A STUDY OF THE ROBUSTNESS OF RAW WAVEFORM BASED SPEAKER EMBEDDINGS UNDER MISMATCHED CONDITIONS

Ge Zhu, Frank Cwitkowitz and Zhiyao Duan

University of Rochester, Rochester, NY, USA
Department of Electrical and Computer Engineering
{ge.zhu,zhiyao.duan}@rochester.edu fcwitkow@ur.rochester.edu

ABSTRACT

In this paper, we conduct a cross-dataset study on parametric and non-parametric raw-waveform based speaker embeddings through speaker verification experiments. In general, we observe a more significant performance degradation of these raw-waveform systems compared to spectral based systems. We then propose two strategies to improve the performance of raw-waveform based systems on cross-dataset tests. The first strategy is to change the real-valued filters into analytic filters to ensure shift-invariance. The second strategy is to apply variational dropout to non-parametric filters to prevent them from overfitting irrelevant nuance features. By combining these strategies, we achieve results comparable to spectral based systems on both the VoxCeleb and VOiCEs datasets. Futhermore, we demonstrate that the learned filters carry little noise compared to existing non-parametric learnable front-ends.

Index Terms— Speaker embedding, filterbank design, raw waveform, robustness, domain mismatch

1. INTRODUCTION

The design and analysis of hand-crafted features has been an active area of research in audio processing. Mel-spectral based features are designed based on human perceptual evidence, but may not necessarily be optimal for extracting speech information. Additionally, there is a possibility of losing important information during the lossy transform. We have also witnessed increasing attention towards data-driven raw waveform-based systems that achieve better performance on phoneme classification [1], source separation [2] and other tasks. Earlier research on sample-level deep neural networks (DNNs) has demonstrated the ability to learn suitable feature embeddings directly from the raw waveform for phone classification [1], music classification [3] and speaker recognition [4]. The performance of these systems is comparable to, and in some cases even surpasses, traditional spectral based methods. On feature interpretability, Tuske *et. al.* [1] showed that DNNs are able to learn bandpass filters purely from the raw waveform without any prior knowledge, and that the first layer can be interpreted as performing a "quasi time-frequency" analysis on audio.

Inspired by these findings, contemporary raw-waveform models typically comprise a modular structure [5, 6, 7, 8, 9, 10]: First, a waveform encoder is used to learn a meaningful representation for audio waveforms and to reduce the dimensionality of feature maps, also referred to as 'wavegrams' [7]. Then, an additional backbone network further processes the wavegram into embeddings. Under this framework, the trainable front-end filterbanks are the key components of raw-waveform based models. Ideally, the filters should only model task-relevant information, while ignoring other nuisance aspects [11]. However, directly learning from densely sampled audio inputs using DNNs without any prior knowledge can lead to over-fitting [12].

There are two main strategies for effectively training a group of meaningful filters from scratch to achieve comparable results to spectral features. These include parametric filterbanks, and non-parametric filterbanks combined with some initialization or regularization strategy. Both filterbank variations are trained together with a respective network architecture. Learnable parametric filterbanks constrain the filters by optimizing only a few parameters, *e.g.*, center frequency and bandwidth [12, 13], of pre-defined parametric functions. With such strong constraints, the learned filters generally follow expected shapes and are easier to interpret. In contrast, non-parametric learnable filterbanks have little to no regularization. In order to mediate this lack of structure, various techniques borrowed from signal processing, such as Gabor initialization [14], multi-scale analysis [10], learnable compression functions [12] and complex convolution [15], are usually applied to the first few layers to avoid overfitting and to speed up convergence.

The performance of raw-waveform based models on cross-domain speech recognition [11, 16] and source separation [2] tasks is known to be more susceptible to mismatch than on in-domain datasets. In this paper, we compare the

This work is partially funded by a gift from Voice Biometrics Group, an NYS Center of Excellence in Data Science award and National Science Foundation grants 1741472 and DGE-1922591.

978-1-6654-0540-9/22/$31.00 ©2022 IEEE

efficacy of raw-waveform speaker embeddings to that of traditional mel-spectrum based methods under different acoustic conditions. We propose several strategies to improve the performance of raw-waveform embeddings on cross-domain tasks, including making use of filter analyticity and variational dropout to learn sparse filter coefficients. Finally, we visualize and analyze the learned filter responses. The complete code for training and inference will also be made available[1].

