EXPLOITING UNCERTAINTIES FOR BINAURAL SPEECH RECOGNITION

Soundararajan Srinivasan[1], *Nicoleta Roman*[2] *and DeLiang Wang*[3]

[1]Biomedical Engineering Department
[3]Department of Computer Science & Engineering and Center for Cognitive Science
The Ohio State University, Columbus, OH 43210, USA

[2]Department of Mathematics
The Ohio State University at Lima, Lima, OH, 45804, USA
{srinivasan.36, roman.45}@osu.edu, dwang@cse.ohio-state.edu

ABSTRACT

Recently several algorithms have been proposed to enhance noisy speech by estimating the signal-to-noise ratio (SNR) within a local time-frequency region based on binaural cues of interaural time and intensity differences (ITD and IID). However, the accuracy of the estimated SNR often varies widely across time and frequency, causing uncertainties in the enhanced speech features. We estimate this uncertainty based on statistics of ITD and IID and show that it can be effectively exploited to improve robust speech recognition. Systematic evaluations using the estimated uncertainty show significant improvement in recognition performance compared to the baseline performance.

Index Terms— computational auditory scene analysis, binaural processing, missing-data recognition, robust speech recognition, uncertainty decoding

1. INTRODUCTION

The performance of automatic speech recognizers (ASRs) degrades rapidly in the presence of noise and other distortions [1]. Speech recognizers are typically trained on clean speech and face a problem of mismatch when used in noisy conditions. While human listeners are able to recognize speech under such adverse conditions, automatic speech recognition remains a challenging problem. Inspired by robustness of the human auditory system, research in computational auditory scene analysis (CASA) has been devoted to building speech separation systems that incorporate known principles of auditory perception. In particular, binaural CASA systems which utilize location information for separation have achieved promising recognition results [2, 3]. Binaural systems typically compare the acoustic signals at the two ears in order to extract the binaural cues of ITD and IID. These cues are correlated with the location of a sound source and hence provide powerful mechanisms for segregating sound sources from different locations.

Typically, CASA systems achieve speech separation by estimating an ideal binary time-frequency (T-F) mask, which is then used as a front-end for missing-data recognition [2, 3, 4]. A T-F unit in the ideal binary mask is labeled 1 if the corresponding T-F unit of the noisy speech contains more speech energy than interference energy; it is labeled 0 otherwise. On the other hand, a frame-based Weiner filter, which utilizes the SNR in a local T-F unit, can effectively enhance noisy speech [5]. Additionally, it has been shown that in a narrow frequency band there exists a systematic relationship between SNR and values of ITD and IID [2]. Motivated by this observation, several binaural systems have also been proposed to estimate this local SNR using ratio or soft masks for speech enhancement [6, 7, 8]. While such systems have also achieved promising results, the accuracy of these algorithms often varies widely across time and frequency.

In this paper, we estimate the uncertainties associated with enhanced speech features using statistics collected for ITD and IID. We show that the estimated uncertainties can be effectively exploited in two different robust speech recognition strategies: Uncertainty decoding and missing-data recognition. In the uncertainty decoding approach, speech enhancement uncertainties contribute to an increase in the variance of acoustic model variables [9]. This results in an integration over all possible speech feature values during the computation of the observation likelihood. We show that the use of estimated speech feature uncertainties in the uncertainty decoder significantly improves recognition performance compared to direct recognition of enhanced speech.

Missing-data recognition makes use of spectro-temporal redundancy in speech to recognize a noisy signal based on its speech dominant T-F units [4]. When parts of an observed speech energy vector, $|X|^2$ are masked by noise or other distortions, $|X|^2$ can be partitioned into its reliable and unreliable constituents as $|X|_r^2$ and $|X|_u^2$. In the marginalization method, the posterior probability is computed using only the reliable constituents by integrating over the unreliable ones. When sound sources are additive, it is suggested in [4] that the true speech value $|\tilde{X}|_u^2$ in the unreliable part may be constrained as $0 \leq |\tilde{X}|_u^2 \leq |X|_u^2$ and therefore used to bound the integral involved in marginalizing the unreliable parts (see Section 3.2). We show that the uncertainties extracted using the binaural cues can also be used to derive tighter bounds and consequently a better recognition score than the marginalization method in [4].

