Learnable Nonlinear Compression for Robust Speaker Verification

Xuechen Liu, Md Sahidullah, Tomi Kinnunen

LEARNABLE NONLINEAR COMPRESSION FOR ROBUST SPEAKER VERIFICATION

Xuechen Liu[1,2]*, Md Sahidullah*[2]*, Tomi Kinnunen*[1]

[1]School of Computing, University of Eastern Finland, Joensuu, Finland
[2]Université de Lorraine, CNRS, Inria, LORIA, F-54000, Nancy, France

ABSTRACT

In this study, we focus on nonlinear compression methods in spectral features for speaker verification based on deep neural network. We consider different kinds of channel-dependent (CD) nonlinear compression methods optimized in a data-driven manner. Our methods are based on power nonlinearities and dynamic range compression (DRC). We also propose multi-regime (MR) design on the nonlinearities, at improving robustness. Results on VoxCeleb1 and VoxMovies data demonstrate improvements brought by proposed compression methods over both the commonly-used logarithm and their static counterparts, especially for ones based on power function. While CD generalization improves performance on VoxCeleb1, MR provides more robustness on VoxMovies, with a maximum relative equal error rate reduction of 21.6%.

Index Terms— Speaker Verification, Nonlinear Compression, Multi-Regime Compression.

1. INTRODUCTION

Automatic speaker verification (ASV) [1], [2] is the task of verifying a person's identity using his or her voice. Modern ASV systems consists of three main components: acoustic feature extractor, speaker embedding extractor, and back-end classifier. In recent years, substantial improvement has been achieved by using *deep neural networks* (DNNs) to implement, especially the last two components. Concerning speaker embedding extractor, statistical models such as *i-vectors* [3] have been replaced by deep models such as *x-vector* with *time-delayed neural network (TDNN)* [4]. As for the back-end, recent studies have replaced *probabilistic linear discriminant analysis* (PLDA) [5] with neural approaches [6].

Concerning features, however, many ASV systems still use *mel-frequency cepstral coefficients* (MFCCs) [7], which are not specialized for ASV and neglects information such as phase and temporal characteristics [8]–[11]. Meanwhile, spectrograms are also widely used [12]–[15]. There are multiple types of spectrograms such as raw one where no filter is applied [12], [16] and more widely-used spectral energies output with mel filters [17], [18]. Even if the spectral representations are usually higher-dimensional (hence, more expressive) than MFCCs, problems of lacking specialization and missing information remain.

There are also attempts to replace hand-crafted features with neural networks [19]–[22]. However, such design may be hard to interpret. Moreover, many state-of-the-art DNN extractors are based on convolutional kernels, whose modeling capability on variabilities across different frequency (*channel* or *subband*) components have been questioned [19], [23]. These potential shortcomings motivate the idea of optimizing signal processing modules of feature extractor, including spectrogram-based features. Such topic has been addressed recently for audio representation learning [24] and ASV [25], but expanding and optimizing nonlinear compression module has received less attention.

This study, motivated by the above, addresses *channel-dependent* (CD) nonlinear compression of spectrogram energies. This is realized, as presented in Section 2, by expanding the nonlinearity from a channel-independent to channel-dependent operation. Similar ideas on mel spectrogram have been effective in keyword spotting [26], audio classification [27], and far-field speaker verification [28]. However, to the best of our knowledge, it has not been applied to various nonlinear compression methods in the task we consider.

Our main contributions are summarized in two folds: 1) We leverage the power of such channel-dependent setting by revisiting two established nonlinear compression methods that have been efficient in previous works and generalized them to be channel-dependent, namely power function and dynamic range compression; 2) In order to capture different level of variabilities and compromise instabilities during the joint optimization, we propose a *multi-regime* (MR) design based on CD.

2. NONLINEAR COMPRESSION IN ACOUSTIC FEATURE EXTRACTION

When using spectrogram to extract features, as illustrated in Fig. 1, we typically apply logarithmic compression to spectral energies. However, logarithm has a singularity at zero. This problem is often addressed by adding a small positive offset: $\log(x + \text{offset})$. Even if it avoids the singularity, the ad-hoc design still lacks specificity to a given task and has unpredictable impacts for different kinds of input [26]. We consider two alternative parameterized methods and further make them to be channel-dependent, which are described below.

