ROBUST SPEECH RECOGNITION BY INTEGRATING SPEECH SEPARATION AND HYPOTHESIS TESTING

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

Missing data methods attempt to improve robust speech recognition by distinguishing between reliable and unreliable data in the time-frequency domain. Such methods require a binary mask which labels time-frequency regions of a noisy speech signal as reliable if they contain more speech energy than noise energy and unreliable otherwise. Current methods for estimating the mask are based mainly on bottom-up speech separation cues such as harmonicity and produce labeling errors that cause a degradation in recognition performance. We propose a two stage recognition system in order to improve mask estimation and produce better recognition results. First, an *n*-best lattice consistent with the speech separation mask is generated. The lattice is then re-scored by expanding the mask using a model-based hypothesis test to determine the reliability of individual time-frequency regions. Systematic evaluations show significant improvement in recognition performance compared to that using speech separation.

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 based systems (e.g. [2]). If samples of the corrupting noise source are available *a priori*, a model for the noise can additionally be trained and noisy speech may be jointly decoded based on the models of speech and noise [3]. However, in many realistic applications, the performance of the above approaches to robust speech recognition is inadequate [4].

Recently a missing data approach to robust speech recognition has been proposed [4]. This method distinguishes between reliable and unreliable data in the spectral or time-frequency (T-F) domain. When speech is contaminated by additive noise, some T-F regions will contain predominantly speech energy (reliable) and the rest are dominated by noise energy. The missing data method treats the latter T-F units as missing or unreliable during recognition (see Section 2.1). The performance of the missing data recognizer is significantly better than the performance of a system using spectral subtraction for speech enhancement followed by recognition of enhanced speech [4].

The missing data recognizer requires a binary T-F mask that provides information about which T-F regions, of the noisy speech signal, are reliable and which are unreliable. Previous studies have shown that the missing data recognizer performs exceedingly well when this mask is known *a priori* [4, 5]. Attempts to estimate such a binary mask through front-end preprocessing using speech separation techniques have been only partly successful. Spectral subtraction is frequently used to generate such binary masks in missing data studies [4, 6]. Noise is assumed to be long-term stationary and its spectrum estimated from frames that do not contain speech (silent frames containing background noise). The noise spectrum is then used to estimate the signal to noise ratio (SNR) in each T-F unit. If the SNR in a T-F unit exceeds a threshold, it is labeled reliable; it is labeled unreliable otherwise. In the presence of non-stationary interference sources, however, the use of spectral subtraction results in a poor estimate of the mask. Methods that primarily utilize the harmonicity of voiced speech have also been proposed to estimate the mask for missing data applications [7, 8, 9]. Hence, they are unable to effectively deal with unvoiced speech. Additionally, accurate estimation of pitch is difficult, if not impossible, when SNR is low. Under these conditions, estimation of the binary mask corresponding to voiced speech may not be reliable too. Thus, the estimation of the binary T-F mask remains a challenging problem.

On the other hand, the human auditory system exhibits a remarkable ability to segregate a target speech source from various interference. According to Bregman [10], this is accomplished via a process termed auditory scene analysis (ASA). ASA involves two types of organization, primitive and schema-driven. Primitive ASA is based on bottom-up cues such as pitch, and spatial location of a sound source. Schema-based ASA is based on top-down use of stored knowledge about auditory inputs, e.g. speech patterns, to supplement primitive analysis. We therefore believe that a top-down approach, using speech models, could be used to refine the mask generated by bottom-up front-end processing to achieve improved recognition results.

In this paper, we present a two-pass missing data recognition system that estimates the binary T-F mask and produces recognition results in the mean time. In the first pass, a mask produced by a speech separation system is used to generate an *n*-best lattice using a missing data recognizer. This represents bottom-up processing. This lattice is then re-scored, to produce the final recognition results, by augmenting the initial mask using the information contained in the states along individual paths. Specifically, we propose a state-based hypothesis test to determine the reliability of each T-F unit. This corresponds to top-down analysis. The

In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2005)*, pp. I. 89-92, 2005

resulting recognition performance is substantially better than that of the conventional ASR and also significantly better that that of the missing data recognizer using the mask produced by speech separation alone.

