A Phonetic Analysis of Speaker Verification Systems through Phoneme selection and Integrated Gradients

Thomas Thebaud, Gabriel Hernandez Sierra, Sarah Flora Samson Juan, Marie Tahon

A Phonetic Analysis of Speaker Verification Systems through Phoneme selection and Integrated Gradients

[†]Thomas Thebaud, [‡]Gabriel Hernández, [◇]Sarah Flora Samson Juan, [⋆]Marie Tahon,

[†]CLSP, Johns Hopkins University, MD, USA
[‡]CENATAV, Cuba
[◇]University of Malaysia Saraway Malaysia
[⋆]LIUM, Le Mans University, France
tthebau1@jhu.edu - gabrielcuba@gmail.com

Abstract

Speaker recognition systems are usually crafted to identify or verify the identity of a given speaker independently of the linguistic content contained in the utterance used. We use two explainability techniques to analyze the impact of phonetic variations on a speaker verification system using VoxCeleb. We use Whisper and the Montreal Forced Aligner (MFA) to transcribe, then segment phonetically the Voxceleb1 test set. Phoneme selection is first used, before computation of the x-vectors, to observe which phonemes are the most discriminative through their impact on EER and MinDCF metrics. Integrated Gradients are then used to show which phonemes yielded the highest gradients comparing two speakers. We find that for the representation of the x-vector in speaker recognition systems, both consonants and vowels are relevant and important to capture the distinctive characteristics of a speaker's voice and generate effective and discriminative representations.

1. Introduction

Speech has been proven to be a reliable way to identify or verify the identity of an individual. The past years have seen an increasing amount of systems leveraging the latest neural technologies, from x-vector [1] to transformers [2], for the speaker verification task. However, most systems prioritize performance over explainability, and as they are neural networks, they often act as black boxes, making the interpretability of their outputs not trivial. In many domains, such as health or forensics, the prediction of speaker identity from its voice is not enough, and it is necessary to include some explanations [3]. Indeed, the current trend for explainable AI is a vital process for transparency of decision-making with machine learning: the user (a doctor, a judge, or a human scientist) has to justify the choice made based on the system output.

If the goal of a speaker verification system is to model the identity of a given speaker from a given utterance independently of its linguistic content, it has been shown that elements such as the linguistic content [4], noise [5, 6] or emotions [7] impact the predictions of the systems. Acoustic units, such as phonemes, diphones, and syllables, provide a foundation for extracting distinctive voice characteristics and creating vocal fingerprints for each individual in the forensic field [8]. The use of phonetic vectors has also been explored for speaker verification [9]. Before the advent of x-vectors, the i-vectors were designed as phonemic bottleneck [1], thus confirming the relevance of this information for speaker modeling.

The exploration of phoneme categories is one important key point towards explaining the information embed in a trained model for both Automatic Speech Recognition (ASR) or Speaker Verification (SV).

For instance, by extracting some intermediate embedding layers, [10] visualized how phoneme categories are represented in a neural network. The pre-trained self-supervised network WavLM [11] has also been explored through phoneme classification. More generally, different approaches have been used to investigated how pre-trained representations embed the phonetic content. Probing consists in extracting intermediate layers and learn a classifier on top of it [12], and the training of a downstream model which classifies phonemes [13].

In the work done, we delve into the explanatory power of acoustic units, particularly in the context of speech embeddings. By exploring the connection between acoustic units and speech embeddings, our goal is to uncover how these representations contribute to the accuracy of text-independent speaker verification systems.

In particular, we expect frame-level embeddings belonging to the same phonetic categories to be very similar for the same speaker. Understanding this connection between acoustic units and speech embeddings is essential for capturing the nuances and patterns that form the basis of speaker verification.

In this paper, we focus on the phonetic content, and propose new techniques to interpret *a posteriori* the behavior of a trained speaker verification system. To do so, we leverage phonemes selection and integrated gradients [14] to explain the impact of various phonemes on a speaker verification system's outputs and performances. The fact that two different techniques yield similar conclusion better comforts it. To the best of our knowledge, it is the first time the integrated gradients' technique has been used on a speaker verification system.

Section 2 presents the related interpretability and explainability works, with a focus on post-hoc and integrated gradients methods. Then, Section 3 details the performed experiments, which results are presented in Section 4 and discussed in Section 5. Section 6 presents our conclusions on the topic.

2. Related Works

This section presents the related works about explainability and interpretability, then an overview of the phoneme selection and segmentation techniques, as well as various integrated gradients techniques.

2.1. Speaker identity and phonetic content

Speaker identity is known to rely not only on voice timber, but also on phonetic content. Many works in the litterature provides clues that phoneme categories such as vowels and nasals are discriminant for speaker identification [15], but specific phonemes such as /s/ also get high speaker verification performances. In another work, fricatives and stops consonants such as /s/, /t/ or /b/ have been found to perform worse than vowels and nasal [16]. Therefore, the way specific speech sounds are pronounced is a relevant speaker characteristic. Indeed, speaker accent is clearly a discriminant feature.

