Extension of uncertainty propagation to dynamic MFCCs for noise robust ASR

Dung Tien Tran, Emmanuel Vincent, Denis Jouvet

EXTENSION OF UNCERTAINTY PROPAGATION TO DYNAMIC MFCCS 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 propagation has been successfully employed for speech recognition in nonstationary noise environments. The uncertainty about the features is typically represented as a diagonal covariance matrix for static features only. We present a framework for estimating the uncertainty over both static and dynamic features as a full covariance matrix. The estimated covariance matrix is then multiplied by scaling coefficients optimized on development data. We achieve 21% relative error rate reduction on the 2nd CHiME Challenge with respect to conventional decoding without uncertainty, that is five times more than the reduction achieved with diagonal uncertainty covariance for static features only.

Index Terms— Automatic speech recognition, noise robustness, uncertainty handling

1. INTRODUCTION

Robust automatic speech recognition (ASR) remains very challenging in scenarios involving nonstationary noise sources overlapping with the target speech [1–4]. Model compensation [5], feature compensation [6] and hybrid compensation techniques [7–10] are the three main types of approaches [11]. Uncertainty propagation [12–15] is a hybrid technique in which the features are considered as a distribution with dynamic covariance matrix instead of point estimates. The mean and the covariance matrix are first *estimated* in the spectral domain using a speech enhancement system and they are *propagated* to the feature domain. This information is then exploited to dynamically adapt the acoustic model on each time frame via uncertainty decoding [16]. In the following, we use uncertainty propagation in combination with multichannel speech enhancement, which typically outperforms single-channel enhancement in real nonstationary noise conditions [4].

The estimation and the propagation of uncertainty have been examined in several previous studies. In the spectral domain, the uncertainty in different time-frequency bins is typically assumed to be independent so that it is represented as a diagonal covariance matrix [12, 15, 17]. In the feature domain, this translates into a full uncertainty covariance matrix over the Mel frequency cepstral coefficients (MFCCs) [17], yet only the diagonal of this matrix is typically retained for decoding [12,13,15]. Moreover, propagation to delta-MFCCs and delta-delta-MFCCs has not been considered to the best of our knowledge. This comes as no surprise, since the above independence assumption is likely to result in erroneous uncertainty estimates over the dynamic MFCCs.

The major contribution of this work is the introduction of a step-by-step procedure to propagate the estimated spectral domain uncertainty to the static MFCCs, to the log-energy, and to their first- and second-order time derivatives. In order to correct the mismatch due to the spectral domain independence assumption, we multiply the estimated covariance matrix by scaling coefficients which are optimized on development data. We evaluate the resulting ASR performance on Track 1 of the 2nd CHiME Challenge [4].

The paper is organized as follows. Section 2 reviews uncertainty estimation in the spectral domain. The proposed uncertainty propagation procedure is described in Section 3. ASR results are reported in Section 4. We conclude in Section 5.

2. UNCERTAINTY ESTIMATION

Let us consider a mixture of J speech and noise sources recorded by I microphones. In the short-time Fourier transform (STFT) domain, the observed multichannel signal $\mathbf{x}_{fn}$ can be modeled as [18]

$$\mathbf{x}_{fn} = \sum_{j=1}^{J} \mathbf{y}_{jfn} \tag{1}$$

where $\mathbf{y}_{jfn}$ is the spatial image of the j-th source, and f and n are the frequency index and the frame index, respectively. The goal of uncertainty estimation is to obtain not only a point estimate of the target speech source $\mathbf{y}_{jfn}$ represented by its *mean* $\widehat{\boldsymbol{\mu}}_{\mathbf{y}_{jfn}}$ but also an estimate of how much the true (unknown) source signal may deviate from it, as represented by

its *covariance matrix* $\widehat{\mathbf{\Sigma}}_{\mathbf{y}_{jfn}}$. This may be achieved by multichannel Wiener filtering as follows [13, 17]:

$$\widehat{\boldsymbol{\mu}}_{\mathbf{y}_{jfn}} = \mathbf{W}_{jfn}\mathbf{x}_{fn} \tag{2}$$

$$\widehat{\mathbf{\Sigma}}_{\mathbf{y}_{jfn}} = (\mathbf{I}_I - \mathbf{W}_{jfn})\, v_{jfn}\mathbf{R}_{jf} \tag{3}$$

where $\mathbf{W}_{jfn} = v_{jfn}\mathbf{R}_{jf}(\sum_{j'} v_{j'fn}\mathbf{R}_{j'f})^{-1}$ is the Wiener filter, $\mathbf{I}_I$ is the identity matrix of size I, and v_{jfn} and $\mathbf{R}_{jf}$ are the short-term power spectrum and the spatial covariance matrix of the source, which may be estimated using a number of alternative speech enhancement techniques [12, 15, 18]. The source spatial images $\mathbf{y}_{jfn}$ are then downmixed into single-channel source signals s_{jfn} as

$$s_{jfn} = \mathbf{u}_f^H \mathbf{y}_{jfn} \tag{4}$$

where $\mathbf{u}_f$ is a steering vector pointing to the source direction and ^{H} denotes conjugate transposition. In the context of the CHiME challenge [4], $\mathbf{u}_f^H = [0.5\ 0.5]$ for all f. The mean and the variance of s_{jfn} are given by

$$\widehat{\mu}_{s_{jfn}} = \mathbf{u}_f^H \widehat{\boldsymbol{\mu}}_{\mathbf{y}_{jfn}} \tag{5}$$

$$\widehat{\sigma}^2_{s_{jfn}} = \mathbf{u}_f^H \widehat{\mathbf{\Sigma}}_{\mathbf{y}_{jfn}} \mathbf{u}_f \tag{6}$$

As an alternative to the STFT, quadratic time-frequency representations often improve enhancement by accounting for the local correlation between channels [18]. The variance of s_{jfn} can still be computed as above but the mean cannot anymore since the mixture is represented by its local covariance matrix $\widehat{\mathbf{R}}_{\mathbf{x}_{fn}}$ instead of $\mathbf{x}_{fn}$. A more general expression may however be obtained for the magnitude of the mean as

$$|\widehat{\mu}_{s_{jfn}}| = \left(\mathbf{u}_f^H \mathbf{W}_{jfn} \widehat{\mathbf{R}}_{\mathbf{x}_{fn}} \mathbf{W}_{jfn}^H \mathbf{u}_f\right)^{1/2}. \tag{7}$$

3. EXTENSION OF UNCERTAINTY PROPAGATION

The mean $\widehat{\mu}_{s_{jfn}}$ and the variance $\widehat{\sigma}^2_{s_{jfn}}$ of the target speech source are propagated step by step to the feature domain for exploitation by the recognizer. We use 39-dimensional feature vectors $\mathbf{c}_n$ consisting of 12 MFCCs, the log-energy, and their first- and second-order time derivatives. For legibility, we remove the index j from now on.

3.1. To the magnitude and the power spectra

The first step is to propagate the uncertainty from the complex-valued spectrum to the magnitude and the power spectra. Let us define the 2×1 vector $\mathbf{v}_{fn} = [|s_{fn}|\ |s_{fn}|^2]^T$. The mean and the covariance matrix of $\mathbf{v}_{fn}$ are given by

$$\hat{\boldsymbol{\mu}}_{\mathbf{v}_{fn}} = \begin{bmatrix} E_1 \\ E_2 \end{bmatrix} \tag{8}$$

$$\widehat{\mathbf{\Sigma}}_{\mathbf{v}_{fn}} = \begin{bmatrix} E_2 - E_1^2 & E_3 - E_1E_2 \\ E_3 - E_1E_2 & E_4 - E_2^2 \end{bmatrix} \tag{9}$$