2. CROSS-DATASET STUDIES

In this section, we present an empirical study comparing several raw-waveform based speaker embeddings with mel-spectrum based models under both matched and mismatched conditions across several speaker verification tasks.

2.1. Datasets

VoxCeleb [17, 18] is a large-scale dataset containing speech spanning a wide range of speakers under uncontrolled acoustic conditions. We use the VoxCeleb2 development partition for training. We also add 100k augmented noisy utterances by adding reverberation, noise, music, and babble to the original speech files following the Kaldi [19] recipe[2]. We use the full VoxCeleb1 dataset, including Vox1-O, Vox1-E and Vox1-H, to perform matched condition tests.

VOiCEs [20], i.e., the Voices Obscured In Complex Environmental Settings corpus, was released with the aim to simulate realistic data under complicated acoustic conditions. It was created by playing Librispeech [21] recordings inside multiple room configurations and re-recording them with 12 different microphones placed at various locations. In addition, pre-recorded background noise plus reverberation or echo were played along with the foreground speech. For evaluating the robustness under mismatched conditions, we used the evaluation partition of this corpus, which consists of 3.6 million trial pairs derived from 11,392 utterances.

2.2. Experimental setup

We select one parametric waveform encoder, SincNet [13], and two non-parametric encoders, multi-scale filters [10] and TDFbank [14], to compare against mel-spectrum based system. All of the speaker embedding systems employ 30 filters of length 400 samples (equivalent to 25ms sampling at 16kHz) with a stride of 5 to extract speech features. Then we feed the output of these three trainable filterbanks to the same backbone network. We model the down-sampling network with sample-level CNN architectures [3, 7, 10]. Specifically, the waveform embeddings output from the learnable filters are first fed into five down-sampling blocks with a decimation rate of 2. Hence, the sequence length of the feature maps is reduced by a factor of 160 in total, equating to 10 ms of hop size. In each down-sampling block, we replace the original dense convolution in [10] with depth-wise separable convolutions, inspired by [8, 22]. In this way, the number of parameters is largely reduced. Finally, speaker embeddings are extracted with time delay neural networks (TDNN).

For the spectral baseline, we use a fixed mel-scaled filterbank, appending the above mentioned backbone network, named ‘x-conv-vector’, for a fair comparison. As a sanity check of the TDNN model’s capability, we also train a vanilla MFCC based model, ‘x-vector (Kaldi)’, and a mel-fbank based model, ‘x-vector’, in PyTorch for reference. In order to eliminate the influence of back-end scoring systems on the final verification results, we simply use cosine similarity for scoring. We also compute the equal error rate (EER) to compare different systems.

2.3. Results and analysis

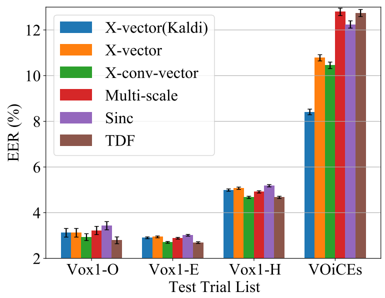


Fig. 1. EER (%) comparison of mel-spectrum based models and raw-waveform based models on different test sets with cosine similarity scoring. Error bars show a 95% confidence interval.

In Fig. 1, we demonstrate EER degradation across datasets for raw-waveform based speaker embeddings. In matched test conditions on VoxCeleb datasets, raw-waveform based speaker embeddings perform on par with the three mel-spectrum based systems. However, in the VOiCEs evaluation dataset, both parametric and non-parametric waveform models lead to degradation compared to spectral based models. It is noted that among all of the six methods, only ‘x-vector (Kaldi)’ performs voice activity detection, making it a less fair baseline. This may also be an important reason for the performance mismatches among mel-spectrum baselines.

