Investigations into Uncertainty Decoding Employing a Discrete Feature Space for Noise Robust Automatic Speech Recognition

Valentin Ion, Reinhold Häb-Umbach

University of Paderborn, Dept. of Communications Engineering, 33098 Paderborn
E-Mail: {ion,haeb}@nt.uni-paderborn.de
Web: www-nt.uni-paderborn.de

Abstract

This paper addresses the robustness of automatic speech recognition to environmental noise. In order to account for reliability of the clean feature estimate we employ the feature posterior density conditioned on observed noisy features to perform uncertainty decoding. We investigate two approaches to estimate the posterior using a discrete feature space, first conditioning only on the current observation, and second on the whole feature sequence of an utterance. Experiments with Aurora 2 showed that the latter provides slightly better performance, as it allows for exploiting the temporal correlations between consecutive features.

1 Introduction

There are numerous studies about the sensitivity of ASR to a mismatch between the acoustic models of the clean speech and the noisy speech features observed in an arbitrary environment. A solution to this problem is to denoise the speech waveform prior to feature extraction by spectral subtraction, or by Wiener filtering [3]. Even with perfect knowledge of the noise power spectral density, the clean speech estimate is still inaccurate since usually the difference between clean and noise phases is unknown. Moreover, in practice, the noise spectra is not perfectly known. Therefore, the resulting clean speech feature estimate is an approximation of the true clean feature and thus, reliable only to some restricted extent. It has been already shown that accounting for the feature reliability, derived from the posterior density of the clean feature conditioned on distorted observations, is beneficial for noise robust [5, 1, 7] and for transmission error robust ASR [4].

The key element of uncertainty decoding approaches is the determination of the clean feature posterior density. In [8], the feature posterior has been evaluated by assessing the effect of the noisy environment on the distributions of the clean speech in log power spectra domain. The noisy speech distribution was assumed normal, and its parameters have been derived form those of the clean speech distribution by applying corrections to its mean and variance vectors. The correction factors have either been learned from stereo-data (SPLICE) or estimated iteratively with an EM algorithm aiming at maximizing the likelihood of the noisy observations. They have been subsequently employed to compute a clean feature point estimate (RATZ), or to directly compensate the acoustic distributions modeled by HMMs in the decoder (STAR). Another approach widely used to compute noisy speech distribution given the clean speech is Vector Taylor Series (VTS) [8, 6] and is based on linearization of the analytical nonlinear relationship between clean, noisy feature, and noise. Thus, for VTS, an estimate of the noise cepstrum at the time of observation have to be available. As a rough approximation of this, the global mean of the noise cepstrum, i.e. estimated on non-speech frames, can be used. However, the noise estimate can be improved as proposed in [8] by a ML reformulation of VTS.

While the feature posterior obtained by these approaches is conditioned only on the current observation, it is expected that conditioning on more observations may improve the performance, as the consecutive clean features are not statistically independent. In [4] we proposed an approach to obtaining the clean feature posterior in a distributed speech recognition scenario, where the transmission errors distort the clean feature. The components of the feature vector are independently corrupted by bit errors which occur sporadically. Thus, unreliable regions of the time-cepstral space tend to be isolated. This allowed to obtain a more informative feature posterior by conditioning it on several past and future observations rather than only on the current observation. The robustness increased especially when the distortion was bursty, as is the case of packet loss during transmission. When the distortion is the environmental noise, a similar situation may occur if the clean speech energy is temporarily below the noise level so that the observed feature approaches the noise cepstrum and becomes uninformative.

This work is aimed at investigating the potential of uncertainty decoding with provision for temporal correlation in the feature sequence. In view of this we attempt to build on the framework presented in [4] which takes advantage of a discrete representation of the feature space by vector quantization. This allows the feature posterior and the a priori knowledge such as cluster probabilities and cluster transition probabilities to take any shape and thus relaxes some usual constrains, e.g. normal probability densities.

The reminder of this paper is organized as follows; The next section shortly reviews the uncertainty decoding. Section 3 describes the observation model involved in feature posterior computation of Section 4. Details about the partitioning of the feature space are given in Section 5. Section 6 contains experimental results and is followed by conclusions.

