SPATIAL-FEATURE-BASED ACOUSTIC SCENE ANALYSIS USING DISTRIBUTED MICROPHONE ARRAY

Keisuke Imoto[†] *and Nobutaka Ono*[‡,†]

[†] SOKENDAI (The Graduate University for Advanced Studies), Kanagawa, Japan
[‡] National Institute of Informatics, Tokyo, Japan
{k-imoto, onono}@nii.ac.jp

ABSTRACT

In this paper we propose a robust and efficient method to utilize the spatial information provided by a distributed microphone array for acoustic scene analysis. In our approach, similarly to the cepstrum, which is widely used as a spectral feature, the logarithm of the amplitude in multichannel observation is converted to a feature vector by a linear orthogonal transformation. Then, the spatial information of the acoustic scene is represented in the spatial feature space. This approach does not require the positions of the microphones and is not sensitive to the synchronization mismatch of channels, both of which make the method suitable for use with a distributed microphone array. Experimental results using real-life environmental sounds show the validity of our approach even when a smaller feature dimension than the original one is used.

Index Terms— Acoustic scene analysis, distributed microphone array, spatial cepstrum, symmetric microphone array, isotropic sound field

1. INTRODUCTION

Considerable research has been conducted on media tagging, surveillance, and automatic life-logging using an acoustic signal, and such research is referred to as acoustic scene analysis or acoustic event detection [1–4]. There are many techniques for analyzing acoustic scenes based on spectral features such as the combination of mel-frequency cepstral coefficients (MFCCs) and a hidden Markov model (HMM) [5, 6] or directly utilizing spectro-temporal information [7, 8]. The intermediate features such as a dictionary of acoustic events [9–11] or the bases captured by the non-negative matrix factorization (NMF) [12] have been also investigated. These techniques utilize the sparsity or other constraints in spectral or temporal domain, and represent acoustic scenes efficiently with less feature dimensions.

In recent years, multichannel signal processing for acoustic scene analysis has attracted increasing attention because of the rapid increase in the use of acoustic sensors such as smart and wearable devices. If many microphones are distributed, they enable us to obtain spatial information, which can be used to recognize acoustic events or acoustic scenes.

The use of position information based on source localization is a straightforward way to use the spatial information provided by multichannel observation. However, even in the single-source case, source localization is not always easy in a real environment because of background noise, reverberation, and reflection by large obstacles such as partitions and desks. In practical use in a distributed microphone array, the positions of microphones are not known in advance. Thus, they have to be estimated before source localization. Moreover, an acoustic event can include multiple sound sources, which introduces other difficulties such as multiple source localization and estimation of the number of sound sources.

In this paper we propose a robust and efficient method to utilize the spatial information provided by a distributed microphone array for acoustic scene analysis. In our approach, the amplitude information in multichannel observation is converted to a feature vector in a similar way to the cepstrum. Then, the spatial information of the acoustic scene is represented in the spatial feature space. This approach does not require the positions of the microphones and is not sensitive to the synchronization mismatch of channels, both of which make the method suitable for use with a distributed microphone array.

The rest of this paper is organized as follows. In section 2, we introduce a method to extract a spatial feature from multichannel observation and discuss its resemblance to the cepstrum. In section 3, we report simulated experiments on spatial feature extraction and evaluate the proposed method by acoustic scene analysis in a real environment. Finally in section 4, we conclude this paper.

2. SPATIAL FEATURE FOR ACOUSTIC SCENE ANALYSIS

Suppose that an acoustic scene is observed using N microphones and let $s_{\omega,\tau,n}$ be the short-time Fourier transform (STFT) representations of a multichannel observation, where ω, τ, and n represent the frequency, time frame, and channel

978-0-9928626-3-3/15/$31.00 ©2015 IEEE

indices, respectively. To extract spatial pattern information steadily, we also assume that the microphone configuration is fixed. In a distributed microphone array, synchronization over channels is a significant issue and the phase information is sometimes unreliable owing to the sampling frequency mismatch. Therefore, in this paper, we focus on only amplitude information in the STFT representation, $a_{\omega,\tau,n} = |s_{\omega,\tau,n}|$, which is more robust to the sampling frequency mismatch.