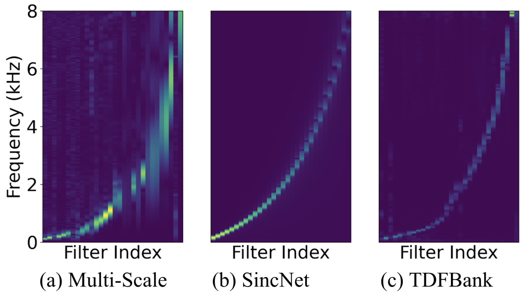
We also visualize learned filter responses after training on the noise augmented VoxCeleb dataset, shown in Fig. 2. We can see that the multi-scale filters and TDFbank have noisier responses compared to SincNet, and the frequency resolution in the higher frequencies of multi-scale filters is worse.

3. ROBUST IMPROVEMENT STRATEGIES

In this section, we propose two strategies and discuss their effect on the robustness of raw-waveform speaker embeddings

[1] https://github.com/gzhu06/TDspkr-mismatch-study

[2] https://github.com/kaldi-asr/kaldi/tree/master/egs/voxceleb/v2



(a) Multi-Scale (b) SincNet (c) TDFBank

Fig. 2. Learned filter responses (normalized by the maximum value for better visualization): (a) multi-scale filters, (2) sinc filters (3) the real part of TDFbank.

under mismatched conditions. Neither strategy introduces additional parameters or computation. The experimental settings and training details are the same as in Sec. 2.2, except we also integrated PLDA scoring for the final comparison of complete speaker verification systems. We adopt the Gaussian PLDA from Kaldi, which was trained on the augmented VoxCeleb-2 training dataset and evaluated on both the VoxCeleb1 and VOiCEs test datasets. Before training, the extracted speaker embeddings were projected onto a 200-dimensional vector with LDA, followed by whitening and length normalization.

3.1. Proposed method

Analytic filterbanks. In the original TDFbank architecture, N real filters and N imaginary filters are initialized into analytic pairs with Gabor wavelets to approximate the mel-filterbanks. Then, a magnitude response is computed using L2 pooling across the output of the real and imaginary pairs. Under the original setting, the weights of the real and imaginary filter components are independently trained without any constraints. As a result, although the initial mel-scale of frequency is mostly preserved after training, the analyticity of the initialization is not preserved. Analytic filters [23] are shift-invariant with respect to time, a desirable property for time-frequency representations. Downsampled convolutions or pooling layers in waveform encoders are not shift-invariant, which compromises their performance on robust classification tasks [24]. A natural way to constrain the analyticity of learned complex filterbanks is to learn only the real component of a filter, and to and infer the imaginary component directly using the Hilbert transform [2, 25]. In this way, the magnitude of the filter response is shift-invariant and the number of filter parameters is essentially halved. Therefore, in this work, we apply the Hilbert transform to obtain the corresponding imaginary filters of real filters. We do this for both the non-parametric and parametric sinc filters.

Sparse variational dropout. Observing the noisy filter responses in Fig. 2, we believe that the non-parametric filters tend to overfit the noisy training data, learning nuisance aspects of the recordings. One way to ease this problem is to regularize the network by dropping irrelevant weights with sparse variational dropout (VD) [26]. VD was originally proposed as a model compression technique to sparsify DNN weights. In this work, we follow our previous work [25] to sparsify filters by applying VD in the first layer of the raw-waveform models.

Dropout can be seen as injecting fixed Bernoulli noise or Gaussian noise into weights during training. Instead of setting a fixed variance as in Gaussian dropout (GD), VD injects an individual multiplicative Gaussian noise $\xi_{ij} \sim N(1, \alpha_{ij})$ to every weight, with the variance α_{ij} consisting of model parameters learned with an approximated KL-divergence measure. By learning an individual variance for every weight, VD is able to induce sparsity across learned weights when $\alpha_{ij} \to \infty$ (equivalent to $p = 1$ in Bernoulli dropout). In such cases, the weights can be ignored or removed from neural networks during inference time.