2 Uncertainty Decoding Framework

This work employs the novel uncertainty decoding (UD) approach which has been exhaustively described in [4] for transmission error robust ASR. However, the focus here is on the computation of the feature posterior density when the distortion occurs by environmental noise rather than channel errors. Figure 1 depicts the processing steps of this approach.

In the first block, the observation model is linearized around the codewords $\mathbf{x}^{(i)}$ in order to evaluate the noisy speech density corresponding to the clean speech of the ith cluster. This is subsequently employed to compute the discrete posterior $P(\mathbf{x}^{(i)}|\mathbf{y})$ where $\mathbf{y}$ denotes either $\mathbf{y}_t$, or

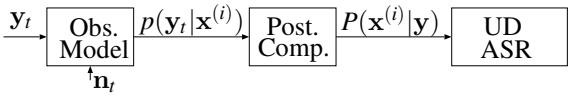

Figure 1: *Uncertainty decoding framework.*

$\mathbf{y}_1^T$, see Section 4. In the decoder, the discrete posterior is used to determine the parameters of a continuous posterior density, which in this work is approximated by a normal distribution. The ASR with uncertainty decoding rule has been presented in [4].

3 Observation Model

The phase-sensitive observation model in cepstral domain introduced by [6] has been employed, and it is briefly reviewed in the following. The feature vectors are the Mel-Frequency Cepstral Coefficients (MFCCs) obtained by applying a set of L Mel-scale filters to the short term power spectrum of the speech, taking the logarithm of them and multiplying this by the discrete cosine transform (DCT) matrix. According to [6], the N-dimensional cepstral vectors of noisy speech $\mathbf{y}$, noise $\mathbf{n}$, and clean speech $\mathbf{x}$ are related by the non-linear equation:

$$\begin{aligned}\mathbf{y} &= \mathbf{x}+\mathbf{\Phi}\log(1+e^{\mathbf{\Phi}^{-1}(\mathbf{n}-\mathbf{x})}+2\alpha e^{\mathbf{\Phi}^{-1}\frac{\mathbf{n}-\mathbf{x}}{2}}) \quad (1)\\ &= \mathbf{x}+g_\alpha(\mathbf{x},\mathbf{n})\end{aligned}$$

Here, $\mathbf{\Phi}$ is the DCT matrix, $\mathbf{\Phi}^{-1}$ its pseudo-inverse. α is a phase-factor vector with L values, one for each Mel filter bank, which is experimentally tuned, see discussion in [6]. Multiplication by α is element-wise. While the phase insensitive model is equivalent to setting all $\alpha = 0$, we accounted for phase by setting $\alpha = 1.5$ in all experiments. Note that the imperfect noise estimate leads itself to uncertainty in the clean feature estimate so that the results do not significantly change using various settings for α.

3.1 Linear Approximation

A linear approximation of (1) around the point $(\mathbf{x}_0,\mathbf{n}_0)$ is given by the first two terms of a VTS expansion [6]:

$$\begin{aligned}\mathbf{y} &\approx \mathbf{x}_0+g_\alpha(\mathbf{x}_0,\mathbf{n}_0)+ \quad (2)\\ &\quad +\frac{\partial\mathbf{y}}{\partial\mathbf{x}}|_{\mathbf{x}_0,\mathbf{n}_0}(\mathbf{x}-\mathbf{x}_0)+\frac{\partial\mathbf{y}}{\partial\mathbf{n}}|_{\mathbf{x}_0,\mathbf{n}_0}(\mathbf{n}-\mathbf{n}_0)\\ &= \mathbf{x}_0+g_\alpha(\mathbf{x}_0,\mathbf{n}_0)+\mathbf{G}(\mathbf{x}-\mathbf{x}_0)+(\mathbf{I}-\mathbf{G})(\mathbf{n}-\mathbf{n}_0)\end{aligned}$$