The rest of the paper is organized as follows. The next section contains a detailed presentation of the proposed binaural system. The uncertainty decoding and missing-data approaches to robust speech recognition are briefly reviewed in Section 3. The proposed systems have been systematically evaluated on a connected digit recognition task and the evaluation results are presented in Sec-

S. Srinivasan is currently at Research and Technology Center, Robert Bosch LLC, Pittsburgh, PA 15212, USA. {soundar.srinivasan@us.bosch.com}.

tion 4. Finally, conclusions and future work are given in Section 5.

2. A BINAURAL FRONT-END FOR AUTOMATIC SPEECH RECOGNTION

We employ a binaural front-end for both uncertainty decoding and missing-data recognition. The uncertainty decoder uses the enhanced speech features obtained using a ratio mask (see Eq. 3) in conjunction with its associated uncertainties. These uncertainties will also be used by the missing-data ASR along with a binary mask (see Section 3.2). The input to the binaural system is a mixture of speech and interference presented at different, but fixed locations. Signals are upsampled from their original frequencies (see Section 4) to 44.1 kHz. Binaural signals are obtained by filtering the monaural signals with measured head-related transfer functions (HRTFs) from a KEMAR dummy head. HRTFs provide location-dependent ITD and IID which can be extracted independently in each T-F unit. The T-F resolution is 20 ms time frames with a 10 ms frame shift, and 512 DFT coefficients. Frames are extracted by applying a running Hamming window to the signal. The ITD/IID estimates are based on the spectral ratio at the two ears:

$$\left(I\hat{T}D, I\hat{I}D\right)(\omega,t) = \left[-\frac{1}{\omega}\angle\frac{X_L(\omega,t)}{X_R(\omega,t)}, \frac{|X_L(\omega,t)|}{|X_R(\omega,t)|}\right] \qquad (1)$$

where $X_L(\omega,t)$ and $X_R(\omega,t)$ are the left and right ear spectral values of the noisy speech signal at frequency ω and time t.

An ideal ratio T-F mask can be computed based on the *a priori* energy ratio $R(\omega,t)$ between target and interference, which is defined as follows:

$$R(\omega,t) = \left[\frac{|S(\omega,t)|^2}{|S(\omega,t)|^2 + |N(\omega,t)|^2}\right], \qquad (2)$$

where $S(\omega,t)$ and $N(\omega,t)$ are the target and noise spectral values at the better ear (the ear with higher SNR). In addition, the ideal binary mask assigns the label 1 to those T-F units whose value of R exceeds 0.5 and assigns the label 0 otherwise.

As seen in Fig. 1, for mixtures of multiple sound sources there exists a strong correlation between the *a priori* energy ratio and the estimated ITD and IID. The scatter plot in Fig. 1A shows the distribution of ITD and *R* for a frequency bin centered at 1 kHz. Similarly, Fig. 1B shows the results for IID at 3.4 kHz. The data is obtained from the training set in [2] consisting of 10 speech signals from the TIMIT database. For the estimation of the binary mask, we employ a nonparametric classification in the joint ITD-IID feature space as used in [2, 6]. In order to estimate the ratio mask, we use ITD below 3 kHz and IID at higher frequencies. At each frequency bin, a mean curve is fitted to the distribution of the ITD/IID and *R* after removing the outliers (outside of 0.2 distance from the median). Thus, for a given $I\hat{T}D|I\hat{I}D(\omega,t)$, the estimated ratio mask $\hat{R}(\omega,t)$ is the corresponding value on the mean curve. The enhanced speech spectral energy $|\hat{S}(\omega,t)|^2$ is computed as:

$$|\hat{S}(\omega,t)|^2 = |X(\omega,t)|^2 \cdot \hat{R}(\omega,t), \qquad (3)$$

where $|X(\omega,t)|^2$ is the spectral energy of the signal at the better ear (see Section 4). In addition, for a given $I\hat{T}D|I\hat{I}D(\omega,t)$, the uncertainty associated with the estimated ratio is given by the variance of *R*, $\hat{\sigma}^2_{\hat{R}(\omega,t)}$. This is then used to obtain the uncertainty associated with the enhanced spectral energy, $\hat{\sigma}^2_{|\hat{S}(\omega,t)|^2}$:

$$\hat{\sigma}^2_{|\hat{S}(\omega,t)|^2} = \hat{\sigma}^2_{\hat{R}(\omega,t)} \cdot |X(\omega,t)|^2. \qquad (4)$$

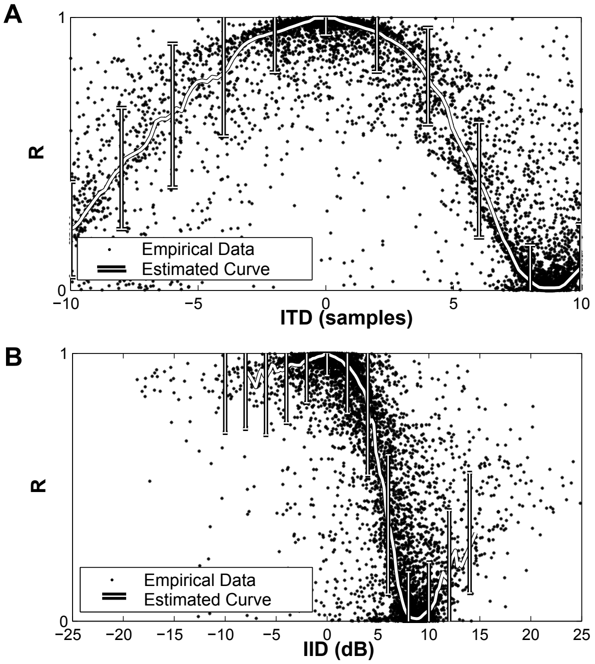

Fig. 1. Relationship between ITD/IID and the energy ratio R. Statistics are obtained with target in the median plane and interference on the right side at $30°$. (A) The scatter plot for the distribution of *R* with respect to ITD for a frequency bin at 1 kHz. The solid white curve shows the mean curve fitted to the data. The vertical bars represent the standard deviation. (B) Corresponding results for IID for a frequency bin at 3.4 kHz.

3. ROBUST SPEECH RECOGNITION WITH UNCERTAIN AND MISSING DATA

In this section, we describe how the use of $\hat{\sigma}^2_{|\hat{S}(\omega,t)|^2}$ can help improve speech recognition using two robust strategies: Uncertainty decoding and missing-data recognition.

3.1. Uncertainty Decoding

The uncertainty decoding method accounts for the imperfections in speech enhancement by integrating the observation likelihood over all possible speech feature values [9]. The observation density of each state in a HMM-based ASR is usually modeled as a mixture of Gaussians. Let $p(z|k,q) = \mathcal{N}(z;\mu_{k,q},\sigma^2_{k,q})$ be the likelihood of observing a clean feature z given state q and mixture k. In this work, z consists of 12 Mel-frequency cepstral coefficients and the log frame energy along with the corresponding delta and acceleration coefficients. Let the enhanced speech value be denoted as $\hat{z}$ and its associated variance as $\sigma^2_{\hat{z}}$. Under these conditions, it is shown in [9] that the new observation likelihood can be computed as

$$\mathcal{N}(\hat{z};\mu_{k,q},\sigma^2_{k,q} + \sigma^2_{\hat{z}}). \qquad (5)$$

