DNN Uncertainty Propagation using GMM-Derived Uncertainty Features for Noise Robust ASR

Karan Nathwani, Emmanuel Vincent, Irina Illina

DNN Uncertainty Propagation using GMM-Derived Uncertainty Features for Noise Robust ASR

Karan Nathwani, Emmanuel Vincent and Irina Illina

***Abstract*—The uncertainty decoding framework is known to improve deep neural network (DNN) based automatic speech recognition (ASR) performance in noisy environments. It operates by estimating the statistical uncertainty about the input features and propagating it to the output senone posteriors by sampling. Unfortunately, this approximate propagation scheme limits the performance improvement. In this work, we exploit the fact that uncertainty propagation can be achieved in closed form for Gaussian mixture acoustic models (GMMs). We introduce new GMM-derived (GMMD) uncertainty features for robust DNN-based acoustic model training and decoding. The GMMD features are computed as the difference between the GMM log-likelihoods obtained with vs. without uncertainty. They are concatenated with conventional acoustic features and used as inputs to the DNN. We evaluate the resulting ASR performance on the CHiME-2 and CHiME-3 datasets. The proposed features are shown to improve performance on both datasets, both for conventional decoding and for uncertainty decoding with different uncertainty estimation/propagation techniques.**

***Keywords*—Robust ASR, DNN acoustic model, GMM-derived uncertainty features, uncertainty decoding**

I. Introduction

Robust automatic speech recognition (ASR) in noisy environments is still a challenging goal. Front-end speech enhancement approaches aim to estimate clean features which are then fed to the back-end [1]. These approaches alone yield limited performance improvement due to the fact that noise is not perfectly compensated and the enhanced features remain distorted. Hence, they are generally used in combination with back-end approaches that retrain or adapt the acoustic model using enhanced data. These back-end approaches compensate the distortion on average over the whole training set. Yet, the ASR performance on a given utterance still depends on the distortion in that specific utterance.

In the recent past, uncertainty decoding has emerged as a promising framework for robust speech recognition. It utilizes the knowledge of the frame-level uncertainty (or variance) of speech distortion during recognition [2]–[4]. The uncertainty can be computed directly in the ASR feature domain [1], [5]–[10] or propagated from the spectral domain to the feature domain [11]–[18]. Uncertainty decoding has been well studied for traditional statistical models such as Gaussian mixture models (GMMs) [2]–[4]. For such models, the expectation of the senone likelihoods with respect to the feature uncertainty distribution can be computed in closed form.

Unlike GMMs, deep neural network (DNN) based acoustic models do not yield such straightforward solutions owing to the presence of nonlinear activations. In order to incorporate uncertainty in DNNs, propagation techniques based on piecewise exponential approximation, Monte Carlo (MC) sampling, or the unscented transform (UT) have been proposed to approximate the expectation of the acoustic scores [19]–[24]. This is followed by incorporating the modified acoustic scores in the decoding algorithm. In [10], we conducted an extensive experimental investigation of DNN uncertainty estimation and propagation techniques on real and simulated datasets in different noise conditions. We showed that these techniques perform well on logmel features, but performance is reduced with more advanced features such as feature-domain maximum likelihood linear regression (fMLLR) [25].

The approximations underlying the above propagation techniques arguably limit the performance of uncertainty decoding for DNN-based acoustic models. In order to address this issue, in this article, we introduce GMM-derived (GMMD) features which account for the impact of frame-level uncertainty on the senone likelihoods. These features correspond to the difference between the GMM log-likelihoods computed by uncertainty decoding and by conventional decoding. They are concatenated with conventional acoustic features and used for DNN training and decoding. The benefit of GMM-derived features has recently been shown in [26]–[28] in the context of speaker adaptation of DNN-based acoustic models. The authors in [29] also used GMM log-likelihoods as input features (without conventional acoustic features) for adaptation to stationary noise. However, the specific GMMD features proposed in this article and their use for uncertainty decoding are new. The proposed method differs from previous DNN uncertainty decoding methods which are all based on numerical approximation of the expectation. We perform extensive experiments to assess the ASR performance achieved with the proposed GMMD features alone or in combination with uncertainty decoding in several nonstationary noise conditions compared to an fMLLR-domain baseline. As a side result, we propose an fMLLR-domain deep neural network based uncertainty (DNNU) estimator and compare it with Delcroix's estimator [7].

