The Speaker and Language Recognition Workshop (Odyssey 2022)
28 June - 1 July 2022, Beijing, China

Time-varying score reliability prediction in speaker identification

Sarah Bakst, Chris Cobo-Kroenke, Aaron Lawson, Mitchell McLaren, Allen Stauffer

STAR Lab
SRI International, Menlo Park, CA
{sarah.bakst,chris.cobo-kroenke,aaron.lawson,mitchell.mclaren,allen.stauffer}@sri.com

Abstract

The present work proposes a method for estimating confidence in speaker identification scores in the time domain. The motivation for this work comes from forensic fingerprinting, where confidence in a captured fingerprint's capacity to successfully identify its source is determined by the clarity of the print, but, crucially, this clarity may not be consistent across the print. Hicklin et al. [1] propose a standard for assessing confidence in different regions of a fingerprint based on this clarity, allowing for nuanced analysis of fingerprint biometric information. Speech audio poses a similar problem for speaker identification (SID), where there may be variability in the reliability of the scores output by the SID system for different time segments of an audio sample. Here we evaluate acoustic characteristics that can be directly measured from audio to evaluate SID score reliability over time segments of equal length.[1]

1. Introduction

Speaker identification (SID) aims to identify an unknown speaker of an audio sample. This is done by assessing whether the speech in the sample of unknown source is sufficiently similar to a known speaker, but the reliability of this assessment is impacted by the quality of the audio. The experiment here sought to identify directly measurable acoustic qualities of audio that predict how SID score reliability may vary throughout an audio sample as audio conditions change. We investigate both speaker-*intrinsic* factors (specific to the speech or speaker, such as amount of voicing or native language) and speaker-*extrinsic* factors (specific to the recording environment, such as channel or background noise).

Previous measures of confidence that can be directly quantified from audio have mainly focused on qualities that are related to clarity of the signal. Signal-to-noise ratio (SNR) has been commonly considered [2, 3, 4, 5, 6], as well as degradation due to reverberation [2], or other forms of signal degradation [2, 3]. Shimmer and jitter [3], the cycle-to-cycle changes in amplitude and frequency (respectively) of a periodic signal, have also been considered as predictors of SID confidence. Additional quantitative measures have included duration [5, 6] and number of speech frames [3]. Models that discarded trials based on some measure of noise tended to show reduced equal error rate (EER). Metadata that are not as directly measurable, such as channel type/mismatch [5, 6], emotional valence/arousal [7], and language [8, 9] have also been shown to improve estimations of confidence.

Further, many of these studies assume that these acoustic qualities pertain to an entire audio file. The present study considers short time segments to determine whether it would be possible to capture how the reliability of SID scores might vary of the course of a sample.

2. Methods

2.1. Data

We developed acoustic quality candidates using the Language and Speaker Recognition (LASR) corpus [10]. Three hundred speakers were recorded both in their native language (100 each of Arabic, Spanish, and Korean) as well as English (varying non-native proficiency and experience). Speakers were recorded both reading a set of texts designed to sample the phonetic space as well as engaging in conversation with another speaker; the interlocutor's speech was omitted from the dataset analyzed here. There were two recording sessions for every speaker in every language and style condition. Seven microphones of varying quality and distance from the speaker (camcorder (Cm), desk mounted (Dm), omnidirectional (Om), studio (Sm), and two[2] telephone microphones (Tm and Tk)) simultaneously recorded each speaker at a sample rate of 16 kHz.

2.2. Investigated factors

2.2.1. Audio quality measures

SNR was hypothesized to predict reliability in SID scores, with cleaner audio resulting in more reliable scores. Two SNR algorithms, the NIST Quick SNR and Wada SNR, are tested here. Kurtosis of LP residuals, a measure of reverberation [11, 12], was hypothesized to predict reliability because distortion caused by reverberation could prevent accurate identification or embedding extraction.

Speech activity detection (SAD) was also hypothesized to predict SID reliability, with a greater proportion of speech hypothesized to result in more reliable scores. We used our own SAD system, which is DNN-based with two hidden layers containing 500 and 100 nodes, respectively. The SAD DNN is trained using 20-dimensional Mel-frequency cepstral coefficients (MFCC) features, stacked with 31 frames. Before training the SAD DNN, the features were mean and variance normalized over a 201-frame window. The output scores are smoothed with a 0.5-sec long window. Finally, the detected

[1]NOTICE This software (or technical data) was produced for the U. S. Government and is subject to the Rights in Data-General Clause 52.227-14, Alt. IV (May 2014) – Alternative IV (Dec 2007)

[2]A third telephone microphone is excluded from the analysis here because it recorded traces of the interlocutor's voice.

