Optimizing Multi-Taper Features for Deep Speaker Verification

Xuechen Liu, Md Sahidullah, Tomi Kinnunen

Optimizing Multi-Taper Features for Deep Speaker Verification

Xuechen Liu, Md Sahidullah, *Member, IEEE*, and Tomi Kinnunen, *Member, IEEE*

***Abstract*—Multi-taper estimators provide low-variance power spectrum estimates that can be used in place of the windowed discrete Fourier transform (DFT) to extract speech features such as mel-frequency cepstral coefficients (MFCCs). Even if past work has reported promising automatic speaker verification (ASV) results with Gaussian mixture model-based classifiers, the performance of multi-taper MFCCs with deep ASV systems remains an open question. Instead of a static-taper design, we propose to optimize the multi-taper estimator jointly with a deep neural network trained for ASV tasks. With a maximum improvement on the SITW corpus of 25.8% in terms of equal error rate over the static-taper, our method helps preserve a balanced level of leakage and variance, providing more robustness.**

***Index Terms*—Multi-taper spectrum, speaker verification**

I. Introduction

FEATURE extractor is a critical component of speech processing systems. It converts a raw waveform into features that feed task-specific models. In many tasks, including automatic speaker verification (ASV) [1], the most widely-used features are the *mel-frequency cepstral coefficients* (MFCCs) computed from a short-term spectral representation—usually, the windowed *discrete Fourier transform* (DFT) [2].

While MFCCs perform well under matched conditions, they lack robustness to data variations. Various methods are available to improve their robustness, such as feature normalization [3], [4] applied *after* MFCC extraction. In addition, the MFCC extractor itself can be improved. In [5], a spectral estimator based on multiple windows was used in place of the single-window DFT. Such *multi-taper* spectrum estimator [6] addresses a specific shortcoming — high variance. Here, a single frame of speech is viewed as a realization of a stationary stochastic process, and 'variance' refers to the variation in power-spectral estimates. Given the ubiquitous role of the power spectrum in different speech front-ends, multi-tapers can also be used to enhance other features, such as *perceptual linear predictive* (PLP) features [7], [8].

A multi-taper power spectrum estimator is simply a weighted average of many windowed DFT power spectrum estimates (*sub-spectra*) obtained with carefully designed window functions (*tapers*) and their associated weights. While there are different approaches to design optimal multi-tapers [6], [9], they typically rely upon assumptions of the stochastic process, rather than the downstream application. In ASV, for instance, we do not extract identity cues from a *single* speech frame — the domain of the short-term power spectrum — but multiple frames, i.e. an utterance. Thus, while application-independent multi-taper design is the standard one, it is plausible that the assumed stochastic process properties are not compatible with the given downstream task or classifier.

Experiments with *Gaussian mixture model* (GMM) based classifiers in [7], [8], [10] indicate that the multi-taper spectrum estimator is a simple yet effective method to improve ASV accuracy. Nonetheless, the community has recently shifted its focus to *deep neural networks* (DNNs; for a comprehensive survey, refer to [11]). This raises a question whether the earlier positive findings can be generalized to modern ASV models, which motivated the present work.

One crucial difference between GMMs and DNNs lies in their ability to model larger temporal contexts. GMMs cannot handle high-dimensional data, as it would require additional dimensionality reduction, diagonal covariances, or limiting the number of frames that can be used for feature extraction. Many DNN architectures (e.g. models with recurrent, dilated convolutional or time-delay layers), however, can cope with an extended temporal context without issues. A successful example is the *time-delay neural network* (TDNN) architecture [12] used in *x-vector* model [13], whose variations and extensions [14], [15] currently form the standard ASV baselines.

The spectrum variance reduction provided by traditional multi-tapers on individual frames might not perfectly combine with a TDNN. We hypothesize that better spectral features could be obtained by optimizing the multi-taper estimator *for an ASV task directly*. While in GMM-based approaches 'features' and 'classifiers' are often treated separately, DNNs enable their joint optimization. Although this is the overall sentiment in *end-to-end* learning [16]–[19], our starting points are DFT-based spectral representation and MFCCs rather than the raw waveform. Therefore, our feature extractor design retains the familiar processing elements and enables one-to-one comparisons with conventional MFCCs obtained either from a single-window DFT, or hand-crafted static multi-taper spectrum estimators. Besides quantitative ASV evaluation on three different datasets, we investigate the spectral and statistical properties of the learned multi-taper estimator.

