Explainability of speaker recognition system via phonetic traits

Ondrej Šuch[a], Muhammad Azeem[a,b], Ali Haidar[a,b], Ondrej Škvarek[c]

[a]Matematický ústav SAV, Ďumbierska 1, Banská Bystrica, Slovakia
[b]Fakulta matematiky, fyziky a informatiky, Univerzita Komenského v Bratislave, Mlynská dolina, Bratislava, Slovakia
[c]Fakulta riadenia a informatiky, Žilinská univerzita v Žiline, Univerzitná 8215/1, Žilina, Slovakia

***Abstract*—Explainability of deep neural networks remains an important research topic in order to provide security guarantees in applications. Explanations of neural networks processing sound are more challenging than those processing images, because texture-based explanations do not map cleanly to human auditory perception. We propose to connect explanations of sound processing networks with phonetic features. To that end we employ statistical techniques relying on binary classification. We show that a probabilistic version of linear discriminant analysis is a good fit for binary classification of two speakers. In order to provide a global understanding of a multi-speaker system, we employ classical multidimensional scaling. Analysis of phonetic properties of the system shows that the majority of variation of the principal axis can be attributed to varying formant frequencies of a cardinal vowel.**

***Index Terms*—Speaker identification, biometry, linear discriminant analysis, explainability of deep neural networks.**

I. Introduction

With the advancement of speech and communication technologies, human voice is quickly becoming the most important interaction interface with automated systems. There are two complementary aspects of human voice that such systems may need to understand:

- The first aspect is *what* was said. This is the subject of automated speech recognition systems which aim to convert an audio signal containing speech into text [1].
- The second aspect is *who* was speaking. This is the aim of systems for speaker verification that aim to verify that the speaker is who he/she claims to be, or speaker identification systems, which aim to identify the speaker among a set of possibilities [2]. The former is usually done in the context of biometric identification, whereas the latter has found important applications in diarization.

Computer comprehension of both aspects of speech has dramatically increased with the arrival of deep neural networks [3], first with convolutional structure [4], and then with transformers [5]. One potential drawback of deep neural networks is that they are generally considered black boxes, whose internal workings are rather poorly understood. While industry-standard evaluation techniques allow for managing the errors of systems based on deep neural networks, enhanced understanding of the networks would increase their trustworthiness and allow for quality comparison.

This work was supported by grant VEGA 2/0056/25 Cycles and edge colorings of cubic graphs, by the EU NextGenerationEU through the Recovery and Resilience Plan for Slovakia under the project 17I04-04-V05-00114 (AI-ONKO), and by Doktogrants APP0766, APP0801 from Slovenská Akadémia Vied.

Lack of interpretability of deep neural networks may be attributed to several intrinsic factors. The first is their sheer complexity. Models performing language processing may have billions of weights, which makes interpretation or fault-finding extremely hard. The second is distributed storage of information in these networks, which makes attribution of behavioral properties to local parts challenging. Neural networks for processing of speech signal face an additional hurdle in that speech is not visible to human eyes. Therefore, for speech neural networks many image-based interpretations do not readily apply. Fortunately, the last factor could be addressed using methods from linguistics.

Interpretation of speech signal has a very long scientific history, and predates the current machine learning revolution by many decades. We can highlight several key research milestones. The first was the development of the concept of formants [6], which allowed researchers to characterize and synthesize vocalic segments (vowels). Another was a better understanding of speech spectra, which, for instance, allowed one to characterize and categorize non-vocalic segments of speech. For modern digital applications the rediscovery of the fast Fourier transform was paramount to efficient computability of the speech spectrum [7]. Another important achievement was the formulation of the source filter theory of speech [8], which underlies all modern VoIP applications. These and related concepts are being studied in phonetics and related fields to this day.

Despite the potential usefulness of phonetic information for explaining automated speech processing systems [9], published literature on the topic has so far been limited [9], [10]. In this note we aim to bring attention to phonetic aspects of speaker identification systems. We will study functional interpretability [11] which circumvents the complexity hurdle posed by the enormous size of underlying neural networks.