10.21437/Odyssey.2022-29

speech segments are padded by a third of a secondSeveral measures were taken: mean and median overal SAD LLR, mean of LLRs greater than 0, mean of LLRs greater than 0.5, and mean of LLRs greater than 1.

2.2.2. *Voice qualities*

Many speaker characteristics that have been identified as useful for speaker identification are properties of source (vocal fold) harmonics, such as harmonics-to-noise ratio (HNR [13]), or how those harmonics are filtered (vocal tract resonances), i.e. formants and their dynamics [14, 15, 16]. Most of these characteristics are more difficult, if not impossible, to measure in unvoiced speech. We therefore hypothesized that regions of audio containing higher proportions of voiced speech would produce more reliable SID scores than regions containing more unvoiced speech (voiceless consonants or whispered speech). We measured several qualities related to voicing: mean voicing probability score (each frame given a binary score reflecting likelihood of voicing), mean autocorrelation peak (as a measure of periodicity), and mean and standard deviation of spectral tilt. Standard deviation of spectral tilt was predicted to highlight audio segments that might contain anomalies such as a prolonged vowel, which may provide less information than multiple syllables. Similarly, a higher standard deviation in spectral tilt would suggest larger a variety of speech segments that might paint a fuller picture of a speaker's voice. All voicing qualities were computed with modifications to *get_f0* [17, 18]. We also investigated several factors which we expected to reflect aspects of the recording environment or equipment: shimmer, jitter, and HNR.

2.3. Analysis

All statistics were done in R [19].

2.3.1. *SID systems*

Speaker recognition experiments in this work are based on speaker embeddings using two different backends: PLDA [20] and DPLDA [21].

We use a multi-bandwidth speaker embeddings network based on 16 kHz PNCC features of 30 dimensions extracted from a bandwidth of 100–7600 Hz using 40 filters and root compression of 1/15. We train on both 8kHz and 16kHz audio; 8kHz audio was upsampled to 16 kHz prior to feature extraction. The architecture of our embeddings extractor DNN follows the Kaldi recipe [22], and specific details on data and our general augmentation process can be found in [23].

The training data for the backend was a one-third sampling of the embeddings training dataset. All embeddings are first transformed by LDA from 512 to 150 dimensions. Mean normalization is then applied to the LDA-reduced embeddings, with the mean calculated from the backend training dataset. Length normalization [24] is then applied before PLDA [25] or DPLDA modeling [21].

2.3.2. *Performance testing of development dataset*

All sound files in the development set were divided into two-second segments. For every language, speaking condition, and microphone, speakers were enrolled in the SID system using segments from one recording session and were tested on the other session. There were no cross-condition trials. For example, an enrolled (first session) sample of speech that came from a native Spanish speaker producing English conversation recorded by the camcorder microphone would only be tested against segments containing native Spanish speakers producing English conversation as recorded by the camcorder microphone in the second session. Each trial produced a log-likelihood ratio (LLR) reflecting the likelihood that the test sound file was produced by the same speaker in the enrolled file. For each comparison of enroll vs. test audio files, the 2-s segments of the enrollment were exhaustively compared to the 2-s segments of the test file. These formed target trials, while the non-target trials were formed by exhaustively comparing each enrollment file to each test file from the non-target speakers. This allowed us to generate an EER for each test file. Candidate acoustic qualities were assessed by their correlation with EER (Pearson's r). EER was also used to assess systematic differences in condition on SID performance. All frame-level acoustic qualities were extracted every 10ms from the 2-s audio segments.

