GMM-BASED SIGNIFICANCE DECODING

Ahmed Hussen Abdelaziz, Steffen Zeiler*, Dorothea Kolossa*, Volker Leutnant†, Reinhold Haeb-Umbach†*

*Insititute of Communication Acoustics, Ruhr-Universität Bochum,
Digital Signal Processing Group, 44801 Bochum, Germany
{Ahmed.HussenAbdelAziz, Steffen.Zeiler, Dorothea.Kolossa}@rub.de
†Department of Communications Engineering, University of Paderborn, Germany
{leutnant,haeb}@nt.uni-paderborn.de

ABSTRACT

The accuracy of automatic speech recognition systems in noisy and reverberant environments can be improved notably by exploiting the uncertainty of the estimated speech features using so-called *uncertainty-of-observation* techniques. In this paper, we introduce a new Bayesian decision rule that can serve as a mathematical framework from which both known and new uncertainty-of-observation techniques can be either derived or approximated. The new decision rule in its direct form leads to the new significance decoding approach for Gaussian mixture models, which results in better performance compared to standard uncertainty-of-observation techniques in different additive and convolutive noise scenarios.

Index Terms— Uncertainty-of-observation, noise robust speech recognition, uncertainty decoding, modified imputation, significance decoding

1. INTRODUCTION

Although statistical models, like the hidden Markov models (HMMs), have shown great success in modeling and recognizing the temporal evolution of the spectral characteristics of speech, noise and reverberation are still great obstacles precluding wider use of automatic speech recognition (ASR) systems. The main reason for dropping performance in real operating conditions is that the distortion introduced by additive and convolutive noise leads to a mismatch between the extracted features and the statistical model. With the intermediary goal of achieving the greatest match between training and test conditions, many front-end and back-end techniques have been developed to achieve an acceptable recognition robustness against noise and reverberation.

Front-end techniques like Wiener filtering [1], used e.g. in the ETSI advanced front end (AFE) [2], stereo piecewise linear compensation for environment (SPLICE) [3], or log-spectral amplitude minimum mean square error (MMSE) estimation [4] can be used to enhance the distorted feature vector $\mathbf{o}_t$, where t denotes the frame index, and compute an estimate $\hat{\mathbf{x}}_t$ of the clean feature vector $\mathbf{x}_t$. These estimates can then be fed into the recognition model and treated as if they were the true clean features. However, front-end approaches do not perfectly compensate the distortion of the features. Therefore, the output features generated by the front-end are neither completely clean nor as distorted as the recorded noisy observations.

The emerging field of probabilistic uncertainty-of-observation techniques takes this uncertainty of estimation into account and considers the output of the front-end not as deterministic but rather as a random variable. This means that the front-end is now required to estimate a conditional probability density function (PDF) $\mathrm{p}\left(\hat{\mathbf{x}}_t|\mathbf{x}_t\right)$ rather than just an enhanced feature vector $\hat{\mathbf{x}}_t$. Fortunately, Bayesian feature enhancement delivers an unbiased estimate $\hat{\mathbf{x}}_t$ for $\mathbf{x}_t$, whose estimation error variance $\mathbf{\Sigma}_{\mathbf{x}_t|\mathbf{o}_{1:t}}$ (the *observation uncertainty*) is equal to the variance of the posterior density $\mathrm{p}\left(\mathbf{x}_t|\mathbf{o}_{1:t}\right)$, where $\mathbf{o}_{1:t} = \mathbf{o}_1, \ldots, \mathbf{o}_t$.

