Audio Engineering Society

Convention Express Paper 245

Presented at the AES 156th Convention
2024 June 15–17, Madrid, Spain

This Express Paper was selected on the basis of a submitted synopsis that has been peer-reviewed by at least two qualified anonymous reviewers. The complete manuscript was not peer reviewed. This Express Paper has been reproduced from the author's advance manuscript without editing, corrections, or consideration by the Review Board. The AES takes no responsibility for the contents. This paper is available in the AES E-Library (http://www.aes.org/e-lib), all rights reserved. Reproduction of this paper, or any portion thereof, is not permitted without direct permission from the Journal of the Audio Engineering Society.

Enhancing Speaker Verification through Trainable Audio Features

Paris Xylogiannis[1,†], Alexandros Emvoliadis[1, †], Nikos Vryzas[1], Lazaros Vrysis[1], and Charalampos Dimoulas[1]

[1]*Multidisciplinary Media & Mediated Communication Research Group (M3C), Aristotle University of Thessaloniki (AUTh)*
[†]*Authors contributed equally*

Correspondence should be addressed to Paris Xylogiannis & Alexandros Emvoliadis (pxylog@jour.auth.gr, alemvoliadis@gmail.com)

ABSTRACT

This study presents a novel pipeline for speaker verification, exploiting the advantages of trainable audio features to enhance identification accuracy and efficiency. Unlike traditional systems that rely heavily on spectrograms and handcrafted features like Mel-Frequency Cepstral Coefficients (MFCCs), our approach incorporates a learnable audio frontend encoder, a speaker embeddings system, and a decoder for signal reconstruction. This pipeline employs audio encoding methods to directly process raw audio inputs, dynamically adapting its filters to extract reliable speaker-specific features. By incorporating trainable features, our system not only captures a broader range of speaker information but also offers flexibility in handling either direct output from the audio encoder or signals reconstructed by the decoder. This capability allows for an in-depth analysis of the impact of direct feature extraction on speaker identification tasks. The proposed pipeline is evaluated on a frequently used speaker verification dataset. Our findings demonstrate significant improvements over the baseline's feature extraction method, highlighting the potential of trainable features in redefining the landscape of speech processing systems.

1 Introduction

Traditional speaker verification systems rely on spectrograms and handcrafted features like MFCCs. Recent studies raise the question of whether these methods, while being interpretable, may not capture the full range of audio characteristics. While these studies has explored trainable features for audio processing tasks, their application in speaker verification and diarization remains underexplored. Existing works on audio-related tasks primarily focus on the computational efficiency and speed of trainable features compared to handcrafted methods. This paper investigates the potential of trainable features specifically for speech-based tasks, considering both their performance and the impact of a reconstruction constraint.

In recent years, state-of-the-art speaker verification

models like X-vectors [1] and ECAPA-TDNN [2] have demonstrated significant performance gains by employing deep neural networks (DNN) to extract speaker embeddings directly from spectral representations. X-vectors are built upon Time Delay Neural Networks (TDNNs), designed specifically to handle the temporal nature of speech input. Typically, systems as such exploit MFCCs or filterbank-based features as input. The produced representation is further processed through layers that operate on short segments of the spectrogram, extracting local patterns. This information is exploited through layers that aggregate frame-level statistics over time, capturing longer-term speaker characteristics. This produces a fixed-dimension 'x-vector'.

LEAF [3] (Learnable Audio Frontend) and its optimized variant, EfficientLEAF [4], represent significant advancements in the field of audio processing by focusing on the trainable features aspect for audio-related tasks. LEAF is designed to serve as a trainable frontend for audio processing tasks, dynamically adapting its filters to optimize feature extraction directly from raw audio signals. This approach examines the capability of learnable audio frontends to capture more robust audio characteristics than traditional handcrafted features. Furthermore, an examination of the potential enhancements of the performance for speaker verification and diarization systems is required. EfficientLEAF further refines this concept by focusing on computational efficiency and reducing resource consumption without compromising the quality of feature extraction. By integrating LEAF or EfficientLEAF as the initial stage in speaker verification pipelines, systems can benefit from features that are directly optimized for the task at hand, potentially leading to more accurate and efficient speaker recognition capabilities. This shift towards trainable features marks a pivotal move away from reliance on static, handcrafted methods, paving the way for more flexible and powerful speech processing systems that can better capture the intricacies of human speech and speaker identities.