The rest of the paper is organized as follows. The next section contains a detailed presentation of the model. The proposed system has been systematically evaluated on a noisy connected digit recognition task and the evaluation results are presented in Section 3. Finally, conclusions and future work are given in Section 4.

2. SYSTEM DESCRIPTION

The proposed system is a two-pass recognition system. In the first pass, we use a mask generated through front-end processing as input to a missing data recognizer to generate an *n*-best lattice. In the second pass, we use a state-based hypothesis test to enhance the mask and produce recognition results at the same time.

2.1. Bottom-up Speech Separation

The input to the system is a mixture of speech and interference, sampled at 20 kHz. Following the original study of Cooke et al. [4], we use an auditory filterbank decomposition of the input signal to generate the feature vectors for recognition. Specifically, the input is first analyzed using a 128 channel gammatone filterbank whose center frequencies are quasi-logarithmically spaced from 80 Hz to 5 kHz. The instantaneous Hilbert envelope at the output of each gammatone filter is smoothed using a first-order filter with 8 ms time constant. The smoothed envelope is then sampled at a frame rate of 10 ms and log compressed. As a result, the input signal is decomposed into a group of T-F units.

The missing data recognizer [4] makes use of spectro-temporal redundancy in speech to recognize a noisy signal based on its speech dominant T-F units. Given an observed speech vector Y, the problem of word recognition is to maximize the posterior $P\left(\omega_i|Y\right)$, where ω_i is a valid word sequence according to the grammar for the recognition task. When parts of Y are masked by noise or other distortions, Y can be partitioned into its reliable and unreliable constituents as Y_r and Y_u. In the marginalization method, the posterior probability using only the reliable constituents is computed by integrating over the unreliable ones [4]. If Y represents spectral energy and sound sources are additive, the unreliable parts can be constrained as $0 \leq Y_u \leq Y$. This bounded marginalization method is shown in [4] to have a better recognition score than the simple marginalization method, and is hence used in all our experiments.

A fundamental requirement of the missing data recognizer is therefore a binary mask that informs it whether a T-F unit is reliable (1) or unreliable (0). This mask is usually generated through front-end processing such as those based on spectral subtraction and harmonicitiy of voiced speech. As stated previously, while the accuracy of the mask produced by such bottom-up speech separation methods is good in limited situations, it may contain large errors under realistic conditions. Our experiments with the missing data recognizer have shown that wrongly labeling unreliable T-F units as reliable, especially, is harmful for recognition. Hence, in the first pass, we use a "conservative" binary T-F mask as input to the missing data recognizer. The conservative mask may be obtained, for example, using spectral subtraction by increasing the SNR threshold above which each T-F unit is labeled 1. This mask is needed to ensure that the first pass retains a small set of viable recognition candidates for subsequent top-down analysis. This will not only reduce the search space during the second pass, but also ensures the less likely recognition candidates are not involved in subsequent mask determination. The output of the system in this stage is a lattice from which an *n*-best hypothesis list can be generated.

2.2. Top-down Hypothesis Testing

In the second stage, we seek to augment the bottom-up mask by the top-down use of speech models. Specifically, we use those states of the hidden Markov model (HMM) speech recognizer that are contained in the *n*-best lattice.