3. Results

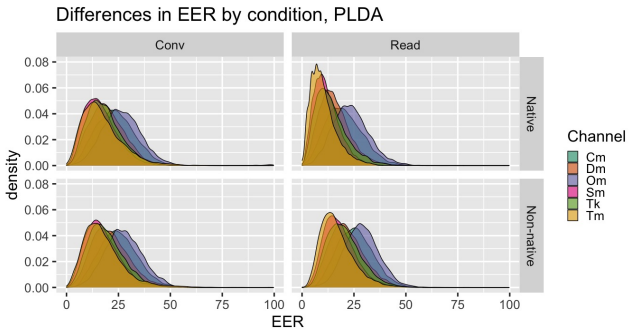

Figure 1: EER distribution by condition for PLDA system

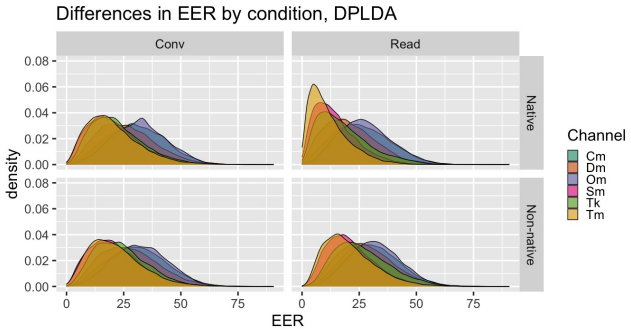

Figure 2: EER distribution by condition for DPLDA system

One possibility for the observed differences in nativeness is that a speaker may produce speech with fewer pauses in the native language, and indeed the mean SAD average LLR was greater for natively produced read speech (3.4) than non-natively produced read speech (3.0), indicating a possibility that there were fewer speech-less frames in the native language condition for reading. However, there was no difference in conversational speech (3.2).

3.1. Metadata indicators of performance

The plots in Figures 1 and 2 show EER differences in performance among the recording conditions, with better performance

(left-shifted distributions) for closer-talking microphones (Sm, Tk, Tm), and better scores for read speech (right columns) rather than conversational (left columns). There were pronounced differences when speakers were producing their native language (top rows) relative to English (non-native, bottom rows), but this is most clearly seen during the read condition, where the difference in average EER was 4% for both systems. These differences persisted across system type (PLDA vs. DPLDA). There was also better performance for read speech than conversational in the native speaking condition only (4% benefit in PDLA, 5% for DPLDA).

3.2. EER prediction

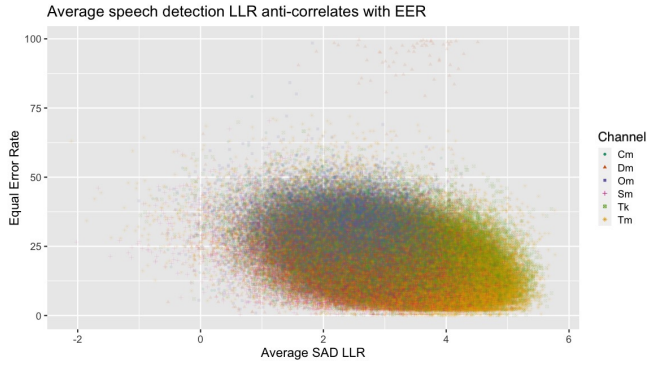

Figure 3: Speech activity predicts EER (PLDA shown)

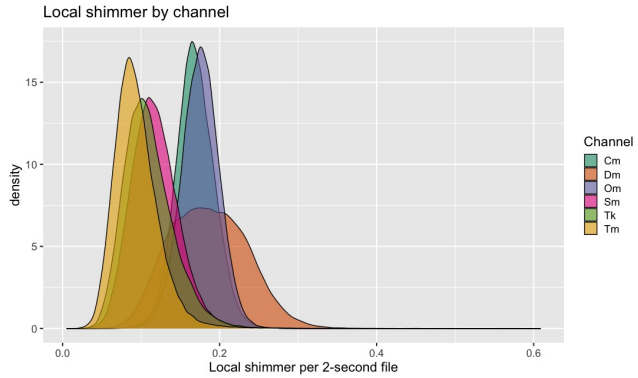

Figure 4: Microphone channels are well-separated by shimmer values.