The PDF $\mathrm{p}(\hat{\mathbf{x}}_t|\mathbf{x}_t)$ describes the generation of the enhanced feature vector $\hat{\mathbf{x}}_t$ through distorting the underlying hidden clean feature vector $\mathbf{x}_t$ with the estimation error $\mathbf{e}_t$, which is often assumed to be Gaussian. Thus,

$$\hat{\mathbf{x}}_t = \mathbf{x}_t + \mathbf{e}_t \tag{1}$$

where

$$\mathrm{p}\left(\mathbf{e}_t\right) = \mathcal{N}\left(\mathbf{e}_t; \mathbf{0}, \mathbf{\Sigma}_{\mathbf{x}_t|\mathbf{o}_{1:t}}\right) \tag{2}$$

and therefore

$$\mathrm{p}\left(\hat{\mathbf{x}}_t|\mathbf{x}_t\right) = \mathcal{N}\left(\hat{\mathbf{x}}_t; \mathbf{x}_t, \mathbf{\Sigma}_{\mathbf{x}_t|\mathbf{o}_{1:t}}\right). \tag{3}$$

Of course, any other estimation algorithm or heuristic about $\mathrm{p}(\mathbf{e}_t)$, e.g. [5], can also be applied here.

A range of uncertainty-of-observation approaches [6–13] has been proposed to exploit the uncertainty and its underlying model. In this paper we introduce a consistent mathematical framework for the newly-introduced significance decoding approach [13] allowing us to extend it to Gaussian mixture model output distributions.

The rest of this paper is organized as follows. At first, a new Bayesian decision rule is introduced in Section 2 as an

alternative to the conventional maximum likelihood rule [14]. Then, a rigorous derivation of significance decoding (SD) using this decision rule is given for the case when the output distribution functions of the emitting states q_t of the HMM are described by a Gaussian mixture model (GMM). The new decision rule can be considered as a general mathematical framework of other previously introduced uncertainty-of-observation techniques as will be discussed in Section 3. In Section 4, the performance characteristics of SD are evaluated and compared to the conventional maximum likelihood, the uncertainty decoding [7] and the modified imputation [11] approaches, using the Grid database [15], where uncertainty decoding and modified imputation are two state-of-the-art uncertainty-of-observation techniques. Finally, the relation of the introduced technique to prior work and conclusions are discussed in Section 5 and Section 6, respectively.

2. SIGNIFICANCE DECODING

2.1. Bayesian Decision Rule

In the presence of noise or other distortions, the clean feature vectors $\mathbf{x}_{1:t}$ are not directly observable. Consequently, the likelihood $\mathcal{L} = \mathrm{p}(\mathbf{x}_t|q_t)$ required by the recognizer can not be computed. The best we can do is to compute the expected value of $\mathrm{p}(\mathbf{x}_t|q_t)$, where for its computation all available information about $\mathbf{x}_t$ should be employed. This means that we have to compute the expected value utilizing the posterior density of $\mathbf{x}_t$, given all observable variables that are statistically related to $\mathbf{x}_t$. Thus, the likelihood $\mathcal{L}$ is to be replaced by

$$E[\mathrm{p}(\mathbf{x}_t|q_t)|\mathbf{o}_{1:t}, q_t] = \int \mathrm{p}(\mathbf{x}_t|q_t)\mathrm{p}(\mathbf{x}_t|\mathbf{o}_{1:t}, q_t)d\mathbf{x}_t, \quad (4)$$

where we restricted ourselves to causal processing, i.e., the posterior is only based on current and past observations.

The expectation in Eq. (4) is reminiscent of the approach taken in the Expectation Maximization (EM) algorithm [16], where the complete data log-likelihood is replaced by the expected value of the complete data log-likelihood, given the observed data.

2.2. Significance Decoding for GMM

In order to evaluate the Bayesian decision rule in (4), first we need to compute the posterior distribution $\mathrm{p}(\mathbf{x}_t|\mathbf{o}_{1:t}, q_t)$. Under the assumption that the enhanced feature vector $\hat{\mathbf{x}}_t$ captures the sufficient statistics of $\mathbf{o}_{1:t}$, we can express $\mathrm{p}(\mathbf{x}_t|\mathbf{o}_{1:t}, q_t)$ as follows [13]:

$$\mathrm{p}(\mathbf{x}_t|\mathbf{o}_{1:t}, q_t) = \mathrm{p}(\mathbf{x}_t|\hat{\mathbf{x}}_t, q_t) = \frac{\mathrm{p}\left(\hat{\mathbf{x}}_t, \mathbf{x}_t|q_t\right)}{\int \mathrm{p}\left(\hat{\mathbf{x}}_t, \mathbf{x}_t|q_t\right) d\mathbf{x}_t}. \quad (5)$$

