Describing the phonetics in the underlying speech attributes for deep and interpretable speaker recognition

Imen Ben-Amor, Jean-François Bonastre, Benjamin O'Brien, Pierre-Michel Bousquet

Describing the phonetics in the underlying speech attributes for deep and interpretable speaker recognition

Imen Ben-Amor, Jean-François Bonastre, Benjamin O'Brien, Pierre-Michel Bousquet

Laboratoire Informatique d'Avignon, EA 4128, Avignon Université,France

name.last-name@univ-avignon.fr

Abstract

Deep neural networks have dominated speaker recognition, with a sharp increase in performance associated with increasingly complex models. This comes at the cost of transparency, which poses serious problems for informed decision making. In response, an intrinsically interpretable scoring approach, BA-LR, was recently presented. This method uses an attribute-based bottom-up representation of speech, linked with a transparent scoring scheme. For the sake of explainability, the present work adds an analysis of the nature of the attributes, by selecting and quantifying the contributions of the phonetic variables that describe it. We propose two methods based on statistical and surrogate models, respectively. The results reveal that the speech attributes are each well described by a set of descriptive variables. This allows us to propose the first transparent scoring scheme in speaker recognition, where the weights of the phonetic variables contributing to each decision item are known.

Index Terms: Speaker recognition, Explainability, speech attributes interpretability, phonetic features

1. Introduction

Automatic speaker recognition (SR) systems typically rely on deep neural networks, in particular ResNet [1] and ECAPA-TDNN [2]. Using models trained on large datasets, these systems efficiently extract speaker embeddings from speech. These embeddings are then compared using similarity metrics to evaluate whether two recordings belong to the same or different speakers. Despite their success in terms of reported error rates, these automatic SR systems do not provide explainable information, as to what is captured in the speaker representation, how it is encoded, and how it is used during the decision making process. This lack of transparency and explainability becomes problematic when considering the judicial domain, as laws evolve, and even more critical in applications such as forensics.

Addressing this issue, recent works in the speech domain used probing classifiers [3] to reveal speaker-related information encoded in the learned representation, such as accent [4], style [5], dialectal/non-dialectal [6], identity, channel, and transcription [7, 8]. Other works analyzed the presence of phonemic information along neural network layers [9, 10, 11, 12]. Such work typically requires finely labeled data, which is a critical and expensive resource that is rarely available. In a different avenue, [13] studied the presence of acoustic features in model layers in order to identify how the network encodes this information. Often inspired by computer vision, explainable AI (XAI) techniques have also been used in speech processing, such as gradient-based techniques [14, 15, 16, 17] and Shapley values [18, 19]. However, these works are unable to propose informed decision schemes, such as which information is captured by SR systems and what is its contribution during decision making process.

The Binary-attribute-based likelihood ratio (BA-LR) estimation is a solution initially proposed for forensic application [20]. It breaks down the SR scoring process into independent sub-processes, where each is dedicated to a specific speech attribute. For a given speech extract, an attribute may be present or absent, which, in turn, is given by a binary value (either 1 or 0). The scoring sub-process outputs a Likelihood Ratio (LR) value calculated for each attribute and for each of the four possible cases (i.e., 00, 01, 10 and 11). Although the BA-LR scoring approach may significantly improve interpretability, it is noteworthy that the attribute extractor utilizes a bottom-up approach and does not provide any additional information on the nature of speech attributes.

The current work aims to fill this gap by exploring the nature of these speech attributes. The prerequisites we set for this work are to not require additional manual labeling, to be able to handle a large amount of data, and to produce a meaningful description for a speech expert. The attributes are intended to be speaker discriminant (i.e., shared only by a subset of speakers). The main principle of the present work consists in comparing two sets of speech extracts: one set composed of speech excerpts where the attribute is detected, i.e., spoken by the group of speakers who share it, and the other set composed of excerpts spoken by other speakers (and where, by definition, the attribute is never present). Through this comparison, our method identifies the phonetic descriptive variables, such as F0, formants, jitter, shimmer, etc, that explain the difference between the two sets, weights them, and provides an in-depth view of the most salient information that characterizes an attribute. Two approaches are presented and evaluated, one employs decision trees as surrogate explainable models and the second relies on a classical statistical test based on a step-wise linear discriminant analysis (SLDA) [21].

The outline of the paper is as follows: Section 2 provides an overview of the BA-LR approach, followed by the proposed interpretability methodology of attributes in Section 3. Sections 4 and 5 present the experimental setup and the results, respectively. Conclusion and future work are summarized in Section 6.

