Parameterized Channel Normalization for Far-field Deep Speaker Verification

Xuechen Liu, Md Sahidullah, Tomi Kinnunen

PARAMETERIZED CHANNEL NORMALIZATION FOR FAR-FIELD DEEP 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

We address far-field speaker verification with deep neural network (DNN) based speaker embedding extractor, where mismatch between enrollment and test data often comes from convolutive effects (e.g. room reverberation) and noise. To mitigate these effects, we focus on two parametric normalization methods: per-channel energy normalization (PCEN) and parameterized cepstral mean normalization (PCMN). Both methods contain differentiable parameters and thus can be conveniently integrated to, and jointly optimized with the DNN using automatic differentiation methods. We consider both fixed and trainable (data-driven) variants of each method. We evaluate the performance on Hi-MIA, a recent large-scale far-field speech corpus, with varied microphone and positional settings. Our methods outperform conventional mel filterbank features, with maximum of 33.5% and 39.5% relative improvement on equal error rate under matched microphone and mismatched microphone conditions, respectively.

Index Terms— acoustic feature extractor, channel normalization, spectrogram, far-field speaker verification.

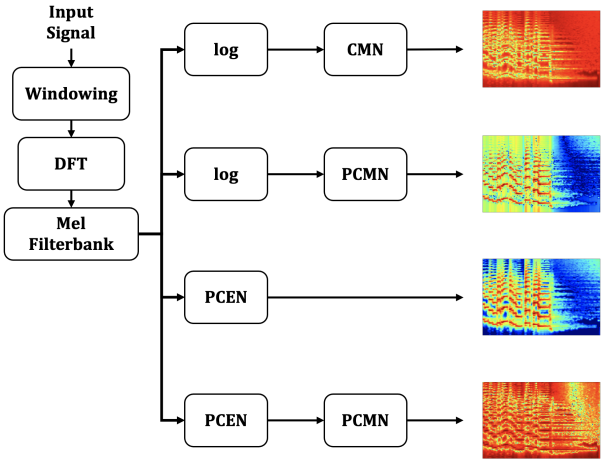


Fig. 1: Proposed feature extractor with channel normalization on mel filterbank, DFT: discrete Fourier transform, PCEN: per-channel energy normalization, CMN: cepstral mean normalization, PCMN: parametric cepstral mean normalization.

1. INTRODUCTION

Automatic speaker verification (ASV) [1] systems aim at verifying the speaker from input speech. ASV systems consist of three main components: acoustic feature extractor, speaker embedding extractor, and backend classifier. Thanks to advances in speaker embedding extraction based on deep neural networks (DNNs), ASV systems have improved substantially from conventional models such as i-vectors [2]. This parallels advances in other speech tasks, such as automatic speech recognition (ASR) [3] and keyword spotting (KWS) [4].

While achieving promising performance under controlled conditions, ASV performance remains substantially low in far-field scenarios, where the user must be authenticated from a distance [5]. Far-field ASV is needed in multiple applications, from virtual assistants to teleconferencing. Given its evolving research value, dedicated datasets and benchmarks have been released. One example is *Voices Obsecured in Complex Environmental Settings* (VOiCES) [5], which formulates a multi-channel simulated distant ASV scenario. Another recent example is Hi-MIA [6], a bilingual corpus (English and Mandarin) focused on smart home scenario with precise definition of microphone types and positions.

Far-field ASV is much more challenging compared to conventional ASV. The factors that impact recognition accuracy can be classified into two main categories: 1) **Environmental variations**, which includes natural room reverberation and additive noises. Common ways to tackle these include masking and de-reverberation techniques [7, 8], along with other speech enhancement techniques; 2) **Intrinsic variations** introduced by microphone array and speakers themselves. Earlier studies have addressed model-wise multi-channel training [9] and adding beamforming front-end [10].