where $E_k = E(|s_{fn}|^k)$ is the k-th order moment of the distribution of $|s_{fn}|$. The distribution of s_{fn} is assumed be complex-valued Gaussian distribution [19]. Therefore, E_k has the following closed form [20]:

$$E_k = \Gamma\left(\frac{k}{2}+1\right)\left(\widehat{\sigma}^2_{s_{fn}}\right)^{\frac{k}{2}} L_{\frac{k}{2}}\left(-\frac{|\widehat{\mu}_{s_{fn}}|^2}{\widehat{\sigma}^2_{s_{fn}}}\right) \tag{10}$$

where Γ is the gamma function and $L_{\frac{k}{2}}$ is the Laguerre polynominal. The first four moments are obtained as

$$E_1 = \Gamma\left(\frac{3}{2}\right)\left(\widehat{\sigma}^2_{s_{fn}}\right)^{\frac{1}{2}} L_{\frac{1}{2}}\left(-\frac{|\widehat{\mu}_{s_{fn}}|^2}{\widehat{\sigma}^2_{s_{fn}}}\right) \tag{11}$$

$$E_2 = \widehat{\sigma}^2_{s_{fn}} + |\widehat{\mu}_{s_{fn}}|^2 \tag{12}$$

$$E_3 = \Gamma\left(\frac{5}{2}\right)\left(\widehat{\sigma}^2_{s_{fn}}\right)^{\frac{3}{2}} L_{\frac{3}{2}}\left(-\frac{|\widehat{\mu}_{s_{fn}}|^2}{\widehat{\sigma}^2_{s_{fn}}}\right) \tag{13}$$

$$E_4 = |\widehat{\mu}_{s_{fn}}|^4 + 4|\widehat{\mu}_{s_{fn}}|^2\widehat{\sigma}^2_{s_{fn}} + 2\widehat{\sigma}^4_{s_{fn}} \tag{14}$$

where $L_{\frac{1}{2}}$ and $L_{\frac{3}{2}}$ are given by [13, 20, 21]

$$L_{\frac{1}{2}}(q) = e^{\frac{q}{2}}\left((1-q)\, I_0(\frac{q}{2}) + qI_1\left(\frac{q}{2}\right)\right) \tag{15}$$

$$L_{\frac{3}{2}}(q) = \frac{1}{3}e^{\frac{q}{2}}\left(\left(2q^2-6q+3\right) I_0\left(\frac{q}{2}\right) + \left(4q-2q^2\right) I_1\left(\frac{q}{2}\right)\right) \tag{16}$$

with I_0 and I_1 denoting order-0 and order-1 Bessel functions.

The full magnitude and power spectra are concatenated into a $2F \times 1$ vector $\mathbf{v}_n = [|s_{1n}| \ldots |s_{Fn}|\ |s_{1n}|^2 \ldots |s_{Fn}^2|]^T$ where F is the number of frequency bins. The mean $\widehat{\boldsymbol{\mu}}_{\mathbf{v}_n}$ and the covariance matrix $\widehat{\mathbf{\Sigma}}_{\mathbf{v}_n}$ of $\mathbf{v}_n$ are obtained by stacking $\hat{\boldsymbol{\mu}}_{\mathbf{v}_{fn}}$ and $\widehat{\mathbf{\Sigma}}_{\mathbf{v}_{fn}}$ in the same order, yielding a block-diagonal covariance matrix with four diagonal blocks.

3.2. To the static MFCCs and to the log-energy

In the second step, uncertainty is propagated to the vector $\mathbf{z}_n$ consisting of the static MFCCs and the log-energy. This vector may be computed using the nonlinear function $\mathcal{F}$

$$\mathbf{z}_n = \mathcal{F}(\mathbf{v}_n) = \bar{\mathbf{L}}\bar{\mathbf{D}}\log\left(\bar{\mathbf{M}}\bar{\mathbf{E}}\mathbf{v}_n\right) \tag{17}$$