Historically, speaker verification systems were based on the Mel Frequency Cepstral Coefficients (MFCC) which captures both phonetic and timber information [17]. The impact of phonetic content on speaker verification has also been investigated on i-vectors [8].

Nowadays, most speaker representations are based on x-vectors [1], which are trained with a pooling layer in order to drastically reduce dependencies on linguistic content. Consequently, whatever the lingusitic content, all utterances from the same speaker should have similar representations. Recently, [18] has demonstrated that x-vectors mostly rely on vowels, nasals and fricatives by analysing multi-head attention.

To put back speaker phonetic variability, a phoneme unit specific TDNN has been shown to perform better than the baseline x-vector representation [19]. Another option is to adapt x-vectors to phonetic information [20].

The experiment presented here investigates with two different techniques if some phoneme categories (especially nasals, vowels and fricatives) are more discriminant than others. We follow the ARPABET[21] classification of phonemes, presented in Table 1.

Category	Symbol
vowels	AA, AE, AH, AO, AW, AY, EH, ER, EY, IH, IY, OW, OY, UH, UW
consonants	
fricative	F, V, TH, DH
stop	P, B, T, D, K, G
nasal	M, N, NG
sibilant	S, Z, SH, ZH
affricate	CH, JH
approximant	W, R, Y
lateral	L

Table 1: Phonetic categories from ARPABET[21] alphabet.

2.2. Post-hoc explainability techniques

Explainability for AI can be addressed at different stages of the process. Pre-hoc explainability intends to understand and describe data with explainable features and statistics. Another stage is to develop explainable-by-design models. The last stage consists of the extraction of post-hoc explanations from a pretrained model by the use of proxy models or perturbation mechanisms. Our scope fall within this last approach.

Most post-hoc approaches are model agnostic, such as LIME [22] or SHAP [23]. SHAP values are defined as the change in the expected model prediction when conditioning on an input feature. When moving towards speech, input features are generally time-frequency representations, also called spectrogram. SHAP values can therefore highlight which part of the spectrogram is responsible for a change in prediction [24].

Still within post-hoc analysis, back-propagation techniques such as Layer-wise Relevant Propagation (LRP) are highly popular in the computer vision fields. LRP consists in distributing the output to each neuron incrementally until the input features are reached. In computer vision, computing heatmaps to visualize relevant features is a popular approach. This has been extended to the audio domain by highlighting important areas on time-frequency maps [25]. In the same attribution method family, the integrated gradients approach [14] has the advantage to be implementation invariant and is sensitive to data perturbation. Integrated gradients are explained more in detail in Section 2.3. Finally, it is also possible to investigate the impact of different characteristics of the data on the predictions of the model. One option is to slightly modify the data with a perturbation mechanism and then evaluate the impact on the prediction [26]. Another approach is to select data which contains a specific characteristic (a color, a phoneme, etc.), then to estimate the impact on the prediction.

In the present experiment, we investigate two post-hoc techniques to evaluate to what extent a trained speaker verification model is sensitive to phonetic input.

2.3. Integrated gradients

Integrated gradients(IG) were first proposed as a visualization technique [14] to show the attention of a neural classifier. Multiple approaches have been proposed to improve characteristics of those IG, such as I-GOS [27] that optimize the heatmap produced so that the classifier only needs the highlighted part to work, or the SmoothTaylor technique [28] that smooth the gradients using a first-order Taylor approximation of the classifier and improve the representation. They have since been generalized to other machine learning techniques [29].

The IG are defined for a classifier F, on the ith dimension, as the integral of the gradients change, following a linear path (using a parameter $\alpha \in [0, 1]$) between a given sample x of class y and a baseline x' that should have a neutral prediction, as shown in Equation 1:

$$IG_i(x) = (x_i - x'_i) \int_{\alpha=0}^{1} \frac{\partial F(x' + \alpha(x - x'))}{\partial x_i} d\alpha \qquad (1)$$

To make it easier to compute, we use Riemann's approximation over m samples, as shown in Equation 2:

$$IG_i(x) := (x_i - x'_i) \sum_{k=0}^{m} \frac{\partial F(x' + \frac{k}{m}(x - x'))}{\partial x_i} \frac{1}{m} \qquad (2)$$

However, to the extent of our knowledge, no previous work has used IG on a speaker verification system, which is what we are proposing in this article.

3. Experimental Setup

The experiments carried out were aimed at verifying the selective use of the phonetic content of the speech for speaker recognition systems.

3.1. Datasets

In this article, we use for all our experiments the dev and test sets from Voxceleb 1 [30] and the dev set from VoxCeleb 2 [31] as our speaker verification protocol.

