BAND-INDEPENDENT MASK ESTIMATION FOR MISSING-FEATURE RECONSTRUCTION IN THE PRESENCE OF UNKNOWN BACKGROUND NOISE

Wooil Kim[1)] and Richard M. Stern[2)]

[1)]Center for Robust Speech Systems, Dept. of Electrical Eng., University of Texas at Dallas, Richardson, TX 75080, USA
[2)] Dept. of Electrical and Computer Engineering and Language Technologies Institute
Carnegie Mellon University, Pittsburgh, PA 15213, USA
wikim@utdallas.edu, rms@cs.cmu.edu

ABSTRACT

An effective mask estimation scheme for missing-feature reconstruction is described that achieves robust speech recognition in the presence of unknown noise. In previous work on Bayesian classification for mask estimation, white noise and colored noise were used for training mask estimators. This paper, which is concerned with both the simulation of a more diverse set of background environments and with mitigating the "sparse training" problem, describes a new Bayesian mask-estimation procedure in which each frequency band is trained independently. The new method employs colored noise for training, which is obtained by partitioning each frequency subband. We also propose a re-evaluation method of voiced/unvoiced decisions to alleviate performance degradation that is caused by errors in pitch detection. Experimental results indicate that the proposed procedure in conjunction with cluster-based missing-feature imputation improves speech recognition accuracy on the Aurora 2.0 database in the presence for all types of background noise considered.

1. INTRODUCTION

Acoustic differences between training environments and the environment in which an actual speech recognition system must work is one of the primary factors that degrades speech recognition accuracy, and the presence of background noise is a major such factor. While many signal-processing schemes have demonstrated reasonable success in the presence of quasi-stationary noise, they are still vulnerable to time-varying noise such as background music, since most methods rely on an estimation of corrupting noise components.

Missing-feature methods have been more effective in coping with the effects of non-stationary noise conditions on speech recognition accuracy (*e.g.* [1][2]). These methods depend mostly on the characteristics of speech that are resistant to noise, rather than on the characteristics of the noise itself. The missing-feature method consists of two steps. The first step is the estimation of a "mask" which determines which parts of a representation of noisy input speech are considered to be unreliable. The second step is to reconstruct the unreliable regions or bypass them for other processing. In this paper we focus on the first step.

Seltzer *et al.* [3] have previously proposed a Bayesian classifier for mask estimation, which was trained on speech corrupted by white noise for the purpose of environment-independent mask estimation. Nevertheless, we found in subsequent work that the use of white noise for training the Bayesian classifier does not in fact provide the desired environment independence in mask estimation. This motivated a training method that employs a combination of colored noises that has been described previously [4].

Our previous work had involved training the speech recognizer in the presence of background noise with different spectral characteristics in different frequency bands, but the multiple frequency bands led to difficulty in obtaining coverage of the possible environments with limited training data. In the algorithm described in this paper we avoid this problem through the use of a mask-estimation algorithm in which each frequency band is independent. We also propose a new method for generating colored noise to train the mask estimator, as well as a re-evaluation method of voiced/unvoiced decisions that mitigates the adverse effects introduced by errors in pitch detection.

This paper is organized as follows. We first review the missing-feature method in Section 2. We discuss the shortcomings of previous techniques and then describe the proposed algorithms in Sections 3 and Section 4. Representative experimental procedures and their results are presented and discussed in Section 5. Finally, in Section 6, we offer some conclusions about our work.

2. MISSING-FEATURE METHOD

2.1. Mask estimation

In the missing-feature approach, it is necessary to determine a "mask" which classifies the spectrum of incoming speech into reliable and unreliable ("missing") regions for missing-feature reconstruction. Reliable regions are defined as the spectro-temporal fragments for which the incoming speech appears to be dominant over the corrupting noise. In unreliable regions, the speech components are assumed to be distorted by the background noise. The reliable regions are assumed to be helpful to speech recognition, while the unreliable regions degrade the performance of recognition. Seltzer *et al.* [3] proposed a Bayesian classifier for mask estimation which makes no assumption about the corrupting background noise.

2.2. Missing-feature reconstruction

Many methods have been proposed for reconstructing missing features, including the cluster-based and correlation-based methods of Raj *et al.* [2]. Based on (maximum *a posteriori* probability (MAP) estimation techniques, they restore unreliable parts of speech spectrogram using both *a priori* characterizations of clean speech and the values of speech features in the reliable regions as indicated by the estimated masks. We employed the cluster-based reconstruction method [2] for our work.