A full table of Pearson's correlation coefficients can be found in the table in Section 6; because this work is exploratory, no p-values are reported because significance may have been inflated due to the sheer number of comparisons. All correlations were weak to moderate. In the results that follow, the first and second r values given corresponds to the PLDA and DPLDA systems, respectively. The strongest correlations with EER were found for average SAD LLR score ($r = -0.4, -0.33$, shown in Figure 3), standard deviation of spectral tilt ($r = -0.31, -0.28$), DC offset ($r = 0.26, 0.19$), and kurtosis of LP residuals ($r = -0.27, -0.26$). HNR had a similar correlation with EER ($r = -0.21, -0.2$) as NISTQuick SNR ($r = -0.22, -0.23$). Shimmer showed a positive relationship with EER ($r = 0.21, 0.21$) All other relationships had

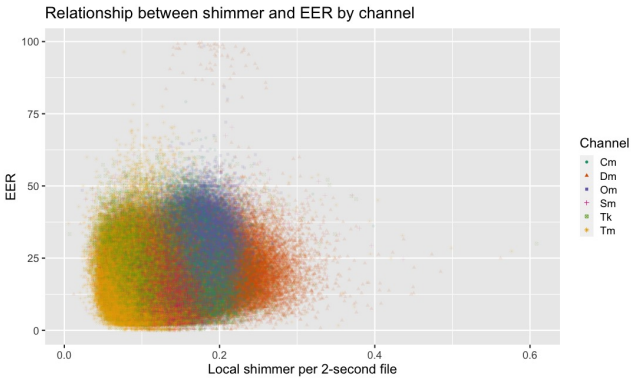

Figure 5: The relationship between EER (PLDA results shown) and shimmer is driven by microphone channel.

correlation coefficients of $|r| < 0.15$.

We compared the power of these acoustic qualities to predict EER in both the PLDA and DPLDA systems. The two systems performed comparably, with an average EER of 24% for DPLDA and 19.5% for PLDA, and scores showed a strong correlation between the two systems ($r = 0.63$).

Some acoustic qualities predicted EER *because* they predicted channel. For example, distributions of local shimmer values were well-defined by microphone channel, as shown in Figure 4. Indeed, there was a relationship between shimmer and EER, but this relationship was almost entirely driven by channel type: within channel, correlations between shimmer and EER ranged from -0.15 (camcorder microphone) to 0.06 (desk microphone), but the correlation over the entire set was 0.21, which appears to be in part driven by higher EER values in channel Om (Figure 5). In this way, these acoustic qualities were capable of making coarse-grained predictions due to recording condition, but they were not able to make finer-grained predictions about which files were likely to show better performance *within* a condition.

3.3. Heteroskedasticity

Several acoustic qualities measured here exhibited a heteroskedastic relationship between the value of the acoustic quality and the EER of the file: in many cases, one extreme of the quality predicted low EER, but the other extreme of the same quality did not necessarily predict high EER. For example, high standard deviation of spectral tilt was only measured in files with low EER, but files with a low standard deviation in spectral tilt could correspond to files with a range of EER values, as shown in Figure 6.In this case, there may have been multiple sources of low standard deviation in spectral tilt, such as environmental (low-fidelity microphone response) or speech content (a single elongated vowel during the audio segment). Other qualities showing a heteroskedastic (and often weak) relationship with EER included shimmer, F0 standard deviation, and kurtosis of LP residuals. For all of these acoustic qualities, one extreme of values predicted *low* EER, but the other extreme did not make predictions about whether the audio segment could be used to give a reliable score or not.

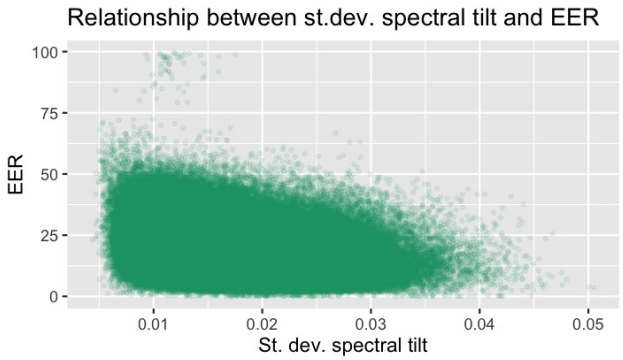

Figure 6: The relationship between standard deviation of spectral tilt and EER (here PLDA is shown) was heteroskedastic, with high values of the quality predicting low EER but low values of the quality having less predictive power.

4. Discussion

Several acoustic qualities tested here showed moderate power in predicting EER of a short segment of audio. Average SAD LLR score relates to the amount of speech that was detected, so it is unsurprising that that it showed a strong relationship with EER. The standard deviation of spectral tilt, which also had one of the stronger correlations with EER, may reflect variability in the representation of the phonetic inventory, meaning that higher values would indicate a greater variety of phone types available in the sample. Low values may indicate a smaller number of phone types, but it could also reflect a particular channel's spectral smearing or fidelity, which could also contribute to a higher EER.