The distribution of the dataset if shown in Table 2. Those

Set	# Speakers	# utterances
VoxCeleb1 dev	1,211	148,642
VoxCeleb1 test	40	4,874
VoxCeleb2 dev	5,994	1,092,009

Table 2: VoxCeleb1 and 2 sets distribution. The *dev* splits were used for training the speaker verification systems, while the *test* split was used for evaluation.

utterances are recorded from various multimedia sources, collected from YouTube, with speakers mainly from the U.S, U.K, Germany, India, and France, speaking English for the most part, including 39% of female speakers.

However, we know that the VoxCeleb1 dataset is not only composed of English speakers [32], which could pose problems with the phoneme segmentation. We found English to represent only 84.96% of the utterances, the VoxCeleb1-test split actually containing sequences with Welsh, Spanish, Norwegian Nynorsk, Georgian, Manx, Urdu, Hindi, Faroese, Malay, Afrikaans, Occitan, Catalan, Haitian, Maori, Assamese, Danish, Telugu, Somali, Galician, French, Bosnian and Dutch languages as well. For the evaluation, we use VoxCeleb test, with the original (O) protocol.

3.2. Phoneme segmentation

Because our aim is to evaluate the impact of the phonetic content, we need to have a phonetic segmentation of our test set, currently VoxCeleb 1 test split.

3.2.1. Transcription

To do so, we first transcribed all utterances using a version of Whisper [33] fine-tuned for time-accurate speech transcription on long-form audios [34]. WhisperX achieve 9.7% Word Error Rate (WER) on the TED-LIUM [35] dataset and 2.2% WER on the Kincaid46 [36] dataset, both being audio recordings in English language. Whisper is originally trained on multilingual data, and is able to transcribe data from multiple languages when prompted. However, we use only English prompts for the whole dataset. We use m-brain implementation of whisperX[1].

To the best of our knowledge, no samples from the VoxCeleb1 test set (O) were used to train the WhisperX model used. WhisperX proposes a temporal alignment of words, but not up to phonemes, which is why we used a second system for the phoneme alignment.

3.2.2. Alignement of phonemes

Once transcribed, each utterance is aligned phoneme-wise to the audio using the Montreal Forced Aligner [37](MFA). The original implementation of the MFA show an accuracy in the phonemes's boundary of 77% with a tolerance of 25ms, and 93% for 50ms tolerance, evaluated on US English on the Buckeye Corpus [38]. The average difference between the boundaries of the gold standard and the predictions is 16.6ms, which is way lower than the length of the windows used for spectrogram computation in the systems we are using We can assume that the MFA will give a boundary on the spectrograms with no more than one frame of error for American English. To the best of our knowledge, no samples from VoxCeleb1 test set (O) were used to train MFA.

[1]https://github.com/m-brain/whisperX

Once transcribed and aligned, we used the utterances of VoxCeleb1 to measure the impact of various phonemes on two different speaker verification systems, using two different techniques. Both systems have been previously trained using Voxceleb1&2 train splits.

3.3. Phoneme selection

The first explainability technique we propose is to measure the variations in the discriminative power of x-vectors when computed from a selected set of phonemes. We employed the pretrained ECAPA-TDNN [39] system from SpeechBrain [40] to analyze the role of linguistic content within utterances in a speaker verification system.

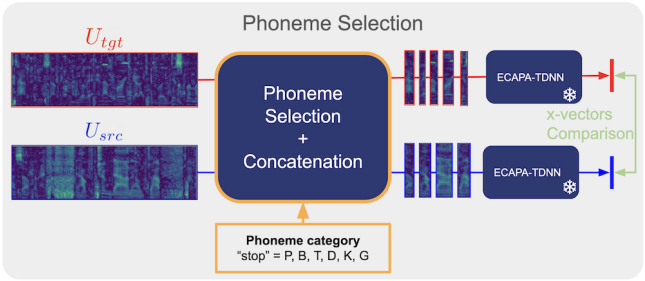

Figure 1: Illustration of the phoneme selection for computing the x-vectors from specific phoneme categories.

The phonetic timestamps define segments in which a single phoneme appears. We also consider as segments non speech part of the signal. As shown in Figure 1, we propose a 5-steps process as described in the followings:

1. The process begins with two utterances (U_{tgt} and U_{src}) and their corresponding phonetic timestamps. As an example (described in Table 3), the U_{tgt} transcription of will be "..., very often I am ..." and its phonetic marks "..., V, EH, R, IY, AO, F, AH, N, AY, EY, M, ...", the U_{src} transcript will be "... to a family ..." and its phonetic marks "... T, AH, AH, F, AE, M, IH, L, IY, ..."
2. At this stage, the phonemes which belong to the analyzed categories are selected. The categories are described below and Table 3 examples the process, providing a visual guide to understanding how phoneme selection is performed within the context of the analysis.
   - **All segments**: full utterance is included.
   - **All-MFA-phonemes**: only segments marked with phonetic content (vowels or consonants) are used.
   - **Common phonemes**: only segments that share phonemes between the target and source utterances are used.
   - **Consonants**: only segments that contain consonants are considered.
   - **Vowels**: only segments that contain vowels are considered.
   - In general, any of the categories present in Table 1 can be used.
3. Once a set of phonemes is selected, we use the phoneme segmentation to extract only the desired segments, and concatenate them as a new utterance.

