Discriminative uncertainty estimation for noise robust ASR

Dung Tien Tran, Emmanuel Vincent, Denis Jouvet

DISCRIMINATIVE UNCERTAINTY ESTIMATION FOR NOISE ROBUST ASR

Dung T. Tran[1,2,3]*, Emmanuel Vincent*[1,2,3]*, Denis Jouvet*[1,2,3]

[1]Inria, Villers-lès-Nancy, F-54600, France
[2]CNRS, LORIA, UMR 7503, Villers-lès-Nancy, F-54600, France
[3]Université de Lorraine, LORIA, UMR 7503, Villers-lès-Nancy, F-54600, France
dung.tran@inria.fr

ABSTRACT

We consider the problem of uncertainty estimation for noise-robust ASR. Existing uncertainty estimation techniques improve ASR accuracy but they still exhibit a gap compared to the use of oracle uncertainty. This comes partly from the highly non-linear feature transformation and from additional assumptions such as Gaussian distribution and independence between frequency bins in the spectral domain. In this paper, we propose a method to rescale the estimated feature-domain full uncertainty covariance matrix in a state-dependent fashion according to a discriminative criterion. The state-dependent and feature index-dependent scaling factors are learned from development data. Experimental evaluation on Track 1 of the 2nd CHiME challenge data shows that discriminative rescaling leads to better results than generative rescaling. Moreover, discriminative rescaling of the Wiener uncertainty estimator leads to 12% relative word error rate reduction compared to discriminative rescaling of the alternative estimator in [1].

***Index Terms*—** Automatic speech recognition, noise robustness, uncertainty handling, discriminative adaptation.

1. INTRODUCTION

In robust speech recognition, uncertainty decoding has attracted a lot of attention recently [2, 3]. The output of the speech enhancement pre-processor is modeled as a Gaussian distribution whose mean is the enhanced feature vector and whose covariance matrix represents the estimated distortion between the enhanced and the clean feature vectors. This uncertainty representation is then used as input to the recognizer. The features are more reliable when their uncertainty tends to be low. Conversely, the features are unreliable when their uncertainty tends to be high. The uncertainty is first computed in the spectral domain [4] then propagated into the feature domain. Because of the non-linear transform applied to the input spectral domain, propagation requires approximate methods such as Vector Taylor series (VTS) [5], moment matching [6] or unscented transform [7]. Due to this approximation and to other simplifying assumptions such as Gaussian distribution and spectral domain independence, the estimated feature uncertainty often underestimates the oracle uncertainty.

To overcome this, the estimated uncertainty can be rescaled by a linear transformation [4, 8–10]. In the past [4, 10], the scaling factors were optimized such that the recaled uncertainty estimates are close to the oracle estimates irrespectively of the resulting state hypotheses. This can be considered as a sub-optimal approach because the same scaling factors are applied to the correct state hypothesis and to the competing state hypotheses. Delcroix et al [8, 9] proposed to train the linear transformation according to a Maximum likelihood (ML) criterion instead. They applied this approach to a diagonal feature uncertainty matrix only and they showed significant improvement. Recently, maximum mutual information (MMI) [11] and boosted MMI (bMMI) [12, 13] were successfully employed for supervised discriminative adaptation of feature means and diagonal uncertainty matrices [1]. In this approach, the diagonal uncertainty matrix was estimated directly in the feature domain as the squared difference between noisy and enhanced features, which was shown to be a poor estimate of uncertainty [4].

In this paper, we propose a method for state-dependent and feature index-dependent discriminative rescaling of the full feature uncertainty covariance matrix. Instead of estimating the uncertainty directly in the feature domain, it is estimated in the spectral domain by using the Wiener filter then propagated via VTS, resulting in a full uncertainty covariance matrix in the feature domain. The discriminative criterion used in this paper is frame-level bMMI which is a soft version of MMI applied within each time frame. In the bMMI criterion, wrong state hypotheses are given more weight than the correct state hypothesis.

The organization of this paper is as follows. In Section 2 we introduce some notations and recall the principle of uncertainty handling. Then in section 3 we introduce the rescaling procedure and the optimization algorithm for the bMMI objective function. Section 4 presents some experimental results on the 2nd CHiME challenge [14] data. Finally in Section 5, we conclude and discuss some future work.