2. BA-LR approach overview

This section presents a description of our BA-LR approach introduced in [20]. We decompose it here into three main steps as shown in Figure 1:

1. A speech extract X is represented by a binary vector, denoted

as "BA-vector", where each coefficient gives the presence (1) or absence (0) of a given speech attribute. The BA-vectors are extracted using a lightly modified X-vector extractor [22], BA-extractor, which is optimized to produce binary representation.

2. During a test scoring of a given pair (X,Y), a partial Likelihood Ratio (LR) is computed separately for each attribute BA_i, using both its behavioral parameters (i.e., T_i, $Dout_i$, Din_i) and its value in X and Y (11, 00, 01, 10). T_i represents the typicality of the attribute, or how frequently it occurs among speakers (i.e., its discriminative power). $Dout_i$ is the probability that an attribute is absent from a speech extract while present in other extracts of the same speaker, while Din_i is the probability of falsely detecting an attribute in an extract, due to noise, for example. Din_i is computed as a combination of a fixed factor Din, and T_i. T_i, $Dout_i$, and Din are estimated during the training phase (Figure 1), which are then used to compute the corresponding partial LR values per attribute BA_i and each test case (11, 00, 01, 10). The final LR is therefore computed as the product of partial LRs, under the assumption that attributes are independent.
3. The training of the BA extractor is driven only by the speaker label associated with the training data. Thus, the extraction of the speech attributes can be considered as a bottom-up process and no meta-information is available on their nature. [20] suggested that providing an acoustic and phonetic description of the attributes would be important for enhancing explainability. However, this step has not been previously done. This third step represents the main focus of the present work, which studies the information captured by each attribute and its connection to speech characteristics.

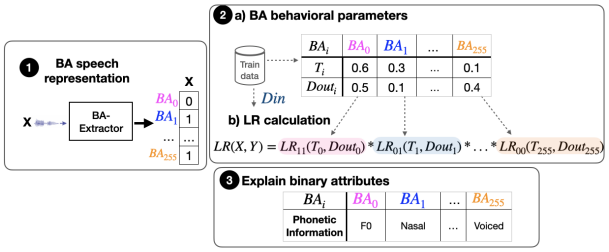

Figure 1: *The overall BA-LR approach methodology following the three steps.*

3. Describing the phonetic nature of the speech attributes

This section presents the core of the current work: a methodology for discovering the nature of the speech attributes of the BA-LR approach. It is important to recall that these attributes are derived from a bottom-up process and that no information is available about their nature. We build our methodology under the assumption that if we can identify phonetic variables that effectively differentiate speech extracts as having (or not) a particular attribute, then these variables are likely to be good descriptors of the attribute.

3.1. Description of the methodology

After extracting BA-vectors for each available speech extract, we apply the following three-steps strategy independently for each attribute, as illustrated in Figure 2:

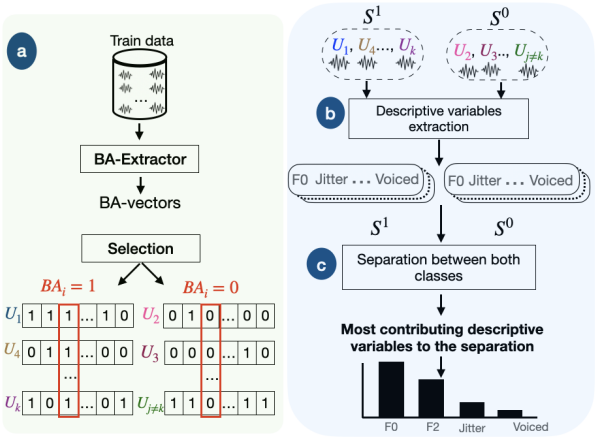

Figure 2: *Methodology of an attribute interpretability following (a), (b), and (c) steps, as applied to each speech attribute*

