



Using gender, phonation and age to interpret automatically discovered speech attributes for explainable speaker recognition

Carole Millot^{1,2}, Clara Ponchard¹, Cédric Gendrot², Jean-François Bonastre^{1,3}, Orane Dufour^{1,4}

¹Inria, France

²Sorbonne Nouvelle, CNRS, Laboratoire de Phonétique et Phonologie, France

³EA 4128, Laboratoire Informatique d'Avignon, France

⁴CNRS & Université de Lorraine, Loria, France

first.last@{inria,sorbonne-nouvelle,loria}.fr

Abstract

Explainability is particularly necessary for speaker recognition because of its potential forensic applications. A recent approach, BA-LR, offers a new level of explainability. It represents a speech utterance by a binary vector where each coefficient indicates the presence or absence of a given speech attribute. However, the interpretation of these attributes remains unclear, as they are found automatically. The aim of this work is to develop a methodology for interpreting these attributes in terms of comprehensible explicit voice traits such as gender, age and perceived phonation type. Our methodology is based on the assumption that if an attribute is useful for a classification task of an explicit voice trait, this means that this attribute encodes all or part of this trait. Our results show that BA-LR attributes encode, at least partially, gender, age and perceived phonation type. These results pave the way for a comprehensive interpretation of BA-LR speech attributes.

Index Terms: speaker recognition, explainability, voice quality, speaker traits, perception

1. Introduction

While automatic speech processing systems have seen their accuracy drastically improve during these past years, they are losing out in terms of interpretability, due to the complexity of the models and datasets. Automatic speaker recognition wishes to verify a speaker's identity using voice recordings and it's no exception to this complexity versus loss rule [1]. The great level of performance is balanced-out by the fact that their outputs cannot be easily explained or interpreted, for instance by breaking them down into interpretable decision factors, highlighting each factor and its contribution to the final decision. This is particularly alarming in the domain of speaker identification, since the systems may be used as part of legal or forensic activities [2], or other "High Risk" activities framed by the AI Act [3].

The BA-LR approach [4] has been proposed to overcome this limitation and enable the use of high-performance deep learning models while offering intrinsically a high level of explainability. It represents audio recordings by a binary vector where each coefficient indicates the presence or absence of a given voice attribute in them. These attributes are automatically determined bottom-up using a deep-learning approach applied on a large training corpus. A major hypothesis of the approach is the fact that the attributes encode speaker specific characteristic shared by a group, small or large, of speakers.

BA-LR offers a good explainability/efficiency (in terms of speaker detection) ratio [5] with about of 3.7% EER on Vox-Celeb database (to be compared with 1.37% for the baseline system). This result was confirmed in mismatched conditions, when the system is trained in one language and condition and is

applied to other languages and conditions [6].

The automatically discovered attributes are partially explained by a set of phonetic descriptive variables in [5]. However, their nature is still not completely understood. Particularly, it is still unclear how to link the descriptive variables used to describe the attributes (like specific spectral tilt measurements or variations of melody) with the proposed hypothesis that an attribute models a discriminating characteristic of the speaker shared by (only) a group of speakers.

In this work, we explore the nature of attributes by attempting to interpret them through three explicit voice traits: gender (representing biological sex), age and a voice quality information, the phonation type. Gender and age help us characterize speakers [7, 8]. Their impact on voice is high: significant variations have been observed between male and female voices due to anatomical differences – women tend to have a higher pitched voice due to their vocal folds being smaller than men's [9, 10]; elderly speakers show increased signs of hoarseness, fatigue and crackling [11] due to the laryngeal muscles deteriorating with time [12]. Voice quality is defined by the "means by which speakers project their identity to the world" [13], and is impacted by multiple acoustic determinants "resulting from the configuration of the vocal apparatus" [14]. According to [15, 16, 17, 18], voice quality has implications for speaker characterization. Phonation types are part of voice quality: they qualify the continuum of possible laryngeal configurations during phonation. Most known phonation types include modal voice, creaky voice (irregular vocal folds vibrations) and breathy voice (incomplete vocal folds closing during phonation) [19]. Our hypothesis is that if an attribute encodes one of the explicit voice traits we have selected, it must play a prominent role in a detection task for that information. To this end, we choose decision trees as simple, explainable classifiers. We train several classifiers using BA vectors as input, and use their performance as an indicator of the presence of any of the explicit voice traits selected in the BA-LR attributes. Before proceeding, we require high-quality labeled training and test corpora for the three parameters we aim to study. Given the limited availability of phonation type labeling resources, we need to create our own dataset.