NIST Quick SNR had a moderate correlation with EER ($r = -0.22$), suggesting that, as expected, the noise on a channel or in the background can hinder SID performance. HNR was similarly correlated with EER, but it was also well-correlated with NIST Quick SNR ($r = 0.64$), suggesting that both measures may have been measuring some of the same qualities in the audio.

Contra the hypothesis that voiced speech may prove more predictive of SID performance than unvoiced speech, there was a negligible negative relationship ($r = -0.06, -0.02$) between mean frame voicing probability and EER. This result may indicate that while there may be many possible measurable qualities of voiced speech that are useful for distinguishing speakers from each other, qualities of unvoiced speech (e.g. sibilant peak frequency) may also be powerful factors.

EER also differed by recording condition, replicating the microphone channel results from earlier studies discussed above. Additionally, the type of speech context (reading or conversational) as well as nativeness may also be important factors, at least in shorter snippets of audio. Better SID performance in native read speech relative to non-native read speech may be explainable by a greater presence of frames containing speech in this condition, perhaps due to increased fluency. The amount of speech did not differ between native and non-native speech in the conversation context, so it remains unclear what is driving the performance difference between speakers in this condition.

5. Conclusion

Several acoustic qualities that can be directly measured from audio were shown to have some ability to predict the reliability of SID scores, even in very short segments of audio. Combining these qualities with audio metadata, such as language spoken, channel, and speaking style could be used to produce nuanced reliability scores that vary throughout an audio file, allowing SID users to extract the parts of a file that are most useful while discarding or down-weighting minimal data. The results here did not differ significantly between the PLDA and DPLDA SID systems, suggesting these results have the potential to be robust to different systems.

6. Appendix

Correlations between acoustic qualities and both EER measurements is shown in Table 1. A full table of correlations among the qualities tested is shown in Table 2.

	DPLDA EER	PLDA EER
DPLDA EER	1.00	0.63
PLDA EER	0.63	1.00
DCOffset	0.19	**0.26**
Kurtosis	**-0.26**	**-0.27**
WadaSNR	-0.13	-0.11
NISTQuickSNR	-0.23	-0.22
SADMnLLR	**-0.33**	**-0.40**
SADMnLLR>0	-0.06	-0.07
SADMn>.5	-0.04	-0.05
SADMnLLR>1	-0.04	-0.05
SADMedLLR	**-0.32**	**-0.39**
stdevF0Hz	0.07	0.06
HNR	-0.20	-0.21
localJitter	0.10	0.11
localShimmer	0.21	0.21
voiProbMn	-0.02	-0.06
acPkMn	-0.11	-0.14
acPkMed	-0.15	-0.17
tiltMn	0.00	-0.01
tiltSd	**-0.28**	**-0.31**

Table 1: Correlation table between both EER measures and all acoustic measures. Correlations stronger than $\pm$ 0.25 are bolded.

7. References

[1] R Austin Hicklin, JoAnn Buscaglia, and Maria Antonia Roberts, "Assessing the clarity of friction ridge impressions," *Forensic science international*, vol. 226, no. 1-3, pp. 106–117, 2013.

[2] Jesús Villalba, Eduardo Lleida, Alfonso Ortega, and Antonio Miguel, "Reliability estimation of the speaker verification decisions using bayesian networks to combine information from multiple speech quality measures," in *Advances in Speech and Language Technologies for Iberian Languages*, Berlin, Heidelberg, 2012, pp. 1–10, Springer.

[3] Jesus Villalba, Alfonso Ortega, Antonio Miguel, and Eduardo Lleida, "Analysis of speech quality measures for the task of estimating the reliability of speaker verification decisions," *Speech Communication*, vol. 78, pp. 42–61, 2016.