2.1. Cepstrum: spectral feature

To extract spectral information, let us consider a frequency-based log amplitude vector such as

$$\mathbf{p}_\tau = \begin{pmatrix} \log \bar{a}_{1,\tau} \\ \log \bar{a}_{2,\tau} \\ \vdots \\ \log \bar{a}_{\Omega,\tau} \end{pmatrix}, \tag{1}$$

where Ω is the number of frequency bins and

$$\bar{a}_{\omega,\tau} = \sqrt{\frac{1}{N}\sum_n a^2_{\omega,\tau,n}} \tag{2}$$

is the average spectrogram over the channels. As another alternative, the largest spectrogram components over channel can be used for extracting spectral information as follows.

$$\bar{a}_{\omega,\tau} = \max_n a_{\omega,\tau,n} \tag{3}$$

The discrete Fourier transform (DFT) of $\mathbf{p}_\tau$ defined as

$$\mathbf{c}_\tau = \mathbf{Z}_\Omega \mathbf{p}_\tau \tag{4}$$

is called the *cepstrum*, where $\mathbf{Z}_\Omega$ is the $\Omega \times \Omega$ DFT matrix. Similarly to the cepstrum, the mel-frequency cepstrum coefficient (MFCC) is defined using the discrete cosine transform (DCT) for mel-frequency representation, which has been widely used as a spectral feature. In both cases, the DFT or DCT matrix acts as a good basis transformation for the log amplitude information, and dimension reduction can be performed by taking lower-order components of $\mathbf{c}_\tau$.

2.2. Spatial cepstrum

Analogous to the definition of the cepstrum, we here consider a channel-based log amplitude vector such as

$$\mathbf{q}_\tau = \begin{pmatrix} \log \tilde{a}_{\tau,1} \\ \log \tilde{a}_{\tau,2} \\ \vdots \\ \log \tilde{a}_{\tau,N} \end{pmatrix}, \tag{5}$$

where

$$\tilde{a}_{\tau,n} = \sqrt{\frac{1}{\Omega}\sum_\omega a^2_{\omega,\tau,n}} \tag{6}$$

is the multichannel power observation at each time frame.

In the frequency case, $\bar{a}_{\omega,\tau}$ represents the logarithm of the amplitude at each subband, which is uniformly spaced on the linear frequency or mel-frequency axis. However, in a spatial case, especially in a distributed microphone array, the microphones are not uniformly placed. Therefore, instead of DFT or DCT, we apply principal component analysis (PCA). Let $\mathbf{R}_q$ be the covariance matrix of $\mathbf{q}_\tau$ given by

$$\mathbf{R}_q = \frac{1}{T}\sum_\tau \mathbf{q}_\tau \mathbf{q}_\tau^{\mathrm{T}}, \tag{7}$$

where T is the number of time frames and $^{\mathrm{T}}$ represents the vector transpose. Because $\mathbf{R}_q$ is a symmetric matrix, the eigenvalue decomposition of $\mathbf{R}_q$ can be represented as

$$\mathbf{R}_q = \mathbf{E}\mathbf{D}\mathbf{E}^{\mathrm{T}}, \tag{8}$$

where $\mathbf{E}$ and $\mathbf{D}$ are the eigenvector matrix and the diagonal matrix in which the diagonal elements are equal to the eigenvalues in descending order, respectively.

Using $\mathbf{E}$, we define the spatial feature as follows.

$$\mathbf{d}_\tau = \mathbf{E}\mathbf{q}_\tau \tag{9}$$

According to PCA, the components of $\mathbf{d}_\tau$ are uncorrelated with each other and we can reduce the feature dimension of $\mathbf{q}_\tau$ without significant loss of information by utilizing only the components whose eigenvalues are large.

If the microphone positions are circularly symmetric and the sound field is isotropic, which means that 1) the acoustic power is identical at all positions and 2) the cross-correlation between two observations does not depend on the angle between their observed positions, the covariance matrix $\mathbf{R}_q$ is a circular matrix and the eigenvalue matrix $\mathbf{E}$ can be an $N \times N$ DFT matrix $\mathbf{Z}_N$ [13, 14]. Then, eq. (9) is exactly equivalent to the definition of the cepstrum. Although this is a special case, we hereafter call $\mathbf{d}_\tau$ the *spatial cepstrum* (*SC*) based on this resemblance.

