A SUPERVISED LEARNING APPROACH TO UNCERTAINTY DECODING FOR ROBUST SPEECH RECOGNITION

Soundararajan Srinivasan

Biomedical Engineering Center
The Ohio State University
Columbus, OH 43210, USA
srinivasan.36@osu.edu

DeLiang Wang

Department of Computer Science &
Engineering and Center for Cognitive Science
The Ohio State University
Columbus, OH 43210, 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 leads to a smearing of this uncertainty. We propose a supervised approach 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.

1. INTRODUCTION

The performance of automatic speech recognizers (ASRs) degrade rapidly in the presence of noise and other distortions [1]. To mitigate the effect of noise on recognition, noisy speech is typically preprocessed by speech enhancement algorithms, such as spectral subtraction (e.g. [2]). However, the accuracy of these algorithms often varies widely across time-frames. It is shown in [3] that the uncertainty resulting from front-end preprocessing can be effectively exploited to improve the recognition results.

Currently the uncertainty associated with enhanced speech features is estimated in either the log Mel-frequency domain or directly in the cepstral domain [3, 4]. 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 [5, 6, 7]. Although signals reconstructed from such masks have been shown to be highly intelligible [5], conventional ASR systems are extremely sensitive to the distortions produced during resynthesis. To minimize the effect of distortions on recognition, these speech enhancement systems have been coupled with a missing-data recognizer [8, 5, 7]. 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 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 [8], a significant disadvantage of the missing data recognizer is that recognition is performed in the spectral or 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 [9]. 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 [10]. This reconstruction is typically based on a trained speech prior.

Although the spectrogram reconstruction method in [10] provides promising results, errors in reconstruction degrade the performance of the ASR. In this paper, we present a two-step supervised learning approach to estimate the uncertainty associated from the reconstructed spectra. 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 multilayer perceptron (MLP). We thus convert the binary uncertainty encoded by the T-F mask into a real-valued uncertainty associated with the reconstructed cepstra. The estimated cepstral-domain uncertainty is utilized by an uncertainty decoder during recognition.

The rest of the paper is organized as follows. The next section briefly reviews the uncertainty decoding framework for robust speech recognition. Section 3 contains a detailed presentation of the proposed method for estimating the uncertainty associated with the reconstructed cepstra. The proposed system has been systematically evaluated on a subset of the Aurora4 noisy speech recognition task and the evaluation results are presented in Section 4. Finally, conclusions and future work are given in Section 5.

2. UNCERTAINTY DECODING

A typical approach for robust speech recognition involves preprocessing the noisy speech signal by speech enhancement algorithms. As discussed in the introduction, the performance of such front-end denoising algorithms is often inconsistent. The uncertainty decoding method accounts for the imperfections in speech enhancement by integrating the observation probability over all possible speech feature values [3]. Hence, the new observation

likelihood is computed as

$$\int_{-\infty}^{\infty} p(z|k,q)p(z|\theta)dz, \quad (1)$$

wheres z is the clean speech feature seen during training and θ denotes the parameter vector characterizing the front-end compensation model. It is suggested in [3] that $p(z|\theta)$ be modeled as $N(z;\hat{z},\Sigma_{\hat{z}})$. The observation density of each state in a HMM-based ASR is usually modeled as a mixture of gaussians. Therefore, $p(z|k,q) = N(z;\mu_{k,q},\Sigma_{k,q})$ is the likelihood of observing z given state q and mixture k. The enhanced speech value is denoted as $\hat{z}$ and the uncertainty due to the enhancement algorithm in given by the variance term $\Sigma_{\hat{z}}$. Under these conditions, it is shown in [3] that the new observation likelihood can be computed as

$$\int_{-\infty}^{\infty} p(z|k,q)p(z|\theta)dz = N(\hat{z};\mu_{k,q},\Sigma_{k,q}+\Sigma_{\hat{z}}). \quad (2)$$

The role of uncertainty associated with the enhanced features can be seen in equation 2 as increasing the variance of the gaussian mixture component. Hence, those enhanced speech features that deviate more from clean ones will contribute less to the overall likelihood. It is shown in [3] that the utilization of this speech feature uncertainty contributes to a significant improvement in the ASR accuracy on a small vocabulary task.