	DC Offset	LPresKurtosis	Wada SNR	NISTQuick SNR	SADmnLLR	SADmnLLR> 0	SADmnLLR> .5	SADmnLLR> 1	SADmedLLR	stdDvF0	HNR	Jitter	Shimmer	MnVoiProb	ACPkMn	ACPkMed	specTltMn	specTltStDv
DCOffset	1.00	-0.31	-0.15	-0.16	-0.22	-0.03	-0.01	0.00	-0.22	0.00	-0.20	0.00	0.25	0.06	-0.03	-0.05	-0.23	-0.34
LPresKurtosis	-0.31	1.00	**0.72**	**0.71**	0.24	0.05	0.03	0.02	0.24	-0.15	**0.46**	-0.22	**-0.53**	-0.34	-0.15	-0.11	0.24	**0.50**
WadaSNR	-0.15	**0.72**	1.00	**0.84**	0.06	0.02	0.02	0.02	0.08	-0.18	**0.48**	-0.26	**-0.54**	**-0.54**	-0.37	-0.34	0.28	**0.45**
NISTQuick SNR	-0.16	**0.71**	**0.84**	1.00	0.25	0.05	0.03	0.03	0.27	-0.25	**0.64**	-0.32	**-0.67**	**-0.47**	-0.21	-0.12	0.19	**0.57**
SADmnLLR	-0.22	0.24	0.06	0.25	1.00	0.17	0.12	0.12	**0.96**	-0.07	0.31	-0.26	-0.30	0.24	0.33	0.39	0.03	0.35
SADmnLLR>0	-0.03	0.05	0.02	0.05	0.17	1.00	**0.47**	0.28	0.16	-0.02	0.06	-0.07	-0.06	0.05	0.06	0.06	0.01	0.06
SADmnLLR>.5	-0.01	0.03	0.02	0.03	0.12	**0.47**	1.00	**0.58**	0.11	-0.01	0.04	-0.07	-0.04	0.05	0.05	0.05	0.00	0.03
SADmnLLR>1	0.00	0.02	0.02	0.03	0.12	0.28	**0.58**	1.00	0.11	0.00	0.05	-0.07	-0.04	0.06	0.05	0.04	-0.01	0.02
SADMedLLR	-0.22	0.24	0.08	0.27	**0.96**	0.16	0.11	0.11	1.00	-0.06	0.31	-0.26	-0.30	0.21	0.31	0.37	0.04	0.34
stdDvF0	0.00	-0.15	-0.18	-0.25	-0.07	-0.02	-0.01	0.00	-0.06	1.00	-0.29	0.33	0.35	0.16	-0.05	-0.08	0.13	-0.15
HNR	-0.20	**0.46**	**0.48**	**0.64**	0.31	0.06	0.04	0.05	0.31	-0.29	1.00	**-0.45**	**-0.75**	-0.03	0.32	0.37	0.00	**0.41**
Jitter	0.00	-0.22	-0.26	-0.32	-0.26	-0.07	-0.07	-0.07	-0.26	0.33	**-0.45**	1.00	**0.49**	0.08	-0.11	-0.14	-0.10	-0.17
Shimmer	0.25	**-0.53**	**-0.54**	**-0.67**	-0.30	-0.06	-0.04	-0.04	-0.30	0.35	**-0.75**	**0.49**	1.00	0.33	0.03	-0.06	-0.27	**-0.47**
MnVoiProb	0.06	-0.34	**-0.54**	**-0.47**	0.24	0.05	0.05	0.06	0.21	0.16	-0.03	0.08	0.33	1.00	**0.84**	**0.75**	**-0.42**	-0.29
ACPkMn	-0.03	-0.15	-0.37	-0.21	0.33	0.06	0.05	0.05	0.31	-0.05	0.32	-0.11	0.03	**0.84**	1.00	**0.93**	**-0.54**	-0.13
ACPkMed	-0.05	-0.11	-0.34	-0.12	0.39	0.06	0.05	0.04	0.37	-0.08	0.37	-0.14	-0.06	**0.75**	**0.93**	1.00	**-0.46**	-0.05
specTltMn	-0.23	0.24	0.28	0.19	0.03	0.01	0.00	-0.01	0.04	0.13	0.00	-0.10	-0.27	**-0.42**	**-0.54**	**-0.46**	1.00	0.17
specTltStDv	-0.34	**0.50**	**0.45**	**0.57**	0.35	0.06	0.03	0.02	0.34	-0.15	**0.41**	-0.17	**-0.47**	-0.29	-0.13	-0.05	0.17	1.00

Table 2: Correlation table for all features tested in this project. Correlations stronger than ± 0.4 are bolded.