2. UNCERTAINTY HANDLING

2.1. Uncertainty estimation

Multichannel speech enhancement techniques typically operate in the spectral domain by means of the short time Fourier transform (STFT) or some auditory-motivated transform. The observed multichannel signal $\mathbf{x}_{fn}$ is assumed to be the mixture of a single-channel target speech signal s_{fn} and a noise signal $\mathbf{b}_{fn}$, with f denoting the frequency index and n the time frame index. Speech enhancement is achieved by applying a multichannel filter, that can be decomposed into a multichannel spatial filter (a.k.a., a beamformer) yielding a single-channel signal x_{fn} followed by a single-channel spectral post-filter [15, 16]. In the following, we employ the Wiener post-filter: the *mean* $\widehat{\mu}_{s_{fn}}$ of s_{fn} is estimated as [5, 7]

$$\widehat{\mu}_{s_{fn}} = \frac{v_{s_{fn}}}{v_{s_{fn}} + v_{b_{fn}}} x_{fn} \tag{1}$$

with $v_{s_{fn}}$ and $v_{b_{fn}}$ the estimated short-term speech and noise power spectra. The uncertainty is then quantified by the posterior variance of the Wiener filter [7]:

$$\widehat{\sigma}^2_{s_{fn}} = \frac{v_{s_{fn}}\, v_{b_{fn}}}{v_{s_{fn}} + v_{b_{fn}}}. \tag{2}$$

2.2. Uncertainty propagation

Uncertainty is propagated to the vector $\mathbf{z}_n$ consisting of the static Mel frequency cepstral coefficients (MFCCs) and the log-energy. This vector may be computed using the nonlinear function $\mathcal{F}$

$$\mathbf{z}_n = \mathcal{F}(\mathbf{v}_n) = \bar{\mathbf{L}}\bar{\mathbf{D}}\log\left(\bar{\mathbf{M}}\bar{\mathbf{E}}\mathbf{v}_n\right) \tag{3}$$

where $\mathbf{v}_n$ is the concatenation of $|s_{fn}|$ and $|s_{fn}|^2$ for all frequency bins f [10]. The matrices $\bar{\mathbf{E}}$, $\bar{\mathbf{M}}$, $\bar{\mathbf{D}}$ and $\bar{\mathbf{L}}$, are expanded versions of the pre-emphasis matrix, the Mel filterbank matrix, the discrete cosine transform (DCT) matrix, and the liftering matrix, respectively. The estimated mean and uncertainty covariance matrix of $\mathbf{z}_n$ are computed by VTS [10]. Both the estimated mean $\widehat{\boldsymbol{\mu}}_{\mathbf{z}_n}$ and covariance $\widehat{\boldsymbol{\Sigma}}_{\mathbf{z}_n}$ of the static MFCC $\mathbf{z}_n$ are then transformed to the full feature vector consisting of static and dynamic MFCCs. It results in an estimated mean $\widehat{\boldsymbol{\mu}}_{\mathbf{c}_n}$ and an estimated uncertainty covariance matrix $\widehat{\boldsymbol{\Sigma}}_{\mathbf{c}_n}$ for each feature vector $\mathbf{c}_n$ [10].

2.3. Generative uncertainty rescaling

The estimated feature-domain uncertainty is often underestimated compared to the oracle uncertainty, that is the squared difference between clean and enhanced features [17]. To overcome this, the estimated uncertainty can be rescaled. Delcroix et al proposed a linear rescaling transformation for the case of diagonal uncertainty covariance matrices [8, 9]. In [10], we extended this approach to full uncertainty covariance matrices as

$$\widehat{\boldsymbol{\Sigma}}^{\text{scaled}}_{\mathbf{c}_n} = \mathbf{Diag}(\mathbf{b})\widehat{\boldsymbol{\Sigma}}_{\mathbf{c}_n}\mathbf{Diag}(\mathbf{b}) \tag{4}$$

where $\widehat{\boldsymbol{\Sigma}}^{\text{scaled}}_{\mathbf{c}_n}$ is the rescaled estimate and $\mathbf{b}$ is a vector of scaling factors (one per feature dimension). We optimized the scaling factors in a state-independent fashion such that the Euclidean (EUC) distance between the diagonal of the rescaled uncertainty covariance matrix and the oracle diagonal uncertainty covariance matrix is minimum.