In this work, we thus focus on improving acoustic feature extractor without increasing computational complexity substantially. Such efforts lead to decent progress in other speech processing tasks. One successful example is *power-normalized cepstral coefficients (PNCCs)* [11], whose efficacy has been demonstrated in speech recognition under various noisy conditions. Multi-taper MFCCs [12] and linear pre-

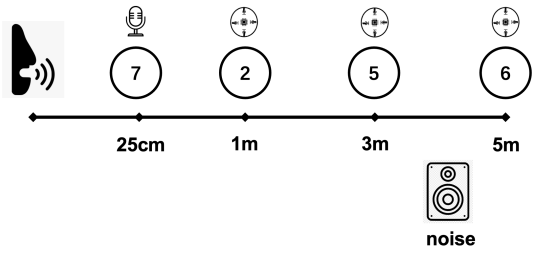


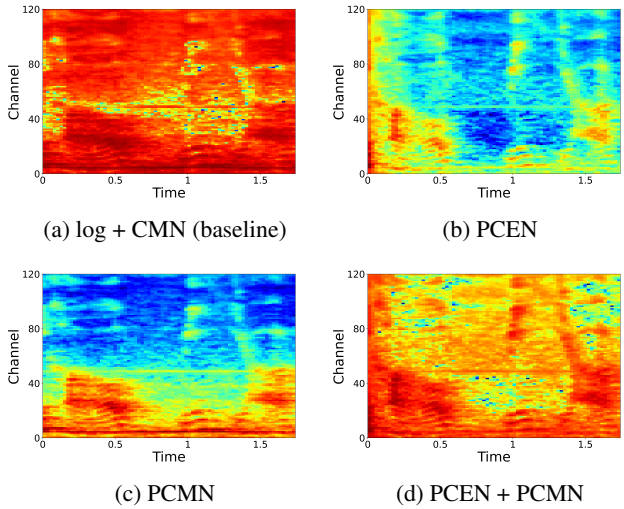
Fig. 2: Recording condition of Hi-MIA, adopted from [6], excluding sources that are not included in our protocol. The distance between noise source and speaker is 4m.

dictive features target robust ASV, but the advantages demonstrated with conventional models [13] do not necessarily generalize to DNN-based ASV [14].

In this work, we extend upon earlier research by adopting two parametric methods that operate on time-frequency spectrogram to enhance feature robustness. The first method, *per-channel energy normalization* (PCEN) [15, 16], replaces the commonly-used logarithmic compression. The second method, *parameterized cepstral mean normalization* (PCMN) [17], generalizes the widely-used parameter-free *cepstral mean normalization* (CMN) by the inclusion of learnable normalization parameters. Even if PCEN and PCMN were addressed in other tasks, to the best of the authors' knowledge, they have not been extensively addressed in DNN-based ASV for far-field applications. Thus, the primary contribution of this work is integration (and joint training of) these compensation methods with modern DNN-based ASV systems. Finally, we also propose a novel channel normalization method that combines the two methods in cascaded fashion. Our experiments are conducted on the recent Hi-MIA corpus [6] with certain advantages elaborated below.

2. HI-MIA DATASET FOR FAR-FIELD SPEAKER VERIFICATION

The work of this paper is conducted on Hi-MIA. It is a dataset which contains 1561 hours of audio from 340 speakers [6] recorded both under clean and smart home conditions in Mandarin and English. It consists of two subsets: *AISHELL-wakeup* and *AISHELL-2019B-eval*, which complement one another in terms of recording conditions. Speech was collected with both close-talking (clean) high-fidelity microphone and 16 types of distributed microphone arrays under real rooms, corresponding to smart home environments. This is one of the advantages of Hi-MIA compared to VOiCES [5], where the source audios are replayed versions of pre-recorded audio. The positional information of different types of devices and noise resources is illustrated in Fig. 2. The publicly available part of the data contains recordings for the close-talking microphone (position 7) and the microphone arrays at positions 2, 5 and 6. For more details about gender and age distribution, refer to [6]. Hi-MIA shares similar recording conditions and contains overlapped speech with the evaluation data of the recent *Far-Field Speaker Verification Challenge 2020* [18].



(a) log + CMN (baseline) (b) PCEN

(c) PCMN (d) PCEN + PCMN

Fig. 3: Different spectrogram representations on a speech utterance from Hi-MIA.

