ONLINE ENVIRONMENTAL ADAPTATION OF CNN-BASED ACOUSTIC MODELS USING SPATIAL DIFFUSENESS FEATURES

Christian Huemmer[1*], *Marc Delcroix*[1], *Atsunori Ogawa*[1], *Keisuke Kinoshita*[1], *Tomohiro Nakatani*[1], *Walter Kellermann*[2]

[1]NTT Communication Science Laboratories, NTT Corporation, Kyoto, Japan
[2]Multimedia Communications and Signal Processing, FAU Erlangen-Nuremberg, Erlangen, Germany
{huemmer,wk}@lnt.de, {marc.delcroix,ogawa.atsunori,kinoshita.k,nakatani.tomohiro}@lab.ntt.co.jp

ABSTRACT

We propose a new concept for adapting CNN-based acoustic models using spatial diffuseness features as auxiliary information about the acoustic environment: the spatial diffuseness features are simultaneously employed as acoustic-model input features and to estimate environmental cues for context adaptation, where one convolutional layer is factorized into several sub-layers to represent different acoustic conditions. This context-adaptive CNN-based acoustic model facilitates an online environmental adaptation and is experimentally verified for the real-world recordings provided by the CHiME-3 task. The best performing setup reduces the average word error rate scores achieved by the baseline system (without using spatial diffuseness features) from 19.4% to 15.9% and 12.2% to 10.7% considering two experimental setups with and without front-end signal enhancement, respectively.

Index Terms— Context adaptation, spatial diffuseness features, CNN-based acoustic model, environmental robustness

1. INTRODUCTION

Since deep learning has become an essential part of designing modern speech recognition systems [1], the mismatch between training and test conditions (e.g., produced by environmental distortions) motivated various concepts for robust automatic speech recognition (ASR). For instance, front-end processing techniques reduce the environmental variability of the recorded signal [2] and can be combined with uncertainty decoding to account for missing information by modeling acoustic features as random variables [3–5]. Furthermore, back-end techniques increase the robustness of ASR systems by adapting all or a subset of acoustic-model parameters. This has been realized without [6–11] or with exploiting auxiliary information [12–18], where the latter has been shown to be especially appealing, as it facilitates an unsupervised adaptation on a small amount of speech data. Following this strategy, auxiliary features have been used as additional input features [12–14], to estimate acoustic-model parameters [15, 16], and to adjust the output of the acoustic model [17, 18].

In this work, we exploit spatial diffuseness features [14] for adapting a convolutional neural network (CNN)-based acoustic model. The spatial diffuseness features have been investigated for deep neural network (DNN)-based ASR systems in [14] and can be extracted online from multi-channel speech recordings: we estimate the diffuseness from the recorded microphone signals in the short-time Fourier transform (STFT) domain and perform a Mel weighting similar to the filterbank feature extraction. This facilitates an online acoustic-model adaptation, which is not possible using auxiliary features like i-vectors [12, 15, 19] or per-utterance noise estimates [13, 17]. Furthermore, the spatial diffuseness features exhibit a similar time-frequency structure as the filterbank features. This is especially appealing for CNN-based acoustic models with filterbank features as inputs, which have been shown to be powerful tools for noise-robust ASR tasks by extracting local information from time-frequency patterns [20, 21].

The diffuseness is a characteristic of the sound field, which can be extracted by assuming background noise and late reverberation to be modeled by a diffuse (spherically isotropic) noise field. Accordingly, spatial diffuseness features provide information about the acoustic environment which is exploited in this paper to combine two concepts for acoustic-model adaptation. First, the spatial diffuseness features are appended to the filterbank features at the input of the acoustic model. This is motivated by the behavior of the human auditory system to exploit environmental information as an indicator for the energy of the desired speech [22]. Second, we extract environmental cues from the spatial diffuseness features to adapt more complex feature representations in upper layers of a CNN-based acoustic model: following the concept of context adaptation [23], one convolutional layer is factorized into several sub-layers representing different acoustic conditions. Each sub-layer is associated with an acoustic context class posterior probability estimated by transforming spatial diffuseness features through a small neural network [23].

The proposed context-adaptive CNN-based acoustic model is experimentally verified using real recordings provided by the 3rd CHiME speech separation and recognition challenge (CHiME-3). In more details, we compare the recognition accuracy achieved with and without incorporating acoustic front-end signal enhancement. The experimental results highlight that, employing spatial diffuseness features as additional input features and at the same time for context adaptation, consistently improves the recognition accuracy of the baseline CNN-based ASR system.