II. Methods

In order to carry out a phonetic explanation of a speaker identification system one needs to decide on

- speaker recognition system, in our case ECAPA-TDNN (cf. Subsection II-A),
- phonetically annotated dataset, in our case TIMIT (cf. Subsection II-B),
- samples provided to the speaker identification neural network (cf. Subsection II-C).

A. Speaker recognition using ECAPA-TDNN

ECAPA-TDNN is a modern, well-performing deep neural network for speaker identification [12]. It uses standard neural network building blocks to provide 192-dimensional embeddings of speaker utterances. For training, the authors used the multilingual VoxCeleb2 dataset [13]. Samples from the dataset were augmented in multiple ways:

- adding noise and babble using MUSAN dataset [14],
- adding reverberations,
- speeding up or slowing down samples,
- reencoding using different codecs.

B. TIMIT dataset

The TIMIT dataset is a collection of recordings from 630 speakers of American English. Each speaker spoke 10 sentences, two of which (coded SA1 and SA2) are common to all speakers. The dataset was annotated by providing words of each sentence, as well as a decomposition of words into phonemes with precise phoneme boundaries. The sentences are sampled at 16kHz.

In 2006 Li Deng et al. published vocal tract resonances (first three formants) for a portion of the TIMIT dataset [15]. This is the so-called VTR dataset.

C. Segment sampling

In order to train binary classifiers, we prepared segments in the following way.

For the binary classifier of Section III, we prepared 640 samples for each speaker based on sentences other than SA1/SA2, all of length 400 ms (chosen to include several phonemes). Of those, 64 samples were clean, and the rest had added noise ranging from 5 dB to 20 dB signal-to-noise ratio:

- white noise at levels 5, 10, 15, 20 dB,
- pink noise at levels 5, 10, 15, 20 dB,
- brown noise at 10 dB.

In Section IV, we used only clean samples of length 400 ms. We prepared 20 segments from each sentence, totaling 200 samples per speaker.

III. Contrasting two speakers

Finer insight into a speaker recognition system can be gained by considering two speakers at a time. The key step is building a probabilistic classification model for embeddings from two speakers, which immediately yields a one-dimensional projection of the very high-dimensional embedding space (via probabilistic class prediction). Thereafter, the projection can be readily visualized.

In the context of speaker identification systems, a popular choice for the probabilistic model is probabilistic linear discriminant analysis (PLDA) [9], [12], [16].

In order to apply PLDA to speaker embeddings one has to resolve two potential difficulties. The first is that speaker embeddings should be compared using angular distance between the embedding vectors, which results from *spherical geometry*. This can be addressed by normalizing embeddings to have unit $L2$ norm and approximating that, in the case of two speakers, the embedding space is nearly linear. This essentially reduces the number of features by one and may manifest as singular behavior while fitting a probabilistic model. This latter difficulty is often resolved using a projection method like PCA [16]. We note that this required selecting one hyperparameter—the dimension of the projection. We chose 185 and we did not encountered numerical issues.

There are other possible alternatives to PLDA, e.g. variants of logistic regression. The plain version of logistic regression usually does not apply because the high quality of the speaker identification network often results in linearly separable data. This can be addressed using either Firth logistic regression (Firth LR) [17], or by using penalized logistic regression (L2 LR) [18]. We compare the predictions of these two models with PLDA in Fig. 1. We can see that while Firth logistic regression is close to penalized logistic regression, PLDA predictions are significantly more confident. Based on the phonetic annotation we can see that the confidence of the prediction rose significantly once the network encountered the /iy/ vowel. This sound is highly distinctive, since it corresponds to an extreme position of the articulators, resulting in a so-called cardinal vowel [19].

