Nonparametric uncertainty estimation and propagation for noise robust ASR

Dung T. Tran, Emmanuel M. Vincent, Denis Jouvet

Nonparametric uncertainty estimation and propagation for noise robust ASR

Dung T. Tran, *Student Member, IEEE,* Emmanuel Vincent, *Senior Member, IEEE,* and Denis Jouvet, *Member, IEEE,*

***Abstract*—We consider the framework of uncertainty propagation for automatic speech recognition (ASR) in highly nonstationary noise environments. Uncertainty is considered as the variance of speech distortion. Yet, its accurate estimation in the spectral domain and its propagation to the feature domain remain difficult. Existing methods typically rely on a single uncertainty estimator and propagator fixed by mathematical approximation. In this paper, we propose a new paradigm where we seek to learn more powerful mappings to predict uncertainty from data. We investigate two such possible mappings: linear fusion of multiple uncertainty estimators/propagators and nonparametric uncertainty estimation/propagation. In addition, a procedure to propagate the estimated spectral-domain uncertainty to the static Mel frequency cepstral coefficients (MFCCs), to the log-energy, and to their first- and second-order time derivatives is proposed. This results in a full uncertainty covariance matrix over both static and dynamic MFCCs. Experimental evaluation on Track 1 of the 2nd CHiME Challenge corpus resulted in up to 29% relative keyword error rate reduction with respect to speech enhancement alone.**

***Index Terms*—Uncertainty estimation, uncertainty decoding, nonparametric estimation, robust speech recognition.**

I. 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], [5] or hybrid techniques [6]–[8]. Uncertainty decoding [9]–[15] 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 posterior distribution with time-varying variance*. This variance or *uncertainty* is then used to dynamically adapt the acoustic model on each time frame for decoding. Decoding rules are available in closed form for hidden Markov models with mixture of Gaussian observation densities (HMM-GMMs) [9] and have recently started being investigated as a promising addition to deep neural network (DNN) based acoustic models [16].

The uncertainty is considered as the variance of speech distortion. It is derived from a parametric model of speech distortion accounting for additive noise or reverberation and it can be computed directly in the feature domain in which ASR operates [2], [10], [17]–[19] or estimated in the spectral domain then propagated to the feature domain [11], [12], [15], [20]–[23]. The latter approach typically performs best, as it allows speech enhancement to benefit from multichannel information in the spectral domain. We adopt this approach, which has led for instance to the best ASR accuracy in a real domestic environment as evaluated by the CHiME Challenge [24].

Most existing spectral-domain uncertainty estimators are fixed by mathematical approximation [11], [12], [15], [20]. Ozerov et al. [21], Astudillo et al. [20], and Adiloğlu et al. [25] showed that the uncertainty can be derived from the Wiener filter given estimates of the speech and noise power spectra, while Nesta et al. [15] considered the variance of the amplitude of each Fourier coefficient resulting from a prior Bernoulli model. In the feature domain, several uncertainty propagators have also been proposed based, e.g., on moment matching [3], unscented transform [11], and vector Taylor series (VTS) [26]. Due to the many approximations involved, e.g., linearization of the logarithm and decorrelation between consecutive speech frames, and to inaccurate estimation of the speech and noise power spectra, the resulting uncertainty is generally underestimated, so that the ASR performance remains lower than with perfect *oracle* uncertainty estimates [9], [21].

In order to address this issue, a few attemps have been made to learn uncertainty from data. Kolossa et al. [11] and Delcroix et al. [10] proposed to map the estimated spectral-domain or feature-domain uncertainty to the the actual uncertainty measured on development data in the same domain by means of a linear or affine mapping. Affine mappings depending on the HMM state were also investigated in [18]. Kallasjoki et al. [27] used Gaussian mixture model (GMM)-based regression to map some heuristic spectral-domain estimates into feature-domain uncertainty. Finally, Srinivasan et al. [22] employed a regression tree to map a spectral-domain binary mask into feature-domain uncertainty. These techniques are applicable to diagonal uncertainty covariance matrices only and they rely on a single mapping, which is applied in one domain only or jointly in the two domains. As a result, the word error rate (WER) reduction resulting from uncertainty decoding with a clean acoustic model and state-of-the-art speech enhancement front-end is typically on the order of 15% or less with respect to speech enhancement alone [13], [27], compared to 60% with the oracle uncertainty [13].

The major contribution of this work is the introduction of a framework to learn more powerful state-independent mappings by linear fusion of multiple uncertainty estimators/propagators and by nonparametric uncertainty estimation/propagation. Separate scalar mappings are used for each frequency bin and for each feature index, which make it possible to learn more complex mappings without overfitting. The mapping coefficients are obtained by minimizing some measure of divergence

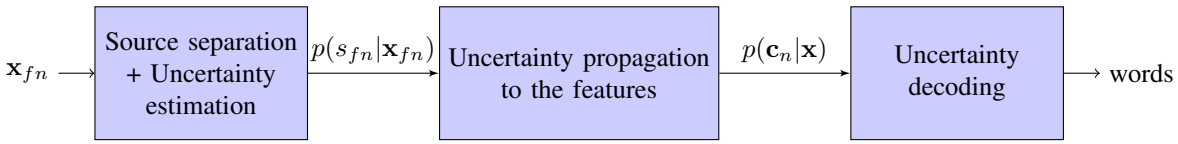

Fig. 1: Schematic diagram of the state-of-the-art uncertainty handling framework.

with respect to oracle uncertainty estimates on development data. We introduce a family of weighted nonnegative matrix factorization (NMF) algorithms to perform this minimization. In addition, we show how to compute full uncertainty covariance matrices on static MFCCs, log-energy, and their time derivatives [28], and how to use them in the proposed mapping framework. This extends our preliminary work [29] which focused on fusion of diagonal uncertainty estimates by unweighted NMF only. We perform an exhaustive evaluation of the impact of the weights and the divergence measure on the resulting ASR accuracy on Track 1 of the 2nd CHiME Challenge corpus using the reference HMM-GMM system provided by the challenge organizers as a baseline.

The paper is organized as follows. Section II introduces the conventional framework for uncertainty handling and the proposed extension to full uncertainty covariance. The proposed fusion and nonparametric estimation techniques are described in Section III and the corresponding estimation algorithms are presented in Section IV. ASR results are discussed in Section V. We conclude in Section VI.

II. Uncertainty handling — Background and extension to full uncertainty covariance

This section presents the uncertainty handling framework in the multichannel case. Fig. 1 shows the general schematic diagram including uncertainty estimation in the spectral domain, uncertainty propagation to the feature domain, and uncertainty decoding of the acoustic model.

