Technical Report TR70, August 2006

Technical Report: OSU-CISRC-8/06-TR70
Department of Computer Science and Engineering
The Ohio State University
Columbus, OH 43210-1277, USA
Web site: **http://www.cse.ohio-state.edu/research/techReport.shtml**
Ftp site: **ftp.cse.ohio-state.edu**
Login: **anonymous**
Directory: **pub/tech-report/2006**
File in pdf format: **TR70.pdf**

Transforming binary uncertainties for robust speech recognition

Soundararajan Srinivasan
Biomedical Engineering Department
The Ohio State University
Columbus, OH, 43210-1277, USA
srinivasan.36@osu.edu

DeLiang Wang
Department of Computer Science and Engineering & Center of Cognitive Science
The Ohio State University
Columbus, OH, 43210-1277, USA
dwang@cse.ohio-state.edu

Abstract

Recently several algorithms have been proposed to enhance noisy speech by estimating a binary mask that can be used to select those time-frequency regions of a noisy speech signal that contain more speech energy than noise energy. This binary mask encodes the uncertainty associated with enhanced speech in the linear spectral domain. The use of the cepstral transformation smears the information from the noise dominant time-frequency regions across all the cepstral features. We propose a supervised approach using regression trees to learn the non linear transformation of the uncertainty from the linear spectral domain to the cepstral domain. This uncertainty is used by a decoder that exploits the variance associated with the enhanced cepstral features to improve robust speech recognition. Systematic evaluations on a subset of the Aurora4 task using the estimated uncertainty shows substantial improvement over the baseline performance across various noise conditions.

I. INTRODUCTION

The performance of automatic speech recognizers (ASRs) degrades rapidly in the presence of noise and other distortions [16]. To mitigate the effect of noise on recognition, noisy speech is typically preprocessed by speech enhancement algorithms, such as spectral subtraction [6], [21]. However, the accuracy of these algorithms often varies widely across time frames. Additionally, the variances of enhanced features also differ even within a time frame. Recently an uncertainty decoding approach to robust speech recognition has been proposed to effectively account for the varied accuracies of features derived from front-end preprocessing [13]. In this approach, computation of the observation probability during recognition involves integration over all possible speech feature values. Specifically, speech enhancement uncertainties contribute to an increase in the variance of acoustic model variables. It is shown in [13] that the uncertainty decoder significantly outperforms the conventional ASR operating on the enhanced speech features.

Currently most algorithms estimate the uncertainty associated with the enhanced speech features in either the log Mel-frequency domain or directly in the cepstral domain [2], [13], [20], [37]. Hence, the uncertainty decoder is coupled with speech enhancement algorithms operating in these domains. However, several speech enhancement algorithms operate in the linear spectral domain. In particular, many recent methods attempt to estimate a binary time-frequency mask that can be used to select those time-frequency (T-F) regions of a noisy speech signal that contain more speech energy than noise energy [1], [15], [30], [31], [42]. Some methods rely on the observation that individual signals in a mixture are sparsely distributed in the time-frequency domain [30], [42]. This enables them to handle a variety of mixing conditions, including those involving more sources than sensors [26]. The use of a binary mask as the computational goal makes only weak assumptions about interference conditions. Further, estimation of the binary mask imposes a lesser demand on the speech enhancement front-end and is often more robust than full-band speech enhancement [35], [39].

Signals reconstructed from such masks have been shown to be substantially more intelligible for human listeners than original mixtures [10], [30]. However, conventional ASR systems are extremely sensitive to the distortions produced during resynthesis. To minimize the effect of distortions on recognition, these speech enhancement systems are currently coupled with a missing-data recognizer [11], [22], [30]. Missing-data ASR attempts to improve robust speech recognition by distinguishing between reliable and unreliable data in the T-F domain. It uses the binary mask generated by speech enhancement algorithms to label the speech-dominant T-F regions as reliable and the rest as unreliable. While the performance of the missing data recognizer is significantly better than the performance of a system using front-end speech enhancement followed by recognition of enhanced speech [11], a significant disadvantage of the missing data recognizer is that recognition is performed in the spectral or the T-F domain. It is well known that recognition using cepstral coefficients yields a superior performance compared to recognition using spectral coefficients under clean speech conditions [12]. In addition, the performance of the missing-data ASR degrades as the vocabulary size increases [35]. Attempts to adapt the missing data method to the cepstral domain have centered around reconstruction or imputation of the missing values in the spectral domain followed by transformation to the cepstral domain [28]. This reconstruction is typically based on a trained speech prior.

Although the spectrogram reconstruction method in [28] provides promising results, errors in mask estimation and subsequent reconstruction degrade the performance of the ASR. In this

paper, we present a two-step, supervised learning approach to estimate the uncertainty associated with the reconstructed cepstra. In the first step, we estimate the uncertainty in the spectral domain by utilizing the statistical information contained in the speech prior used in spectrogram reconstruction. In the second step, this uncertainty is transformed to the cepstral domain using a nonlinear regression model. Specifically, we employ a nonparametric learning approach using regression trees to directly estimate the uncertainties associated with the static, dynamic, and acceleration cepstral coefficients. We thus convert the binary uncertainty encoded by the T-F mask into a real-valued uncertainty associated with the cepstral features. The estimated cepstral-domain uncertainty is utilized by an uncertainty decoder during recognition. We show that the resulting system improves the recognition performance over that of the conventional ASR across various noise conditions.

The rest of the paper is organized as follows. The next section briefly reviews the uncertainty decoding framework for robust speech recognition. Section III contains a detailed presentation of the proposed method for estimating the uncertainty associated with the reconstructed cepstra. The method has been systematically evaluated on a subset of the Aurora4 noisy speech recognition task and the evaluation results are presented in Section IV. This section also contains a performance comparison of the missing-data ASR and the uncertainty decoder on a digit recognition task. Finally, conclusions and future work are given in Section V.

II. UNCERTAINTY DECODING

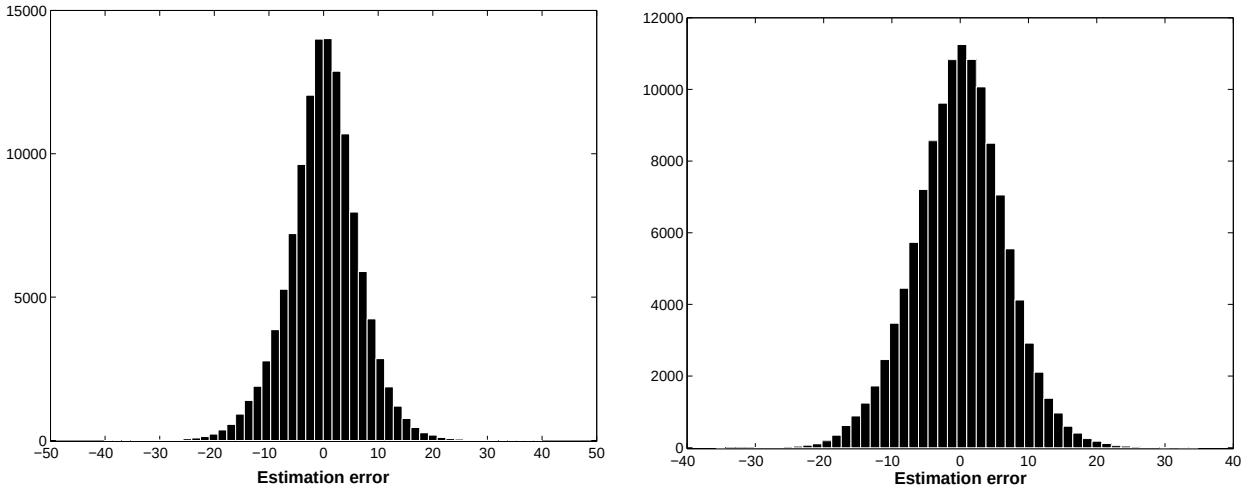
A typical approach for robust speech recognition involves preprocessing a noisy speech signal by a speech enhancement algorithm to produce an estimate of the clean speech features. These features are then used directly in the evaluation of the acoustic model probability in ASR systems. As discussed in the introduction, the performance of front-end denoising algorithms is often inconsistent. This inconsistency could potentially change the mean and the variance of the features extracted. Conventional ASR systems are especially sensitive to changes in the variance of the features derived from the output of speech enhancement algorithms [8]. The uncertainty decoding method accounts for such distortions in speech enhancement by integrating the probability of the observed features over all possible speech feature values [13]. Hence, the new observation likelihood is computed as