3.2. Results

Comparison. In this experiment, we evaluate the proposed strategies with the same experimental setup as in Sec. 2. We can see that the 'Multi-scale', 'Sinc' and 'TDF' baselines in Table 1 show more degradation on the VOiCEs test set compared to spectral baselines, which is consistent with the conclusion of Sec. 2. By comparing 'TDF' and 'Sinc' with their corresponding analytic versions, we find that 'Sinc+$\mathcal{H}$' only achieves a marginal improvement over the 'Sinc' baseline on VoxCeleb but a slight degradation on VOiCEs, whereas 'TDF+$\mathcal{H}$' significantly outperforms the 'TDF' baseline on VOiCEs and yields comparable results on VoxCeleb. This shows that the analyticity constraint helps non-parametric filters learn robust representations, but this is not the case for parametric filters. This may be because the benefit of filter analyticity is mainly on learning transient components, which cannot be well modeled in the sinc filters anyway. Comparing 'TDF' and 'TDF+VD', we can also observe a significant improvement on VOiCEs, without compromising the performance on VoxCeleb, with the help of VD. Among all of the raw-waveform based systems, 'TDF+$\mathcal{H}$+VD' achieves the best results on the out-of-domain test set, with both VD and analytic filters helping to boost the performance. Compared with the three spectral based models, it achieves comparable results to 'x-conv-vector' with a similar model size and training strategy. Note that x-vector (Kaldi) achieves similar performance to that reported in [27] on the VOiCEs dataset; this further validates our TDNN backbone implementation.

Ablation study. In order to better demonstrate the effectiveness of each component without the influence of the scoring backend, we conducted several ablation studies using cosine similarity, as shown in Table 2. The improvement of applying analytic filters is consistent with the PLDA backend results in Table 1. Different from results in Table 1, we find that 'TDF+$\mathcal{H}$+VD' outperforms 'TDF+$\mathcal{H}$' when cosine similarity is used. Similarly, 'TDF+$\mathcal{H}$+VD' outperforms 'x-

Table 1. EER (%) comparison on different test sets. All models are trained on the noise augmented VoxCeleb2 training set and scored with PLDA backend. A statistical significance test is performed using a bootstrap procedure: an absolute value of 0.05 of EER difference for Vox1-E and Vox1-H is outside the 95% confidence interval for all methods, while for Vox1-O and VOiCEs the EER difference has to be larger than 0.15 and 0.13 respectively. $\mathcal{H}$ refers to filter analyticity through Hilbert transform.

System	Feature	VoxCeleb-O		VoxCeleb-E		VoxCeleb-H		VOiCEs	
		EER	min-DCF	EER	min-DCF	EER	min-DCF	EER	min-DCF
x-vector (Kaldi)	MFCC	2.26	0.256	2.37	0.279	4.14	0.408	**6.79**	**0.553**
x-vector	Mel-fbank	2.37	0.264	2.42	0.280	4.18	0.406	8.14	0.658
x-conv-vector	Mel-fbank	**2.04**	**0.241**	**2.17**	**0.252**	**3.79**	**0.379**	7.10	0.581
Multi-scale	Waveform	2.28	0.273	2.38	0.285	4.17	0.408	8.54	0.705
Sinc		2.37	0.287	2.32	0.278	4.02	0.400	8.55	0.682
Sinc+$\mathcal{H}$		2.15	0.270	2.28	0.271	3.91	0.396	8.90	0.669
TDF		**1.98**	**0.230**	**2.19**	**0.249**	**3.85**	**0.383**	8.38	0.663
TDF+$\mathcal{H}$		2.01	0.261	2.27	0.263	3.98	0.396	7.46	**0.621**
TDF+VD		1.98	0.235	2.30	0.264	4.05	0.385	7.68	0.626
TDF+$\mathcal{H}$+VD		1.99	0.266	2.26	0.253	3.93	0.385	**7.40**	0.633

Table 2. EER (%) comparison on different test sets. All models are trained on the augmented VoxCeleb2 training set and scored with cosine similarity.