- The extracts are grouped into two sets (Fig. 2-a). The first set, denoted "S_1", groups the extracts where the considered attribute is present. The second set, denoted "S_0", contains the extracts from speakers other than those present in S_1 and where the attribute has a value of 0. In other words, S_1 contains positive examples of the attribute pronounced by the set of speakers who share this attribute, while S_0 presents negative examples pronounced by other speakers, who never had the attribute. Finally, some randomly selected extracts are eliminated from S_0 to balance the number of extracts in the two sets in order to avoid bias during the selection process.
- The second step, in Fig. 2-b, is dedicated to describing the speech extracts. For this, we choose a set of M descriptive variables, which can be of any type, as long as they can be computed automatically from speech. For this specific work, we opted to use descriptive phonetic variables, such as F0, formants, jitter, shimmmer, etc., but other variable types are also possible, such as phonemic [23, 24, 25] or language-related ones [26]. The values of the variables are then extracted for each speech extract.
- The last step, in Fig. 2-c, is to select the relevant descriptive variables for the attribute in question and weight them. As explained earlier, the strategy relies on selecting the variables that best explain the difference between sets S_1 and S_0. For this purpose, we propose two solutions. The first, issued from the field of XAI, consists in training an intrinsically explainable classifier, called a surrogate model, to separate examples of sets S_1 and S_0. It uses the descriptive variables extracted from each training example as features. The most influential variables are the ones that best describe the concerned speech attribute. To ensure the relevance of this choice, we use a second method based on a classical statistical test, which is less powerful, but provides a simpler solution. Both approaches are detailed later in this section.

3.2. Surrogate model

Good candidates for surrogate models are those that are inherently explainable, capable of giving sufficient separability performance between class 1 and class 0 examples, and easy to train (i.e., stable and not too costly in terms of computing power). We choose a decision tree as surrogate model, since it is known to be fast, simple, and explainable by nature. Specifi-

cally, we use TreeExplainer from SHAP toolkit[1], which computes Shapley values [27]. Shapley values are used to estimate feature contributions to the classifier predictions following Equation 1, where X_j is a descriptive variable, BA_i is the attribute described, and ShapMean(X_j) is the average of Shapley values obtained for X_j across all instances using the BA_i surrogate model.

$$\text{Contribution}_{BA_i}(X_j) = \frac{ShapMean(X_j)}{\sum_{k=1}^{M}(ShapMean(X_k))} \quad (1)$$

3.3. Statistical test

We select step-wise linear discriminant analysis (SLDA) [21] as a statistical method for identifying a linear combination of the explanatory variables that characterize or separate the examples of the two classes 1 and 0. The algorithm begins with an empty list of explanatory variables and then adds variables to it with the highest discriminant power based on Wilks' Lambda value (Equation 2) until the p-value achieves a threshold of 0.01.

$$\text{Wilk's lambda} = \frac{det(A)}{det(A+B)} \quad (2)$$

where det is the determinant, A is the within class covariance matrix, and B is the between class covariance matrix.

4. Experimental protocol

In this section, we describe the datasets and protocols used for the experimental validation[2] of our proposals.

4.1. Data set and protocol

We use four corpora for the experiments: VoxCeleb1&2 [1], SITW [28] and VOiCES [29], as summarized in Table 1. During steps 1 and 2 of our BA-LR approach (as shown in Figure 1), VoxCeleb2 (Vox2) is used for training the BA-extractor and computing behavioral parameters related to attributes (T_i, $Dout_i$, Din). It is also used to discover the nature of attributes (step 3 in Figure 1).

VoxCeleb1 (Vox1), SITW, and VOiCES are used for testing only, and have no intersection with Vox2 in terms of speakers. For testing, we use the evaluation protocol of SITW and VOiCES, whereas for Vox1 we select the first ten utterances of each of the 1251 speakers, resulting in 56,295 target pairs (45 per speaker). We balance the number of non-target/target trials for Vox1, SITW, and VOiCES, by randomly selecting a subset of non-target trials.

Table 1: *Description of data sets*

	Train	**Test**		
	Vox2	**Vox1**	**SITW**	**VOiCES**
# of speakers	5,994	1,251	180	100
# of utterances	1,021175	153,516	2883	11,392
# of trials		112,558	7,316	72,886

4.2. Baseline Vs. BA-LR description

To provide a point of comparison, we utilize a ResNet-based baseline [1, 30] with 256-dimensional embeddings as the seed for the BA-LR extractor. The BA-LR extractor adds a thresholding layer dedicated to producing sparse representations. After binarization, 256-dimensional vectors are obtained, which are then pruned to eliminate zero-activity coefficients. This results in a set of 205 BAs in each BA-vector.

4.3. Descriptive variables extraction

We use the open-source audio feature extraction toolkit OpenSmile[3] in order to compute the descriptive variables. We use the eGeMAPS [31], pre-defined set of 88 descriptive variables, which contains 18 low-level descriptors (LLD)[4], 7 LLDs of cepstral and dynamic parameters, their corresponding functionals[5], and several spectral parameters.