A. Multichannel source separation

Let us consider a mixture of J speech and noise sources recorded by I microphones. In the complex short-time Fourier transform (STFT) domain, the observed multichannel signal $\mathbf{x}_{fn}$ can be modeled as [30]

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

where $\mathbf{y}_{jfn}$ is the so-called spatial image of the j-th source, and f and n are the frequency index and the frame index, respectively. Each source image is assumed to follow a zero-mean complex-valued Gaussian model

$$p(\mathbf{y}_{jfn}) = \mathcal{N}(\mathbf{y}_{jfn}; \mathbf{0}, v_{jfn}\mathbf{R}_{jf}) \tag{2}$$

whose parameters v_{jfn} and $\mathbf{R}_{jf}$ are the short-term power spectrum and the spatial covariance matrix of the source, respectively, which may be estimated using a number of alternative speech enhancement techniques [11], [15], [30]. Once estimated, these parameters are used to derive an estimate of the target speech source by multichannel Wiener filtering

$$\widehat{\boldsymbol{\mu}}_{\mathbf{y}_{jfn}} = \mathbf{W}_{jfn}\mathbf{x}_{fn} \tag{3}$$

with

$$\mathbf{W}_{jfn} = v_{jfn}\mathbf{R}_{jf}\Big(\sum_{j'} v_{j'fn}\mathbf{R}_{j'f}\Big)^{-1}. \tag{4}$$

The source spatial image is then downmixed into a single-channel source signal estimate $\widehat{\mu}_{s_{jfn}}$ as

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

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 [24], $I = 2$ and $\mathbf{u}_f^H = [0.5\ 0.5]$ for all f.

As an alternative to the STFT, quadratic time-frequency representations often improve enhancement by accounting for the local correlation between channels [30]. Expression (3) is not applicable anymore in that case since the mixture signal is represented by its empirical covariance matrix $\widehat{\mathbf{R}}_{\mathbf{xx}_{fn}}$ instead of $\mathbf{x}_{fn}$. A more general expression may however be obtained for the magnitude of the mean as

$$|\widehat{\mu}_{s_{jfn}}| = \Big(\mathbf{u}_f^H \mathbf{W}_{jfn} \widehat{\mathbf{R}}_{\mathbf{xx}_{fn}} \mathbf{W}_{jfn}^H \mathbf{u}_f\Big)^{1/2} \tag{6}$$

which is enough for subsequent feature computation.

B. Uncertainty estimation

The goal of uncertainty estimation is to obtain not only a point estimate of the target speech source s_{jfn} represented by the *mean* $\widehat{\mu}_{s_{jfn}}$ of its posterior distribution $p(s_{jfn}|\mathbf{x}_{fn})$ but also an estimate of how much the true (unknown) source signal may deviate from it, as represented by its posterior *variance* $\widehat{\sigma}^2_{s_{jfn}}$. Three state-of-the-art estimators may be considered.

1) Kolossa's estimator: Kolossa et al. [11] assumed the uncertainty to be proportional to the squared difference between the enhanced signal and the mixture

$$(\widehat{\sigma}^{\mathrm{Kol}}_{s_{jfn}})^2 = \alpha|\widehat{\mu}_{s_{jfn}} - x_{fn}|^2 \tag{7}$$

where $x_{fn} = \mathbf{u}_f^H \mathbf{x}_{fn}$ is the downmixed mixture signal and the scaling factor α is found by minimizing the Euclidean distance between the estimated uncertainty and the oracle uncertainty defined hereafter in Section IV-A.

2) Wiener estimator: Astudillo [12] later proposed to quantify uncertainty by the posterior variance of the Wiener filter. In the multichannel case, the posterior covariance matrix of $\mathbf{y}_{jfn}$ is given by [21]

$$\widehat{\boldsymbol{\Sigma}}_{\mathbf{y}_{jfn}} = (\mathbf{I}_I - \mathbf{W}_{jfn})\, v_{jfn}\mathbf{R}_{jf} \tag{8}$$

with $\mathbf{I}_I$ the identity matrix of size I. The variance of s_{jfn} is then easily derived as

$$(\widehat{\sigma}^{\mathrm{Wie}}_{s_{jfn}})^2 = \mathbf{u}_f^H \widehat{\boldsymbol{\Sigma}}_{\mathbf{y}_{jfn}} \mathbf{u}_f. \tag{9}$$

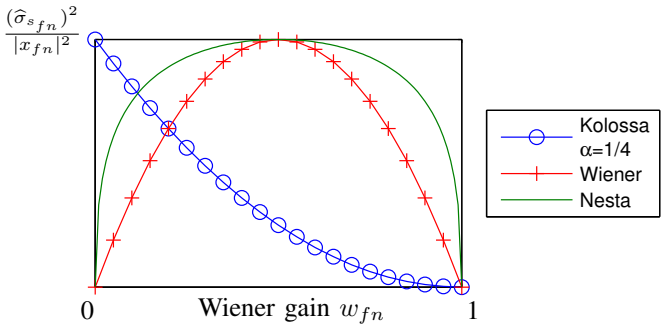

Fig. 2: Behavior of the uncertainty estimators. The notion of Wiener gain is defined later in Section III-A2.

3) Nesta's estimator: Recently, Nesta et al. [15] obtained a different estimate based on a binary speech/noise predominance model[1]:

$$(\widehat{\sigma}^{\text{Nes}}_{s_{jfn}})^2 = \widehat{p}_{jfn}(1-\widehat{p}_{jfn})|x_{fn}|^2 \tag{10}$$

where $\widehat{p}_{jfn} = \sqrt{v^{\text{tar}}_{jfn}}/(\sqrt{v^{\text{tar}}_{jfn}} + \sqrt{v^{\text{noi}}_{jfn}})$ and $v^{\text{tar}}_{jfn} = \mathbf{u}_f^H(v_{jfn}\mathbf{R}_{jf})\mathbf{u}_f$ and $v^{\text{noi}}_{jfn} = \mathbf{u}_f^H(\sum_{j'\neq j} v_{j'fn}\mathbf{R}_{j'f})\mathbf{u}_f$ are the prior variances of the target speech source j and the other sources, respectively. The behavior of the three estimators is illustrated in Fig. 2.

C. Propagation

From now on, we process one target speech source only and we drop index j for notation convenience. The posterior mean $\widehat{\mu}_{s_{fn}}$ and variance $\widehat{\sigma}^2_{s_{fn}}$ of the target speech source are propagated step by step to the feature domain for exploitation by the recognizer. At each step, the posterior is approximated as a Gaussian and represented by its mean and variance [12]. We use 39-dimensional feature vectors $\mathbf{c}_n$ consisting of 12 MFCCs, the log-energy, and their first- and second-order time derivatives. Propagation is achieved in three steps illustrated in Fig. 3.

The first step propagates the mean and the variance of s_{fn} to the magnitude domain. Since s_{fn} is complex-valued Gaussian, the k-th order moment M_k of $|s_{fn}|$ can be derived in closed form as [31]

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

where Γ is the gamma function and $L_{\frac{k}{2}}$ is the Laguerre polynominal.

[1]This formula was initially defined for the variance of $|s_{jfn}|$ [15], however we found it beneficial to use it for the variance of s_{jfn} instead.

