker similarity scores, but otherwise we cannot optimize or interact with it. The acoustic features, however, are selected by us based on hypotheses on the type of variation expected in given data. Our work is reminiscent of *score calibration* [4] where ASV score is adjusted with the aid of external quality signals as side information. Nonetheless, besides using different methodology [5], our perspective is on **explanatory analysis of a ASV system on a given evaluation corpus**, rather than on improving predictive performance. Other related research includes probing information in speaker embeddings [6, 7]. Different from these studies that are either specific to a given type of speaker embedding or require training new classifiers in the embedding space, we model the detection score in terms of explanatory variables. The latter consists of acoustic-phonetic measures available in a public-domain toolkit [8].

We focus on modeling target trials. The ideal ASV score for a same-speaker (target) trial is as large number as possible — optimally, plus infinity. Acoustic mismatch between enrollment and test data may lower the ASV score and consequently lead to falsely rejected (missed) target speaker. In an access control context, miss implies user inconvenience and in a forensic context it implies falsely declaring that the perpetrator is not present in a given trace sample. Using an up-to-date x-vector system and the large-scale VoxCeleb dataset [9] that consists of 'found data' quality celebrity recordings, we aim to identify what types of acoustic mismatches are likely to contribute to increased target speaker misses.

We extend upon our recent work [5] in terms of speech database size, qualities, and the selected acoustic features. In [5] we used a self-collected (now publicly available) AVOID corpus of 60 Finnish speakers. The speakers were asked to purposefully modify their voices to sound like *old* and *child* speakers, so as to purposefully reinforce large variation between enrollment and test data. Indeed, the standard Kaldi x-vector system was shown to severely degrade. For instance, *equal error rate* (EER) of male speakers increased from $\sim 1.6\%$ (modal-modal) to $\sim 25\%$ (modal-intended child). The degradation was associated/explained by differences in F0 and formants. Nonetheless, one may argue that in contemporary communication context, we do not attempt to disguise our identity or perform caricaturic voice acting. Nonetheless, the authors have observed substantial variation in speaking styles and background audio qualities in VoxCeleb data through informal listening. It is therefore plausible that target speakers may get easily missed on VoxCeleb data, too. Thus, another aim of our work is to address generalizability of our earlier findings [5] (for acted voice data) to contemporary speech present in the VoxCeleb dataset.

2. Analysis methodology

We provide a brief summary of the interpretative model presented in [5]. An important aspect of ASV systems reliability is to understand the factors that affect its accuracy. Can the score

http://dx.doi.org/10.21437/Interspeech.2020-2715

provided by ASV systems be explained by changes in acoustic measures of compared speech segments? To address this question, we model our data using a statistical regression technique specially design for repeated measures known as *linear mixed effect model* (LME) [10]. In general, regression models seek to relate a *dependent variable* to a set of *predictors* or *independent variables*.

2.1. Dependent and Predictor Variables

Let $\mathcal{U} = (\mathcal{U}_\text{e}, \mathcal{U}_\text{t})$ denote a pair of enrollment and test utterances. An ASV system produces a *log-likelihood ratio* (LLR) score (*dependent variable*, y) between the two utterances as,

$$y = \log \frac{p(\mathcal{U}|H_0, \boldsymbol{\theta}_\text{asv})}{p(\mathcal{U}|H_1, \boldsymbol{\theta}_\text{asv})}, \tag{1}$$

where H_0 and H_1 represent the target (same-speaker) and non-target (different-speaker) hypotheses, respectively, and $\boldsymbol{\theta}_\text{asv}$ encapsulates all the ASV parameters. In our case, (1) represents LLR score from a *probabilistic linear discriminant analysis* (PLDA) back-end classifier [3], while the two utterances are represented using their x-vector [2] speaker embeddings. The higher the value of y, the more confident the ASV system is that the speakers in the two utterances are the same.

While y serves as the response variable, our predictor variables, x, are formed by *acoustic distances* of the form $x = |\varphi(f(\mathcal{U}_\text{e})) - \varphi(f(\mathcal{U}_\text{t}))|$. Here $f(\cdot)$ is a short-term (frame-level) feature extractor that converts a speech utterance into a sequence of scalar features, and $\varphi(\cdot)$ is a fixed summary statistics operator. By including different features and summary operators, we come up with a vector of D acoustical predictors, $\boldsymbol{x} = (x_1, \ldots, x_D)$ for any utterance pair $(\mathcal{U}_\text{e}, \mathcal{U}_\text{t})$. In this work, $\varphi \in \{\texttt{mean}, \texttt{std}\}$ consists of mean and standard deviation while the features include various standard speech features (see Table 1).