The remainder of this paper is as follows. Section II provides some background on uncertainty decoding and propagation. Section III introduces the proposed GMMD uncertainty features. The experimental setup and the results are presented in Sections IV and V. We conclude in Section VI.

K. Nathwani is with IIT Jammu, Jammu 181 121, India. E. Vincent is with Inria, Villers-lès-Nancy, F-54600, France. I. Illina is with Université de Lorraine, LORIA, UMR 7503, Vandœuvre-lès-Nancy, F-54506, France. This work was performed while K. Nathwani was with Inria. We acknowledge the support of Bpifrance (FUI voiceHome). Experiments presented in this paper were carried out using the Grid'5000 testbed, supported by a scientific interest group hosted by Inria and including CNRS, RENATER and several Universities as well as other organizations (see `https://www.grid5000.fr`).

II. DNN UNCERTAINTY DECODING

A. Conventional vs. Uncertainty Decoding

In clean conditions, given a time sequence of clean feature vectors $\mathbf{y}_t$, a DNN-based acoustic model outputs the posterior

$$p_{\text{DNN}}(s_i|\mathbf{y}_t) \tag{1}$$

for all hidden Markov model (HMM) states (senones) s_i indexed by i at all times t. Decoding is achieved by applying dynamic programming to the pseudo log-likelihoods $\log p_{\text{DNN}}(s_i|\mathbf{y}_t) - \log p(s_i)$.

In noisy conditions, noisy features $\mathbf{z}_t$ are observed instead and enhanced features $\widehat{\mathbf{y}}_t$ can be obtained by means of any speech enhancement technique. Conventional decoding treats the enhanced features as clean and simply replaces $\mathbf{y}_t$ by $\widehat{\mathbf{y}}_t$ in (1). Alternatively, the uncertainty decoding framework assumes that the (unknown) clean features $\mathbf{y}_t$ are Gaussian distributed with mean $\widehat{\mathbf{y}}_t$ and diagonal covariance $\widehat{\boldsymbol{\sigma}}^2_{\mathbf{y}_t}$ [2]–[4] (see Section IV-C for estimation details):

$$p_{\text{unc}}(\mathbf{y}_t|\widehat{\mathbf{y}}_t, \widehat{\boldsymbol{\sigma}}^2_{\mathbf{y}_t}) = \mathcal{N}(\mathbf{y}_t; \widehat{\mathbf{y}}_t, \widehat{\boldsymbol{\sigma}}^2_{\mathbf{y}_t}). \tag{2}$$

$\widehat{\boldsymbol{\sigma}}^2_{\mathbf{y}_t}$ represents the variance of the residual speech distortion after enhancement. To account for this uncertainty, the clean posteriors (1) are replaced by their expectation [21]:

$$p_{\text{DNN}}(s_i|\widehat{\mathbf{y}}_t, \widehat{\boldsymbol{\sigma}}^2_{\mathbf{y}_t}) = \mathbb{E}_{p_{\text{unc}}}[p_{\text{DNN}}(s_i|\mathbf{y}_t)]. \tag{3}$$

Decoding is then achieved as above.

B. Uncertainty Propagation

The expectation can be computed as follows [21]. MC consists of drawing random samples from the feature uncertainty distribution (2), which are input to the DNN. The average of the outputs approximates the posterior expectation. Alternatively, UT consists of drawing samples from this distribution in a deterministic fashion and associating them with weights. The samples are passed through the DNN and the weighted average of the outputs approximates the posterior expectation.

III. PROPOSED GMM-DERIVED UNCERTAINTY FEATURES

A. GMMD Feature Computation

In order to overcome the approximate nature of MC and UT, we propose to use complementary GMMD uncertainty features during training and decoding. The procedure for computing these features is shown in Fig. 1.