System	Vox1-O	Vox1-E	Vox1-H	Voices
x-vector (Kaldi)	3.12	2.9	4.99	8.41
x-vector	3.12	2.94	5.07	10.78
x-conv-vector	2.93	2.7	**4.67**	10.45
TDF	2.79	**2.69**	4.67	12.74
TDF+VD	3.01	2.79	4.81	11.10
TDF+$\mathcal{H}$	**2.72**	2.81	4.86	10.72
TDF+$\mathcal{H}$+BD	3.06	2.77	4.83	11.69
TDF+$\mathcal{H}$+GD	2.98	2.73	4.83	11.29
TDF+$\mathcal{H}$+VD	**2.72**	2.72	4.72	**10.32**

conv-vector' slightly on VOiCEs. These differences suggest that by dropping filter weights through VD, the final learned speaker embeddings tend to become less Gaussian, hence yield worse results with the PLDA backend. We also experimented with different dropout techniques, the results shown in the last four rows in Table 2. we can observe that BD and GD are not helpful in improving robustness compared to 'TDF+$\mathcal{H}$' baseline, while VD achieves better verification results in all of the in-domain and out-of-domain tasks.

Filter visualization. In Fig. 3, we visualize several learned non-parametric filters at different frequency bands under different training settings for TDF based methods. When trained on the noisy dataset, the learned filters are less regular and contain more noise-like shapes than filters trained on the clean dataset. With the help of VD, the learned filter at 345Hz is similar to the one trained without noise, and only the center weights of the filters at 2258Hz and 7937Hz are retained. The nuisance picked up from the noise are not present in the filters. Although there is no significant improvement on EER over the baseline with VD, this verifies that during training, raw waveform models tend to capture nuisance information from noisy data, and proves that dropping out the corresponding weights does not affect the final performance.

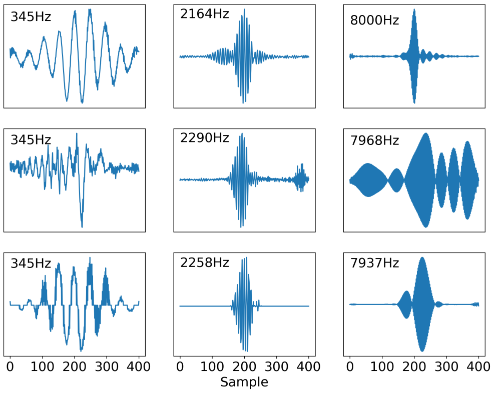


Fig. 3. Examples of learned filters with their maximum response frequency labeled. Top row: 'TDF+$\mathcal{H}$' filters trained on clean VoxCeleb. Middle row: 'TDF+$\mathcal{H}$' filters trained on noise augmented VoxCeleb. Bottom row: 'TDF+$\mathcal{H}$+VD' filters trained on noise augmented VoxCeleb.

4. CONCLUSION

In this paper, we performed a systematic empirical study of multiple parametric and non-parametric raw-waveform based speaker embeddings. Compared to mel-spectrum baselines, these raw-waveform based methods yield similar results on in-domain tests, but show a more significant degradation on cross-domain tests. In order to bridge this performance gap, we proposed to apply filter analyticity to promote shift-invariance of the learned filters and variational dropout on non-parametric filters to discard task irrelevant information during training. Finally, we observed a significant improvement for non-parametric raw-waveform based embeddings with respect to cosine similarity and PLDA backends, achieving similar performance to the mel-spectrum baselines. Future work includes investigating how shift-invariance or analytic filters influence learned representations and whether more general stochastic layers will prevent over-fitting.

5. REFERENCES

[1] Zoltán Tüske, Pavel Golik, Ralf Schlüter, and Hermann Ney, "Acoustic modeling with deep neural networks using raw time signal for lvcsr," in *Fifteenth annual conference of the international speech communication association*, 2014.

[2] Manuel Pariente, Samuele Cornell, Antoine Deleforge, and Emmanuel Vincent, "Filterbank design for end-to-end speech separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6364–6368.

[3] Taejun Kim, Jongpil Lee, and Juhan Nam, "Sample-level cnn architectures for music auto-tagging using raw waveforms," in *IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 366–370.

[4] Hannah Muckenhirn, Mathew Magimai Doss, and Sébastien Marcell, "Towards directly modeling raw speech signal for speaker verification using cnns," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4884–4888.

[5] Jee-weon Jung, Hee-soo Heo, ju-ho Kim, Hye-jin Shim, and Ha-jin Yu, "Rawnet: Advanced end-to-end deep neural network using raw waveforms for text-independent speaker verification," *Proc. Interspeech*, pp. 1268–1272, 2019.