2.2. Mixed effects model

In LME models [10], predictors that are common to all observations are known as *fixed effects*. They are represented by means of contrast. In our model, these are the acoustic distances for each single target trial. Factors that are considered as a sample of a population, in turn, are known as *random effects*. The random effects in our model are the speakers. The model reflects variations associated with the speakers, as a variable with zero mean and unknown variance.

To be more specific, our model is defined as:

$$y_{ij} = \boldsymbol{\beta}^t \boldsymbol{x}_{ij} + b_i + \varepsilon_{ij}, \tag{2}$$

where y_{ij} is the LLR score for the jth trial of target speaker i, $\boldsymbol{\beta}^t \boldsymbol{x}_{ij}$ is the fixed effect part (acoustic distances and their weights), b_i is the per-speaker *random effect* and ε_{ij} is the residual. The assumption for a random speaker effect and the residual error is that they are independent of each other and follow a normal distribution: $b_i \sim \mathcal{N}(0, {\sigma_b}^2)$ and $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$.

3. Experimental setup

3.1. VoxCeleb corpus

VoxCeleb is a publicly available large-scale dataset of speech extracted from celebrities' YouTube videos [9]. VoxCeleb1 contains over 100,000 utterances from 1251 celebrities with 55% male speakers. *VoxCeleb2*, in turn, contains over 6000 celebrities (61% male). VoxCeleb2 is mainly used as a training set for ASV systems evaluations. The audio material can be considered as real-world *found data* including a variety of background noises, audio quality from different processing and recording devices, and speech style variations. It mostly consists of interviews in radio and TV programs, theaters, and red carpet. In the present study, we analyze the speaker variation of the entire VoxCeleb1's dataset, with speech from all the 1251 speakers, 561 female and 690 male, comprising 121,350 and 168,571 same speaker trials respectively. In contrast to our previous study where speakers were asked to disguise their voices [5], the speech variations in the VoxCeleb dataset correspond to the circumstances in which they are performed — whether a live-show interview with an audience, a radio or TV program in a formal or informal atmosphere.

3.2. ASV system

X-vector embedding [2] is based on speaker-discriminative training of a deep neural network model with a long temporal context. The x-vector system uses 30 mel-frequency cepstral coefficients (MFCCs) as input features, extracted from 25 ms frames, mean-normalized over a sliding window of three seconds. Non-speech frames are discarded with an energy-based speech activity detection. For speaker similarity scoring, *probabilistic linear discriminant analysis* (PLDA) is used as back-end [11]. In practice, we use the pre-trained x-vector recipe in Kaldi [12] trained on augmented VoxCeleb2 dataset [2]. Scoring this system on VoxCeleb1 (VoxCeleb1-E trial list) results in equal error rate (EER) of 2.54%.

3.3. Acoustic features

Table 1: *The mixed effect model uses a total of 23 predictor features, formed from the following combinations of features and their long-term statistical summary measures.*

Acoustic features, f		
F0	Fundamental frequency	F0
VQ	Loudness	
	Jitter	
	Shimmer	
	log Harmonic-to-noise-Ratio	HNR
	Spectral tilt	H1 – H2
		H1 – A3
Formant	Formant frequencies,	F1 to F4
	formant bandwidths,	B1 to B4
	formant amplitudes	A1 to A4
Spectral f.	Spectral flux	
Temporal	Voiced segments per second	
	Voiced segments length	
	Unvoiced segments length	

The selected acoustic features presented in Table 1 were extracted automatically using the openSMILE toolkit [8], which implements feature extraction at the frame level and provides summarization through statistical functionals at the utterance level. It has been used to serve applications such as emotion recognition, speaker trait analysis, and speaker recognition. Automatic extraction of features allows analysis of large datasets (such as VoxCeleb) for which phonetic annotations are not available. Even if feature extraction is performed without supervision (such as hand-made post-corrections), we expect a