The rest of this paper is organized as follows: section 2 describes the data-gathering method used to create our dataset. Section 3 presents the experiments and the corresponding results. Section 4 discusses these results and their implications.

2. Dataset and data preparation

This section presents the selected data as well as the labeling in terms of perceived phonation type that we carried out. Some additional details on data preparation are also provided.

2.1. Dataset description

This work requires a dataset possessing a great variety of speakers, with various ages and phonation types. Thus, we use several French-speaking corpora to obtain equitably distributed gender, age and phonation type classes: parts of PTSVOX [20], CommonVoice (CV) [21] and PFC [22] as well as data gathered by the fifth author. Table 1 shows the number of speakers and extracts per corpora. Speaking style can be spontaneous, prepared (discourse or interview) or read. Available metadata include the speakers' gender ("female" or "male") and age ("young" under 30 years old, "middle-aged" between 30 and 60 years old and "old" over 60 years old).

Silences are removed from the recording using the Silero VAD system¹ and the recordings are cut into three-second extracts, mimicking the BA-LR training process.

Table 1: Number of speakers, mean extracts per speaker and female-to-male balance per corpora

Corpus	Speakers	Mean extracts/speaker	F/M%
PTSVOX	14	9	44%
CV	33	5	22%
PFC	11	7	39%
Other	21	7	32%
Total	79	7	34%

2.2. Phonation type annotation

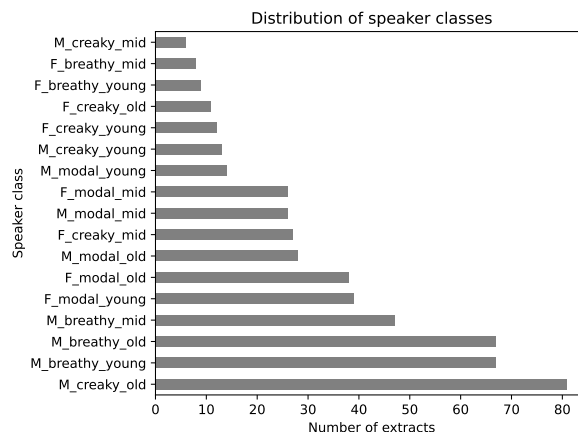
We perceptually annotated the audio extracts using three types of phonation: creaky, modal and breathy. The two steps data selection and annotation methodology is described in the following paragraphs.

We start by an initial step wishing to collect prototypical excerpts for voices for the three types of phonation. "Prototypical" audio excerpts possess a prototypical phonation type persisting in the speech extract. Three expert judges carried out the perceptual labeling of the data. First, one judge annotated all speech excerpts and assigned each speaker a phonation type perceived as dominant. Next, 15 speakers were randomly selected for each phonation type, resulting in a total of 45 speakers. For each speaker, five speech excerpts were randomly chosen, forming a dataset of 225 speech excerpts. Then, the two other judges annotated perceptually the data. The final decisions were taken by consensus between the three judges, using decision rules drafted during the discussion like: "phonation types related to pathological voices are deemed prototypical, while voices possessing both creaky and breathy voice characteristics at the same time are deemed non prototypical". This first data gathering enabled the collection of a set of 173 audio extracts prototypical of one phonation type, including 43 with creaky voice and 12 with breathy voice.

After this first phase we have a set of manually labelled prototypical examples of the three phonation types as well as annotation rules. To extend the number of creaky and breathy audio extracts, we automatically select a set of other speech excerpts close to the speech examples linked to the phonation labels we're looking for, and then ask the judges to label them using the same procedure as before. The speech utterances selection is done using a combination of three acoustic measures known to be discriminant for creaky and/or breathy voices, H1*-H2*

¹<https://github.com/snakers4/silero-vad>

Figure 1: Bar chart showing proportion of all possible voice traits combination ("M" = male, "F" = female, "mid" = middle-aged).



[23], H2KH5K [24] and Cepstral Peak Prominence (CPP) [25]. The measures are extracted using PraatSauce², a porting of the VoiceSauce software [26] on Praat [27].

