AUDITORY-MODELING INSPIRED METHODS OF FEATURE EXTRACTION FOR ROBUST AUTOMATIC SPEECH RECOGNITION

Zhinian Jing and Mark Hasegawa-Johnson

Department of Electrical and Computer Engineering,
University of Illinois, Urbana, IL 61801, USA

ABSTRACT

This paper proposes a technique of extracting robust feature vectors for ASR. The technique is inspired by work related to auditory modeling. It involves first filtering the speech signal through a bank of band-pass filters, which are based on a model of the human cochlea. Autocorrelation functions (ACF) are computed on the filters' outputs. Then the individual ACFs are scaled by their corresponding voice indices (VIs), which use information related to the pitch. A summed ACF is then obtained by summing the individual ACFs across the bands. Feature vectors are then computed using standard cepstral analysis, by treating the summed ACF as a regular ACF. Finally, frame indices (FIs) weigh the feature vectors in the time domain. The effectiveness of the proposed techniques, compared to LPCC and MFCC, are demonstrated by comparing the results obtained from simple recognition experiments.

INTRODUCTION

Human beings perform significantly better than today's best ASR systems in noisy environment. Thus, perfect imitation of human speech recognition (HSR) processes would certainly improve ASR, and it seems reasonable to assume that selective imitation of some HSR processes (like auditory processing) might also improve ASR. In fact, imitation of certain HSR properties has already achieved wide use in ASR. Examples include mel-scale frequency warping [1] and RASTA modulation filtering [2]. Ghitza [2] proposed the EIH model that imitates firing patterns of auditory nerve fibers.

The goal of this research is to draw upon the knowledge of auditory modeling in search of a method that generates intrinsically more robust feature vectors. However, no attempts are made to invent a completely new paradigm. Instead, short-time spectral analysis is still the basis, and attempts are made to improve the spectral analysis such that it is more robust to common types of noise.

In essence, the proposed technique examines the speech signal in frequency domain, and weights the frequency bands according to the amount of speech information available in them. This method is based on the simple observation that additive noise affects the frequency bands differently. The frequency bands with relatively high SNR should be emphasized in computing the feature vectors.

METHOD DESCRIPTION

An overview of the system is given in the block diagram of Figure 1.

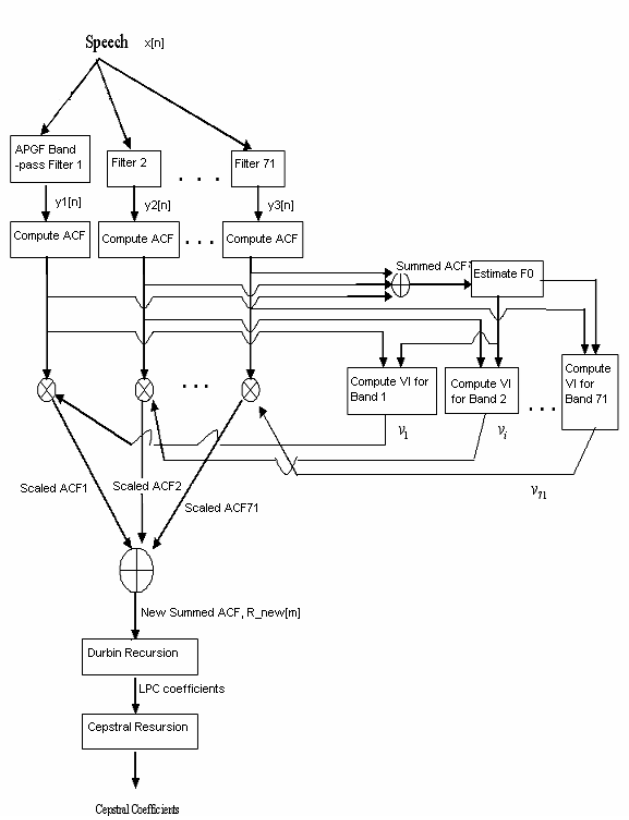


Figure 1. Block diagram of the proposed Voice Index method of extracting ASR features.

As with LPCC and MFCC, input to the system is a digitized speech signal, which is framed and windowed in

30 ms blocks with 10 ms increments, typical of short-time speech processing. Then each speech block is filtered through a filter bank of overlapping band-pass filters. The pre-emphasis stage may be achieved with the filter gains of the particular filter bank used. The filters do not broaden in response to signal level changes, but otherwise are designed to accurately simulate the frequency selectivity properties of the human cochlea. The filters are implemented with the all-pole gamma-tone filters (APGFs), which are derived by discarding zeros from the well-known gamma-tone filters [3]. The filters' parameters are designed after Robert and eriksson's model, [4]. In [4], 120 filters are used to include characteristic frequencies (CFs) up to 20 kHz, which corresponds to 71 filters for CFs from about 100 Hz to 4 kHz. Figure 2(a) plots the magnitude response of the individual filters. Figure 2(b) shows the total root-sum-square response of the filters is smooth and has some attenuation at low frequencies.