The distributions of the log spectra of clean speech are modeled by Gaussian mixture densities with K clusters. A noisy speech vector $S(t)$ is considered to have reliable components $S_0(t)$ and missing components $S_m(t)$. We can determine the cluster membership k of $S(t)$ by its *a posteriori* probability. If $S(t)$ has unreliable elements, their probability could be calculated by integrating them out:

$$\begin{aligned}\hat{k}_{S(t)} &= \arg\max_k \{P(S(t)\mid k)P(k)\} \\ &= \arg\max_k \left\{ P(k) \int_{-\infty}^{Y_m(t)} P(S(t)\mid k)\, dS_m(t) \right\}\end{aligned} \quad (1)$$

where $Y_m(t)$ indicates the observed values of the unreliable parts. Finally, the unreliable parts $S_m(t)$ are restored using bounded MAP estimation based on the observations in the reliable regions $S_0(t)$, the Gaussian model of the cluster determined by (1), and the upper bound of $Y_m(t)$.

$$\hat{S}_m(t) = \arg\max_{S_m} \left\{ P(S_m(t) \mid S_0(t), \mu_{\hat{k}_{S(t)}}, \Sigma_{\hat{k}_{S(t)}}, S_m(t) \le Y_m(t)) \right\} \quad (2)$$

3. BAND-INDEPENDENT MASK ESTIMATION

3.1. Background and motivation

In our previous work we simulated various kinds of background noise by increasing the number of partitions N of the frequency band [4]. Theoretically, all kinds of noise observed in nature could be characterized by increasing N to half of size of the discrete Fourier transform (DFT) and training under a variety of SNRs. Unfortunately, the frequency of occurrence of each colored noise decreases as the number of kinds of noise increases when the amount of training data is limited. We believe that this is reflected in the failure to observe continued reductions in error rate as the number of partitions N increased beyond a point.

In this paper we propose a new approach to cope with the problem of limited data caused by increasing the number of frequency partitions. In our previous work the features of the mask classifier in a particular frequency band were affected by adjacent frequency bands, and the resulting data-insufficiency problem was exacerbated as we attempted to model an increased number of types of background noise. For these reasons we developed a new Bayesian classifier in which independent processing is performed in each frequency band. With this "band-independent" classifier, we only need to apply the various kinds of spectral events to a particular frequency band in order to simulate the various types of background noise considered. Therefore, we expect that the proposed method would enables us to characterize the spectral patterns of a number of different background environments while using a relatively small number of combinations of colored noise.

3.2. Design of the mask estimators

3.2.1. Subband cepstral coefficients

Cepstral coefficients, which provide an efficient characterization of the short-time spectral envelope, are used as features for the mask estimators, just as they are for the speech recognition system itself. Nevertheless, in our mask estimators, cepstral coefficients are derived separately for each of the spectral regions spanned by each Mel-filter channel. In other words, we obtain "subband cepstral coefficients" by taking the logarithm of the spectrum of the signal emerging from each Melfilter channel, applying the discrete cosine transform (DCT), retaining the upper N^{th} coefficients. In our experiments, a fifth-order subband cepstrum was calculated for every Mel-filter channel, along with its first derivative.

3.2.2. Spectral flatness measure (SFM)

The SFM indicates which tonal component (if any) is dominant in a given signal frame, and it has usually been used as a measure for determining which segments of an utterance are voiced or unvoiced [5]. The SFM can be calculated using the ratio of the geometric and arithmetic averages of the spectrum as in (3):

$$SFM = \left(\prod_N X(n) \right)^{1/N} \Big/ \frac{1}{N} \sum_N X(n) \quad (3)$$

The SFM is expected to represent the amount of contamination by the background noise and it is computed from the log spectrum selected by each Mel-filter channel in this paper.

Together with the subband cepstrum and SFM features, the Comb Filter Ratio (CFR) feature used in our previous work is also used as one of features in the voiced speech regions for the proposed mask estimator, which is based on Bayesian classification as in our previous work. The acoustic models of the estimator use a 12^{th}-order feature vector for voiced speech and an 11^{th}-order vector for unvoiced. Both feature vectors are trained using a new type of colored noise to be described in the next section. The mask estimators in each frequency band work independently.

3.3. Partitioned band-by-band training

In this section we describe a modified method for generating colored noise for training the band-independent mask estimator. In this proposed scheme, each Mel frequency sub-band is partitioned into N parts instead of partitioning the entire frequency band. We refer to this approach as the "intra-band partition method" for generating colored noise.