II. Optimized Multi-taper Spectral Estimator

A. Multi-taper

Let $\boldsymbol{x} = [x(0), \ldots, x(N-1)]$ to denote a short speech frame of N samples. The windowed DFT [2] is defined by

$$\hat{S}(f) = \left| \sum_{t=0}^{N-1} w(t)x(t)e^{-i2\pi tf/N} \right|^2, \quad (1)$$

This project was partially funded by Academy of Finland (project 309629) and Inria Nancy Grand Est.

X. Liu and M. Sahidullah are with Université de Lorraine, CNRS, Inria, LORIA, F-54000, Nancy, France (e-mail: xuechen.liu@inria.fr, md.sahidullah@inria.fr).

X. Liu and T. Kinnunen are with School of Computing, University of Eastern Finland, FI-80101, Joensuu, Finland (e-mail: tomi.kinnunen@uef.fi).

where $\hat{S}(f)$ denotes the estimated power spectrum and $f = 0, \ldots, N-1$ is the DFT frequency bin. Additionally, $w(t)$ is the window function (taper) — in this work, the *Hamming* window with $w(t) = 0.54 - 0.46\cos(2\pi t/N)$ for $0 \leq t < N$ (and $w(t) = 0$ for other t). The primary purpose of the window is to reduce *spectral leakage* compared to the rectangular window, also known as 'no windowing'. Nonetheless, the variance of the spectrum estimates provided by (1) remains high. The multi-taper spectral estimator [6] aims at reducing the variance through multiple, weighted DFT power spectrum estimates:

$$\hat{S}(f) = \sum_{j=1}^{K} \lambda(j) \left| \sum_{t=0}^{N-1} w_j(t) x(t) e^{-i2\pi t f/N} \right|^2, \qquad (2)$$

where K is the number of tapers, $j = 1, \ldots, K$ denotes the taper index and $\boldsymbol{w}_j = [w_j(0), \ldots, w_j(N-1)]$ represents j^{th} taper, with its associated weight $\lambda(j) > 0$. Windowed DFT in (1) is obtained as a special case with $K = 1$, $\lambda = 1$ and $\boldsymbol{w}_1$ set as the Hamming window.

Averaging reduces variance by making the resulting spectrum less susceptible to small within-frame data variation. After a suitable set of tapers has been selected, their number (K) can be selected to trade-off between two conflicting criteria of variance reduction (high K) and bias reduction (low K). A high value of K implies a 'rigid' spectrum that smears detail but provides a (statistically) stable representation; a low K, in turn, retains more detail but is susceptible to perturbations. The choice of K typically requires some experimentation in a given downstream task.

In previous work, the tapers and their weights were set in a hand-crafted manner. A detailed study in GMM-based speaker recognition was conducted in [5], where three different types of taper designs were considered. Based on preliminary ASV experiments, in this work, we focus on *sine-weighted cepstral estimator* (SWCE) [9], where both the tapers and their weights are provided by closed-form equations:

$$\begin{aligned} w_j(t) &= \sqrt{2/(N+1)} \sin\left[2\pi j t/(N+1)\right] \\ \lambda(j) &= \sin\left[2\pi j/(N+1)\right] / \sum_{k=0}^{K} \sin\left[2\pi k/(N+1)\right]. \end{aligned} \qquad (3)$$

Readers may refer to [6] for further details on multi-tapers, which are out of our scope. In general, the tapers and their weights are designed to uncorrelate the estimation errors between the sub-spectra. As noted above, however, the tapers are designed for short-term stationary signals (frames). Hence, such design ignores both the temporal context and interaction with the downstream model, which is DNN-based ASV here.