$$\int_{-\infty}^{\infty} p(z|M)p(z|\theta)dz, \tag{1}$$

wheres z is the clean speech feature seen during training and M denotes a parameterized model of the observation density. Following the suggestion in [13] we assume that the front-end compensation model, parameterized by θ, can be characterized as:

$$p(z|\theta) = N(z; \hat{z}, \Sigma_{\hat{z}}), \tag{2}$$

where $\hat{z}$ is the enhanced feature. The model in (2), therefore, states that the error in the estimation of the clean speech feature, $z - \hat{z}$, is Gaussian distributed with zero mean and a variance of $\Sigma_{\hat{z}}$. For many speech enhancement algorithms, this is a valid assumption. Fig. 1 shows the histograms of the deviation of two estimated cepstral coefficients from the true ones. The speech samples are derived from the clean and noisy development portions of the Aurora4 database [24]. The noise source corresponds to a restaurant environment. The speech enhancement algorithm



(a) (b)

Fig. 1. Histograms of the errors in the estimation of clean speech features using a speech enhancement algorithm. Statistics are obtained using clean speech and speech corrupted with restaurant noise from the Aurora4 database. (a) Histogram of the estimation error for the 4^{th} order cepstral coefficient. (b) Corresponding results for the 11^{th} order cepstral coefficient.

used is a spectral subtraction algorithm (see Section IV). Fig. 1(a) shows the error distribution corresponding to the 4^{th} order cepstral coefficient. Similarly, Fig. 1(b) shows the error distribution corresponding to the 11^{th} order cepstral coefficient. Note that the distributions can be well approximated by zero mean Gaussians.

The observation density in each state of a hidden Markov model (HMM) based ASR is usually modeled as a mixture of Gaussian densities. Therefore,

$$p(z|k,q) = N(z; \mu_{k,q}, \Sigma_{k,q}) \tag{3}$$

is the likelihood of observing z given state q and mixture k; $\mu_{k,q}$ and $\Sigma_{k,q}$ are the mean and the variance of the Gaussian mixture component. When noisy speech is processed by unbiased speech enhancement algorithms, it is shown in [13] that the observation likelihood should be computed as

$$\int_{-\infty}^{\infty} p(z|k,q)p(z|\theta)dz = N(\hat{z}; \mu_{k,q}, \Sigma_{k,q} + \Sigma_{\hat{z}}). \tag{4}$$

The role of the uncertainty associated with the enhanced features can be seen in (4) as that of increasing the variance of individual Gaussian mixture components. Hence, those enhanced speech features that deviate more from clean ones will contribute less to the overall likelihood. For example, from Fig. 1 we can observe that the variance of the error distribution corresponding to the 4^{th} order cepstral coefficient is smaller than that of the 11^{th} order coefficient. Hence, the observation likelihood extracted from the former can be expected to contribute more to the final acoustic model score.

It is shown in [2], [4], [13] that the utilization of the speech feature uncertainty contributes to a significant improvement in the ASR accuracy on small vocabulary tasks. The performance improvement is particularly substantial when the variance of the enhanced features is known *a priori* [13]. Hence, an accurate estimate of the speech feature uncertainty is critical for realizing the full benefits from uncertainty decoding.

III. LEARNING CEPSTRAL UNCERTAINTY FROM SPECTRUM

Current methods for estimating the uncertainty involve the use of speech enhancement algorithms operating in log-Mel frequency or cepstral domains [13], [20]. However, a large class of speech enhancement algorithms use other frequency representations such as auditory frequency (e.g. [15]), discrete Fourier transform (DFT) (e.g. [6]), etc. In particular, several recent algorithms perform speech enhancement by attempting to estimate a binary mask that can be used to select speech-dominant T-F regions of a noisy speech signal [30], [42]. Specifically, the T-F units in a noisy mixture are selectively weighted (1 or 0) in order to enhance the desired signal. While the subjective intelligibility of such enhanced signals is high [10], [30], the speech features extracted for use in ASR suffer from distortion due to the mismatch arising from the noise dominant T-F units. To mitigate the effect of these distortions on recognition, these algorithms have been typically coupled to a missing-data ASR [22], [30]. The missing-data ASR treats the noise-dominant T-F units as missing or unreliable and marginalizes them during recognition. As mentioned in the introduction, this constrains the recognition to be performed in the spectral or the T-F domain. The use of cepstral transformation smears the information from the noisy T-F units across all the cepstral features, preventing its effective marginalization.

To utilize the advantage of cepstral features for recognition, it is suggested in [28] that the information in the noise-dominant T-F regions be first reconstructed using a speech prior. This allows for the subsequent use of the cepstral transformation. While promising recognition results are reported in [28], as mentioned in Section I, the ASR performance is sensitive to errors in mask estimation and reconstruction. Estimation of these errors would enable their use in the uncertainty decoder for improved recognition results. Hence, we propose a two-step method for estimating the uncertainty associated with reconstructed cepstra. In the first-step, we estimate the uncertainty associated with the reconstructed spectra by utilizing the statistical information contained in the speech prior used in reconstructing the speech information in the noise-dominant T-F units. In the second step, a non linear regression is performed to transform the estimated spectral-domain variance into the cepstral domain. We use the non-parametric method of decision trees [7] for the regression operation.

A. Estimating the Uncertainty of Reconstructed Spectra

The noisy input is first analyzed using a short T-F decomposition. The T-F resolution is 20 ms time frames with a 10 ms frame shift and 257 discrete Fourier transform coefficients. Frames are extracted by applying a running Hamming window to the signal. This signal is then processed by a speech segregation algorithm that estimates an ideal binary mask. A T-F unit in the ideal binary mask is 1 if in the corresponding T-F unit, the noisy speech contains more speech energy than interference energy; it is 0 otherwise. The ideal binary mask may be obtained *a priori* from premixing speech and noise. In practice, the ideal binary mask is not obtainable from a noisy signal, but can be estimated using speech separation algorithms. A binary T-F mask thus

estimated is used in conjunction with the spectrogram reconstruction approach to derive features for recognition.

In the spectrogram reconstruction approach, a noisy spectral vector Y at a particular frame is partitioned into reliable and unreliable constituents as Y_r and Y_u, where $Y = Y_r \cup Y_u$ [28]. The reliable features are the T-F units labeled 1 (speech-dominant) in the binary T-F mask while the unreliable features are the ones labeled 0 (noise-dominant). Assuming that the reliable features Y_r approximate well the true ones X_r, a Bayesian decision is then employed to estimate the remaining components X_u given the reliable ones and a prior speech model. As in [28], we model the speech prior as a mixture of Gaussians,

$$p(X) = \sum_{k=1}^{M} p(k)p(X|k), \tag{5}$$

where $M = 1024$ is the number of mixture components, k is the mixture index, $p(k)$ is the mixture weight, and $p(X|k) = N(X; \mu_k, \Sigma_k)$. The binary mask is also used to partition the mean and covariance of each mixture into their reliable and unreliable components as:

$$\mu_k = \begin{bmatrix} \mu_{r,k} \\ \mu_{u,k} \end{bmatrix} \ , \ \Sigma_k = \begin{bmatrix} \Sigma_{rr,k} & \Sigma_{ru,k} \\ \Sigma_{ur,k} & \Sigma_{uu,k} \end{bmatrix}. \tag{6}$$

Note that $\Sigma_{ru,k}$ and $\Sigma_{ur,k}$ denote the cross-covariance between the reliable and unreliable components.