In the intra-band partition method for colored noise, a particular subband is partitioned into N parts. N types of colored noise are generated from each subband, each of which has energy only in a narrow-band corresponding to each partition. To generate each narrow-band colored noise, a method that is similar to spectral subtraction is employed, rather than the bandpass filtering used in [4]. Specifically, components of white noise are removed outside the frequency region of interest, and a resulting narrow-band colored noise is obtained in the time domain through inverse Fourier transformation.

This produces N types of colored noise which each have a narrow frequency range in a particular subband. We then generate 2^N types of colored noise to train the mask estimator by combining the N narrow-band colored noise that are obtained. The combination is selected randomly for each time frame (typically 30, 60, or 300 ms) for time-varying noise samples and the same combination is used over the entire duration if stationary noise samples are desired. A noise-corrupted speech database for training the band-independent mask estimator is produced by adding the colored noise obtained as described above to a clean speech database at various SNRs. In our work, 20, 15, 10, 5, and 0 dB are used.

Figure 1 compares the proposed method to the method used in our previous work for generating colored noise. As the figure shows, the entire frequency band is partitioned into four parts in

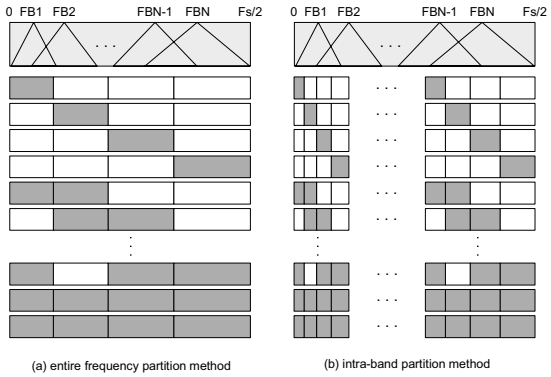

Fig. 1. Comparison of the methods used for generating colored noise. (a) entire frequency band partition method, (b) intra-band partition method.

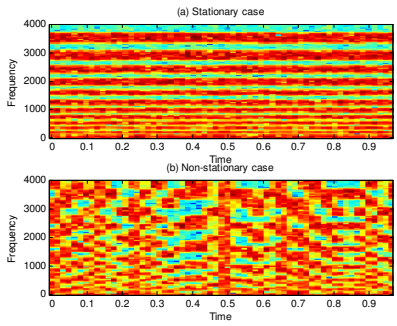

Fig 2. Spectrograms of colored noise generated as described in this paper: (a) stationary case, (b) non-stationary case.

the previous scheme while each Mel-filter band is split into four regions in the proposed method. We can obtain 2^N combinations of colored noise in both of methods. However, the proposed scheme has the effect of partitioning the entire frequency band into 4 times the number of Mel-filter banks M, or 2^{4M} combinations, since the mask estimator corresponding to each subband is trained independently. This means that we can simulate a greater number of spectral patterns with a relatively small number of combinations. Figure 2 shows examples of colored noise generated by the proposed method, in which each subband is split into four parts.

4. RE-EVALUATOIN OF VOICED DECISIONS

In our previous work, we proposed a restoration method of voiced frames in cases in which the misclassification of voiced frames led to errors in mask estimation [4]. In the present work we perform a similar type of re-evaluation method, but for frames that are classified unvoiced as well as voiced.

The proposed method is based on a simple classifier using Gaussian mixture models (GMMs) with general MFCCs as the feature vectors. GMMs for voiced speech, unvoiced speech, and silence were trained using the noise-corrupted speech database as described above. The training database was made by adding the colored noise to the clean speech database, which was generated by partitioning the entire spectrum as in Fig. 1a. Figure 3 presents a diagram of the entire re-evaluation method employed in this work and especially the left dashed-line box shows the procedure of unvoiced frame restoration scheme proposed in this paper. We found that it was more effective in reducing insertion errors to consider the frames which are re-evaluated as silence to be unreliable in mask estimation.

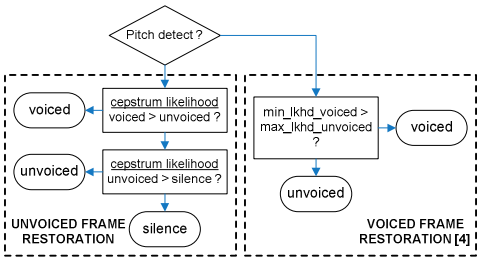

Fig. 3. Diagram of the re-evaluation method for voiced/unvoiced decisions.