The use of a conservative criterion during bottom-up mask generation ensures that the probability of each reliable T-F unit being dominated by speech is high. Hence, during top-down processing we only analyze those T-F units labeled 0 by the first stage. As, the number of unreliable T-F units under low SNR conditions is very high, a state-based analysis of each unreliable T-F unit is computationally prohibitive. Therefore, we only analyze those T-F units which have a high probability of being labeled 1. For this purpose, we use spectral subtraction. The spectrum of noise is estimated as the average spectrum of the first 10 frames of the noisy speech spectrum [4]. The noise spectrum is then used to estimate the local SNR in each T-F unit. Then each T-F unit labeled 0 by bottom-up speech separation is now be labeled using a threshold δ as

$$label = \begin{cases} 2 & \text{if } SNR_{local} \geq \delta \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

The choice of δ represents a trade-off between increasing the computation time of top-down analysis and possibly reducing the number of speech dominant T-F units. The optimal value of δ is dependent on SNR [11]. For simplicity we set δ to be a constant and use the value of $\delta = 0$ dB as suggested in [5]. Not all T-F units labeled 2 correspond to speech due to the limitations of spectral subtraction stated previously. Therefore, top-down processing is needed to remove noise dominant T-F units from among those labeled 2.

Note that those T-F units labeled 1 by bottom-up processing are not affected by spectral subtraction based labeling. Therefore, we now have a three-way mask. The lattice generated by the first pass is now re-scored using the missing data recognizer and the three way mask. During re-scoring, each active state independently analyzes the T-F units labeled 2. The observation density of each state in a HMM based ASR, which is usually modeled as a mixture of gaussians with diagonal covariance, models a particular class of speech signal. This information could therefore be used to verify whether the observed value in a T-F unit is consistent with a speech state. This corresponds to top-down processing. It has been suggested that the observed value may be used to generate "evidence" and "counter-evidence" measures for speech in a particular state [5]. In this paper, we use the two measures to construct a hypothesis test for each T-F unit labeled 2 as follows: If a T-F unit satisfies the inequality in (2), we label it as 1; it is labeled 0 otherwise.

$$\frac{p(y_i|k,q)}{p(y_i)} > \int_0^{y_i} p(x_i|k,q)\frac{1}{\alpha}dx, \tag{2}$$

where

$$\alpha = \int_0^{y_i} p(x_i)dx \tag{3}$$

is the normalization factor. y_i and x_i are respectively the observed spectral energy and speech spectral energy seen during training in a frequency channel i. $p(y_i|k,q)$ is the likelihood of observing y_i given state q and mixture k. The prior $p(x_i)$, used to normalize the probabilities in (2) is modeled empirically using a mixture of gaussians based on the data used in training of the ASR. The LHS of (2) therefore models the evidence for speech in state q. The RHS provides the probability that the observed value is speech corrupted with additive noise, or counter-evidence for speech. As stated previously, since y_i represents spectral energy, under additive noise conditions, the range for the true speech value is $[0 \ldots y_i]$. In the absence of any knowledge about the noise level, equation 2 represents a conservative decision. This is consistent with our observation regarding the wrong labeling of unreliable T-F units as 1. Recall that the top-down mask refinement is simultaneous with lattice re-scoring.

Thus, we first use bottom-up cues to generate a conservative mask and subsequently refine this mask using top-down processing. Barker et al. [5] have also proposed a top-down approach using a speech recognizer to identify the T-F units that correspond to speech signal in a noisy mixture. The primarily difference in our approach is in the use of bottom-up cues in our initial mask generation. This mask drastically helps reduces the search space for top-down analysis. Additionally, the prior distribution in their probabilistic framework is modeled as an uniform distribution which is suspected to cause a bias toward labeling the T-F units as unreliable [5].

3. EXPERIMENTAL RESULTS

We have evaluated the system on a speaker-independent recognition of connected digits. This task is used in [4, 5]. 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 [12] 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 232 utterances from 3 speakers different from the speakers in the training set. The speech prior is modeled empirically using a mixture of 230 gaussians using all the utterances in the training data. A HMM toolkit, HTK [13] is used for training. During testing, the decoder is modified to incorporate our mask-generation and missing data recognition. To test the robustness of the two-pass recognizer system on the aforementioned task, noise is added at a range of SNRs from -5 dB to 10 dB in steps of 5 dB. The noise source is factory noise from the NOISEX corpus [14], which is also used in [4]. Factory noise is chosen as it has energy in the formant regions, therefore posing challenging problems during recognition. It is also impulsive, making it difficult to estimate its spectrum using spectral subtraction methods [4].