Vocoders, short for "voice encoders," are a category of audio processing tools that analyze and synthesize the human voice by separating the signal into its component parts. Typically, a vocoder consists of two main components: an analysis module and a synthesis module. The analysis module extracts important features from a source audio signal, typically separating the spectral characteristics of the sound from the temporal aspects. In the context of speaker verification and diarization, vocoders can be instrumental in reconstructing voices from sparse features, allowing for a more stabilized training process regarding the learnable audio frontend.

In this work, we introduce an innovative pipeline for enhancing speaker verification by harnessing the potential of trainable audio features. This pipeline integrates three key components: a learnable audio frontend encoder, a speaker verification system, and a vocoder to introduce a reconstruction constraint to the training process of the feature extraction module. Initially, the pipeline employs an encoder to process raw audio inputs, exploiting trainable filters to extract informative, speaker-specific features directly from the audio signal. The pipeline is uniquely flexible: the speaker embeddings model can receive either the direct output of the audio frontend encoder or a signal reconstructed by the vocoder. We investigate the performance of our proposed system on a well known speaker verification dataset.

The rest of the paper is structured as follows:

1. Section 2 details the employed components that form the proposed system.
2. Section 3 describes the conducted experiments and the obtained results
3. Lastly, Section 4 concludes the paper.

2 Methods

2.1 Learnable Audio Frontend

LEAF consists of multiple layers utilizing learnable filters, including a learnable compression layer. Temporal downsampling is also performed through trainable low-pass filtering. However, the convolutional layers (Gabor filterbank and temporal pooling) and the normalization step, make the learnable feature extraction process significantly slower compared to the traditional spectrogram generation. In the present study, we use



Fig. 1: The EfficientLEAF model pipeline.

EfficientLEAF, a model introduced as a computational improvement upon LEAF. As described in the original paper, EfficientLEAF adapts convolution window sizes and strides dynamically for subsets of filters and replaces the sequentially computed normalization, Per-Channel Energy Normalization (PCEN) [5], with a learnable logarithmic compression, temporal median subtraction and temporal batch normalization.

2.2 Speaker Embeddings

The primary objective of this work is to investigate and enhance the learnable front-end audio processing capabilities in tasks related to speaker verification. The proposed pipeline includes both trainable audio features and a decoder module. Utilizing a lightweight speaker embedding model is important to balance and reduce computational demands during training. Furthermore, it is crucial to ensure the system's efficiency during inference in real-world applications where processing speed is a critical factor that influences usability on low-resource devices.

Given these considerations, we employ the X-vector architecture for speaker embeddings generation. X-vectors, although computationally inexpensive, have demonstrated robust performance in various speaker-related tasks, making them a suitable choice for the proposed system. The model's architecture for 24-dimensional features is presented in Table 1. The baseline system with fixed audio representations is trained using filterbanks as input features.

Table 1: Architecture details of the X-vector speaker embedding model

Layer	Layer context	Total context	Input × output
frame1	[t - 2, t + 2]	5	120×512
frame2	{t - 2, t, t + 2}	9	1536×512
frame3	{t - 3, t, t + 3}	15	1536×512
frame4	{t}	15	512×512
frame5	{t}	15	512×1500
stats pooling	{0, T}	T	1500×3000
segment6	{0}	T	3000×512

2.3 Mel-GAN model

MelGAN [6] is a generative adversarial network [7] designed for efficient and high-quality audio synthesis, particularly in generating mel-spectrograms from audio signals. Unlike traditional vocoders [8, 9] that often require complex and computationally expensive algorithms, MelGAN offers a lightweight, faster alternative that can operate in real-time with minimal latency. This makes MelGAN an ideal choice for applications requiring immediate audio synthesis, such as real-time speech processing systems. The MelGAN architecture comprises a generator and multiple discriminators functioning at different scales. The generator is responsible for converting noise vectors into audio waveforms, aiming to produce outputs indistinguishable from real audio samples. It uses transposed convolutions to upscale the input features progressively, synthesizing audio at the waveform level from the input mel-spectrogram features. On the other hand, the discriminators evaluate the authenticity of the generated audio at different resolutions, providing robust feedback to the generator for improvement through adversarial training. This multi-scale approach allows MelGAN to capture and reproduce the finer details of audio textures and patterns effectively.

