Fusion of Multiple Uncertainty Estimators and Propagators for Noise Robust ASR

Dung Tran, Emmanuel Vincent, Denis Jouvet

FUSION OF MULTIPLE UNCERTAINTY ESTIMATORS AND PROPAGATORS 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

Uncertainty decoding has been successfully used for speech recognition in highly nonstationary noise environments. Yet, accurate estimation of the uncertainty on the denoised signals and propagation to the features remain difficult. In this work, we propose to fuse the uncertainty estimates obtained from different uncertainty estimators and propagators by linear combination. The fusion coefficients are optimized by minimizing a measure of divergence with oracle estimates on development data. Using the Kullback-Leibler divergence, we obtain 18% relative error rate reduction on the 2nd CHiME Challenge with respect to conventional decoding, that is about twice as much as the reduction achieved by the best single uncertainty estimator and propagator.

Index Terms— Noise robust ASR, uncertainty handling

1. INTRODUCTION

Automatic speech recognition (ASR) remains challenging in everyday nonstationary noise environments [1]. Robust ASR approaches [2] may be classified as model compensation [3], feature compensation [4] or hybrid techniques [5–7]. Uncertainty decoding [8–14] has emerged as a promising hybrid technique whereby speech enhancement is applied to the input noisy signal and the enhanced features are not considered as point estimates but as a *Gaussian distribution with time-varying variance* or *uncertainty* that is used to dynamically adapt the acoustic model on each time frame for decoding. Uncertainty decoding may be used with feature-domain or spectral-domain enhancement. We adopt the latter approach, as it benefits from multichannel information and it has led to the best ASR accuracy in a real domestic environment as evaluated by the CHiME Challenge [15]. Following [10, 11, 14], we *estimate* the uncertainty in the spectral domain and we subsequently *propagate* it to the feature domain.

Various uncertainty estimators on the spectral domain have been proposed based on statistical models or on heuristics [10, 11, 14]. Several feature domain uncertainty propagators have also been found based on moment matching [3], on the unscented transform [10], or on vector Taylor series [16]. While the latter were shown to be accurate [11], we found the former to be somewhat inaccurate experimentally so that the ASR performance remains lower than the one that would be achieved with perfect *oracle* uncertainty estimates [8, 17].

In this work, we introduce a fusion framework to improve uncertainty estimates by linearly combining different uncertainty estimators and propagators. The fusion coefficients are obtained by minimizing some measure of divergence with oracle uncertainty estimates on development data. We evaluate the impact on ASR performance for different divergences on Track 1 of the 2nd CHiME Challenge [15].

The paper is organized as follows. Section 2 introduces a number of existing uncertainty estimators and propagators. The fusion framework and the estimation of the fusion coefficients are described in Section 3. ASR results are discussed in Section 4. We conclude in Section 5.

2. BACKGROUND

2.1. Spectral domain 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 [4, 10]. In the following, we employ the Wiener post-filter: the *mean* $\widehat{\mu}_{s_{fn}}$ of s_{fn} is estimated as [11, 17]

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

with $v_{s_{fn}}$ and $v_{b_{fn}}$ the estimated short-term speech and noise power spectra. The goal of uncertainty estimation is to quantify how much the true (unknown) value of s_{fn} deviates from

$\widehat{\mu}_{s_{fn}}$ using its *variance* denoted as $\widehat{\sigma}^2_{s_{fn}}$. In the following, we use the terms variance and uncertainty interchangeably.

2.1.1. Kolossa's estimator

Kolossa et al. [10] assumed the uncertainty to be proportional to the squared difference between the enhanced signal and the mixture

$$\widehat{\sigma}^2_{s_{fn}} = \alpha |\widehat{\mu}_{s_{fn}} - x_{fn}|^2 \tag{2}$$

where the scaling factor α is found by minimizing the Euclidean distance between the estimated uncertainty and the oracle uncertainty $\sigma^2_{s_{fn}}$ defined hereafter in (10).