3. PARAMETERIZED CHANNEL NORMALIZATION

3.1. Per-channel Energy Normalization

As a dynamic compression technique, PCEN [15] addresses the singularity problem of logarithmic compression at zero, which is non-trivial and not robust to environmental and loudness variations from speakers. With t and f being time and frequency indices, PCEN is formulated as:

$$\text{PCEN}[t,f] = \left(\frac{E[t,f]}{(M[t,f]+\epsilon)^{\alpha}} + \delta\right)^{r} - \delta^{r} \qquad (1)$$

Here, $E[t,f]$ notes the input spectrogram energies. The PCEN operation in eq. 1 consists of two parts: *automatic gain control* (AGC) and *dynamic range compression* (DRC). AGC is represented by the term $G[t,f] = E[t,f]/(M[t,f]+\epsilon)^{\alpha}$, where $M[t,f] = (1-s)M[t-1,f]+sE[t,f]$ denotes temporally integrated energies, computed using a first-order infinite impulse response (IIR) filter with pre-set smoothing coefficient s. For all experiments in this paper, s is the reciprocal to number of mel filters (40), following [15]. The main control parameter is $\alpha \in (0,1]$, which models the degree of compression. ϵ is a small number to avoid division by zero. Note that computation of $G[t,f]$ can be re-formulated to subtraction at logarithmic domain, followed by an exponential operation.

After obtaining the AGC-controlled energies, PCEN spectrogram is obtained by DRC, which is expressed as $(G[t,f]+\delta)^r - \delta^r$, where the positive bias term $\delta > 1$ and the exponent $r \in (0, 1]$ are the main control parameters. They are designed to compress the loudness variations in the signal, reflecting earlier work on speech restoration [19].

Fig. 3a and 3b compare the baseline and PCEN normalized mel spectrograms on a speech utterance. PCEN performs extensive compression of the dynamic range in specific parts of the signal while keeping part of the pattern. On the other hand, it has enhancement effect on speech onsets, which may be helpful in improving robustness. More work on analyzing compression characteristics of PCEN can be found in [16].

3.2. Parametric Cepstral Mean Normalization

Conventional CMN is a parameter-free and blind estimator which does not have interaction with other speech modules. It is expressed as below:

$$\hat{X}_t[i] = X_t[i] - \mu_t[i] \tag{2}$$

where t and i are the time frame and feature indices, respectively. $\mu_t[i] = \sum_{m=t-N}^{t} X_m[i]/N$ stores cepstral mean values with a sliding window of length $N + 1$. As a replacement, PCMN [17] is formulated as:

$$\hat{X}_t[i] = \beta[i]X_t[i] - (\alpha[i]\mu_t[i] + \mu_0[i]) \tag{3}$$

where β, α and μ_0 are the additional parameters. Compared to conventional CMN, this parameterized version allows the interaction between the normalizer and other learnable modules such as DNN speaker embedding extractor. It decides whether performing cepstral mean subtraction or not by varying the gain parameters $\alpha[i] \in [0, 1]$.

Similar to PCEN, PCMN can also be effective on handling speech characteristics [17]. However, it may treat such information in a different way. Fig. 3c shows the PCMN-processed mel spectrogram. Interestingly, the low-frequency speech components are generally preserved while high-frequency energies are subtracted to a lower level by PCMN. The compressed part becomes not as flat as observed from PCEN. This might be beneficial in preserving speech patterns as PCEN while creating more distinction and may also smear useful speech information. We take such a potential advantage and hypothesize that directly temporal modeling via PCMN can also lead to better ASV performance.

3.3. Trainable channel normalization

While related parameters can be primarily set in hand-crafted fashion, both PCEN and PCMN have data-driven variants in their original works [15, 17], making joint optimization via back propagation possible along with the DNN speaker embedding extractor.

For PCEN, as seen from eq. 1, the parameters α, δ, r are designed to be differentiable. Therefore, they can be generalized as frequency-dependent or even time-frequency dependent. Following [15], we generalize them to be frequency-dependent and perform the joint optimization: $\boldsymbol{\alpha} = \alpha(f), \boldsymbol{\delta} = \delta(f)$, and $\boldsymbol{r} = r(f)$, where f denotes frequency bin index.

For PCMN, from [17] the temporal dependency the method integrates can be leveraged by using a linear projection layer $\boldsymbol{W} \cdot \boldsymbol{Y}_t + \boldsymbol{b}$, where $\boldsymbol{Y}_t = [\boldsymbol{X}_{t-10}, ..., \boldsymbol{X}, ..., \boldsymbol{X}_{t+10}]$ are spliced cepstral input, $\boldsymbol{W} \sim (\boldsymbol{\alpha}, \boldsymbol{\beta})$ and $\boldsymbol{b}$ are corresponding weights and bias. The weights contain the frequency-dependent learnable values of $\alpha[i]$ and $\beta[i]$ and the bias can be regarded as a frequency-dependent variant of $\mu_0[i]$.