B. Data-driven multi-taper

As explained above, we attempted to learn the multi-taper spectrum estimator jointly with a downstream system — specifically, a DNN-based speaker embedding extractor. To this end, we treated the taper weights $\boldsymbol{\lambda} = \lambda(j), j = 1, ..., K$ as part of the neural network parameters. They were updated jointly with the TDNN parameters, denoted by $\boldsymbol{W}$. As an example, using first-order gradient descent, the model parameters are updated by,

$$\begin{aligned} \boldsymbol{W}_{nk} &\leftarrow \boldsymbol{W}_{nk} - \eta_n * \frac{\partial J_{\text{loss}}}{\partial \boldsymbol{W}_{nk}} \\ \boldsymbol{\lambda}_{nk} &\leftarrow \boldsymbol{\lambda}_{nk} - \eta_n * \frac{\partial J_{\text{loss}}}{\partial \boldsymbol{\lambda}_{nk}}, \end{aligned} \qquad (4)$$

where J_{loss} is the objective function of the neural network (here, additive angular softmax [20] between the output of the network and speaker labels), and η denotes the learning rate; n and k are epoch and iteration indices, respectively. The gradient with respect to network and taper weights is computed based on chain rule. In this study, we used Adam optimizer [21].

We optimized only for the taper *weights*, while the tapers $\boldsymbol{w}_j$ remained static, which allowed the weights to be treated as 'leaf' scalar nodes in the computational graph during optimization, making optimization efficient without introducing additional training parameters and neural layers. Moreover, it avoided introducing an excessive number of additional taper parameters ($K \cdot N$, where N is the number of spectral bins), which could have made the joint learning challenging in terms of finding an optimal solution. Our preliminary ASV experiments with learnable taper functions (represented by the vectors $\boldsymbol{w}_j$) indicated less promising result; hence, this direction was not pursued further.

In addition to the choice of the number of tapers K, other important design considerations include weight initialization and their non-negativity constraints. Following [22], we considered two initialization approaches. The first used random weights from a standard normal distribution, and the latter initialized the weights using (3). Note that since (2) is a power spectrum estimator, we require $\lambda(j) > 0$ as a constraint. Inspired by previous works on different types of tapers [8], we additionally constrained the sum of weights to unity ($\sum_{j=1}^{K} \lambda(j) = 1$). To this end, the updated weights are processed with rectified linear unit (ReLU) [23] activation function to enforce positivity, followed by length normalization $\boldsymbol{\lambda} \leftarrow \boldsymbol{\lambda}/\|\boldsymbol{\lambda}\|_1$ to enforce the latter constraint. The choice of ℓ_1 norm is inspired in part by sparsity considerations. We refer to such constraint by simply noting the ReLU function in the following sections.

III. Experimental Setup

Data. We reported ASV performance on three different datasets. The first one was *Voxceleb1-test*, a 40-speaker *test* partition from Voxceleb1 following [24]. We also included *eval* partition of *speakers in the wild* (SITW) corpus [25], under core-core condition (*SITW-EVAL*) and logical access (LA) scenario of ASVspoof 2019 [26] with only bonafide imposters from its ASV protocol (*ASVSpoof2019-LA*). The three datasets have diverse qualities; *ASVSpoof2019-LA* has the lowest acoustic mismatch between the enrollment and the test data due to its relatively clean and highly-controlled recording conditions. The other two datasets are more similar to one another, as both contain audio extracted from videos.

TABLE I: Speaker verification results on different evaluation sets.

					Voxceleb1-test		SITW-EVAL		ASVSpoof2019-LA	
Type	Taper	Num. tapers	Weight init.	Weight constraint	EER(%)	MinDCF	EER(%)	MinDCF	EER(%)	MinDCF
Static	DFT	-	-	-	2.01	0.2344	2.93	0.2901	1.52	0.149
	SWCE	8	-	-	2.12	0.2663	3.72	0.3214	1.32	0.1546
Data-driven	SWCE	8	Gaussian	None	2.23	0.2364	3.19	0.3321	1.60	**0.1405**
		8	Gaussian	*ReLU*	**1.96**	0.2473	3.11	0.2939	1.38	0.1500
		20	Gaussian	*ReLU*	2.01	0.2503	2.95	0.2864	1.34	0.2403
		8	SWCE	None	**1.96**	**0.2209**	**2.78**	**0.2862**	**1.20**	0.1570
		8	SWCE	*ReLU*	2.12	0.2596	2.87	0.2932	1.50	0.1485
		20	SWCE	*ReLU*	2.33	0.3213	3.74	0.3583	1.45	0.1561
		2	Gaussian	None	1.98	0.2497	3.21	0.3003	1.34	0.1325
		2	SWCE	None	**1.95**	0.2559	3.06	0.2969	1.42	0.1377

