Environment-Independent Mask Estimation for Missing-Feature Reconstruction

Wooil Kim[1,2], Richard M. Stern[1] and Hanseok Ko[2]

[1)]Department of Electrical and Computer Engineering and School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA

[2)]Department of Electronics and Computer Engineering, Korea University, Seoul, Korea

{wikim,rms}@cs.cmu.edu, hsko@korea.ac.kr

10.21437/Interspeech.2005-248

Abstract

In this paper, we propose an effective mask-estimation method for missing-feature reconstruction in order to achieve robust speech recognition in unknown noise environments. In previous work, it was found that training a model for mask estimation on speech corrupted by white noise did not provide environment-independent recognition accuracy. In this paper we describe a training method based on bands of colored noise that is more effective in reflecting spectral variations across neighboring frames and subbands. We also achieved further improvement in recognition accuracy by reconsidering frames that appeared to be unvoiced in the initial pitch analysis. Performance is evaluated using the Aurora 2.0 database in the presence of various types of noise maskers. Experimental results indicate that the proposed methods are effective in estimating masks for missing-feature reconstruction while remaining more independent of the noise conditions.

1. Introduction

The presence of background noise typically causes differences between training and testing conditions, which can significantly degrade the recognition accuracy of speech recognition systems. Various schemes for reducing these differences have been developed over the last several decades and they have demonstrated reasonable success in the presence of stationary noise. Nevertheless, these approaches are still vulnerable to the effects of time-varying noise such as background music, since most of these schemes are primarily based on the 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. These methods depend mostly on the characteristics of speech that are resistant to noise, rather than the characteristics of the noise itself. This enables (in principle) an improvement in accuracy that is independent of the specific nature of the masking noise, even when the noise is transient in nature [1-3].

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 bypass or reconstruct the unreliable regions. In this paper, we focus on the first step. Seltzer *et al.* [2] 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. We have found, however, that the use of white noise for training the Bayesian classifier does not in fact provide the desired environment independence in mask estimation, for reasons that we believe are related to the inability of white noise to reflect spectral variations of practical noise environments realistically over time and frequency. For this reason we propose a new training method that employs a combination of colored noises. In addition, we propose a modified decision method that provides some additional improvement in accuracy for voiced frames that are improperly recognized as being unvoiced in the first-pass pitch detection process.

This paper is organized as follows. We first review the missing-feature method in Section 2. We then describe the proposed approaches in Sections 3 and 4. Representative experimental procedures and results are presented and discussed in Section 5. Finally, in Section 6, we summarize our findings.

2. Overview of the missing-feature method

2.1. Mask estimation

The missing-feature approach requires that we determine a "mask" which classifies the spectrum into reliable and unreliable ("missing") regions for missing-feature reconstruction. Seltzer *et al.* [2] proposed a Bayesian classifier for the mask estimation, employing speech features which make no assumption about the corrupting noise signal. In the spirit of their work, we designed a Bayesian classifier for mask estimation using the following several features.

2.1.1. Comb filter ratio (CFR)

In voiced speech, the ratio of the energy at harmonics of the fundamental pitch to the energy at the spectral valleys (at intervening frequencies) can be a reasonable measure of noise in the presence in speech signal. To calculate the energy at the harmonics, we used a comb filter represented by the following transfer function.

$$H_{comb}(z) = z^{-p} / \left(1 - g z^{-p}\right) \tag{1}$$

where p and g are the pitch period and a parameter that controls the sharpness of the filter shape respectively. For energy outside the harmonic frequencies, a shifted version of the comb filter was used. We can obtain the CFR for every Mel-filter-bank output using the filtered speech signal.

2.1.2. Subband energy to fullband energy ratio

The ratio of subband energy to fullband energy can indicate how much a particular band of speech is corrupted by the background noise. Generally, this ratio decreases as the signal becomes more corrupted.

2.1.3. *Subband energy to fullband/subband noise floor ratio*

The ratio of energy to a floor value of noise is another measure of noise corruption. To estimate the noise floor, we used the minimum-statistics method which is originally designed for spectral subtraction [4].

2.1.4. *Flatness*

Spectral shape along adjacent frames and subbands also indicates the amount of corrupting noise. We define spectral "flatness" as the average difference between energy in a particular frame and spectral band and in neighboring frames and bands.

2.2. Missing-feature reconstruction

The cluster-based and correlation-based methods have been proposed previously by Raj *et al.* [3]. They restore unreliable parts of speech representations using the known distributions of speech sounds and the reliable regions as indicated by the masks obtained using Bayesian detection and estimation. In this paper, we employ the cluster-based reconstruction method.