In the optimization of above trainable front-ends, we employ *kernelized initialization*, where the parameters are not selected from a specific distribution such as normal one [20], but are migrated from common practical knowledge such as the workable hand-crafted counterparts. It has been applied and demonstrated to be effective for data-driven MFCCs [21]. Kernel initialization sets a starting point for further learning and adaptation via back-propagation. The exact values of related parameters are addressed in the next section.

4. EXPERIMENTS

4.1. Data

We conducted all experiments on the Hi-MIA dataset. Training of the speaker embedding extractor was conducted using AISHELL-2 [22], following the original Hi-MIA protocol [6]. The training data contains 1991 speakers recorded using an iOS device. The data was further augmented using room impulse response (RIR) [23] and noise sources from the MUSAN dataset [24].

Evaluation was conducted on the *test* partition of Hi-MIA. Our protocol defines two trial sets and six trial lists in total, based on recording conditions illustrated in Fig. 2:

- *Matched microphone*, where the type of microphone between each enrollment and test utterances for each trial pair are the same. Enrollment data are recorded at position 2 while test utterances can originate from either 2, 5 or 6. This results in Ma-2, Ma-5, Ma-6 in Table 2.
- *Mismatched microphone*, where the enrollment utterances are recorded using the close-talk microphone at position 7 while the test utterances are from the microphone arrays at position 2, 5 and 6. This results in Mis-2, Mis-5, Mis-6 in table 2.

The protocol of full trial lists are available as an open-sourced Kaldi[1] recipe.

[1] https://github.com/kaldi-asr/kaldi/tree/master/egs/hi_mia/v1

Module	Values
PCEN	$\alpha = 0.98, \delta = 2.0, r = 0.5$ [15]
PCMN	$\beta = 1.0, \alpha = 0.5, \mu_0 = 0.0$

Table 1: Setting of parameter values for related modules for fixed front-ends. They are also used for kernel initialization if applicable.

Trial ID	#Enroll	#Test	#Target	#Nontarget
Ma-2	Arr, Pos. 2	Arr, Pos. 2	35910	35250
Ma-5	Arr, Pos. 2	Arr, Pos. 5	34920	35175
Ma-6	Arr, Pos. 2	Arr, Pos. 6	34575	35175
Mis-2	Mic, Pos. 7	Arr, Pos. 2	34845	35010
Mis-5	Mic, Pos. 7	Arr, Pos. 5	35625	34920
Mis-6	Mic, Pos. 7	Arr, Pos. 6	34860	35655

Table 2: Statistics for Hi-MIA sub-trials. Arr: Microphone array, Mic: close-talking microphone. Pos: position IDs where audios are recorded, referred from Fig. 2.

4.2. System Configuration

Front-ends. For all ASV systems considered the number of mel filters was set to 40, which is same as the input feature dimension for speaker embedding network. We consider mel filterbank with logarithmic compression and CMN post processor as our baseline, as illustrated in Fig. 1. Additionally, we combine PCEN and PCMN, where PCEN replaces the logarithmic non-linearity. We refer to front-ends with fixed and trainable (adaptive) components, respectively, as *fixed front-ends* and *adaptive front-ends*. Details of exact parameter values for fixed front-ends and kernel initialization is shown in Table 1. For kernel initialization, the values of vector-wise parameters are set as constant across all elements.

Speaker embedding extractor. Extended x-vector based on time-delayed neural network (E-TDNN) [25] is used as speaker embedding extractor. It is one of the descendants from *x-vector* which have reached promising performance using mel cepstral features. We introduce two modifications from original configuration: 1) For pooling layer, we acquired attentive statistics pooling [26]; 2) For loss function we make use of additive margin softmax [27]. For inference, we extract 512-dimensional embedding vector for each utterance from the first fully-connected layer after the pooling layer.

Backend. For all the experiments, we train corresponding probabilistic linear discriminant classifiers (PLDA) using the speaker embeddings from the *train* partition provided by Hi-MIA open-sourced protocol. Embeddings are processed via mean subtraction, length normalization, and centering using a 200-dimensional linear discriminant analyzer (LDA) before being fed to the PLDA.