4. Using our pre-trained ECAPA-TDNN system, we extract the x-vector associated to each concatenated utterance.
5. Finally, we compare pairs of x-vectors using cosine similarity (a commonly used distance metric for speaker embedding extracted from ECAPA-TDNN systems [39]) and compute the minDCF and Equal Error Rate.

This process allows us to compare how much information is conveyed by each phoneme category through two different metrics. Another option would have been to extract x-vector for each segment, however, we know that x-vectors are not robust enough to be computed on very short speech segments which are of the same order of magnitude as a frame. That is the reason why we decided to concatenate shared segments. Anyway, longer segments always contain more information, so we have to randomly select a subset of the phonemes to be able to compare two categories occupying a different percent of time in the dataset.

All segments (full utterance)	U_{tgt} : ..., very often I am ... U_{src} :... to a family ...
All-MFA-phonemes	V, EH, R, IY, AO, F, AH, N, AY, EY, M T, AH, AH, F, AE, M, IH, L, IY
Common phonemes	AH, F, M, IY AH, AH, F, M, IY
Consonants	V, R, F, N, M T, F, M, L
Vowels	EH, IY, AO, AH, AY, EY AH, AH, AE, IH, IY

Table 3: An example of the process of selecting phonetic marks for each proposed category is illustrated.

All experiments are evaluated on the VoxCeleb1 Original test set [30]. Performance evaluation will be based on two metrics: Equal Error Rate (EER) and minimum normalized detection cost (minDCF) with $P_{target} = 10^{-2}$ and $C_{FA} = C_{Miss} = 1$, as for the NIST SRE speaker verification challenges [41]. These metrics will allow us to measure the effectiveness of the speaker verification system for each phonetic category chosen.

3.4. Integrated gradients for speaker verification

Integrated gradients [27] have been used previously for classification tasks, as a visualization technique to show which areas of an input were used to predict a given class. Here, we propose to modify the initial technique and adapt it to a speaker verification system. Our model will highlight the areas on a time-frequency representation from a source utterance U_{src} that would impact its similarity to a target utterance U_{tgt}. Then, we use the computed IG, in addition to the previously computed phoneme segmentation, to measure and compare the power of the IG over each category of phonemes.

3.4.1. Computation of the Integrated Gradients

As shown in Figure 2, the proposed process is the following:

1. We train a ResNet34 system (similar to the one presented in [42]) using the Hyperion toolkit[2] with the VoxCeleb1&2 train sets.

[2]https://github.com/hyperion-ml/hyperion

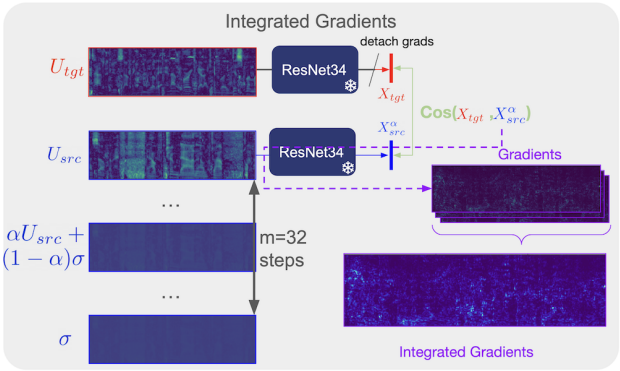

Figure 2: Computation of integrated gradients on a frozen speaker verification system, given a couple of utterances U_{src} and U_{tgt} from the same speaker, using $m = 32$ steps for the Riemann's integration.

2. For each source utterance U_{src} from the original test split of VoxCeleb1, we find a random target utterance from the same speaker, of a different session U_{tgt}.
3. For both utterances, we compute their x-vectors X_{src} and X_{tgt}, and the **cosine similarity** between both.
4. Being from the same speaker, the similarity should be 1, so we compute the gradient of the difference between the similarity and 1, related to the source utterance $\nabla(cosine(X_{src}, X_{tgt}) - 1)$.
5. To approximate the Riemann's integral with 32 steps, we compute 32 times the gradients, as shown in Equation 2, using a linear interpolation of the source utterance and a Gaussian noise σ of the same dimensions.
6. Once integrated, we have the integrated gradients for a given source utterance U_{src}.

