A NEW UNCERTAINTY DECODING SCHEME FOR DNN-HMM HYBRID SYSTEMS WITH MULTICHANNEL SPEECH ENHANCEMENT

Christian Huemmer[1]*, Andreas Schwarz*[1]*, Roland Maas*[1]*, Hendrik Barfuss*[1]
Ramón Fernández Astudillo[2]*, and Walter Kellermann*[1]

[1]Multimedia Communications and Signal Processing,
University of Erlangen-Nuremberg, Erlangen, Germany
[2]Spoken Language Systems Laboratory, INESC-ID-Lisboa, Lisboa, Portugal
{huemmer,schwarz,maas,barfuss,wk}@lnt.de, ramon@astudillo.com

ABSTRACT

Uncertainty decoding combines a probabilistic feature description with the acoustic model of a speech recognition system. For DNN-HMM hybrid systems, this can be realized by averaging the DNN outputs produced by a finite set of feature samples (drawn from an estimated probability distribution). In this article, we employ this sampling approach in combination with a multi-microphone speech enhancement system. We propose a new strategy for generating feature samples from multichannel signals, based on modeling the spatial coherence estimates between different microphone pairs as realizations of a latent random variable. From each coherence estimate, a spectral enhancement gain is computed and an enhanced feature vector is obtained, thus producing a finite set of feature samples, of which we average the respective DNN outputs. In the experimental part, this new uncertainty decoding strategy is shown to consistently improve the recognition accuracy of a DNN-HMM hybrid system for the 8-channel REVERB Challenge task.

Index Terms— uncertainty decoding, multichannel speech enhancement, DNN-based acoustic model

1. INTRODUCTION

In recent years, deep neural networks (DNNs) have emerged as an effective method for discriminating between context-dependent phonetic units in acoustic models of state-of-the-art automatic speech recognition (ASR) systems. For instance, DNN-HMM hybrid systems include one DNN to directly map an observed feature vector to the posterior likelihoods of the context-dependent Hidden Markov Model (HMM) states [1]. Although DNN-based acoustic modeling has been shown to achieve remarkable results for different recognition tasks [2, 3], the ASR performance is still degraded by environmental distortions in adverse acoustic environments [4, 5]. This is why a variety of different adaptation schemes for environmental robustness have been proposed which either adapt the feature vectors (e.g., speech enhancement [6, 7, 8]) or the back-end parameters (e.g., DNN adaptation [9, 10, 11]).

Uncertainty decoding bridges front-end processing and back-end adaptation by combining a probabilistic feature description with the acoustic model of an ASR system. For DNN-based acoustic models, especially uncertainty decoding based on numerical sampling has been shown to be promising using the following strategy [12, 13]: The DNN outputs produced by a finite set of feature samples (drawn from an estimated probability distribution) are averaged to approximate the posterior likelihoods of the context-dependent HMM states.

As the main contribution of this paper, we propose a new strategy to generate feature samples by modeling estimated parameters as realizations of a latent random variable. To put this idea into practice, we consider a DNN-HMM hybrid system with multichannel speech enhancement (beamformer and postfilter) in the short-time Fourier transform (STFT) domain. The coherence estimates at several microphone pairs (used to calculate the postfilter gains) are modeled as realizations of the latent spatial coherence, so that we perform separate coherence-based postfilters (one for each microphone pair) and average the resulting DNN outputs. This new uncertainty decoding scheme accounts for the variance of the coherence estimation and is shown to consistently improve the recognition accuracy of a DNN-HMM hybrid system for the 8-channel REVERB Challenge task [14].

This article is structured as follows: First, the DNN-HMM hybrid system is introduced in Section 2 as a baseline system. After this, we propose the uncertainty decoding scheme and its application to the DNN-HMM hybrid system in Section 3. Finally, the experimental part (Section 4) is followed by concluding remarks (Section 5).

The authors would like to thank the Deutsche Forschungsgemeinschaft (contract number KE 890/4-2) and the Foundation for Science and Technology (project UID/CEC/50021/2013 and grant SFRH/BPD/68428/2010) for supporting this work.

R. Maas was with the University of Erlangen-Nuremberg while the work has been conducted. He is now with Amazon, Seattle, WA.