Using the measurements collected from the prototypical examples in the first step, we define rules to select the second round candidates using a simple strategy: the rules should work for all the prototypical examples. At the end, all the speech extracts with a negative H1*-H2* are selected as creaky candidates and the speech extracts with both a CPP measure < 12 and a H2KH5K measure < 10 are selected as breathy candidates. It gives a total of 512 new candidates which are sent to the three judges to check the perceived phonation type. Out of these extracts, 200 are unanimously classified as breathy or creaky, leading to a total of 538 extracts including the initial ones. The final distribution of extracts is shown in Table 2. The combination of the three voice traits is called the "combined" condition; its distribution can be seen Figure 1. Notably, the dataset lacks recordings of older female speakers with a breathy voice.

Table 2: Number of extracts per corpora depending on age and phonation type (Cr = creaky voice, Mo = modal, Br = breathy)

Corpus	-30 _{yo}	30-60 _{yo}	+60 _{yo}	Cr	Mo	Br
PTSVOX	123	0	0	15	53	55
CV	31	90	43	29	55	80
PFC	0	14	65	4	33	44
Other	0	36	117	102	30	21
Total	154	130	225	150	171	200

2.3. BA vector extraction

As mentioned in the introductory section, we use the binary attribute (BA) vector representation of the speech extracts. It represents a speech utterance by a BA vector where a coefficient, $BA(i)$, indicates whether the speech attribute i is present or not in the extract. We use the BA extractor presented in [4], inspired by the Resnet extractor presented in [28]. It is trained

²<https://github.com/kirbyj/pratsauce>

using VoxCeleb2 [29], an English-speaking database of more than one million recordings produced by more than 6000 speakers (training data is given to the system in the form of 2 – 3 seconds audio extracts from the recordings). A BA vector of dimension 206 is extracted for each speech extract. No adaptation or fine-tuning of the BA extractor was done.

3. Experiments

This section presents the different experiments proposed in this article and the corresponding results.

Table 3: *Training and testing datasets*

Set	Nb. speakers	Nb. extracts
Train	63	433
Test	16	86

3.1. Gender, age and phonation type classification

As indicated in the introduction, we use decision trees as surrogate models to study the presence contained in the BA-LR attributes of the three explicit voice traits, namely gender, age and phonation type. We use `DecisionTreeClassifier` from the Python library Scikit-learn [30], and the “gini” criterion to find the best attribute to use at each node. The data distribution for train and test sets is described in Table 3. The sets are obtained using Scikit-Learn’s `GroupShuffleSplit` so as to prevent the same speaker from appearing in both train and test sets. The splitting separates data in 80% for train set and 20% for test set.

We first train a decision tree classifier for each of the three voice traits (see section 2.1 for details): gender (two classes), age (three classes) and phonation type (three classes), separately. Then, we train a model combining all the informations (17 classes), denoted “combined” (see Figure 1). Maximum depth is specified for each model in Table 4.

Table 4: *Performance of decision tree models on the test set, in terms of accuracy, precision and recall (using macro averaging). The model depth on the train set is also provided*

Task	Depth	Acc.	Precision	Recall
Gender	8	0.86	0.86	0.87
Age	11	0.45	0.47	0.49
Phonation type	10	0.49	0.55	0.53
Combined	14	0.13	0.12	0.17

3.1.1. Results for age, gender and phonation type

For gender classification, a depth of 8 is sufficient to obtain a 100% accuracy on the train data. Testing phase reveals an accuracy of 86% (see Table 4). For age classification, a depth of 11 is required to attain 100% accuracy on the train data and the model obtains an accuracy of 45% on the test set. Extracting the tree’s wrong predictions and confusion matrix reveals that half of the wrong predictions are on the middle-aged class.

For phonation types, the accuracy on test set is 49% and the depth of the model is 10. Phonation type with the most wrong predictions is the creaky voice (44% of wrong predic-

tions), while the most well-predicted is modal voice (23% of wrong predictions).

3.1.2. Results for the “combined” condition

The accuracy for the “combined” condition, where all the labels are used, is 13% on the test set (with a depth of 14 and 100% accuracy on the train set). Although lower than the accuracy obtained for individual classification tasks, this performance compares well with the accuracy of a random system, which is around 6% for a task of 17 classes.

3.2. Estimation of the contribution of each attribute

We choose decision trees as classifiers for their simplicity, but also for the high level of explainability they allow. To evaluate which attributes contributed most to the model’s decision, we use the `TreeExplainer` method from the SHapley Additive exPlanations (SHAP) toolbox [31]. We do this independently for gender, age and phonation type classifiers.