$\mathbf{G}$ and $\mathbf{I}-\mathbf{G}$ denote the Jacobian matrices with respect to the variable $\mathbf{x}$ and $\mathbf{n}$, respectively, and $\mathbf{I}$ is the identity matrix. The matrix $\mathbf{G}$ is evaluated as:

$$\mathbf{G} = \mathbf{I}-\mathbf{\Phi}\mathbf{\Lambda}\mathbf{\Phi}^{-1}, \quad (3)$$

in which Λ is a diagonal matrix $\mathbf{\Lambda} = diag(\lambda)$ with:

$$\lambda = \frac{e^{\mathbf{\Phi}^{-1}(\mathbf{n}_0-\mathbf{x}_0)}+\alpha e^{\mathbf{\Phi}^{-1}\frac{\mathbf{n}_0-\mathbf{x}_0}{2}}}{1+e^{\mathbf{\Phi}^{-1}(\mathbf{n}_0-\mathbf{x}_0)}+2\alpha e^{\mathbf{\Phi}^{-1}\frac{\mathbf{n}_0-\mathbf{x}_0}{2}}}. \quad (4)$$

3.2 Distribution of the noisy feature

The idea of our approach is to partition the clean cepstral feature space and model the distribution within each resulted region, which is identified by a codeword $\mathbf{x}^{(i)}$, by a Gaussian density $\mathcal{N}(\mathbf{x};{\mu_\mathbf{x}}^{(i)},{\mathbf{\Sigma}_\mathbf{x}}^{(i)})$.

The assumption that the observation model is linear around $\mathbf{x}^{(i)}$, allows for computing the distribution of the noisy feature after the distortion process as a Gaussian distribution, $\mathcal{N}(\mathbf{y};{\mu_\mathbf{y}}^{(i)},{\mathbf{\Sigma}_\mathbf{y}}^{(i)})$. The linear transformation parameters depend on the current noise estimation $\mu_{\hat{\mathbf{n}}}$ and its estimation error variance $\mathbf{\Sigma}_{\hat{\mathbf{n}}}$. Using (2) at the point $({\mu_\mathbf{x}}^{(i)},\mu_{\hat{\mathbf{n}}})$ they become:

$$\begin{aligned}{\mu_\mathbf{y}}^{(i)} &= {\mu_\mathbf{x}}^{(i)}+g_\alpha({\mu_\mathbf{x}}^{(i)},\mu_{\hat{\mathbf{n}}}), \quad (5)\\ {\mathbf{\Sigma}_\mathbf{y}}^{(i)} &= \mathbf{G}{\mathbf{\Sigma}_\mathbf{x}}^{(i)}\mathbf{G}'+(\mathbf{I}-\mathbf{G})\mathbf{\Sigma}_{\hat{\mathbf{n}}}(\mathbf{I}-\mathbf{G})', \quad (6)\end{aligned}$$

with $\mathbf{G}'$ being the transposed matrix of $\mathbf{G}$.

4 Feature Posterior Computation

The distribution of the noisy feature corresponding to a codeword $\mathbf{x}^{(i)}$, estimated for each time instance t, can be employed to infer the posterior probability of each region of the clean feature space at that time. Most authors have proposed the straight forward approach, labeled UD0 in the following, which neglects the correlation between the current and past/future clean features. By doing so, the posterior is conditioned only on $\mathbf{y}_t$. In the approach investigated in our work (UD1) the correlation is modelled by the transition probability $P(\mathbf{x}_t^{(i)}|\mathbf{x}_{t-1}^{(j)})$, e.g. from the codeword j to i. This is used to infer the posterior conditioned on all observations $\mathbf{y}_1\ldots\mathbf{y}_T$, i.e. $P(\mathbf{x}^{(i)}|\mathbf{y}_1^T)$.

In the first step, common to both approaches, the likelihood of the observation $\mathbf{y}_t$ is evaluated for each cluster:

$$p(\mathbf{y}_t|\mathbf{x}^{(i)}) = \mathcal{N}(\mathbf{y};{\mu_\mathbf{y}}^{(i)},{\mathbf{\Sigma}_\mathbf{y}}^{(i)})|_{\mathbf{y}=\mathbf{y}_t} \quad (7)$$

In case of UD0, the posterior probability of each cluster i is obtained by applying the Bayes' rule:

$$P(\mathbf{x}^{(i)}|\mathbf{y}_t) = \frac{p(\mathbf{y}_t|\mathbf{x}^{(i)})P(\mathbf{x}^{(i)})}{\sum_{j=1}^{N}p(\mathbf{y}_t|\mathbf{x}^{(j)})P(\mathbf{x}^{(j)})} \quad (8)$$