Evaluation. Results are reported in terms of equal error rate (EER %). We also provide detection error trade-off (DET) curves of fixed front-ends for analysis.

5. RESULTS

The results on fixed and adaptive front-ends are presented in Table 3 and 4, respectively. Prefix 'a' added to the beginning of component indicates the adaptive variants. Recall that the mismatched microphone scenario contains additional mismatch between high-fidelity close talking microphone and the microphone array. The enrollment utterances are always recorded with microphone distance of 25cm from the speaker. Therefore, it is expected that for test utterances at same position, corresponding EERs are generally higher. This is what we indeed observe in both scenarios. In the following, we discuss results for matched and mismatched scenarios one by one.

5.1. Matched Microphone

Fixed front-ends. All the proposed variants outperform the baseline when there is a large distance between the enrollment and test recordings (Ma-5, Ma-6). Lowest EERs come from log+PCMN, where the maximum relative improvement of 33.5% over baseline comes from the furthest microphone position. Meanwhile, when the sources come from the same position, systems with PCEN do not outperform the baseline. Nonetheless, they do reach slightly lower EERs when the testing microphone moves from position 5 to 6. This indicates the potential of PCEN in normalizing room acoustic differences. Combining PCEN and PCMN gives slightly worse numbers than PCEN for all three trials. The comparative difference between PCEN and PCMN may be attributed to suboptimal parameter values: for computational reasons, we chose the values based on earlier recommendations from [16] and [17].

Adaptive front-ends. Somewhat disappointingly, results for the adaptive front-ends indicates generally worse performance compare to their fixed counterparts. The only system that returns satisfying performance for certain conditions is adaptive PCEN, which outperforms fixed PCEN by relatively 2.5% in Ma-2. This indicates sensitivity of data-driven methods, especially there is mismatch between training and test data. There are meanwhile numbers of points can be improved related to engineering issues and domain adaptation on those parameter values, left as a future work. As a starting point, we see that kernelized initialization of the parameters gives slight relative improvement for most cases. Maximum improvement made by kernelized initialization is from aPCMN in Ma-5, by relatively 16%. Such finding furthers the potential importance of hand-crafted knowledge to robust acoustic features.

5.2. Mismatched Microphone

Fixed front-ends. The improvement brought by fixed front-ends agree with ones in the matched microphone scenario. PCMN with logarithmic nonlinearity retains the lowest EERs overall, with a maximum relative improvement of 46.6% in

		Hi-MIA matched mic			Hi-MIA mismatched mic		
Non-linearity	Post norm.	Ma-2	Ma-5	Ma-6	Mis-2	Mis-5	Mis-6
log	CMN	3.65	7.43	8.51	8.16	11.01	12.84
PCEN	-	4.05	6.93	6.87	6.45	10.7	12.34
log	PCMN	**3.27**	**5.56**	**5.66**	**4.35**	**6.66**	**9.23**
PCEN	PCMN	4.32	7.23	7.61	6.18	9.87	11.74

Table 3: EER (%) results on Hi-MIA for fixed front-ends.

			Hi-MIA matched mic			Hi-MIA mismatched mic		
Non-linearity	Post norm.	Kernel init.	Ma-2	Ma-5	Ma-6	Mis-2	Mis-5	Mis-6
aPCEN	-	no	4.35	8.25	9.28	7.44	12.64	15.4
aPCEN	-	yes	**3.94**	8.05	9.11	**5.85**	12.47	15.67
log	aPCMN	no	4.45	8.85	9.14	8.94	13.24	14.58
log	aPCMN	yes	4.31	7.43	8.9	6.17	11.59	16.09
aPCEN	aPCMN	no	4.35	8.25	9.29	7.44	13.64	16.4
aPCEN	aPCMN	yes	4.26	8.65	9.65	7.37	13.52	16.25
aPCEN(no DRC)	-	yes	4.17	8.55	8.97	7.07	13.02	16.08
aPCEN(no AGC)	-	yes	4.60	**6.66**	**7.48**	6.34	**9.90**	**10.95**

Table 4: EER (%) results on Hi-MIA for adaptive front-ends, including ablation study for adaptive PCEN.

the Mis-2 condition. At the same time, for the two PCEN variants, better verification performance compared to the baseline on all cases is observed. Different from matched microphone condition, here the combination of PCEN and PCMN outperforms PCEN-only front-end, especially at the furthest microphone condition (Mis-6, relatively 4.8%).