5. Experimental results & Discussions

In this section, we first evaluate the overall performance of our implementation of BA-LR approach in terms of speaker discrimination and generalization capabilities. Then, we address the core of this work: the description of the nature of the speech attributes automatically discovered by the proposed methodology in terms of descriptive phonetic variables.

Table 2: *SR performance in terms of EER and Cllr$_{min/act}$.*

Dataset	**X-vectors**		**BA-vectors**	
	EER	**Cllr$_{min/act}$**	**EER**	**Cllr$_{min/act}$**
Vox1	1.37	0.05/0.82	3.7	0.14/0.31
SITW	1.4	0.05/0.82	3.5	0.13/0.28
VOiCES	3.96	0.15/0.87	4.7	0.18/0.46

5.1. Speaker recognition performance

Table 2 shows the equal error rate (EER) and Cllr$_{min/act}$ [32] obtained using our BA-LR system and a X-vector baseline, for the three different test sets. BA-LR achieves a comparable EER for all corpora, with a loss of 1.72% (absolute) in average compared to the baseline[6]. This seems to be relatively moderate with respect to the well-known trade-off between performance and explainability, which, when required, is often observed. The BA-LR approach shows good generalization capability, as indicated by its performance on all test sets, especially for VOiCES, which differs significantly from the training set in terms of both speech content and recording conditions.

5.2. Attribute automatic description

We first use the surrogate model (Section 3.2) solution in order to build automatic descriptions of the different attributes in terms of descriptive variables (Section 4.3). The process is performed independently for each attribute. Figure 4 shows the accuracy obtained for all BAs surrogate models on both Vox2

[1]https://github.com/slundberg/shap

[2]https://github.com/LIAvignon/BA-LR

[3]https://github.com/audeering/opensmile

[4]These include frequency related parameters, such as pitch, jitter, formants, energy related parameters shimmer, loudness, and harmonics-to-noise ratio.

[5]The 20^{th}, 50^{th}, and 80^{th} percentile, the range between the 20^{th} and 80^{th} percentiles, and the mean and standard deviation of the slope of rising/falling signal.

[6]The BA-LR binary representation is ≈40 times smaller than the x-vector one, which opens up an avenue for performance improvement

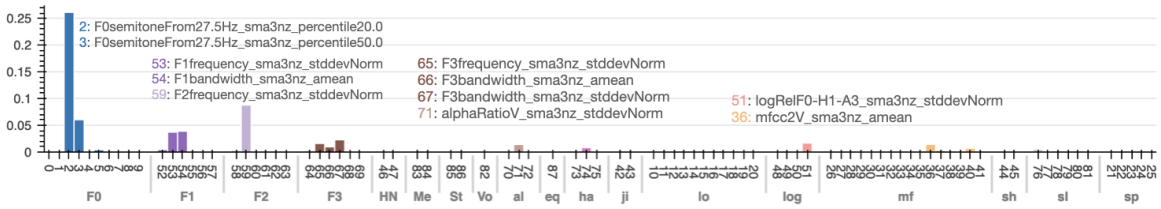

Figure 3: *Descriptive variable contributions to the BA_9 surrogate model, grouped by families.*

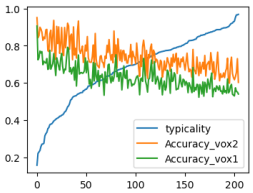

Figure 4: *Accuracy of BAs surrogate models on Vox2 (train) and Vox1 (test) sets with their associated typicality values.*

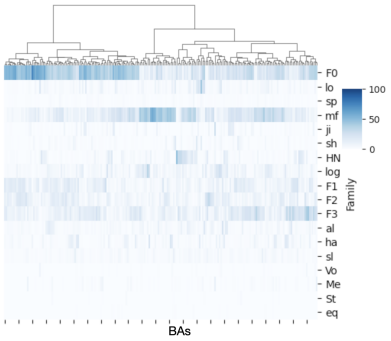

Figure 5: *Contribution of family variables to BAs (%).*

(train set) and Vox1 (test set), ranked by the typicality of the BAs (from lowest to highest).