TABLE II: Spectral analysis statistics on input synthetic signal for different multi-taper estimators. The number of tapers for systems covered in this table is 8.

Type	Taper	Weight init./constraint	Distance.	Width(Hz)
Static	DFT	-	0.21	39.26
	SWCE	-	1.04	314.15
Data-driven	SWCE	Gaussian/None	0.31	196.35
		Gaussian/*ReLU*	0.29	196.35
		SWCE/None	0.36	196.35
		SWCE/*ReLU*	0.34	235.62

The speaker embedding extractors were trained on the VoxCeleb. We used the *dev* partitions [24], [27] consisting of 7205 speakers after removing speakers overlapped with SITW.

Feature extractors. MFCCs obtained with conventional Hamming window-based DFT and with SWCE-based multi-taper spectrum estimator [9] formed our baselines. For the proposed data-driven multi-taper MFCCs, we contrasted the two initialization methods explained above. We also addressed the impact of the proposed ReLU update rule and noted the results for the selected number of tapers. In all cases, the acoustic features that feed the neural network were 40 MFCCs computed with the same number of the mel filters.

Speaker embedding extractor. *Extended x-vector* based on *time-delayed neural network* (E-TDNN) [14] served as the speaker embedding extractor, which showed promising performance over the basic x-vector model [13]. In addition, we replaced the conventional statistics pooling in the original network with attentive statistics pooling [28] and trained the network using the *additive angular margin softmax* loss function [20]. We extracted a 512-dimensional speaker embedding from the first fully-connected layer after statistics pooling for each input utterance.

Classifier back-end. For each corpus, we trained a probabilistic linear discriminant (PLDA) back-end classifier using speaker embeddings from the trained speaker embedding extractor. The extracted embeddings were length-normalized and centered using a 200-dimensional linear discriminant analyzer (LDA) prior to the PLDA estimation. The subspace size of PLDA was the same as that of LDA.

Metrics. We evaluated the speaker verification performance with *equal error rate* (EER) and minimum detection cost function (minDCF). Target speaker prior for minDCF was $p_{\text{tar}} = 0.01$. *Detection error trade-off* (DET) curves on SITW-EVAL were also provided. We used Kaldi [29] to compute EERs and minDCFs and BOSARIS [30] to draw DET plots.

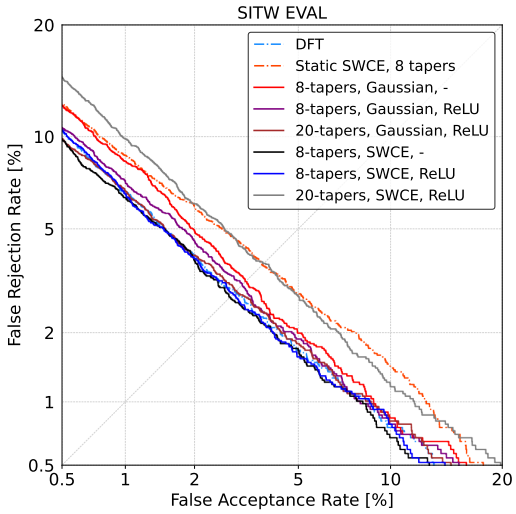

Fig. 1: DET curves on SITW-EVAL. Proposed spectral estimators are marked with (num. tapers, weight init., weight constraint) from Table I.

IV. Experimental Results

A. Speaker Verification Results

Table I shows the results for the baseline MFCCs, static multi-taper MFCCs, and the proposed data-driven variants. Similar to our previous findings in [31] using a smaller-scale training set, static SWCE with eight tapers did not outperform baseline MFCCs on the *Voxceleb1-test*. However, it yielded slightly lower EER on *ASVSpoof2019-LA* than the baseline, which indicated its potentiality in test conditions with a lower mismatch between enrollment and test.