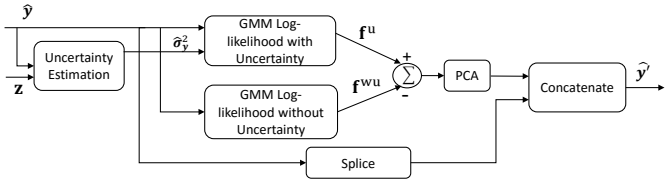

Fig. 1. Procedure for computing GMMD features and concatenating them with spliced acoustic features. $\mathbf{z}$, $\widehat{\mathbf{y}}$, and $\widehat{\boldsymbol{\sigma}}^2_{\mathbf{y}}$ denote the acoustic features of noisy speech, the acoustic features of enhanced speech, and the estimated uncertainty, respectively.

Considering a GMM-based acoustic model with the same set of states as the DNN-based acoustic model, the log-likelihood of state s_i can be computed from the enhanced features as

$$f^{\text{wu}}_{it} = \log p_{\text{GMM}}(\widehat{\mathbf{y}}_t|s_i) = \log\left(\sum\nolimits_j \pi_{ij}\mathcal{N}(\widehat{\mathbf{y}}_t; \boldsymbol{\mu}_{ij}, \boldsymbol{\sigma}^2_{ij})\right) \tag{4}$$

where $\boldsymbol{\mu}_{ij}$, $\boldsymbol{\sigma}^2_{ij}$, and π_{ij} denote the mean, the variance, and the weight of the j-th component of state s_i. Taking the uncertainty into account, the logarithm of the expectation of the likelihood of the clean features is given by [2]–[4]

$$f^{\text{u}}_{it} = \log(\mathbb{E}_{p_{\text{unc}}}[p_{\text{GMM}}(\mathbf{y}_t|s_i)]) \tag{5}$$

$$= \log\left(\sum\nolimits_j \pi_{ij}\mathcal{N}(\widehat{\mathbf{y}}_t; \boldsymbol{\mu}_{ij}, \boldsymbol{\sigma}^2_{ij} + \widehat{\boldsymbol{\sigma}}^2_{\mathbf{y}_t})\right). \tag{6}$$

This equation is classically used for GMM uncertainty decoding, e.g., [8], [11], [17]. In order to leverage it for DNN uncertainty decoding, we stack the log-likelihoods without and with uncertainty into I-dimensional vectors $\mathbf{f}^{\text{wu}}_t = [f^{\text{wu}}_{1t}, \ldots, f^{\text{wu}}_{It}]$ and $\mathbf{f}^{\text{u}}_t = [f^{\text{u}}_{1t}, \ldots, f^{\text{u}}_{It}]$, with I the number of states. We define the GMMD uncertainty features as the difference

$$\text{GMMD}_t = \mathbf{f}^{\text{u}}_t - \mathbf{f}^{\text{wu}}_t. \tag{7}$$

These features are computed in closed form and they represent the change in the GMM acoustic scores due to the uncertainty. The difference $\mathbf{f}^{\text{u}}_t - \mathbf{f}^{\text{wu}}_t$ is more invariant to other variabilities than $\mathbf{f}^{\text{wu}}_t$ itself, which is expected to facilitate learning.

B. Use in Conventional or Uncertainty Decoding

GMMD features are high-dimensional. In practice, we reduce their dimension by principal component analysis (PCA) [30]. The resulting features are concatenated with the spliced enhanced features $\widehat{\mathbf{y}}_t$ and used during training and decoding. Decoding can be achieved either in a conventional fashion or by uncertainty decoding. In the latter case, the uncertainty over the concatenated features $[\widehat{\mathbf{y}}_{t-\tau}, \ldots, \widehat{\mathbf{y}}_t, \ldots, \widehat{\mathbf{y}}_{t+\tau}, \text{GMMD}_t]$ is equal to $[\widehat{\boldsymbol{\sigma}}^2_{\mathbf{y}_{t-\tau}}, \ldots, \widehat{\boldsymbol{\sigma}}^2_{\mathbf{y}_t}, \ldots, \widehat{\boldsymbol{\sigma}}^2_{\mathbf{y}_{t+\tau}}, \mathbf{0}]$.