2.1.2. Wiener estimator

Astudillo [11] later proposed to quantify uncertainty by the posterior variance of the Wiener filter:

$$\widehat{\sigma}^2_{s_{fn}} = \frac{v_{s_{fn}} v_{b_{fn}}}{v_{s_{fn}} + v_{b_{fn}}}. \tag{3}$$

2.1.3. Nesta's estimator

Recently, Nesta et al. [14] obtained a different estimate based on a binary speech/noise predominance model[1]:

$$\widehat{\sigma}^2_{s_{fn}} = \widehat{p}_{fn}(1 - \widehat{p}_{fn})|x_{fn}|^2 \tag{4}$$

where $\widehat{p}_{fn} = \sqrt{v_{s_{fn}}}/(\sqrt{v_{s_{fn}}} + \sqrt{v_{b_{fn}}})$. The behavior of the three estimators is illustrated in Figure 1.

2.2. Feature domain uncertainty propagation

The estimated means and variances of s_{fn} are stacked into a mean vector $\widehat{\boldsymbol{\mu}}_{\mathbf{s}_n}$ and a diagonal covariance matrix $\widehat{\boldsymbol{\Sigma}}_{\mathbf{s}_n}$ in each time frame n and they are propagated to the features. We use 39-dimensional feature vectors $\mathbf{c}_n$ consisting of 12 Mel-frequency cepstral coefficients (MFCCs), the log-energy, and their first- and second-order time derivatives.

2.2.1. VTS propagator

Vector Taylor series (VTS), which was first introduced for another purpose in [16], consists of linearizing the MFCC transform by its first-order Taylor expansion and of propagating uncertainty through this linear transform [17]. Denoting by $\mathbf{E}$ the diagonal matrix of pre-emphasis coefficients, by $\mathbf{M}$ the Mel-filterbank matrix, by $\mathbf{D}$ the discrete cosine transform (DCT) matrix, and by $\mathbf{L}$ the diagonal matrix of lifter coefficients, we obtain the mean $\widehat{\boldsymbol{\mu}}_{\text{MFCC}_n}$ and the covariance matrix $\widehat{\boldsymbol{\Sigma}}_{\text{MFCC}_n}$ of the MFCCs as [18]

$$\widehat{\boldsymbol{\mu}}_{\text{MFCC}_n} = \mathbf{LD}\log(\mathbf{ME}|\widehat{\boldsymbol{\mu}}_{\mathbf{s}_n}|) \tag{5}$$

$$\widehat{\boldsymbol{\Sigma}}_{\text{MFCC}_n} = \mathbf{LDDiag}(1/\widehat{\boldsymbol{\mu}}_{\text{MEL}_n})\widehat{\boldsymbol{\Sigma}}_{\text{MEL}_n}(\mathbf{LDDiag}(1/\widehat{\boldsymbol{\mu}}_{\text{MEL}_n}))^T \tag{6}$$

[1] This formula was initially defined for the variance of $|s_{fn}|$ [14], however we found it beneficial to use it for the variance of s_{fn} instead.

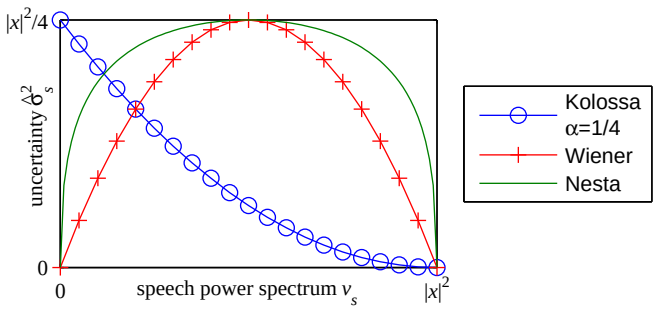

Fig. 1. Behavior of the uncertainty estimators.