The joint probability $\mathrm{p}\left(\hat{\mathbf{x}}_t, \mathbf{x}_t|q_t\right)$ in (5) factors into:

$$\begin{aligned}\mathrm{p}\left(\hat{\mathbf{x}}_t, \mathbf{x}_t|q_t\right) &= \mathrm{p}\left(\hat{\mathbf{x}}_t|\mathbf{x}_t, q_t\right)\mathrm{p}\left(\mathbf{x}_t|q_t\right) \\ &= \mathrm{p}\left(\hat{\mathbf{x}}_t|\mathbf{x}_t\right)\mathrm{p}\left(\mathbf{x}_t|q_t\right). \end{aligned} \quad (6)$$

The second line in (6) was obtained by assuming that the estimate $\hat{\mathbf{x}}_t$ is independent of the HMM state q_t, as is typically the case in front-end processing methods.

Conventionally, the output distribution of the emitting states q_t is modeled by a GMM of K mixtures via

$$\mathrm{p}\left(\mathbf{x}_t|q_t\right) = \sum_{\kappa=1}^{K} \omega_{q_\kappa} \mathcal{N}\left(\mathbf{x}_t; \boldsymbol{\mu}_{q_\kappa}, \boldsymbol{\Sigma}_{q_\kappa}\right) \quad (7)$$

with $\boldsymbol{\mu}_{q_\kappa}$, $\boldsymbol{\Sigma}_{q_\kappa}$ and ω_{q_κ} as the mean, the covariance matrix and the weight of the κ^{th} mixture of the q_t^{th} state, respectively. By applying (3) and (7) to (6), we obtain the following expression for the joint probability

$$\begin{aligned}&\mathrm{p}\left(\hat{\mathbf{x}}_t, \mathbf{x}_t|q_t\right) = \mathrm{p}\left(\mathbf{x}_t|q_t\right)\mathrm{p}\left(\hat{\mathbf{x}}_t|\mathbf{x}_t\right) = \\ &\sum_{\kappa=1}^{K} \omega_{q_\kappa} \mathcal{N}\left(\mathbf{x}_t; \boldsymbol{\mu}_{q_\kappa}, \boldsymbol{\Sigma}_{q_\kappa}\right) \mathcal{N}\left(\hat{\mathbf{x}}_t; \mathbf{x}_t, \boldsymbol{\Sigma}_{\mathbf{x}_t|\mathbf{o}_{1:t}}\right). \end{aligned} \quad (8)$$

The multiplication of the Gaussian functions in (8) can be reformulated [17] to

$$\begin{aligned}\mathrm{p}\left(\hat{\mathbf{x}}_t, \mathbf{x}_t|q_t\right) = \sum_{\kappa=1}^{K} &\omega_{q_\kappa} \mathcal{N}\left(\mathbf{x}_t; \tilde{\boldsymbol{\mu}}_{q_\kappa}, \tilde{\boldsymbol{\Sigma}}_{q_\kappa}\right) \times \\ &\mathcal{N}\left(\hat{\mathbf{x}}_t; \boldsymbol{\mu}_{q_\kappa}, \boldsymbol{\Sigma}_{q_\kappa} + \boldsymbol{\Sigma}_{\mathbf{x}_t|\mathbf{o}_{1:t}}\right), \end{aligned} \quad (9)$$