It is shown in [11], [28] that a good estimate of X_u is its expected value conditioned on X_r:

$$E_{X_u|X_r}(X_u) = \sum_{k=1}^{M} p(k|X_r)\hat{X}_{u,k}, \tag{7}$$

where $p(k|X_r)$ is the *a posteriori* probability of the k'th mixture given the reliable data and $\hat{X}_{u,k}$ is the expected value of X_u given the k'th mixture. $p(k|X_r)$ is estimated using the Bayesian rule and the marginal distribution $p(X_r|k) = N(X_r; \mu_{r,k}, \Sigma_{rr,k})$ as:

$$x(k|X_r) = \frac{p(k)p(X_r|k)}{\sum_{k=1}^{M} p(k)p(X_r|k)}. \tag{8}$$

The expected value in the unreliable T-F units corresponding to the k'th mixture can be computed as shown in [14] as:

$$\hat{X}_{u,k} = \mu_{u,k} + \Sigma_{ur,k}\Sigma_{rr,k}^{-1}(X_r - \mu_{r,k}). \tag{9}$$

Besides estimating the speech spectral value in the unreliable T-F units, we are also interested in computing the uncertainty in our estimates. The variance associated with the reconstructed spectral vector $\hat{X}$ can also be computed in a similar fashion to the computation of the mean in (7) as:

$$\begin{aligned} \hat{\Sigma}_{\hat{X}} = \sum_{k=1}^{M} p(k|X_r)\{ & \left(\begin{bmatrix} X_r \\ \hat{X}_{u,k} \end{bmatrix} - \mu_k \right) \\ & \times \left(\begin{bmatrix} X_r \\ \hat{X}_{u,k} \end{bmatrix} - \mu_k \right)^T + \begin{bmatrix} 0 & 0 \\ 0 & \hat{\Sigma}_{u,k} \end{bmatrix} \}, \end{aligned} \tag{10}$$

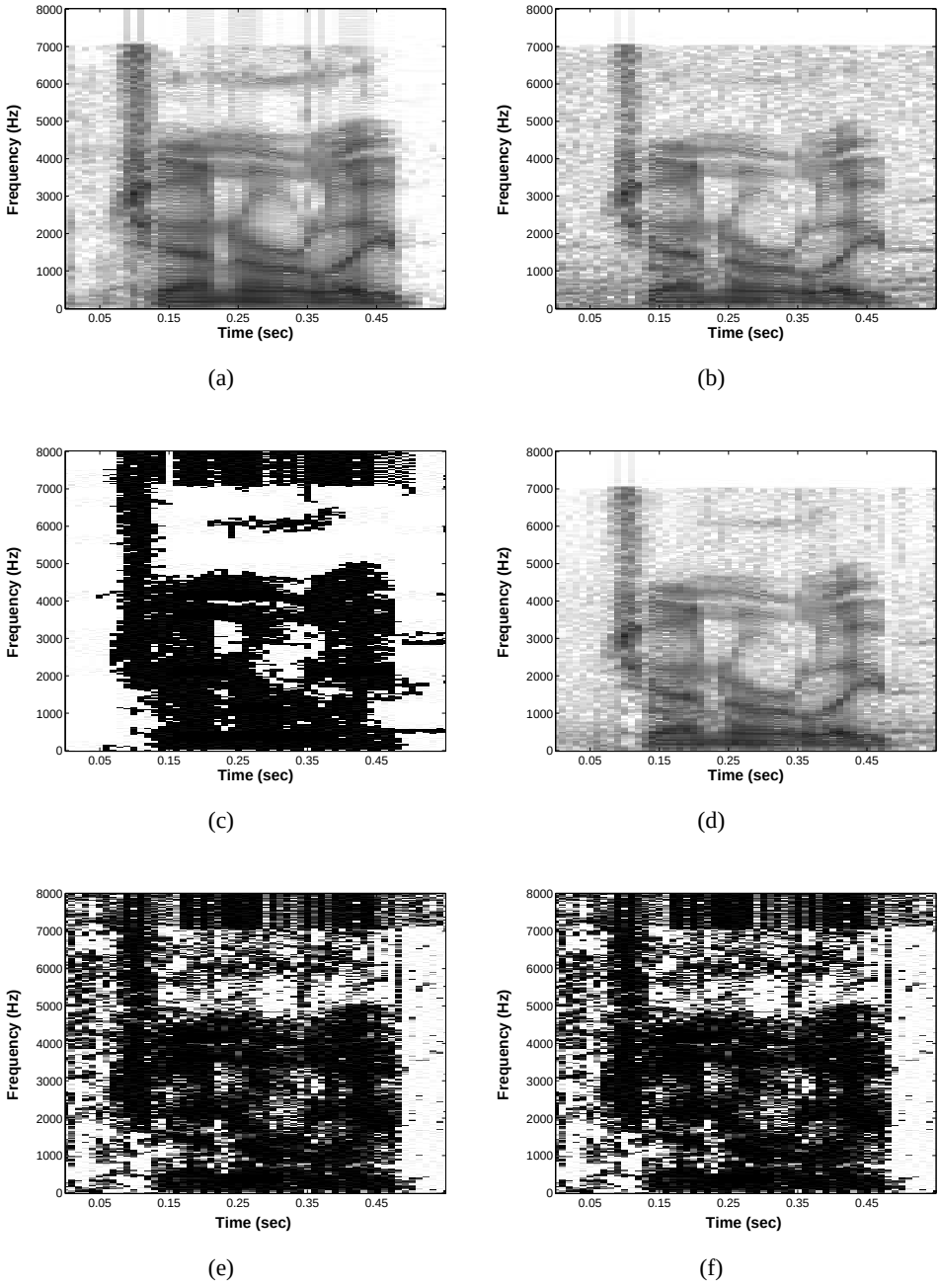


Fig. 2. Comparison between estimated and ideal uncertainties of the reconstructed spectra. (a) The spectrogram of a clean speech utterance. (b) The spectrogram of a mixture of the clean speech utterance and restaurant noise. (c) The binary T-F mask produced by spectral subtraction. Reliable T-F units are marked black and unreliable white. (d) The spectrogram obtained from (b) by applying the mask in (c) followed by reconstruction of the unreliable T-F units. (e) The true uncertainty associated with the reconstructed spectrogram in (d). (f) The corresponding estimated uncertainty.

where

$$\hat{\Sigma}_{u,k} = \Sigma_{uu,k} - \Sigma_{ur,k}\Sigma_{rr,k}^{-1}\Sigma_{ru,k}. \quad (11)$$

Fig. 2 shows the comparison between ideal and estimated uncertainties. Fig. 2(a) shows a spectrogram of a clean speech utterance from the Aurora4 database. Fig. 2(b) shows the spectrogram of a mixture of the speech and restaurant noise from the same database. A binary T-F mask, estimated using a spectral subtraction-based algorithm (see Section IV), is shown in Fig. 2(c). The reliable T-F units in this mask are black and the unreliable ones white. This mask is applied to the mixture as explained above and the results are presented in Fig. 2(d). Note that the application of the binary mask and the spectrogram reconstruction algorithm results in a significant reduction of noise in the mixture, especially in the mid- and high-frequency regions. However, the enhancement is not perfect and deviations from clean speech exist. Fig. 2(e) and Fig. 2(f) show the ideal and the estimated uncertainties associated with the reconstructed spectrogram in Fig. 2(d), respectively. The ideal uncertainty is computed as the squared difference between the spectral energies of the enhanced and the clean speech utterances. We use the diagonal components of $\hat{\Sigma}_{\hat{X}}$ as the estimate of the uncertainty associated with the reconstructed spectral vector $\hat{X}$. Observe that the estimated uncertainty is similar to the ideal uncertainty, especially in those time frames that contain voice activity. The cesptra $\hat{z}$ derived from $\hat{X}$ is used as input to the ASR in the experiments reported in Section IV. Note that no information about the noise source is used in the estimation of $\hat{\Sigma}_{\hat{X}}$.