In the remainder of this paper, we propose the context-adaptive CNN-based acoustic model in Section 2, illustrate relations to prior work in Section 3 and show experimental results for the CHiME-3 task in Section 4. An outlook to future work in Section 5 is followed by concluding remarks in Section 6.

2. CONTEXT-ADAPTIVE CNN-BASED ACOUSTIC MODEL

In this section, we provide details about the proposed context-adaptive CNN-based acoustic model in Fig. 1 including feature extraction (Section 2.1), acoustic-model topology (Section 2.2) and the concept of context adaptation (Section 2.3).

*Christian Huemmer is with the Chair of Multimedia Communications and Signal Processing, FAU Erlangen-Nuremberg[2]. This work was done during his stay as a visiting researcher at NTT.

2.1. Acoustic features

We exploit filterbank and spatial diffuseness features illustrated for one speech recording in Fig. 2. As first part of both feature extractions, the microphone signals are transformed into the complex-valued STFT domain using a Hamming window of length 25 ms and a frame shift of 10 ms (DFT length 512).

Filterbank features: The magnitude of a single-channel STFT-domain signal $X(t, f)$ at time t and frequency f is weighted by a Mel filterbank (using 80 Mel bands) and a logarithmic function. This is realized for a reference microphone or the output of an acoustic front-end signal enhancement scheme in Section 4.

Diffuseness features [24]: Two STFT-domain microphone signals $X_1(t, f)$ and $X_2(t, f)$ are modeled to be produced by a superposition of plane waves emitted by the desired speaker and (spherically isotropic) diffuse noise capturing background noise and late reverberation. Based on this signal model, the spatial coherence function of the diffuse noise is given as

$$\Gamma_{\mathrm{n}}(f) = \frac{\sin(2\pi f \frac{d}{c})}{2\pi f \frac{d}{c}}, \tag{1}$$

where d is the microphone distance and c is the speed of sound. To extract spatial diffuseness features, we estimate the auto- and cross-power spectra of the microphone signals

$$\hat{\Phi}_{x_p x_q}(t, f) = \lambda \hat{\Phi}_{x_p x_q}(t, f) + (1 - \lambda) X_p(t, f) X_q^*(t, f), \tag{2}$$

where $q = 1, 2$, $p = 1, 2$ and $\lambda = 0.68$ [24]. This is followed by calculating the spatial coherence

$$\hat{\Gamma}_{\mathrm{x}}(t, f) = \frac{\hat{\Phi}_{x_1 x_2}(t, f)}{\sqrt{\hat{\Phi}_{x_1 x_1}(t, f) \hat{\Phi}_{x_2 x_2}(t, f)}}, \tag{3}$$

which is used to estimate the coherent-to-diffuse power ratio (CDR)

$$\mathrm{CDR}(t, f) = g\left(\Gamma_{\mathrm{n}}(f), \hat{\Gamma}_{\mathrm{x}}(t, f)\right). \tag{4}$$

Various CDR estimators $g(\cdot)$ have been proposed in the literature (see overview article [24]), where we choose (25) in [24] as it has been shown to be unbiased and not requiring direction of arrival (DOA) information. As the CDR is equivalent to a signal-to-noise ratio for a spatial signal model, we reduce the dynamic range by performing a mapping between 0 and 1. This leads to the diffuseness

$$\mathrm{DIFF}(t, f) = (1 + \mathrm{CDR}(t, f))^{-1},$$

which is weighted by a Mel filterbank (80 Mel bands) to determine the spatial diffuseness features. Note that this 2-channel feature extraction is adapted to a 6-channel scenario in Section 4 by selecting a reference microphone and averaging the time-frequency dependent diffuseness (2.1) estimated for 5 different microphone pairs.

It should be highlighted that filterbank and spatial diffuseness features are extracted online using the same Mel weighting and that each spatial diffuseness feature thus provides additional environmental information to a corresponding filterbank feature [24].

2.2. CNN-based acoustic model

As inputs of the CNN-based acoustic model in Fig. 1, we create feature maps of size 19×80 by appending the 80 filterbank and spatial diffuseness features of the current time frame with the respective features from 9 previous and successive time frames.