3.4.2. Explaining the integrated gradients

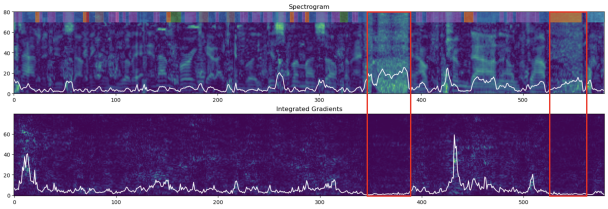

Figure 3: Spectrogram of an utterance from VoxCeleb1 (top), it's associated integrated gradients (bottom), and the phoneme segmentation colored by category of phonemes (upper part). Red boxes show parts with no phonemes detected. In white is the power of the spectrogram and the integrated gradients.

Figure 3 shows side to side the spectrogram of an utterance, with the associated integrated gradients, and the phonetic segmentation (upper part). We can see that the gradients are not uniform temporally nor by frequencies, and that when no phonemes were detected, there are almost no gradients.

As integrated gradients are not uniform, we are looking into what impacts their distribution, or which area of the spectrogram yield more information for speaker recognition systems. We look at three aspects that could impact the system:

1. The power: the correlation between the power of the spectrogram and the power of the integrated gradients.

2. The linguistic aspect: Comparing various phonemes and categories of phonemes.
3. The inter-speaker variations: comparing the previous quantities on average vs per speaker.

Power of the integrated gradients Once the Integrated Gradients have been computed, we are measuring which parts of the utterances were used the most, by looking at the power of the spectrograms for each frame. Our first experiment measures the cross-correlation between the power of the spectrogram and the integrated gradients. As we saw that the areas where no phonemes were detected yielded no gradients, we also compute the same correlation using only the areas with MFA phonemes.

Phoneme selection for the Integrated Gradients Each utterance is sliced by phoneme, following the segmentation produced in Section 3.2. To avoid biases linked to the various powers linked to each phoneme, the power of each segment is then computed to normalize the gradients by the average power of the spectrogram. Then, the average power of the integrated gradients for each phoneme is computed. Averaging will compensate for the inner variability in the length of each category of phonemes.

Inter-speakers variations Once we computed the gradients power per phonemes, we measured the variability of those powers between the 40 speakers of voxceleb1, to show if there is a difference in which phonemes are used per each speaker for their identification.

We show and explore the various powers per phonemes, categories of phonemes and speakers in Section 4.2.

4. Results

4.1. Phoneme selection performances

The performance of the systems obtained by selecting various sets of phonemes is presented in Table 4.

To ensure consistency in the use of timestamps for each phoneme, it is necessary to verify how close the x-vectors obtained using all phonemes (All-MFA-phonemes) for each utterance are in comparison to the construction of x-vectors without phonetic marks (All segments). The two first lines of Table 4 shows the baseline result, as well as the same metrics computed only on MFA-phonemes. The use of the phonetic content for each utterance obtains approximately the same performance as the baseline, 0.91 vs 0.81 of EER. This result may show that the temporal alignment of the phonemes for each utterance is correct and that the voice regions used contain the discriminatory information of each speaker.

By using only "common phonemes", the speaker verification achieves an EER of 2.35 using only 45 % of the time for each comparison. This approach is compared with the "All segments*" and "All-MFA-phonemes*" using a random selection method to adjust the time in each experiment, obtaining significantly better performance. This result shows that it is possible to obtain and utilize the common phonetic information between the compared expressions in text-independent speaker recognition systems, achieving greater effectiveness using only a small percentage of total duration.

From lines "All Consonants" and "All vowels" of Table 4, we show that both consonants and vowels are important in identifying and distinguishing a person's voice. In terms of proportion, about 60% of the letters in the English language are consonants, while about 40% are vowels. This distribution is reflected in words and everyday speech, where more consonants are used than vowels. Reflecting this are the results obtained using the consonants and vowels separately to obtain the x-vectors. With consonants, 4.17% EER is achieved using 46.5 percent of the time of VoxCeleb1 utterances, while with the vowels 6.22 % EER is achieved using only 37 percent of the time. If we equate the average time used (37%) for both consonants and vowels, then the EER between the two categories is very similar.

Category	↓ EER	↓ minDCF	% of time
All segments	0.81	0.05	100
All MFA-phonemes	0.91	0.04	83.6
All segments*	0.93	0.09	83.6
Common phones	2.35	0.18	45
All-MFA-phonemes*	2.82	0.24	45
All segments*	3.49	0.29	45
All Consonants	4.17	0.30	46.5
All Vowels	6.22	0.37	37.9
All Consonants*	6.41	0.41	37.9
fricative	36.05	0.967	5.95
stop	43.13	0.999	14.50
nasal	34.52	0.991	8.53
sibilant	38.40	0.992	7.81
affricate	27.77	0.222	0.98
approximant	34.18	0.988	5.68
lateral	33.13	0.682	3.06

Table 4: Performances on classification tasks, using various phoneme selections. When we want to compare systems using the same amount of time, random phonemes are removed until we reach the same amount, the lines are shown by using a star (*).