Monaural CASA systems that compute an ideal binary mask have been used as front-ends for the missing-data recognizer previously [8]. A T-F unit in the ideal binary mask is labeled 1 if the corresponding T-F unit of noisy speech contains more speech energy than interference energy; it is labeled 0 otherwise. This mask may be obtained *a priori* , from premixing speech and noise. While several systems estimate this mask well in low-frequencies, they perform poorly in high-frequencies (for an exception, see [15]). Additionally, under noisy and reverberant conditions, high-frequency components of speech are more corrupted than low-frequency ones. Hence, to reveal the potential for our top-down processing stage, we set high-frequency components (above 1 kHz) of the ideal-binary mask to 0 and use it as our bottom-up, speech separation (initial) mask for the first pass. Here we use 2 best tokens in each state to generate the *n*-best lattice. Our experiments indicate that using 2 tokens in each state is sufficient to mostly retain the true hypothesis while maintaining a reasonable compuational cost. The top-down processing stage will be used to identify reliable regions above 1 kHz. Table 1 summarizes the performance of the proposed system when using the ideal binary mask below 1 kHz ("Integrated Mask"). Performance is measured in terms of word-level recognition accuracy at various SNRs. For comparison, we also show the performance of the missing data recognizer when using the ideal binary mask at all frequencies ("IBM"), which represents the ceiling performance for the proposed approach. Additionally, the performance of the missing data recognizer when using the initial mask is also shown ("Initial Mask").

Table 1. Digit recognition accuracy (%) of the proposed sytem and the missing data recognizer using the initial mask

System	*SNR (dB)*			
	-5	**0**	**5**	**10**
Initial Mask	18.9	67.4	72.8	71.9
Integrated Mask	18.5	74.9	79.6	85.8
IBM	72.4	85.7	92.7	96.2

At SNRs > -5 dB, the proposed system shows significant performance improvement over the missing-data recognizer when using the ideal binary mask below 1 kHz. This indicates that the top-down hypothesis testing stage of our system is able to correctly identify the reliable T-F units in the high frequency region. A slight degradation in performance is observed at -5 dB. This may due to the poor quality of the *n*-best list generated during the first pass with the missing data recognizer. Our results also confirm previous findings that show that missing data techniques achieve high accuracy of recognition when the ideal binary mask is available.

We now present results using a monaural CASA system that is able to handle both high and low frequencies of speech [15]. 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 [15]. Provided a target pitch contour can be estimated, this segregation mechanism produces a binary mask which selects T-F units where speech dominates the interference. The system shows a robust performance when tested with a variety of noise intrusions. Hence, we use the system to generate our initial mask. For input to the system in [15], a pitch estimate is derived from Praat [16].

Fig. 1 summarizes the performance of the proposed system when using the mask generated by the system in [15] as the initial mask for the first pass. Across all SNR conditions, the proposed system shows significant improvement over the performance of the missing data recognizer with the initial mask. For e.g, at 10 dB, a reduction in word-error-rate (WER) of 47% is obtained. Note at