In case of UD1, a Forward-Backward recursion evaluates the forward and backward probabilities $P(\mathbf{y}_1^t,\mathbf{x}^{(i)})$ and $P(\mathbf{x}^{(i)}|\mathbf{y}_{t+1}^T)$. The discrete feature posterior is:

$$P(\mathbf{x}^{(i)}|\mathbf{y}_1^T) = \frac{P(\mathbf{y}_1^t,\mathbf{x}^{(i)})P(\mathbf{x}^{(i)}|\mathbf{y}_{t+1}^T)}{\sum_{j=1}^{N}P(\mathbf{y}_1^t,\mathbf{x}^{(j)})P(\mathbf{x}^{(j)}|\mathbf{y}_{t+1}^T))} \quad (9)$$

Using (8) or (9) we approximate a continuously-valued posterior density, $p(\mathbf{x}_t|\mathbf{y})$ (here we used $\mathbf{y}$ to denote either $\mathbf{y}_t$ or $\mathbf{y}_1^t$) with diagonal covariance in the form of:

$$p(\mathbf{x}_t|\mathbf{y}) = \mathcal{N}(\mathbf{x}_t;\mu_{\mathbf{x}_t|\mathbf{y}},\mathbf{\Sigma}_{\mathbf{x}_t|\mathbf{y}}) \quad (10)$$

where the Gaussian parameters are given by:

$$\begin{aligned}\mu_{\mathbf{x}_t|\mathbf{y}} &= \sum_{i=1}^{N}P(\mathbf{x}^{(i)}|\mathbf{y}){\mu_\mathbf{x}}^{(i)} \quad (11)\\ \mathbf{\Sigma}_{\mathbf{x}_t|\mathbf{y}} &= \sum_{i=1}^{N}P(\mathbf{x}^{(i)}|\mathbf{y})(\mu_{\mathbf{x}_t|\mathbf{y}}-{\mu_\mathbf{x}}^{(i)})^2 \quad (12)\end{aligned}$$

The recognition is then carried out with the complete posterior density (UD0 or UD1) by employing the uncertainty decoding rule. In the experimental part of this work,

we present also results obtained by classical plug-in decoding rule, with the minimum mean squared error (MMSE) point estimate $\hat{\mathbf{x}}_t|_{MMSE} = \mu_{\mathbf{x}_t|\mathbf{y}}$ of the clean feature, approach labeled MMSE0 and MMSE1, respectively.

5 Partitioning approach

The clustering of the clean cepstral space has to adhere to some contradictory constraints. On the one hand, the regions delimited by the cluster borders must be small enough to ensure that the quantization error does not significantly deteriorates the recognition accuracy, and the linear approximation holds. On the other hand, this may result in a very large number of clusters and poses difficulties due to computation complexity of (6) and (7) which linearly increases with the number of clusters N.

As the observation model encompasses the full cepstral vector with 13 components, a split vector quantization scheme, as in [2, 3], cannot be directly applied here. In some preliminary investigations, in order to evaluate the VQ approach, we employed a feature vector with only 6 cepstral (static) components for recognition. Although the noise robustness showed considerable improvement compared to the performance of the standard front-end [2] with the same number of vector components, this non-standard choice of vector dimensionality may work properly only for small vocabulary recognition. A much better quantization scheme that we found has been to consider each Gaussian component of the HMM state conditioned mixture as a quantization cluster. Details of both approaches are given in the following.

5.1 Vector Quantization (VQ)

In this approach, the 6-dimensional cepstral space $(c_0, c_1, \ldots, c_5)$ has been partitioned using the Lloyd-Max algorithm. The criteria was to minimize the mean-squared quantization error based on a Mahanalobis distance measure. The training was performed on the clean training set of Aurora 2.

We carried out some experiments with the goal to evaluate the performance degradation which occurs due to smaller vector dimensionality and due to the rough quantization. For comparison, note that e.g. [2] employs split VQ with a total of 42 bits for a 14 components vector.

Table 1 gives the recognition accuracy on the clean test set obtained with this setup with different settings for the number of clusters (1024 and 265). The column labeled MFCC-13 shows results without quantization, and using a 39-components feature vector (13 static MFCCs and their first and second-order derivatives). For the rest of experiments the feature vector had 18 components; 6 static MFCCs and their derivatives. The acoustic models were those of the standard Aurora 2 recognizer, i.e. word-level HMMs, 3 Gaussian components per state for each word and 6 components for silence model.