with

$$\begin{aligned}\tilde{\boldsymbol{\mu}}_{q_\kappa} = &\boldsymbol{\Sigma}_{q_\kappa}\left(\boldsymbol{\Sigma}_{\mathbf{x}_t|\mathbf{o}_{1:t}} + \boldsymbol{\Sigma}_{q_\kappa}\right)^{-1}\hat{\mathbf{x}}_t + \\ &\boldsymbol{\Sigma}_{\mathbf{x}|\mathbf{o}_{1:t}}\left(\boldsymbol{\Sigma}_{\mathbf{x}_t|\mathbf{o}_{1:t}} + \boldsymbol{\Sigma}_{q_\kappa}\right)^{-1}\boldsymbol{\mu}_{q_\kappa}, \end{aligned} \quad (10)$$

and

$$\tilde{\boldsymbol{\Sigma}}_{q_\kappa} = \boldsymbol{\Sigma}_{\mathbf{x}_t|\mathbf{o}_{1:t}}\left(\boldsymbol{\Sigma}_{\mathbf{x}_t|\mathbf{o}_{1:t}} + \boldsymbol{\Sigma}_{q_\kappa}\right)^{-1}\boldsymbol{\Sigma}_{q_\kappa}. \quad (11)$$

The denominator of (5) is obtained by noting that the integral over the entire range of a PDF evaluates to one:

$$\begin{aligned}&\int \sum_{\kappa=1}^{K} \omega_{q_\kappa} \mathcal{N}\left(\mathbf{x}_t; \tilde{\boldsymbol{\mu}}_{q_\kappa}, \tilde{\boldsymbol{\Sigma}}_{q_\kappa}\right) \times \\ &\qquad \mathcal{N}\left(\hat{\mathbf{x}}_t; \boldsymbol{\mu}_{q_\kappa}, \boldsymbol{\Sigma}_{q_\kappa} + \boldsymbol{\Sigma}_{\mathbf{x}_t|\mathbf{o}_{1:t}}\right) d\mathbf{x}_t \\ &= \sum_{\kappa=1}^{K} \omega_{q_\kappa} \mathcal{N}\left(\hat{\mathbf{x}}_t; \boldsymbol{\mu}_{q_\kappa}, \boldsymbol{\Sigma}_{q_\kappa} + \boldsymbol{\Sigma}_{\mathbf{x}_t|\mathbf{o}_{1:t}}\right). \end{aligned} \quad (12)$$

Now, inserting (9) and (12) in (5), we obtain the following expression for the sought posterior:

$$\mathrm{p}(\mathbf{x}_t|\hat{\mathbf{x}}_t, q_t) = \sum_{\kappa=1}^{K} \tilde{\omega}_{q_\kappa} \mathcal{N}\left(\mathbf{x}_t; \tilde{\boldsymbol{\mu}}_{q_\kappa}, \tilde{\boldsymbol{\Sigma}}_{q_\kappa}\right), \quad (13)$$

with

$$\tilde{\omega}_{q_\kappa} = \frac{\omega_{q_\kappa} \mathcal{N}\left(\hat{\mathbf{x}}_t; \boldsymbol{\mu}_{q_\kappa}, \boldsymbol{\Sigma}_{q_\kappa} + \boldsymbol{\Sigma}_{\mathbf{x}_t|\mathbf{o}_{1:t}}\right)}{\sum_{m=1}^{K} \omega_{q_m} \mathcal{N}\left(\hat{\mathbf{x}}_t; \boldsymbol{\mu}_{q_m}, \boldsymbol{\Sigma}_{q_m} + \boldsymbol{\Sigma}_{\mathbf{x}_t|\mathbf{o}_{1:t}}\right)}. \quad (14)$$

Applying Equations (7) and (13) to (4) and re-arranging the products of Gaussians, again in accordance with [17], yields the general SD likelihood