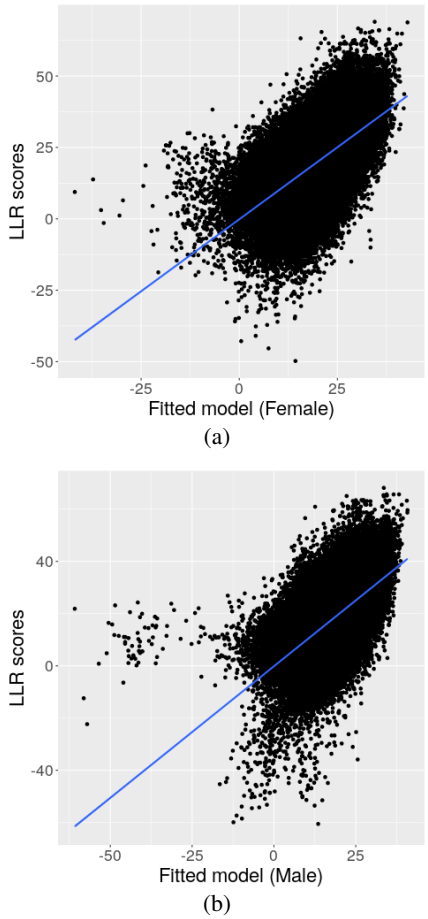

Figure 1: *Correlation for fitted model values and LLR scores (x-vector) with Pearson correlation r=0.60 for female and* $r = 0.58$ *for male speakers' trials*

reliable summary of the prominent acoustic variations presented in the VoxCeleb data. The selected feature extraction parameters are based on earlier work in the analysis of voice production changes related to affective states. This feature set is known as the *extended Geneva Minimalistic Acoustic Parameter Set* (eGeMAPS) [13]. The selected set of 23 acoustic features can be grouped as follows: F0, voice quality (VQ), formant, spectral flux and temporal.

3.4. Modeling the effect of acoustic variations

We investigate the effect that acoustic feature variation at the speaker level have on LLR score of target trials. Each trial was represented by the absolute difference of the mean and standard deviation of the features in Table 1. The two summary statistics (mean and standard deviation) for all the segments in the trial list were further standardized prior to the distance (absolute difference) computation. The original acoustic features have varied ranges and this normalization ensures that none of them dominate the distance computation.

For the LME model, the acoustic feature distances were used as the fixed effects. They were used both individually, and as groups of features. In the latter case, we simply sum up the distance values within a given feature group (example: all features within the voice quality group) to form a new predictor.

Table 2: *Parameter of the mixed effects model of x-vector ASV system's scores and acoustic feature groups variations with speakers as random effect.* r *used for feature group ranking.*

Male speakers				
Fixed effects:				
	Estimate	Std. error	t-value	r
β_0: Intercept	28.36	0.19	149.3	
β_1: F0	-1.02	0.02	-47.81	0.54
β_2: VQ	-0.36	0.006	-56.72	0.54
β_3: Formant 1	-0.20	0.01	-15.41	0.52
β_4: Formant 2	-0.15	0.01	-10.04	0.51
β_5: Formant 3	-0.16	0.01	-11.11	0.51
β_6: Formant 4	-0.31	0.01	-30.60	0.50
β_7: Temporal	-0.01	0.009	-1.56	0.48
β_8: Spectral flux	-0.29	0.007	-38.43	0.47
Random effects:				
	Variance			
Speaker: σ_b^2	4.67^2			
Residual:σ^2	9.01^2			

Female speakers				
Fixed effects:				
	Estimate	Std. error	t-value	r
β_0: Intercept	32.60	0.21	155.30	
β_1: F0	-1.01	0.03	-38.03	0.52
β_2: Formant 3	-0.37	0.02	-22.31	0.52
β_3: VQ	-0.25	0.008	-32.56	0.52
β_4: Formant 2	-0.41	0.02	-23.96	0.52
β_5: Formant 1	-0.29	0.01	-19.55	0.51
β_6: Formant 4	-0.32	0.01	-26.43	0.51
β_8: Spectral flux	-0.29	0.009	-32.51	0.47
β_7: Temporal	-0.13	0.01	-12.36	0.47
Random effects:				
	Variance			
Speaker: σ_b^2	4.5^2			
Residual:σ^2	9.3^2			

As random effects, we defined intercepts for each speaker. In this exploratory model, we seek to identify the feature variation that better explains the LLR score per trial and formulate speaker level interpretation of this relation. We first verified that our dependent variable, the LLR score, is approximately normally distributed, which is an assumption in our model. Visual inspection of density and quantile-quantile plots showed that even without a perfect normality, the assumption was met reasonably well for our model. We use the *lme4 package* [10] to fit the linear mixed effects model, using Wald's F-test to obtained the significance test.