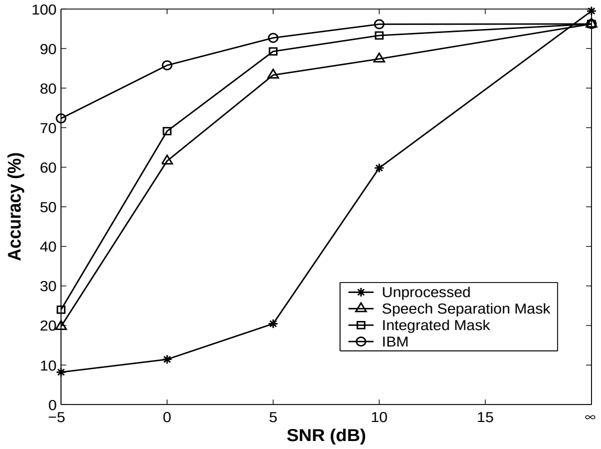

Fig. 1. Performance of the proposed system and the missing data recognizer using the mask produced by the speech separation system in [15]. Integrated Mask refers to the performance of the proposed system. IBM and Speech Separation Mask refer to the performance of the missing-data recognizer using the ideal binary mask and mask from [15] respectively. For comparison, the performance of the conventional ASR without the use of any front-end processing is also shown.

SNRs > 0 dB, the performance of the proposed system is close to that of the missing data recognizer when using the ideal binary mask. Also note that the performance of both the proposed system and the missing data recognizer is substantially better than that of a conventional MFCC based ASR with no preprocessing.

4. CONCLUSION

The missing data recognizer shows excellent performance when the ideal binary is accurately estimated. Mask estimated by speech separation systems may contain large errors even under moderate amounts of additive noise. In this paper, we have presented a two-pass missing data recognition system that refines the mask generated through front-end processing and provides significant reduction in the WER compared to that of the missing data recognizer when using the speech separation mask. It is known that lattice re-scoring can be an order of magnitude faster than normal recognition [13]. So due to the small size of the lattice generated in our first pass, only a small increase in the computation time is observed for our system over that of the missing data recognizer.

A significant advantage of our system is that a noise model is not required. Hence, it is applicable under various noise conditions. In our experiments we have used the ideal binary mask below 1 kHz and a mask derived from a CASA system as bottom-up speech separation masks. However, masks produced by other speech separation methods could also be used for the same. In using a fixed threshold with spectral subtraction to generate candidate T-F units during the second pass, the system is unable to produce adequate T-F units labeled 2 at very low SNRs. This may be the cause for the relatively small decrease in WER obtained under these conditions. Future work will address this issue and attempt to improve the performance at very low SNRs.

ACKNOWLEDGMENTS. This research was supported in part by an AFOSR grant (FA9550-04-1-0117) and an NSF grant (IIS-0081058). We thank J. Barker for help with the speech fragment decoder.

5. REFERENCES

[1] Y. Gong, “Speech recognition in noisy environments: A survey,” *Speech Communication*, vol. 16, pp. 261–291, 1995.

[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] M. J. F. Gales and S. J. Young, “Robust continuous speech recognition using parallel model combination,” *IEEE Trans. on Speech, and Audio Processing*, vol. 4, pp. 352–359, 1996.

[4] 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.

[5] 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.

[6] A. Drygajlo and M. El-Maliki, “Speaker verification in noisy environment with combined spectral subtraction and missing data theory,” in *Proc. ICASSP ’98*, 1998, vol. 1, pp. 121–124.

[7] M. L. Seltzer, B. Raj, and R. M. Stern, “Classifier-based mask estimation for missing feature methods of robust speech recognition,” in *Proc. ICSLP ’00*, 2000, pp. 538–541.

[8] G. J. Brown, J. Barker, and D. L. Wang, “A neural oscillator sound separator for missing data speech recognition,” in *Proc. IJCNN ’01*, 2001, pp. 2907–2912.

[9] H. Van Hamme, “Robust speech recognition using cepstral domain missing data techniques and noisy masks,” in *Proc. ICASSP ’04*, 2004, vol. 1, pp. 213–216.

[10] A. S. Bregman, *Auditory scene analysis*, The MIT Press, Cambridge, MA, 1990.

[11] 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.

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

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

[14] 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.

[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] P. Boersma and D. Weenink, “Praat: doing Phonetics by Computer, Version 4.0.26,” *http://www.fon.hum.uva.nl/praat*, 2002.