5. EXPERIMENTAL RESULTS

We evaluated the proposed methods following the procedures specified by the Aurora 2.0 evaluation [6]. HMMs were trained for speech recognition (using the HTK package) and GMMs were trained for cluster-based missing-feature reconstruction using a training database that contained 8,440 utterances of clean speech. The multi-condition testing database was generated by combining clean speech from Set A in Aurora 2.0 with four types of noise samples: white noise, car noise, speech babble, and background music. The white noise and speech babble were obtained from NOISEX92 and the car noise was from Aurora 2.0. Background music was obtained from the prelude parts of ten Korean pop songs which have various types of intensity and speed. Speech and noise were combined with five SNRs; 20, 15, 10, 5, and 0 dB. Each of these 20 noise conditions is represented by 1,001 samples.

The performance of a baseline system was first evaluated at 5-dB SNR, and these results are shown in Figure 4. In this figure results obtained using the method described in this paper are compared to spectral subtraction (SS), cluster-based missing-feature restoration using masks derived from Oracle knowledge (missingO), and a combination of the latter two methods. These test conditions were used for all remaining experiments in this paper.

Figure 5 presents the recognition accuracy obtained using the band-independent mask estimation that is described in this paper. These results were obtained by using the mask estimation described above in conjunction with cluster-based missing-feature reconstruction [2]. The "Ex-multi" condition indicates multi-style training while excluding the testing condition. The curves show the dependence of recognition accuracy on the number of partitions used in each band while training the mask estimator. While there is some variability, greatest accuracy is obtained for 2 to 4 bands per channel. The fact that accuracy is comparable to or better than that obtained using "ex-multi" training confirms that the proposed method is effective for training in unknown background noise.

Figure 6 compares the recognition accuracy obtained using mask estimators trained using the intra-band partition method described in this paper to that obtained with mask estimators that had been trained by partitioning the entire spectrum as in [4]. The complete Aurora 2.0 testing database was used in obtaining these data. Four bands per Mel channel were used in the current implementation while 12 bands were used in partitioning the entire spectrum; both of these numbers provided best results in pilot data. It is seen that except for the case of speech babble, training the mask estimator using intra-band partitioning provides better recognition accuracy than training the estimator by portioning the entire spectrum. Improvement is especially evident at low SNRs with white noise and car noise and over all SNRs for background music. The significant improvement in background music, which

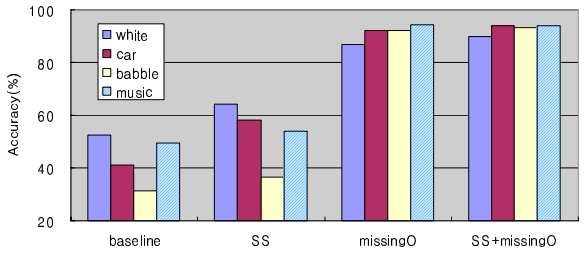

Fig. 4. Word accuracy of the baseline system at 5 dB SNR.

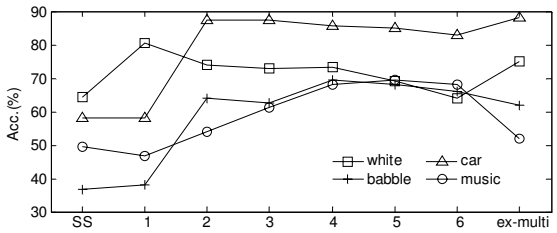

Fig. 5. Dependence of word accuracy using the proposed method on the number of partitions within each Mel frequency channel.

is especially problematic for recognition, provides additional validation for the approaches described. We believe that these approaches are successful in part because the mask estimator can be more completely trained with a small number of examples.

Figure 7 compares the recognition accuracy obtained with and without re-evaluation of the voiced/unvoiced decisions. It can be seen that the use of re-evaluation improved recognition accuracy for all background noises and all SNRs except for car noise at lower SNRs, and re-evaluation provided clear benefit in the case of the speech babble condition, which did not improve through the use of the new type mask estimation alone. Re-evaluation was effective especially because it reduced the number of insertion errors in speech recognition.

We also note that the proposed method usually outperforms the Vector Taylor Series (VTS) algorithm, which is known for its good performance in quasi-stationary noise [7]. Although the approach described in this paper produces slightly worse accuracy than VTS in white noise or car noise at high SNRs, it is better than VTS at lower SNRs. We believe that the accuracy of the present algorithm at high SNRs is adversely affected by insertion errors in recognition that are caused by errors in pitch-detection.