Figure 2 shows GMM and DNN posteriorgrams for one real CHiME-3 utterance. The DNN posteriorgrams are obtained by conventional decoding. The GMM posteriorgram without uncertainty and the DNN posteriorgram without GMMD features both differ significantly from the ground truth. The GMM posteriorgram with uncertainty is closer to the ground truth, but has nonzero probability for several other states. The DNN posteriorgram with GMMD features matches best with the ground truth compared to any other posteriorgram, and it finds the transition from `SIL` to `AA1l`.

IV. EXPERIMENTAL SETUP

A. Datasets

We evaluated the proposed features on the CHiME-2 [31] and CHiME-3 [32] datasets. The CHiME-2 dataset was created by convolving clean Wall Street Journal (WSJ0) utterances with binaural room impulse responses and adding real domestic background noise at six signal-to-noise ratios (SNRs) from -6 to +9 dB. The training set contains 7138 noisy utterances

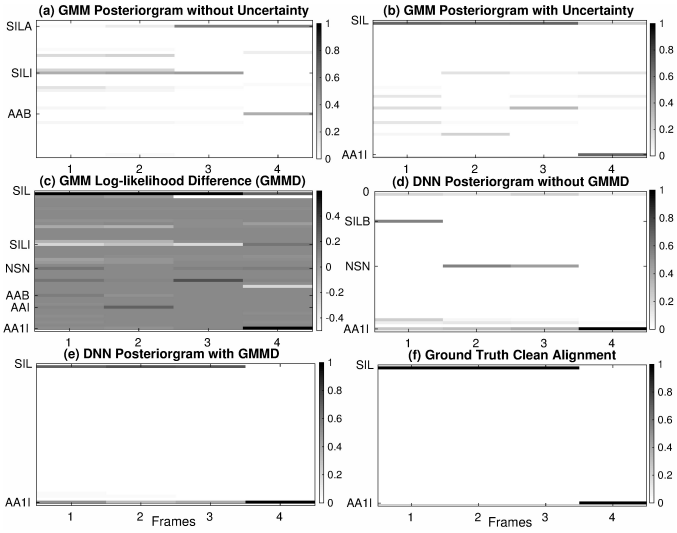

Fig. 2. Posteriorgrams on 4 successive time frames of a CHiME-3 utterance: (a) GMM conventional decoding of fMLLR features, (b) GMM uncertainty decoding of fMLLR features, (d) DNN conventional decoding of fMLLR features, (e) DNN conventional decoding of fMLLR+GMMD features. The GMMD features and the ground truth computed from clean features are shown in (c) and (f). Low probability states are hidden. See Section IV for details.

from 83 speakers. The development and test sets contain 2460 and 1980 noisy utterances from 10 and 8 speakers, respectively.

The CHiME-3 dataset provides real and simulated noisy WSJ0 utterances acquired by a tablet fitted with 6 microphones in four nonstationary noise environments: bus (BUS), café (CAF), pedestrian area (PED), and street (STR). For training, 1600 real and 7138 simulated utterances from 87 speakers were used. The development set contains 1640 real and 1640 simulated utterances from 4 speakers, and the test set 1320 real and 1320 simulated utterances from 4 speakers. The results are reported on the real development and test sets only hereafter.

B. Speech Enhancement and ASR Baseline

The uncertainty estimators considered in the following are applicable to any speech enhancement technique. As an example, we enhanced all training, development, and test data via multichannel nonnegative matrix factorization [33] using the FASST toolbox [34] with identical settings to [10], [17], [21], [22]. This allows comparison with these earlier studies.