The second step propagates the resulting means and variances to the static MFCCs

$$\mathbf{z}_n = \mathbf{Diag}(\mathbf{l})\mathbf{D}\log(\mathbf{M}\,\mathbf{Diag}(\mathbf{e})|\mathbf{s}_n|) \tag{12}$$

where $|\mathbf{s}_n| = [|s_{1n}|,\ldots,|s_{Fn}|]^T$ with F the number of frequency bins, $\mathbf{Diag}(.)$ is the diagonal matrix built from its vector argument, $\mathbf{e}$, $\mathbf{M}$, $\mathbf{D}$, and $\mathbf{l}$ are the $F\times 1$ vector of pre-emphasis coefficients, the $26\times F$ Mel filterbank matrix, the 12×26 discrete cosine transform (DCT) matrix, and the 12×1 vector of liftering coefficients, respectively. The propagation can be achieved by various techniques including the unscented transform (UT) and moment matching (MM), also known as the log-normal transform [3], [11], [12]. The mean and variance of the log-energy are also separately computed and concatenated with those of the MFCCs, yielding the mean $\widehat{\mu}_{z_n}$ and the variance $\widehat{\sigma}^2_{z_n}$ of the static feature vector.

In the third step, the uncertainty about the static features is propagated 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 $\bar{\mathbf{z}}_n = [\mathbf{z}^T_{n-4}\ \mathbf{z}^T_{n-3}\ldots\mathbf{z}^T_{n+4}]^T$. 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

$$\mathbf{c}_n = (\mathbf{A}\otimes\mathbf{I}_C)\bar{\mathbf{z}}_n \tag{13}$$

where $\otimes$ is the Kronecker product, $\mathbf{I}_C$ is the identity matrix of size $C=13$, and the matrix $\mathbf{A}$ is given by [32]

$$\mathbf{A} = \frac{1}{100}\begin{bmatrix} 0 & 0 & 0 & 0 & 100 & 0 & 0 & 0 & 0 \\ 0 & 0 & -20 & -10 & 0 & 10 & 20 & 0 & 0 \\ 4 & 4 & 1 & -4 & -10 & -4 & 1 & 4 & 4 \end{bmatrix} \tag{14}$$

The mean and the covariance matrix of the posterior distribution $p(\mathbf{c}_n|\mathbf{x})$ are derived as

$$\widehat{\boldsymbol{\mu}}_{\mathbf{c}_n} = (\mathbf{A}\otimes\mathbf{I}_C)\widehat{\boldsymbol{\mu}}_{\bar{\mathbf{z}}_n} \tag{15}$$

$$\widehat{\boldsymbol{\Sigma}}_{\mathbf{c}_n} = (\mathbf{A}\otimes\mathbf{I}_C)\widehat{\boldsymbol{\Sigma}}_{\bar{\mathbf{z}}_n}(\mathbf{A}\otimes\mathbf{I}_C)^T \tag{16}$$

where $\widehat{\boldsymbol{\mu}}_{\bar{\mathbf{z}}_n}$ and $\widehat{\boldsymbol{\Sigma}}_{\bar{\mathbf{z}}_n}$ are obtained by concatenating $\widehat{\boldsymbol{\mu}}_{\mathbf{z}_{n-4}},\ldots,\widehat{\boldsymbol{\mu}}_{\mathbf{z}_{n-4}}$ into a column vector and $\widehat{\boldsymbol{\Sigma}}_{\mathbf{z}_{n-4}},\ldots,\widehat{\boldsymbol{\Sigma}}_{\mathbf{z}_{n+4}}$ into a block-diagonal matrix. Only the diagonal of $\widehat{\boldsymbol{\Sigma}}_{\mathbf{c}_n}$ is retained.

D. Uncertainty decoding

The likelihood of the acoutic model given an observation is modified by marginalizing over the clean data as

$$p(\mathbf{x}_n|q) = \int_{\mathbf{c}_n} \frac{p(\mathbf{c}_n|\mathbf{x}_n)p(\mathbf{x}_n)}{p(\mathbf{c}_n)} p(\mathbf{c}_n|q)d\mathbf{c}_n \tag{17}$$

where $p(\mathbf{c}_n|q)$ is the clean acoustic model for state q. For low distortion levels, this can be approximated as [9], [20]

$$p(\mathbf{c}_n|q) \approx \int_{\mathbf{c}_n} p(\mathbf{c}_n|\mathbf{x}_n)p(\mathbf{c}_n|q)d\mathbf{c}_n. \tag{18}$$

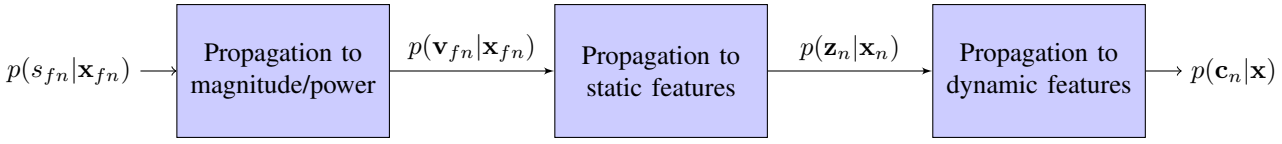

Fig. 3: Schematic diagram of uncertainty propagation from the complex-valued STFT domain to the feature domain.

In the case when $p(\mathbf{c}_n|q)$ is a GMM with M components with weights, means, and diagonal covariance matrices denoted as ω_m, $\boldsymbol{\mu}_m$, and $\boldsymbol{\Sigma}_m$, respectively, the modified likelihood (18) can be computed in closed form as [9]

$$p(\mathbf{c}_n|q) \approx \sum_{m=1}^{M} \omega_m \mathcal{N}(\widehat{\boldsymbol{\mu}}_{\mathbf{c}_n}; \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m + \widehat{\boldsymbol{\Sigma}}_{\mathbf{c}_n}). \quad (19)$$

E. Extension to a full uncertainty covariance matrix

In practice, we found the restriction to a diagonal uncertainty covariance matrix to limit ASR performance. This can be explained by the fact that source separation errors are often localized in the time-frequency plane, which results in correlation of uncertainties across MFCC coefficients. We now explain how to compute a full uncertainty covariance for both static and dynamic MFCCs, following the same three steps as in Fig. 3.

In the first step, we do not only compute the scalar moments of the magnitude and power spectra but also their cross-moments. 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

$$\hat{\boldsymbol{\mu}}_{\mathbf{v}_{fn}} = \begin{bmatrix} M_1 \\ M_2 \end{bmatrix} \quad (20)$$

$$\widehat{\boldsymbol{\Sigma}}_{\mathbf{v}_{fn}} = \begin{bmatrix} M_2 - M_1^2 & M_3 - M_1 M_2 \\ M_3 - M_1 M_2 & M_4 - M_2^2 \end{bmatrix} \quad (21)$$

where M_k is defined in (11). The full magnitude and power spectra are concatenated into a $2F \times 1$ vector $\mathbf{v}_n = [|s_{1n}| \dots |s_{Fn}|\ |s_{1n}|^2 \dots |s_{Fn}^2|]^T$. The mean $\widehat{\boldsymbol{\mu}}_{\mathbf{v}_n}$ and the covariance matrix $\widehat{\boldsymbol{\Sigma}}_{\mathbf{v}_n}$ of $\mathbf{v}_n$ are obtained by stacking $\hat{\boldsymbol{\mu}}_{\mathbf{v}_{fn}}$ and $\widehat{\boldsymbol{\Sigma}}_{\mathbf{v}_{fn}}$ in the same order, yielding a block-diagonal covariance matrix with four diagonal blocks.