2.4. Uncertainty decoding

At the decoding stage, since the clean data $\mathbf{c}_n$ are not exactly known, one cannot directly compute the log-likelihood. We assume that the acoustic emission probability of each state is modeled by a Gaussian mixture model (GMM) trained on clean data. The log-likelihood of each state is hence modified by marginalizing over clean data as [16, 18]

$$p(\mathbf{c}_n|q) = \sum_m w_{q,m}\mathcal{N}(\widehat{\boldsymbol{\mu}}_{\mathbf{c}_n}|\boldsymbol{\mu}_{q,m}, \boldsymbol{\Sigma}_{q,m} + \widehat{\boldsymbol{\Sigma}}_{\mathbf{c}_n}) \tag{5}$$

where m is the component index and $w_{q,m}$, $\boldsymbol{\mu}_{q,m}$, and $\boldsymbol{\Sigma}_{q,m}$ are the weights, means, and covariances of all Gaussian components for state q.

3. DISCRIMINATIVE UNCERTAINTY RESCALING

We now consider discriminative rescaling of the estimated uncertainty covariance matrix using the bMMI criterion. Focusing on the case of a diagonal uncertainty covariance and state-dependent rescaling factors first, the rescaled uncertainty for state q, feature i, and time frame n is given by

$$(\widehat{\sigma}^{\text{scaled}}_{q,c_{i,n}})^2 = b^2_{q,i}(\widehat{\sigma}_{c_{i,n}})^2 \tag{6}$$

where $(\widehat{\sigma}_{c_{i,n}})^2$ is the i-th diagonal element of $\widehat{\boldsymbol{\Sigma}}_{\mathbf{c}_n}$ and $b_{q,i}$ is the i-th element of the state-dependent scaling vector $\mathbf{b}_q$. We denote the resulting rescaled diagonal uncertainty covariance matrix as $\widehat{\boldsymbol{\Sigma}}^{\text{scaled-diag}}_{q,\mathbf{c}_n}$.

The goal is to find the vector $\mathbf{b}_q$ so as to maximize the log-likelihood ratio of the correct recognition hypotheses w.r.t. the incorrect recognition hypotheses at the frame level. The frame-level bMMI criterion is given by:

$$F_{bMMI} = \sum_n \log\left(\frac{p(\mathbf{c}_n|q_n^{\text{true}}, \mathbf{b}_{q_n^{\text{true}}})p(q_n^{\text{true}})}{\sum_{q_n} p(\mathbf{c}_n|q_n, \mathbf{b}_{q_n})p(q_n)e^{\epsilon A(q_n, q_n^{\text{true}})}}\right) \tag{7}$$

where q_n are the hypothesized states and q_n^{true} is the correct state. The term $A(q_n, q_n^{\text{true}})$ is equal to 0 if q_n is the correct state q_n^{true} and to 1 otherwise and ϵ is a boosting factor to be chosen.

The derivative of F_{bMMI} w.r.t. $b_{q,i}$ is given by

$$\frac{\partial F_{bMMI}}{\partial b_{q,i}} = \sum_n \left(\frac{\partial \log(p(\mathbf{c}_n|q_n^{\text{true}}, \mathbf{b}_{q_n^{\text{true}}}))}{\partial b_{q,i}} - \sum_{q_n} \gamma_{q_n} \frac{\partial \log(p(\mathbf{c}_n|q_n, \mathbf{b}_{q_n}))}{\partial b_{q,i}} \right) \quad (8)$$

where

$$\gamma_{q_n} = \frac{p(\mathbf{c}_n|q_n, \mathbf{b}_{q_n}) p(q_n)}{\sum_{q'_n} p(\mathbf{c}_n|q'_n, \mathbf{b}_{q'_n}) p(q'_n) e^{\epsilon A(q'_n, q_n^{\text{true}})}} \quad (9)$$

are the normalized boosted state posteriors obtained using the forward-backward algorithm. Only the terms for which $q_n^{\text{true}} = q$ or $q_n = q$ are eventually nonzero in (8). Computing the corresponding derivatives, we obtain

$$\frac{\partial F_{bMMI}}{\partial b_{q,i}} = \sum_{m,n} (\mathbb{1}_{q_n^{\text{true}}=q} - \gamma_{q_n=q}) \xi_{q,m,n} \delta_{q,m,i,n} \theta_{q,m,i,n} \quad (10)$$

with

$$\xi_{q,m,n} = \frac{w_{q,m} \mathcal{N}(\widehat{\boldsymbol{\mu}}_{\mathbf{c}_n}; \boldsymbol{\mu}_{q,m}, \boldsymbol{\Sigma}_{q,m} + \widehat{\boldsymbol{\Sigma}}^{\text{scaled-diag}}_{q,\mathbf{c}_n})}{\sum_{m'} w_{q,m'} \mathcal{N}(\widehat{\boldsymbol{\mu}}_{\mathbf{c}_n}; \boldsymbol{\mu}_{q,m'}, \boldsymbol{\Sigma}_{q,m'} + \widehat{\boldsymbol{\Sigma}}^{\text{scaled-diag}}_{q,\mathbf{c}_n})} \quad (11)$$