For condition, namely *Voxceleb1-test*, data-driven multi-taper systems did not show a substantial advantage over the Hamming window but demonstrated consistent improvement over static multi-taper in most cases. SWCE weight initialization yielded the lowest EERs and minDCFs. For *ASVSpoof2019-LA*, best system outperformed static SWCE by relatively 9.1% on EER and 5.7% on minDCF. The data-driven taper exhibited a noticeable performance gain on *SITW-EVAL* with a 25.8% relative EER improvement compared to

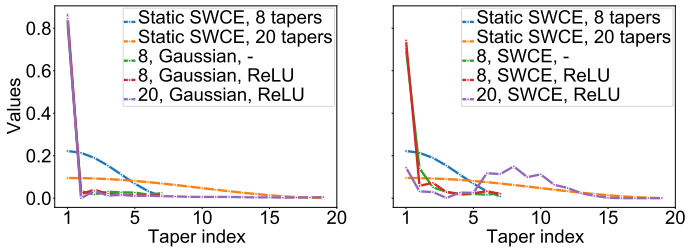

Fig. 2: Weights of different static and data-driven multi-taper estimators. Proposed spectral estimators are marked with (num. tapers, weight init., weight constraint) from Table I.

static SWCE, which indicate that data-driven approaches can improve robustness.

Adding ReLU generally improved the ASV performance of systems with Gaussian weight initialization, which was not the case for SWCE kernel initialization, where such addition degraded performance in all categories except for minDCF for *ASVSpoof2019-LA*. Our proposed updating approach might over-fit the weights since kernel initialization already sets a proper starting point for learning. This can be observed for both $K = 8$ and $K = 20$. Moreover, we found from pilot experiments that using larger number of tapers (e.g. 32) degraded the performance[1], which may due to over-smoothed spectrum. With increased computational time, we limited our experiments to maximum of $K = 20$ tapers.

The above findings are in line with the DET curve (Fig. 1) in most operating points. The advantage of learned 8-taper systems compared to static ones is apparent in regions with less constraint on false alarms, especially with kernel initialization.

B. Analysis

Two analytic studies appeared to be of interest: 1) a study on learned taper weight values, compared to hand-crafted settings; 2) a study on spectral leakage. The former can give an alternative design choice for multi-taper estimators, while the latter can bridge the statistical importance of the estimators and deep ASV performance.

Learned taper weights. Figure 2 shows the weight values of different learned spectral estimators, including the baseline static SWCE. Among all 8-taper learned multi-taper estimators, weight values are heavily concentrated on the first two tapers, with the remaining weights being close to zero. In light of the DNN-based ASV results, a lower number of tapers may be better for robustness. To further validate this hypothesis, we conducted two additional experiments with two tapers only with Gaussian and SWCE initialization, respectively. The results of those two experiments are shown at the end of Table I. They show that using such a low number of tapers does not exacerbate ASV performance. In fact, the one with SWCE weight initialization reached the lowest EER on *Voxceleb1-test*, outperforming static SWCE by 8.0%.

Spectral leakage. To measure relative leakage experimentally, we generated synthetic signal, by $s_n(t) = \sin(2\pi n f_s t / N_{\text{FFT}})$, where f_s is sampling rate (16 kHz here),

[1]EER and minDCF are 3.56%/0.4091 for Gaussian/*ReLU* and 2.82%/0.3443 for SWCE/*ReLU* on VoxCeleb1-test for $K = 32$.