2.1. Power Function

The concept of applying power nonlinearity to compress the signal amplitude is inspired by human-auditory processing [29], [30]. By using X and Y to denote the input and output magnitude spectra respectively, power nonlinearity is expressed as:

$$Y[t, f] = X[t, f]^{1/\alpha}, \tag{1}$$

where α is known as *temperature coefficient* for the compression. t and f are the time and channel indices, respectively, for spectrogram energies. Experimentally, two particular values of α have been popular in speech front-ends. The first one is $\alpha = 3$, known as *cube-root* [31]–[33]. The other one is $\alpha = 15$, known as *power-law* [34]. Setting higher values of α can provide better recognition performance in the presence of white noise, while lower values may be required for maintaining accuracy for cleaner speech [11].

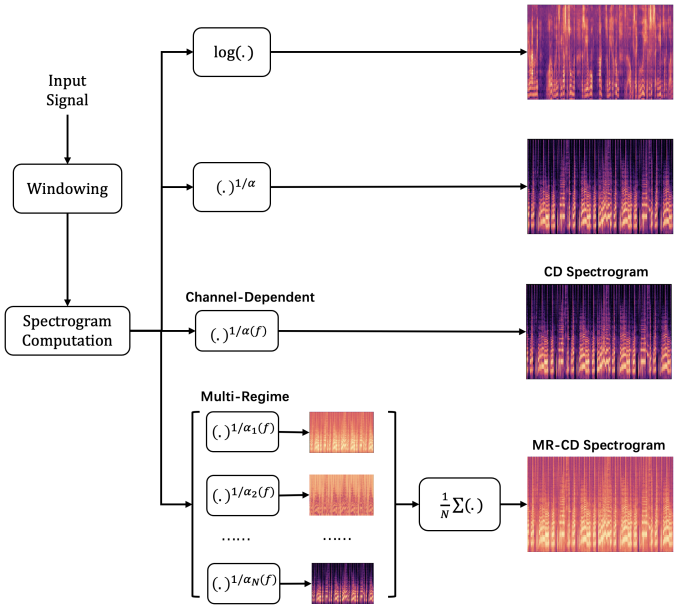

Fig. 1: Feature extraction with log and proposed compression methods, using power function as example.

2.2. Dynamic Range Compression

Power nonlinearity neither addresses foreground-background noise nor other variations. These problems can be addressed by applying dynamic range compression (DRC). It was proposed and applied originally to far-field keyword spotting as part of *per-channel energy normalization* (PCEN) [26], [35]. PCEN has also been applied recently in audio event detection [36]. Using the same notation as above, the DRC operation is defined by:

$$Y[t,f] = (X[t,f] + \delta)^r - \delta^r, \tag{2}$$

where $\delta > 0$ is a positive bias and r is the exponential offset. This method bears resemblance to *spectral subtraction* [37] in speech denoising. In the context of PCEN, it is applied to the spectral energies processed with *automatic gain control* (AGC) [38].. In this study, under the framework of DNN-based ASV, we investigate the efficiency of DRC by directly applying it to spectral energies.

3. PROPOSED METHOD

3.1. Channel-Dependent Design

While the control parameters in Eq. (1) and (2) can be set by hand, this may lead to suboptimal performance in a recognition task. Related prior studies on channel-dependent compression utilize information such as loudness and signal-to-noise ratio (SNR) [39], which motivates data-driven settings. From the equations, we can see that the parameters α, δ, and r are differentiable. Therefore, we propose to optimize them as part of the neural network by generalizing to their channel-dependent (CD) counterparts: $\boldsymbol{\alpha} = \alpha(f), \boldsymbol{\delta} = \delta(f), \boldsymbol{r} = r(f)$, where f is the channel index. This design follows the proposal from [26]. The generalized parameters are then jointly optimized with the neural network. Furthermore, we employ *kernelized initialization* where the parameters are initialized from their static counterparts [25].

3.2. Multi-Regime Design

Learnable parameters are tuned and selected by the training data during the learning process, thus might be suboptimal if domain mismatch between training and testing data is large, due to joint training which may let the parameters suffer from the overparameterized DNN models [40]. CD generalization of the parameters with kernel initialization may scrutinize such problem by larger search space and proper starting point, but it still may fail to have a wide-enough coverage of different level of speech variabilities. This is especially the case when the DNN has large number of layers, which may cause the problem of vanishing gradient when being back-propagated to first early layers, then nonlinearities [41].