In the following, we provide details about the CNN-based acoustic model in Fig. 1. Note that the topology and parameters of this acoustic model have been optimized for the CHiME-3 task using filterbank features at the input. By doing so, we provide a strong baseline ASR system for the experimental evaluation in Section 4.
A convolutional layer transforms N input feature maps to M output feature maps using $N \cdot M$ matrices $\mathbf{W}_{n,m}$ of size $P \times Q$ and M bias values b_m, where $m = 1, ..., M$ and $n = 1, ..., N$. As a consequence of this, each convolutional layer is characterized by the set of parameters $\{N, M, P, Q\}$ and denoted as

Conv$(N, M, P \times Q)$.

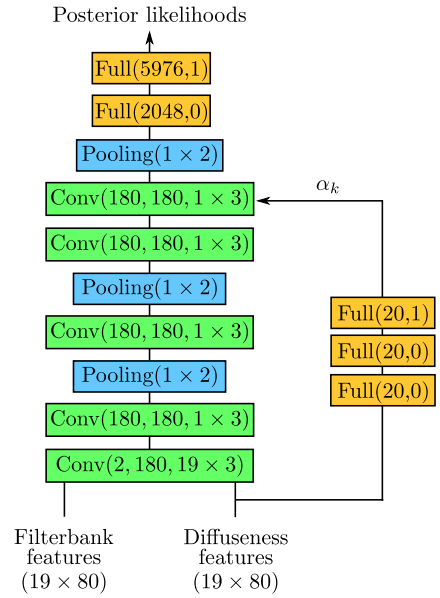

Fig. 1: Overview of the CNN-based neural network with context adaptation in the fifth convolutional layer.

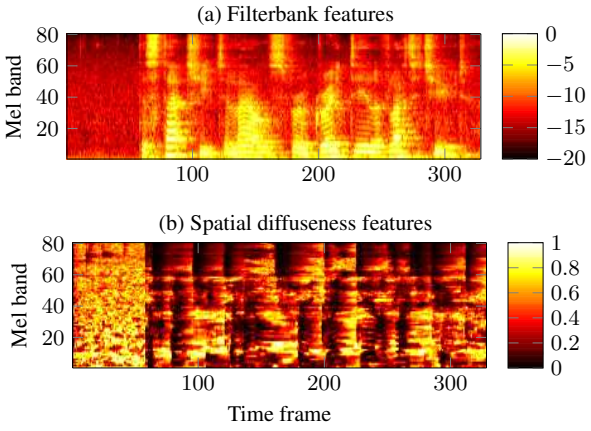

Fig. 2: Acoustic features for one speech utterance recorded with a 2-channel microphone array in a noisy and reverberant environment (Filterbank features in (a) extracted from one microphone signal).

As illustrated for the special case of Conv$(2, 2, 2 \times 5)$ in Fig. 3, one element y_m of the mth output feature map is calculated as follows:

$$y_m = \sigma(z_m), \quad z_m = \sum_{n=1}^{N} \langle \mathbf{W}_{n,m}, \mathbf{X}_n \rangle_F + b_m, \tag{5}$$

where $\mathbf{X}_n$ is a local input patch of size $P \times Q$ and $\sigma(\cdot)$ denotes the sigmoid function. Furthermore, the Frobenius inner product $\langle \cdot, \cdot \rangle_F$ in (5) sums up the element-wise product of two equally-sized matrices and thus realizes one instant of the two-dimensional convolution operator for the current input patch $\mathbf{X}_n$ (shifted during convolution). As illustrated in Fig. 1, the CNN-based acoustic model additionally contains pooling layers

$$\textit{Pooling}(P \times Q)$$

to extract the maximum value of local input patches of size $P \times Q$. Furthermore, fully-connected layers

$$\textit{Full}(L, R)$$

include an affine transformation and L hidden nodes with sigmoid $(R = 0)$ or softmax $(R = 1)$ nonlinearities. As outputs of the CNN-based acoustic model in Fig. 1, we estimate the posterior likelihoods of 5976 context-dependent HMM states.