Hence, the uncertainty associated with the enhanced features increases the variance of the Gaussian mixture component. Therefore, those enhanced speech features that deviate more from clean ones will contribute less to the overall likelihood. Recall that $\hat{\sigma}^2_{|\hat{S}(\omega,t)|^2}$ is the variance associated with the enhanced spectral features. Since in our work the uncertainty decoder uses cepstral features, this variance needs to be transformed into the cepstral domain. Hence, we perform a non-linear regression to transform the estimated spectral-domain variance into the cepstral domain as proposed in [10]. Specifically, we use a multilayer perceptron (MLP) to transform $\hat{\sigma}^2_{|\hat{S}(\omega,t)|^2}$ into $\sigma^2_{\hat{z}}$, the variance associated with the enhanced cepstra. For each frame, the input to the perceptron consists of $\hat{\sigma}^2_{|\hat{S}(\omega,t)|^2}$ corresponding to that frame supplemented by the enhanced cepstra in that frame and in one frame before and after. The desired MLP output is set to be the squared difference between the enhanced and clean cepstra [9]. The details of our MLP training can be found in [10]. We train the MLP using the mixtures of speech signals mentioned in Section 2 and not the signals used in our recognition experiments (see Section 4).

3.2. Missing-data Recognition

The missing-data ASR uses the binary mask produced by speech separation systems to partition an input spectral energy vector into its reliable and unreliable components. The missing data ASR treats the T-F regions labeled 0 as unreliable data during recognition and marginalizes the unreliable components in the computation of the observation likelihood. It is suggested in [4] that under additive and uncorrelated noise conditions, the true speech energy $|\tilde{X}|^2_u$, in the unreliable part may be constrained as $0 \leq |\tilde{X}|^2_u \leq |X|^2_u$. This constraint is then used as bounds on the integral used in marginalizing the unreliable features:

$$p(y|k,q) = p(y_r|k,q)\int_0^{|X|^2_u} p(y_u|k,q)dy_u, \qquad (6)$$

where $p(y|k,q)$ is modeled as a Gaussian, y_r and y_u correspond to clean spectral energies in the reliable and unreliable parts respectively and $y = y_r \cup y_u$. We propose to strengthen the limits in (6) by using the estimated standard deviation, $\hat{\sigma}_{|\hat{S}(\omega,t)|^2}$ as:

$$p(y|k,q) = p(y_r|k,q)\int_{|\hat{S}(\omega,t)|^2-\hat{\sigma}_{|\hat{S}(\omega,t)|^2}}^{|\hat{S}(\omega,t)|^2+\hat{\sigma}_{|\hat{S}(\omega,t)|^2}} p(y_u|k,q)dy_u. \qquad (7)$$

Fig. 2 shows how the proposed upper bound, $|\hat{S}\,(\omega,t)\,|^2 + \hat{\sigma}_{|\hat{S}(\omega,t)|^2}$ and the the proposed lower bound $|\hat{S}\,(\omega,t)\,|^2 - \hat{\sigma}_{|\hat{S}(\omega,t)|^2}$ together offer tighter bounds for clean speech energy in a noisy frame than the original upper bound of noisy spectral energy $|X|^2_u$ and the lower bound of 0. The data corresponds to a mixture of clean speech and factory noise at an SNR of 0 dB.

4. EXPERIMENTAL RESULTS

We have evaluated the binaural system on a speaker-independent recognition of connected digits task. This task is also used in [4]. Thirteen (1-9, silence, short pause between words, zero and oh) word-level models are trained. All except the short pause model have 10 states, whose output distribution is modeled as a mixture of 10 Gaussians [4]. The short pause model has only three states. The TIDigits database's male speaker data [11] is used for both training and testing. 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 different from the speakers in the training set. The signals in this database are sampled at 20 kHz. A HMM toolkit, HTK [12] is used for training. During testing, the decoder is modified to incorporate uncertainty decoding and missing-data recognition. The noise source is factory noise from the NOISEX corpus [13], which is also used in [4]. Factory noise is chosen as it has energy in the formant regions, therefore posing challenging problems for ASR. Noise is added at a range of SNRs from -10 dB to 10 dB in steps of 5 dB. In all our experiments, the target speech source is in the median plane and the noise source on the right side at 30°, making the left ear the better ear in terms of SNR.