The distributions of the log spectra of clean speech are modeled by Gaussian mixture densities with K clusters. Consider, for example, a noisy speech vector $S(t)$ with unreliable (*i.e.* missing) components $S_m(t)$ and reliable components $S_o(t)$. The cluster membership k of $S(t)$ is nominally determined by its *a posteriori* probability. $S(t)$ has unreliable elements, and these elements must be obtained by integrating them out:

$$\begin{aligned} \hat{k}_{S(t)} &= \arg\max_k \{P(S(t) \mid k) P(k)\} \\ &= \arg\max_k \left\{ P(k) \int_{-\infty}^{Y_m(t)} P(S(t) \mid k) dS_m(t) \right\} \end{aligned} \quad (2)$$

where $Y_m(t)$ represents 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, the Gaussian model of the cluster as determined by (2), and the upper bound of $Y_m(t)$.

$$\hat{S}_m(t) = \arg\max_{S_m} \left\{ P(S_m(t) \mid S_0(t), \mu_{\hat{k}_{S(t)}}, \Sigma_{\hat{k}_{S(t)}}, S_m(t) \le Y_m(t)) \right\} \quad (3)$$

3. Training using colored-noise combinations

In the work of Seltzer *et al.*, white noise was used for training the classifier for mask estimation in order to reduce the effects of unknown noise [2][5]. The acoustic models for the mask estimation were trained using a speech database that was corrupted by white noise. The trained classifier was then applied to factory noise and music noise conditions without any prior information about the test conditions. Observed performance was comparable to the matched training conditions where the types of maskers used in training and testing the Bayesian mask estimator are identical.

Our own results obtained using matched training conditions as well as multi-condition training produced comparable results to those of Seltzer *et al.* Nevertheless, we found that the use of white-noise backgrounds to train the Bayesian mask classifier failed to provide robust recognition accuracy that remained good in the presence of all types of corrupting noise.

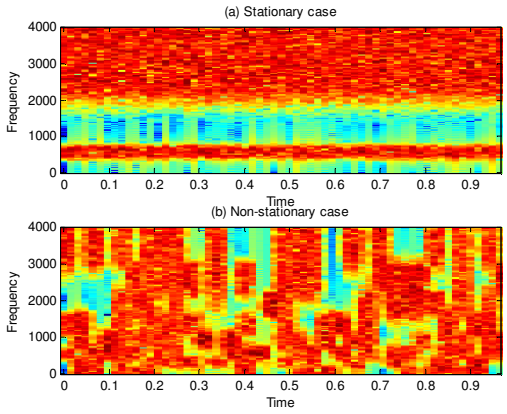

Figure 1: Spectrogram of 8-subband colored noise.

We believe that training the Bayesian mask classifier on white noise does not provide good recognition accuracy over all types of noise (at least in our experiments) because realistic noise signals do not corrupt every spectral band of the speech signal evenly. White noise should be predictive of results for other types of noise if the mask estimates obtained at every subband were in fact totally independent from those in adjacent frames and other subbands, utilizing such corrupting property of white noise for the purpose of an environment-independent model looks reasonable.

Unfortunately, independent mask estimates across frames and spectral bands are not generally obtained in practice. For example, the first subband energy feature (Sec. 2.1.2) uses a value for the noise floor that is typically estimated from the neighboring frames. In addition, the flatness feature (Sec. 2.1.4) directly exploits the spectral differences compared to the neighboring frames and subbands.

In other words, we believe that the spectral variations across adjacent frames and subbands can influence the features obtained from a particular subband. Therefore, in order to obtain environment-independent models for the Bayesian mask estimator, we must incorporate the spectral variations across frames and subbands into each subband model, which in effect simulates the occurrences of various kinds of noise conditions.

In this paper, we propose a training method that uses combinations of colored noises for the purpose of generating an environment-independent model that can be used for mask estimation. We incorporate the effects of spectral variation across adjacent frames and subbands by training the acoustic models for mask estimation on speech databases that have been corrupted by various random combinations of colored noise.

The colored noise samples are obtained by first partitioning the entire spectrum into N adjacent frequency bands that increase in bandwidth as the center frequency increases according to the Mel scale. A set of 10^{th}-order Butterworth bandpass filters is realized with center frequencies and bandwidths corresponding to the N frequency bands described above. A particular colored-noise sample is obtained by passing white noise through a subset of the

Butterworth filters that is chosen at random. For a particular colored noise sample, a new random selection of bandpass filters is obtained every 30 ms, 60 ms, 300 ms, or never. (The first three durations between changes in filter selection are intended to represent non-stationary noise conditions, while the last condition is intended to represent stationary noise.) Figure 1 shows examples of spectrograms obtained from typical noise samples with 8 subbands. The combination of subbands actually used changes every 30 ms in the lower panel and never in the upper panel.