2.3. Context-adaptive convolutional layer

The fundamental idea of context adaptation is to factorize one or several layers of the acoustic model into K sub-layers representing different speaker or environmental characteristics. In more details, the context-adaptive convolutional layer consists of K sets of weight matrices and bias terms $\{\mathbf{W}_{n,m,k}, b_{m,k}\}$, where $k = 1, ..., K$, to calculate the input of the sigmoid function in (5) following

$$z_m = \sum_{k=1}^{K} \alpha_k \left(\sum_{n=1}^{N} \langle \mathbf{W}_{n,m,k}, \mathbf{X}_n \rangle_F + b_{m,k} \right). \tag{6}$$

Note that this can be rewritten into (5) by exploiting the linearity of the operators in (6) and the substitution

$$\mathbf{W}_{n,m} = \sum_{k=1}^{K} \alpha_k \mathbf{W}_{n,m,k} \ , \quad b_m = \sum_{k=0}^{K-1} \alpha_k b_{m,k}. \tag{7}$$

As a consequence, the factorization leads to one convolutional layer depending on the factor class posterior probabilities α_k. To estimate the weights α_k, we transform the spatial diffuseness features of the current context window (± 9 time frames) through a neural network with 3 fully-connected layers of 20 nodes (this topology has been optimized for the CHiME-3 task). The resulting context-adaptive CNN-based acoustic model is shown in Fig. 1.

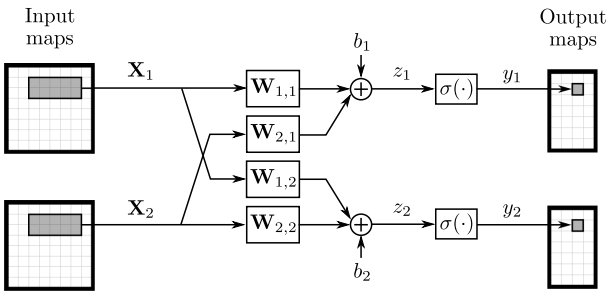

Fig. 3: Convolutional layer *Conv.* $(2, 2, 2 \times 5)$.

3. RELATION TO PRIOR WORK

The spatial diffuseness features have been investigated for DNN-based acoustic models in [14] and can be extracted online from multi-channel speech recordings (see Section 2): the diffuseness is estimated in the STFT domain and weighted by the same Mel filterbank used for the filterbank feature extraction. This leads to a similar time-frequency structure of filterbank and spatial diffuseness features (compare Figs. 2a) and 2b)), which is well-suited for CNN-based acoustic models extracting local information from time-frequency patterns [20, 21]. The properties of the spatial diffuseness features are different compared to previously-used auxiliary features like i-vectors [12, 15, 19], speaker codes [25, 26], per-utterance noise estimates [13, 17] or bottleneck features [18, 27]. To the best of our knowledge, spatial diffuseness features have not been exploited for CNN-based acoustic models so far.

The concept of context-adaptation for CNN-based acoustic models has been proposed and compared to different factorization techniques in [23]. In contrast to this previous work, we focus on environmental robustness instead of speaker adaptation. Furthermore, using auxiliary features at the acoustic-model input and at the same time for context-adaptation has not been considered so far and thus represents a new concept for adapting CNN-based acoustic models.

4. EXPERIMENTS

4.1. Experimental setup

The experimental results are produced using a trigram language model and the CHiME-3 corpus consisting of simulated data and real recorded speech of different speakers talking to a tablet device in four different environments: "BUS" (bus), "CAF" (cafeteria), "PED" (pedestrian) and "STR" (street). The multi-condition training set consists of 1600 real and 7138 simulated utterances (~ 18 hours of speech). We perform the parameter optimization on the development set (3280 utterances) and realize the final scoring using the evaluation test set (2640 utterances).

Acoustic-model training: The first layer of the auxiliary network is initialized to perform an arithmetic averaging over 4 neighboring elements in each 80-dimensional feature vector and the context window of 19 frames. This initialization has shown to be beneficial for mapping the 19×80 diffuseness feature map to a very small number of 20 hidden nodes in the first layer of the auxiliary network. Besides this, the context-adaptive acoustic model in Fig. 1 is randomly initialized and directly fine-tuned (cross-entropy criterion) without pretraining. For this, we used a batch size of 128, a momentum of 0.9 and an initial learning rate of 0.08 which was gradually decreased in the 25 training epochs when the frame accuracy did not improve for the cross-validation set. Furthermore, dropout regularization was used for the fully connected layers at the output of the acoustic model in Fig. 1.