$$\delta_{q,m,i,n} = 1 - \frac{(\widehat{\mu}_{c_{i,n}} - \mu_{q,m,i})^2}{\sigma^2_{q,m,i} + b^2_{q,i} \sigma^2_{c_{i,n}}} \quad (12)$$

$$\theta_{q,m,i,n} = -\frac{b_{q,i} \sigma^2_{c_{i,n}}}{\sigma^2_{q,m,i} + b^2_{q,i} \sigma^2_{c_{i,n}}}. \quad (13)$$

The gradient is then averaged over all utterances.

Assuming that the training data are so-called "stereo data" consisting of aligned clean and noisy signals, the correct state hypothesis is computed by forced alignment of the clean model on the clean training data. The scaling factors $b_{q,i}$ are initialized using the state-independent EUC criterion, as explained in Section 2.3 and detailed in [10]. The bMMI objective function is then optimized using gradient ascent by

$$b_{q,i} \leftarrow b_{q,i} + \eta \frac{\partial F_{bMMI}}{\partial b_{q,i}} \quad (14)$$

where η is the step size. After convergence, the rescaled diagonal and full uncertainty covariance matrices are given by (6) and by $\widehat{\boldsymbol{\Sigma}}^{\text{scaled}}_{q,\mathbf{c}_n} = \mathbf{Diag}(\mathbf{b_q}) \widehat{\boldsymbol{\Sigma}}_{\mathbf{c}_n} \mathbf{Diag}(\mathbf{b}_q)$, respectively. The whole procedure can also be applied in a state-independent fashion.

4. EXPERIMENTS

We assess the proposed method on Track 1 of the 2nd CHiME Challenge [14]. The task considers the problem of recognizing commands being spoken in a noisy living room from recordings made using a binaural manikin. The target utterances are taken from the small-vocabulary Grid corpus. Speech consists of 6-word utterances of the form <command> <color> <preposition> <letter> <digit> <adverb>. Each utterance has been convolved with a set of binaural room impulse responses (BRIRs) simulating speaker movements and reverberation. The utterances are read by 34 speakers and mixed with real domestic background noise at 6 different signal-to-noise ratios (SNRs). The task is to report the 'letter' and 'digit' keywords and performance is measured by keyword accuracy. The training set contains 500 clean (reverberated but noiseless) utterances corresponding to 0.14 hour per speaker. The development set and the test set each contain 600 utterances corresponding to 0.16 hour per SNR.

4.1. Experimental setup

Speech enhancement was applied to the development and test datasets using the Flexible Audio Source Separation Toolbox (FASST) [19] with same settings as in [10]. Speaker-dependent acoustic models with diagonal GMM densities were trained from the clean training set using the HTK baseline provided by the challenge organizers [14]. They consist of conventional left-to-right HMMs with a total of 250 states each modeled by a GMM consisting of 7 Gaussian components.

We estimated the optimal scaling factors both in a state-independent and state-dependent way. The step size η was fixed to 0.01. Training data for bMMI was collected randomly from the development dataset and consists of 300 utterances for each SNR level which corresponds to 5 min. The optimal boosting factor ϵ was found to be 0.1. We used 50 iterations of gradient ascent. For comparison, we also evaluated the performance resulting from the state-independent EUC criterion in [10] and from the state-dependent EUC criterion.

Uncertainty decoding was performed using the HTK baseline with Astudillo's patch[1] for diagonal uncertainty covariances and with our own patch for full uncertainty covariances[2].

4.2. Experimental results

The resulting state-dependent scaling factors are shown in Fig. 1. Most of them look similar to each other. However, certain states such as $q = 8$ appear to be associated with larger uncertainty and some other states such as $q = 209$ with smaller uncertainty. It can also be seen that the scaling factors are larger for lower-order MFCCs and the log-energy and their derivatives than for higher-order MFCCs and their derivatives.