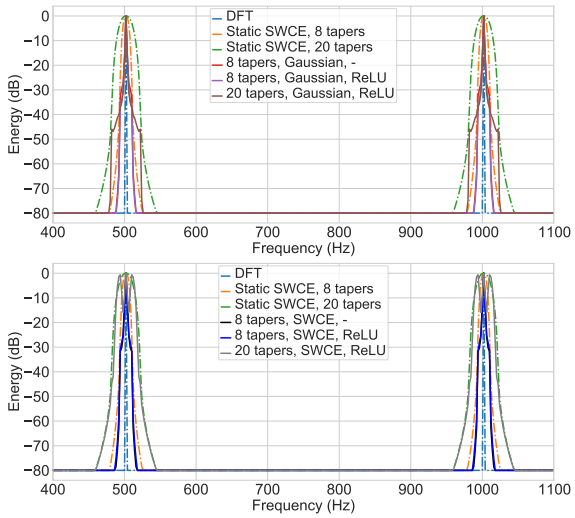

Fig. 3: Spectral representation from sinusoids. Proposed spectral estimators are marked with (num. tapers, weight init., weight constraint) from Table I.

n is the frequency bin index that made to unit amplitude of spectral energy, sin denotes sinusoidal function and N_{FFT} is the number of FFT bins (512 here). The final signal is the sum of $s_n(t)$ of different frequencies. We measured the performance of different estimators by two metrics. The first one was the spectral difference from ground truth measured through the Itakura-Saito distance [32]. Second, we defined and measured the attenuation width where the spectral energies were sufficiently low. This is expressed as $w = (n_{\text{right}} - n_{\text{left}}) * f_s / N_{\text{FFT}}$ in Hz, where n_{right} and n_{left} are edge points where the spectral power is 80 dB below unity gain. We chose 500 Hz and 1 kHz as center frequencies, referring to the average values of the first two formant frequencies [33]. Spectra returned by all estimators including two baselines are visualized in Fig. 3 and Table II shows the leakage statistics of the static and the 8-taper systems with respect to ground truth. The leakage indicated in the figure shows that the better-performed data-driven systems return a certain level of leakage that lies between their static counterparts and ground truth. Numbers returned by best-performing systems lies in between static SWCE and ground truth representation, indicating that while lower spectral leakage is expected, a trade-off between certain levels of leakage and perturbation must occur in order to reach better ASV performance.

V. Conclusion

We re-evaluated static multi-taper spectral estimator for speaker verification with DNN-based speaker embedding extractor and proposed optimization schemes, enabling joint learning of taper weight values with the DNN. We then investigated the effect of kernel initialization using the static counterparts. The proposed optimized multi-taper features show promising speaker verification performance and a high level of robustness on varieties of speech corpora. Further analysis shows that the learned multi-taper implicitly maintains a decent trade-off between spectral leakage and variance, corresponding to an improved ASV performance.

REFERENCES

[1] J. H.L. Hansen and T. Hasan, "Speaker recognition by machines and humans: A tutorial review," *IEEE Signal Processing Magazine*, vol. 32, no. 6, pp. 74–99, 2015.

[2] F.J. Harris, "On the use of windows for harmonic analysis with the discrete Fourier transform," *Proceedings of the IEEE*, vol. 66, no. 1, pp. 51–83, 1978.

[3] X. Huang, A. Acero, H. Hon, and R. Reddy, *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*, Prentice Hall PTR, USA, 1st edition, 2001.

[4] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 578–589, 1994.

[5] T. Kinnunen, R. Saeidi, F. Sedlak, K. A. Lee, J. Sandberg, M. Hansson-Sandsten, and H. Li, "Low-variance multitaper MFCC features: A case study in robust speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 7, pp. 1990–2001, 2012.

[6] D.J. Thomson, "Spectrum estimation and harmonic analysis," *Proceedings of the IEEE*, vol. 70, no. 9, pp. 1055–1096, 1982.

[7] M. J. Alam, T. Kinnunen, P. Kenny, P. Ouellet, and D. O'Shaughnessy, "Multitaper MFCC and PLP features for speaker verification using i-vectors," *Speech Communication*, vol. 55, no. 2, pp. 237–251, 2013.

[8] M. Alam, P. Kenny, and D. O'Shaughnessy, "Low-variance multitaper mel-frequency cepstral coefficient features for speech and speaker recognition systems," *Cognitive Computation*, vol. 5, 12 2013.

[9] M. Hansson-Sandsten and J. Sandberg, "Optimal cepstrum estimation using multiple windows," in *Proc. ICASSP*, 2009, pp. 3077–3080.