The relatively high accuracy values obtained on train set (between 0.6 and 0.97) and the small difference in accuracy between train and test sets (0.11 on average) indicate a good selection and fit of the surrogate models. A strong inverse correlation is also observed between typicality and accuracy, indicating that the attributes that carry the greatest power to discriminate between speakers are also the best represented by the descriptive variables.

Figure 5 illustrates the contributions of different families[7] of variables to each attribute. The attributes are ordered by hierarchical clustering. The differences along the attributes (i.e, BAs)(x-axis) reinforce the hypothesis that the attributes do not encode the same phonetic information. The figure also reveals that some variable families are on average more important than others, even if some exceptions are present for some attributes.

We then apply the SLDA (Section 3.3) method and measure the similarity between the variables selected by both methods.

[7]The families are: F0 (pitch); F1, F2, F3, (formants); HN (Harmonics-to-noise ratio); Me and St (mean and std of Voice/unvoiced); Vo (VoicedSegmentperSecond); al (alpha ratio); eq (equivalent-sound-level); Ha (Hammarberg index); ji (jitter); lo (loudness); log (logRelF0-H1-A3); mf (MFCCs 1-4); sh (shimmer); sl (slope); sp (spectral-flux).

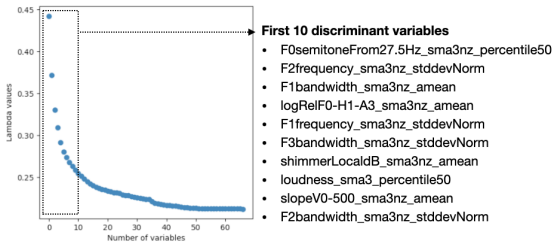

Figure 6: *Lambda values as a function of the number of selected variables for attribute BA_9 (and the first 10 variables)*

For that, we use the surrogate models to select the variables representing 75% of the contributions using Shapley values and evaluate the intersection with the variables selected by SLDA method. Achieving ≈80% of convergence between both methods strengthens confidence in the attribute description that we obtained. This also shows the value of the surrogate model approach, which allows for more accurate quantification of information via Shapley values. This convergence is illustrated in Figures 3 and 6 for a given BA.

6. Conclusion

This study has built on a newly introduced approach for speaker recognition, BA-LR, which highlights the presence or absence of specific speech attributes in speech extracts via the use of a transparent process to decompose the scoring process by attribute. While the BA-LR framework provides an interesting approach to interpreting speaker recognition, its bottom-up process for discovering speech attributes leaves the nature of these attributes unknown, resulting in a lack of explicability. This work has addressed this gap by describing the nature of the attributes contributing to the LR value. To achieve this goal, we have described speech excerpts and attributes by a set of phonetic variables that are easy to extract automatically. We have also proposed the use of decision trees as explainable surrogate models, trained to separate the set of speech excerpts having a given attribute from those that do not. From the surrogate models, we were able to select and weight the most important descriptive variables for each attribute using Shapley values, which differ across attributes. We have evaluated the robustness of our approach by comparing it with a traditional SLDA, and we have achieved clear convergence in terms of variables selection. Overall, our work has opened a new perspective on explainable speaker recognition, represented by the complete architecture described in this paper. This approach departs from traditional SR systems, by taking full advantage of them. However, it is important to note that the integration of human expertise is still required to interpret the combination of phonetic variables in terms of high-level speech features.

7. References

[1] A. Nagrani, J. S. Chung, W. Xie, and A. Zisserman, "Voxceleb: Large-scale speaker verification in the wild," *Computer Speech Language*, 2020.

[2] B. Desplanques, J. Thienpondt, and K. Demuynck, "Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," *Interspeech*, 2020.

[3] S. A. Chowdhury, N. Durrani, and A. Ali, "What do end-to-end speech models learn about speaker, language and channel information? a layer-wise and neuron-level analysis," in *Journal of Computer Speech and Language 2021*, 2021.

[4] A. Prasad and P. Jyothi, "How accents confound: Probing for accent information in end-to-end speech recognition systems," in *58th Annual Meeting of the Association for Computational Linguistics*, 2020.

[5] Z. Elloumi, L. Besacier, O. Galibert, and B. Lecouteux, "Analyzing learned representations of a deep asr performance prediction model," in *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 2018.

[6] S. A. Chowdhury, A. Ali, S. Shon, and J. Glass, "What does an end-to-end dialect identification model learn about non-dialectal information?" in *Proc. INTERSPEECH 2020*.

[7] D. Raj, D. Snyder, D. Povey, and S. Khudanpur, "Probing the information encoded in x-vectors," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2019.