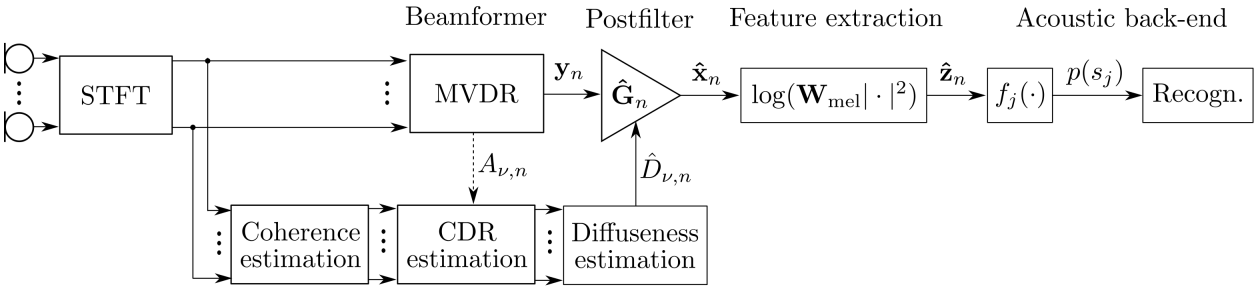

Fig. 1. DNN-HMM hybrid system with MVDR beamforming and coherence-based postfiltering in the STFT domain. The length-257 output vector of the postfilter $\hat{\mathbf{x}}_n$ is transformed into the logmelspec domain (vector $\hat{\mathbf{z}}_n$ of length 24) and passed through the nonlinear function $f_j(\cdot)$ to estimate the posterior likelihood $p(s_j)$ of the j-th context-dependent HMM state s_j.

2. DNN-HMM HYBRID SYSTEM WITHOUT UNCERTAINTY DECODING

In this part, we briefly introduce the DNN-HMM hybrid system which will be employed in Section 3 to integrate the proposed uncertainty decoding scheme. As shown in Fig. 1, the acoustic front-end is realized in the STFT domain by combining an MVDR beamformer (spatial filtering stage) with a coherence-based spectral enhancement method as postfilter (for dereverberation and noise suppression).

MVDR beamformer: As first part of the acoustic front-end, the STFT-domain microphone signals are processed by an MVDR beamformer to reduce background noise and reverberation. The MVDR beamformer design is based on the assumption of free-field sound-wave propagation and the constraint that a plane wave coming from the desired look direction (here estimated using the SRP-PHAT pseudo spectrum [15]) can pass the system without distortions [16].

Postfilter: As second front-end component, we realize a coherence-based postfilter with diagonal gain matrix (diag$\{\cdot\}$ creates a diagonal matrix)

$$\hat{\mathbf{G}}_n = \text{diag}\{(1-\hat{D}_{1,n}), ..., (1-\hat{D}_{257,n})\} \tag{1}$$

at time n producing the length-257 vector (DFT length 512)

$$\hat{\mathbf{x}}_n = [\hat{x}_{1,n}, ..., \hat{x}_{257,n}]^{\text{T}} \tag{2}$$

(with complex-valued coefficients $\hat{x}_{\nu,n}$, where $\nu = 1, ..., 257$) by a linear transformation of the input vector $\mathbf{y}_n$:

$$\hat{\mathbf{x}}_n = \hat{\mathbf{G}}_n \mathbf{y}_n. \tag{3}$$

The estimated diffuseness values $0 \leq \hat{D}_{\nu,n} \leq 1$ in (1) are determined as described in [17]: For each microphone pair, indexed by $l = 1, ..., L$, we estimate the time- and frequency-dependent complex-valued coherence $\Gamma_{\nu,n}^{(l)}$ (using auto- and cross-power spectra estimates, see (1) in [18]) and estimate the respective coherent-to-diffuse power ratio (CDR) $CDR_{\nu,n}^{(l)}$ by inserting the spatial coherence function of a diffuse (spherically isotropic) sound field