In the proposed pipeline, MelGAN receives as input the generated spectrotemporal representation that EfficientLEAF produces. Thus, the generator is conditioned and generates raw audio waveforms. Both real and generated waveforms serve as input to the discriminator part. By incorporating speaker characteristics into the training process, the MelGAN can better model voice nuances, ensuring that the synthesized speech not only has high fidelity but also retains speaker-specific features. MelGAN is trained upon the Least Square GANs loss, as formulated above for the discriminator and the generator accordingly:

$$L_{Disc} = \frac{1}{2}E_{\text{real}}\left[(D(x)-1)^2\right] + \frac{1}{2}E_{\text{fake}}\left[D(G(z))^2\right] \tag{1}$$

$$L_{Gen} = \frac{1}{2}E_{\text{fake}}\left[(D(G(z))-1)^2\right] \tag{2}$$

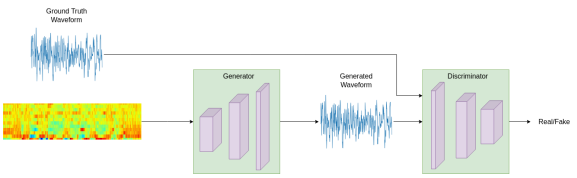


Fig. 2: The Mel-GAN model pipeline.

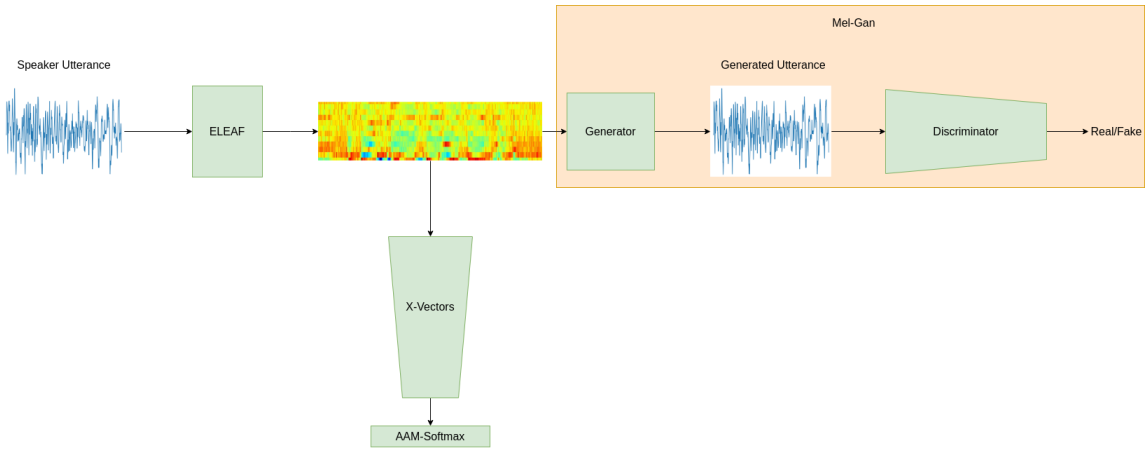


Fig. 3: Our end-to-end trainable proposed pipeline.

while the model also employs a feature loss to stabilize the training procedure:

$$L_{Feats} = \sum_{i=1}^{L} \frac{1}{N_i} \|D^{(i)}(x) - D^{(i)}(G(z))\|_1 \qquad (3)$$

2.4 Overall Pipeline

Finally, the proposed pipeline is detailed within this section. Raw audio is passed through the EfficientLEAF module, which is responsible for generating spectrotemporal representations. These serve as input to both the X-vectors and the MelGAN models. The system is trained in an end-to-end fashion. Thus, a combination of losses is adopted to guide the system to its optimal solution. Specifically, we omit the softmax layer from the original X-vector model and replace it with the Additive Angular Margin Softmax (AAM-Softmax) [10] which has been exploited in numerous studies from face verification to speaker verification [11]. Moreover, we have incorporated the losses employed for the MelGAN training. The proposed pipeline is depicted in Figure 3.

3 Experiments & Results

For our experiments we considered the LibriSpeech dataset [12], a corpus of approximately 1000 hours of read English speech with sampling rate of 16 kHz for each conducted experiments. Specifically, we have trained the X-vectors along with the AAM-Softmax as a baseline model. We have also set the EfficientLEAF just before the X-vector speaker embedding model, as a trainable features extraction component. Finally, the MelGAN model is introduced as an attempt to stabilize the representations produced from the Learnable Audio Frontend component.

Regarding the Speaker Embeddings model, apart from the 24 mel filter configurations, we have also introduced a model that receives a representation composed of 80 frequency components. With this setup, we focus on the examination of the behavior of each component in our system, when it comes to an increased information capacity with respect to the backbone model. We use multiple data augmentation techniques during training to enhance model robustness, incorporating audio from the MUSAN [13] and RIR [14] datasets. Moreover, we apply SpecAugment [15] on the learnable features of the samples to mitigate overfitting.