B. Transforming Spectral Uncertainty into Cepstral Domain

In the second step, we use a set of regression trees to transform $diag\ \{\hat{\Sigma}_{\hat{X}}\}$ into $\hat{\Sigma}_{\hat{z}}$, the estimated variance associated with the reconstructed cepstra. Regression tree is a flexible and easy-to-interpret tool for non-parametric and multivariate regression analysis. It is a particularly attractive option if a parametric form of relationship between the predictor and the dependent variables is unavailable from domain knowledge. Since regression trees are fairly well documented in the literature [7], we only provide a brief overview here. A regression tree performs a histogram analysis of the regression surface. In essence, this involves the use of a binary decision tree to partition the input space using a sequence of yes/no questions that form the leaf nodes of the tree. Depending on the answers, the tree is traversed until a terminal node is reached. The terminal nodes contain the values of the dependent variable. In regression analysis, a question is chosen so that the answer partitions the predictor variable space in such a manner as to minimize the weighted sample variance of the dependent variable at that node [7].

For each frame, the input to the regression tree consists of $diag\ \{\hat{\Sigma}_{\hat{X}}\}$ corresponding to that frame. It is also found to be useful to supplement the spectral domain variance by the reconstructed cepstra in that frame and in one frame before and after. The desired output, as suggested in [13], is a diagonal matrix formed by the squared difference between the reconstructed and clean cepstra. The feature vectors used in the recognition experiments reported in Section IV consist of 12 Mel-frequency cepstral coefficients and the log frame energy along with the corresponding delta and acceleration coefficients. Hence, the uncertainties corresponding to a 39 dimensional output feature are required. We estimate the uncertainty corresponding to static, delta and acceleration coefficients independently. Note that the cepstral transform approximately orthogonalizes the spectral features [34]. While it is certainly possible to use the same transformation used in the computation of the difference coefficients to compute their

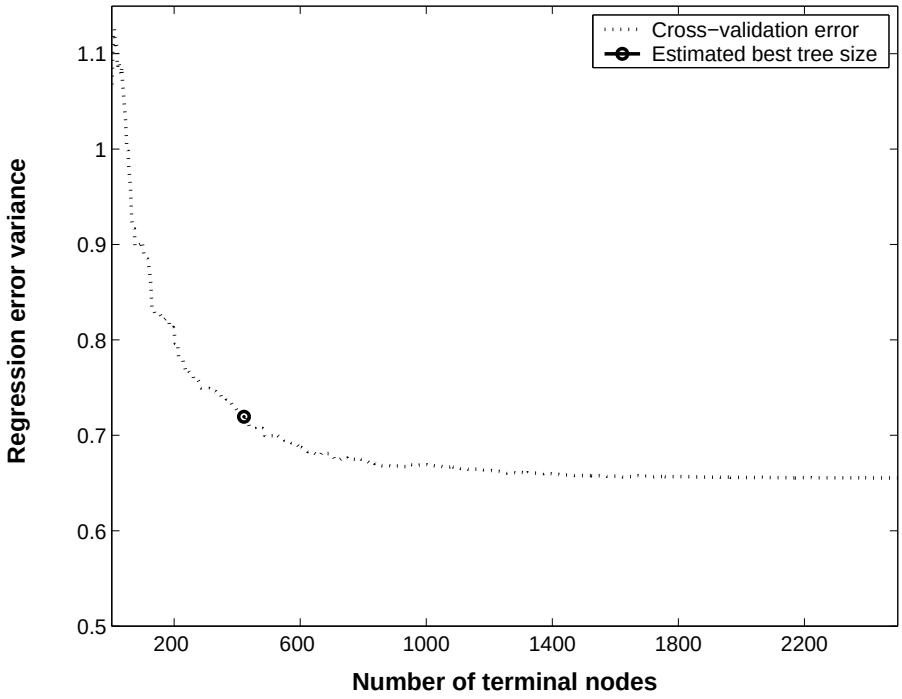


Fig. 3. Choosing the best tree size by cross-validation. The plot shows how the regression error variance changes as the number of terminal node increases on the cross-validation dataset. The plot also shows the number of terminal nodes for the best tree chosen. The best tree has an error variance one standard deviation higher than the minimum variance.

uncertainties from the static ones [13], [18], it may be advantageous to estimate them directly. It is shown in [41] that the difference features are more robust than static features. However, the computation of the uncertainties of the dynamic coefficients from the static ones using typical linear transforms would result in increased uncertainty of the static ones. Additional justification for the independence modeling assumption using difference coefficients can be found in [40]. Hence, we train a separate tree for each output dimension using the same input feature set.

Two parameters critical to the successful use of regression trees are the minimum splitting threshold and the tree size. The minimum splitting threshold refers to the minimum number of training samples in a terminal node for it to be a valid one [32]. In our experiments, we set the minimum splitting threshold to 10. To avoid over-fitting, a 10-fold cross-validation is used to find the best tree size [7], [32]. Fig. 3 shows how the regression error variance changes on the cross-validation data as the tree size increases. The data corresponds to regressing the uncertainty of the 1^{st} order cepstral coefficient. Note that the regression error variance decreases as the number of terminal nodes is increased. However, to avoid over-fitting, we choose the best tree size as the one that has an error variance one standard deviation higher than the minimum regression error variance. Fig. 3 also shows the tree size and the cross-validation error variance of the best tree so chosen.

Fig. 4(a) and Fig. 4(b) show the true and the estimated cepstral uncertainties for the same

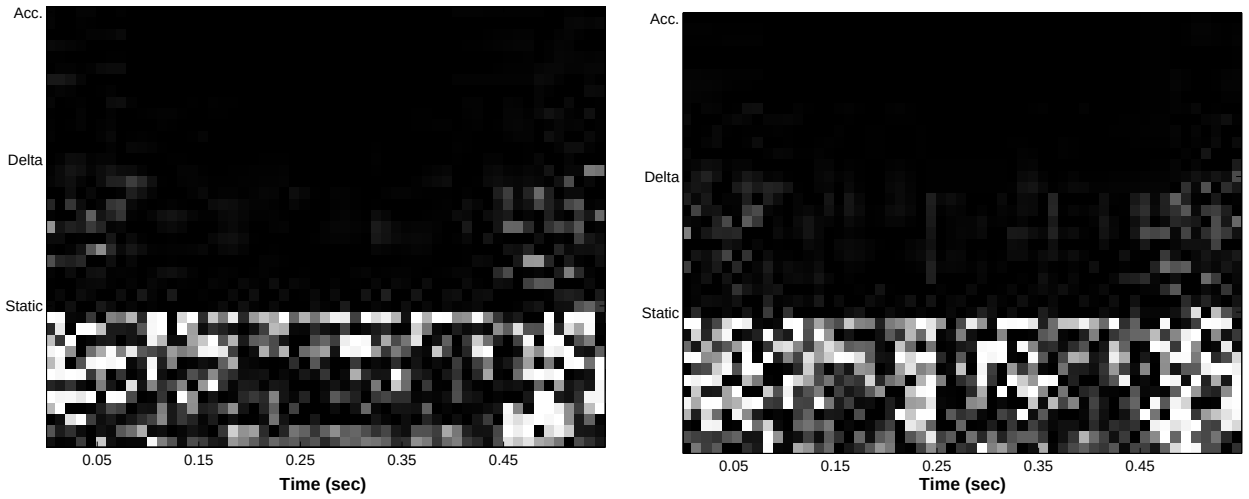


Fig. 4. Comparison between true and estimated cepstral domain uncertainties. (a) The true uncertainty associated with the static, the dynamic and the acceleration cepstral coefficients. (b) The corresponding estimated uncertainty.