A. LDA model assumptions

Recall that PLDA assumes that two classes (in our case two speakers) are normally distributed in the feature space with equal covariance matrices. These assumptions are usually hard to check in high-dimensional data. We will provide several pieces of evidence supporting PLDA as a good choice of binary classification method. First, the training method of ECAPA-TDNN, which clusters together the embeddings of one speaker around a single class weight vector, is consistent with one aspect of normality - unimodality of the distribution. Secondly, we compute statistical tests for normality and homoscedasticity for the *one-dimensional* projection via log-odds. The illustrative results in Tables I and II show the projection passes both tests with one borderline outcome. Finally, we also provide a visualization of the log-odds distribution via a Q-Q plot in Fig. 2. Again, the good match with the normal distribution bodes well for using PLDA+PCA as a classification method.

IV. Contrasting samples from all speakers

In order to analyze global properties of speaker identification system we can use classical multidimensional scaling (MDS) [20]. This statistical method finds a low-dimensional representation of instances if a matrix of pairwise distances

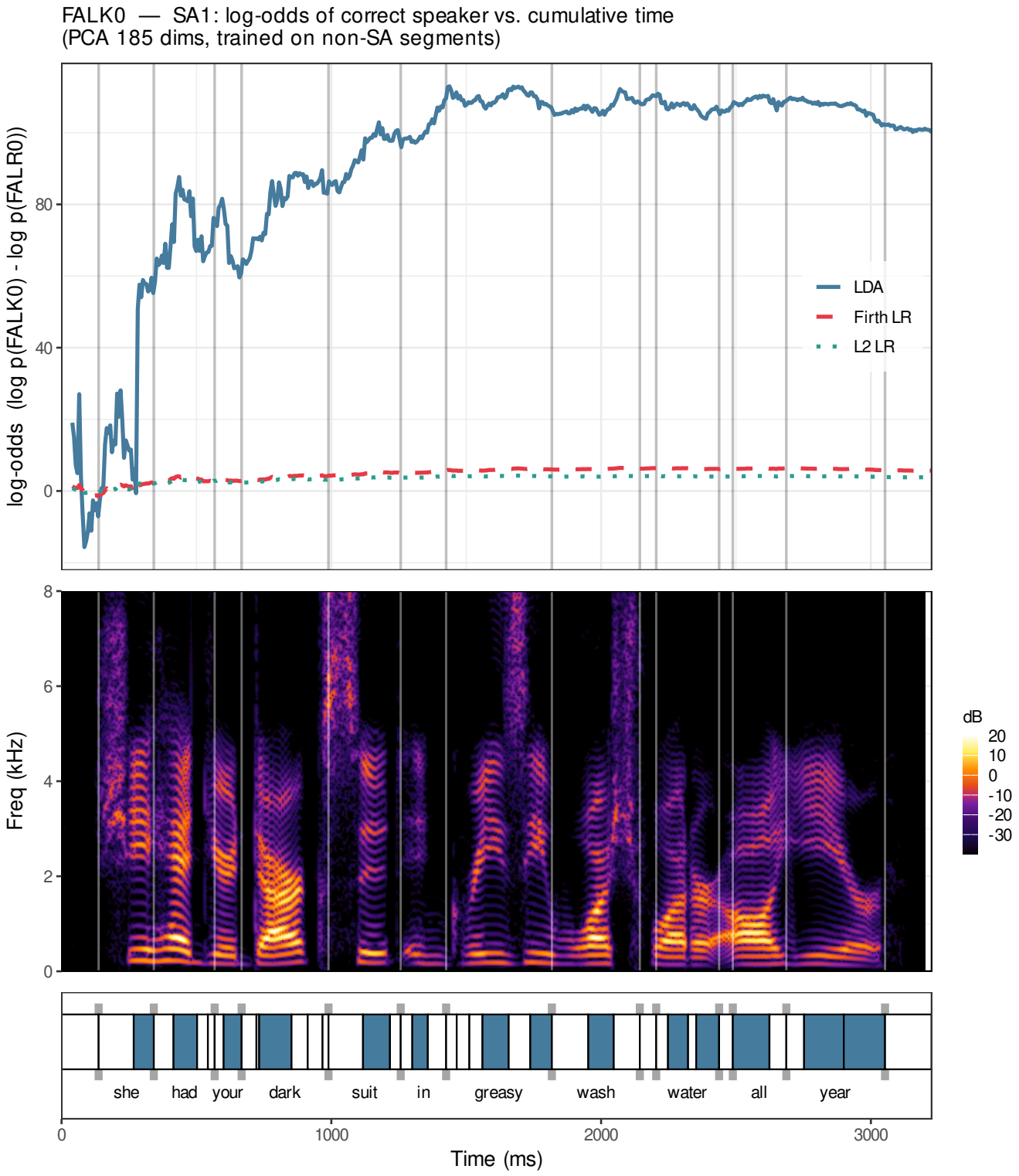

Fig. 1: Log-odds of the correct speaker (FALK0 vs. FALR0) as a function of cumulative listening time for sentence SA1, for three classifiers: LDA, Firth logistic regression, and L2-penalized logistic regression (ridge, penalty selected by 10-fold cross-validation). All models trained on PCA-reduced (185 components) L2-normalized ECAPA-TDNN embeddings of non-SA segments (i.e., from sentences other than SA1/SA2). Vertical lines mark word boundaries; filled rectangles indicate vowels.

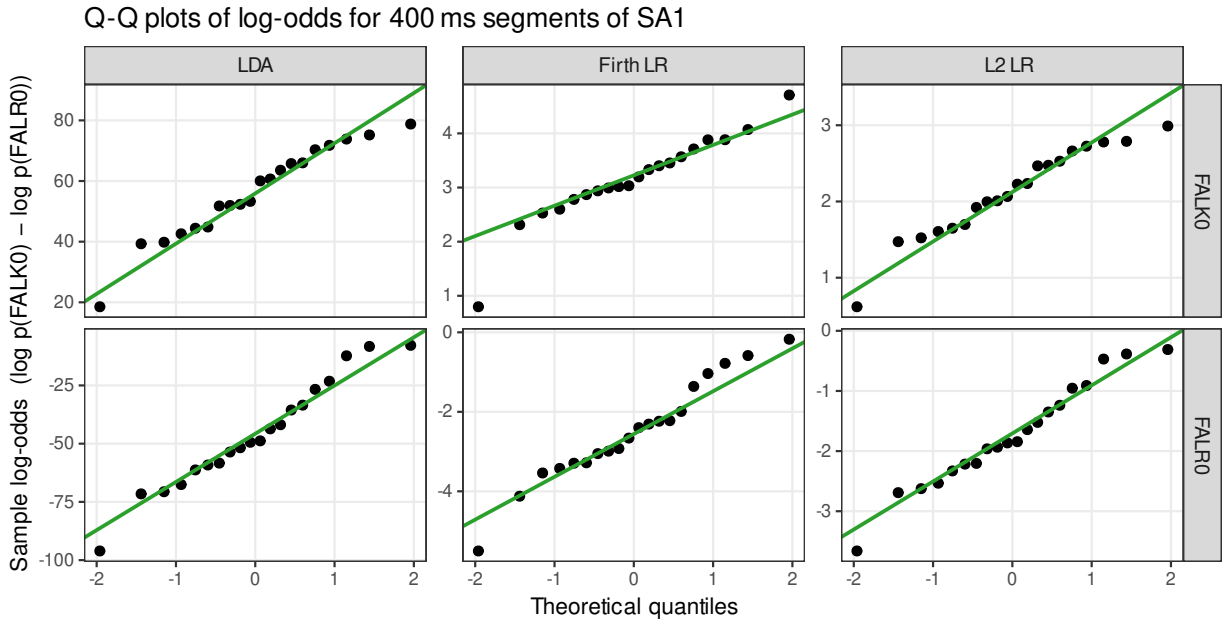

Fig. 2: Q-Q plots (against standard normal) of log-odds $\log p(\text{FALK0} \mid \mathbf{x}) - \log p(\text{FALR0} \mid \mathbf{x})$ for 20 random 400 ms segments of sentence SA1. Rows: speakers. Columns: probabilistic models. The green line is the normal reference line.