To calculate the SC, the positions of the microphones are not required, enabling convenient distributed microphone array processing. Also, owing to the resemblance to the original cepstrum, we can apply cepstral mean normalization (CMN) [15] to the SC to compensate for the mismatch in the microphone sensitivity, as well as other techniques applicable to the cepstrum domain [16].

Additionally, we can apply this spatial feature extraction method at each frequency bin without average the feature and we can consider the same approach for spectro-spatially concatenated observation vectors, which are included in our future work.

3. EXPERIMENTS

3.1. Representation of spatial pattern by spatial cepstrum

We show an example of the spatial correlation of a multichannel observation and a spatial pattern represented by the SC in

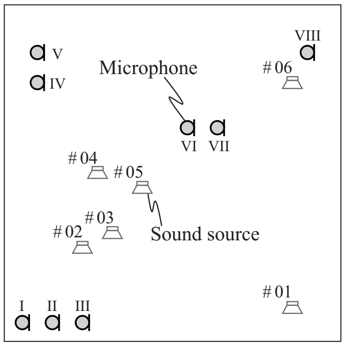


Fig. 1. Microphone and sound source arrangements in simulated experiment

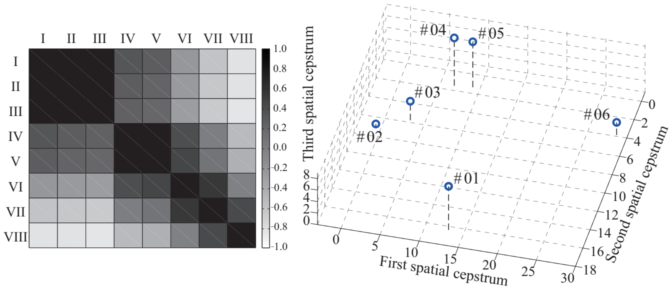


Fig. 2. Spatial correlation of log amplitude (left) and spatial feature represented by spatial cepstrum in 3-dimensions (right)

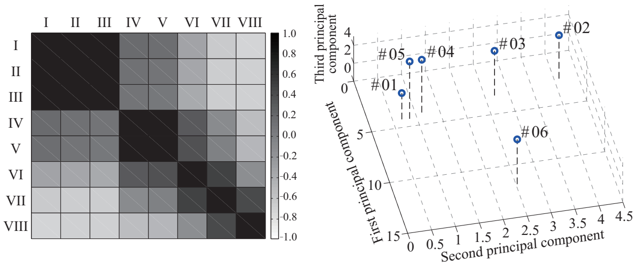


Fig. 3. Spatial correlation of amplitude (left) and spatial feature represented by its principal components in 3-dimensions (right)

a simulated experiment.

Figure 1 shows the arrangement of microphones and loudspeakers in our experiment. In this experiment, each loudspeaker played a fixed-length stationary 1 kHz pure tone in order without the overlap.

Figure 2 shows the spatial correlation between channels (left) and the spatial feature representation of each sound source by the SC (right). For better visualization, the normalized correlation such as eq. (10) is shown where $r(i, j)$ is the (i, j) entry of the covariance matrix of $\mathbf{q}_{\tau,n}$ given by eq. (7).

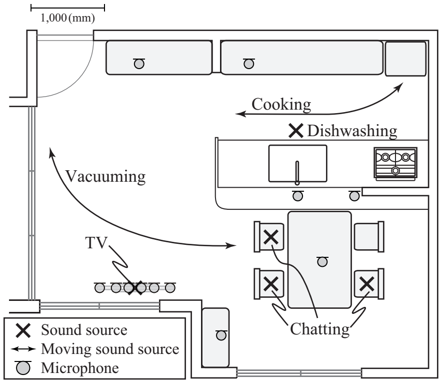


Fig. 4. Microphone and sound source arrangements in experiment using real-life environmental sounds