[1] http://www.astudillo.com/ramon/research/stft-up/

[2] http://full-ud-htk.gforge.inria.fr/

Method	depend on state	Test set							Development set						
		-6 dB	-3 dB	0 dB	3 dB	6 dB	9 dB	Average	-6 dB	-3 dB	0 dB	3 dB	6 dB	9 dB	Average
no uncertainty	no	73.75	78.42	84.33	89.50	91.83	92.25	85.01	73.25	78.02	84.33	89.25	91.75	92.18	84.80
EUC (diagonal uncertainty) [10]	no	78.67	79.50	86.33	90.17	92.08	93.75	86.75	78.25	79.17	85.92	89.87	91.80	93.41	86.40
EUC (full uncertainty) [10]	no	81.75	81.83	88.17	90.50	92.67	93.75	88.11	80.63	81.87	87.35	90.57	92.33	93.75	87.75
bMMI estimation (full uncertainty)	no	82.75	83.33	88.17	90.50	92.75	93.50	88.50	82.50	83.17	88.00	90.28	92.17	93.17	88.21
EUC (full uncertainty)	yes	82.00	82.75	88.25	90.75	92.67	93.50	88.32	81.67	83.00	88.17	90.33	91.75	93.00	87.99
bMMI estimation [1]	yes	79.92	82.00	87.17	90.67	92.92	93.42	87.68	80.50	80.51	85.82	90.58	91.50	93.52	87.07
bMMI estimation (diagonal uncertainty)	yes	82.50	83.44	88.50	90.28	92.17	93.50	88.40	81.50	82.64	88.00	90.75	91.83	93.42	88.01
bMMI estimation (full uncertainty)	yes	83.50	84.08	88.75	91.33	93.03	94.51	**89.20**	82.75	83.50	88.17	91.75	93.00	93.67	**88.80**

Table 1. ASR performance expressed in terms of keyword accuracy (in %). Average accuracies have a 95% confidence interval of $\pm 0.8\%$

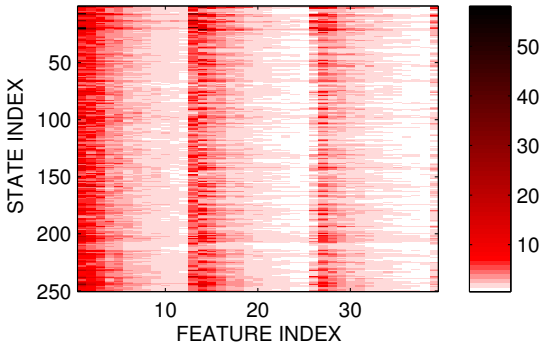

Fig. 1. State-dependent scaling factors trained via bMMI.

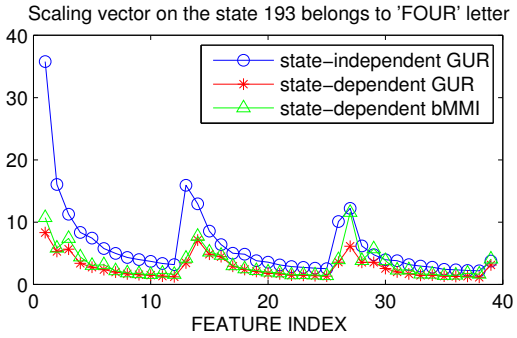

Fig. 2. Optimal scaling factors for state $q = 193$ which belongs to the digit 'four'.

The scaling factors obtained using state-independent EUC, state-dependent EUC, and state-dependent bMMI are compared in Fig. 2 for one particular state. The scaling factors trained by bMMI tend to be smaller than with EUC for most feature indexes, with large differences for certain feature indexes.

The resulting ASR performance figures are listed in Table 1. The baseline without uncertainty propagation achieved 85.01% keyword accuracy. Full uncertainty covariance matrices outperformed diagonal uncertainty matrices in all experiments. More precisely, full uncertainty covariance matrices improved the relative word error rate (WER) by 10% and 7% with EUC and bMMI, respectively.

State-dependent scaling factors also improved performance compared to state-independent scaling factors. The achieved improvements correspond to 2% and 6% relative WER reduction with EUC and bMMI, respectively.

Finally, the bMMI criterion outperformed the EUC criterion for both state-dependent and state-independent rescaling by 3% and 8% relative, respectively. The proposed bMMI approach outperformed the bMMI approach in [1] by 12% relative, due in part to the use of the Wiener uncertainty estimator and to that of a full uncertainty covariance matrix[3]. An even greater improvement could be obtained in the future by considering only keywords in the expression of the bMMI criterion.

5. CONCLUSION

In this paper, we proposed a method for discriminatively rescaling the estimated full feature uncertainty matrix at the frame level. The resulting rescaled uncertainty covariance matrix was confirmed to yield better ASR accuracy and improved by 12% relative compared to [1]. Our results are also among the top three for Track 1 of the 2nd CHiME Challenge [14] and the best ones to our knowledge without using other features than MFCCs or a multi-stream speech recognizer. In future work, we will seek to develop a method to estimate the inter-frame correlation between uncertainties and test our approach on a medium vocabulary task. Using DNNs to estimate parameters of speech and noise would also be very promising.