Therefore, inspired by the design of *multiple* feature maps in image processing [42] and audio event detection [36], we use a *multi-regime* (MR) design by passing the spectrum to multiple submodules, with shared compression algorithm, but different initialized parameters, as shown in Fig. 1. The output of each module is averaged to form the input for further operations. Using power function as an example:

$$Y[t,f] = \frac{1}{N}\sum_{i=1}^{N}(X[t,f])^{1/\alpha_i(f)} \tag{3}$$

We define the initial values by defining minimum and maximum to create N evenly spaced values: $\alpha_i = \alpha_{\min} + (\alpha_{\max} - \alpha_{\min}) * i/(N-1), i = 1, ..., N$, where $\alpha_{\max}$ and $\alpha_{\min}$ are maximum and minimum reference values and N denotes number of intermediate spectrograms generated. The setup for this work is shown in Table 1. For all cases in this work, $N = 3$. Further tuning of number of intermediates and parameter search is left as future work.

Method	CD	MR-CD
cube root [31]	$\alpha = 3$	$\alpha_{\max} = 3, \alpha_{\min} = 1$
power law [33]	$\alpha = 15$	$\alpha_{\max} = 15, \alpha_{\min} = 1$
DRC	$\delta = 2.0$ $r = 0.5$	$\delta_{\max} = 2.0, \delta_{\min} = 1.0$ $r_{\max} = 1.0, r_{\min} = 0.0$

Table 1: Parameter settings for kernel initialization. DRC values in CD are from [26] while for MR-CD it was hand-crafted based on [35] and pilot experiments.

4. EXPERIMENTAL PROTOCOL

Data. For all experiments, we train the DNN speaker embedding extractor using the *dev* set of VoxCeleb2 [16], which consists of 5994 speakers. We report the performance of different methods on two evaluation sets: 1) The two test sets from VoxCeleb1 [12] following [43], known as *VoxCeleb1-E* and *VoxCeleb1-H*. 2) The recent *VoxMovies* [44], which overlaps with VoxCeleb1 in terms of speakers and contains various levels of mismatch between the enrollment and test utterances. It consists of five trial sets, denoted E-1 (easiest) through E-5 (hardest). Besides condition-specific results, we also report the *pooled* performance over the all five sets.

Features. We use raw magnitude spectrogram obtained using *short-time Fourier transform* (STFT) as the time-frequency representation, to which different compression methods are applied, as illustrated in Fig. 1. The number of frequency bins $N_{\text{STFT}} = 512$ for all systems. The sampling rate is 16 kHz, and the STFT is computed using a 25 ms Hamming window every 10 ms. Additionally,

		VoxCeleb1-E		VoxCeleb1-H		VoxMovies					
Method	Design	EER(%)	minDCF	EER(%)	minDCF	E-1	E-2	E-3	E-4	E-5	*pooled*
log	-	2.23	0.2676	4.43	0.5371	10.55	16.82	15.17	24.73	19.55	20.64
$\log(x + \text{offset})$	-	2.41	0.2920	4.93	0.6219	13.3	16.49	14.41	25.44	18.38	19.00
cube-root	-	1.93	0.2377	4.02	0.5090	10.68	14.74	13.00	25.21	16.61	18.21
	CD	**1.91**	**0.2329**	**3.84**	**0.5071**	10.84	14.79	**12.93**	25.71	16.67	18.33
	MR-CD	2.24	0.2876	4.52	0.5421	10.92	**12.99**	13.77	**15.86**	15.79	**14.27**
power-law	-	2.14	0.2542	4.38	0.5239	11.09	15.15	13.58	25.96	17.13	18.70
	CD	2.14	0.2505	4.29	0.5186	**10.25**	13.44	13.23	21.76	15.82	16.39
	MR-CD	2.78	0.3141	5.31	0.5683	12.07	14.27	15.35	18.05	16.92	15.80
DRC	-	2.25	0.2598	4.60	0.5629	11.09	14.07	14.14	22.20	**15.16**	16.89
	CD	2.67	0.2993	5.20	0.6408	11.95	14.07	14.45	19.78	16.24	16.38
	MR-CD	2.90	0.3526	5.81	0.6442	12.96	15.48	16.70	19.55	16.94	17.03

Table 2: Speaker verification results on VoxCeleb and VoxMovies. 'CD' means channel-dependent and 'MR' means multi-regime setups. Rows with slight grey shades are regarded as baselines. '-' indicates the system with static nonlinearity without learning involved. For VoxMovies all results are reported in EER(%).

we include a system where the logarithm is factorized by an offset as part of the baseline, as mentioned in Section 2: $\log(x + \text{offset})$. The offset is parameterized by an exponential function $\text{offset} = exp(\boldsymbol{\beta})$, where $\boldsymbol{\beta} = \beta(f)$ is CD and initialized with normal distribution.