noisy mixture as used in generating the results shown in Fig. 2, respectively. The brightness of a pixel is related to the degree of the uncertainty of the corresponding cepstral coefficient. Greater brightness indicates higher uncertainty. Fig. 4(a) shows the true uncertainties corresponding to the static, the delta and the acceleration (Acc.) coefficients. The figure supports the conclusion in [41] that the dynamic cepstral coefficients are more reliable than the static ones. Fig. 4(b) shows the uncertainties estimated using the set of 39 regression trees. Notice that the estimated uncertainties approximate well the true ones.

IV. EXPERIMENTAL RESULTS

We have evaluated the proposed method of uncertainty estimation in conjunction with the uncertainty decoder on the Aurora 4, 5000 word closed-vocabulary recognition task [24]. This task is based on the Wall Street Journal (WSJ0) database [25], which is created by recording speakers reading articles from the Wall Street Journal. Aurora4 consists of several test sets corresponding to different noise sources digitally added to the clean speech recordings. The signal to noise ratio (SNR) is randomly chosen from 5 dB to 15 dB, with an average SNR of 10 dB. This database also includes other test sets that incorporate microphone and sampling rate variations. As the focus of this paper is on noise robustness, we consider only a subset of the Aurora4 task. This subset corresponds to training and testing on the recordings from the Sennheiser microphone at 16 kHz and processed by a P.341 filter [24]. The use of the P.341 filter simulates the transmission characteristics for wideband telephony [17]. In particular, 7138 utterances from 83 speakers in the “training_clean_sennh” set are used for training the acoustic model and the speech prior used in reconstruction (see Section III-A). The acoustic model

TABLE I

WER (%) OF UNCERTAINTY DECODING AND RECOGNITION WITH RECONSTRUCTED CEPSTRA WHEN USING THE SPECTRAL SUBTRACTION MASK ON THE AURORA4 TASK. FOR COMPARISON, BASELINE RECOGNITION RESULTS AND RESULTS OBTAINED USING THE ETSI ADVANCED FEATURE EXTRACTION ALGORITHM ARE ALSO SHOWN.

System	*Test Set*					
	2	**3**	**4**	**5**	**6**	**7**
Baseline	57.5	55.4	55.4	63.0	54.1	65.9
Enhanced Speech	23.3	47.3	54.6	50.4	50.5	49.1
UD	22.1	43.5	51.5	49.5	47.6	46.9

consists of state-tied, cross-word triphone-based HMMs. The observation density in each state is modeled using a mixture of 4 Gaussians [23]. We use the same bigram language model and the CMU pronunciation dictionary-based lexicon [9] as used in generating the baseline results on Aurora4 [23]. Testing is performed on noisy utterances from 6 different noise sources: car, babble, restaurant, street, airport and train. These noisy utterances correspond to test sets 2-7 respectively. We use the standard "short test set definitions" consisting of 166 test utterances for each noise condition. This set gives results representative of the complete test set [24]. The number of speakers in the test set is 8. Training and testing on clean speech are performed using the toolkit and scripts developed for Aurora4 [23]. For testing on the noisy datasets, the decoder in [23] is modified to incorporate the uncertainty decoding method. The word error rate (WER) under clean speech conditions is 10.5%.

For training the regression trees (Section III-B), we use only a 40-utterance development-subset corresponding to one of the noise sources, restaurant noise. Note that for robust speech recognition, it is desirable to utilize as little *a priori* information about noise as possible. Hence, we avoid using other noise sources in training the set of regression trees. To obtain the reconstructed spectra during the regression learning, we use ideal binary T-F masks (see Section III-A). As the Aurora4 corpus does not separately provide the noise source used to construct the noisy test sets, the noise signal is estimated from the mixture and the clean speech signals by assuming that speech and noise are uncorrelated in the mixture. The noise signal is then estimated by subtracting the clean speech signal from the mixture signal. Finally, the enhanced (reconstructed) cepstra $\hat{z}$ and its associated variance $\Sigma_{\hat{z}}$, estimated using the method described in Section III, are used in (4) to perform uncertainty decoding in the following experiments.

Spectral subtraction is frequently used to generate binary T-F masks in missing data studies [3], [11]. Hence, we first report results using binary masks generated by spectral subtraction. The spectrum of noise is estimated as the average spectrum of the first and the last 50 frames of the noisy speech spectrum. The noise spectrum is then used to estimate the local SNR in each T-F unit. As in [11], a T-F unit is labeled speech-dominant in the mask if the local SNR exceeds a threshold. The choice of this threshold represents a trade-off between providing more T-F units with reliable labels to the spectrogram reconstruction algorithm (Section III-A) and preventing wrong labeling of T-F units [29]. The optimal value is also dependent on the SNR [29], [33]. For simplicity we set this threshold to a constant. The value of 5 dB is found to give the best

recognition performance on the development set and is used for all the test sets. Additionally, as suggested in [28] the estimated noise spectrum is used to "clean" the reliable T-F units by subtracting the noise energy from the mixture energy.

Table I summarizes the performance of the uncertainty decoder ("UD") on the reconstructed cepstra by utilizing the estimated uncertainty. Performance is measured in terms of percentage WER. For comparison, we also show the performance of the conventional decoder on the reconstructed cepstra ("Enhanced Speech"). Additionally, the baseline performance of the conventional decoder on the noisy data is also shown ("Baseline"). As can be seen from Table I, across all noise conditions, the performance of the uncertainty decoder using the estimated uncertainty shows significant improvement over that of the conventional ASR on the reconstructed cepstra. The average reduction in error rate is 5.2%. Moreover, large improvement over the baseline performance is obtained, with an average error rate reduction of 27.34%. Notice that the system is able to generalize well across noise conditions not seen during the regression tree training.

We now present results using the masks generated by a computational auditory scene analysis (CASA) system [15]. This system is a voiced speech separation system based on two main stages: segmentation and grouping. In segmentation, the input signal is decomposed into a collection of contiguous T-F units that are dominated by one sound source. During grouping, those segments that likely belong to the same source are grouped together based on common periodicity. In the low-frequency range, the system generates segments based on temporal continuity and cross-channel correlation, and groups them based on periodicity similarity. For high-frequencies, the signal envelope fluctuates at the pitch rate and amplitude modulation rates are used for grouping [15]. Provided that the speech pitch contour can be estimated, this segregation mechanism produces a binary mask that labels those T-F units where speech dominates the interference. The CASA system shows a robust performance when tested with a variety of noise intrusions. For input to the system in [15], a pitch estimate is derived from the noisy speech signal using Praat [5]. The system in [15] uses an auditory filterbank decomposition of the input signal. For consistency with the DFT decomposition used in our spectrogram reconstruction, this mask is mapped into the DFT domain prior to reconstruction by labeling the corresponding DFT bins. Note that the system in [15] segregates only voiced speech. Hence, if a valid pitch is not detected in a particular frame, we use the mask obtained by spectral subtraction in those frames. Table II shows the performance of the uncertainty decoder when using the combined mask from [15] and spectral subtraction. As before, across all SNR conditions, significant improvement over the performance of the conventional ASR on the enhanced speech is obtained when using the estimated variance. The average reduction in WER is 7.6%. Note that under non-stationary noise conditions (e.g. restaurant, test set 4), the performances of both the conventional ASR and uncertainty decoder using the combined mask are significantly better than their performance when using the spectral subtraction mask alone. However, Mask estimation based on spectral subtraction appears to be better for the stationary noise conditions in the present study. This is due to the inability of the voice separation system to recover the inharmonic components of speech in the voiced frames. On the other hand, under non-stationary noise conditions, this drawback is more than offset by improved segregation of harmonic components.