where $\bar{\mathbf{E}}$, $\bar{\mathbf{M}}$, $\bar{\mathbf{D}}$ 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. More specifically, these matrices are defined as

$$\bar{\mathbf{E}} = \begin{bmatrix} \mathbf{Diag}(\mathbf{e}) & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_F \end{bmatrix} \quad \bar{\mathbf{M}} = \begin{bmatrix} \mathbf{M} & \mathbf{0} \\ \mathbf{0} & \mathbf{J}_F \end{bmatrix} \tag{18}$$

$$\bar{\mathbf{D}} = \begin{bmatrix} \mathbf{D} & \mathbf{0} \\ \mathbf{0} & 1 \end{bmatrix} \qquad \bar{\mathbf{L}} = \begin{bmatrix} \mathbf{Diag}(\mathbf{l}) & \mathbf{0} \\ \mathbf{0} & 1 \end{bmatrix} \tag{19}$$

where $\mathbf{I}_F$ is the identity matrix of size F, $\mathbf{J}_F$ is a $1 \times F$ vector of ones, $\mathbf{Diag}(.)$ is the diagonal matrix built from its vector

argument, $\mathbf{e}$ and $\mathbf{l}$ are the vectors of pre-emphasis and liftering coefficients, and $\mathbf{M}$ and $\mathbf{D}$ are the usual Mel filterbank and DCT matrices, respectively. Following the improvement demonstrated by vector Taylor series (VTS) over other techniques in [17], $\mathcal{F}$ is approximately linearized by its first-order VTS expansion [22] as $\mathbf{z}_n \approx \mathcal{F}\left(\mathbf{v}_n^0\right) + \mathcal{J}_{\mathcal{F}}\left(\mathbf{v}_n^0\right)\left(\mathbf{v}_n - \mathbf{v}_n^0\right)$. The mean and the covariance of $\mathbf{z}_n$ are therefore computed as

$$\widehat{\boldsymbol{\mu}}_{\mathbf{z}_n} = \mathcal{F}(\widehat{\boldsymbol{\mu}}_{\mathbf{v}_n}) = \bar{\mathbf{L}}\bar{\mathbf{D}}\log\left(\bar{\mathbf{M}}\bar{\mathbf{E}}\widehat{\boldsymbol{\mu}}_{\mathbf{v}_n}\right) \tag{20}$$

$$\widehat{\boldsymbol{\Sigma}}_{\mathbf{z}_n} = \mathcal{J}_{\mathcal{F}}\left(\widehat{\boldsymbol{\mu}}_{\mathbf{v}_n}\right)\widehat{\boldsymbol{\Sigma}}_{\mathbf{v}_n}\mathcal{J}_{\mathcal{F}}\left(\widehat{\boldsymbol{\mu}}_{\mathbf{v}_n}\right)^T \tag{21}$$

with the Jacobian matrix $\mathcal{J}_{\mathcal{F}}\left(\widehat{\boldsymbol{\mu}}_{\mathbf{v}_n}\right)$ given by

$$\mathcal{J}_{\mathcal{F}}\left(\widehat{\boldsymbol{\mu}}_{\mathbf{v}_n}\right) = \bar{\mathbf{L}}\bar{\mathbf{D}}\mathbf{Diag}\left(1/(\bar{\mathbf{M}}\bar{\mathbf{E}}\widehat{\boldsymbol{\mu}}_{\mathbf{v}_n})\right)\bar{\mathbf{M}}\bar{\mathbf{E}} \tag{22}$$

where the division is performed element-wise. The static MFCCs are subject to cepstral mean normalization [23]. For large enough number of time frames N, we treat the mean of the MFCCs over time as a deterministic quantity. Therefore, the mean MFCC vectors $\widehat{\boldsymbol{\mu}}_{\mathbf{z}_n}$ are normalized as usual while the covariance matrices are unchanged.