4.2. Integrated Gradients Powers per Phonemes

Correlations between spectrogram power and integrated gradients First, we compared the cross-correlation between the power of the spectrogram and the power of the gradients. The correlation computed for each time frame is $29.8\% \pm 3.6\%$ while measuring for each time frame including a detected phoneme, it slightly raises to $30.5\% \pm 3.6\%$. There is indeed a correlation between the power of the spectrogram and the intensity of the gradients, so we normalize the integrated gradients by the power of the spectrogram in the next experiments.

Phoneme-wise integrated gradients Then, after having computed the powers for each phoneme, normalized by the time used by each slice, we average them by phonetic categories, as shown in Table 1. Table 5 shows the average power of the Integrated Gradients per phoneme category.

Speaker-wise integrated gradients In Figure 4 we can see the average power for each phoneme category, for each of the 40 speakers in the VoxCeleb1 test dataset.

We can see here that there is a visible inter-speaker variation of the integrated gradients power: the model pays attention at different phonemes for different speakers. However, the conclusions linked to the averages presented in Table 5 are still valid: nasals, vowels and fricatives contain generally more power while sibilant and affricate contain less.

To compare both analysis methods, we use Pearson's coefficient correlation between the EER of selected phonemes and the power of integrated gradients. However, we can not show that there is a correlation between them (p values ≥ 0.18).

Category	Average Power ↑	% of time
All segments	0.8326	100 %
All MFA-phonemes	0.4477	83.6 %
All but MFA-phonemes	0.5191	16.4%
All Vowels	0.4762	37.9 %
All Consonants	0.4256	46.5 %
fricative	0.4944	5.95 %
stop	0.4213	14.50 %
nasal	0.4602	8.53 %
sibilant	0.3220	7.81 %
affricate	0.2339	0.98 %
approximant	0.4596	5.68 %
lateral	0.5054	3.06 %

Table 5: Average power of the Integrated Gradients for all the segments, only for the detected phonemes, and by phoneme categories.

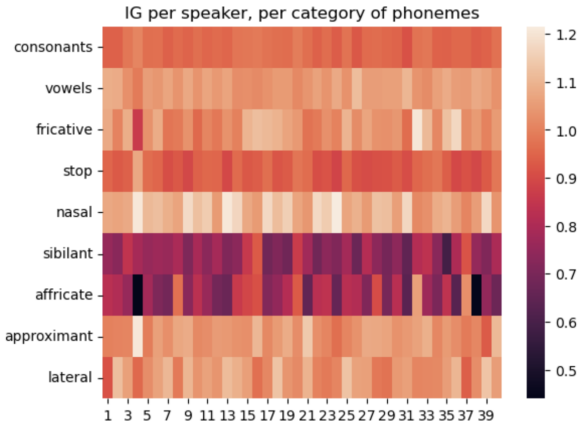

Figure 4: Average power for each phoneme category, for each of the 40 speakers in the VoxCeleb1 dataset. The power has been normalized per speaker for a better comparison.

5. Discussion

The techniques presented all have their limits, and we are discussing them in this section.

The first limit is linked to the variety of languages considered: We use, for the phoneme segmentation, systems that have been proven to give their best results only for English. If Whisper is a multilingual system, and the MFA is too, both have their best results on the English language, and 15% of the dataset evaluated is composed of other languages, some being considered as underresourced languages.

The second limit is the phoneme-by-phoneme approach. Even if we assumed that the segmentation was perfect, cutting between phonemes extract a whole portion of any phonetic analysis: how does the transitions (especially co-articulation) between phonemes affect the system? We might be destroying the performances of the system when pruning too much, only because of the transitions that are ignored here.

Third there is the non-articulated sounds, that are being ignored by the whole analysis, which actually seem to contain some information, as removing them rises the EER from 0.81% to 0.91%.

6. Summary and Conclusion

In this article, we explore two explainability techniques to perform a phonetic analysis of speaker verification systems. Using Whisper to transcribe VoxCeleb1, then the MFA to align the phonemes, we obtained a phoneme segmentation of our testing dataset, which we used to perform experiments on various categories of phonemes.

First, we explored how selecting only certain phonemes would impact the performances of a pretrained ECAPA-TDNN speaker verification system. We show that it is possible to use common phonetic information between compared utterances in text-independent speaker recognition systems, implying greater effectiveness using only a small percentage of utterance time. Furthermore, for the speaker recognition systems based on the x-vector representation, both consonants and vowels are relevant and crucial. They play a significant role in capturing the distinctive voice characteristics of a speaker and generating effective and discriminative representations.

