

SMOOTHED SPECTRAL SUBTRACTION FOR A FREQUENCY-WEIGHTED HMM IN NOISY SPEECH RECOGNITION

Hiroshi Matsumoto and Noboru Naitoh

Dept. of Electrical & Electronic Eng., Faculty of Engineering, Shinshu University
500 Wakasato, Nagano-shi, Nagano 380, Japan

ABSTRACT

This paper proposes improved methods of smoothed spectral subtraction to enhance the recognition performance of a frequency-weighted HMM (HMM-FW) in very noisy environments. The conventional spectral subtraction tends to produce discontinuity in estimated power spectra. This distortion is undesirable for HMM-FW which uses group delay spectra as feature vectors. In order to remove this distortion, this paper proposes two frequency smoothing methods in log-spectral domain: (1) a low-pass filtering by DCT, and (2) a weighted minimum mean square error method (WMSE) which fits cosine series to an estimated log-power spectrum. The results show that the smoothers are very effective under very noisy conditions, especially for the frequency-weighted HMM. The WMSE method combined with HMM-FW achieves the highest recognition accuracies, for instance, improving recognition rate from 68% to 88% at -6dB SNR of car noise.

1. INTRODUCTION

The approaches to noise robust speech recognition are broadly classified into speech enhancement and robust pattern matching. In HMM-based pattern matching, adaptation methods such as PMC [1] are very successful at moderate noise levels. However, under severe noisy condition, the discrimination between classes may be reduced due to the variance of the signal spectra. Another approach is robust HMMs [2],[3]. As such an HMM, we previously proposed a frequency-weighted HMM [3]. This HMM has been proved to be robust to additive noise over a wide range of SNR due to the use of both the group-delay spectra and the fixed covariances derived from the frequency-weighting coefficients. While this HMM is not robust to very noisy conditions, it can be combined with speech enhancement techniques to achieve high robustness.

Among speech enhancement techniques, the spectral subtraction (SS) has been proved to be effective and efficient noise reduction method, especially for filter-bank-based speech recognizers [4],[5]. However, since the spectral subtraction independently processes each spectral component, the estimated spectra tend to have discontinuity in low SNR

frequency regions. Thus, this distortion may affect the performance of HMM-FW since it utilizes group-delay spectra.

In order to remove this distortion, this paper proposes two frequency smoothing methods in log-spectral domain: (1) a low-pass filtering by discrete cosine transform (DCT), and (2) a weighted minimum mean square error method (WMSE) which fits cosine series to an estimated log-power spectrum. These smoothers intend to interpolate low SNR components from high SNR ones.

These smoothing methods are combined with speech recognizer based on a grand variance HMM [6] as well as HMM-FW, and are evaluated through word recognition tests using the NOISEX-92 database down-sampled to 8 kHz. For three types of noises, the performances of SS with or without the proposed smoothers are also compared with nonlinear spectral subtraction [7].

2. FREQUENCY-WEIGHTED HMM

In this study, we use a p -channel filter bank in speech analysis. In the frequency-weighted HMM [3], the pseudo-group-delay spectrum x^d is used as an observation vector to utilize its robustness to noises [8],[9]. x^d is defined by the inverse DCT of frequency-weighted cepstral coefficients:

$$x^d = C \text{diag}[1, 2, \dots, p] x^c, \quad (1)$$

where C represents the $(p \times p)$ discrete cosine transform matrix.

In the frequency-weighted HMM all the covariances are replaced to the inverse of a frequency-weighting matrix $\Sigma_F = \text{diag}[w_1, w_2, \dots, w_p]$. The frequency-weighting coefficients w_k is the smoothed and compressed power spectrum, and are derived from the grand mean μ_G^d of the pseudo-group-delay spectra as follows:

$$w_k = \eta \exp\{\nu \tilde{\mu}_k^l\}, \quad (2)$$

where $\tilde{\mu}_k^l$ is the k th component of the smoothed log power spectrum given by

$$\tilde{\mu}^l = C^T L C \mu_G^d \quad (3)$$

with $L = \text{diag}[1, 1/2, \dots, 1/q, 0, \dots, 0]$, and ν is a compression factor and η a scale factor [10]. Therefore, the covariance in this model is not estimated statistically, but instead is derived based on a prior knowledge on the perceptual importance in frequency domain and/or expected variance due to degradation of speech.