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 $\mathcal{F}$

$$\mathbf{z}_n = \mathcal{F}(\mathbf{v}_n) = \bar{\mathbf{L}}\bar{\mathbf{D}} \log\left(\bar{\mathbf{M}}\bar{\mathbf{E}}\mathbf{v}_n\right) \quad (22)$$

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

$$\bar{\mathbf{E}} = \begin{bmatrix} \mathbf{Diag}(\mathbf{e}) & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_F \end{bmatrix} \quad \bar{\mathbf{M}} = \begin{bmatrix} \mathbf{M} & \mathbf{0} \\ \mathbf{0} & \mathbf{J}_F \end{bmatrix} \quad (23)$$

$$\bar{\mathbf{D}} = \begin{bmatrix} \mathbf{D} & \mathbf{0} \\ \mathbf{0} & 1 \end{bmatrix} \qquad \bar{\mathbf{L}} = \begin{bmatrix} \mathbf{Diag}(\mathbf{l}) & \mathbf{0} \\ \mathbf{0} & 1 \end{bmatrix} \quad (24)$$

where $\mathbf{I}_F$ is the identity matrix of size F, $\mathbf{J}_F$ is a $1 \times F$ vector of ones, and $\mathbf{e}$, $\mathbf{M}$, $\mathbf{D}$, and $\mathbf{l}$ are defined as in Section II-C. Following the improvement demonstrated by VTS over UT and MM in [21], $\mathcal{F}$ is approximately linearized by its first-order VTS expansion around its mean [26]

$$\mathbf{z}_n \approx \mathcal{F}\left(\widehat{\boldsymbol{\mu}}_{\mathbf{v}_n}\right) + \mathcal{J}_{\mathcal{F}}\left(\widehat{\boldsymbol{\mu}}_{\mathbf{v}_n}\right)\left(\mathbf{v}_n - \widehat{\boldsymbol{\mu}}_{\mathbf{v}_n}\right). \quad (25)$$

The mean and the covariance of $\mathbf{z}_n$ are therefore computed as

$$\widehat{\boldsymbol{\mu}}_{\mathbf{z}_n} = \mathcal{F}(\widehat{\boldsymbol{\mu}}_{\mathbf{v}_n}) = \bar{\mathbf{L}}\bar{\mathbf{D}} \log\left(\bar{\mathbf{M}}\bar{\mathbf{E}}\widehat{\boldsymbol{\mu}}_{\mathbf{v}_n}\right) \quad (26)$$

$$\widehat{\boldsymbol{\Sigma}}_{\mathbf{z}_n} = \mathcal{J}_{\mathcal{F}}\left(\widehat{\boldsymbol{\mu}}_{\mathbf{v}_n}\right) \widehat{\boldsymbol{\Sigma}}_{\mathbf{v}_n} \mathcal{J}_{\mathcal{F}}\left(\widehat{\boldsymbol{\mu}}_{\mathbf{v}_n}\right)^T \quad (27)$$

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

$$\mathcal{J}_{\mathcal{F}}\left(\widehat{\boldsymbol{\mu}}_{\mathbf{v}_n}\right) = \bar{\mathbf{L}}\bar{\mathbf{D}}\,\mathbf{Diag}\left(1/(\bar{\mathbf{M}}\bar{\mathbf{E}}\widehat{\boldsymbol{\mu}}_{\mathbf{v}_n})\right)\bar{\mathbf{M}}\bar{\mathbf{E}} \quad (28)$$

where the division is performed element-wise. The static MFCCs are subject to cepstral mean normalization [32]. 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.

In the third step, the mean $\widehat{\boldsymbol{\mu}}_{\bar{\mathbf{c}}_n}$ and the covariance matrix $\widehat{\boldsymbol{\Sigma}}_{\mathbf{c}_n}$ of the full feature vector $\mathbf{c}_n$ as obtained as in (15)–(16) and the full matrix $\widehat{\boldsymbol{\Sigma}}_{\mathbf{c}_n}$ is retained. Note that the covariance matrices $\boldsymbol{\Sigma}_m$ of the clean model remain diagonal, however.

III. Proposed fusion and nonparametric estimation framework

After having reviewed the state of the art and the proposed extension to full uncertainty covariance, we now present the proposed fusion and nonparametric estimators. The learning of the corresponding fusion weights and kernel weights is addressed later in Section IV. We first focus on uncertainty estimation in the spectral domain and then on uncertainty propagation to the feature domain.

A. Fused/nonparametric uncertainty estimation

Looking back at Fig. 2, we see that the three spectral-domain uncertainty estimators introduced in Section II-B 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 learn the optimal estimator from data.

1) Fusion: A first idea is to fuse multiple uncertainty estimators by linear combination in order to obtain a more accurate estimator. This is a form of early fusion. In the following, we assume that the fusion weights depend on frequency f but that they are independent of the signal-to-noise ratio and the HMM state. Indeed, the signal-to-noise ratio in each time-frequency bin is typically unknown and the uncertainty represents the variance of speech distortion, which depends on the speech enhancement technique but not on the HMM-GMM subsequently used for decoding.

Denoting by E the number of estimators, the fused estimator $(\widehat{\sigma}_{s_{fn}}^{\text{fus}})^2$ can be expressed as

$$(\widehat{\sigma}_{s_{fn}}^{\text{fus}})^2 = \sum_{e=1}^{E} w_{s_f}^e \, (\widehat{\sigma}_{s_{fn}}^e)^2 \quad (29)$$

where $(\widehat{\sigma}_{s_{fn}}^e)^2$ is one of the original estimators in (7), (9), (10), and $w_{s_f}^e$ are the fusion coefficients. The fusion coefficients are constrained to be nonnegative so that the fused estimator is always nonnegative. Stacking the original uncertainty estimates over all time frames into a $E \times N$ matrix $\widehat{\boldsymbol{\Lambda}}_{s_f}$ and the fused estimates into a $1 \times N$ vector $\widehat{\boldsymbol{\lambda}}_{s_f}^{\text{fus}}$ for each frequency f, where

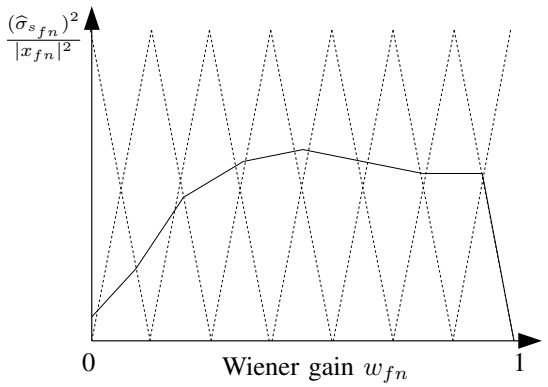

Fig. 4: Example triangular kernels and resulting mapping.