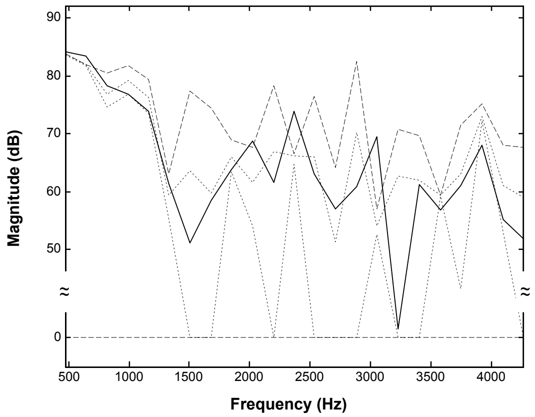

Fig. 2. An illustration of tighter bounds for clean speech spectrum using estimated uncertainties. The solid curve shows the clean speech spectral energies in a noisy frame. The proposed upper and lower bounds for the estimate of clean speech are shown using the dotted lines. The original upper bound in [4] is the noisy spectral energy and the original lower bound is 0. These are shown as dashed lines. Note that the proposed bounds are much tighter compared to the original bounds.

We first report results of using uncertainty decoding on the enhanced speech (using the ratio mask). Recall that the uncertainty decoder operates in the cepstral domain and utilizes the enhanced cepstra $\hat{z}$ and the cepstral domain variance, $\sigma^2_{\hat{z}}$. Fig. 3 summarizes the performance of the uncertainty decoder using the estimated uncertainty ("Enhanced Speech + Estimated UD"). Performance is measured in terms of word-level recognition accuracy at various SNRs. For comparison, we also show the performance of the conventional ASR (without uncertainty) on the enhanced cepstra ("Enhanced Speech") and that of the uncertainty decoder using the ideal uncertainty ("Enhanced Speech + Ideal UD"). Ideal uncertainty is computed as the squared difference between the enhanced and clean cepstra as in [9]. Additionally, the baseline performance of the conventional ASR on the noisy data ("Baseline") and that obtained by using an advanced front-end feature extraction algorithm ("ETSI AFE"), which is standardized by the European Telecommunication Standards Institute (ETSI) [14], are also shown. Across all SNRs, the performance of the uncertainty decoder using the estimated uncertainty shows significant improvement over that of the conventional ASR on the enhanced cepstra. At SNR greater than or equal to 0 dB, the performance of the uncertainty decoder using the estimated uncertainty is close to its performance using the ideal uncertainty. Moreover, substantial improvement over the baseline performance and ETSI advanced feature extraction algorithm (at most SNRs) are also obtained.

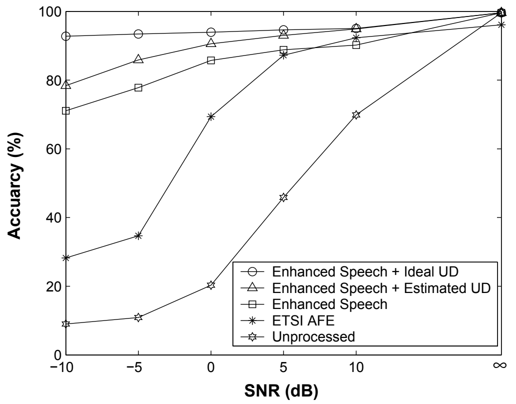

Fig. 3. Performance of uncertainty decoding with estimated and ideal variances and recognition with enhanced cepstra. For comparison, the performance of the conventional ASR without the use of any front-end processing and with processing by the ETSI advanced feature extraction algorithm are also shown.