$$\Gamma_{\text{diff},\nu}^{(l)} = \text{sinc}(2\pi f_\nu d^{(l)}/c) \tag{4}$$

into (5), where Re$\{\cdot\}$ is the real part, c is the speed of sound, f_ν is the center frequency of the ν-th STFT band and $d^{(l)}$ is the microphone distance at the l-th microphone pair. Note that the index l has been omitted in (5) for notational convenience and that also other CDR estimates could be applied (see overview in [19]). To account for the impact of the beamformer on the diffuse noise component, we weight the CDR $CDR_{\nu,n}^{(l)}$ with a correction factor $A_{\nu,n}$ (see (3.28) in [20]) and derive L microphone-pair specific diffuseness estimates

$$D_{\nu,n}^{(l)} = (1 + A_{\nu,n}\ CDR_{\nu,n}^{(l)})^{-1}. \tag{6}$$

Finally, the averaged diffuseness values

$$\hat{D}_{\nu,n} = \frac{1}{L}\sum_{l=1}^{L} D_{\nu,n}^{(l)} \tag{7}$$

are then inserted into (1) to calculate the postfilter gain matrix $\hat{\mathbf{G}}_n$.

Feature extraction: As shown in Fig 1, the output of the Wiener filter is transformed into the lower-dimensional domain using the mel-filterbank matrix $\mathbf{W}_{\text{mel}}$ (of dimensions 24×257) and the natural logarithm $\log(\cdot)$:

$$\hat{\mathbf{z}}_n = \log\left(\mathbf{W}_{\text{mel}}|\hat{\mathbf{x}}_n|^2\right). \tag{8}$$

$$CDR_{\nu,n} = \frac{\Gamma_{\text{diff},\nu}\ \text{Re}\{\Gamma_{\nu,n}\} - |\Gamma_{\text{diff},\nu}|^2 - \sqrt{\Gamma_{\text{diff},\nu}^2\,\text{Re}\{\Gamma_{\nu,n}\}^2 - \Gamma_{\text{diff},\nu}^2\,|\Gamma_{\nu,n}|^2 + \Gamma_{\text{diff},\nu}^2 - 2\,\Gamma_{\text{diff},\nu}\,\text{Re}\{\Gamma_{\nu,n}\} + |\Gamma_{\nu,n}|^2}}{|\Gamma_{\nu,n}|^2 - 1} \tag{5}$$

Acoustic back-end: In our implementation, the nonlinear function $f_j(\cdot)$ in Fig. 1 captures per-utterance mean and variance normalization, dynamic extension (delta and acceleration coefficients), context extension (± 5 frame splicing) as well as the DNN (6 hidden layers with 2048 sigmoid activation functions, output layer with 3463 elements). Thus, the posterior likelihood $p(s_j)$ of the j-th context-dependent HMM state s_j is given by the nonlinear transformation of $\hat{\mathbf{z}}_n$:

$$p(s_j|\hat{\mathbf{z}}_n) = f_j(\hat{\mathbf{z}}_n). \tag{9}$$

3. DNN-HMM HYBRID SYSTEM WITH UNCERTAINTY DECODING

3.1. General idea of the uncertainty decoding scheme

Modeling acoustic features as random variables is a common way to account for missing information (e.g., caused by environmental distortions). Assuming the probability distribution $p(\mathbf{z}_n)$ of the feature vector $\mathbf{z}_n$ and its mathematical relation to the j-th DNN output $f_j(\mathbf{z}_n)$ to be known, uncertainty decoding combines the probabilistic feature description with a DNN-based acoustic model by estimating the posterior probability of the context-dependent HMM state s_j following [12]:

$$p(s_j) = \mathcal{E}\{f_j(\mathbf{z}_n)\}. \tag{10}$$

However, the nonlinear structure of $f_j(\mathbf{z}_n)$ (e.g., due to the DNN activation functions) precludes an (exact) closed-form solution of (10) and motivates two kinds of approximations: First, assuming the DNN node activations to be statistically independent an approximate solution can be found based on a piecewise function. However, the assumption of independence and the lack of a solution for the soft-max limits the accuracy of the approximation [13, 21]. Second, the mathematical expectation in (10) can be approximated using numerical sampling techniques. This is computationally more efficient than linearizing $f_j(\mathbf{z}_n)$, because reasonable improvements in the recognition accuracy can already be achieved for a small number of samples [12, 13]. To give an example for a numerical sampling strategy, consider the concept of random sampling which can be summarized as follows [12]: We draw a finite set of feature samples $\mathbf{z}_n^{(l)}$ ($l = 1, ..., L$) from the probability distribution $p(\mathbf{z}_n)$ and average the DNN outputs $f_j(\mathbf{z}_n^{(l)})$ to approximate $p(s_j)$:

$$p(s_j) \approx \frac{1}{L}\sum_{l=1}^{L} f_j(\mathbf{z}_n^{(l)}). \tag{11}$$

This sampling scheme has been shown to improve the recognition accuracy of DNN-HMM hybrid systems by using the posterior distribution of a single-channel Wiener filter for estimating the PDF $p(\mathbf{z}_n)$ [12].

In this article, we propose a new approach for applying the approximation in (11) with the goal to capture parameter estimation errors in DNN-HMM hybrid systems with multichannel speech enhancement (beamformer and postfilter): The coherence estimates between different microphone pairs are interpreted as realizations of the latent spatial coherence. Thus, we implement one coherence-based postfilter for each microphone pair to extract several feature vector realizations for applying the DNN-output averaging in (11). This new strategy replaces the averaging of diffuseness estimates following (7) by averaging the posterior likelihoods at the DNN output according to (11). An overview of the difference between DNN-HMM hybrid system with and without uncertainty decoding is shown in Fig. 2. Note that this uncertainty decoding scheme combines the variance of the parameter estimation with the acoustic model of the DNN-HMM hybrid system.

3.2. Incorporation into the DNN-HMM hybrid system

In this section, we incorporate the proposed uncertainty decoding scheme into the DNN-HMM hybrid system of Section 2. As fundamental idea, the microphone-pair specific coherence estimates $\Gamma_{\nu,n}^{(l)}$ are modeled to be realizations of the latent time- and frequency-dependent coherence. Instead of averaging the diffuseness estimates obtained from the coherence estimates as in (7), we realize L separate coherence-based postfilters:

$$\begin{aligned} \mathbf{x}_n^{(l)} &= \mathbf{G}_n^{(l)}\mathbf{y}_n \\ \mathbf{G}_n^{(l)} &= \text{diag}\{(1 - D_{1,n}^{(l)}), ..., (1 - D_{M,n}^{(l)})\}. \end{aligned} \tag{12}$$

The L obtained feature vectors

$$\mathbf{z}_n^{(l)} = \log\left(\mathbf{W}_{\text{mel}}|\mathbf{x}_n^{(l)}|^2\right), \tag{13}$$

are then used for applying the DNN-output averaging in (11).

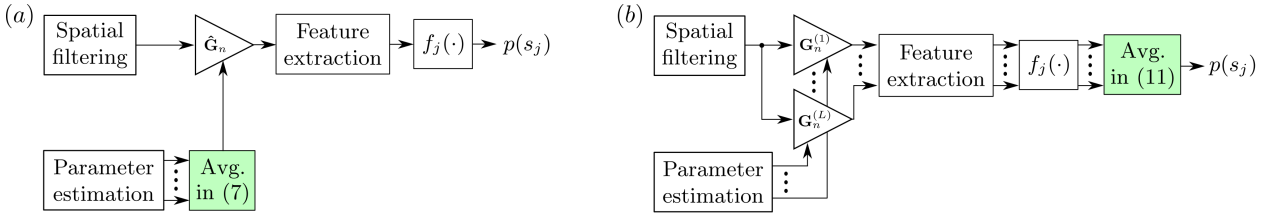

Fig. 2. Conceptional difference between the DNN-HMM hybrid system (a) without and (b) with uncertainty decoding.