N is the number of frames, (29) can be written in matrix form as

$$\widehat{\boldsymbol{\lambda}}_{s_f}^{\text{fus}} = \mathbf{w}_{s_f} \widehat{\boldsymbol{\Lambda}}_{s_f} \tag{30}$$

where $\mathbf{w}_{s_f}$ is the $1 \times E$ vector of fusion coefficients.

In order to compensate for possible additive bias in the original uncertainty estimates, we also add a nonnegative frequency-dependent bias. This is simply achieved by adding a row of ones to the matrix $\widehat{\boldsymbol{\Lambda}}_{s_f}$ and a corresponding coefficient in $\mathbf{w}_{s_f}$ for the bias value.

2) Nonparametric mapping: Although the fused uncertainty estimator potentially improves over the original fixed estimators, its shape remains constrained by these original estimators. This motivates us to learn the full shape of the estimator from data in a nonparametric fashion. To do this, we express the uncertainty in the same way as in (29), but we replace the existing estimators $(\widehat{\sigma}^e_{s_{fn}})^2$ by *kernels* $(\widehat{\sigma}^e_{s_{fn}})^2$.

As it appears from Section II-B and Fig. 2, most uncertainty estimators share two properties. First, the estimated uncertainty is proportional to the mixture power spectrum. Second, they can be expressed as a function of the Wiener gain, that is the ratio of the speech power spectrum and the mixture power spectrum. In the multichannel case, we define the Wiener gain w_{fn} as

$$w_{fn} = \frac{1}{I} \operatorname{tr}(\mathbf{W}_{fn}) \tag{31}$$

where $\mathbf{W}_{fn}$ is the multichannel Wiener filter defined in (4). By property of the multichannel Wiener filter, w_{fn} is real-valued and between 0 and 1.

Based on these two properties, we define the kernels as

$$(\widehat{\sigma}^e_{s_{fn}})^2 = |x_{fn}|^2 \mathrm{b}^e(w_{fn}) \tag{32}$$

where $\mathrm{b}^e(.)$ are a set of normalized kernel functions on $[0, 1]$ indexed by $e \in \{1, \ldots, E\}$. In the following, we choose triangular kernels

$$\mathrm{b}^e(w_{fn}) = (E-1)\max(0, 1-|(E-1)w_{fn} - (e-1)|) \tag{33}$$

which results in a piecewise linear mapping. The weights $w^e_{s_f}$ encode the value of the mapping when $w_{nf} = (e-1)/(E-1)$. Fig. 4 shows the shape of the kernels and the resulting mapping. The number of kernels E governs the precision of the uncertainty estimates. A bigger E potentially increases accuracy, but too large E results in overfitting.

B. Fused/nonparametric uncertainty propagation with diagonal covariance

The estimated spectral-domain uncertainties are propagated to the feature domain by VTS. Although this results in better feature-domain uncertainties than UT or MM, we found experimentally these feature-domain uncertainties to be underestimated. This may be due to the initial assumption that spectral-domain uncertainties are independent across time-frequency bins, as well as to the approximations involved in VTS. The estimation of the correlation of uncertainties across time-frequency bins appears to be a difficult far-end goal. Therefore, we propose to learn from data how to correct the estimated uncertainties. Let us consider first the case of a diagonal uncertainty covariance matrix $\widehat{\boldsymbol{\Sigma}}_{\mathbf{c}_n}$.

1) Rescaling: A first way of correcting the underestimation is to rescale the coefficients of one estimated uncertainty covariance matrix as [10]

$$\widehat{\boldsymbol{\Sigma}}^{\text{scaled}}_{\mathbf{c}_n} = \mathbf{Diag}(\mathbf{g}) \widehat{\boldsymbol{\Sigma}}_{\mathbf{c}_n} \tag{34}$$

where $\mathbf{g}$ is a 39×1 vector of nonnegative scaling coefficients.

2) Fusion: One may also keep several spectral-domain uncertainty estimates by applying the fused estimator in (29) with different values of the fusion weights $w^e_{s_f}$ and propagate each of them to the feature domain, yielding P feature-domain uncertainty estimates $(\widehat{\sigma}^p_{c_{in}})^2$ indexed by p for each feature index i. These uncertainty estimates may then be fused similarly to above. In the following, we assume that the fusion weights depend on the feature index i but that they are independent of the signal-to-noise ratio and the HMM state. The fused uncertainty propagator is obtained as

$$\widehat{\boldsymbol{\lambda}}_{c_i}^{\text{fus}} = \mathbf{w}_{c_i} \widehat{\boldsymbol{\Lambda}}_{c_i} \tag{35}$$

where $\widehat{\boldsymbol{\lambda}}_{c_i}^{\text{fus}}$ is the $1 \times N$ vector of fused estimates, $\mathbf{w}_{c_i}$ is the $1 \times P$ vector of fusion coefficients and $\widehat{\boldsymbol{\Lambda}}_{c_i}$ is a $P \times N$ matrix whose elements are $(\widehat{\sigma}^p_{c_{in}})^2$. This expression generalizes (34) to the case of multiple feature-domain uncertainty estimates.

3) Nonparametric mapping: Finally, we can estimate the uncertainty nonparametrically by applying (35) where $(\widehat{\sigma}^p_{c_{in}})^2$ are a set of kernels indexed by $p \in \{1, \ldots, P\}$. In the following, we choose triangular kernels defined as

$$(\widehat{\sigma}^p_{c_{in}})^2 = (P-1)\max(0, 1-|(P-1)(\bar{\sigma}_{c_{in}})^2 - (p-1)|) \tag{36}$$

where $(\bar{\sigma}_{c_{in}})^2$ is the result of linearly normalizing the nonparametric feature-domain uncertainty estimator $(\widehat{\sigma}_{c_{in}})^2$ to the interval $[0, 1]$ for each feature index i.

C. Fused/nonparametric uncertainty propagation with full covariance

To exploit the full benefit of uncertainty decoding, a full uncertainty covariance matrix is needed. The extension of (35) to full covariance matrices is not trivial, however. Therefore, we first estimate the rescaling or fusion weights from the diagonal $\mathbf{diag}(\widehat{\boldsymbol{\Sigma}}_{\mathbf{c}_n})$, where $\mathbf{diag}(.)$ is the vector consisting of

the diagonal entries of its matrix argument, and we apply them to the full matrix $\widehat{\mathbf{\Sigma}}_{\mathbf{c}_n}$ using the following heuristic approach. We compute a vector of equivalent rescaling coefficients as

$$\mathbf{g} = \frac{\widehat{\mathbf{\Sigma}}_{\mathbf{c}_n}^{\text{fus}}}{\widehat{\mathbf{\Sigma}}_{\mathbf{c}_n}} \tag{37}$$

where the division is performed element-wise, and we apply them to the full matrix as

$$\widehat{\mathbf{\Sigma}}_{\mathbf{c}_n}^{\text{fus}} = \mathbf{Diag}(\mathbf{g})^{1/2} \widehat{\mathbf{\Sigma}}_{\mathbf{c}_n} \mathbf{Diag}(\mathbf{g})^{1/2}. \tag{38}$$

This approach is applicable to the three methods presented above (rescaling, fusion, and nonparametric mapping) and it ensures that the positive semi-definiteness of the full covariance matrix is preserved.