6. CONCLUSIONS

In this paper, we have described an effective method of mask estimation for missing-feature algorithms that obtains robust performance of speech recognition under unknown noise environments. We describe a new way of training a Bayesian classifier for mask estimation using colored noise that is generated by partitioning each Mel subband. A re-evaluation method for reducing the performance degradation due to incorrect voiced/unvoiced decisions was also proposed. Experimental results demonstrate that the proposed mask estimation scheme is effective in improving speech recognition accuracy under various kinds of unknown noise environments.

7. ACKNOWLEDGMENT

This work was supported by the National Science Foundation (Grant IIS-0420866) and the Post-doctoral Fellowship Program of Korea Research Foundation (D00142).

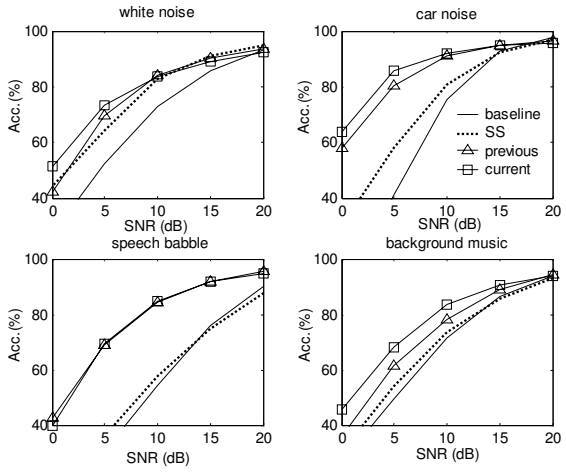

Fig. 6. Comparison of recognition accuracy obtained using the proposed method with the recognition accuracy that had been obtained previously.

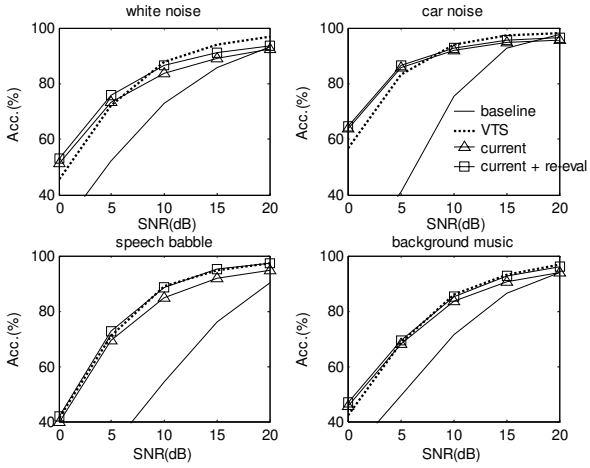

Fig. 7. Comparison of recognition accuracy obtained with and without re-evaluation method for voiced/unvoiced decisions.

8. REFERENCES

[1] M. Cooke, P. Green, L. Josifovski, and A. Vizinho, "Robust automatic speech recognition with missing and unreliable acoustic data," *Speech Communication*, 34(3): 267-285, 2001.

[2] B. Raj, M. L. Seltzer, and R. M. Stern, "Reconstruction of missing features for robust speech recognition," *Speech Communication*, 43(4): 275-296, 2004.

[3] M. L. Seltzer, B. Raj, and R. M. Stern, "A Bayesian classifier for spectrographic mask estimation for missing-feature speech recognition," *Speech Communication*, 43(4): 379-393, 2004.

[4] W. Kim, R. M. Stern, and H. Ko, "Environment-Independent Mask Estimation for Missing-Feature Reconstruction," *Proc. of Interspeech2005*, pp.2637-2640, Sep. 2005.

[5] J. D. Johnston, "Transform coding of audio signals using perceptual noise criteria," *IEEE Journal on Selected Areas in Communications*, vol.6, pp.314-323, Feb. 1988.

[6] H. G. Hirsch & D. Pearce, "The AURORA Experimental Framework for the Performance Evaluations of Speech Recognition Systems under Noisy Conditions," *ISCA ITRW ASR2000*, Sep. 2000.

[7] P. J. Moreno, B. Raj, and R. M. Stern, "Data-Driven Environmental Compensation for Speech Recognition: A Unified Approach," *Speech Communication*, 24: 267-85, 1998.