Table 1. *WER scores for the REVERB challenge evaluation test set with DNN trained on clean data.*

	SimData							RealData		
	$T_{60} \approx 0.25$ s		$T_{60} \approx 0.5$ s		$T_{60} \approx 0.75$ s			$T_{60} \approx 0.7$ s		
	Near	Far	Near	Far	Near	Far	Avg.	Near	Far	Avg.
No PF, no uncertainty decod.	6.2	7.4	7.4	10.6	8.1	12.5	8.7	27.1	28.4	27.8
With PF, no uncertainty decod.	6.1	7.2	7.4	10.4	8.0	12.0	8.5	27.0	28.2	27.6
With PF and uncertainty decod.	**6.0**	**6.9**	**6.9**	**9.8**	**7.8**	**11.4**	**8.1**	**25.0**	**27.0**	**26.0**

4. EXPERIMENTS

The experimental verification of the proposed uncertainty decoding scheme is based on the 8-channel REVERB challenge task (circular microphone array with a diameter of 0.2 m) using the WSJ0 trigram 5k language model [14]. We employ the Kaldi toolkit [22] as ASR back-end system and train a GMM-HMM baseline system on the clean WSJCAM0 Cambridge Read News REVERB corpus [23] (details on the feature extraction in [12]) to determine the state-frame alignment employed for DNN training: A generative pretraining using the contrastive divergence algorithm (on restricted Boltzmann machines) is followed by discriminative fine-tuning using the mini-batch stochastic gradient descent approach (based on the cross-entropy criterion) [1]. We consider DNN training using the multi-condition training set (7861 utterances) provided by the REVERB challenge [14].

The evaluation test set consists of ~ 5000 environmentally-distorted utterances and is split into two categories: First, the utterances of the clean WSJCAM0 Cambridge Read News REVERB corpus are artificially corrupted ("SimData") using measured impulse responses ($T_{60} \approx 0.25$ s, 0.5 s and 0.7 s), recorded noise sequences (added to the microphones signals with 20 dB signal-to-noise ratio) and source-microphone spacings of 0.5 m ("Near") and 2 m ("Far"). Second, multichannel recordings ("RealData") in a reverberant ($T_{60} \approx 0.7$ s) and noisy environment are considered with source-microphone spacings of 1 m ("Near") and 2.5 m ("Far").

Table 1 shows the word error rate (WER) scores for the evaluation test set, where training and decoding were performed using the MVDR beamformer with ("With PF") or without ("No PF") coherence-based postfiltering. We observe that the postfilter improves the recognition accuracy of the DNN-HMM hybrid system and that incorporating the uncertainty decoding scheme (without retraining the DNN) leads to further reductions of the WER scores. Note that this performance gain is achieved for artificially corrupted evaluation data as well as for real recordings. Furthermore, the recognition accuracy consistently increases especially in scenarios with large speaker-microphone distances and for real-world recordings.

With respect to the computational complexity, it should be emphasized that the decoder remains unchanged besides the transformation of multiple feature samples through the non-linear function $f_j(\cdot)$. In our implementation, the posterior likelihoods at the DNN outputs are estimated in Matlab (using a GPU of type NVIDIA Tesla K20m). All remaining decoding steps are performed with the Kaldi Toolkit (4 parallel jobs running locally on an AMD Phenom II 1090T with 2.8 GHz). Based on this setup, the incorporation of the proposed uncertainty decoding scheme into the DNN-HMM hybrid system (with coherence-based postfilter) increases the average decoding time by only 28 %.

5. CONCLUSIONS

We proposed a new uncertainty decoding scheme for DNN-HMM hybrid systems with multichannel speech enhancement (beamformer and postfilter). To capture front-end estimation errors, the estimated spatial coherence values (used for calculating the postfilter gains) are considered as samples drawn from an unknown probability distribution. As a consequence, we realize one coherence-based postfilter for each microphone pair and thus produce a finite set of feature samples, of which we average the respective DNN outputs to approximate the posterior likelihoods of the context-dependent HMM states. This uncertainty decoding scheme is experimentally verified using a DNN-HMM hybrid system with MVDR beamforming and coherence-based postfiltering, where consistent improvements in the recognition accuracy could be achieved for the 8-channel REVERB Challenge task.

It should be emphasized that the proposed uncertainty decoding scheme is also applicable to various DNN-based speech recognition systems. On the one hand, other front-end parameters (e.g., the time-differences of arrival) can be modeled as realizations of a latent random variable. On the other hand, the transformation of the acoustic features to the posterior likelihoods of the HMM states is not restricted to the non-linear function considered in this article (e.g., one could also employ a convolutional DNN).