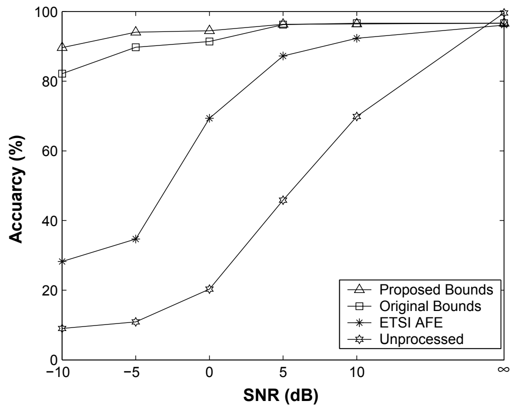

Fig. 4. Performance of missing-data recognition with proposed and original bounds for marginalization. For comparison, the performance of the conventional ASR without the use of any front-end processing and with processing by the ETSI advanced feature extraction algorithm are also shown.

We now present results of the missing-data recognition using the binary mask produced by the binaural system. Note that the missing-data ASR operates using the noisy spectral energy feature vectors. Fig. 4 summarizes the performance of the missing-data ASR by using the bounds in (7) (“Proposed Bounds”). For comparison, we also show the performance of the missing-data ASR using the bounds from [4] (“Original Bounds”). Additionally, the baseline performance of the conventional ASR on the noisy data (“Baseline”) and that obtained by using the ETSI advanced front-end algorithm (“ETSI AFE”) are also shown. At SNR less than or equal to 0 dB, the use of the proposed bounds leads to significant improvement over that of the original bounds. For example at SNR$= -10$ dB, a reduction in word-error-rate (WER) of 42% is obtained. Note also that the missing-data recognition yeilds substantial improvement over the baseline performance as well as recogniton on the ETSI features.

5. CONCLUSION

We have proposed a method for estimating the uncertainties resulting from imperfections in speech enhancement by a binaural CASA system. Using the uncertainty decoder in [9], we have shown that the estimated uncertainty can yield significant reductions in WER compared to conventional recognition on the enhanced cepstra. Additionally, the uncertainties are used to derive tighter bounds for the bounded marginalization method, resulting in improved missing-data recognition. Note that our estimation of the speech feature uncertainty is independent of speech and noise signals used in ASR evaluation, which is desirable for robust speech recogntion.

ACKNOWLEDGMENTS. 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 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] 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.

[3] 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 Commun.*, vol. 43, pp. 361–378, 2004.

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

[5] R. McAulay and M. L. Malpass, “Speech enhancement using a soft-decision noise supression filter,” *IEEE Trans. Acoust., Speech, and Signal Process.*, vol. 28, pp. 137–145, 1980.

[6] S. Srinivasan, N. Roman, and D. L. Wang, “On binary and ratio time-frequency masks for robust speech recognition,” in *Proc. ICSLP ’04*, 2004, pp. 2541–2544.

[7] H-M. Park and R. M. Stern, “Spatial separation of speech signals using continuously-variable masks estimated from comparisons of zero crossings,” in *Proc. ICASSP ’06*, 2006, vol. IV, pp. 1165–1168.

[8] Y-I. Kim, S. J. An, and R. M. Kil, “Zero-crossing based binaural mask estimation for missing data speech recognition,” in *Proc. ICASSP ’06*, 2006, vol. V, pp. 89–92.

[9] 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. Speech, and Audio Process.*, vol. 13, pp. 412–421, 2005.

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

[11] R. G. Leonard, “A database for speaker-independent digit recognition,” in *Proc. ICASSP ’84*, 1984, pp. 111–114.

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

[13] 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,” Technical Report, Speech Research Unit, Defense Research Agency, Malvern, UK, 1992.

[14] STQ-AURORA, “Speech Processing, Transmission and Quality Aspects (STQ); Distributed speech recognition; Advanced front-end feature extraction algorithm; Compression algorithms,” in *ETSI ES 202 050 V1.1.4*. European Telecommunications Standards Institute, 2005-11.