IV. Learning of fusion/nonparametric coefficients

The uncertainty estimators presented in the previous section rely on a set of weights. We propose to learn these weights on development data for which the true speech signal is known such that the resulting uncertainty estimates are as close as possible to the oracle uncertainty.

A. Oracle estimators

The oracle uncertainty refers to the best possible uncertainty that can be estimated when the clean data is known. It can be computed in the spectral domain as

$$(\sigma_{s_{fn}})^2 = |\widehat{\mu}_{s_{fn}} - s_{fn}|^2 \tag{39}$$

and in the feature domain in the diagonal case as

$$(\sigma_{c_{in}})^2 = |\widehat{\mu}_{c_{in}} - c_{in}|^2 \tag{40}$$

where s_{fn} and c_{in} are the clean complex-valued STFT coefficients and the clean features, respectively[2].

B. Weighted divergence measures

For the three proposed approaches (rescaling, fusion, and nonparametric mapping), we optimize the weights on development data by minimizing some measure of divergence between the estimated uncertainties and the oracle uncertainties. There are many possible choices of divergences, including the well-known Itakura-Saito (IS), Kullback-Leibler (KL), and squared Euclidean (EUC) divergences, which belong to the family of β-divergences with $\beta = 0$, 1, or 2, respectively [33], and more general Bregman divergences. These divergences can be characterized by two main properties: their shape, i.e., how they penalize underestimation and overestimation with respect to each other, and their scale, i.e., how they vary with respect to the scale of the input.

The scale property is particularly important in our context since the scale of speech spectra is extremely variable from one frame to another and the scale of features is extremely variable from one feature index to another. We therefore consider the minimization of the following weighted β-divergence measures

$$\mathbf{w}_{s_f} = \arg\min_{\mathbf{w}_{s_f} \geq 0} \sum_n |x_{fn}|^{\alpha-2\beta} d_\beta \left((\sigma_{s_{fn}})^2 | (\mathbf{w}_{s_f} \widehat{\mathbf{\Sigma}}_{s_f})_n \right) \tag{41}$$

$$\mathbf{w}_{c_i} = \arg\min_{\mathbf{w}_{c_i} \geq 0} \sum_n (\tilde{\sigma}_{c_i})^\alpha d_\beta \left((\sigma_{c_{in}})^2 | (\mathbf{w}_{c_i} \widehat{\mathbf{\Sigma}}_{c_i})_n \right) \tag{42}$$

where $(\sigma_{s_{fn}})^2$ and $(\sigma_{c_{in}})^2$ are the oracle uncertainties in time frame n, $(\mathbf{w}_{s_f} \widehat{\mathbf{\Sigma}}_{s_f})_n$ and $(\mathbf{w}_{c_i} \widehat{\mathbf{\Sigma}}_{c_i})_n$ are the estimated uncertainties in that time frame,

$$d_\beta(x|y) = \frac{1}{\beta(\beta-1)} (x^\beta + (\beta-1)y^\beta - \beta x y^{\beta-1}) \tag{43}$$

is the β-divergence between two scalars, and $\tilde{\sigma}_{c_i}$ is the standard deviation of the features defined by

$$\tilde{\sigma}_{c_i} = \sqrt{\frac{1}{N}\sum_n c_{in}^2 - \left(\frac{1}{N}\sum_n c_{in}\right)^2}. \tag{44}$$

The exponent α governs the scale of the divergence. For instance, the value $\alpha = 0$ corresponds to a scale-invariant spectral-domain divergence[3]. The spectral-domain divergences corresponding to $\alpha = 1$ and $\alpha = 2$ scale with the magnitude and the squared magnitude of the signal, respectively.

C. Multiplicative update rules

The optimization problems (41) and (42) are instances of weighted nonnegative matrix factorization (NMF) [34]. The fusion coefficients are found by applying the following iterative multiplicative updates [33]:

$$\mathbf{w}_{s_f} \leftarrow \mathbf{w}_{s_f} \odot \frac{\left(\Gamma_{s_f} \odot (\mathbf{w}_{s_f} \widehat{\mathbf{\Sigma}}_{s_f})^{\beta-2} \odot \mathbf{\Sigma}_{s_f}\right) (\widehat{\mathbf{\Sigma}}_{s_f})^T}{\left(\Gamma_{s_f} \odot (\mathbf{w}_{s_f} \widehat{\mathbf{\Sigma}}_{s_f})^{\beta-1}\right) (\widehat{\mathbf{\Sigma}}_{s_f})^T} \tag{45}$$

$$\mathbf{w}_{c_i} \leftarrow \mathbf{w}_{c_i} \odot \frac{\left(\Gamma_{c_i} \odot (\mathbf{w}_{c_i} \widehat{\mathbf{\Sigma}}_{c_i})^{\beta-2} \odot \mathbf{\Sigma}_{c_i}\right) (\widehat{\mathbf{\Sigma}}_{c_i})^T}{\left(\Gamma_{c_i} \odot (\mathbf{w}_{c_i} \widehat{\mathbf{\Sigma}}_{c_i})^{\beta-1}\right) (\widehat{\mathbf{\Sigma}}_{c_i})^T} \tag{46}$$

where $\odot$ denotes element-wise multiplication and powers are computed element-wise, $\mathbf{\Sigma}_{s_f}$ and $\mathbf{\Sigma}_{c_i}$ are the $1 \times N$ vectors of oracle uncertainties, Γ_{s_f} is the $1 \times N$ vector with entries $|x_{fn}|^{\alpha-2\beta}$, and Γ_{c_i} is the $1 \times N$ vector whose entries are all equal to $(\tilde{\sigma}_{c_i})^\alpha$.

The coefficients $\mathbf{w}_{s_f}$ and $\mathbf{w}_{c_i}$ estimated on the development data are then applied to the test data.

V. Experimental evaluation

We assess the proposed fusion framework on Track 1 of the 2nd CHiME Challenge [24]. 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). The task is to report

[2] In the case when a quadratic time-frequency representation is used instead of the STFT, the spectral-domain oracle uncertainty can still be computed by expressing (39) in terms of the empirical mixture covariance matrix $\widehat{\mathbf{R}}_{\mathbf{xx}_{fn}}$ and the correlation vector $\widehat{\mathbf{r}}_{\mathbf{x}s_{fn}}$

[3] Note that, according to (32) and (43), the β-divergence in (41) scales with $|x_{fn}|^{2\beta}$ hence the normalized β-divergence scales with $|x_{fn}|^\alpha$.