Second, we adapted a visualization technique, the integrated gradients, for speaker verification, and measured which areas of a given utterance attracted more of the *attention* from a pretrained ResNet34 speaker verification system. We found out that there was a high cross correlation between the power of a spectrogram and the power of the gradients, thus we normalized the gradients by the power of the spectrogram. Then, we showed that the variability between different phonemes categories also transfers to the integrated gradients, we found out the same conclusions from the phoneme selection technique on broad categories of phonemes, even if the various consonants categories behave more erratically. We also show that if there are inter-speaker variations, the amount of information present for each speaker in each category of phonemes still stays similar.

Overall, we proposed two original and complementary approaches for phonetic investigation, allowing to better explain the phonetic impact on various speaker verification systems. Due to technical reasons, both approaches have been initially explored independently. The fact that similar results are obtained with different speaker models strengthen our conclusion.

Our segmentation performance is hindered by the presence of multiple languages within the test dataset, leading to imperfections. Therefore, our future efforts aim to enhance our methodology by incorporating multilingual analysis, utilizing language-specific segmentation models tailored to each language's phonetic characteristics. Additionally, we seek to delve into bi-phone analysis to more accurately capture the effects of transitions between phonemes, thereby improving the precision of our segmentation approach.

7. Acknowledgments

This work was performed using HPC resources from GENCI–IDRIS (Grants 2022-AD011012565 and AD011012527). This project has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No 101007666. The research reported here was conducted at the 2023 Frederick Jelinek Memorial Summer Workshop on Speech and Language Technologies, hosted at Le Mans University (France) and sponsored by Johns Hopkins University.

8. References

[1] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur, "X-Vectors: Robust DNN Embeddings For Speaker Recognition," in *ICASSP*, 2018.

[2] Junyi Peng, Oldřich Plchot, Themos Stafylakis, Ladislav Mosner, Lukáš Burget, and Jan "Honza" Černocký, "Improving Speaker Verification with Self-Pretrained Transformer Models," in *Proc. INTERSPEECH 2023*, 2023, pp. 5361–5365.

[3] J. P. Campbell, W. Shen, W. M. Campbell, R. Schwartz, J.-F. Bonastre, and D. Matrouf, "Forensic speaker recognition," *IEEE*, Mar. 2009.

[4] Tianchi Liu, Rohan Kumar Das, Maulik Madhavi, Shengmei Shen, and Haizhou Li, "Speaker-Utterance Dual Attention for Speaker and Utterance Verification," in *Proc. Interspeech 2020*, 2020, pp. 4293–4297.

[5] David Snyder, Guoguo Chen, and Daniel Povey, "Musan: A music, speech, and noise corpus," *arXiv preprint arXiv:1510.08484*, 2015.

[6] Chunlei Zhang, Meng Yu, Chao Weng, and Dong Yu, "Towards robust speaker verification with target speaker enhancement," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6693–6697.

[7] R. Pappagari, T. Wang, J. Villalba, N. Chen, and N. Dehak, "X-Vectors Meet Emotions: A Study On Dependencies Between Emotion and Speaker Recognition," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2020, pp. 7169–7173, ISSN: 2379-190X.

[8] Moez Ajili, Jean-François Bonastre, Waad Ben Kheder, Solange Rossato, and Juliette Kahn, "Phonetic content impact on Forensic Voice Comparison," in *2016 IEEE Spoken Language Technology Workshop (SLT)*, San Diego, France, Dec. 2016, pp. 210–217, IEEE.

[9] Yi Liu, Liang He, Jia Liu, and Michael T. Johnson, "Speaker Embedding Extraction with Phonetic Information," in *Proc. Interspeech 2018*, 2018, pp. 2247–2251.

[10] Odette Scharenborg, Sebastian Tiesmeyer, Mark Hasegawa-Johnson, and Najim Dehak, "Visualizing Phoneme Category Adaptation in Deep Neural Networks," in *Proc. Interspeech 2018*, 2018, pp. 1482–1486.

[11] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei, "WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing," *IEEE Journal of Selected Topics in Signal Processing*, pp. 1–14, 2022.

[12] Danni Ma, Neville Ryant, and Mark Liberman, "Probing acoustic representations for phonetic properties," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 311–315.

[13] Dan Wells, Hao Tang, and Korin Richmond, "Phonetic Analysis of Self-supervised Representations of English Speech," in *Interspeech 2022*. Sept. 2022, pp. 3583–3587, ISCA.

[14] Mukund Sundararajan, Ankur Taly, and Qiqi Yan, "Axiomatic attribution for deep networks," in *International conference on machine learning*. PMLR, 2017, pp. 3319–3328.

[15] J.P. Eatock and J.S. Mason, "A quantitative assessment of the relative speaker discriminating properties of phonemes," in *Proceedings of ICASSP '94. IEEE International Conference on Acoustics, Speech and Signal Processing*, 1994, vol. i, pp. I/133–I/136 vol.1.

[16] R. Kashyap, "Speaker recognition from an unknown utterance and speaker-speech interaction," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 6, pp. 481–488, 1976.