Front-end enhancement: We evaluate the recognition accuracy achieved by the proposed context-adaptive CNN-based acoustic model using filterbank features extracted from (a) a reference microphone signal or (b) the output of the acoustic front-end in [2], where the latter includes weighted prediction error (WPE)-based dereverberation and minimum variance distortionless response (MVDR) beamforming[1]. It should be mentioned that the results here are not directly comparable with our system submitted to the CHiME-3

[1]Note that the speech enhancement front-end we use in this paper is not using online processing.

Table 1: WER scores (in %) for the real recordings of the CHiME-3 development and evaluation test set obtained with the CNN-based acoustic model in Fig. 1. We compare the recognition accuracy achieved by exploiting spatial diffuseness features in addition to the filterbank features as acoustic-model inputs and at the same time as auxiliary information for context adaptation.

Input features	Context adaptation	Front-end	Dev set	Eval set				
			Avg.	**Avg.**	BUS	CAF	PED	STR
Filterbank features	-	-	11.5	**19.4**	26.8	20.7	16.1	14.2
Filterbank + Spatial diffuseness features	-	-	9.9	**16.6**	22.9	15.3	14.7	13.5
Filterbank + Spatial diffuseness features	✓	-	9.7	**15.9**	22.0	14.6	14.0	13.2
Filterbank features	-	✓	7.3	**12.2**	17.6	11.1	10.2	9.9
Filterbank + Spatial diffuseness features	-	✓	7.3	**11.4**	14.9	10.0	10.6	10.3
Filterbank + Spatial diffuseness features	✓	✓	7.1	**10.7**	14.3	9.5	9.6	9.3

challenge [2] because we use a simpler system that, e.g., does not use recurrent neural network (RNN)-based language-model rescoring or two-pass unsupervised acoustic-model adaptation.
For both baseline systems (with and without front-end speech enhancement), the diffuseness features are estimated online from the recorded microphone signals (see Section 2.1) and applied for performing the acoustic-model adaptation illustrated in Fig. 1. Note that we employ the same feature extraction (and thus acoustic front-end signal enhancement) during training and testing.

4.2. Experimental results

Besides the topology of the auxiliary network in Fig. 1, we optimized the parameters for the factorization, where the best performance was achieved using the uppermost convolutional layer for context adaptation (as shown in Fig. 1). Furthermore, we noticed that the recognition accuracy slightly improves when using $K = 3$ instead of $K = 2$ classes in (6).

Results without acoustic front-end: The word error rate (WER) scores achieved without front-end speech enhancement are shown in the upper part of Table 1. It is obvious that appending spatial diffuseness features to the filterbank features consistently improves the recognition accuracy of the CNN-based ASR system. Moreover, simultaneously performing context adaptation leads to further reductions of the WER scores. Note that incorporating spatial diffuseness features is especially effective in the closed environments of cafeteria and bus.

Results with acoustic front-end: The recognition accuracy achieved using front-end signal enhancement for extracting filterbank features is illustrated in the lower part of Table 1. We notice that exploiting spatial diffuseness features as additional input features leads to a reduction of the WER scores in the cafeteria and bus environment. Interestingly, the slightly decreased recognition accuracy for the pedestrian and street environment is compensated by simultaneously exploiting spatial diffuseness features for context adaptation. This confirms the proposed concept for adapting CNN-based acoustic models by exploiting spatial diffuseness features as acoustic-model inputs and at the same time for context adaptation.

Finally, it should be emphasized that the spatial diffuseness is a characteristic of the sound field [14]. As a consequence, the proposed context-adaptive CNN-based acoustic model in Fig. 1 can be applied for different microphone array geometries [28] and arrays consisting of directional microphones [29], as long as (at least) one additional microphone is placed in an appropriate distance to the reference microphone [24].

5. FUTURE WORK

The proposed context-adaptive CNN-based acoustic model has been experimentally verified using matched conditions between training and testing. It is of great interest to exploit spatial diffuseness features for ASR systems using training on unprocessed speech and front-end signal enhancement during decoding [2]. For this purpose, the impact of the acoustic front-end on the diffuse noise field has to be taken into account [30].

In this paper, we exploit the same context window of spatial diffuseness features as acoustic-model inputs and for context-adaptation. It appears to be promising for future work to exploit a larger temporal window for context adaptation, e.g., by including a recurrent layer into the auxiliary network of Fig. 1.