Each configuration is trained on 360 hours of speech data for 80 epochs. We have used Adam optimizer with an initial learning rate of $1e-4$, decaying every 20 epochs. We have evaluated the performance of each configuration using the Accuracy and the Equal Error Rate (EER), which are commonly used in classification and verification tasks, on the Librispeech clean speech test set.

4 Conclusions and Discussion

Through Table 2 it is evident that the configurations that employ the 80-mel filter backbone shows better

Table 2: EER and Accuracy performance of all systems on the Librispeech clean test set.

Architecture	24-channels		80-channels	
	EER (%) ↓	**Accuracy (%) ↑**	**EER (%) ↓**	**Accuracy (%) ↑**
X-vectors	26.88	31.52	11.85	31.01
ELEAF	4.94	33.66	3.58	57.67
ELEAF-MelGAN	**4.92**	**36.12**	**3.57**	**58.90**

performance than the 24-mel filter backbone. However, both backbone models perform poorly when operate as standalone components. The introduction of the EfficientLEAF module enhances performance regarding both evalution metrics. Finally, when employing the MelGAN component, the system outperforms any other configuration regardless of the capacity of the backbone.

Building upon these foundational assessments, further investigations will explore the potential discrepancies between machine-generated and human-perceived audio cues. This will involve a deeper quantitative and qualitative analysis to understand the perceptual thresholds at which differences become significant. Additionally, comparative studies will be conducted to assess the fidelity and precision of learnable generated representations against standard mel-scaled spectrograms. We will also explore the refinement of speech features through exploiting the signal produced by the vocoder as input to speaker embedding models.Such explorations will not only enrich our comprehension of machine learning in audio analysis, but also potentially contribute to more advanced audio processing technologies in the future.

References

[1] Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., and Khudanpur, S., "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 5329–5333, IEEE, 2018.

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

[3] Zeghidour, N., Teboul, O., Quitry, F. d. C., and Tagliasacchi, M., "LEAF: A learnable frontend for audio classification," *arXiv preprint arXiv:2101.08596*, 2021.

[4] Schlüter, J. and Gutenbrunner, G., "Efficientleaf: A faster learnable audio frontend of questionable use," in *2022 30th European Signal Processing Conference (EUSIPCO)*, pp. 205–208, IEEE, 2022.

[5] Wang, Y., Getreuer, P., Hughes, T., Lyon, R. F., and Saurous, R. A., "Trainable frontend for robust and far-field keyword spotting," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5670–5674, IEEE, 2017.

[6] Kumar, K., Kumar, R., De Boissiere, T., Gestin, L., Teoh, W. Z., Sotelo, J., De Brebisson, A., Bengio, Y., and Courville, A. C., "Melgan: Generative adversarial networks for conditional waveform synthesis," *Advances in neural information processing systems*, 32, 2019.

[7] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y., "Generative adversarial nets," *Advances in neural information processing systems*, 27, 2014.

[8] Donahue, C., McAuley, J., and Puckette, M., "Adversarial audio synthesis," *arXiv preprint arXiv:1802.04208*, 2018.

[9] Engel, J., Resnick, C., Roberts, A., Dieleman, S., Norouzi, M., Eck, D., and Simonyan, K., "Neural audio synthesis of musical notes with wavenet autoencoders," in *International Conference on Machine Learning*, pp. 1068–1077, PMLR, 2017.

[10] Deng, J., Guo, J., Xue, N., and Zafeiriou, S., "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4690–4699, 2019.

[11] Lin, Y., Qin, X., and Li, M., "Cross-Domain ArcFace: Learnging Robust Speaker Representation Under the Far-Field Speaker Verification," *Proc. The 2022 Far-field Speaker Verification Challenge (FFSVC2022)*, pp. 6–9, 2022.

[12] Panayotov, V., Chen, G., Povey, D., and Khudanpur, S., "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 5206–5210, IEEE, 2015.

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

[14] Ko, T., Peddinti, V., Povey, D., Seltzer, M. L., and Khudanpur, S., "A study on data augmentation of reverberant speech for robust speech recognition," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 5220–5224, IEEE, 2017.

[15] Park, D. S., Chan, W., Zhang, Y., Chiu, C.-C., Zoph, B., Cubuk, E. D., and Le, Q. V., "Specaugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, 2019.