3. 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 [3, 4]. However, a large class of speech enhancement algorithms use various other frequency representations such as auditory frequency (e.g. [11]), discrete Fourier transform (DFT) (e.g. [2]) etc. In particular, several recent algorithms perform speech enhancement by attempting to estimate a binary mask that can be used to select the speech-dominant T-F regions of a noisy speech signal [5, 6]. Specifically, the T-F units in the noisy mixture are selectively weighted (1 or 0) in order to enhance the desired signal. To mitigate the effect of distortions arising from the noise-dominant T-F units on recognition, these algorithms have been typically coupled to a missing-data ASR that treats the these T-F units as missing or unreliable during recognition [5, 7]. As mentioned in the introduction, this constrains the recognition to be performed in the spectral or T-F domain. To utilize the superiority of cepstral features for recognition, it is suggested in [10] that 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 [10], errors in reconstruction contribute to a degradation in ASR performance. Estimation of the reconstruction errors would enable their use in the uncertainty decoder to further improve the 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 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. Since MLP is well known as a universal function approximator [12], we use it for the regression operation.

3.1. Estimating the Uncertainty of Reconstructed Spectra

The noisy input is first decomposed into 256 DFT coefficients every frame. Each frame is 25ms long with 15 ms frame shift. Frames are extracted by applying a running Hamming window to the signal. In the spectrogram reconstruction approach, a noisy spectral vector Y at a particular frame is partitioned into its reliable and unreliable constituents as Y_r and Y_u [10]. The reliable features are the T-F units labeled speech-dominated in the binary T-F mask (produced by a speech enhancement algorithm) while the unreliable features are the ones labeled 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 [10], we model the speech prior as a mixture of gaussians,

$$p(X) = \sum_{k=1}^{M} p(k)p(X|k), \quad (3)$$

where $M = 1024$ is the number of mixtures, k is the mixture index, $p(k)$ is the mixture weight and $p(X|k) = N(X;\mu_k,\Sigma_k)$. The mean and covariance of each mixture can also be partitioned into their reliable and unreliable components as

$$\mu_k = \begin{bmatrix}\mu_{r,k}\\ \mu_{u,k}\end{bmatrix},\ \Sigma_k = \begin{bmatrix}\Sigma_{rr,k} & \Sigma_{ru,k}\\ \Sigma_{ur,k} & \Sigma_{uu,k}\end{bmatrix}. \quad (4)$$

Note that $\Sigma_{ru,k}$ and $\Sigma_{ur,k}$ denote the cross-covariance between the reliable and unreliable components. It is shown in [8, 10] that a good estimate of X_u is its expected value conditioned on X_r

$$E_{X_u|X_r}(X_u) = \sum_{k=1}^{M} p(k|X_r)\hat{X}_{u,k}, \quad (5)$$

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 from the marginal distribution $p(X_r|k) = N(X_r;\mu_{r,k},\Sigma_{rr,k})$. The conditional mean corresponding to the k'th mixture is then given by

$$\hat{X}_{u,k} = \mu_{u,k} + \Sigma_{ur,k}\Sigma_{rr,k}^{-1}(X_r - \mu_{r,k}). \quad (6)$$

The variance associated with the reconstructed spectral vector $\hat{X}$ can also be computed as

$$\hat{\Sigma} = \sum_{k=1}^{M} p(k|X_r)\{\left(\begin{bmatrix}X_r\\ \hat{X}_{u,k}\end{bmatrix} - \mu_k\right) \times \left(\begin{bmatrix}X_r\\ \hat{X}_{u,k}\end{bmatrix} - \mu_k\right)^T + \begin{bmatrix}0 & 0\\ 0 & \hat{\Sigma}_{u,k}\end{bmatrix}\}, \quad (7)$$

as shown in [13], where

$$\hat{\Sigma}_{u,k} = \Sigma_{uu,k} - \Sigma_{ur,k}\Sigma_{rr,k}^{-1}\Sigma_{ru,k}. \quad (8)$$

We use $\hat{\Sigma}$ as the estimate of the uncertainty associated with$\hat{X}$. The cesptra $\hat{z}$ derived from $\hat{X}$ is used as input to the ASR in the experiments reported in Section 4. Note that no information about the noise source is used in the estimation of $\hat{\Sigma}$.