TABLE I: Shapiro–Wilk normality test for log-odds of 400 ms SA1 segments. W is the test statistic; p is the two-sided p-value ($n = 20$).

Classifier	Speaker	W	p
LDA	FALK0	0.9553	0.454
LDA	FALR0	0.9716	0.787
Firth LR	FALK0	0.9298	0.153
Firth LR	FALR0	0.9706	0.767
L2 LR	FALK0	0.9489	0.351
L2 LR	FALR0	0.9715	0.787

TABLE II: Homoscedasticity tests for log-odds of 400 ms SA1 segments, comparing variance between FALK0 and FALR0 ($n = 20$ per group). p-values reported.

Classifier	Bartlett	Levene
LDA	0.077	0.112
Firth LR	0.046	0.062
L2 LR	0.098	0.123

is known. For distance we shall use Mahalanobis distance [21]. We estimate the distance d using Fisher discriminant ratio FDR

$$d^2 = FDR = \frac{(m_1 - m_2)^2}{s_1^2 + s_2^2},$$

where m_1, m_2, and s_1, s_2 are the means and dispersions of log-odds of a probabilistic model.

We have seen in Fig. 1 that the magnitude of the predictions varies significantly among LDA, and two kinds of logistic regression. Interestingly, FDR for all three methods is approximately the same as shown in Table III. This is caused by very high correlation among the log-odds of predictions of these three models, as confirmed by the Pearson correlations in Table IV. In view of the results of Section III-A we opt to use LDA.

TABLE III: Fisher discriminant ratio for log-odds of 400 ms SA1 segments (FALK0 vs. FALR0, $n = 20$ per group).

Classifier	FDR
LDA	13.718
Firth LR	13.852
L2 LR	13.656

TABLE IV: Pearson correlations between log-odds of the three classifiers across all 40 SA1 test segments (both speakers).

	LDA	Firth LR	L2 LR
LDA	1.0000	0.9868	0.9998
Firth LR		1.0000	0.9875
L2 LR			1.0000

The resulting multidimensional scaling plot is shown in Fig. 3. We can see that the first axis generally splits the speakers by sex. The differences in acoustic properties of male vs. female voices have been thoroughly studied. For instance, females have shorter vocal tracts and higher fundamental frequency.

A. Analysis of first MDS coordinate

Let us try to explain what lies behind the variation observed in the first MDS coordinate. For that purpose we use phonetic traits of the /iy/ cardinal vowel, which we have shown to significantly affect the embeddings of the speaker identification system. Also, the corpus contains more /iy/ measurements than for the other vowels (e.g. /aa/, /uw/). To quantify how well /iy/ phonetic traits explain speaker discriminability, we regressed the first MDS coordinate of VTR speakers on four predictors:

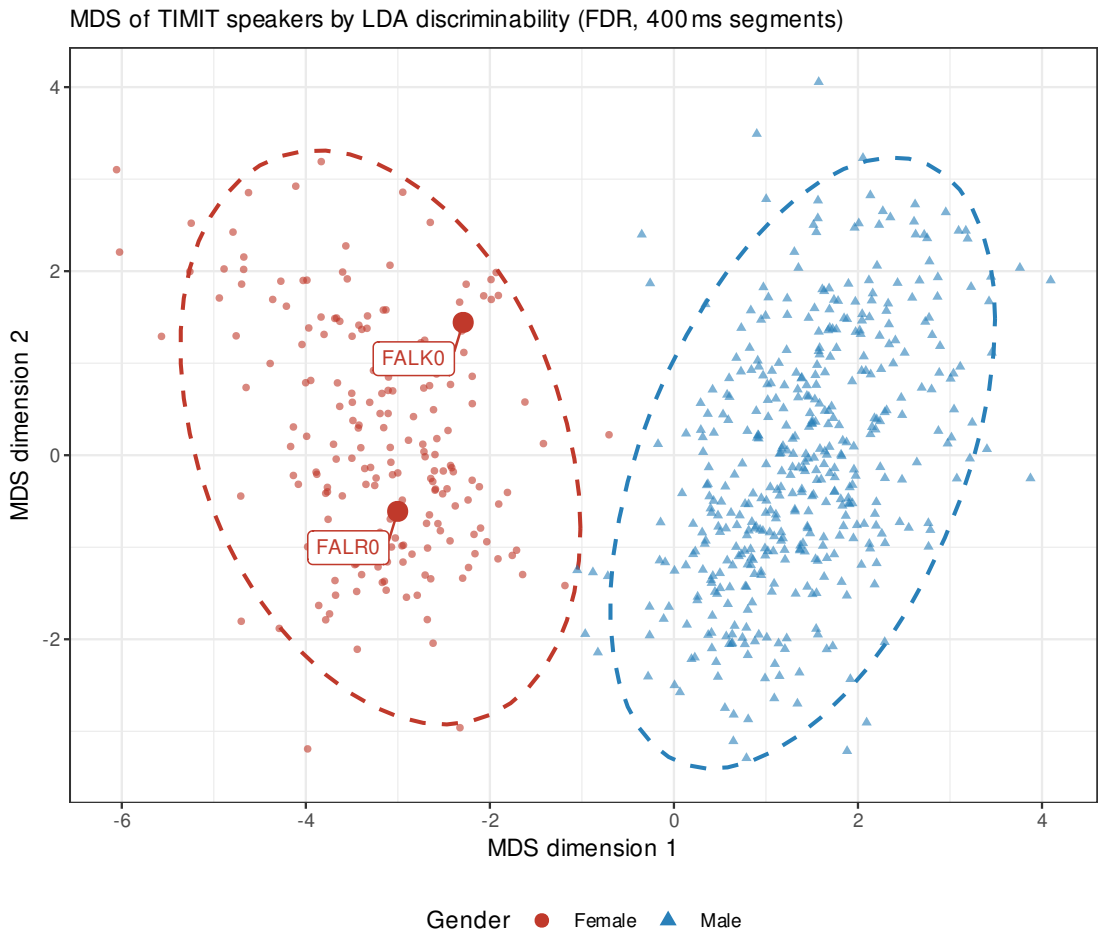

Fig. 3: Multidimensional scaling (MDS) of 630 speakers using Fisher discriminant ratio (LDA, 400 ms segments) as squared distances. Female speakers are shown in red, male in blue. FALK0 and FALR0 (the speakers analyzed in this paper) are labeled. Dashed ellipses show the 95% confidence region for each sex, assuming a bivariate normal distribution of the MDS coordinates within that group.

- the mean formant frequencies F_1, F_2, F_3 of /iy/ (in kHz) per VTR corpus annotations, and
- duration (ms) of the vowel in *she* in sentence SA1, from TIMIT phone annotations. The most commonly annotated vowel is /iy/, present in 554 of the 630 speakers.

For model selection we used Bayes information criterion (BIC) [22]. The results over all $2^4 = 16$ predictor subsets are shown in Table V. The best model is $F_1 + F_2$ ($R^2 = 0.681$); adding duration of *she* or F_3 each cost $\Delta\text{BIC} = 4.4$, providing no evidence of improvement.

We can see that the linear model using the first two formants explains 68% of variance of the first MDS dimension.

V. Discussion

Let us review the key findings we obtained:

- In our pairwise classification example, LDA gave significantly different confidence from logistic regression. Therefore one should pay close attention to a binary model's calibration.
- We have observed no significant departure from normality in LDA projection.
- We introduced MDS for visualization of speaker population.
- We linked formant values with the first MDS coordinate, showing that ECAPA-TDNN learns representation that can be interpreted phonetically.
- We showed that duration plays no role in the main MDS coordinate. This is consistent with the absence of LSTM units in the network architecture [23].