In our previous work on coherence-based spectral enhancement [31], we experienced that exploiting DOA-dependent instead of DOA-independent CDR estimation in (4) leads to an improved recognition accuracy of a DNN-based ASR system. It seems intuitive that combining the advantage of different CDR estimators further increases the effectiveness of spatial diffuseness features in improving the performance of CNN-based ASR systems.

Finally, it is of interest to incorporate a neutral cluster [32] into the factorization in (6) to further increase the robustness to unseen acoustic conditions.

6. CONCLUSIONS

This paper presents a new concept for the online environmental adaptation of CNN-based acoustic models: spatial diffuseness features are extracted online from the recorded microphone signals and simultaneously exploited as acoustic-model inputs and to perform context adaptation. Following the latter strategy, one convolutional layer is factorized into several sub-layers to represent different acoustic conditions. Each sub-layer is associated with a factor class posterior probability estimated by transforming spatial diffuseness feature through a small neural network. The proposed context-adaptive CNN-based acoustic model is experimentally verified using the real recordings of the CHiME-3 corpus. It is shown that exploiting spatial diffuseness features as auxiliary information at the acoustic-model input and for context adaptation consistently improves the recognition accuracy of a baseline ASR system with and without front-end signal enhancement.

7. REFERENCES

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

[2] T. Yoshioka, N. Ito, M. Delcroix, A. Ogawa, K. Kinoshita, M. Fujimoto, C. Yu, W.J. Fabian, M. Espi, T. Higuchi, and others, "The NTT CHiME-3 system: Advances in speech enhancement and recognition for mobile multi-microphone devices," in *Proc. ASRU*. 2015, pp. 436–443, IEEE.

[3] C. Huemmer, R. Maas, A. Schwarz, R. F. Astudillo, and W. Kellermann, "Uncertainty decoding for DNN-HMM hybrid systems based on numerical sampling," in *Proc. INTERSPEECH*. 2015, pp. 3556–3560, ISCA.

[4] A.H. Abdelaziz, S. Watanabe, J.R. Hershey, E. Vincent, and D. Kolossa, "Uncertainty propagation through deep neural networks," in *Proc. INTERSPEECH*. 2015, pp. 3561–3565, ISCA.

[5] C. Huemmer, A. Schwarz, R. Maas, H. Barfuß, R. F. Astudillo, and W. Kellermann, "A new uncertainty decoding scheme for DNN-HMM hybrid systems with multichannel speech enhancement," in *Proc. ICASSP*. 2016, pp. 5760–5764, IEEE.

[6] J. Neto, L. Almeida, M. Hochberg, C. Martins, L. Nunes, S. Renals, and T. Robinson, "Speaker-adaptation for hybrid HMM-ANN continuous speech recognition system," in *Proc. EUROSPEECH*. 1995, pp. 2171–2174, ISCA.

[7] R. Gemello, F. Mana, S. Scanzio, P. Laface, and R. de Mori, "Adaptation of hybrid ANN/HMM models using linear hidden transformations and conservative training," in *Proc. ICASSP*. 2006, pp. 1189–1192, IEEE.

[8] F. Seide, G. Li, X. Chen, and D. Yu, "Feature engineering in context-dependent deep neural networks for conversational speech transcription," in *Proc. ASRU*. 2011, pp. 24–29, IEEE.

[9] K. Yao, D. Yu, F. Seide, H. Su, L. Deng, and Y. Gong, "Adaptation of context-dependent deep neural networks for automatic speech recognition," in *Proc. SLT*. 2012, pp. 366–369, IEEE.

[10] D. Yu, K. Yao, H. Su, G. Li, and F. Seide, "KL-divergence regularized deep neural network adaptation for improved large vocabulary speech recognition," in *Proc. ICASSP*. 2013, pp. 7893–7897, IEEE.

[11] P. Swietojanski and S. Renals, "Learning hidden unit contributions for unsupervised speaker adaptation of neural network acoustic models," in *Proc. SLT*. 2014, IEEE.

[12] G. Saon, H. Soltau, D. Nahamoo, and M. Picheny, "Speaker adaptation of neural network acoustic models using i-vectors," in *Proc. ASRU*. 2013, pp. 55–59, IEEE.

[13] M.L. Seltzer, D. Yu, and Y. Wang, "An investigation of deep neural networks for noise robust speech recognition," in *Proc. ICASSP*. 2013, pp. 7398–7402, IEEE.