[4] Jonas Richiardi, Plamen Prodanov, and Andrzej Drygajlo, "Speaker verification with confidence and reliability measures," in *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*. IEEE, 2006, vol. 1, pp. I–I.

[5] Mark C. Huggins and John. J. Grieco, "Confidence metrics for speaker identification," in *Seventh International Conference on Spoken Language Processing.*, 2002.

[6] William M Campbell, Douglas A Reynolds, Joseph P Campbell, and KJ Brady, "Estimating and evaluating confidence for forensic speaker recognition," in *Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*. IEEE, 2005, vol. 1, pp. I–717–720.

[7] Srinivas Parthasarathy and Carlos Busso, "Predicting speaker recognition reliability by considering emotional content," in *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2017.

[8] Liang Lu, Yuan Dong, Xianyu Zhao, Jiqing Liu, and Haila Wang, "The effect of language factors for robust speaker recognition," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2009, pp. 4217–4220.

[9] Gianni Fenu, Giacomo Medda, Mirko Marras, and Giacomo Meloni, "Improving fairness in speaker recognition," in *Proceedings of the 2020 European Symposium on Software Engineering*, 2020, pp. 129–136.

[10] Steven D. Beck, Reva Schwartz, and Hirotaka Nakasone, "A bilingual multi-modal corpus for language and speaker recognition (LASR) services," in *Odyssey: The Speaker and Language Recognition Workshop*, 2004.

[11] B.W. Gillespie, H.S. Malvar, and D.A.F. Florencio, "Speech dereverberation via maximum-kurtosis subband adaptive filtering," in *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)*, 2001, vol. 6, pp. 3701–3704 vol.6.

[12] Kohei Hayashida, Masato Nakayama, Takanobu Nishiura, Yoichi Yamashita, Toshiharu Horiuchi, and Tsuneo Kato, "Close/distant talker discrimination based on kurtosis of linear prediction residual signals," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 2327–2331.

[13] Soo Jin Park, Gary Yeung, Jody Kreiman, Patricia A Keating, and Abeer Alwan, "Using voice quality features to improve short-utterance, text-independent speaker verification systems," in *Interspeech*, 2017, pp. 1522–1526.

[14] Carol Y Espy-Wilson, Sandeep Manocha, and Srikanth Vishnubhotla, "A new set of features for text-independent speaker identification," in *Ninth International Conference on Spoken Language Processing*, 2006.

[15] Lei He, Yu Zhang, and Volker Dellwo, "Between-speaker variability and temporal organization of the first formant," *The Journal of the Acoustical Society of America*, vol. 145, no. 3, pp. EL209–EL214, 2019.

[16] Kirsty McDougall and Francis Nolan, "Discrimination of speakers using the formant dynamics of/u/in british english," in *Proceedings of the International Congress of Phonetic Sciences*, 2007, pp. 1825–1828.

[17] David Talkin, *Speech Coding and Synthesis*, chapter A robust algorithm for pitch tracking (RAPT), Elsevier Science, 1995.

[18] Luciana Ferrer, Harry Bratt, Colleen Richey, Horacio Franco, Victor Abrash, and Kristin Precoda, "Classification of lexical stress using spectral and prosodic features for computer-assisted language learning systems," *Speech Communication*, Feb 2015.

[19] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2020, https://www.R-project.org.

[20] Mitchell McLaren, Aaron Lawson, Luciana Ferrer, Nicolas Scheffer, and Yun Lei, "Trial-based calibration for speaker recognition in unseen conditions," in *Proc. Odyssey*, 2014, pp. 19–25.

[21] Luciana Ferrer, Mitchell McLaren, and Niko Brümmer, "A speaker verification backend with robust performance across conditions," *Computer Speech & Language*, vol. 71, no. 101258, 2022.

[22] *NIST SRE 2016 Xvector Recipe*, https://david-ryan-snyder.github.io/2017/10/04/model_sre16_v2.html, 2017.

[23] ML McLaren, Diego Castan, Mahesh Kumar Nandwana, Luciana Ferrer, and Emre Yilmaz, "How to train your speaker embeddings extractor," 2018.

[24] Daniel Garcia-Romero and Carol Y Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," in *Twelfth annual conference of the international speech communication association*, 2011.

[25] Simon JD Prince and James H Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *2007 IEEE 11th International Conference on Computer Vision*. IEEE, 2007, pp. 1–8.