Future research may focus on interpretation of additional MDS coordinates. One may also investigate the dependence of our results on the experimental setup (the choice of the network, corpus or segment preparation).

Our work may have applications to determining effective

TABLE V: BIC model selection: predicting the first MDS coordinate for 177 VTR speakers from mean /iy/ formants (F_1, F_2, F_3 in kHz) and duration of the vowel in *she* in SA1 (dur, ms). Models ranked by BIC; ΔBIC relative to best model.

Predictors	R^2	adj. R^2	BIC	ΔBIC
F_1, F_2	0.681	0.677	639.6	0.0
F_1, F_2, dur	0.682	0.677	644.0	4.4
F_1, F_2, F_3	0.682	0.677	644.0	4.4
F_1, F_2, F_3, dur	0.684	0.676	648.4	8.8
F_1, F_3	0.603	0.599	678.1	38.6
F_1, F_3, dur	0.604	0.597	683.1	43.5
F_2	0.509	0.507	710.5	70.9
F_2, dur	0.518	0.513	712.5	72.9
F_2, F_3	0.513	0.508	714.3	74.7
F_2, F_3, dur	0.521	0.513	716.4	76.8
F_3	0.443	0.440	732.9	93.3
F_3, dur	0.449	0.442	736.3	96.7
F_1	0.412	0.409	742.5	102.9
F_1, dur	0.412	0.406	747.6	108.0
(none)	0.000	0.000	831.3	191.8
dur	0.003	-0.003	836.1	196.5

TABLE VI: Coefficients and 95% confidence intervals for the two best BIC models predicting the first MDS coordinate from /iy/ formants ($n = 177$ VTR speakers). Significance: $^*p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

	$F_1 + F_2$	$F_1 + F_2 + F_3$
Predictor	Estimate [95% CI]	Estimate [95% CI]
Intercept	18.2 [16.3, 20.1]***	17.6 [15.2, 20]***
F_1 (kHz)	−17.2 [−20.7, −13.7]***	−17.6 [−21.2, −13.9]***
F_2 (kHz)	−5.73 [−6.67, −4.8]***	−6.47 [−8.42, −4.53]***
F_3 (kHz)	—	0.854 [−1.11, 2.82]

vocal tract length [24]. A recent perceptual study [25] showed that listeners make a joint judgement about vocal tract length and vowel identity. If the first MDS coordinate is highly correlated with the effective vocal tract length, as hinted by the strong correlation with formant frequencies of the /iy/ sound, it could provide a fully automated way to estimate the length. At present, in phonetics studies one determines the effective vocal tract length through onerous manual formant determination [26]–[30].

ACKNOWLEDGMENT

We thank Matej Paprnák for inspiring discussions that led to this work. Claude Code (https://claude.com/claude-code), was used to implement the experiments, and statistical analyses.

REFERENCES

[1] D. Yu and L. Deng, *Automatic speech recognition*. Springer, 2016, vol. 1.

[2] D. A. Reynolds, "Speaker identification and verification using gaussian mixture speaker models," *Speech communication*, vol. 17, no. 1-2, pp. 91–108, 1995.

[3] L. Deng, G. Hinton, and B. Kingsbury, "New types of deep neural network learning for speech recognition and related applications: An overview," in *2013 IEEE international conference on acoustics, speech and signal processing*. IEEE, 2013, pp. 8599–8603.

[4] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. J. Lang, "Phoneme recognition using time-delay neural networks," in *Backpropagation*. Psychology Press, 2013, pp. 35–61.

[5] L. Dong, S. Xu, and B. Xu, "Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 5884–5888.

[6] L. Hermann, "Phonophotographische Untersuchungen," *Pflügers Archiv European Journal of Physiology*, vol. 58, no. 5, pp. 264–279, 1894.

[7] J. W. Cooley and J. W. Tukey, "An algorithm for the machine calculation of complex fourier series," *Mathematics of computation*, vol. 19, no. 90, pp. 297–301, 1965.