6. REFERENCES

[1] G. Hinton, L. Deng, D. Yu, A.-R. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath G. Dahl, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal Process. Magazine*, vol. 29, no. 6, pp. 82–97, Nov. 2012.

[2] G.E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Trans. Audio, Speech and Lang. Process.*, vol. 20, no. 1, pp. 30–42, Jan. 2012.

[3] A.-R. Mohamed, G. Hinton, and G. Penn, "Understanding how deep belief networks perform acoustic modelling," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Process. (ICASSP)*. 2012, pp. 4273–4276, IEEE.

[4] M. Delcroix, Y. Kubo, T. Nakatani, and A. Nakamura, "Is speech enhancement pre-processing still relevant when using deep neural networks for acoustic modeling?," in *INTERSPEECH*, Aug. 2013, pp. 2992–2996.

[5] T. Yoshioka and M.J.F. Gales, "Environmentally robust ASR front-end for deep neural network acoustic models," *Computer Speech Language*, vol. 31, no. 1, pp. 65–86, 2015.

[6] I. Cohen, "Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 5, pp. 466–475, Sep. 2003.

[7] A. Krueger and R. Haeb-Umbach, "Model-based feature enhancement for reverberant speech recognition," *IEEE Trans. Audio, Speech and Lang. Process.*, vol. 18, no. 7, pp. 1692–1707, Sep. 2010.

[8] M.J.F. Gales and Y.-Q. Wang, "Model-based approaches to handling additive noise in reverberant environments," in *Proc. Joint Workshop Hands-free Speech Comm. Microphone Arrays (HSCMA)*. 2011, pp. 121–126, IEEE.

[9] X. Li and J. Bilmes, "Regularized adaptation of discriminative classifiers," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Process. (ICASSP)*. 2006, vol. 1, pp. 237–240, IEEE.

[10] H. Liao, "Speaker adaptation of context dependent deep neural networks," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Process. (ICASSP)*. 2013, pp. 7947–7951, IEEE.

[11] 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. IEEE Int. Conf. Acoustics, Speech, Signal Process. (ICASSP)*. 2013, pp. 7893–7897, IEEE.

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

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

[14] K. Kinoshita, M. Delcroix, T. Yoshioka, T. Nakatani, A. Sehr, W. Kellermann, and R. Maas, "The REVERB challenge: A common evaluation framework for dereverberation and recognition of reverberant speech," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoustics (WASPAA)*. 2013, pp. 1–4, IEEE.

[15] C.H. Knapp and G.C. Carter, "The generalized correlation method for estimation of time delay," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 1, pp. 320–327, 1976.

[16] H.L. Van Trees, *Detection, Estimation, and Modulation Theory, Optimum Array Processing*, Detection, Estimation, and Modulation Theory. Wiley, 2004.

[17] A. Schwarz, C. Huemmer, R. Maas, and W. Kellermann, "Spatial diffuseness features for DNN-based speech recognition in noisy and reverberant environments," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Process. (ICASSP)*. 2014, pp. 4380–4384, IEEE.

[18] A. Schwarz and W. Kellermann, "Unbiased coherent-to-diffuse ratio estimation for dereverberation," in *Proc. Int. Workshop Acoustic Echo, Noise Control (IWAENC)*. 2014, pp. 6–10, IEEE.

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

[20] K.U. Simmer, J. Bitzer, and C. Marro, "Post-filtering techniques," in *Microphone Arrays*, M. Brandstein and D. Ward, Eds., pp. 39–60. Springer Berlin Heidelberg, Jan. 2001.

[21] R.F. Astudillo and J.P. Neto, "Propagation of uncertainty through multilayer perceptrons for robust automatic speech recognition," in *INTERSPEECH*, 2011, pp. 461–464.

[22] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, and others, "The Kaldi speech recognition toolkit," 2011.

[23] T. Robinson, J. Fransen, D. Pye, J. Foote, S. Renals, P. Woodland, and S. Young, "WSJCAM0 Cambridge read news for REVERB LDC2013E109," Web Download, 2013.