3.5. Metrics

To evaluate the feature distances in terms of their added information to the model, we compared the correlation of the fitted model of individual feature distances with the LLR scores using Pearson correlation. The feature distances were ranked based on how their inclusion in the model increased the correlation of fitted model and LLR score. Also different models with groups of feature distances were compared using a standard likelihood test *ANOVA*. The *Akaike information criterion* (AIC) [14] value was used to compare the models and identified the model with better fit. The AIC value decreases with better models.

Table 3: *Correlation of fitted models from individual differences from Formant features (mean and standard deviation (SD) and LLR scores*

	Male								Female							
	Formant 1		Formant 2		Formant 3		Formant 4		Formant 1		Formant 2		Formant 3		Formant 4	
μ	F1	0.490	A2	0.480	A3	0.481	A4	0.482	F1	0.472	A2	0.474	A3	0.475	B4	0.478
	A1	0.477	F2	0.474	F3	0.463	B4	0.471	A1	0.470	F2	0.459	F3	0.462	A4	0.476
	B1	0.465	B2	0.469	B3	0.466	F4	0.457	B1	0.460	B2	0.455	B3	0.454	F4	0.463
σ	F1	0.474	A2	0.462	A3	0.466	A4	0.468	F1	0.455	F2	0.466	F3	0.463	F4	0.474
	B1	0.462	F2	0.460	F3	0.464	F4	0.468	A1	0.450	A2	0.455	A3	0.459	B4	0.471
	A1	0.459	B2	0.454	B3	0.451	B4	0.459	B1	0.443	B2	0.442	B3	0.450	A4	0.462

Table 4: *Correlation of fitted models from individual voice quality (VQ) feature (mean and standard deviation) and LLR scores*

	Male		Female	
μ	HNR	0.498	H1-A3	0.493
	H1-A3	0.497	HNR	0.478
	Loudness	0.487	Loudness	0.474
	H1-H2	0.470	H1-H2	0.456
	Shimmer	0.457	Shimmer	0.450
	Jitter	0.454	Jitter	0.444
σ	Loudness	0.473	Loudness	0.464
	HNR	0.471	HNR	0.458
	Shimmer	0.460	Shimmer	0.456
	H1-H2	0.453	H1-H2	0.440
	H1-A3	0.452	H1-A3	0.440
	Jitter	0.448	Jitter	0.437

4. Results

We analyze the change of acoustical features to explain the LLR score associated with the target trial's enrollment and test utterances of VoxCeleb1 data separated by gender. We fitted linear mixed effect models with the sum of feature distances corresponding to the feature group variation. The eight feature group distances models were fitted with speakers as the random effects. We compared the feature group models using the Pearson correlation between fitted values of the model and the LLR scores, a higher correlation coefficient (r) indicated the order in which the feature group were added the final model. Table 2 presents the regression coefficients for the final models for female and male speakers separately. The r coefficient was used for the ranking of the feature group in the model.

F0 is the feature group distance that contributes first to our explanatory model. It is worth mention that this feature group consist only of two measures, the F0 mean and standard deviation distances in semitone scale. While other feature groups consist of six to twelve feature distances with exception of spectral flux that also includes two distance measurements.

Figures 1 shows the correlation between the fitted model values and the x-vector's LLR scores. Visual inspection of residual plots did not reveal obvious deviations from homoscedasticity or normality. AIC and p-values were obtained by maximum likelihood ratio tests. Both gender models have a similar correlation coefficient, r of 0.6 for female and 0.58 for male speakers. The lower correlation coefficient is expected considering the variability not dependent on the speaker effect is high with residual error variation of 9.3^2 for females and 9.01^2 for males. The variation corresponding to the speaker effect is similar for female and male speakers, 4.5^2 and 4.67 respectively. Since all the trials' LLR scores were used in this exploratory model it is expected that some observations could be consider as "outliers" enabling the identification of a group of speakers' trials to further analyze.