For each dataset, we trained a GMM acoustic model with 40 fMLLR features and a DNN model with 40 fMLLR features with $\tau = 5$ left and right context frames via Kaldi [35]. Both acoustic models were trained on the enhanced training set. The targets were obtained by forced alignment using a GMM acoustic model trained and applied on clean data (for simulated training data) or on enhanced data (for real training data). The GMM-based model for CHiME-3 has 15025 Gaussians.

The DNNs follow the architecture in [10], [21], [22]: a 440-dimensional input layer and seven 2048-dimensional hidden layers. The output layer consists of 2000 states for CHiME-2 and 1978 for CHiME-3. For decoding, we used the challenges' baseline trigram language model and 5k vocabulary.

TABLE I. AVERAGE WER (%) ON THE CHIME-2 AND CHIME-3 TEST SETS. MC/UT PERTAIN TO DNN ACOUSTIC MODELS ONLY.

Uncertainty		**CHiME-2 test set**			**CHiME-3 real test set**		
		GMM	**DNN**		**GMM**	**DNN**	
Est.	**Prop.**	fMLLR	fMLLR	fMLLR +GMMD	fMLLR	fMLLR	fMLLR +GMMD
None		44.22	17.62	14.76	23.98	19.72	18.05
DU	MC	41.14	17.52	14.78	25.03	19.04	17.98
	UT		17.72	14.91		21.97	19.81
DNNU	MC	41.87	17.41	**13.89**	23.71	18.27	**17.22**
	UT		17.46	14.28		19.29	17.92

No rescoring (using, e.g., neural network language models or sequence-level minimum Bayes risk) was performed.

We assess the ASR performance in terms of the word error rate (WER). For the CHiME-2 development and test sets, the 95% confidence interval is about ± 0.4 %. For CHiME-3, the confidence interval is about $\pm 0.3\%$ for the development set and $\pm 0.5\%$ for the test set. In the following tables, for each test condition, the best choice of features, acoustic model, and uncertainty estimation/propagation technique is shown in bold.

C. Uncertainty Estimation and Representation

We considered two uncertainty estimators. First is Delcroix's uncertainty (DU) estimator [7] $\widehat{\boldsymbol{\sigma}}^2_{\mathbf{y}_t} = \alpha(\widehat{\mathbf{y}}_t - \mathbf{z}_t)^2$ which is proportional to the elementwise squared difference between the enhanced and the noisy features. Following [21], we set $\alpha = 0.4$. As second estimator, we propose a DNNU estimator that takes a 80-dimensional input vector $[\mathbf{z}_t, \widehat{\mathbf{y}}_t - \mathbf{z}_t]$ consisting of the noisy features and their difference with the enhanced features and outputs a 40-dimensional uncertainty vector $\widehat{\boldsymbol{\sigma}}^2_{\mathbf{y}_t}$. This estimator is trained on enhanced simulated training data, using the elementwise squared difference $(\widehat{\mathbf{y}} - \mathbf{y})^2$ between the enhanced and the clean features as a target. It differs from the neural network-based uncertainty estimator in [10] in two aspects. First, we use 3 hidden layers instead of 2. Second, we use fMLLR domain inputs and outputs instead of logmel.

For uncertainty propagation, we drew 3 samples from the feature uncertainty distribution for both MC and UT.

We computed the GMMD features as explained in Section III using DNNU in (6), irrespective of whether DU or DNNU were used for possibly subsequent DNN uncertainty decoding. We reduced their dimension to 100 via PCA before concatenation with the 440-dimensional spliced fMLLR features. We denote the concatenated features as fMLLR+GMMD. For comparison, we also considered the 40-dimensional uncertainty vectors estimated by DU and DNNU themselves as features instead of GMMD. We denote the resulting concatenated features as fMLLR+DU and fMLLR+DNNU.

V. RESULTS AND DISCUSSION

A. Overall Results

Table I presents the average results on the CHiME-2 and CHiME-3 test sets for both GMM and DNN acoustic models without uncertainty (called "None") and with various uncertainty estimation and propagation techniques. We observed a similar behavior on the CHiME-2 and CHiME-3 development sets (not shown here). We can make the following observations.