$$\begin{aligned}\mathcal{L}_g^{\mathrm{SD}} &:= E[\mathrm{p}(\mathbf{x}_t|q_t)|\mathbf{o}_{1:t}, q_t] \\ &= \int \sum_{\kappa=1}^{K}\sum_{\nu=1}^{K} \tilde{\omega}_{q_\kappa}\omega_{q_\nu}\mathcal{N}\left(\mathbf{x}_t; \tilde{\boldsymbol{\mu}}_{q_\kappa}, \tilde{\boldsymbol{\Sigma}}_{q_\kappa}\right) \times \\ &\qquad \mathcal{N}\left(\mathbf{x}_t; \boldsymbol{\mu}_{q_\nu}, \boldsymbol{\Sigma}_{q_\nu}\right) d\mathbf{x}_t \qquad (15) \\ &= \sum_{\kappa=1}^{K}\sum_{\nu=1}^{K} \tilde{\omega}_{q_\kappa}\omega_{q_\nu}\mathcal{N}\left(\tilde{\boldsymbol{\mu}}_{q_\kappa}; \boldsymbol{\mu}_{q_\nu}, \boldsymbol{\Sigma}_{q_\nu} + \tilde{\boldsymbol{\Sigma}}_{q_\kappa}\right). \qquad (16)\end{aligned}$$

2.3. Limiting cases

For sake of simplicity, in the following we assume that the output distribution function of the emitting states is Gaussian, i.e. $\mathrm{p}\left(\mathbf{x}_t|q_t\right) = \mathcal{N}\left(\mathbf{x}_t; \boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q\right)$.

The first limiting case we discuss here is the uncertainty flooding, where the feature stream to be recognized is extremely distorted, so $\boldsymbol{\Sigma}_{\mathbf{x}_t|\mathbf{o}_{1:t}} \to \infty$. In this case, the observation $\mathbf{o}_{1:t}$ and thus the estimate $\hat{\mathbf{x}}_t$ becomes independent of $\mathbf{x}_t$. Then, Eq. (4) simplifies to

$$\begin{aligned}&E\left[\mathrm{p}\left(\mathbf{x}_t|q_t\right)|\mathbf{o}_{1:t}, q_t\right]|_{\boldsymbol{\Sigma}_{\mathbf{x}_t|\mathbf{o}_{1:t}}\to\infty} \\ &\quad= \int \mathrm{p}\left(\mathbf{x}_t|q_t\right)\mathrm{p}\left(\mathbf{x}_t|q_t\right) d\mathbf{x}_t \\ &\quad= \int \mathcal{N}\left(\mathbf{x}_t; \boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q\right)\mathcal{N}\left(\mathbf{x}_t; \boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q\right) d\mathbf{x}_t \\ &\quad= \mathcal{N}\left(\boldsymbol{\mu}_q; \boldsymbol{\mu}_q, 2\boldsymbol{\Sigma}_q\right). \qquad (17)\end{aligned}$$

Thus, even if a frame is completely uninformative, the likelihood is still dependent on the HMM state and thus the frame contributes to the classification. However, the variance is doubled, and thus the importance of the frame for classification is deemphasized.

Another limiting case occurs when the feature stream is extremely reliable in the absence of noise or distortions. Then, $\boldsymbol{\Sigma}_{\mathbf{x}_t|\mathbf{o}_{1:t}} = 0$, so that the posterior PDF $\mathrm{p}\left(\mathbf{x}_t|\mathbf{o}_{1:t}, q_t\right)$ approaches a Dirac delta impulse at the position of the clean speech feature vector. Plugging this into (4) immediately reveals the ordinary likelihood:

$$\begin{aligned}&E\left[\mathrm{p}\left(\mathbf{x}_t|q_t\right)|\mathbf{o}_{1:t}, q_t\right]|_{\boldsymbol{\Sigma}_{\mathbf{x}_t|\mathbf{o}_{1:t}}=0} \\ &\quad= \mathcal{N}\left(\mathbf{x}_t; \boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q\right) =: \mathcal{L}^{\mathrm{ML}}. \qquad (18)\end{aligned}$$

3. RELATION TO OTHER UNCERTAINTY DECODING RULES

In this section, we show that other uncertainty-of-observation decoding rules reported previously can be related to the general decoding rule of SD given in (4).