where $\widehat{\boldsymbol{\mu}}_{\text{MEL}_n} = \mathbf{ME}\widehat{\boldsymbol{\mu}}_{|\mathbf{s}_n|}$, $\widehat{\boldsymbol{\Sigma}}_{\text{MEL}_n} = \mathbf{ME}\widehat{\boldsymbol{\Sigma}}_{|\mathbf{s}_n|}(\mathbf{ME})^T$, $\widehat{\boldsymbol{\mu}}_{|\mathbf{s}_n|}$ and $\widehat{\boldsymbol{\Sigma}}_{|\mathbf{s}_n|}$ are the mean and the covariance matrix of $|\mathbf{s}_n|$ which are derived from $\widehat{\boldsymbol{\mu}}_{\mathbf{s}_n}$ and $\widehat{\boldsymbol{\Sigma}}_{\mathbf{s}_n}$ using the statistics of the Rice distribution [11], $\mathbf{Diag}(\cdot)$ transforms a vector into a diagonal matrix, ^{T} denotes matrix transposition, and the logarithm, the division, and the magnitude are taken element-wise. Note that $\widehat{\boldsymbol{\mu}}_{\text{MFCC}_n}$ is deterministically computed, so it does not depend on the chosen uncertainty estimator. Similar calculations are performed for the log-energy and for the mean and the covariance matrix of the dynamic coefficients as detailed in [18]. Cepstral mean subtraction is applied and only the diagonal of the covariance matrix is eventually retained [11]. The resulting mean and variance of the ith feature in frame n are denoted as $\widehat{\mu}_{c_{in}}$ and $\widehat{\sigma}^2_{c_{in}}$, respectively.

2.2.2. MM and UT propagators

Alternative uncertainty propagation techniques include the unscented transform (UT) and moment matching (MM), also known as the log-normal transform, which provide other formulas to propagate $\widehat{\boldsymbol{\mu}}_{|\mathbf{s}_n|}$ and $\widehat{\boldsymbol{\Sigma}}_{|\mathbf{s}_n|}$ through the logarithm of the MFCC transform [3, 10, 11]. Both $\widehat{\mu}_{c_{in}}$ and $\widehat{\sigma}^2_{c_{in}}$ then depend on the chosen uncertainty estimator.

3. PROPOSED FUSION FRAMEWORK

Uncertainty decoding has the potential to improve ASR performance on noisy data close to that on clean data as shown by oracle experiments [8, 17]. Yet, the improvement observed in practical scenarios is typically lower due to the inaccuracy of the uncertainty estimators. Improving existing estimators is therefore crucial to unleash its full potential.

Figure 1 shows that the three estimators introduced in Section 2.1 have different behaviors. Kolossa's estimator decreases when the speech power spectrum increases. The two other estimators reach a maximum when the power spectra of speech and noise are equal but Nesta's estimator increases more quickly than the Wiener estimator.

Motivated by this observation, we propose to fuse multiple spectral domain uncertainty estimators in order to obtain

more accurate estimators. The fused uncertainty estimates are propagated to the feature domain using one or more uncertainty propagators and the resulting feature domain estimates are further fused in order to obtain a more accurate propagator. Both fusions are achieved by linear combination.

3.1. Fusion of uncertainty estimators

In the spectral domain, fusion is performed separately in each frequency bin f. Denoting by E the number of estimators and by N the number of time frames, it can be expressed as

$$(\widehat{\sigma}^{\text{fus}}_{s_{fn}})^2 = \sum_{e=1}^{E} w^e_{s_f} \, (\widehat{\sigma}^{e}_{s_{fn}})^2 \tag{7}$$

where $(\widehat{\sigma}^{e}_{s_{fn}})^2$ is one of the original estimators in (2), (3), (4), $w^e_{s_f}$ are fusion coefficients, and $(\widehat{\sigma}^{\text{fus}}_{s_{fn}})^2$ is the fused estimator. The fusion coefficients are constrained to be nonnegative so that the fused estimator is always nonnegative. Stacking the original uncertainty estimates into a $E \times N$ matrix $\widehat{\mathbf{\Sigma}}_{s_f}$ and the fused estimates into a $1 \times N$ vector $\widehat{\mathbf{\Sigma}}^{\text{fus}}_{s_f}$ for each frequency f, (7) can be written in matrix form as

$$\widehat{\mathbf{\Sigma}}^{\text{fus}}_{s_f} = \mathbf{w}_{s_f} \widehat{\mathbf{\Sigma}}_{s_f} \tag{8}$$

where $\mathbf{w}_{s_f}$ is the $1 \times E$ vector of fusion coefficients. These coefficients are optimized on development data for which the true speech signal is known by solving the optimization problem

$$\mathbf{w}_{s_f} = \arg \min_{\mathbf{w}_{s_f} \geq 0} \mathrm{D}\left(\mathbf{\Sigma}_{s_f} | \mathbf{w}_{s_f} \widehat{\mathbf{\Sigma}}_{s_f}\right) \tag{9}$$

where D is a divergence measure [19] such as the Itakura-Saito (IS) divergence, the Kullback-Leibler (KL) divergence, or the squared Euclidean distance, and $\mathbf{\Sigma}_{s_f}$ is the $1 \times N$ vector of oracle uncertainty estimates computed by [10]

$$\sigma^2_{s_{fn}} = |\widehat{\mu}_{s_{fn}} - s_{fn}|^2 \tag{10}$$

where s_{fn} is the true speech signal. Note that, in that case, N represents all time frames of all development samples.