The 5 and 10 best attributes in terms of contribution are presented in Table 5, along with the total percentage contribution for both. On average, the top-five attributes contribute 41% to classification, with phonation type classifier’s best attributes having the lowest total contribution (31%). Top-ten attributes increase mean contribution by 11% for classification tasks, with age classifier’s best attributes reaching 60% contribution. Interestingly, two attributes (144 and 193) are shared between the top-ten lists for different classification tasks.

Table 5: *Top-five followed by top-ten BA attributes in terms of contribution and total contribution, for each classification task*

	Gender	Age	Phonation type
	BA59	BA154	BA231
	BA151	BA45	BA206
	BA219	BA135	BA193
	BA228	BA23	BA210
	BA121	BA196	BA88
\sum top-5	42%	47%	31%
	BA144	BA230	BA192
	BA240	BA197	BA185
	BA229	BA147	BA144
	BA53	BA188	BA136
	BA193	BA157	BA111
\sum top-10	53%	60%	39%

In order to determine if the voice traits are encoded by all or only a part of the attributes, we rerun the experiments using only the five or the ten most important attributes to represent the data. We train new decision trees on these data and recount the accuracy as we did previously. Results are summarized in Table 6. For top-five attributes, gender classification accuracy increases from 86% to 88%. Interestingly, the needed depth decreases to only 2. Age classification task also sees an increase in accuracy – from 45% to 48%. For phonation type classification, we observe a large degradation of the test accuracy from 49% to become 36%. The depth needed to obtain the best accuracy on training data (67%) is 5. Shifting the focus on top-ten attributes, we report an increase in accuracy compared to top-five case. Gender classification accuracy increases to 91% when for age classification we see an increase of 12% for accuracy,

Table 6: Performance of decision tree models on the test set, using only the top-five (left) and top-ten (right) attributes as features. Accuracy, precision and recall (using macro averaging) are provided, as well as the model depth and accuracy for the train set

Task	Depth		Acc.		Precision		Recall		Acc.TR	
Gender	2	4	0.88	0.91	0.89	0.91	0.88	0.91	0.91	0.94
Age	5	6	0.48	0.60	0.50	0.62	0.48	0.61	0.72	0.80
Phonation type	5	10	0.36	0.41	0.35	0.44	0.46	0.44	0.67	0.89
Combined	7	11	0.16	0.20	0.16	0.20	0.19	0.22	0.88	0.97

ending up at 60%. Depth increases to 6 by one point to obtain 80% for best accuracy on training data. Finally, accuracy on test data for phonation type classification is still lower than using all attributes, but it increases from 36% to 41% while doubling the depth needed (training accuracy is 91%). It is interesting to determine whether the three sets of best contributors can be used together for the “combined” condition. Using these 15 attributes, we obtain an accuracy of 16% on the test set, which is 3% more compared to the accuracy obtained using full BA vectors (see Table 4). When using top-ten attributes (resulting in 30 attributes in total), accuracy increases to 20% for the test set (to be compared with 13% for all attributes).

Looking at the tree depths, the decision rules for gender classification involve only four attributes for the top-10 (and 2 for the top-5), with an accuracy of 91%. For phonation type, the rules incorporate 10 attributes for an accuracy of only 41%, again for the top-10. This difference shows that gender, a categorical label with a “golden rule”, is directly encoded into the BA attributes whereas phonation type, a continuous label obtained from perceptual evaluation, seems to be only indirectly integrated into them.

4. Discussion

Explainability is a key factor in speaker recognition, particularly for forensic and investigative applications. When the best-performing systems proposed in the literature are unable to provide an explanation of their decision process, a recent alternative, BA-LR, proposes an approach that is inherently explicable, while offering a good explainability/efficiency ratio. BA-LR is based on a speech representation in which an utterance is modeled by a binary vector indicating the presence or absence in it of a set of speech attributes. In BA-LR, the decision process allows a high level of explainability due to the estimation of an LLR independently for each speech attribute, and LLR values depend solely on the presence or absence of the attribute, combined with certain statistical parameters estimated using a reference population. However, speech attributes are discovered automatically and, although an initial description using phonetic parameters has been achieved, their nature is far from interpretable.

The aim of the present work is to develop a methodology for interpreting these attributes in terms of comprehensible explicit voice traits. Our methodology is based on the assumption that if an attribute is useful for a classification task of an explicit voice trait, it means that this attribute encodes all or part of this explicit voice trait.