[14] A. Schwarz, C. Huemmer, R. Maas, and W. Kellermann, "Spatial diffuseness features for DNN-based speech recognition in noisy and reverberant environments," in *Proc. ICASSP*. 2014, pp. 4380–4384, IEEE.

[15] Y. Miao, H. Zhang, and F. Metze, "Speaker adaptive training of deep neural network acoustic models using i-vectors," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 11, pp. 1938–1949, 2015.

[16] M. Delcroix, K. Kinoshita, T. Hori, and T. Nakatani, "Context adaptive deep neural networks for fast acoustic model adaptation," in *Proc. ICASSP*. 2015, pp. 4535–4539, IEEE.

[17] J. Li, J.-T. Huang, and Y. Gong, "Factorized adaptation for deep neural network," in *Proc. ICASSP*. 2014, pp. 5537–5541, IEEE.

[18] Y. Qian, T. Tan, D. Yu, and Y. Zhang, "Integrated adaptation with multi-factor joint-learning for far-field speech recognition," in *Proc. ICASSP*. 2016, pp. 5770–5774, IEEE.

[19] O. Abdel-Hamid and H. Jiang, "Rapid and effective speaker adaptation of convolutional neural network based models for speech recognition," in *Proc. INTERSPEECH*. 2013, pp. 1248–1252, ISCA.

[20] T.N. Sainath, B. Kingsbury, G. Saon, H. Soltau, A. Mohamed, G.E. Dahl, and Ramabhadran B., "Deep convolutional neural networks for large-scale speech tasks," *IEEE Signal Process. Mag.*, vol. 64, pp. 39–48, 2015.

[21] T. Yoshioka, K. Ohnishi, F. Fang, and T. Nakatani, "Noise robust speech recognition using recent developments in neural networks for computer vision," in *Proc. ICASSP*. 2016, pp. 5730–5734, IEEE.

[22] J.F. Culling, B.A. Edmonds, and K.I. Hodder, "Speech perception from monaural and binaural information," *Journal Acoustical Society America*, vol. 119, no. 1, pp. 559–565, 2006.

[23] M. Delcroix, K. Kinoshita, T. Hori, and T. Nakatani, "Context adaptive neural networks for rapid adaptation of deep CNN based acoustic models," in *INTERSPEECH*. 2016, pp. 1573–1577, ISCA.

[24] A. Schwarz and W. Kellermann, "Coherent-to-diffuse power ratio estimation for dereverberation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 6, pp. 1006–1018, 2015.

[25] O. Abdel-Hamid and H. Jiang, "Fast speaker adaptation of hybrid NN/HMM model for speech recognition based on discriminative learning of speaker code," in *Proc. ICASSP*. 2013, pp. 7942–7946, IEEE.

[26] S. Xue, O. Abdel-Hamid, H. Jiang, and L. Dai, "Direct adaptation of hybrid DNN/HMM model for fast speaker adaptation in LVCSR based on speaker code," in *Proc. ICASSP*. 2014, pp. 6339–6343, IEEE.

[27] Y. Miao, L. Jiang, H. Zhang, and F. Metze, "Improvements to speaker adaptive training of deep neural networks," in *Proc. SLT*. 2014, pp. 165–170, IEEE.

[28] D.P. Jarrett, O. Thiergart, E.A.P. Habets, and P.A. Naylor, "Coherence-based diffuseness estimation in the spherical harmonic domain," in *Proc. IEEEI*. 2012, pp. 1–5, IEEE.

[29] O. Thiergart, T. Ascherl, and E. A. P. Habets, "Power-based signal-to-diffuse ratio estimation using noisy directional microphones," in *Proc. ICASSP*. 2014, pp. 7440–7444, IEEE.

[30] K.U. Simmer, J. Bitzer, and C. Marro, "Post-filtering techniques," in *Microphone Arrays*, Digital Signal Processing, pp. 39–60. Springer Berlin Heidelberg, Jan. 2001.

[31] H. Barfuss, C. Huemmer, A. Schwarz, and W. Kellermann, "Robust coherence-based spectral enhancement for speech recognition in adverse real-world environments," preprint available: http://arxiv.org/pdf/1604.03393v2.pdf, 2016.

[32] T. Tan, Y. Qian, M. Yin, Y. Zhuang, and K. Yu, "Cluster adaptive training for deep neural network," in *Proc. ICASSP*. 2015, pp. 4325–4329, IEEE.