3. SMOOTHING FOR SPECTRAL SUBTRACTION

3.1 Spectral Subtraction

In a standard spectral subtraction, the clean speech power spectrum $\hat{x} = [\hat{x}_1, \hat{x}_2, \dots, \hat{x}_p]^T$ is estimated from the noisy speech power spectrum $x = [x_1, x_2, \dots, x_p]^T$ by

$$\hat{x}_k = \max \{x_k - \alpha \hat{N}_k, \gamma \hat{N}_k\}, \quad (4)$$

where \hat{N}_k is the estimated noise power spectrum in the k -th channel, and α the overestimation factor, and $\gamma \hat{N}_k$ the spectral floor[5]. This spectral subtraction doesn't take account of the correlation between neighboring spectral components. Therefore, the estimated spectra tend to have discontinuity in low SNR components due to both flooring and variation of noise. This distortion may affect the performance of HMM-FW since the spectral discontinuities produce spectral peaks in pseudo-group-delay spectra.

3.2 Spectral Smoothing

To suppress this distortion caused by spectral subtraction, the low SNR components of the estimated log-power spectrum $\hat{x}^l = \log \hat{x}$ are required to be smoothed while preserving the high SNR components. We propose a new estimate of log-power spectrum \hat{x}^l defined by a weighted sum of the non-smoothed and smoothed log-power spectra, \hat{x}^i and \hat{x}^s , as follows:

$$\hat{x}^l = W \hat{x}^s + (I - W) \hat{x}^i, \quad (5)$$

$$W = \text{diag}[W_1, \dots, W_p]. \quad (6)$$

The weighting coefficient W_k is a monotonically increasing function of the estimated SNR and is bounded between 0 and 1. Therefore, the low SNR components of \hat{x}^l are replaced by those of the smoothed log-power spectrum \hat{x}^s derived from \hat{x}^i . In this study, we choose the transfer function of the estimated Wiener filter as such weighting coefficients:

$$W_k = \max \left\{ \frac{x_k - \beta \hat{N}_k}{x_k}, W_{min} \right\}, \quad (7)$$

where β is the overestimation factor to estimate the Wiener filter, and W_{min} is a value close to zero. β is not always the same as α in equation (4).

The smoothed log-power spectrum \hat{x}^s is given by DCT of

the $(r \times 1)$ liftered cepstral vector \bar{x}_r^c , which will be described later, as follows:

$$\bar{x}^l = C_r \bar{x}_r^c, \quad (8)$$

where C_r represents the first r columns of the matrix C . The number of cepstral components for liftering, r , is determined depending on the number of reliable spectral components in \hat{x} . In this study, r is set to the number of the coefficients W_k 's which are greater than a threshold θ . Thus, the smoothness is controlled by the smoothing parameters θ and β . \bar{x}_r^c is estimated based on one of the following two methods.

(1) The DCT method

In this method, the vector \bar{x}_r^c is simply given by the lower r components of the following cepstral vector \bar{x}^c :

$$\bar{x}^c = \frac{1}{p} C^T \log \hat{x}. \quad (9)$$

(2) The weighted MSE method (WMSE)

Whereas in the DCT method every spectral component of \hat{x} equally contributes to \bar{x}^c , this method takes the reliability of each spectral component into account using W . The unknown cepstral vector \bar{x}_r^c is estimated to minimize

$$J(\bar{x}_r^c) = (\hat{x}^l - \bar{x}_r^c)^T W (\hat{x}^l - \bar{x}_r^c). \quad (10)$$

The estimated \bar{x}_r^c is given by the normal equation,

$$C_r^T W C_r \bar{x}_r^c = p C_r^T W \hat{x}^l. \quad (11)$$

Therefore, unlike the DCT method, the cepstral vector \bar{x}_r^c in this method is altered depending on the values of β and θ even if r is held unchanged.

4. EVALUATION

4.1 Database and Speech Analysis

In speech analysis, a 15-channel uniform filter bank system with a flat composite spectrum was implemented at a sampling rate of 8kHz using the same design procedure as described by Dautrich, Rabiner, and Martine [11]. The spacing of channel center frequencies and the band width of each bandpass filter were set to 250Hz (8kHz/32) and 300Hz, respectively. First, the speech signal from the NOISEX-92 database was down-sampled to 8kHz and preemphasized with $(1 - 0.98z^{-1})$. The output of each bandpass filter was followed by a square law detector and a moving average filter of 12 points. The channel outputs were sampled at every 10ms.

Three stationary back ground noises were used: car, white, and Lynx helicopter noises. The degraded speech by car and Lynx noises were used from the NOISEX-92 database. On the other hand, the white noise was generated in a computer, and was added to clean speech so that the global SNR for

each word is equal to a predetermined value.

Each of the HMM models was trained using 10 repetitions of noise-free samples. Another set of ten utterances was used for testing. The Vitarbi algorithm was used for testing. The beginning and end points were fixed to those in the label files. Thus, only substitution errors were scored.

4.2 Baseline System

In the evaluations, the recognition system uses a grand variance HMM (HMM-GT) [6] as well as the frequency-weighted HMM for comparative experiments. In HMM-GT all the covariances for every word model are fixed to the 'grand diagonal covariance' Σ_G^d over all the training speech. HMM-GT for each word was also used as the initial model for HMM-FW. The structure of both HMM-FW and HMM-GT is a left-to-right model of 26 states with a single Gaussian component.