[6] Jee weon Jung, Seung bin Kim, Hye jin Shim, Ju ho Kim, and Ha-Jin Yu, "Improved RawNet with Feature Map Scaling for Text-Independent Speaker Verification Using Raw Waveforms," in *Proc. Interspeech*, 2020, pp. 1496–1500.

[7] Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D Plumbley, "Panns: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2880–2894, 2020.

[8] Sangwook Han, Jaeuk Byun, and Jong Won Shin, "Time-domain speaker verification using temporal convolutional networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6688–6692.

[9] Wei-Wei Lin and Man-Wai Mak, "Wav2spk: A simple dnn architecture for learning speaker embeddings from waveforms.," in *Proc. Interspeech*, 2020, pp. 3211–3215.

[10] Ge Zhu, Fei Jiang, and Zhiyao Duan, "Y-Vector: Multiscale Waveform Encoder for Speaker Embedding," in *Proc. Interspeech*, 2021, pp. 96–100.

[11] Erfan Loweimi, Peter Bell, and Steve Renals, "On the robustness and training dynamics of raw waveform models.," in *Proc. Interspeech*, 2020, pp. 1001–1005.

[12] Neil Zeghidour, Olivier Teboul, Félix de Chaumont Quitry, and Marco Tagliasacchi, "Leaf: A learnable frontend for audio classification," *ICLR*, 2021.

[13] Mirco Ravanelli and Yoshua Bengio, "Speaker recognition from raw waveform with sincnet," in *IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 1021–1028.

[14] Neil Zeghidour, Nicolas Usunier, Iasonas Kokkinos, Thomas Schaiz, Gabriel Synnaeve, and Emmanuel Dupoux, "Learning filterbanks from raw speech for phone recognition," in *IEEE international conference on acoustics, speech and signal Processing (ICASSP)*. IEEE, 2018, pp. 5509–5513.

[15] Junyi Peng, Xiaoyang Qu, Jianzong Wang, Rongzhi Gu, Jing Xiao, Lukáš Burget, and Jan Černockỳ, "Icspk: Interpretable complex speaker embedding extractor from raw waveform," *Proc. Interspeech*, pp. 511–515, 2021.

[16] Purvi Agrawal and Sriram Ganapathy, "Interpretable representation learning for speech and audio signals based on relevance weighting," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2823–2836, 2020.

[17] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman, "VoxCeleb: A Large-Scale Speaker Identification Dataset," in *Proc. Interspeech*, 2017, pp. 2616–2620.

[18] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman, "VoxCeleb2: Deep Speaker Recognition," in *Proc. Interspeech 2018*, 2018, pp. 1086–1090.

[19] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely, "The Kaldi speech recognition toolkit," in *IEEE Workshop on Automatic Speech Recognition and Understanding*. Dec. 2011, IEEE Signal Processing Society.

[20] Colleen Richey, Maria A Barrios, Zeb Armstrong, Chris Bartels, Horacio Franco, Martin Graciarena, Aaron Lawson, Mahesh Kumar Nandwana, Allen Stauffer, Julien van Hout, et al., "Voices obscured in complex environmental settings (voices) corpus," *arXiv preprint arXiv:1804.05053*, 2018.

[21] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.

[22] Yi Luo and Nima Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

[23] JL Flanagan, "Parametric coding of speech spectra," *The Journal of the Acoustical Society of America*, vol. 68, no. 2, pp. 412–419, 1980.

[24] Richard Zhang, "Making convolutional networks shift-invariant again," in *International conference on machine learning*. PMLR, 2019, pp. 7324–7334.

[25] Frank Cwitkowitz, Mojtaba Heydari, and Zhiyao Duan, "Learning sparse analytic filters for piano transcription," *arXiv preprint arXiv:2108.10382*, 2021.

[26] Dmitry Molchanov, Arsenii Ashukha, and Dmitry Vetrov, "Variational dropout sparsifies deep neural networks," in *International Conference on Machine Learning*. PMLR, 2017, pp. 2498–2507.

[27] Weiwei Lin, Man-Wai Mak, and Lu Yi, "Learning mixture representation for deep speaker embedding using attention," in *Proc. Odyssey 2020 The Speaker and Language Recognition Workshop*, 2020, pp. 210–214.