3.2. Transforming Spectral Uncertainty into Cepstral Domain

In the second step, we use a MLP to transform $\hat{\Sigma}$ into $\Sigma_{\hat{z}}$, the variance associated with the reconstructed cepstra. For each frame, the input to the perceptron consists of $\hat{\Sigma}$ corresponding to that frame supplemented by the reconstructed cepstra in that frame and in one frame before and after. The desired MLP output is set to be the squared difference between the reconstructed and clean cepstra. We train a one-hidden-layer (373-800-39) MLP [12]. The number of neurons in the hidden layer is varied from 200 to 2000 during an initial training phase. The MLP with 800 neurons performed as well as any of the larger ones and hence is used in the experiments reported in Section 4. The feature vectors used in the recognition experiments reported below consist of 12 Mel-frequency cepstral coefficients and the log frame energy along with the corresponding delta and acceleration coefficents. Hence, the output layer has 39 neurons. Note that we jointly estimate the uncertainty corresponding to static, delta and acceralation coefficients. The transfer function of the hidden and output layers neurons are hyperbolic tangent sigmoid and linear respectively. The MLP is trained using backpropagation, optimized by the scaled conjugate gradient method [12]. The network is trained for 100 epochs and a 10-fold cross-validataion was used to avoid over-fitting.

4. EXPERIMENTAL RESULTS

We have evaluated the proposed method of uncertainty estimation in the conjunction with the uncertainty decoder on the Aurora 4, 5000 word closed-vocabulary recognition task [14]. Aurora4 consists of several test sets corresponding to different noise sources digitally added to the clean speech recordings at a randomly chosen signal to noise ratio (SNR) from 5 dB to 15 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 that corresponds to training and testing on the Sennheiser microphone at 16 kHz and processed by a P.341 filter [14]. In particular, 7138 utterances from the "training_clean_sennh" set are used in training the cross-word triphone-based acoustic models with 4 gaussians per state [15] and the speech prior used in reconstruction (see Section 3.1). We use the same bigram language model and the lexicon used in generating the baseline results on Aurora4 [15]. 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 [14]. Training and testing are performed using the toolkit and scripts developed for Aurora [15]. The recognition accuracy on clean speech is 85.5%. For training the MLP (Section 3.2), we use only a 40 utterance development-subset corresponding to one of the noise sources, street 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 MLP. To obtain the reconstructed spectra during the MLP training, we use *a priori* binary T-F masks that retain those T-F units of the noisy speech signal whose energy is within 3 dB of the corresponding clean speech energy as suggested in [8]. Finally, the enhanced (reconstructed) cepstra $\hat{z}$ and its associated variance $\Sigma_{\hat{z}}$, estimated using the method described in Section 3, are used in equation 2 to perform uncertainty decoding in the following experiments.

Table 1. WER (%) of uncertainty decoding and recognition with reconstructed cepstra when using the spectral subtraction mask on the Aurora4 task. For comparison, baseline recognition results are also shown.

System	*Test Set*					
	2	**3**	**4**	**5**	**6**	**7**
Baseline	58.4	58.9	53.8	62.4	56.9	65.7
Enhanced Speech	39.4	56.7	50.6	59.5	52.8	53.6
UD	28	43.2	48.2	56.7	47.6	45

Spectral subtraction is frequently used to generate binary T-F masks in missing data studies [8]. 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 25 frames of the noisy speech spectrum. The noise spectrum is then used to estimate the local SNR in each T-F unit. As in [8], a T-F unit is labeled speech-dominant in the mask if the local SNR exceeds 7.7 dB. Table 1 summarizes the performance of the uncertainty decoder ("UD") on the reconstructed cepstra by utilizing the estimated uncertainty. Performance is measured in terms of word error rate (WER). For comparison, we also show the performance of the conventional decoder on the reconstructed cepstra ("Enhanced Speech") [10]. Additionally, the baseline performance of the conventional decoder on the noisy data is also shown ("Baseline"). 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. Moreover, substantial improvement over the baseline performance is also obtained. Notice that the system is able to generalize well across noise conditions not seen during the MLP training.