TABLE II. WER (%) PER SNR CONDITION ON THE CHIME-2 TEST & DEVELOPMENT SETS USING DNN ACOUSTIC MODELS.

Features	Uncertainty		Test Set						Development Set					
	Est.	Prop.	-6 dB	-3 dB	0 dB	3 dB	6 dB	9 dB	-6 dB	-3 dB	0 dB	3 dB	6 dB	9 dB
fMLLR	None		29.28	21.69	17.78	14.06	12.20	10.73	36.43	29.57	25.03	21.10	17.45	15.58
	DU	MC	29.18	21.43	17.78	14.30	11.75	10.69	35.85	28.60	24.28	20.30	16.86	14.95
		UT	29.12	21.86	17.88	14.63	11.79	11.05	37.12	29.81	24.94	21.14	18.00	15.19
	DNNU	MC	28.93	21.76	17.47	13.83	11.81	10.68	36.07	28.45	24.26	19.47	15.44	14.70
		UT	29.03	21.74	17.53	13.93	11.79	10.76	36.22	28.37	24.79	19.45	15.49	14.60
fMLLR+DU	None		28.27	20.86	17.59	13.27	11.97	11.26	34.96	28.39	24.22	19.74	16.98	14.54
fMLLR+DNNU	None		28.66	20.83	17.58	13.24	12.57	11.60	35.01	29.23	23.86	19.96	16.35	15.07
fMLLR+GMMD	None		25.85	18.59	14.74	11.43	9.29	8.70	32.45	24.69	21.34	17.11	14.24	12.33
	DU	MC	26.45	18.95	14.50	11.24	8.97	8.61	32.25	24.43	21.58	17.24	14.23	12.49
		UT	26.38	19.10	14.61	11.62	9.29	8.50	32.69	25.69	21.68	17.19	14.29	13.03
	DNNU	MC	**25.08**	**17.99**	**13.21**	**10.79**	**8.58**	**7.72**	**31.78**	**23.71**	**20.09**	**16.40**	**13.39**	**11.48**
		UT	25.41	18.31	13.86	11.03	9.07	8.03	32.14	24.19	20.52	16.71	13.61	11.80

- In all configurations, DNN acoustic models greatly improve the ASR performance compared to GMM. In the following, we consider DNN acoustic models applied on fMLLR features without uncertainty as our baseline.
- Compared to the baseline, uncertainty decoding based on DNNU and MC improves the WER by 5% and 1% relative for CHiME-2 and CHiME-3 respectively.
- Alternatively, using fMLLR+GMMD features provides an improvement of 16% and 8% relative. This means GMMD features better handle uncertainty than classical DNN uncertainty propagation techniques.
- Combining fMLLR+GMMD, DNNU and MC improves the WER by 21% and 13% relative compared to the baseline. This means GMMD features and DNN uncertainty decoding are complementary. To our knowledge, this is the largest improvement obtained for DNN uncertainty decoding on top of an fMLLR-domain baseline.
- The proposed DNNU estimator outperforms the DU estimator in all but one (GMM acoustic model on CHiME-2) situations. Also, MC-based uncertainty propagation systematically outperforms UT-based propagation for all features, acoustic models, and uncertainty estimators.

B. Impact of SNRs

Table II presents the WER on the CHiME-2 development and test sets as a function of the SNR. In addition to the systems in Table I, we also show the results obtained using fMLLR+DU or fMLLR+DNNU features. Our main findings still hold: for all SNRs, DNN uncertainty decoding based on fMLLR+GMMD, DNNU and MC outperforms other configurations. The improvement relative to the baseline increases with the SNR, from 14% at -6 dB to 28% at 9 dB on the test set. Also, with fMLLR+GMMD features, DNNU outperforms DU in all cases and MC often performs slightly better than UT. Without uncertainty decoding, fMLLR+GMMD features also systematically improve performance compared to fMLLR features only, while fMLLR+DU and fMLLR+DNNU provide a smaller, less consistent improvement.

C. Impact of Noise Environments

The impact of the noise environments on the WER in the real CHiME-3 development and test sets is shown in Table III.