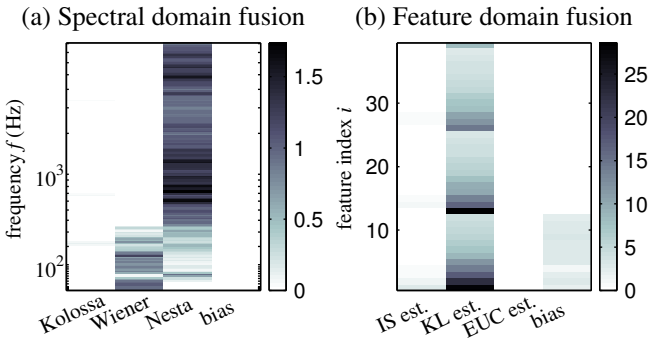

Fig. 5: Learned fusion coefficients (a) $\mathbf{w}_{s_f}$ and (b) $\mathbf{w}_{c_i}$ with $\alpha = 0$ and $\beta = 1$. IS est., KL est., and EUC est. refer to the feature-domain estimators resulting from spectral-domain fusion with $\alpha = 0$ and $\beta = 0$, 1, or 2, respectively, and "bias" refers to the additive bias as explained in Section III-A1.

the letter and digit tokens, which are the two most difficult words in the utterances. Performance is measured in terms of keyword accuracy, that is the percentage of tokens recognized correctly. The training set contains 500 noiseless reverberated utterances corresponding to 0.14 hour per speaker. The development set and the test set contain 600 utterances each, corresponding to 0.16 hour per SNR.

A. Experimental setup

Speech enhancement is applied to the development and test datasets using the Flexible Audio Source Separation Toolbox (FASST) [30] with the same settings as in [28], which were optimized on the development set. A quadratic time-frequency representation on the Equivalent Rectangular Bandwith (ERB) scale was used. Speaker-dependent acoustic models are trained from the reverberated noiseless training set using the HTK baseline provided by the challenge organizers [24]. Decoding is performed using the HTK baseline with Astudillo's patch[4] for diagonal uncertainty covariances and with our own patch[5] for full uncertainty covariances. These patches dynamically adapt the GMM observation probabilities as described in Section II-D.

B. Estimated fusion/nonparametric coefficients

Fig. 5 represents the optimal fusion coefficients estimated on the development set for Kolossa's, Wiener, and Nesta's estimators. The resulting spectral-domain estimator is a scaled version of Nesta's at higher frequencies and a combination of Wiener and Nesta's at lower frequencies, while the resulting feature-domain estimator is mostly a scaled version of the KL-fused estimator with some additive bias on the static features.

Fig. 6 illustrates the nonparametric mappings learned from the development set. Contrary to Wiener and Nesta's estimators, the learned spectral-domain mapping has an asymmetric shape and it varies with frequency. It is interesting to note that the mapping value at very low frequencies remains large for

[4]http://www.astudillo.com/ramon/research/stft-up/
[5]http://full-ud-htk.gforge.inria.fr/

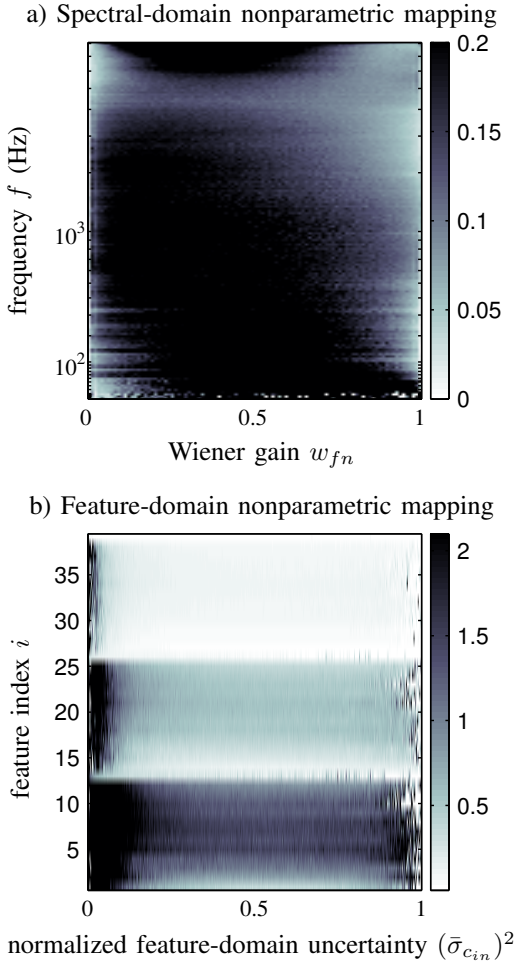

Fig. 6: Learned nonparametric mappings (a) $\mathbf{w}_{s_f}$ with $\alpha = 2$, $\beta = 1$, $E = 200$, and (b) $\mathbf{w}_{c_i}$ with $\alpha = 0$, $\beta = 1$, $P = 400$.

Wiener gain values close to 1, which is consistent with the fact that there is no speech at these frequencies. The learned feature-domain mapping is more difficult to interpret, as it not monotonously increasing with respect to $(\bar{\sigma}_{c_{in}})^2$ as one would expect. Nevertheless, the learned uncertainty is larger for static MFCCs than for delta and delta-delta MFCCs, which is consistent with the fact that the value range is larger for the former.

C. ASR results

We now evaluate the impact of the proposed uncertainty estimators on keyword accuracy. In all experiments, the average accuracies on all development or all test data have a 95% confidence interval on the order of $\pm 0.8\%$.

1) Full uncertainty covariance: Table I assesses our baseline system with Wiener uncertainty estimation and VTS uncertainty propagation. Similar results were observed with Nesta's estimator (not shown in the table). After source separation, the accuracy with conventional decoding (no uncertainty) is 85.01% on average over all SNRs in the test set. State-of-the-art uncertainty decoding with diagonal uncertainty covariance increases accuracy to 86.29%. The proposed system using

Uncertainty covariance matrix	Uncertain features	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	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	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	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**
full	static	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	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	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**

TABLE I: Baseline keyword accuracy (in %) achieved with Wiener+VTS.

			Test set							Development set						
estimation	propagation	cov.	-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
Wiener	VTS + scaling	diag.	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
fusion	VTS		78.33	80.17	85.92	90.08	92.08	94.17	86.97	78.33	80.17	85.75	89.92	92.50	93.50	86.69
fusion	fusion		80.50	82.17	88.25	91.33	92.50	93.58	88.05	80.00	81.92	87.25	91.50	92.25	93.08	87.66
nonparametric	VTS		80.00	81.92	87.25	91.50	92.25	93.08	87.66	79.75	81.67	87.17	89.75	91.58	93.50	87.23
nonparametric	nonparametric		81.75	83.50	88.33	91.08	92.75	93.00	**88.40**	80.83	82.00	88.25	90.50	92.67	93.50	**87.95**
Wiener	VTS + scaling	full	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
fusion	VTS		81.00	81.50	87.33	91.00	93.50	94.92	88.20	80.33	81.33	87.17	91.08	92.25	93.50	87.68
fusion	fusion		83.17	84.33	89.75	91.17	93.33	93.33	89.18	83.33	83.25	88.42	91.50	93.17	93.17	88.73
nonparametric	VTS		82.33	82.58	88.00	92.00	93.33	93.92	88.69	81.42	82.00	87.92	91.75	92.50	93.75	88.22
nonparametric	nonparametric		83.78	84.92	88.42	91.25	93.75	94.42	**89.42**	83.00	83.50	88.67	92.08	93.00	93.75	**89.00**

TABLE II: Keyword accuracy (in %) achieved with various fusion or nonparametric mapping schemes. This is to be compared to the baseline Wiener+VTS performance in Table I.

full uncertainty covariance on all features achieves 87.00% accuracy that is 13% relative error rate reduction compared to conventional decoding. The results systematically improve when modeling the uncertainty over both static and dynamic features. Overall, this validates the benefit of full uncertainty covariance over both static and dynamic features.