To show the ceiling performance of the proposed method, we also report results obtained using the ideal binary T-F masks. These masks are generated in a similar fashion to those used in our regression tree training. For comparison, recognition results using the ideal uncertainties ("Ideal UD") are also shown. Ideal uncertainty is computed as the squared difference between

TABLE II

WER (%) OF UNCERTAINTY DECODING AND RECOGNITION WITH RECONSTRUCTED CEPSTRA WHEN USING THE COMBINED VOICE SEPARATION AND SPECTRAL SUBTRACTION MASK.

System	*Test Set*					
	2	**3**	**4**	**5**	**6**	**7**
Enhanced Speech	31.1	45.5	50.4	51.6	53.2	53.2
UD	27.5	42	46.9	51.5	47.1	49.2

TABLE III

WER (%) FROM UNCERTAINTY DECODING WITH ESTIMATED AND IDEAL VARIANCE AND RECOGNITION WITH RECONSTRUCTED CEPSTRA WHEN USING THE IDEAL BINARY MASK.

System	*Test Set*					
	2	**3**	**4**	**5**	**6**	**7**
Enhanced Speech	14.7	22	25.2	29	19.6	26
Estimated UD	14	20	22	24.9	17.5	25.7
Ideal UD	14	20.1	22.2	25.1	16.8	24.9

the reconstructed and clean cepstra as suggested in [13]. Table III shows that the performance of the uncertainty decoder using the estimated uncertainty ("Estimated UD") is close to its performance using the ideal uncertainty. This indicates that the proposed approach estimates the uncertainty associated with the reconstructed cepstra accurately. Notice that even with the use of ideal binary masks, the uncertainly decoder can still improve recognition results compared to the conventional ASR; the average reduction in error rate is 8.75%. Note that for test sets 3-5, the performance of uncertainty decoder using the ideal uncertainties is slightly worse compared to its performance using the estimated ones. However, the performance difference is statistically insignificant.

It can also be seen from Table III that use of the ideal binary mask results in an excellent performance for both the conventional ASR and the uncertainty decoder. This supports that use of the ideal binary mask as a computational goal for speech separation systems (see also [39]).

A. Comparison of Regression Trees and Multilayer Perceptrons in Learning the Cepstral Domain Uncertainty

In an earlier study [36], we used a multilayer perceptron (MLP) to transform the spectral domain uncertainty into the variance associated with the reconstructed cepstra. Since MLP is well known as a universal function approximator [27], it can also be used for learning this transformation. Specifically, we trained a one-hidden-layer (374-800-39) MLP [27]. The input and the output features are the same as those described in Section III-B. The transfer functions of the hidden and output layer neurons are tangent hyperbolic sigmoid and linear respectively. The

TABLE IV
COMPARISON OF WER (%) USING TWO DIFFERENT TRANFORMATION METHODS.

Transformation Method	*Test Set*					
	2	**3**	**4**	**5**	**6**	**7**
Regression Tree	22.1	43.5	51.5	49.5	47.6	46.9
MLP	22.7	45.5	51.5	51.9	47.6	49.4

MLP is trained using backpropagation, optimized by the scaled conjugate gradient method [27]. The network is trained for 100 epochs and a 10-fold cross-validation is used to avoid over-fitting.

Table IV compares the WER of the uncertainty decoder using regression trees and MLP to transform the uncertainty from the spectral domain to the cepstral domain. For both methods, the enhanced speech was produced using the binary masks generated by spectral subtraction. From Table IV, we can see that the performance of the two methods is similar. Hence, both methods are suitable for learning the uncertainty transformation. For this task, the regression tree is slightly better perhaps due to its non-parametric property which enables it to make minimal assumptions about the nature of the regression surface.

B. Comparison of Uncertainty Decoding with Missing-data Recognition.

As mentioned in Sections I and III, for robust speech recognition, speech segregation systems that compute a binary T-F mask have been coupled to a missing-data ASR. While previous studies have shown that the performance of the missing-data ASR degrades as the vocabulary size increases [28], [35], here we investigate whether uncertainty decoding can be a valid alternative to missing-data recognition even on a small vocabulary task. The specific missing-data method used is the bounded marginalization algorithm which is known to provide the best recognition results on small vocabulary tasks [11], [28]. In the marginalization method, the posterior probability using only the reliable constituents is computed by integrating over the unreliable ones [11]. Feature vectors for the missing-data ASR are the spectral energies extracted as described in Section III-A. The bounded marginalization method uses the knowledge that the true value of the spectral energy in the unreliable parts lies between 0 and the observed spectral energy. These bounds are used as limits on the integral involved in marginalizing the posterior probability over the unreliable features.

We evaluate the two recognition approaches on a speaker independent connected digit recognition task. The grammar for this task allows for the repetition of one or more digits. This is the same task used in the original study in [11]. Thirteen (1-9, a silence, very short pause between words, zero and oh) word-level models are trained for both recognizers. All except the short pause model have 8 emitting states. The short pause model has a single emitting state, tied to state 4 of the silence model. The output distribution in each state is modeled as a mixture of 10 Gaussians, as suggested in [11]. The TIDigits database's male speaker dataset is used for both training and testing [19]. Specifically, the models are trained using 4235 utterances in the training set of this database. Testing is performed on a subset of the testing set consisting of 461 utterances from 6 speakers. All test speakers are different from the speakers in the training