4.1. Ranking of features in terms of their explanatory power

The feature ranking was based on the highest Pearson correlation between the model fitted with the feature groups and the LLR scores as shown in column r in Table 2. To analyze the importance of variations for independent feature's distances, models were fitted with each feature and the correlation to LLR scores was use to compare them. As mentioned in the previous section, F0 distance was the individual most important feature in the exploratory models. Then we analyzed the features in the voice quality and formant groups. Table 3 shows the ranking of formant features (frequency, bandwidth and amplitude) distances. For both genders, amplitudes and frequencies provide more information to the model in their respective formant group for mean and standard deviation of the feature distance.

Similar analysis was carried out for the voice quality features. Harmonic-to-noise-ratio and harmonic variation H1-A3 provided more information to the model as shown in the correlation of the fitted models and the LLR score presented in Table 4. The ranking was nearly consistent for both genders, being shimmer and jitter the lowest ranked feature distances in this feature group.

5. Conclusion

Why does a given automatic speaker verification system miss (reject) a target speaker? Ideally this should not happen in the first place, but when it does, it is useful to analyze the reasons. This may suggest ideas for future improvements of the recognition technology itself, inform users of the limitations of a given recognizer, or suggest ways of composing new evaluation corpora based on *found data*. No automatic speaker verification system is (or will likely ever be) completely immune to mismatch across enrollment and test data.

We approached the question from the perspective of regression analysis using a linear mixed effects model. The modeled variable is the LLR score of a speaker recognition system (here, x-vector PLDA) while the predictor variables consist of enrollment-vs-test distances in the first-order (mean) and second-order (standard deviation) statistics of selected acoustic features. We extended our previous work [5] in terms of the database and the acoustic features.

Overall, the acoustic variation impacts strongly the score of the ASV system. We found correlations up to ≈ 0.6 of the fitted model and the LLR score. Interestingly, our analysis confirms an important finding noted in [5] for a completely different corpus (but the same, Kaldi x-vector system): F0 mismatch plays a key role. Unsurprisingly, differences in formants and voice quality parameters contribute to degraded score, too.

6. References

[1] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech Communication*, vol. 52, no. 1, pp. 12 – 40, 2010.

[2] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5329–5333.

[3] S. J. D. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *IEEE 11th International Conference on Computer Vision, ICCV 2007, Rio de Janeiro, Brazil, October 14-20, 2007*. IEEE Computer Society, 2007, pp. 1–8. [Online]. Available: https://doi.org/10.1109/ICCV.2007.4409052

[4] M. I. Mandasari, R. Saeidi, M. McLaren, and D. A. van Leeuwen, "Quality measure functions for calibration of speaker recognition systems in various duration conditions," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 11, pp. 2425–2438, 2013.

[5] R. González Hautamäki, V. Hautamäki, and T. Kinnunen, "On the limits of automatic speaker verification: Explaining degraded recognizer scores through acoustic changes resulting from voice disguise," *The Journal of the Acoustical Society of America*, vol. 146, no. 1, pp. 693–704, 2019. [Online]. Available: https://doi.org/10.1121/1.5119240

[6] S. Wang, Y. Qian, and K. Yu, "What does the speaker embedding encode?" in *Proc. Interspeech 2017*, 2017, pp. 1497–1501. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2017-1125

[7] D. Raj, D. Snyder, D. Povey, and S. Khudanpur, "Probing the information encoded in x-vectors," in *IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2019, Singapore, December 14-18, 2019*. IEEE, 2019, pp. 726–733.

[8] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in openSMILE, the munich open-source multimedia feature extractor," in *MM 2013 - Proceedings of the 2013 ACM Multimedia Conference*, 10 2013, pp. 835–838.

[9] A. Nagrani, J. S. Chung, W. Xie, and A. Zisserman, "Voxceleb: Large-scale speaker verification in the wild," *Computer Science and Language*, 2019.

[10] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015.

[11] S. J. D. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *Proc. of International Conference on Computer Vision (ICCV)*. IEEE, 2007, pp. 1–8.

[12] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, Dec. 2011, iEEE Catalog No.: CFP11SRW-USB.

[13] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong, "The geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2016.

[14] H. Akaike, "A new look at the statistical model identification," *IEEE Transactions on Automatic Control*, vol. 19, no. 6, pp. 716–723, Dec 1974.