[17] D.A. Reynolds and R.C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 1, pp. 72–83, Jan. 1995, Conference Name: IEEE Transactions on Speech and Audio Processing.

[18] B. Shaik Mohammad Rafi, Sreekanth Sankala, and K. Sri Rama Murty, "Relative significance of speech sounds in speaker verification systems," *Circuits Syst Signal Process 42*, p. 5412–5427, 2023.

[19] Xianhong Chen and Changchun Bao, "Phoneme-unit-specific time-delay neural network for speaker verification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1243–1255, 2021.

[20] Y. Liu, L. He, J. Liu, and et al., "Introducing phonetic information to speaker embedding for speaker verification.," in *Journal of Audio Speech and Music Processing*, 2019, vol. 19.

[21] John S Garofolo, Lori F Lamel, William M Fisher, Jonathan G Fiscus, and David S Pallett, "Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1," *NASA STI/Recon technical report n*, vol. 93, pp. 27403, 1993.

[22] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin, ""Why Should I Trust You?": Explaining the Predictions of Any Classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA, Aug. 2016, KDD '16, pp. 1135–1144, Association for Computing Machinery.

[23] Scott M Lundberg and Su-In Lee, "A Unified Approach to Interpreting Model Predictions," *NIPS*, pp. 4768–4777, 2017.

[24] Sunit Sivasankaran, Emmanuel Vincent, and Dominique Fohr, "Explaining Deep Learning Models for Speech Enhancement," in *Interspeech 2021*. Aug. 2021, pp. 696–700, ISCA.

[25] Sören Becker, Marcel Ackermann, Sebastian Lapuschkin, Klaus-Robert Müller, and Wojciech Samek, "Interpreting and Explaining Deep Neural Networks for Classification of Audio Signals," *arXiv:1807.03418 [cs, eess]*, Oct. 2019, arXiv: 1807.03418.

[26] Thomas Fel, Melanie Ducoffe, David Vigouroux, Rémi Cadène, Mikael Capelle, Claire Nicodème, and Thomas Serre, "Don't lie to me! robust and efficient explainability with verified perturbation analysis," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 16153–16163.

[27] Zhongang Qi, Saeed Khorram, and Fuxin Li, "Visualizing deep networks by optimizing with integrated gradients.," in *CVPR Workshops*, 2019, vol. 2, pp. 1–4.

[28] Gary SW Goh, Sebastian Lapuschkin, Leander Weber, Wojciech Samek, and Alexander Binder, "Understanding integrated gradients with smoothtaylor for deep neural network attribution," in *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 2021, pp. 4949–4956.

[29] John Merrill, Geoff Ward, Sean Kamkar, Jay Budzik, and Douglas Merrill, "Generalized integrated gradients: A practical method for explaining diverse ensembles," *arXiv preprint arXiv:1909.01869*, 2019.

[30] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: a large-scale speaker identification dataset," in *INTERSPEECH*, 2017.

[31] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman, "Voxceleb2: Deep speaker recognition," *arXiv preprint arXiv:1806.05622*, 2018.

[32] Andrew Brown, Jaesung Huh, Joon Son Chung, Arsha Nagrani, Daniel Garcia-Romero, and Andrew Zisserman, "Voxsrc 2021: The third voxceleb speaker recognition challenge," *arXiv preprint arXiv:2201.04583*, 2022.

[33] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever, "Robust speech recognition via large-scale weak supervision," in *International Conference on Machine Learning*. PMLR, 2023, pp. 28492–28518.

[34] Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman, "Whisperx: Time-accurate speech transcription of long-form audio," *INTERSPEECH 2023*, 2023.

[35] Anthony Rousseau, Paul Deléglise, and Yannick Esteve, "Ted-lium: an automatic speech recognition dedicated corpus.," in *LREC*, 2012, pp. 125–129.

[36] J Kincaid, "Which automatic transcription service is the most accurate?—2018," 2018.

[37] Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger, "Montreal forced aligner: Trainable text-speech alignment using kaldi.," in *Interspeech*, 2017, vol. 2017, pp. 498–502.

[38] Mark A Pitt, Keith Johnson, Elizabeth Hume, Scott Kiesling, and William Raymond, "The buckeye corpus of conversational speech: Labeling conventions and a test of transcriber reliability," *Speech Communication*, vol. 45, no. 1, pp. 89–95, 2005.

[39] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck, "Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," *arXiv preprint arXiv:2005.07143*, 2020.

[40] Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba, Jianyuan Zhong, et al., "Speechbrain: A general-purpose speech toolkit," *arXiv preprint arXiv:2106.04624*, 2021.

[41] M Przybocki and A Martin, "The nist year 2003 speaker recognition evaluation plan," 2003.

[42] Hossein Zeinali, Shuai Wang, Anna Silnova, Pavel Matějka, and Oldřich Plchot, "But system description to voxceleb speaker recognition challenge 2019," *arXiv preprint arXiv:1910.12592*, 2019.