Table 1. Typical sounds in each acoustic scene

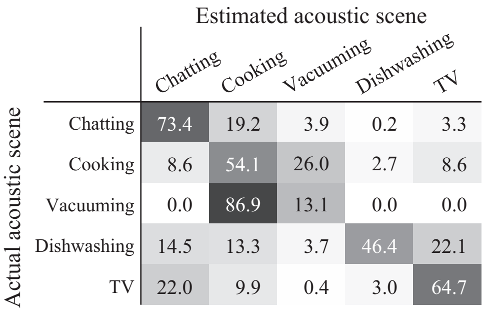
Acoustic scene	Typical sound
Chatting	voices, coughing
Cooking	cutting, sizzling, running water, clattering dishes
Vacuuming	whine of cleaner, footsteps
Dishwashing	running water, clattering dishes
TV	voices, music, sound effects, cheering

$$S(i, j) = \frac{r(i, j)}{\sqrt{r(i, i)r(j, j)}} \tag{10}$$

The figure indicates that the SC can represent the relative positions of sound sources without using the positions of microphones. For comparison, the spatial correlation between channels by using the amplitude vectors without taking the logarithm (left) and the spatial feature representation of each sound source by the principal component (right) are shown in Fig. 3. In this case, it appears that the relative position is not represented accurately, and thus the distance between sound sources #1 and #2-6 is tended to underestimate compared to the distance in real space. This indicates that taking the logarithm is reasonable for extracting the spatial pattern in analogy with a spectrum feature.

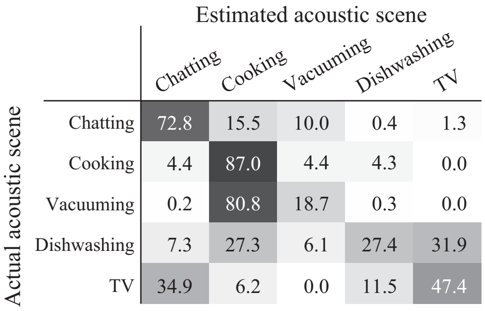
3.2. Acoustic scene recognition using real-life environmental sounds

We evaluated the performance of the SC by acoustic scene recognition using a real-life environmental sound dataset recorded in a living room. Twelve microphones were distributed as shown in Fig. 4 and the recorded sounds were roughly synchronized using a trigger sound. The dataset consists of 52.1 min. of recordings that involve five categories of acoustic scenes: "chatting," "cooking," "vacuuming," "washing dishes," and "watching TV," which are manually labeled.



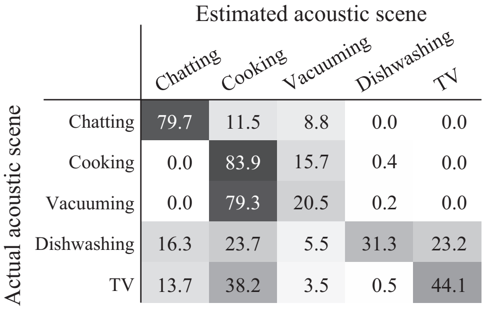
Actual acoustic scene \ Estimated acoustic scene	Chatting	Cooking	Vacuuming	Dishwashing	TV
Chatting	73.4	19.2	3.9	0.2	3.3
Cooking	8.6	54.1	26.0	2.7	8.6
Vacuuming	0.0	86.9	13.1	0.0	0.0
Dishwashing	14.5	13.3	3.7	46.4	22.1
TV	22.0	9.9	0.4	3.0	64.7

Fig. 5. Acoustic scene estimation accuracy with 1024-dimensional cepstrum feature extracted by using largest sound pressure over channels in terms of recall (%)



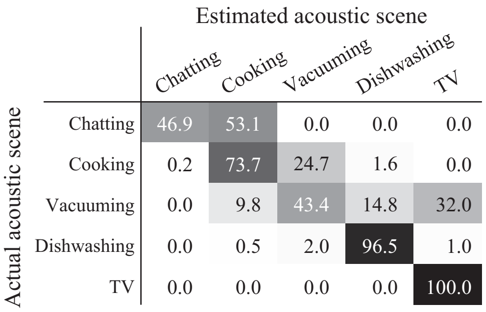
Actual acoustic scene \ Estimated acoustic scene	Chatting	Cooking	Vacuuming	Dishwashing	TV
Chatting	72.8	15.5	10.0	0.4	1.3
Cooking	4.4	87.0	4.4	4.3	0.0
Vacuuming	0.2	80.8	18.7	0.3	0.0
Dishwashing	7.3	27.3	6.1	27.4	31.9
TV	34.9	6.2	0.0	11.5	47.4

Fig. 6. Acoustic scene estimation accuracy with 1024-dimensional cepstrum feature extracted by using averaged sound pressure over channels in terms of recall (%)