3.2. Fusion of uncertainty propagators

Several fused uncertainty estimators corresponding to different choices of divergence are retained. The resulting spectral domain uncertainty estimates $\widehat{\mathbf{\Sigma}}^{\text{fus}}_{s_f}$ are then propagated to the feature domain using one or more propagators such as VTS, MM, or UT, yielding P feature domain uncertainty estimates $(\widehat{\sigma}^{p}_{c_{in}})^2$ indexed by p. Assuming that the corresponding means $\widehat{\mu}^{p}_{c_{in}}$ are identical for all p (for instance, when using VTS only), these uncertainty estimates are in turn stacked into a $P \times N$ matrix $\widehat{\mathbf{\Sigma}}_{c_i}$ for each feature index i and a fused uncertainty propagator is obtained as

$$\widehat{\mathbf{\Sigma}}^{\text{fus}}_{c_i} = \mathbf{w}_{c_i} \widehat{\mathbf{\Sigma}}_{c_i} \tag{11}$$

where $\widehat{\mathbf{\Sigma}}^{\text{fus}}_{c_i}$ is the $1 \times N$ vector of fused estimates and $\mathbf{w}_{c_i}$ is the $1 \times P$ vector of fusion coefficients. This equation still holds when the corresponding means differ, except that one mean $\widehat{\mu}^{\text{ref}}_{c_{in}}$ is chosen as a reference and the entries of $\widehat{\mathbf{\Sigma}}_{c_i}$ are corrected for the squared bias as $(\widehat{\sigma}^{p}_{c_{in}})^2 + (\widehat{\mu}^{p}_{c_{in}} - \widehat{\mu}^{\text{ref}}_{c_{in}})^2$. In either case, the fusion coefficients are optimized as

$$\mathbf{w}_{c_i} = \arg \min_{\mathbf{w}_{c_i} \geq 0} \mathrm{D}\left(\mathbf{\Sigma}_{c_i} | \mathbf{w}_{c_i} \widehat{\mathbf{\Sigma}}_{c_i}\right) \tag{12}$$

where $\mathbf{\Sigma}_{c_i}$ is the $1 \times N$ vector of oracle feature domain uncertainties computed from the true features c_{in} as [8]

$$\sigma^2_{c_{in}} = (\widehat{\mu}_{c_{in}} - c_{in})^2. \tag{13}$$

3.3. Additive bias compensation

In order to compensate for a possible additive bias in the original uncertainty estimates, we do not only scale them by the fusion coefficients but we also add a nonnegative frequency- or feature-dependent bias. This is simply achieved by adding a row to the matrices $\widehat{\mathbf{\Sigma}}_{s_f}$ and $\widehat{\mathbf{\Sigma}}_{c_i}$ whose elements are equal to 1. The optimal bias is then found as the corresponding coefficient of $\mathbf{w}_{s_f}$ or $\mathbf{w}_{c_i}$.

3.4. Estimation of the fusion coefficients

The optimization problems (9) and (12) are instances of nonnegative matrix factorization (NMF) [20]. The IS divergence, the KL divergence, and the squared Euclidean distance belong to the more general family of β-divergences with $\beta = 0$, 1, or 2, respectively [19]. The fusion coefficients are found by applying the following iterative multiplicative updates [19]:

$$\mathbf{w}_{s_f} \leftarrow \mathbf{w}_{s_f} \odot \frac{\left((\mathbf{w}_{s_f} \widehat{\mathbf{\Sigma}}_{s_f})^{\beta-2} \odot \mathbf{\Sigma}_{s_f}\right) (\widehat{\mathbf{\Sigma}}_{s_f})^T}{(\mathbf{w}_{s_f} \widehat{\mathbf{\Sigma}}_{s_f})^{\beta-1} (\widehat{\mathbf{\Sigma}}_{s_f})^T} \tag{14}$$