3.1. Observation Uncertainty

If $\mathrm{p}(\mathbf{x}_t|\mathbf{o}_{1:t}, q_t)$ is approximated by $\mathrm{p}(\mathbf{x}_t|\mathbf{o}_t)$ in (4), we get the decision rule of the so-called *observation uncertainty* (OU)

$$\begin{aligned}&\int \mathrm{p}(\mathbf{x}_t|q_t)\mathrm{p}(\mathbf{x}_t|\mathbf{o}_{1:t}, q_t)d\mathbf{x}_t \approx \\ &\qquad\qquad \int \mathrm{p}(\mathbf{x}_t|q_t)\mathrm{p}(\mathbf{x}_t|\mathbf{o}_t)d\mathbf{x}_t =: \mathcal{L}^{(\mathrm{OU})}. \qquad (19)\end{aligned}$$

This rule has been proposed by [6], and it was considered a heuristics, *which does not appear to arise from any mathematical framework* [18]. Equation (19) shows that it can be viewed as an approximation of SD, obtained under the assumption that the posterior of the clean features $\mathrm{p}(\mathbf{x}_t|\mathbf{o}_{1:t}, q_t)$ is unaffected by the model state q_t.

3.2. Modified Imputation

Equation (16) shows that the complexity of computing $\mathcal{L}_g^{\mathrm{SD}}$ increases quadratically with the number of mixtures K. However, a simplified version $\mathcal{L}_s^{\mathrm{SD}}$ of the SD likelihood can be obtained by neglecting all terms in (16) with $\kappa \neq \nu$ and simplifying the GMM weights to be only the output distribution weights ω_{q_κ}:

$$\mathcal{L}_s^{\mathrm{SD}} = \sum_{\kappa=1}^{K} \omega_{q_\kappa}\mathcal{N}\left(\tilde{\boldsymbol{\mu}}_{q_\kappa}; \boldsymbol{\mu}_{q_\kappa}, \boldsymbol{\Sigma}_{q_\kappa} + \tilde{\boldsymbol{\Sigma}}_{q_\kappa}\right). \qquad (20)$$

A further simplification is possible when the increase of the variance is neglected in (20). Then, the so-called *modified imputation* (MI) rule [11] is obtained

$$\mathcal{L}^{\mathrm{MI}} = \sum_{\kappa=1}^{K} \omega_{q_\kappa}\mathcal{N}\left(\tilde{\boldsymbol{\mu}}_{q_\kappa}; \boldsymbol{\mu}_{q_\kappa}, \boldsymbol{\Sigma}_{q_\kappa}\right). \qquad (21)$$

4. EXPERIMENTS AND RESULTS

4.1. Dataset

We have chosen signals uttered by five speakers (two male and three female) from the Grid database [15] to evaluate the new significance decoding rule defined in (16). As a training set, we have considered about 950 clean utterances (95%) of each speaker of the first four speakers. The rest of the utterances of the four speakers together with 50 utterances of the fifth speaker who did not appear in the training set have been considered as the clean test set.

The noisy test set is generated by distorting the clean test set with three different types of additive noise: white noise [19], speech babble [19] and office noise [20], each in the range of 0 to 15 dB. In the last test condition, using office noise, the speech signals have been convolved with the room impulse response (RIR) of an office [21] before the noise was added. The office dimensions are about $6.10 \times 4.30 \times 3.20$

m^3 and it has a reverberation time (T_{60}) of about 0.7 s. The RIR has been measured at a distance of 50 cm between loudspeaker and microphone.

4.2. Experimental setup

The training and test sets have been downsampled to $f_s = 8$ kHz and enhanced using a Wiener filter. The noise power estimate needed for the Wiener filter has been obtained using improved minima controlled recursive averaging (IMCRA) [22]. The variance of the Wiener filter in the STFT domain is used as the observation uncertainty $\mathbf{\Sigma}_{\mathbf{x}_t|\mathbf{o}_{1:t}}$ [23].

The mean and the variance extracted from the Wiener filter are propagated through the feature extraction stages [23]. The features are chosen to be the 13 static mel-frequency cepstral coefficients as well as the 26 delta and acceleration coefficients described in the ETSI advanced front end (AFE) [2].