We now present results using a monaural computational auditory scene analysis (CASA) system [11]. This system is a voiced speech separation system based on two main stages: 1) segmentation and 2) 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 are likely to belong to the same source are grouped together. 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 [11]. Provided the speech pitch contour can be estimated, this segregation mechanism produces a binary mask that selects T-F units where speech dominates the interference. The system shows a robust performance when tested with a variety of noise intrusions. For input to the system in [11], a pitch estimate is derived from the noisy speech signal using Praat [16]. The system in [11] 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. Further, if a valid pitch is not detected in a particular frame, we use the mask obtained by spectral subtraction in those frames. Table 2 shows the performance of the uncertainty decoder when using the combined mask from [11] and spectral subtraction. As before, across all SNR conditions, significant improvement over the performance of the conventional

Table 2. WER (%) of uncertainty decoding and recognition with reconstructed cepstra when using the combined voiced and spectral subtraction mask on the Aurora4 task.

System	*Test Set*					
	2	**3**	**4**	**5**	**6**	**7**
Enhanced Speech	40.5	50.7	47.6	54.7	50.3	52.3
UD	25.2	40.7	42.1	54.3	46.7	48.9

Table 3. WER (%) from uncertainty decoding with estimated and ideal variance and recognition with reconstructed cepstra when using the *a priori* mask.

System	*Test Set*					
	2	**3**	**4**	**5**	**6**	**7**
Enhanced Speech	23.4	33.6	36.1	35.7	31.1	41.8
Estimated UD	19	27.7	30.3	25.6	24.2	35.2
Ideal UD	18	21.2	26.3	19.9	23.3	32.4

ASR on the enhanced speech [10] is obtained when using the estimated variance. Note that under non-stationary noise conditions (e.g. babble), the performances of both the conventional ASR and uncertainty decoder are significantly better than their performance when using the spectral subtraction mask alone.

To show the ceiling performance of the proposed method, we also report the results obtained using *a priori* binary T-F masks. These masks are generated in a similar fashion to those used in our MLP training. For comparison, recognition results using the ideal uncertainty ("Ideal UD") are also shown. Ideal uncertainty is computed as the squared difference between the reconstructed and clean cepstra as in [3]. Table 3 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 the ability of the proposed approach to estimate the uncertainty associated with the reconstructed cepstra accurately.

5. CONCLUSION

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 paradigm in [3] on the Aurora4 task, we have shown that the estimated uncertainty can yield significant reductions in WER compared to conventional recognition on the enhanced cepstra. We have also obtained substantial improvement over baseline ASR performance.

A key advantage of the proposed method is that it does not assume a noise model. Our MLP training 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 4, the system is able to generalize across noise sources not seen during the MLP training. Hence, the proposed method can be used in conjunction with CASA systems that do no require noise conditions be known *a priori* for robust speech recognition.

ACKNOWLEDGMENTS. This research was supported in part by an AFOSR grant (FA9550-04-1-0117) and an AFRL grant via Veridian. We thank A. Acero and M. L. Seltzer for helpful suggestions.

6. REFERENCES

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

[2] 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.

[3] 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.

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

[5] 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.

[6] 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.

[7] 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.

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

[9] 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.

[10] 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.

[11] 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.

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

[13] D. Williams, X. Liao, Y. Xue, and L. Carin, "Incomplete-data classification using logistic regression," in *Proc. The 22nd International Machine Learning Conference*, L. D. Raedt and S. Wrobel, Eds. 2005, ACM Press.

[14] N. Parihar and J. Picone, "Analysis of the aurora large vocabulary evalutions," in *Proc. Eurospeech '03*, 2003, pp. 337–340.

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

[16] P. Boersma and D. Weenink, "Praat: doing Phonetics by Computer, Version 4.0.26," 2002.