Speaker embeddings. We use x-vector with *extended TDNN* to generate speaker embeddings, following the design choice from [45] with two main modifications: 1) We replace the statistics pooling layer with attentive statistics pooling [46]; 2) Instead of multi-class cross-entropy, we use *additive angular softmax* [47] as the training objective. We set the scaling factor $s = 30$ and the margin $m = 0.2$. We extract the embedding vectors from the first fully-connected layer after the pooling layer. The extracted vectors are centered and projected via a 150-dimensional *linear discriminant classifier* (LDA).

Evaluation. For both VoxCeleb1 and VoxMovies, we train *probabilistic LDA* (PLDA) classifier using VoxCeleb1. We report ASV performance in terms of *equal error rate* (EER) and *minimum detection cost function* (minDCF). For minDCF, the target speaker prior is $p_{\text{tar}} = 0.01$ and detection costs were $C_{\text{fa}} = C_{\text{miss}} = 1.0$.

5. RESULTS

5.1. Speaker Verification

The results are presented in Table 2. Let us first focus on VoxCeleb1. The EERs for both of the two power functions are improved from both the logarithm baseline and their static counterparts (marked as '-' in the 'Design' column of the table) by the CD design. As part of the baseline, applying CD exponential offset on the logarithm compression degrades the performance for both test sets. The lowest EER on both test sets is obtained using *cube-root* with CD, outperforming the baseline logarithm by 14.3% and 13.3%, respectively. This indicates the usefulness of CD. Nevertheless, the same design does not work well with DRC, which contradicts the findings reported in [26] with mel spectrogram (in a different task, though). This indicates that in ASV with spectrogram input, DRC may not combine well with CD compression. For all compression methods, the MR-CD design degrades performance and fails to show substantial improvement over the logarithm. One reason could be suboptimal parameter initialization as in [26], where the DRC parameters are set for far-field keyword spotting.

On the other hand, the trend is different for VoxMovies with more severe mismatch. Both *pooled* and condition-specific indicate that improvements from CD generalization are modest, as opposed with the observations from VoxCeleb1. For *cube-root*, CD actually degrades the performance for pooled and individual trial sets apart from E-3, where 14.7% relative EER reduction is obtained over the logarithm. However, for *power-law* CD improves upon its static counterpart, with lowest EER on E-1 across all systems.

Generalization via both MR and CD brings substantial improvements on nonlinearities based on power function. Lowest EER of pooled, E-2 and E-4 is obtained using *cube-root* with MR and CD. Its pooled performance outperforms its static counterpart by relatively 21.6%. We notice the same for *power-law*, whose pooled performance with MR and CD outperform its static version by relatively 15.5%. This indicates usefulness of MR in enhancing the robustness.

Nevertheless, the behavior of DRC is different from the power function. While its static setting reduced the EER from the logarithm by relative 18.2%, applying CD results in only slight relative EER reduction (3.1%) on pooled results and does not lead to better performance for the individual trial sets apart from E-4. Generalizing it with MR degrades the performance, which agrees with our observations for VoxCeleb. However, its static setting reaches the best performance across all methods. This indicates the parameters for DRC (bias and offset) being not suitable to cope with CD and MR, at least not within the DRC framework itself. Further investigation is needed on its parameterization.

5.2. Representation Analysis

We illustrate the learnt temperature representation of the two power nonlinear functions from VoxCeleb2 *dev* set (as described in Section 4) in Fig. 2. Note that for power function, larger magnitude of temperature parameters will result in higher compressing effect.

As shown in the figure, applying only CD on power functions casts more compression on both low and high frequency regions (higher temperature values imply more aggressive compression, according to Eq.(1)). Meanwhile, applying MR results in relatively less compression on some of the middle frequency components as well as low frequency regions, while relatively maintaining its pattern on high frequency components. Interestingly, we see that both power function methods result in similar numerical range, even if

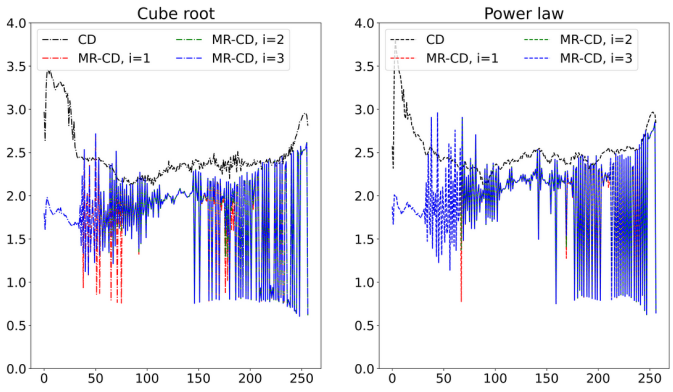

Fig. 2: Learnt temperature values of nonlinear compression based on power function. The x-axis denotes frequency (channel) bin index and y-axis measures the values.

their initialized values are very different (Table 1).