4. Voiced-frame restoration

Many features used for mask estimation in voiced frames (such as comb filter ratio) are based on pitch information. In noisy environments, however, the initial pitch extraction process can cause some voiced frames to be misidentified as unvoiced frames. For these frames, the classifier designed for unvoiced speech would normally be used for mask estimation even though the frame is voiced, which leads to incorrect results.

In this section, we describe a procedure that reconsiders the nature of frames that had initially been determined to be unvoiced and that correctly identifies some of these frames as voiced. The proposed restoration method uses the identical Bayesian classifier that is used for mask estimation. Since frames where pitch is not detected have a high probability of being unvoiced speech or non-speech, we reclassify an "unvoiced" frame as "voiced" only in the strict condition that the lowest likelihood computed from the voiced model is greater than the highest likelihood obtained from the unvoiced model.

More specifically, the voiced-frame restoration is accomplished by estimating conditional probabilities on a frequency-by-frequency basis. In each frequency band that is used to extract log spectral coefficients in conventional Mel-frequency analysis, four probabilities are computed:

$$P_1 = P[\text{reliable band} \mid \text{voiced frame}]$$
$$P_2 = P[\text{unreliable band} \mid \text{voiced frame}]$$
$$P_3 = P[\text{reliable band} \mid \text{unvoiced frame}]$$
$$P_4 = P[\text{unreliable band} \mid \text{unvoiced frame}]$$

In each frequency band, an overall probability of the frame being voiced is obtained from the *minimum* of P_1 and P_2 above, and the overall probability of the frame being unvoiced is obtained from the *maximum* of P_3 and P_4. The overall probabilities of being voiced and being unvoiced are multiplied together across all frequency bands to produce the final probability from which the ultimate voiced/unvoiced decision is made. Once a given frame is determined to be voiced speech, the mask estimation proceeds on that basis.

5. Experimental results

We evaluated the performance of the procedures described in the previous two sections using the Aurora 2.0 evaluation procedure. The Aurora 2.0 procedure uses 23 Mel-filterbanks for feature extraction, which means that the cluster-based missing-feature method was applied using 23 log-spectral coefficients per analysis frame.

The recognizer was trained using the standard Aurora 2.0 training database that contains 8,440 utterances of clean speech. A testing database of degraded speech was obtained by combining clean speech from the Aurora 2.0 database with four types of noise samples: white noise, car noise, speech babble, and background music. The white noise and car noise represent stationary noise conditions, and they were obtained from the NOISEX92 and Aurora 2.0, respectively. Speech babble and background music stand for the non-stationary environments; they were obtained from the Aurora 2.0 and CDs of Mozart (KV545, K265, K525, K622, and KV618). The test database included speech that was corrupted by each of the four types of noise at 5 kinds of SNRs, (20, 15, 10, 5, and 0 dB). Each of these 20 noise conditions is represented by 1,001 samples of degraded speech.

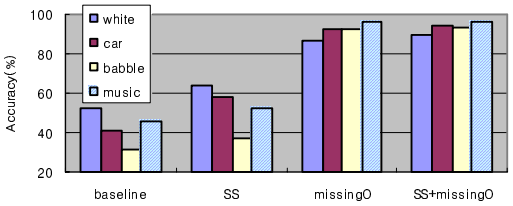

Figure 2: Accuracy of baseline system at 5-dB SNR.

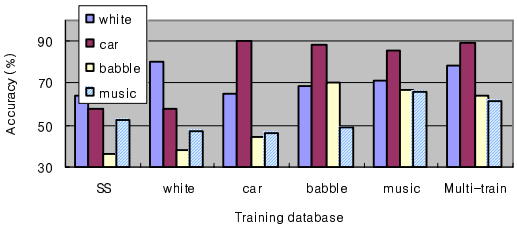

Figure 3: Accuracy using environment-specific masks.

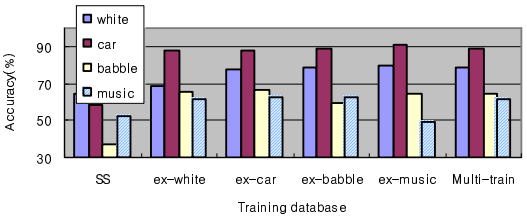

Figure 4: Accuracy using masks derived from exclusive multi-condition training.

We first examined the baseline system's accuracy at 5 dB SNR, as shown in Figure 2[1]. Results are compared using no noise compensation ("baseline"), spectral subtraction ("SS"), cluster-based missing-feature restoration using masks derived from Oracle knowledge ("missingO"), and a combination of the latter two methods. 16-component Gaussian mixtures with full covariance matrices were used for modeling the log spectral coefficients. Best performance was obtained by combining spectral subtraction with missing-feature restoration. This combination was used for all remaining

[1] Word accuracy in these experiments is 100% minus the NIST word error rate, which includes a penalty for insertions.

experiments in this paper, with the spectral subtraction preceding the missing-feature algorithms.