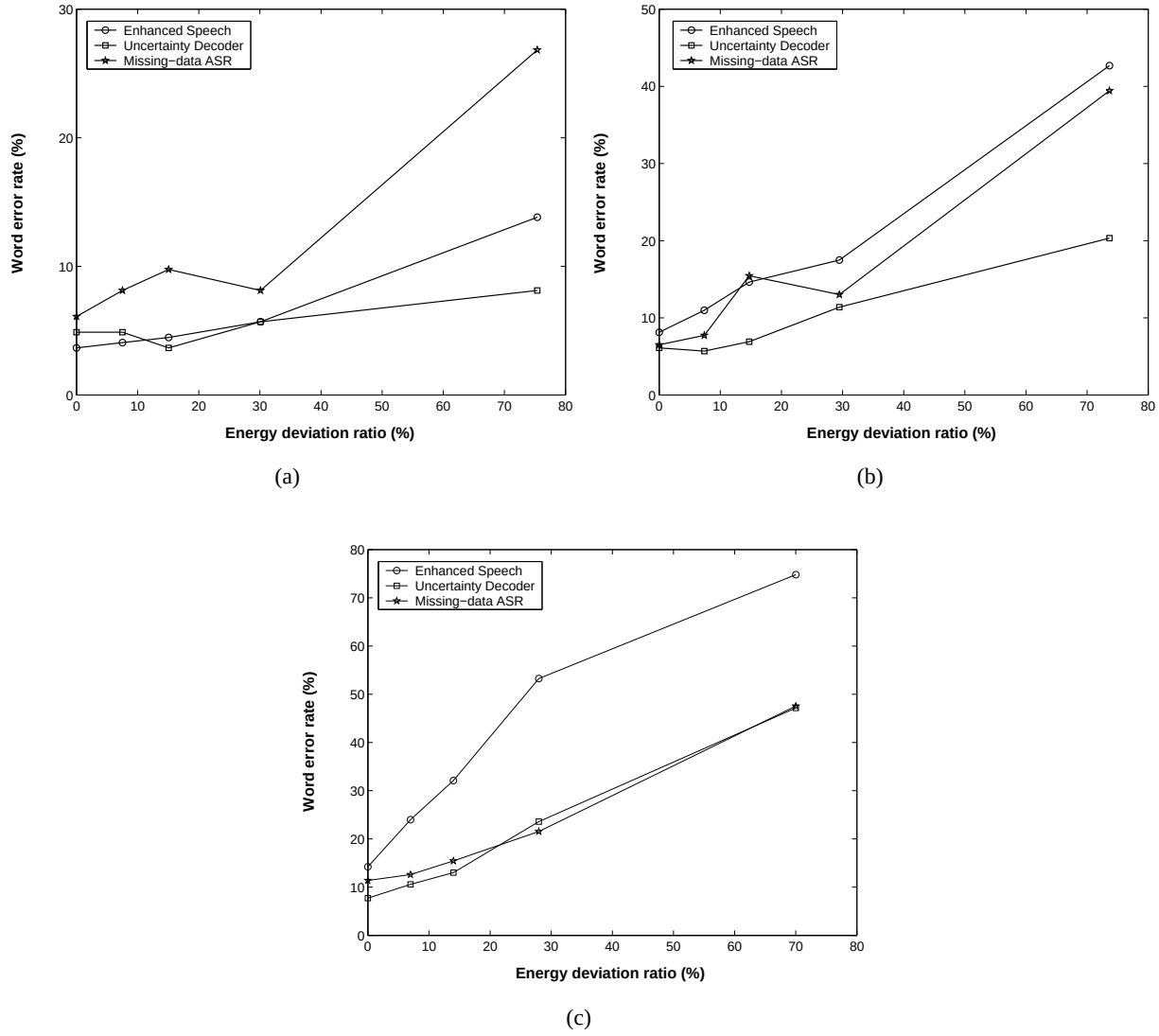


Fig. 5. Comparison of conventional recognition of enhanced speech, uncertainty decoding and missing-data recognition. The figures show the WER with respect to deviations from the ideal binary mask. (a) WER at 10 dB SNR. (b) WER at 5 dB SNR. (c) WER at 0 dB SNR.

set. The signals in this database are sampled at 20 kHz. The noise source is factory noise from the NOISEX corpus [38], which is also used in [3], [11]. Factory noise is chosen as it has energy in formant regions, therefore posing challenging problems for recognition. An HMM toolkit, HTK [43], is used for training. During testing, the decoder is modified to incorporate the uncertainty decoding and the bounded marginalization methods.

For both methods, the binary T-F mask used is the ideal binary mask. As pointed out earlier, this mask needs to be estimated in practice. Hence, we investigate how robust the two recognizers

are to deviations from the ideal binary mask. Specifically, we randomly flip a certain number of 1's and 0's in the mask. The percentage deviation is measured using the fraction of 1's flipped, which takes on the values of 0%, 5%, 10%, 20% and 50%. Since the spectral energy in a T-F unit has a large dynamic range, we additionally calculate the energy deviation as the ratio of the total energy corresponding to the flipped bits to the total energy corresponding to T-F units labeled 1 in the ideal binary mask. The resulting masks are used directly by the missing-data ASR. For use by the uncertainty decoder we reconstruct the speech spectral energy in the missing T-F units and derive cepstral features along with the associated uncertainties as described in Section III. Fig. 5(a-c) summarizes the performance of the two recognizers at SNR values of 10 dB, 5 dB and 0 dB respectively. Additionally the performance of the conventional ASR on the reconstructed cepstra is also shown. Performance is given in terms of WER across various energy deviation ratios. To better illustrate the differences between the two recognition methods, the error rates in Fig. 5 are plotted to different scales for the three different SNRs. Fig. 5(a) shows that at 10 dB SNR, both the conventional ASR and the uncertainty decoder outperform the missing-data recognizer. The uncertainty decoder also outperforms the missing-data ASR at 5 dB SNR as shown in Fig. 5b. Fig. 5(c) shows that the performance of the uncertainty decoder and the missing-data recognizer are comparable at the 0 dB SNR condition. Hence, the proposed uncertainty decoding approach gives a strong alternative to the missing-data approach for robust speech recognition using binary T-F masks. Note that across all SNRs the uncertainty decoding outperforms the coventional recognition of the reconstruted cepstra.

V. CONCLUDING REMARKS

We have proposed a general solution to the problem of estimating the uncertainty of cepstral features derived from the output of front-end preprocessing algorithms that use a binary T-F mask for speech enhancement. Using the uncertainty decoding approach in [13] on the Aurora4 task, we have shown that the estimated uncertainty yields significant reductions in WER compared to conventional recognition on the enhanced cepstra. We have also obtained substantial improvements over the baseline ASR performance. Furthermore, our experiments on the digit recognition task suggest that the proposed method provides a valid alternative to the missing-data approach for robust speech recognition.

The principal advantage of the proposed method is that it neither requires that noise conditions be known *a priori* nor assumes a noise model. Our training of regression trees requires a limited amount of aligned clean and noisy speech data, corresponding to one of the noise sources used in the evaluation. However, as seen in Section IV, the system is able to generalize across noise sources not seen during training. We wish to emphasize that the exact choice of the noise source used in learning the uncertainties is not crucial for the performance. In an earlier study [36], for example, we used a different noise source, street noise, but the resulting performance was very similar. Hence, the proposed method can be used in conjunction with CASA systems that do not require noise conditions known *a priori* for robust speech recognition.

An alternative approach for estimating the uncertainties associated with the reconstructed cepstra is given in [18]. The variance of the static coefficients is approximated using the unscented transform. The variance of the dynamic coefficients is estimated using the same linear transformation employed in obtaining the dynamic features. As described in Section III-B, this approach is not optimal. A key advantage of the proposed method is the direct estimation

of uncertainties corresponding to the static, the delta and the acceleration coefficients. This enables us to exploit the differences in the *a priori* accuracies of the static and the dynamic coefficients [41].

For learning the transformation of spectral domain uncertainties to cepstral ones, we used the ideal binary T-F mask. This transformation was then applied to the masks generated using spectral subtraction and a CASA system. Although the resulting uncertainty estimates provide promising results, additional improvements may be obtained by training the regression trees directly on the output of particular speech enhancement algorithms. Future work will address this issue.

ACKNOWLEDGMENT

This research was supported in part by an AFOSR grant (FA9550-04-1-0117), an AFRL grant via Veridian and an NSF grant (IIS-0534707). We thank A. Acero and M. L. Seltzer for helpful suggestions. A preliminary version of this work was presented in 2006 ICASSP.

REFERENCES

[1] S. Araki, S. Makino, H. Sawada, and R. Mukai, "Underdetermined blind separation of convolutive mixtures of speech with directivity pattern based mask and ICA," in *Proc. Fifth International Conference on Independent Component Analysis '04*, 2004, pp. 898–905.

[2] J. A. Arrowood and M. A. Clements, "Using observation uncertainty in HMM decoding," in *Proc. International Conference on Spoken Language Processing '02*, 2002, pp. 1561–1564.