The Java Audio-visual SPEech Recognizer (JASPER) [24] has been used in training and testing. A set of 52 HMMs (51 words of the Grid database + silence) has been trained using the clean training set. The HMMs are whole-word left-to-right linear models, with three states per phoneme. Each state has a GMM output distribution with four mixture components and diagonal covariance matrices.

4.3. Results

Table 1 compares the performance of SD defined in (16) with maximum likelihood defined in (18), uncertainty decoding according to [7], and modified imputation defined in (21) in terms of word accuracy , which is defined by:

$$\%Accuracy = \frac{N - D - S - I}{N} \cdot 100. \qquad (22)$$

In (22) (N), (S), (I) and (D) indicate the number of reference labels, substitutions, insertions and deletions, respectively.

The results shown in Table 1 indicate that SD outperforms all other approaches, with clear gains in almost every test condition. When the clean test set is confined to the four speakers used in training, we notice that the recognition accuracy is exactly the same for all decoding approaches as expected. However, SD outperforms the other approaches when a fifth speaker, unseen in the training data, is added to the clean test set. This indicates the adaptation capability of SD compared to the other approaches.

5. RELATION TO PRIOR WORK

Existing probabilistic uncertainty-of-observation approaches introduce the feature uncertainty provided by the front-end to the back-end classifier in one of two ways. One way, used by the uncertainty decoding approach [7] and its variants, e.g [8–10], is to use the uncertainty to dynamically compensate the statistical model parameters so that the model parameters better match the corrupted features. The second way is using the uncertainty to carry out model-based enhancement on the features themselves, as done by the imputation approaches like modified imputation [11].

The significance decoding approach shown here exploits the observation uncertainty to dynamically compensate both the features and the model parameters, by first carrying out model-based feature enhancement, and then dynamically adjusting the model variance to account for the residual observation uncertainty.

Table 1. Word accuracies result when the decoding rules of maximum likelihood (ML), uncertainty decoding (UD), modified imputation (MI) and significance decoding (SD) are used.

Noise Type	SNR [dB]	ML	UD	MI	SD
Babble	15	83.48	84.82	87.87	**88.30**
	10	66.31	68.37	73.33	**75.74**
	5	45.32	47.09	52.91	**56.17**
	0	25.53	27.45	35.18	**35.46**
White	15	69.65	76.52	87.73	**88.09**
	10	55.04	64.26	78.30	**78.58**
	5	42.77	48.79	65.74	**66.45**
	0	30.28	32.41	**50.35**	48.94
Office noise with reverberation	15	59.93	62.13	67.02	**67.59**
	10	45.32	49.43	57.80	**59.22**
	5	29.86	31.49	37.30	**38.01**
	0	22.77	23.12	**24.54**	24.47
Clean, 5 Spk.	-	95.46	95.46	95.39	**95.82**
Clean, 4 Spk.	-	**98.58**	**98.58**	**98.58**	**98.58**
Average	-	55.02	57.85	65.15	**65.82**

6. CONCLUSIONS

In this paper, we have presented a novel uncertainty-of-observation decoding approach, which extends the definition of the significance decoding rule introduced in [13] to Gaussian mixture models. The significance decoding rule is deduced based on the known concept [16] of replacing the needed but uncomputable observation likelihood by its conditional expectation given all relevant observable parameters.

Using this general form of significance decoding provides a notable ASR performance gain over a wide range of additive and convolutive noise conditions.

7. ACKNOWLEDGEMENTS

This work has been supported in part by the Ministry of Economic Affairs and Energy of the State of North Rhine- Westphalia, Grant IV.5-43-02/2-005-WFBO-009.

8. REFERENCES

[1] P. Vary and R. Martin, *Digital Speech Transmission*. John Wiley & Sons, 2006.

[2] *Speech Processing, Transmission and Quality Aspects (STQ); Distributed speech recognition; Advanced front-end feature extraction algorithm; Compression algorithms*, ETSI, ES.202.050 Std., 2003.