In this article, we explored three explicit voice traits: gender (representing biological sex), age and phonation type.

Using a BA vector extractor trained on an English database, VoxCeleb2, we built decision-tree classifiers for the gender (two classes), age (three classes), phonation (three classes) and combined (first three together, 17 classes) classification tasks. With

accuracy rates on test data of 86% for gender, 49% for phonation and 45% for age (compared with 50%, 33% and 33% for the corresponding random systems), we demonstrate the links between BA attributes and the three explicit speech features we have selected, paving the way for a comprehensive interpretation of BA-LR speech attributes. This finding is confirmed by the “combined” task, which achieved an accuracy of around three times the performance of the random system.

Next, we determined the attributes that contribute most to the decision for the three individual tasks – first the top-five, then the top-ten. Interestingly, we observed a fairly complete separation between the three subsets (only two attributes are present for two classification tasks, among the ten most contributing attributes), which we believe indicates that the main voice traits are each encoded by a specific subset of BA-LR attributes. Reducing the feature set to selected attributes improved performance for all classification tasks, except for phonation type. Using the top-ten attributes instead of the top-five shows that gender classification is already saturated, while age classification benefits from five additional attributes. This result reinforces the hypothesis we have just stated. For phonation type, a limited set of attributes reaching maximum performance could not be identified among all attributes.

This difficulty in isolating a specific subset of attributes for phonation type can likely be explained by the nature of the variable itself. Unlike gender, which is the only categorical variable, phonation type is continuous, meaning that it requires a threshold for analysis. However, establishing a clear threshold to classify an extract as creaky or breathy has proven challenging. Notably, breathiness is always present in a speaker’s voice, while creaky voice can be intermittently replaced by modal voice, making a three-second sample potentially insufficient for detection. In addition, the BA vector extractor used in this study has its own characteristics, such as input stride and statistical clustering, which may influence its ability to detect creaky or breathy voices – especially when these phonation types are only present in specific parts of the utterance.

Overall, these results are particularly interesting as the BA vector extractor was not designed to characterize the explicit voice traits targeted here, but it does rely in part on them in its decision-making process. It is likely, in view of our results, that information on age and phonation type is embedded in some attributes, but coded differently, perhaps decomposed into finer acoustic features.

For future work, it would be interesting to use a larger dataset to improve the accuracy of the results. We would also to extend our work to other explicit voice traits like nasality or regional accent, in order to interpret more BA-LR attributes. Finally, the integration of phonation information directly into the BA-LR learning process seems interesting: this would allow specific attributes to be dedicated to their detection, and to have some control over what is encoded in the attributes.

5. Acknowledgements

This thesis is co-funded by the Defence Innovation Agency. We would like to thank Imen ben Amor for her work on the BA-LR, without which this work could not be possible.