Table 1: Word accuracies [%] on the Aurora 2 test set A. "noisy" denotes the average over SNR 0-20 dB.

	MFCC-13	MFCC-6	VQ-1024	VQ-256
clean	99.29	99.16	98.59	98.22
noisy	61.32	46.59	42.08	38.44

While in a clean scenario the VQ and the dimensionality reduction produced a relatively small performance degradation, in the noisy scenario, the robustness has been drastically reduced mainly due to the dimensionality reduction, i.e from 61.32% to 46.59%.

VQ while keeping the original dimensionality has not been investigated since it would require much more clusters and thus, is not of practical interest.

5.2 HMM Mixture Quantization (MQ)

The idea of using the acoustic mixtures as quantization clusters is based on the fact that the MFCC vectors are not uniformly spread over the cepstral space. They tend to be grouped in clusters corresponding to each acoustic unit. Therefore, it is not necessary to project the whole clean space onto the noisy one, but only those regions to that most of the features of an acoustic units belong.

A very simple method which we employed in this work was to train HMM acoustic models for recognition using standard ML training with feature vectors consisting of 13 static MFCC components. The resulting Gaussian mixture densities were indexed and constituted a pool of clusters. The quantization is performed as follows; given a feature vector, the Gaussian density which maximizes its likelihood is chosen from the pool and its index represents the codeword. The dequantization occurs by retrieving the real-valued Gaussian mean vector with that index.

The word accuracies obtained using the standard recognizer (see Section 5.1) and the quantized/dequantized features are shown in Table 2. We used 546, 361, and 184 mixtures. MFCC-13 are the baseline results without quantization. While in the clean environment, using a comparable number of clusters, the performance is slightly degraded, in the noisy environment MQ is more robust, i.e. only 10% degradation compared with 30% for VQ.

Table 2: Word accuracies [%] on the Aurora 2 test set A. "noisy" denotes the average over SNR 0-20 dB.

	MFCC-13	MQ-546	MQ-361	MQ-184
clean	99.29	98.29	97.61	96.95
noisy	61.32	53.81	52.82	50.82

Note that the results of this section were obtained without employing the proposed enhancement technique. The results in the clean environment reveal the upper bound of word accuracy attainable with each particular quantization scheme.

6 Robust ASR Experiments

In this section we present the results of our investigations with regard to improving the noise robustness by using the clean speech estimate and its variance at decoding. As stated in Section 3, the approach proposed in this work requires an estimate of the noise cepstrum. Since the goal was not necessarily to develop an enhanced noise estimation scheme, we used two simple methods.

The first one is artificial, in the sense that the true noise cepstrum is in fact available, i.e. for each utterance we subtracted the clean waveform from the noisy one and computed the cepstral vector, but, however, this was averaged (moving average - MAVG) over 20 consecutive noise cepstra and then used as noise estimate. The squared standard deviation over this interval was considered as the variance of the estimate. This should mimic the uncertainty in deter-

mining the instantaneous value of noise cepstrum, which usually occurs performing a real estimation.

The second estimation method used the first 20 noisy features of each utterance as estimation interval (global average - GAVG). This is a global rough estimation of the time variant noise cepstrum, however it has been widely used in noise robustness evaluations on Aurora 2.

As a reference, the word accuracies on Aurora 2 noisy set A, obtained with the noise robust front-end standardized by ETSI [3] (AFE) was 87.87%. However, note that while AFE reaches 99.32% in the clean conditions, the accuracy of our approach is limited to the values given in Tables 1 and 2.

Table 3 shows the word accuracies using the MAVG noise estimation. UD0 and UD1 denote uncertainty decoding while MMSE0 and MMSE1 were obtained with classical decoding using the expectation of the feature posterior, see Section 4.

Table 3: Word accuracies [%] with MAVG

	MMSE0	UD0	MMSE1	UD1
MQ-546	86.07	87.33	-	-
MQ-361	84.50	85.15	85.48	85.94
MQ-184	80.62	81.46	81.67	82.25
VQ-256	82.88	82.05	69.00	70.19

Table 4 shows the word accuracies obtained using the noise estimation on first 20 non-speech frames.