[3] J. Droppo, L. Deng, and A. Acero, "Evaluation of the SPLICE algorithm on the Aurora2 database," in *Proc. Eurospeech*, Aalborg, Denmark, September 2001, pp. 217–220.

[4] P. C. Loizou, *Speech enhancement: theory and practice*. CRC Taylor and Francis, 2007.

[5] M. Delcroix, S. Watanabe, and T. Nakatani, *Robust Speech Recognition of Uncertain or Missing Data*. Springer, 2011, ch. Variance Compensation for Recognition of Reverberant Speech with Dereverberation Preprocessing, pp. 225–255.

[6] J. A. Arrowood and M. A. Clements, "Using observation uncertainty in HMM decoding," in *Proc. International Conference on Spoken Language Processing*, Denver, Colorado, Sepember 2002.

[7] 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.

[8] J. Droppo, A. Acero, and L. Deng, "Uncertainty decoding with SPLICE for noise robust speech recognition," in *Proc. The International Conference on Acoustics, Speech, and Signal Processing*, vol. I, Orlando, Florida, USA, May 2002, pp. 57–60.

[9] H. Liao and M. J. F. Gales, "Joint uncertainty decoding for noise robust speech recognition," in *Proc. Interspeech*, Lisbon, Portugal, Sep. 2005.

[10] V. Ion and R. Haeb-Umbach, "A novel uncertainty decoding rule with applications to transmission error robust speech recognition," *IEEE Trasactions on Audio, Speech and Language Processing*, vol. 16, no. 5, pp. 1047–1060, 2008.

[11] D. Kolossa, A. Klimas, and R. Orglmeister, "Separation and robust recognition of noisy, convolutive speech mixtures using time-frequency masking and missing data techniques," in *Proc. IEEE Workshop on Application of Signal Processing to Audio and Acoustics*, Mohonk Mountain House, New Paltz, New York, USA, October 2005, pp. 82–85.

[12] A. H. Abdelaziz, S. Zeiler, and D. Kolossa, "Audio-visual speech recognition for uncertain acoustical observations," in *Proc. ITG Fachtagung Sprachkommunikation*, Braunschweig, Germany, September 2012.

[13] A. H. Abdelaziz and D. Kolossa, "Decoding of uncertain features using the posterior distribution of the clean data for robust speech recognition," in *Proc. Interspeech*, Portland, Oregon, USA, Sptember 2012.

[14] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, February 1989.

[15] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421–2424, November 2006.

[16] J. Bilmes, "A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models," ICSI, Tech. Rep. TR-97-021, 1997.

[17] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*. Wiley-Interscience, 2001, ch. Maximum-Likelihood and Bayesian parameter estimation, pp. 95–97.

[18] H. Liao and M. Gales, "Issues with uncertainty decoding," *Speech Communication*, vol. 50, no. 4, pp. 265–277, April 2008.

[19] Institute for Perception-TNO and Speech Research Unit-RSRE, retrieved November 2012. [Online]. Available: http://spib.rice.edu/spib/data/signals/noise/

[20] "Sound ideas," The General Sound Effects Library, Series 6000, retrieved November 2012. [Online]. Available: http://www.sound-ideas.com/6000.html

[21] A. Kitzig, "Niederrhein University room impulse response package information," Niederrhein University of Applied Sciences, Tech. Rep., June 2010.

[22] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Transactions on Speech and Audio*, vol. 11, no. 5, p. 466–475, 2003.

[23] R. F. Astudillo, D. Kolossa, and R. Orglmeister, "Accounting for the uncertainty of speech estimates in the complex domain for minimum mean square error speech enhancement," in *Proc. Interspeech*, Brighton, United Kingdom, 2009.

[24] A. Vorwerk, S. Zeiler, D. Kolossa, R. F. Astudillo, and D. Lerch, *Robust Speech Recognition of Uncertain or Missing Data*. Springer, 2011, ch. Use of Missing and Unreliable Data for Audiovisual Speech Recognition, pp. 345–373.