[10] M. J. Alam, P. Kenny, P. Dumouchel, and D. O'Shaughnessy, "Robust feature extractors for continuous speech recognition," in *2014 22nd European Signal Processing Conference (EUSIPCO)*, 2014, pp. 944–948.

[11] Z. Bai and X. Zhang, "Speaker recognition based on deep learning: An overview," *Neural Networks*, vol. 140, pp. 65–99, 2021.

[12] V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *Proc. INTERSPEECH*, 2015.

[13] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *Proc. ICASSP*, 2018, pp. 5329–5333.

[14] D. Snyder, D. Garcia-Romero, G. Sell, A. McCree, D. Povey, and S. Khudanpur, "Speaker recognition for multi-speaker conversations using x-vectors," in *Proc. ICASSP*, 2019, pp. 5796–5800.

[15] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification," in *Proc. INTERSPEECH*, 2020, pp. 3830–3834.

[16] J. Jung, H. Heo, I. Yang, S. Yoon, H. Shim, and H. Yu, "D-vector based speaker verification system using raw waveform CNN," in *Proceedings of the 2017 International Seminar on Artificial Intelligence, Networking and Information Technology (ANIT 2017)*. 2017/12, pp. 126–131, Atlantis Press.

[17] J. Jung, H. Heo, I. Yang, H. Shim, and H. Yu, "A complete end-to-end speaker verification system using deep neural networks: From raw signals to verification result," in *Proc. ICASSP*, 2018, pp. 5349–5353.

[18] H. Muckenhirn, M. Magimai-Doss, and S. Marcell, "Towards directly modeling raw speech signal for speaker verification using CNNs," in *Proc. ICASSP*, 2018, pp. 4884–4888.

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

[20] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 4685–4694.

[21] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR (Poster)*, 2015.

[22] X. Liu, M. Sahidullah, and T. Kinnunen, "Learnable MFCCs for speaker verification," in *2021 IEEE International Symposium on Circuits and Systems (ISCAS)*, 2021, pp. 1–5.

[23] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th International Conference on International Conference on Machine Learning*, Madison, WI, USA, 2010, ICML'10, p. 807–814, Omnipress.

[24] A. Nagrani, J. Chung, and A. Zisserman, "VoxCeleb: A large-scale speaker identification dataset," in *Proc. INTERSPEECH*, 2017, pp. 2616–2620.

[25] M. McLaren, L. Ferrer, Diego Castán L., and A. Lawson, "The speakers in the wild (SITW) speaker recognition database," in *Proc. INTERSPEECH*, 2016, pp. 818–822.

[26] A. Nautsch, X. Wang, N. Evans, T. H. Kinnunen, V. Vestman, M. Todisco, H. Delgado, M. Sahidullah, J. Yamagishi, and K. A. Lee, "ASVspoof 2019: Spoofing countermeasures for the detection of synthesized, converted and replayed speech," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 3, no. 2, pp. 252–265, 2021.

[27] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep speaker recognition," in *Proc. INTERSPEECH*, 2018.

[28] K. Okabe, T. Koshinaka, and K. Shinoda, "Attentive statistics pooling for deep speaker embedding," in *Proc. INTERSPEECH*, 2018, pp. 2252–2256.

[29] D. Povey et al, "The kaldi speech recognition toolkit," in *Proc. ASRU*. 2011, IEEE Signal Processing Society.

[30] N. Brümmer and E. Villiers, "The BOSARIS toolkit: Theory, algorithms and code for surviving the new DCF," *ArXiv*, vol. abs/1304.2865, 2013.

[31] X. Liu, M. Sahidullah, and T. Kinnunen, "A comparative re-assessment of feature extractors for deep speaker embeddings," in *Proc. INTERSPEECH*, 2020, pp. 3221–3225.

[32] F Itakura and S Saito, "Analysis synthesis telephony based on the maximum likelihood method," in *Proc. 6th of the International Congress on Acoustics*, 1968, pp. C17–C20.

[33] J.C. Catford, *A Practical Introduction to Phonetics*, Oxford University Press, New York, 1988.