Before recognition experiments, the parameters involved in the base-line system were optimized. As to the frequency-weighted HMM, the normalized scale $\bar{\eta}$, which is defined by $\{|W^{-1}|/|\Sigma_G^d|\}^{1/p}$, was set to 100 for all the noise conditions, and the compression factor ν in equation (2) is set to 1.0 for white noise and to 0.0 for car and Lynx noises [10].

As to spectral subtraction, on the basis of the preliminary experiments, the overestimation factor α and the flooring parameter γ in equation (4) were set to 2 and 0.001, respectively, over all the noise conditions. The noise estimate \hat{N}_k was obtained by averaging noise spectra over 30 frames preceding each word. In addition to the standard spectral subtraction, the nonlinear spectral subtraction (NSS) was implemented to compare with the smoothed spectral subtraction proposed here.

4.3 Effect of Smoothing

First, the effects of the smoothing parameters β and θ were examined using W_{min} . Figure 1 shows the recognition performance of the HMM-FW based system with three values of β as a function of the smoothing parameter θ at two SNRs for each type of noises. In these plots, the value of $\theta = 0$ implies no smoothing, and thus corresponds to the conventional SS. From this figure, the following general trends emerge. In both methods, the maximum recognition score for each β tends to increase with larger value of β for the noises whose spectra have steeper falling slope, such as for car noise. As to θ , the larger the value of β , the smaller the optimum value of θ to give the highest score. In the DCT method, the optimum values of θ are mostly smaller than those in the WMSE method, and thus the smoother is less effective in severe SNR conditions, especially for white noise. In the HMM-GT based system, the above trends were also found. Table 1 summarizes the globally optimal values of β and θ for each type of noises and for both smoothing methods, though their optimal values are not always consistent across SNRs and the types of HMMs.

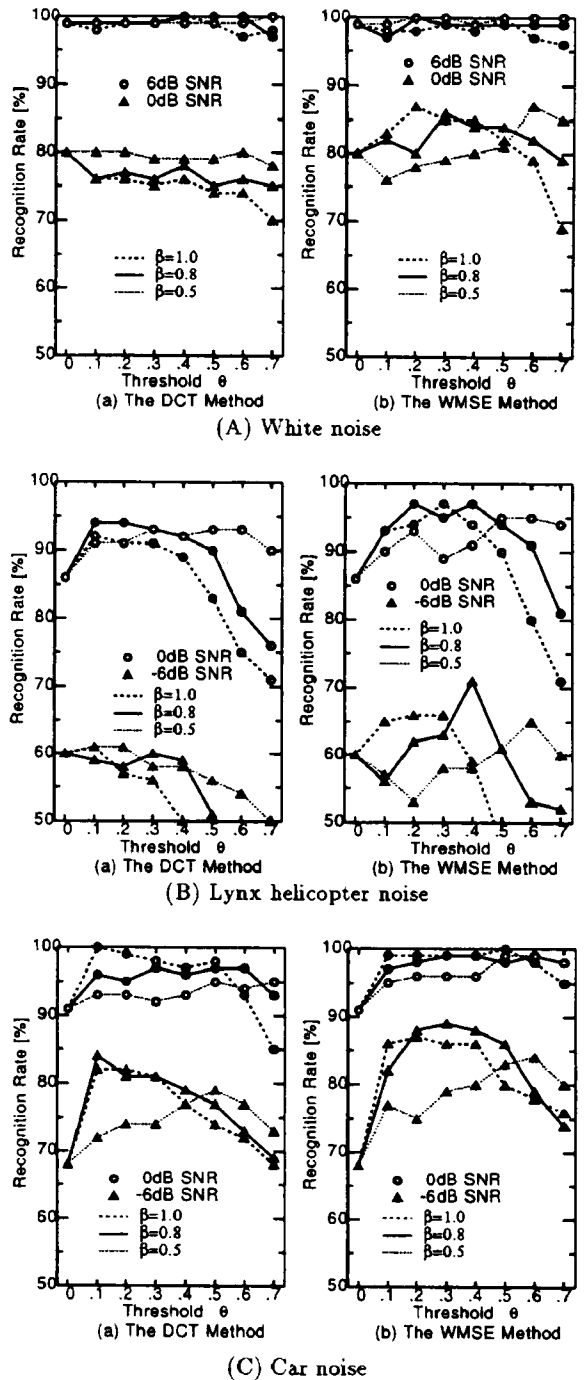


Figure 1: Effects of the smoothing parameters θ and β in both DCT and WMSE smoothers on the performance of the HMM-FW based recognizer at two SNR conditions of white, Lynx, and car noises.