$$\mathbf{w}_{c_i} \leftarrow \mathbf{w}_{c_i} \odot \frac{\left((\mathbf{w}_{c_i} \widehat{\mathbf{\Sigma}}_{c_i})^{\beta-2} \odot \mathbf{\Sigma}_{c_i}\right) (\widehat{\mathbf{\Sigma}}_{c_i})^T}{(\mathbf{w}_{c_i} \widehat{\mathbf{\Sigma}}_{c_i})^{\beta-1} (\widehat{\mathbf{\Sigma}}_{c_i})^T} \tag{15}$$

where $\odot$ denotes element-wise multiplication and powers are computed element-wise. The fusion coefficients estimated on the development data are then applied to the test data.

4. EXPERIMENTS

We assess the proposed fusion framework on Track 1 of the 2nd CHiME Challenge [15]. The target utterances are 6-word sequences of the form <command> <color> <preposition> <letter> <digit> <adverb>. The utterances are read by 34 speakers and mixed with real domestic background noise at 6 different signal-to-noise ratios (SNRs). Intending to increase difficulty, the task is to report the letter and digit tokens and performance is measured as the percentage of tokens recognized correctly. The training set contains 500 noiseless reverberated utterances corresponding to 0.14 hour per speaker.

Uncertainty		Test set							Development set						
estimation	propagation	-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
none	none	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
Kolossa	VTS	75.25	79.83	85.42	89.92	92.25	93.88	86.08	74.58	79.54	85.12	89.73	92.15	93.56	85.78
Wiener	VTS	76.50	79.08	85.83	89.92	92.00	93.75	86.18	76.13	78.68	85.56	89.68	91.75	93.50	85.88
Nesta	VTS	77.58	80.00	85.33	89.33	92.33	94.08	86.44	77.00	79.52	85.17	89.33	92.15	93.78	86.16
KL fusion	VTS	78.33	80.17	85.92	**90.08**	92.08	**94.17**	86.79	78.01	80.07	85.75	89.96	91.67	**93.82**	86.55
	KL fusion	**81.33**	**81.92**	**88.17**	89.58	**92.42**	93.08	**87.75**	**79.25**	**81.67**	**86.92**	**90.58**	**92.25**	93.33	**87.33**

Table 1. Keyword accuracy (in %) before and after fusion.

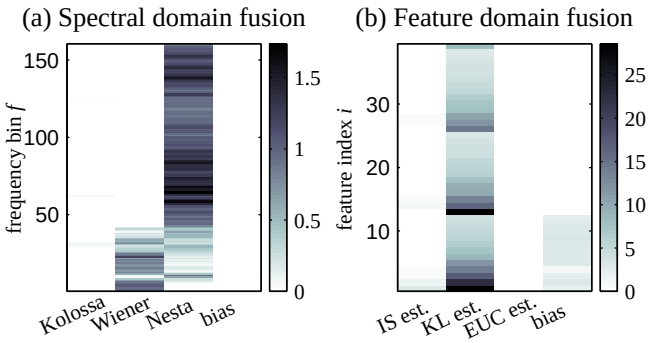

Fig. 2. Estimated fusion coefficients (a) $\mathbf{w}_{s_f}$ and (b) $\mathbf{w}_{c_i}$ using the KL divergence.

Divergence	Estimation	Propagation
IS	85.16	86.49
KL	**86.55**	**87.33**
EUC	86.18	86.72

Table 2. Keyword accuracy (in %) for several divergences.

The development set and the test set each contain 600 utterances corresponding to 0.16 hour per SNR.

4.1. Experimental setup

Speech enhancement is applied to the development and test datasets using the Flexible Audio Source Separation Toolbox (FASST) [21] with the following settings optimized on the development set. An auditory-motivated equivalent rectangular bandwidth (ERB) time-frequency representation is used with 160 bands and half-overlapping 32 ms time frames. The number of noise sources is set to 2. The power spectra of speech and noise are modeled by NMF with 32 components and their spatial covariance matrices are modeled as full-rank [21].

Speaker-dependent acoustic models are trained from the training set using the HTK baseline provided by the challenge organizers [15]. Decoding is performed using the HTK baseline with Astudillo's uncertainty decoding patch[2]. This patch dynamically adapts the Gaussian mixture model (GMM) observation probabilities according to Deng's rule [8] so as to account for the estimated feature domain uncertainty.