[3] J. P. Barker, M. P. Cooke, and D. P. W. Ellis, "Decoding speech in the presence of other sources," *Speech Communication*, vol. 45, pp. 5–25, 2005.

[4] M. C. Benitez, J. C. Segura, A. D. Torre, J. Ramirez, and A. Rubio, "Including uncertainty of speech observations in robust speech recognition," in *Proc. International Conference on Spoken Language Processing '04*, 2004, pp. 137–140.

[5] P. Boersma and D. Weenink, "Praat: doing Phonetics by Computer, Version 4.0.26," 2002. [Online]. Available: http://www.fon.hum.uva.nl/praat

[6] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. ASSP-27, no. 2, pp. 113–120, 1979.

[7] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and regression trees*. New York, NJ: Chapman & Hall, 1984.

[8] C. Breithaupt and R. Martin, "Statistical analysis and performance of DFT domain noise reduction filters for robust speech recognition," in *Proc. Interspeech '06*, 2006, to appear.

[9] Carnegie Mellon University. The CMU Pronouncing Dictionary. [Online]. Available: http://www.speech.cs.cmu.edu/cgi-bin/cmudict

[10] P. S. Chang, "Exploration of behavioral, physiological, and computational approaches to auditory scene analysis," Master's thesis, Department of Computer Science & Engineering, The Ohio State University, 2004. [Online]. Available: http://www.cse.ohio-state.edu/pnl/theses/Chang_MSThesis04.pdf

[11] M. Cooke, P. Green, L. Josifovski, and A. Vizinho, "Robust automatic speech recognition with missing and unreliable acoustic data," *Speech Communication*, vol. 34, pp. 267–285, 2001.

[12] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. ASSP-28, no. 4, pp. 357–366, 1980.

[13] 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 Trans. on Speech, and Audio Processing*, vol. 13, pp. 412–421, 2005.

[14] Z. Ghahramani and M. I. Jordan, "Supervised learning from incomplete data via an EM approach," in *Advances in Neural Information Processing Systems 6*, J. D. Cowan, G. Tesauro, and J. Alspector, Eds. San Francisco, CA: Morgan Kaufmann Publishers, 1993, pp. 120–127.

[15] G. Hu and D. L. Wang, "Monaural speech segregation based on pitch tracking and amplitude modulation," *IEEE Trans. on Neural Networks*, vol. 15, pp. 1135–1150, 2004.

[16] X. Huang, A. Acero, and H. Hon, *Spoken Language Processing*. Upper Saddle River, NJ: Prentice Hall PTR, 2001.

[17] ITU-T, "Recommendation P.341," in *Transmission characteristics for wideband (150-7000 Hz) digital hands-free telephony terminals*. The International Telecommunication Union, 2005.

[18] 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 Applications of Signal Processing to Audio and Acoustics '05*, 2005, pp. 82–85.

[19] R. G. Leonard, "A database for speaker-independent digit recognition," in *Proc. ICASSP '84*, 1984, pp. 111–114.

[20] H. Liao and M. J. F. Gales, "Joint uncertainty decoding for noise robust speech recognition," in *Proc. Interspeech '05*, 2005, pp. 3129–3132.

[21] R. Martin, "Statistical methods for the enhancement of noisy speech," in *Speech Enhancement*, J. Benesty, S. Makino, and J. Chen, Eds. NY: Springer, 2005, ch. 3, pp. 43–65.

[22] K. J. Palomaki, G. J. Brown, and D. L. Wang, "A binaural processor for missing data speech recognition in the presence of noise and small-room reverberation," *Speech Communication*, vol. 43, pp. 361–378, 2004.

[23] N. Parihar and J. Picone, "DSR front end LVCSR evaluation," in *Aurora Working Group*. European Telecommunications Standards Institute, 2002.

[24] ——, "Analysis of the aurora large vocabulary evalutions," in *Proc. Eurospeech '03*, 2003, pp. 337–340.

[25] D. Paul and J. Baker, "The design of wall street journal-based CSR corpus," in *Proc. International Conference on Spoken Language Processing '92*, 1992, pp. 899–902.

[26] M. S. Pedersen, D. L. Wang, J. Larsen, and U. Kjems, "Separating underdetermined convolutive speech mixtures," in *Proc. 6th International Conference on Independent Component Analysis and Blind Source Separation '06*, 2006, pp. 674–681.

[27] J. C. Principe, N. R. Euliano, and W. C. Lefebvre, *Neural and adaptive systems*. New York, NY: John Wiley and Sons, Inc., 2000.

[28] B. Raj, M. L. Seltzer, and R. M. Stern, "Reconstruction of missing features for robust speech recognition," *Speech Communication*, vol. 43, pp. 275–296, 2004.

[29] P. Renevey and A. Drygajlo, "Detection of reliable features for speech recognition in noisy conditions using a statistical criterion," in *Proc. Consistent & Reliable Acoustic Cues for Sound Analysis Workshop '01*, 2001, pp. 71–74.

[30] N. Roman, D. L. Wang, and G. J. Brown, "Speech segregation based on sound localization," *J. Acoust. Soc. Am.*, vol. 114, pp. 2236–2252, 2003.

[31] S. T. Roweis, "Factorial models and refiltering for speech separation and denoising," in *Proc. Eurospeech '03*, 2003, pp. 1009–1012.

[32] S. J. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 2nd ed. Upper Saddle River, NJ: Prentice Hall, 2003.

[33] M. L. Seltzer, J. Droppo, and A. Acero, "A harmonic-model-based front end for robust speech recognition," in *Proc. Eurospeech '03*, 2003, pp. 1277–1280.

[34] M. L. Shire, "Discriminant training of front-end and acoustic modeling stages to heterogeneous acoustic environments for multi-stream automatic speech recognition," Ph.D. dissertation, University of California, Berkely, 2000.

[35] S. Srinivasan, N. Roman, and D. L. Wang, "On binary and ratio time-frequency masks for robust speech recognition," in *Proc. International Conference on Spoken Language Processing '04*, 2004, pp. 2541–2544.

[36] S. Srinivasan and D. L. Wang, "A supervised learning approach to uncertainty decoding for robust speech recognition," in *Proc. ICASSP '06*, vol. I, 2006, pp. 297–300.

[37] V. Stouten, H. V. Hamme, and P. Wambacq, "Model-based feature enhancement with uncertainty decoding for noise robust ASR," *Speech Communication*, 2006, in press.

[38] A. P. Varga, H. J. M. Steeneken, M. Tomlinson, and D. Jones, “The NOISEX-92 study on the effect of additive noise on automatic speech recogonition,” Speech Research Unit, Defense Research Agency, Malvern, UK,” Technical Report, 1992.

[39] D. L. Wang, “On ideal binary mask as the computational goal of auditory scene analysis,” in *Speech separation by humans and machines*, P. Divenyi, Ed. Norwell, MA: Kluwer Academic, 2005, pp. 181–197.

[40] C. K. I. Williams, “How to pretend that correlated variables are independent by using difference observations,” *Neural Computation*, vol. 17, pp. 1–6, 2005.

[41] C. Yang, F. K. Soong, and T. Lee, “Static and dynamic spectral features: Their noise robustness and optimal weights,” in *Proc. ICASSP ’05*, vol. I, 2005, pp. 241–244.

[42] O. Yilmaz and S. Rickard, “Blind separation of speech mixtures via time-frequency masking,” *IEEE Trans. on Signal Processing*, vol. 52, pp. 1830–1847, 2004.

[43] S. Young, D. Kershaw, J. Odell, V. Valtchev, and P. Woodland, *The HTK Book (for HTK Version 3.0)*. Microsoft Corporation, 2000.