[8] S. Wang, Y. Qian, and K. Yu, "What does the speaker embedding encode?" *interspeech*, 2017.

[9] T. Nagamine, M. L. Seltzer, and N. Mesgarani, "Exploring how deep neural networks form phonemic categories," in *Interspeech*, 2015.

[10] Y. Belinkov, A. Ali, and J. Glass, "Analyzing phonetic and graphemic representations in end-to-end automatic speech recognition," *Interspeech 2019*, 2019.

[11] Y. Belinkov and J. Glass, "Analyzing hidden representations in end-to-end automatic speech recognition systems," *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017.

[12] S. Shon, H. Tang, and J. R. Glass, "Frame-level speaker embeddings for text-independent speaker recognition and analysis of end-to-end model," *2018 IEEE Spoken Language Technology Workshop (SLT)*, 2018.

[13] G. Beguš and A. Zhou, "Interpreting intermediate convolutional layers of generative cnns trained on waveforms," in *IEEE/ACM transactions on audio, speech, and language processing*, 2022.

[14] A. Krug, R. Knaebel, and S. Stober, "Neuron activation profiles for interpreting convolutional speech recognition models," *32nd Conference on Neural Information Processing Systems (NeurIPS 2018)*, 2018.

[15] A. Krug, M. Ebrahimzadeh, J. Alemann, J. Johannsmeier, and S. Stober, "Analyzing and visualizing deep neural networks for speech recognition with saliency-adjusted neuron activation profiles," *Electronics*, 2021.

[16] K. Markert, R. Parracone, M. Kulakov, P. Sperl, C.-Y. Kao1, and K. B¨ottinger, "Visualizing automatic speech recognition – means for a better understanding?" in *ISCA Symposium on Security and Privacy in Speech Communication*, 2022.

[17] P. Li, L. Li, A. Hamdulla, and D. Wang, "Reliable visualization for deep speaker recognition," *Interspeech*, 2022.

[18] A. R. Syed and M. I. Mandel, "Data valuation for acoustic models in automatic speech recognition," in *NeurIPS 2020 Workshop on Dataset Curation and Security*, 2020.

[19] S. Sivasankaran, E. Vincent, and D. Fohr, "Explaining deep learning models for speech enhancement," in *Interspeech*, 2021.

[20] I. Ben-Amor and J.-F. Bonastre, "Ba-lr: Binary-attribute-based likelihood ratio estimation for forensic voice comparison," in *International Workshop on Biometrics and Forensics (IWBF)*, 2022.

[21] K. Stapor, "Better alternatives for stepwise discriminant analysis," *Folia Oeconomica*, vol. 1, no. 311, 2016.

[22] 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)*. IEEE, 2018, pp. 5329–5333.

[23] M. Ajili, J.-F. Bonastre, W. B. Kheder, S. Rossato, and J. Kahn, "Phonetic content impact on forensic voice comparison," in *IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2016.

[24] Y. Lei, N. Scheffer, L. Ferrer, and M. McLaren, "A novel scheme for speaker recognition using a phonetically-aware deep neural network," in *IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2014.

[25] S. Wang and J. Rohdin, "On the usage of phonetic information for text-independent speaker embedding extraction," in *Interspeech*, 2019.

[26] L. Lu, Y. Dong, X. Zhao, J. Liu, and H. Wang, "The effect of language factors for robust speaker recognition," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2009, pp. 4217–4220.

[27] S. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *31st Conference on Neural Information Processing Systems (NIPS 2017)*, 2017.

[28] M. McLaren, L. Ferrer, D. Castan, and A. Lawson, "The speakers in the wild (sitw) speaker recognition database," *Interspeech*, 2016.

[29] M. K. Nandwana, J. V. Hout, M. McLaren, C. Richey, A. Lawson, and M. A. Barrios, "The voices from a distance challenge 2019 evaluation plan," *Interspeech*, 2019.

[30] M. Mohammadamini, D. Matrouf, J.-F. Bonastre, S. Dowerah, R. Serizel, and D. Jouvet, "Learning noise robust resnet-based speaker embedding for speaker recognition," in *Odyssey The Speaker and Language Recognition Workshop*, 2022.

[31] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*, 2010.

[32] N. Brümmer and J. Du Preez, "Application-independent evaluation of speaker detection," *Computer Speech & Language*, vol. 20, no. 2-3, pp. 230–275, 2006.