6. CONCLUSION

In this work, we have investigated the alternative nonlinearities for spectrogram compression and their dynamic, channel-dependent variants. We have extended their representation via channel-wise manner and utilized a multi-regime design based on it. Initialization on relevant parameters has been based on the corresponding static known values. We have evaluated the performance of proposed extended dynamic compression methods for different degree of mismatch conditions. Results demonstrates the efficacy of the proposed methods on power nonlinearities, with a maximum of 21.6% pooled EER reduction on VoxMovies. Future work may focus on: 1) extending the framework with other types of spectrogram; 2) exploring more advanced design and appropriate initialization and tuning methods, especially for DRC.

7. ACKNOWLEDGEMENTS

This work was partially supported by Inria Nancy Grand Est.

8. REFERENCES

[1] J. Hansen and T. Hasan, "Speaker recognition by machines and humans: A tutorial review," *IEEE Sig. Proc. Mag.*, vol. 32, no. 6, pp. 74–99, 2015.

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

[3] N. Dehak, P. J. Kenny, R. Dehak, *et al.*, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[4] D. Snyder *et al.*, "X-vectors: Robust DNN embeddings for speaker recognition," in *Proc. ICASSP*, 2018, pp. 5329–5333.

[5] S. Ioffe, "Probabilistic linear discriminant analysis," in *ECCV 2006*, 2006, pp. 531–542.

[6] S. Ramoji *et al.*, "NPLDA: A deep neural plda model for speaker verification," in *Proc. Odyssey 2020*, 2020, pp. 202–209.

[7] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[8] E. Loweimi, J. Barker, O. S. Torralba, *et al.*, "Robust source-filter separation of speech signal in the phase domain," in *Proc. Interspeech*, 2017, pp. 414–418.

[9] L. D. Alsteris and K. K. Paliwal, "Short-time phase spectrum in speech processing: A review and some experimental results," *Digital Signal Processing*, vol. 17, no. 3, pp. 578–616, 2007.

[10] Z. Wu *et al.*, "Synthetic speech detection using temporal modulation feature," in *Proc. ICASSP*, 2013, pp. 7234–7238.

[11] C. Kim and R. Stern, "Power-normalized cepstral coefficients (PNCC) for robust speech recognition," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 24, no. 7, pp. 1315–1329, 2016.

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

[13] S. Shon *et al.*, "VoiceID loss: Speech enhancement for speaker verification," in *Proc. Interspeech*, 2019, pp. 2888–2892.

[14] C. Zhang *et al.*, "Towards robust speaker verification with target speaker enhancement," in *Proc. ICASSP*, 2021, pp. 6693–6697.

[15] W. Xie *et al.*, "Utterance-level aggregation for speaker recognition in the wild," in *Proc. ICASSP*, 2019, pp. 5791–5795.

[16] J. Chung *et al.*, "Voxceleb2: Deep speaker recognition," in *Proc. Interspeech*, 2018, pp. 1086–1090.

[17] D. Garcia-Romero, G. Sell, and A. Mccree, "MagNetO: X-vector magnitude estimation network plus offset for improved speaker recognition," in *Proc. Odyssey 2020 The Speaker and Language Recognition Workshop*, 2020, pp. 1–8.

[18] B. Desplanques *et al.*, "ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification," in *Proc. Interspeech*, 2020, pp. 3830–3834.

[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] N. Zeghidour, N. Usunier, I. Kokkinos, *et al.*, "Learning filterbanks from raw speech for phone recognition," in *Proc. ICASSP*, 2018, pp. 5509–5513.

[21] J.-w. Jung, H.-s. Heo, h. Kim, *et al.*, "Rawnet: Advanced end-to-end deep neural network using raw waveforms for text-independent speaker verification," *Proc. Interspeech*, pp. 1268–1272, 2019.