3.3. To the full feature vector

In the third step, we propagate the uncertainty about the static features to the full feature vector. The static features in the 4 preceding 4 frames, in the current frame, and in the following 4 frames are concatenated into a column vector $\bar{\mathbf{z}}_n = [\mathbf{z}_{n-4}^T\ \mathbf{z}_{n-3}^T \dots \mathbf{z}_{n+4}^T]^T$. The full feature vector $\mathbf{c}_n = [\mathbf{z}_n\ \Delta\mathbf{z}_n\ \Delta^2\mathbf{z}_n]$ can be expressed in matrix form as

$$\mathbf{c}_n = (\mathbf{A} \otimes \mathbf{I}_C)\bar{\mathbf{z}}_n \tag{23}$$

where $\otimes$ is the Kronecker product, $\mathbf{I}_C$ the identity matrix of size $C = 13$, and the matrix $\mathbf{A}$ is given by [23]

$$\mathbf{A} = \frac{1}{100}\begin{bmatrix} 0 & 0 & 0 & 0 & 100 & 0 & 0 & 0 & 0 \\ 0 & 0 & -20 & -10 & 0 & 10 & 20 & 0 & 0 \\ 4 & 4 & 1 & -4 & -10 & -4 & 1 & 4 & 4 \end{bmatrix}. \tag{24}$$

The mean and the covariance matrix of $\mathbf{c}_n$ are derived as

$$\widehat{\boldsymbol{\mu}}_{\mathbf{c}_n} = (\mathbf{A} \otimes \mathbf{I}_C)\widehat{\boldsymbol{\mu}}_{\bar{\mathbf{z}}_n} \tag{25}$$

$$\widehat{\boldsymbol{\Sigma}}_{\mathbf{c}_n} = (\mathbf{A} \otimes \mathbf{I}_C)\widehat{\boldsymbol{\Sigma}}_{\bar{\mathbf{z}}_n}(\mathbf{A} \otimes \mathbf{I}_C)^T \tag{26}$$

where $\widehat{\boldsymbol{\mu}}_{\bar{\mathbf{z}}_n}$ and $\widehat{\boldsymbol{\Sigma}}_{\bar{\mathbf{z}}_n}$ are obtained by concatenating $\widehat{\boldsymbol{\mu}}_{\mathbf{z}_{n-4}}, \dots, \widehat{\boldsymbol{\mu}}_{\mathbf{z}_{n-4}}$ into a column vector and $\widehat{\boldsymbol{\Sigma}}_{\mathbf{z}_{n-4}}, \dots, \widehat{\boldsymbol{\Sigma}}_{\mathbf{z}_{n+4}}$ into a block-diagonal matrix. Either the full uncertainty covariance matrix $\widehat{\boldsymbol{\Sigma}}_{\mathbf{c}_n}$ or its diagonal $\mathbf{diag}(\widehat{\boldsymbol{\Sigma}}_{\mathbf{c}_n})$ are then exploited to dynamically adapt the recognizer using Deng's uncertainty decoding rule [16].

3.4. Uncertainty scaling

The spectral domain uncertainty estimates in Section 2 rely on the assumption that uncertainty is independent across time-frequency bins. This assumption is not satisfied in practice, so that it translates into biased feature domain uncertainty estimates. The estimation of uncertainty across time-frequency bins appears to be a difficult far-end goal. In this work, we propose a simpler approach to compensate for this bias by scaling the coefficients of the uncertainty covariance matrix. More precisely, the diagonal covariance matrix and the full covariance matrix are scaled as

$$\mathbf{diag}(\widehat{\boldsymbol{\Sigma}}_{\mathbf{c}_n}^{\text{scaled}}) = \mathbf{Diag}(\mathbf{b})\mathbf{diag}(\widehat{\boldsymbol{\Sigma}}_{\mathbf{c}_n}) \tag{27}$$

$$\widehat{\boldsymbol{\Sigma}}_{\mathbf{c}_n}^{\text{scaled}} = \mathbf{Diag}(\mathbf{b})^{1/2}\widehat{\boldsymbol{\Sigma}}_{\mathbf{c}_n}\mathbf{Diag}(\mathbf{b})^{1/2} \tag{28}$$