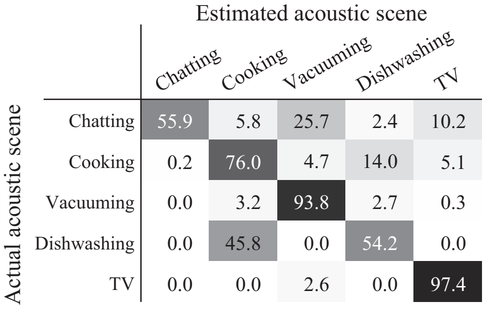
Actual acoustic scene \ Estimated acoustic scene	Chatting	Cooking	Vacuuming	Dishwashing	TV
Chatting	79.7	11.5	8.8	0.0	0.0
Cooking	0.0	83.9	15.7	0.4	0.0
Vacuuming	0.0	79.3	20.5	0.2	0.0
Dishwashing	16.3	23.7	5.5	31.3	23.2
TV	13.7	38.2	3.5	0.5	44.1

Fig. 7. Acoustic scene estimation accuracy with 12-dimensional MFCCs extracted by using largest sound pressure over channels in terms of recall (%)



Actual acoustic scene \ Estimated acoustic scene	Chatting	Cooking	Vacuuming	Dishwashing	TV
Chatting	46.9	53.1	0.0	0.0	0.0
Cooking	0.2	73.7	24.7	1.6	0.0
Vacuuming	0.0	9.8	43.4	14.8	32.0
Dishwashing	0.0	0.5	2.0	96.5	1.0
TV	0.0	0.0	0.0	0.0	100.0

Fig. 8. Acoustic scene estimation accuracy with 12-dimensional SC in terms of recall (%)



Actual acoustic scene \ Estimated acoustic scene	Chatting	Cooking	Vacuuming	Dishwashing	TV
Chatting	55.9	5.8	25.7	2.4	10.2
Cooking	0.2	76.0	4.7	14.0	5.1
Vacuuming	0.0	3.2	93.8	2.7	0.3
Dishwashing	0.0	45.8	0.0	54.2	0.0
TV	0.0	0.0	2.6	0.0	97.4

Fig. 9. Acoustic scene estimation accuracy with three-dimensional SC in terms of recall (%)

Table 2. Average estimation accuracy in terms of F-score (%) and feature dimension

Method	Feature dim.	Average F-score
Cepstrum (maximum)	1024	51.3%
Cepstrum (averaged)	1024	52.0%
MFCC (maximum)	12	55.3%
Spatial cepstrum	12	70.7%
Spatial cepstrum	3	70.1%

We randomly separated the sound dataset into 9,333 sound clips for training and 3,162 sound clips for evaluation. The sampling frequency was 48,000 Hz. Each acoustic scene typically included the sounds listed in Table 1, and none of the acoustic scene overlapped with each other in all recordings.

The cepstrum, MFCCs, and SC were extracted from the recorded sounds with respect to each sound clip. Acoustic scenes were then modeled and recognized using Gaussian mixture models (GMMs). For calculating the cepstrum and MFCCs, 1) selecting the largest amplitude and 2) averaging the amplitudes over all channels were investigated. Then a 2,048 point FFT was applied to each sound clip and we obtained 1,024-dimensional cepstrum features and 12-dimensional MFCCs. After calculating the cepstrum and MFCCs, we applied CMN with respect to each channel. Each acoustic scene was modeled by eight Gaussian components with diagonal covariance.

Figures 5, 6, 7, 8, and 9 show confusion matrices of acoustic scene recognition accuracy in terms of recall. These results indicate that the spatial pattern extracted by the SC enables the acoustic scene to be recognized effectively as well as when using the cepstrum or MFCC. The results also indicate that the SC is robust in the case of acoustic scenes involving movement such as "vacuuming" and "cooking." The average F-scores and the feature dimensions are listed in Table 2. The experimental results indicate that acoustic scenes can be recognized precisely even when an SC with a smaller feature dimension than the original one is used. From these results, we conclude that the proposed method achieves effective and efficient acoustic scene analysis by using distributed microphone signals.