6. References

- [1] Z. Bai and X.-L. Zhang, "Speaker recognition based on deep learning: An overview," *Neural Networks*, vol. 140, pp. 65–99, 2021.
- [2] J. P. Campbell, W. Shen, W. M. Campbell, R. Schwartz, J.-F. Bonastre, and D. Matrouf, "Forensic speaker recognition," *IEEE Signal Processing Magazine*, vol. 26, no. 2, pp. 95–103, 2009.
- [3] F. Sovrano, S. Sapienza, M. Palmirani, and F. Vitali, "Metrics, explainability and the european ai act proposal," *J*, vol. 5, no. 1, pp. 126–138, 2022.
- [4] I. Ben Amor and J.-F. Bonastre, "Ba-lr: Binary-attribute-based likelihood ratio estimation for forensic voice comparison," in *2022 International workshop on biometrics and forensics (IWBIF)*. IEEE, 2022, pp. 1–6.
- [5] 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 *Proceedings of Interspeech 2023*, 2023.
- [6] I. Ben Amor, J.-F. Bonastre, and D. van der Vloed, "Forensic speaker recognition with ba-lr: calibration and evaluation on a forensically realistic database," in *Odyssey 2024*, 2024.
- [7] M. Candini, E. Zamagni, A. Nuzzo, F. Ruotolo, T. Iachini, and F. Frassinetti, "Who is speaking? implicit and explicit self and other voice recognition," *Brain and cognition*, vol. 92, pp. 112–117, 2014.
- [8] 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.
- [9] J. Hillenbrand, L. A. Getty, M. J. Clark, and K. Wheeler, "Acoustic characteristics of american english vowels," *The Journal of the Acoustical society of America*, vol. 97, no. 5, pp. 3099–3111, 1995.
- [10] L. A. Ramig and R. L. Ringel, "Effects of physiological aging on selected acoustic characteristics of voice," *Journal of Speech, Language, and Hearing Research*, vol. 26, no. 1, pp. 22–30, 1983.
- [11] N. D. Gregory, S. Chandran, D. Lurie, and R. T. Sataloff, "Voice disorders in the elderly," *Journal of Voice*, vol. 26, no. 2, pp. 254–258, 2012.
- [12] K. H. Tarafder, P. G. Datta, and A. Tariq, "The aging voice," *Bangabandhu Sheikh Mujib Medical University Journal*, vol. 5, no. 1, pp. 83–86, 2012.
- [13] J. Kreiman, D. Vanlancker-Sidtis, and B. R. Gerratt, "Defining and measuring voice quality," in *ISCA Tutorial and Research Workshop on Voice Quality: Functions, Analysis and Synthesis*, 2003.
- [14] R. J. Podesva and P. Callier, "Voice quality and identity," *Annual review of applied Linguistics*, vol. 35, pp. 173–194, 2015.
- [15] J. D. Laver, "Voice quality and indexical information," *British Journal of Disorders of Communication*, vol. 3, no. 1, pp. 43–54, 1968.
- [16] E. Gold and P. French, "International practices in forensic speaker comparisons: second survey," *International Journal of Speech, Language and the Law*, vol. 26, no. 1, pp. 1–20, 2019.
- [17] S. J. Park, G. Yeung, J. Kreiman, P. A. Keating, and A. Alwan, "Using voice quality features to improve short-utterance, text-independent speaker verification systems," in *Interspeech*, 2017, pp. 1522–1526.
- [18] J. Přibíl, A. Přibílová, and J. Matoušek, "Evaluation of speaker identification based on voice gender and age conversion," *Journal of Electrical Engineering*, vol. 69, no. 2, pp. 138–147, 2018.
- [19] M. Gordon and P. Ladefoged, "Phonation types: a cross-linguistic overview," *Journal of phonetics*, vol. 29, no. 4, pp. 383–406, 2001.
- [20] A. Chanclu, L. Georgeton, C. Fredouille, and J.-F. Bonastre, "Ptsvox: une base de données pour la comparaison de voix dans le cadre judiciaire," in *Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Volume 1: Journées d'Études sur la Parole*, 2020, pp. 73–81.
- [21] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," *arXiv preprint arXiv:1912.06670*, 2019.
- [22] J. Durand, B. Laks, and C. Lyche, "Le projet pfc (phonologie du français contemporain): une source de données primaires structurées," *Phonologie, variation et accents du français*, pp. 19–61, 2009.
- [23] P. Keating, C. Esposito, M. Garellek, S. Khan, and J. Kuang, "Phonation contrasts across languages," in *Poster presented at the 12th Conference on Laboratory Phonology*, 2010.
- [24] J. Kreiman and M. Garellek, "Perceptual importance of the voice source spectrum from h2 to 2 khz," *The Journal of the Acoustical Society of America*, vol. 130, no. 4_Supplement, pp. 2570–2570, 2011.
- [25] J. Hillenbrand, R. A. Cleveland, and R. L. Erickson, "Acoustic correlates of breathy vocal quality," *Journal of Speech, Language, and Hearing Research*, vol. 37, no. 4, pp. 769–778, 1994.
- [26] Y. Shue, P. Keating, C. Vicens, and K. Yu, "Voicesauce: A program for voice analysis," in *Proceedings of ICPhS 2011*, 2011, pp. 1846–1849.
- [27] P. Boersma, "Praat, a system for doing phonetics by computer," *Glott. Int.*, vol. 5, no. 9, pp. 341–345, 2001.
- [28] M. MohammadAmini, D. Matrouf, J.-F. Bonastre, S. Dowerah, R. Serizel, and D. Jouvét, "Learning noise robust resnet-based speaker embedding for speaker recognition," in *The Speaker and Language Recognition Workshop (Odyssey 2022)*, 2022, pp. 41–46.
- [29] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: a large-scale speaker identification dataset," in *Proceedings of Interspeech 2017*, 2017, pp. 2616–2620.
- [30] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
- [31] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee, "From local explanations to global understanding with explainable ai for trees," *Nature Machine Intelligence*, vol. 2, no. 1, pp. 2522–5839, 2020.