where $\mathbf{b}$ is a 39×1 vector of nonnegative scaling coefficients (one per feature). Note that (28) preserves the positive-definiteness of the full covariance matrix. The scaling coefficients are optimized on development data for which the true speech signal is known then they are used for test data. The *oracle* (perfect) uncertainty covariance matrix is defined as [17]: $\boldsymbol{\Sigma}_{\mathbf{c}_n} = \left(\widehat{\boldsymbol{\mu}}_{\mathbf{c}_n} - \mathbf{c}_n\right)\left(\widehat{\boldsymbol{\mu}}_{\mathbf{c}_n} - \mathbf{c}_n\right)^T$. Where $\mathbf{c}_n$ is the true feature vector. The optimal coefficients are found by minimizing some measure of divergence D [24] between the scaled diagonal covariance matrix and the oracle diagonal covariance matrix:

$$\mathbf{b} = \arg\min_{\mathbf{b}} \mathrm{D}\left(\mathbf{diag}\left(\boldsymbol{\Sigma}_{\mathbf{c}_n}\right)|\mathbf{Diag}(\mathbf{b})\mathbf{diag}(\widehat{\boldsymbol{\Sigma}}_{\mathbf{c}_n})\right). \tag{29}$$

In the following, we employ the squared Euclidean distance, so that the scaling coefficients are found in closed form. Fig. 1 depicts the scaling coefficients estimated on the development data of the 2nd CHiME Challenge, which are subsequently applied to the test data. All scaling coefficients are larger than 1, which supports the claim in [25] that Wiener-based spectral domain uncertainty estimates are systematically underestimated.

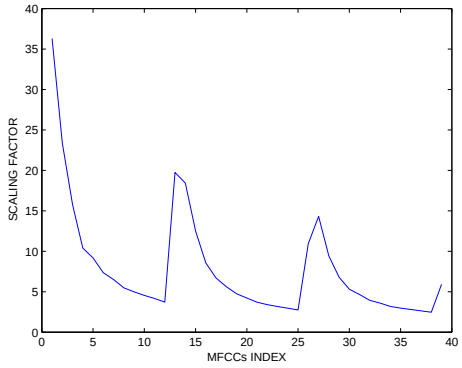

Fig. 1. Optimal scaling coefficients.

4. EXPERIMENTS

We assess our uncertainty propagation procedure on Track 1 of the 2nd CHiME Challenge [4]. Speech consists of 6-word

Uncertainty covariance matrix	Uncertain features	Scaling	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			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
diagonal	static	no	75.00	79.00	84.75	90.13	91.92	93.67	85.74	74.93	78.75	84.83	89.92	91.83	92.18	85.41
	dynamic	no	75.00	79.00	84.92	90.33	91.92	92.33	85.58	74.67	78.92	84.75	89.50	91.93	92.48	85.37
	all	no	76.93	79.17	85.92	90.00	92.00	93.75	86.29	76.13	78.75	85.56	89.68	91.75	93.50	85.89
	static	yes	76.50	79.25	85.67	90.17	92.58	92.58	86.13	77.00	78.51	85.82	89.58	91.50	93.52	85.98
	dynamic	yes	76.50	79.25	85.50	90.00	91.92	92.67	86.00	75.92	78.00	85.75	89.75	91.83	92.42	85.61
	all	yes	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**
full	static	no	76.75	79.33	85.50	90.33	92.33	93.67	86.31	76.40	79.33	85.50	89.75	91.92	92.38	85.88
	dynamic	no	76.75	79.17	85.75	90.33	92.00	93.83	86.30	76.17	79.25	85.50	89.75	91.92	92.55	85.85
	all	no	77.92	80.75	86.75	90.50	92.92	93.75	87.00	77.92	79.81	86.51	89.93	92.92	93.75	86.80
	static	yes	77.42	79.50	86.67	90.33	92.83	94.17	86.82	77.81	79.64	86.00	90.16	92.17	93.00	86.46
	dynamic	yes	77.92	80.00	86.75	90.17	92.17	93.50	86.75	77.86	79.92	86.17	89.83	91.93	92.42	86.35
	all	yes	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**