The remainder of the experiments used missing-feature recognition based on blind mask estimation. The Bayesian mask estimator was trained using noise that was presented at seven SNRs (clean, 20, 15, 10, 5, 0 and -5 dB). Different classifiers were designed for mask estimation of voiced and unvoiced speech frames, where each frame could in general include both reliable and unreliable frequency bands. The five features described in Section 2 were used for mask estimation in voiced frames and the same features except for CFR were used for the unvoiced frames. We used a pitch-detection algorithm based on histograms [5]. The mask-estimation features are modeled by 16 Gaussian mixtures with diagonal covariance. The performance of a given mask estimation method is inferred by the recognition accuracy that it provides.

Figure 3 describes the recognition accuracy obtained at 5-dB SNR when masker was trained on a single type of noise. Unsurprisingly, greatest accuracy was obtained when testing conditions matched the type of noise with which the mask was trained. Multi-style training of the mask estimator ("multi-train") appears to provide comparable accuracy. Recognition accuracy was fairly high when car noise was used for the testing data under most conditions. We believe that this is a consequence of the predominantly low-frequency spectral profile of car noise, which happens to be similar to the average spectrum of babble and music noise as well. We also measured the recognition accuracy was trained in multi-style fashion, but excluding the specific noise used in the training data Figure 4. Except for car speech, this training causes accuracy to degrade compared to results in Figure 3 obtained when training and testing conditions are matched.

From the pilot works above, we suggest that the database used to train the mask estimator should reflect the spectral variations that occur in the test conditions. While multi-condition training could be a good solution, it may not be effective in totally unknown environments.

Figure 5 shows the recognition accuracy obtained using the colored-noise mask training method for mask estimation described in Sec. 3. Recognition accuracy is plotted as a function of the number of subbands of colored noise used for training. It appears that training on colored noise with a small number of bands is best when testing with white noise and car noise, while training on colored noise with 8-12 subbands is best when testing in speech babble or music noise. We believe that these differences reflect the fact that the spectra of white noise and car noise tend to be smoother than those of speech babble or music noise. The proposed training method provides accuracy that is comparable to that obtained in the "ex-multi" condition, and it has the advantage that it is not necessary to collect noise samples under multiple conditions.

The plots in Figure 6 show evaluation results for the entire test set as a function of SNR. "Mask1" refers to results obtained using masks trained on colored noise with 12 subbands, while "Mask2" refers to results obtained when the voiced-frame restoration method (Sec. 4) is implemented as well. Except for babble noise at low SNRs, the restoration method for voiced frames provides significant improvement, and in stationary noise cases such as white noise and car noise, the relative improvement can be as great as 12.3%. These results show that the restoration method is effective in estimating the correct mask by restoring the voiced frames when pitch-detection fails due to noise corruption.

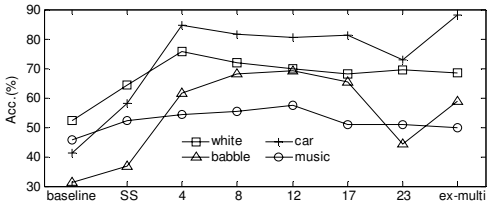

Figure 5: Performance as the number of subbands for combinations of colored noise used in training.

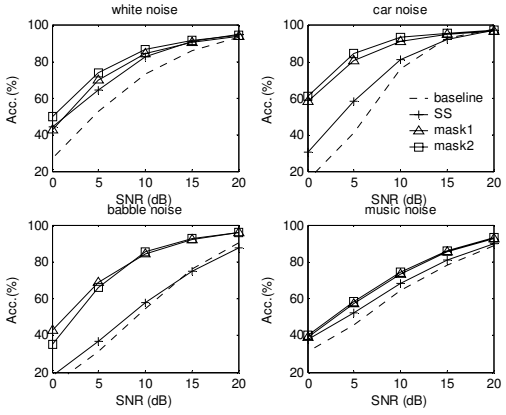

Figure 6: Accuracy over the entire test set.

6. Conclusions

In this paper, we have described an effective method of environment-independent mask estimation for missing-feature algorithms. The proposed method employs model training using a combination of colored noise and a decision step for restoration of voiced frames. The experimental results show the proposed training method is useful for the environment-independent mask estimation and the decision procedure for voiced frame is effective to accurately estimate the mask.

7. Acknowledgements

This work was supported by the Post-doctoral Fellowship Program of Korea Science & Engineering Foundation (KOSEF).

8. References

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

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

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

[4] Martin, R., "Spectral subtraction based on minimum statistics," *EUSIPCO-94*, pp.1182-1185, 1994.

[5] Seltzer, M. L., *Automatic Detection of Corrupted Speech Features for Robust Speech Recognition,* M.S. thesis, Carnegie Mellon University, 2000.