6. REFERENCES

[1] M. Delcroix, S. Watanabe, T. Nakatani, and A. Nakamura, "Discriminative approach to dynamic variance adaptation for noisy speech recognition," in *Proc. HSCMA*, 2011.

[3]Note that the performance reported in this paper differs from [1] due to the use of a different speech enhancement system.

[2] D. Kolossa and R. Haeb-Umbach, Eds., *Robust Speech Recognition of Uncertain or Missing Data - Theory and Applications*, Springer Verlag, 2011.

[3] R. Astudillo and D. Kolossa, "Uncertainty propagation," in *Robust Speech Recognition of Uncertain or Missing Data - Theory and Applications*, D. Kolossa and R. Haeb-Umbach, Eds., pp. 35–62. Springer, 2011.

[4] D. T. Tran, E. Vincent, and D. Jouvet, "Fusion of multiple uncertainty estimators and propagators for noise robust ASR," in *Proc. ICASSP*, 2014.

[5] A. Ozerov, M. Lagrange, and E. Vincent, "Uncertainty-based learning of acoustic models from noisy data," *Computer Speech and Language*, vol. 27, no. 3, pp. 874–894, Feb. 2013.

[6] M. Gales, *Model Based Techniques for Noise Robust Speech Regcognition*, Ph.D. thesis, Cambridge University, 1995.

[7] R. Astudillo, *Integration of Short-Time Fourier Domain Speech Enhancement and Observation Uncertainty Techniques for Robust Automatic Speech Recognition*, Ph.D. thesis, TU Berlin, 2010.

[8] M. Delcroix, K. Kinoshita, T. Nakatani, S. Araki, A. Ogawa, T. Hori, S. Watanabe, et al., "Speech recognition in living rooms: Integrated speech enhancement and recognition system based on spatial, spectral and temporal modeling of sounds," *Computer Speech and Language*, vol. 27, no. 3, pp. 851–873, May 2013.

[9] M. Delcroix, T. Nakatani, and S. Watanabe, "Static and dynamic variance compensation for recognition of reverberant speech with dereverberation preprocessing," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 2, pp. 324–334, Jan 2009.

[10] D. T. Tran, E. Vincent, and D. Jouvet, "Extension of uncertainty propagation to dynamic MFCCs for noise-robust ASR," in *Proc. ICASSP*, 2014.

[11] E. McDermott, S. Watanabe, and A. Nakamura, "Discriminative training based on an integrated view of mpe and mmi in margin and error space," in *Proc. ICASSP*, 2010.

[12] D. Povey, D. Kanevsky, B. Kingsbury, B. Ramabhadran, G. Saon, and K. Visweswariah, "Boosted MMI for model and feature-space discriminative training," in *Proc. ICASSP*, 2008.

[13] Y. Tachioka, S. Watanabe, J. L. Roux, and J. R. Hershey, "Discriminative methods for noise robust speech recognition: A chime challenge benchmark," in *Proc. CHiME*, 2013.

[14] E. Vincent, J. Barker, S. Watanabe, J. Le Roux, F. Nesta, and M. Matassoni, "The second 'CHiME' speech separation and recognition challenge: An overview of challenge systems and outcomes," in *Proc. ASRU*, 2013.

[15] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amppitude estimator," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 32, no. 6, pp. 1109–1121, Jul 1984.

[16] D. Kolossa, R. Astudillo, E. Hoffmann, and R. Orglmeister, "Independent component analysis and time-frequency masking for multi speaker recognition," in *EURASIP Journal on Audio, Speech, and Music Processing*, 2010, vol. 2010, Article ID 651420.

[17] L. Deng, J. Wu, J. Droppo, and A. Acero, "Dynamic compensation of HMM variances using the feature enhancement uncertainty computed from a parametric model of speech distortion," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 13, no. 3, pp. 412 – 421, May 2005.

[18] L. Deng, "Front-end, back-end, and hybrid techniques for noise-robust speech recognition," in *Robust Speech Recognition of Uncertain or Missing Data - Theory and Applications*, pp. 67–99. Springer, 2011.

[19] A. Ozerov, E. Vincent, and F. Bimbot, "A general flexible framework for the handling of prior information in audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1118 – 1133, May 2012.