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

utterances 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). The task is to report the letter and digit keywords and performance is measured by keyword accuracy. The training set contains 500 noiseless reverberated 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 is applied to the development and test datasets using the Flexible Audio Source Separation Toolbox (FASST) [18] with the following settings optimized on the development set. A quadratic time-frequency representation on the auditory-motivated equivalent rectangular bandwidth (ERB) scale is used with 160 bands and half-overlapping 32 ms frames. The number of noise sources is set to 2. The power spectra of speech and noise are modeled by nonnegative matrix factorization (NMF) with 32 components and their spatial covariance matrices are modeled as full-rank [18]. Speaker-dependent acoustic models with diagonal Gaussian mixture model (GMM) densities are trained from the training set using the HTK baseline provided by the challenge organizers [4]. Uncertainty decoding is performed using the HTK baseline with Astudillo's patch[1] for diagonal uncertainty covariances and with our own patch for full uncertainty covariances.

4.2. Experimental results

ASR accuracies are reported in Table 1. Similar trends are observed on the development and the test data. On average over all SNRs in the test set, the baseline accuracy with conventional decoding (no uncertainty) is 85.01%. State-of-the-art uncertainty decoding with diagonal uncertainty covariance on static features (and no uncertainty on dynamic features) increases accuracy to 85.74%, that is 4% relative error rate reduction with respect to the baseline. Using the full uncertainty covariance, modeling the uncertainty over the dynamic features, and/or scaling the estimated uncertainties systematically improve the average performance. The best system using full uncertainty covariance on all features achieves 88.11% accuracy. This corresponds to 21% relative error rate reduction with respect to the baseline, that is five times more than the reduction achieved with diagonal uncertainty covariance for static features.

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

5. CONCLUSION

We presented a procedure for estimating the uncertainty about both static and dynamic MFCCs in the context of noise robust ASR based on uncertainty decoding. The estimated uncertainty is scaled by minimizing some measure of divergence with oracle uncertainty estimates on development data. The results demonstrate the benefit of modeling the uncertainty over both static and dynamic features, of scaling these estimates, and of using the full uncertainty covariance. In future work, we will seek to develop a method to estimate the inter-frame correlation between uncertainties.

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 under-

standing, part 1," *IEEE Signal Processing Magazine*, vol. 26, no. 3, pp. 75–80, May 2009.

[2] M. Wölfel and J. McDonough, *Distant Speech Recognition*, Wiley, 2009.

[3] T. Virtanen, R. Singh, and B. Raj, Eds., *Techniques for Noise Robustness in Automatic Speech Recognition*, Wiley, 2012.

[4] 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.

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

[6] C. Kim and R. Stern, "Power-normalized cepstral coefficients (PNCC) for robust speech recognition," in *Proc. ICASSP*, 2012, pp. 4101–4104.

[7] 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.

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

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] H. Kallasjoki, S. Keronen, G. J. Brown, J. F. Gemmeke, U. Remes, and K. J. Palomäki, "Mask estimation and sparse imputation for missing data speech recognition in multisource reverberant environments," in *Proc. CHiME*, 2011, pp. 58–63.

[15] 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.

[16] 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.

[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] 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.

[19] 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.

[20] I. Gradshteyn and I. Ryzhik, *Table of Intergral, Series and Product*, 1995.

[21] S. Rice, "Mathematical analysis of random noise," *Bell System Technical Journal*, vol. 23, 1944.

[22] 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.

[23] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK book*, 2002.

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

[25] K. Adiloğlu and E. Vincent, "Variational Bayesian inference for source separation and robust feature extraction," Tech. Rep. RT-0428, Inria, Aug. 2012.