2) Fusion and nonparametric mapping: Table II shows the results achieved with fusion or nonparametric mapping with the optimal values of α and β. Similar trends are seen for diagonal and full uncertainty covariances. In the following, we comment the latter only.

Starting from the above Wiener+VTS baseline, the average accuracy on the test set improves to 88.11% by feature-domain uncertainty rescaling. This is already a significant improvement, which confirms that the uncertainties estimated by state-of-the-art techniques must be rescaled in order to match the actual uncertainty in the data.

By fusing Kolossa's, Wiener, and Nesta's uncertainty estimators, performance improves to 88.20%. Further fusing the IS-fused estimator, the KL-fused estimator and the EUC-fused estimator in the feature domain yields 89.18% accuracy, that is 28% relative error rate reduction compared to conventional decoding and 9% with respect to rescaling.

Finally, using a nonparametric mapping in both the spectral and the feature domain resulted in 89.42% keyword accuracy, that is 29% relative error rate reduction compared to conventional decoding and 2% with respect to fusion. This is about twice larger than the improvements due to uncertainty decoding reported in the state of the art, that are typically on the order of 15% relative or less compared to conventional decoding [18], [27]. These results are also among the top three for Track 1 of the 2nd CHiME Challenge [24] and the best ones to our knowledge without using other features than MFCCs or a multi-stream speech recognizer.

Method		fusion			nonparametric		
cov.	β \ α	0	1	2	0	1	2
diag.	0	85.16	85.94	86.47	86.68	87.16	87.20
	1	86.55	86.65	**86.69**	86.92	87.18	**87.23**
	2	86.18	86.20	86.53	86.50	87.00	87.15
full	0	86.74	87.00	87.16	88.00	88.14	88.25
	1	87.58	87.63	**87.68**	87.93	88.12	**88.29**
	2	87.16	87.23	87.33	87.78	88.04	88.15

TABLE III: Keyword accuracy (in %) on the development set for various weighted divergence choices in the spectral domain.

Method		fusion			nonparametric		
cov.	β \ α	0	1	2	0	1	2
diag.	0	86.49	86.02	85.91	86.85	86.34	86.17
	1	**87.33**	87.00	86.79	**87.95**	87.64	87.20
	2	86.72	86.58	86.27	87.22	87.00	86.68
full	0	88.57	88.51	88.49	88.73	88.65	88.41
	1	**88.73**	88.66	88.56	**89.00**	88.82	88.63
	2	88.63	88.62	88.16	88.84	88.60	88.56

TABLE IV: Keyword accuracy (in %) on the development set for various weighted divergence choices in the feature domain.

3) Impact of weighted divergence choice: Tables III and IV complete these results by evaluating the choice of the divergence parameters α and β. In either case, this choice does not have a significant impact on the resulting keyword accuracy. The best choices appear to be weighted KL-divergences, namely $\alpha = 2$ and $\beta = 1$ in the spectral domain, and $\alpha = 0$ and $\beta = 1$ in the feature domain.

VI. Conclusion

We proposed a framework to improve the accuracy of uncertainty estimates in the context of uncertainty decoding by fusion of multiple uncertainty estimators or by nonparametric mapping both in the spectral and in the feature domain. The fusion weights and the nonparametric mappings are learned from development data, which makes it possible to address some of the shortcomings of state-of-the-art uncertainty estimators based on fixed mathematical approximations. Experiments on the 2nd CHiME Challenge data showed that nonparametric uncertainty estimation and propagation results in a significant reduction of keyword error rate of 29% relative compared to conventional decoding (without uncertainty) with a GMM-HMM baseline.

Future work will consider discriminative criteria for learning the nonparametric mappings. This work is also expected to have some impact on DNN acoustic models, for which uncertainty decoding has recently started being investigated [16].

Acknowledgment

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

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*. Springer, 2011, pp. 67–99.

[3] M. Gales, "Model based techniques for noise robust speech regcognition," Ph.D. dissertation, 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] C. Kim and R. Stern, "Power-normalized cepstral coefficients (PNCC) for robust speech recognition," in *Proc. Interspeech*, 2009, pp. 1231–1234.

[6] 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.

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

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

[9] 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.

[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] 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*, vol. 2010, 2010, article ID 651420.

[12] R. Astudillo, "Integration of short-time fourier domain speech enhancement and observation uncertainty techniques for robust automatic speech recognition," Ph.D. dissertation, TU Berlin, 2010.

[13] 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.

[14] 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.

[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] B. Li and K. C. Sim, "An ideal hidden-activation mask for deep neural networks based noise-robust speech recognition," in *Proc. ICASSP*, 2014, pp. 200 – 204.

[17] A. Krueger and R. Haeb-Umbach, "Model based feature enhancement for automatic speech recognition in reverberant environments," in *Proc. ICASSP*, 2013, pp. 126–130.

[18] M. Delcroix, S. Watanabe, T. Nakatani, and A. Nakamura, "Cluster-based dynamic variance adaptation for interconnecting speech enhancement pre-processor and speech recognizer," *Computer Speech and Language*, vol. 27, no. 1, pp. 350–368, 2013.

[19] H. Liao, "Uncertainty decoding for noise robust speech recognition," Ph.D. dissertation, Cambridge University, 2007.

[20] R. F. Astudillo and R. Orglmeister, "Computing MMSE estimates and residual uncertainty directly in the feature domain of ASR using STFT domain speech distortion models," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 5, pp. 1023–1034, May 2013.

[21] 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.

[22] S. Srinivasan and D. Wang, "Transforming binary uncertainties for robust speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 2130–2140, Sep 2007.

[23] 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.

[24] 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.

[25] K. Adiloğlu and E. Vincent, "A general variational Bayesian framework for robust feature extraction in multisource recordings," in *Proc. ICASSP*, 2012, pp. 273 – 276.

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

[27] H. Kallasjoki, J. F. Gemmeke, and K. J. Palomki, "Estimating uncertainty to improve exemplar-based feature enhancement for noise robust speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 2, pp. 368 – 380, Feb 2014.

[28] D. T. Tran, E. Vincent, and D. Jouvet, "Extension of uncertainty propagation to dynamic MFCCs for noise-robust ASR," in *Proc. ICASSP*, 2014, pp. 5507–5511.

[29] ——, "Fusion of multiple uncertainty estimators and propagators for noise-robust ASR," in *Proc. ICASSP*, 2014, pp. 5512–5516.

[30] 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.

[31] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series and Products*. Academic Press, 1995.

[32] S. J. Young, G. Evermann, M. J. F. Gales, T. Hain, D. Kershaw *et al.*, *The HTK Book, version 3.4*, University of Cambridge, 2006.

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

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