TABLE III. WER (%) PER NOISE ENVIRONMENT ON THE REAL CHIME-3 TEST & DEVELOPMENT SETS USING DNN ACOUSTIC MODELS.

Features	Uncert.		Test Set				Development Set			
	Est.	Prop.	BUS	CAF	PED	STR	BUS	CAF	PED	STR
fMLLR	None		24.38	17.23	26.93	10.34	11.02	9.44	7.17	8.50
	DU	MC	22.15	16.73	27.24	10.07	11.19	9.61	7.63	8.13
		UT	27.71	20.00	29.12	11.07	11.55	10.68	8.05	9.50
	DNNU	MC	20.35	16.48	26.36	**9.89**	10.34	**9.18**	6.83	8.08
		UT	22.60	17.22	27.04	10.30	10.38	9.38	7.03	**8.07**
fMLLR +DU	None		22.82	18.01	27.26	10.98	10.98	9.41	7.70	8.41
fMLLR +DNNU	None		22.99	18.36	26.67	10.91	11.04	9.50	7.67	8.39
fMLLR +GMMD	None		20.32	15.94	25.67	10.30	10.67	9.61	6.93	8.18
	DU	MC	20.25	15.85	25.45	10.39	10.90	9.70	7.33	8.51
		UT	23.79	17.10	27.19	11.18	11.27	10.02	7.69	8.64
	DNNU	MC	**19.55**	**15.12**	**24.33**	9.91	**10.22**	9.39	**6.75**	8.16
		UT	20.09	15.81	25.29	10.49	**10.22**	9.62	6.88	8.17

DNN uncertainty decoding based on fMLLR+GMMD, DNNU and MC always improves the WER compared to the baseline. In 5 out of 8 test conditions, it is the best configuration. In the 3 remaining conditions, the difference with the best configuration is not statistically significant. For all features, DNNU systematically outperforms DU and MC systematically outperforms UT. Finally, without uncertainty decoding, fMLLR+GMMD features improve the WER compared to the baseline in 7 out of 8 test conditions, while fMLLR+DU and fMLLR+DNNU improve it in 4 and 3 conditions only, respectively.

VI. CONCLUSION

We proposed GMMD features as additional inputs to DNN acoustic models for noise-robust ASR. These features encode the impact of the uncertainty, that is the variance of the residual speech distortion, on the acoustic scores. Experiments on both simulated and real data showed that they systematically and significantly improve performance compared to fMLLR features alone, or to using the estimated uncertainty itself as a feature. We also proposed a DNNU estimator that successfully operates in the fMLLR domain and showed that applying uncertainty propagation to fMLLR+GMMD features further improves the WER, up to 21% and 13% relative compared to conventional decoding on fMLLR features without uncertainty. In the future, we will assess our approach with other enhancement techniques and explore joint optimization of the feature dimension reduction matrix and the DNN acoustic model.

REFERENCES

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

[2] J. A. Arrowood and M. A. Clements, "Using observation uncertainty in HMM decoding," in *Proc. Interspeech*, 2002, pp. 1561–1564.

[3] N. Becerra Yoma and M. Villar, "Speaker verification in noise using a stochastic version of the weighted Viterbi algorithm," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 3, pp. 158–166, 2002.

[4] L. Deng, 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 Speech and Audio Processing*, vol. 13, no. 3, pp. 412–421, 2005.

[5] H. Liao and M. Gales, "Joint uncertainty decoding for noise robust speech recognition," in *Proc. Interspeech*, 2005, pp. 3129–3132.

[6] V. Stouten and P. Wambacq, "Model-based feature enhancement with uncertainty decoding for noise robust ASR," *Speech Communication*, vol. 48, no. 11, pp. 1502–1514, 2006.

[7] 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, 2009.

[8] 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 & Language*, vol. 27, no. 1, pp. 350–368, 2013.

[9] L. Lu, K. Chin, A. Ghoshal, and S. Renals, "Joint uncertainty decoding for noise robust subspace Gaussian mixture models," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 9, pp. 1791–1804, 2013.