[8] G. Fant, "T. Chiba and M. Kajiyama, pioneers in speech acoustics," *Journal of the Phonetic Society of Japan*, vol. 5, no. 2, pp. 4–5, 2001.

[9] Y. Liu, L. He, J. Liu, and M. T. Johnson, "Introducing phonetic information to speaker embedding for speaker verification," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2019, no. 1, p. 19, 2019.

[10] I. Ben-Amor, J.-F. Bonastre, B. O'Brien, and P.-M. Bousquet, "Describing the phonetics in the underlying speech attributes for deep and interpretable speaker recognition," in *Interspeech 2023*, 2023.

[11] Y. Zhang, P. Tiňo, A. Leonardis, and K. Tang, "A survey on neural network interpretability," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 5, no. 5, pp. 726–742, 2021.

[12] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification," *arXiv preprint arXiv:2005.07143*, 2020.

[13] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," *arXiv preprint arXiv:1806.05622*, 2018.

[14] D. Snyder, G. Chen, and D. Povey, "MUSAN: A music, speech, and noise corpus," *arXiv preprint arXiv:1510.08484*, 2015.

[15] L. Deng, X. Cui, R. Pruvenok, J. Huang, S. Momen, Y. Chen, and A. Alwan, "A database of vocal tract resonance trajectories for research in speech processing," in *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, vol. 1, 2006, pp. I–I.

[16] S. Ioffe, "Probabilistic linear discriminant analysis," in *European Conference on Computer Vision*. Springer, 2006, pp. 531–542.

[17] D. Firth, "Bias reduction of maximum likelihood estimates," *Biometrika*, pp. 27–38, 1993.

[18] G. James, D. Witten, T. Hastie, R. Tibshirani *et al.*, *An introduction to statistical learning: with applications in R*. Springer, 2013, vol. 103.

[19] D. Jones, *An English pronouncing dictionary*. Psychology Press, 2003, vol. 3.

[20] J. C. Gower, "Some distance properties of latent root and vector methods used in multivariate analysis," *Biometrika*, vol. 53, no. 3-4, pp. 325–338, 1966.

[21] P. C. Mahalanobis, "On the generalized distance in statistics," *Sankhyā: The Indian Journal of Statistics, Series A (2008-)*, vol. 80, pp. S1–S7, 2018.

[22] A. E. Raftery, "Choosing models for cross-classifications." *American sociological review*, vol. 51, no. 1, 1986.

[23] S. Hochreiter and J. Schmidhuber, "LSTM can solve hard long time lag problems," *Advances in neural information processing systems*, vol. 9, 1996.

[24] S. Barreda and N. Silbert, "Writing up experiments: An investigation of the perception of apparent speaker characteristics from speech acoustics," in *Bayesian Multilevel Models for Repeated Measures Data*. Routledge, 2023, pp. 427–456.

[25] A. Persson, S. Barreda, and T. F. Jaeger, "Comparing accounts of formant normalization against us english listeners' vowel perception," *The Journal of the Acoustical Society of America*, vol. 157, no. 2, pp. 1458–1482, 2025.

[26] S. Barreda, "Vowel normalization as perceptual constancy," *Language*, vol. 96, no. 2, pp. 224–254, 2020.

[27] ——, "Vowel normalization and the perception of speaker changes: An exploration of the contextual tuning hypothesis," *The Journal of the Acoustical Society of America*, vol. 132, no. 5, pp. 3453–3464, 2012.

[28] A. Anikin, S. Barreda, and D. Reby, "A practical guide to calculating vocal tract length and scale-invariant formant patterns," *Behavior Research Methods*, vol. 56, no. 6, pp. 5588–5604, 2024.

[29] S. Barreda, "Normalization, essentialization, and the erasure of social and linguistic variation," *Journal of Phonetics*, vol. 110, p. 101409, 2025.

[30] S. Barreda and T. M. Nearey, "A regression approach to vowel normalization for missing and unbalanced data," *The Journal of the Acoustical Society of America*, vol. 144, no. 1, pp. 500–520, 2018.