Adaptive front-ends. For Mis-2 condition with 75cm distance between the close talking microphone and the arrays, most adaptive front-ends outperform the baseline. Lowest EER is obtained by aPCEN with kernel initialization in Mis-2, outperforming baseline by relatively 28.3%. Nonetheless, for the other two cases, no adaptive front-end feature extractor gives lower EER than baseline and the performance gap between adaptive and fixed systems is noticeable. Also, while still giving relatively lower EER in Mis-5 by a maximum of 16.5% (aPCMN), kernel initialization degrades the performance for all adaptive systems in Mis-6 condition.

5.3. Analysis with DET plots

DET curves of fixed front-ends for PCEN and PCMN on Ma-2 and Mis-2 are illustrated in Fig. 4. The DET curves reiterate what has been reflected in the EERs. Interestingly, however, when the false alarm rate gets high, performances of different front-ends approach to each other. Especially for the combination of logarithmic compression and PCMN: in fact, under Ma-2 it is outperformed by the baseline as can be noticed from the figure. Therefore, it may not be a good option for systems that are less strict on false alarms.

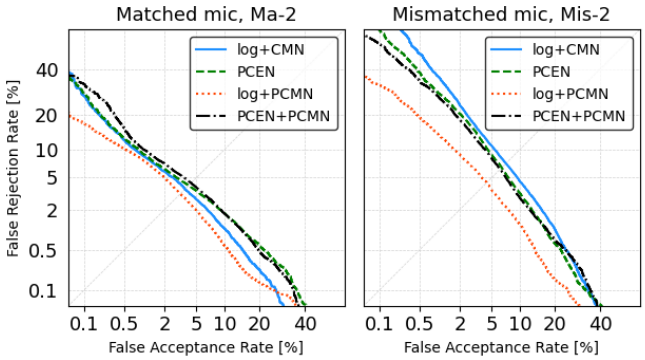


Fig. 4: DET curves for fixed front-ends.

5.4. Analysis with Adaptive Components

Ablation study on PCEN. As noted earlier, PCEN consists of two main components: AGC and DRC. As our last experiment, we address their effect by removing one of them from the adaptive pipeline. The results are shown at the last two rows of of Table 4. By removing DRC the filterbank energies are divided by its filtered variant without further compression. Compared to full adaptive PCEN, the results become worse. Meanwhile, by removing AGC, the energies are directly compressed without the division. In this case, performance of aPCEN is substantially improved and outperforms baseline in all conditions, except for Ma-2 with least mismatch between enrollment and test utterances. Results on Mis-6 indicates re-

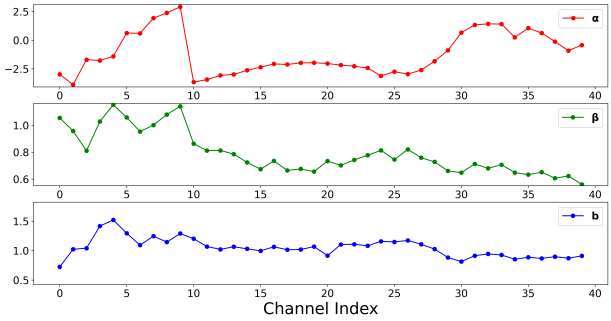


Fig. 5: Values of learnt weights $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and bias $\boldsymbol{b}$ for system with log and aPCMN (with kernel initializaion). Number of channel indices is $i = 40$ for all of them.

versed gap between baseline and outperform it by relatively 14.7% EER. Such observations may unveil the possible disadvantage cast by AGC and the potential benefits brought by DRC as a non-linearity.

Learnt values of PCMN. Finally, we show the weights and bias values of the best-performed system with adaptive PCMN involved in Fig. 5, where we combine logarithmic compression, adaptive PCMN and kernel initialization. As is evident, the learnt normalizers are different for each channel (unlike in the conventional non-parametric CMN). Further interpretation of the learnt normalization operations, and particularly their dependency on the training data, is deferred to a future work.