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

4.2. Experimental results

Preliminary experiments showed that VTS-based uncertainty propagation outperforms MM and UT for all uncertainty estimators. Only the results of VTS are hence reported hereafter. Table 1 shows the results before and after KL fusion. Similar trends are observed on the development and the test data. On average over all SNRs, the baseline test set accuracy with conventional decoding (no uncertainty) is 85.01%. Nesta's uncertainty estimator outperforms the other individual estimators and it achieves 86.44% accuracy, that is 10% relative error rate reduction with respect to the baseline. Both fusions are achieved by SNR independent linear combination. By fusing all uncertainty estimators, performance further improves to 86.79%. Fig. 2a indicates that the optimal estimator is a scaled version of Nesta's at higher frequencies and a mixture of Wiener and Nesta's at lower frequencies. Finally, fusing the IS-fused estimator, the KL-fused estimator and the EUC-fused estimator in the feature domain yields 87.75% accuracy, that is 18% relative error rate reduction compared to the baseline. Fig. 2b indicates that mostly a scaled version of the KL-fused estimator is retained and that it is compensated for an additive bias on the static features. Table 2 completes these results by showing that the KL divergence performs better than the other two divergences for both fusion stages.

5. CONCLUSION

We proposed a fusion framework to improve the accuracy of uncertainty estimates in the context of uncertainty decoding. Experiments on the 2nd CHiME Challenge data showed that minimizing the KL divergence between the fused uncertainties and the oracle uncertainties results in a significantly reduction of error rate by 18% relative to conventional decoding, compared to 10% only for the best single uncertainty estimator and propagator. In the future, we aim to generalize the proposed linear framework into a nonlinear fusion framework and with larger set of divergences.

6. ACKNOWLEDGMENT

This work has been partly realized thanks to the support of the Région Lorraine and the CPER MISN TALC project.

7. REFERENCES

[1] J. M. Baker, L. Deng, J. Glass, S. Khudanpur, C.-H. Lee, N. Morgan, and D. O'Shaughnessy, "Research developments and directions in speech recognition and understanding, part 1," *IEEE Signal Processing Magazine*, vol. 26, no. 3, pp. 75–80, May 2009.

[2] 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.

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

[4] 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.

[5] L. Deng, A. Acero, M. Plumpe, and X. D. Huang, "Large vocabulary speech recognition under adverse acoustic environments," in *Proc. ICSLP*, 2000, pp. 806–809.

[6] M. Cooke, "Robust automatic speech recognition with missing and unreliable acoustic data," *Speech Communication*, vol. 34, no. 3, pp. 267–285, June 2001.

[7] H. Liao and M. J. F. Gales, "Adaptive training with joint uncertainty decoding for robust recognition of noisy data," in *Proc. ICASSP*, 2007, vol. 4, pp. 389–392.

[8] 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.

[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. 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.

[11] 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.

[12] 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.

[13] R. F. Astudillo, D. Kolossa, A. Abad, S. Zeiler, R. Saeidi, P. Mowlaee, J. P. da Silva Neto, and R. Martin, "Integration of beamforming and uncertainty-of-observation techniques for robust ASR in multi-source environments," *Computer Speech and Language*, vol. 27, no. 3, pp. 837–850, May 2013.

[14] F. Nesta, M. Matassoni, and R. Astudillo, "A flexible spatial blind source extraction framework for robust speech recognition in noisy environments," in *Proc. CHiME*, 2013, pp. 33–40.

[15] 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.

[16] P. J. Moreno, B. Raj, and R. M. Stern, "A vector Taylor series approach for environment-independent speech recognition," in *Proc. ICASSP*, 1996, vol. 2, pp. 733 – 736.

[17] 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.

[18] D. T. Tran, E. Vincent, and D. Jouvet, "Extension of uncertainty propagation to dynamic MFCCs for noise-robust ASR," submitted.

[19] R. Kompass, "A generalized divergence measure fon nonnegative matrix factorization," *Neural Computation*, vol. 19, no. 3, pp. 780–791, Mar. 2007.

[20] D. D. Lee and H. S. Seung, "Learning the parts of objects with nonnegative matrix factorization," *Nature*, vol. 401, pp. 788–791, 1999.

[21] 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.