Table 4: Word accuracies [%] with GAVG

	MMSE0	UD0	MMSE1	UD1
MQ-546	82.42	82.99	-	-
MQ-361	82.49	83.14	80.42	80.93
MQ-184	74.45	78.03	72.74	73.06
VQ-256	76.48	77.11	71.08	71.08

For almost all quantization schema, the results of both tables reveal that accounting for estimation variance by UD instead of using the plug-in rule with MMSE estimate, yields improvement.

The rows MQ-184 and MQ-361 of Table 3 confirm our expectations that considering the temporal correlations (MMSE1 and UD1) is beneficial to increasing robustness. This is however a technical challenge since it requires the estimation of transition probabilities $P(\mathbf{x}_t^{(i)}|\mathbf{x}_{t-1}^{(j)})$ which can be a large matrix ($N \times N$). Although the performance improves by increasing the number of partitions N, the highest value which we used was 546. Given the amount of available training data of Aurora 2, the matrix estimation has not been possible at this value and thus the experiments MMSE1 and UD1 could not have been performed.

An interesting discussion arises from comparing the results of Tables 3 and 4; While the performance degradation by using a less accurate noise estimation is normal, the dramatical loss with MMSE1 and UD1 (in Table 4) may seem unexpected. However, this can be explained as follows; In case of the noise estimation using the first 20 non-speech frames, the estimation variance is higher. That is, there are more occurrences in which the true instantaneous noise cepstrum significantly differs from the estimated one, here the temporal mean value over an interval. At time instances where the true value is well above the estimate, the estimated clean feature is inaccurate, i.e. still noisy. Moreover, it has a small variance, as the SNR evaluated with the poor noise estimate is higher than the real one. Thus, the estimated clean feature is still noisy but its contribution at decoding is not diminished since its variance is low. While without considering correlations this affects only the current frame, the effect is amplified when the dependency of succeeding features on the current one is considered. This fact is confirmed also by VQ-256 where MMSE1 and UD1, even with the more accurate moving average noise estimate performed worse than MMSE0 and UD0. As already mentioned, VQ-256 is more sensitive to noise due to lower feature dimensionality.

7 Conclusions

We described an approach to obtaining the clean feature posterior using a picewise linearization of the observation model in the cepstral domain. The partitioning of the clean feature space allowed us to estimate the transition probabilities between the clusters and employ them to compute a posterior density conditioned on a sequence of observed noisy features, rather than only on the current observation. The decoding using this posterior has been shown to be more robust, however, the improvement is dependent of the accuracy of the noise estimation.

Acknowledgements

This work was supported by Deutsche Forschungsgemeinschaft under contract number HA 3455/2-3.

References

[1] J. Droppo, A. Acero, and L. Deng. Uncertainty decoding with SPLICE for noise robust speech recognition. In *Proc. of ICASSP, Orlando, Florida, May 2002*.

[2] ES.201.108. Speech processing, Transmission and Quality aspects (STQ); Distributed speech recognition;Front-end feature extraction algorithm; Compression algorithms. *ETSI*, April 2000.

[3] ES.202.050. Speech processing, Transmission and Quality aspects (STQ); Distributed speech recognition; Advanced front-end feature extraction algorithm; Compression algorithms. *ETSI*, Oct 2002.

[4] V. Ion and R. Haeb-Umbach. A novel uncertainty decoding rule with applications to transmission error robust speech recognition. *IEEE Trans. on Audio, Speech, and Language Processing*, 16:1047–1060, 2008.

[5] Trausti T. Kristjansson and Brendan J. Frey. Accounting for uncertainty in observations: A new paradigm for robust automatic speech recognition. In *Proc. of ICASSP*, 2002.

[6] Jinyu Li, Li Deng, Dong Yu, Yifan Gong, and A. Acero. HMM adaptation using a phase-sensitive acoustic distortion model for environment-robust speech recognition. In *Proc of. ICASSP*, 2008.

[7] H. Liao and Gales M. J. F. Issues with uncertainty decoding for noise robust automatic speech recognition. *Speech Communication*, 50:265–277, 2008.

[8] P. Moreno. *Speech Recognition in Noisy Environments*. PhD thesis, Carnegie Mellon University, 1996.