4. CONCLUSION AND FUTURE WORK

We proposed a robust and efficient method for extracting the spatial information provided by a distributed microphone array. Inspired by the cepstrum, we defined the spatial cepstrum by PCA of the log amplitudes in multichannel observation and showed that it can be equivalent to the original definition of the cepstrum in a special case. Experimental results using real-life environmental sounds indicated that the SC is efficient for recognizing acoustic scenes precisely even when an SC with a smaller feature dimension is used. In future work, we will combine the spatial and spectral features and evaluate the performance of acoustic scene analysis.

5. ACKNOWLEDGEMENTS

This work was supported by a Grant-in-Aid for Scientific Research (B) (Japan Society for the Promotion of Science (JSPS) KAKENHI Grant Number 25280069).

REFERENCES

[1] D. Giannoulis, E. Benetos, D. Stowell, M. Rossignol, M. Lagrange, and M. D. Plumbley, "Detection and classification of acoustic scenes and events: an *ieee* aasp challenge," *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* (*WASPAA*), 2013.

[2] Y. Ohishi, D. Mochihashi, T. Matsui, M. Nakano, H. Kameoka, T. Izumitani, and K. Kashino, "Bayesian semi-supervised audio event transcription based on Markov Indian buffet process," *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing* (*ICASSP*), pp. 3163–3167, 2013.

[3] Y. Peng, C. Lin, M. Sun, and K. Tsai, "Healthcare audio event classification using hidden Markov models and hierarchical hidden markov models," *Proc. IEEE Int. Conf. on Multimedia and Expo* (*ICME*), pp. 1218–1221, 2009.

[4] A. Harma, M. F. McKinney, and J. Skowronek, "Automatic surveillance of the acoustic activity in our living environment," *Proc. IEEE Int. Conf. on Multimedia and Expo* (*ICME*), 2005.

[5] A. Eronen, V. T. Peltonen, J. T. Tuomi, A. P. Klapuri, S. Fagerlund, T. Sorsa, G. Lorho, and J. Huopaniemi, "Audio-based context recognition," *IEEE Trans. Audio Speech Lang. Process.*, pp. 321–329, 2006.

[6] A. Mesaros, T. Heittola, A. Eronen, and T. Virtanen, "Acoustic event detection in real life recordings," *Proc. 18th European Signal Processing Conf.* (*EUSIPCO*), pp. 1267–1271, 2010.

[7] C. V. Cotton and D. P. Ellis, "Spectral vs. spectro-temporal features for acoustic event detection," *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* (*WASPAA*), pp. 69–72, 2011.

[8] S. Chu, S. Narayanan, and C. C. Kuo, "Environmental sound recognition with time-frequency audio features," *IEEE Trans. Audio Speech Lang. Process.*, vol. 17, pp. 1142–1158, 2009.

[9] S. Kim, S. Narayanan, and S. Sundaram, "Acoustic topic models for audio information retrieval," *in Proc. 2009 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* (*WASPAA 2009*), pp. 37–40, 2009.

[10] K. Imoto, Y. Ohishi, H. Uematsu, and H. Ohmuro, "Acoustic scene analysis based on latent acoustic topic and event allocation," *Proc. IEEE Int. Workshop on Machine Learning for Signal Processing* (*MLSP*), 2013.

[11] K. Imoto and N. Ono, "Acoustic scene analysis from acoustic event sequence with intermittent missing event," *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing* (*ICASSP*), pp. 156–160, 2015.

[12] S. Kim, S. Narayanan, and S. Sundaram, "Non-negative matrix factorisation applied to auditory scenes classification," *ATIAM (ParisTech)*, 2011.

[13] H. Shimizu, N. Ono, K. Matsumoto, and S. Sagayama, "Isotropic noise suppression in the power spectrum domain by symmetric microphone arrays," *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* (*WASPAA*), pp. 54–57, 2007.

[14] N. Ito, H. Shimizu, N. Ono, and S. Sagayama, "Diffuse noise suppression using crystal-shaped microphone arrays," *IEEE Trans. Audio Speech Lang. Process.*, pp. 2101–2110, 2011.

[15] F. H. Liu, R. M. Stern, X. Huang, and A. Acero, "Efficient cepstral normalization for robust speech recognition," *Proc. Workshop on Human Language Technology*, pp. 69–74, 1993.

[16] O. Viilli and K. Laurila, "Cepstral domain segmental feature vector normalization for noise robust speech recognition," *Speech Communication*, pp. 133–147, 1998.