[10] K. Nathwani, J. Morales-Cordovilla, S. Sivasankaran, I. Illina, and E. Vincent, "An extended experimental investigation of DNN uncertainty propagation for noise robust ASR," in *Proc. HSCMA*, 2017, pp. 26–30.

[11] D. Kolossa, R. F. Astudillo, E. Hoffmann, and R. Orglmeister, "Independent component analysis and time-frequency masking for speech recognition in multitalker conditions," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2010, no. 1, pp. 1–13, 2010.

[12] R. F. 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] 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, 2013.

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

[15] A. H. Abdelaziz, S. Zeiler, D. Kolossa, V. Leutnant, and R. Haeb-Umbach, "GMM-based significance decoding," in *Proc. ICASSP*, 2013, pp. 6827–6831.

[16] D. T. Tran, E. Vincent, and D. Jouvet, "Fusion of multiple uncertainty estimators and propagators for noise robust ASR," in *Proc. ICASSP*, 2014, pp. 5512–5516.

[17] ——, "Nonparametric uncertainty estimation and propagation for noise robust ASR," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 11, pp. 1835–1846, 2015.

[18] K. Adiloğlu and E. Vincent, "Variational Bayesian inference for source separation and robust feature extraction," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 10, pp. 1746–1758, 2016.

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

[20] R. F. Astudillo, A. Abad, and I. Trancoso, "Accounting for the residual uncertainty of multi-layer perceptron based features," in *Proc. ICASSP*, 2014, pp. 6859–6863.

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

[22] Y. Tachioka and S. Watanabe, "Uncertainty training and decoding methods of deep neural networks based on stochastic representation of enhanced features," in *Proc. Interspeech*, 2015, pp. 3541–3545.

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

[24] C. Huemmer, A. Schwarz, R. Maas, H. Barfuss, R. F. Astudillo, and W. Kellermann, "A new uncertainty decoding scheme for DNN-HMM hybrid systems with multichannel speech enhancement," in *Proc. ICASSP*, 2016, pp. 5760–5764.

[25] K. Nathwani, E. Vincent, and I. Illina, "Consistent DNN uncertainty training and decoding for robust ASR," in *Proc. ASRU*, 2017.

[26] N. A. Tomashenko and Y. Y. Khokhlov, "Speaker adaptation of context dependent deep neural networks based on MAP-adaptation and GMM-derived feature processing." in *Proc. Interspeech*, 2014, pp. 2997–3001.

[27] ——, "GMM-derived features for effective unsupervised adaptation of deep neural network acoustic models." in *Proc. Interspeech*, 2015, pp. 2882–2886.

[28] N. Tomashenko, Y. Khokhlov, and Y. Estève, "On the use of Gaussian mixture model framework to improve speaker adaptation of deep neural network acoustic models," in *Proc. Interspeech*, 2016, pp. 3788–3792.

[29] S. Kundu, K. C. Sim, and M. Gales, "Incorporating a generative front-end layer to deep neural network for noise robust automatic speech recognition," in *Proc. Interspeech*, 2016, pp. 2359–2363.

[30] X. Jiang, "Linear subspace learning-based dimensionality reduction," *IEEE Signal Processing Magazine*, vol. 28, no. 2, pp. 16–26, 2011.

[31] 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, pp. 162–167.

[32] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'CHiME' speech separation and recognition challenge: Analysis and outcomes," *Computer Speech & Language*, vol. 46, pp. 605–626, 2017.

[33] 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, 2012.

[34] Y. Salaün, E. Vincent, N. Bertin, N. Souviraà-Labastie, X. Jaureguiberry, D. T. Tran, and F. Bimbot, "The flexible audio source separation toolbox version 2.0," in *ICASSP Show & Tell*, 2014.

[35] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlíček, Y. Qian, P. Schwarz, J. Silovský, G. Stemmer, and K. Veselý, "The Kaldi speech recognition toolkit," in *Proc. ASRU*, 2011, pp. 1–4.