[22] W. Lin and M. Mak, "Wav2Spk: A simple DNN architecture for learning speaker embeddings from waveforms," in *Proc. Interspeech*, 2020, pp. 3211–3215.

[23] E. Loweimi, P. Bell, and S. Renals, "On learning interpretable CNNs with parametric modulated kernel-based filters," in *Proc. Interspeech*, 2019, pp. 3480–3484.

[24] M. Won, S. Chun, O. Nieto, *et al.*, "Data-driven harmonic filters for audio representation learning," in *Proc. ICASSP*, 2020, pp. 536–540.

[25] 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.

[26] Y. Wang *et al.*, "Trainable frontend for robust and far-field keyword spotting," in *Proc. ICASSP*, 2017, pp. 5670–5674.

[27] C. Ick and B. McFee, "Sound event detection in urban audio with single and multi-rate pcen," in *Proc. ICASSP*, 2021, pp. 880–884.

[28] X. Liu, M. Sahidullah, and T. Kinnunen, "Parameterized channel normalization for far-field deep speaker verification," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU) (to appear)*, 2021.

[29] R. M. Stern, A. Acero, F.-H. Liu, *et al.*, "Signal processing for robust speech recognition," in *Automatic Speech and Speaker Recognition: Advanced Topics*, C.-H. Lee, F. K. Soong, and K. K. Paliwal, Eds. Boston, MA: Springer US, 1996, pp. 357–384.

[30] C. Kim and R. M. Stern, "Power-normalized cepstral coefficients (PNCC) for robust speech recognition," in *Proc. ICASSP*, 2012, pp. 4101–4104.

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

[32] M. J. Alam *et al.*, "Multitaper MFCC and PLP features for speaker verification using i-vectors," *Speech Communication*, vol. 55, no. 2, pp. 237–251, 2013.

[33] S. O. Sadjadi and J. H. Hansen, "Mean hilbert envelope coefficients (MHEC) for robust speaker and language identification," *Speech Communication*, vol. 72, pp. 138–148, 2015.

[34] C. Kim, "Signal processing for robust speech recognition motivated by auditory processing," Ph.D. dissertation, Carnegie Mellon University, School of Computer Science, 2010.

[35] V. Lostanlen *et al.*, "Per-channel energy normalization: Why and how," *IEEE Sig. Pro. Lett.*, vol. 26, no. 1, pp. 39–43, 2019.

[36] C. Ick and B. McFee, "Sound event detection in urban audio with single and multi-rate PCEN," in *Proc. ICASSP*, 2021, pp. 880–884.

[37] J. Porter and S. Boll, "Optimal estimators for spectral restoration of noisy speech," in *Proc. ICASSP*, vol. 9, 1984, pp. 53–56.

[38] J. Prez *et al.*, *Automatic Gain Control: Techniques and Architectures for RF Receivers*. Springer-Verlag New York, 2011.

[39] B. Nasersharif and A. Akbari, "A framework for robust MFCC feature extraction using SNR-dependent compression of enhanced mel filter bank energies," in *Proc. Interspeech*, 2006.

[40] C. Liu, L. Zhu, and M. Belkin, "On the linearity of large non-linear models: When and why the tangent kernel is constant," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, *et al.*, Eds., vol. 33, Curran Associates, Inc., 2020, pp. 15 954–15 964.

[41] E. Loweimi, P. Bell, and S. Renals, "On the robustness and training dynamics of raw waveform models," in *Proc. Interspeech*, 2020, pp. 1001–1005.

[42] R. C. Gonzalez and R. E. Woods, *Digital image processing*. Upper Saddle River, N.J.: Prentice Hall, 2008.

[43] J. Chung, A. Nagrani, E. Coto, *et al.*, "VoxSRC 2019: The first voxceleb speaker recognition challenge," in *ISCA archive*, 2019.

[44] A. Brown, J. Huh, A. Nagrani, *et al.*, "Playing a part: Speaker verification at the movies," in *Proc. ICASSP*, 2021, pp. 6174–6178.

[45] D. Snyder *et al.*, "Speaker recognition for multi-speaker conversations using x-vectors," in *Proc. ICASSP*, 2019, pp. 5796–5800.

[46] K. Okabe *et al.*, "Attentive statistics pooling for deep speaker embedding," in *Proc. Interspeech*, 2018, pp. 2252–2256.

[47] F. Wang, J. Cheng, W. Liu, *et al.*, "Additive margin softmax for face verification," *IEEE Signal Processing Letters*, vol. 25, no. 7, pp. 926–930, 2018.