6. CONCLUSION

We addressed far-field speaker verification problem with mismatch conditions introduced by room reverberation and acoustic noise. We proposed new feature extractors to alleviate the negative impact due to these mismatches on verification performance by introducing two parameterized techniques on channel normalization: PCEN and PCMN. Our results on the recent Hi-MIA dataset confirm the efficacy of the introduced methods, especially for fixed PCMN. Our ablation study indicated the potential of DRC from PCEN.

In future research, we plan continue to explore both methods and their ingredients, especially compression parts such as DRC as a non-linearity itself and better integration on PCEN and PCMN with factorization, in order to compensate the high mismatch created by microphone and the factors.

7. ACKNOWLEDGEMENTS

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

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

[3] G. Hinton, L. Deng, D. Yu, G. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[4] G. Chen, C. Parada, and G. Heigold, "Small-footprint keyword spotting using deep neural networks," in *Proc. ICASSP*, 2014, pp. 4087–4091.

[5] C. Richey, M. A. Barrios, Z. Armstrong, C. Bartels, H. Franco, M. Graciarena, A. Lawson, M. K. Nandwana, A. Stauffer, J. Hout, P. Gamble, J. Hetherly, C. Stephenson, and K. Ni, "Voices obscured in complex environmental settings (VOICES) corpus," Tech. Rep., 2018.

[6] X. Qin, H. Bu, and M. Li, "HI-MIA: A far-field text-dependent speaker verification database and the baselines," in *Proc. ICASSP*, 2020, pp. 7609–7613.

[7] H. Taherian, Z. Wang, and D. Wang, "Deep learning based multi-channel speaker recognition in noisy and reverberant environments," in *Proc. Interspeech 2019*, 2019, pp. 4070–4074.

[8] X. Qin, D. Cai, and M. Li, "Far-field end-to-end text-dependent speaker verification based on mixed training data with transfer learning and enrollment data augmentation," in *Proc. Interspeech 2019*, 2019, pp. 4045–4049.

[9] D. Cai, X. Qin, and M. Li, "Multi-Channel Training for End-to-End Speaker Recognition Under Reverberant and Noisy Environment," in *Proc. Interspeech 2019*, 2019, pp. 4365–4369.

[10] L. Mošner, O. Plchot, J. Rohdin, and J. Černocký, "Utilizing VOiCES dataset for multichannel speaker verification with beamforming," in *Proc. Odyssey 2020*, 2020, pp. 187–193.

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

[12] T. Kinnunen et al., "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.

[13] X. Jing, J. Ma, J. Zhao, and H. Yang, "Speaker recognition based on principal component analysis of LPCC and MFCC," in *2014 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC)*, 2014, pp. 403–408.

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

[15] Y. Wang, P. Getreuer, T. Hughes, R. F. Lyon, and R. A. Saurous, "Trainable frontend for robust and far-field keyword spotting," in *Proc. ICASSP*, 2017, pp. 5670–5674.

[16] V. Lostanlen, J. Salamon, M. Cartwright, B. McFee, A. Farnsworth, S. Kelling, and J.P. Bello, "Per-Channel Energy Normalization: Why and how," *IEEE Signal Processing Letters*, vol. 26, no. 1, pp. 39–43, 2019.

[17] O. Kalinli, G. Bhattacharya, and C. Weng, "Parametric cepstral mean normalization for robust speech recognition," in *Proc. ICASSP*, 2019, pp. 6735–6739.

[18] X. Qin, M. Li, H. Bu, W. Rao, R. K. Das, S. Narayanan, and H. Li, "The INTERSPEECH 2020 far-field speaker verification challenge," in *Proc. Interspeech 2020*, 2020, pp. 3456–3460.

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

[20] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*. 2010, pp. 249–256, PMLR.

[21] 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.

[22] J. Du, X. Na, X. Liu, and H. Bu, "AISHELL-2: transforming mandarin ASR research into industrial scale," *CoRR*, vol. abs/1808.10583, 2018.

[23] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *Proc. ICASSP*, 2017, pp. 5220–5224.

[24] D. Snyder, G. Chen, and D. Povey, "MUSAN: A music, speech, and noise corpus," *CoRR*, vol. abs/1510.08484, 2015.

[25] 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.

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

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