After the filtering stage, an ACF is computed for each and every filtered signal,

$$r_i[m] = \sum_{n=1}^{N} y_i[n] \cdot y_i[n+m], \; m \in \{100,250\}. \quad (1)$$

Then a summed ACF is obtained by summing the individual ACFs,

$$R[m] = \sum_{i=1}^{71} r_i[m]. \quad (2)$$

It can be easily shown that the Fourier transform of R[m]

$$R[m] \xleftrightarrow{Fourier} |X(w)|^2 \cdot \sum_i |H_i(w)|^2, \quad (3)$$

where X(w) is the Fourier transform of the input speech signal, and $H_i(w)$ is the frequency response of ith filter.

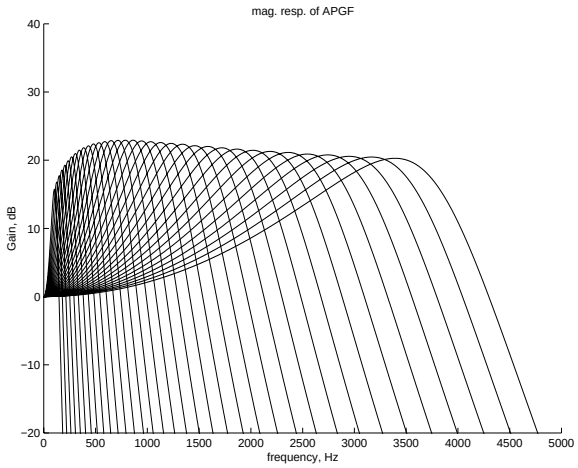
Referring to Figure 2(b), it is obvious that R[m] is a valid autocorrelation function of the input speech.

VOICE INDEX

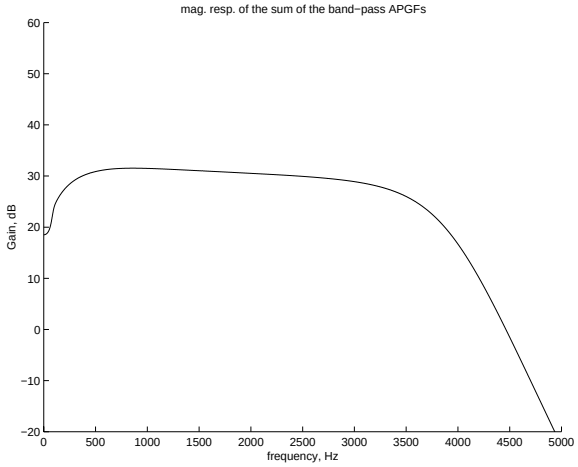
The goal of the VI approach is to weigh each frequency band with a "voice index" parameter, VI, based on the amount of speech information present in that band. The weighed ACFs are added together to produce a summed ACF, $R_v[m]$, that has a better speech signal to noise ratio than the standard $R[m]$:

$$R_v[m] = \sum_{i=1}^{71} v_i \cdot r_i[m], \quad (4)$$

$$R_v[m] \xleftrightarrow{Fourier} |X(w)|^2 \cdot \sum_i v_i \cdot |H_i(w)|^2, \quad (5)$$



(A)



(B)

Figure 2. (a) Magnitude response of the filter bank (every other filter is shown). (b) Total root-sum-square response of the filter bank.

For this idea of spectral weighing to work, one has to find an effective method of computing the VI. In this paper, the role of the fundamental frequency (F0) or pitch is exploited to solve for the VI. The VI of the ith band v_i is computed as follows. First a summed ACF, $R[m]$, is computed by summing the unweighted individual ACFs $r_i[m]$, as in Equation (2). Then F0 is estimated from $R[m]$ by

$$F0 = f_s \cdot \frac{1}{M}, \quad (6)$$

where $M = \arg\max_m \{R[m]\}$, for $m = 100..250$ and f_s is the sampling frequency. This method of computing pitch is shown to be valid in [5] and [6]. Several ways of

computing v_i are explored. One particularly simple and effective approach is

$$v_i = \frac{r_i[M]}{r_i[0]}. \quad (7)$$

Equation (7) is inspired by the work in [6], where the presence of a local maximum at $r_i[M]$ determines whether the particular band is part of a set of ACFs belonging to a particular vowel.

FRAME INDEX

The FI method refers to a weighing scheme that modifies the time domain computation of distance between a pair of utterances. Let $D(S,T)$ denote the total distance between a test utterance and a template utterance, $D(S,T) = \sum_{j=1}^{N} d_j$, where d_j denotes the distance between jth pair of frames. With the FI, denoted by f_j, the total distance becomes

$$D(S,T) = \sum_{j=1}^{N} d_j \cdot f_j . \quad (8)$$

The idea of the FI is very simple. Over the duration of a noise corrupted speech utterance, the SNR varies frame by frame. During a strongly voiced frame, the SNR is much higher than the overall SNR of the entire utterance, so the spectrum is not much affected by noise. On the other hand, during a silent or an unvoiced frame, the spectrum could be almost entirely due to noise. Experiments show the frame SNR could range anywhere from $-\infty$ to around 40 dB, for a digit embedded in white noise with a global SNR of 0 dB. The FI simply places more emphasis on the frames that are less corrupted by noise.

Computing the FI without prior knowledge of the noise spectrum is not easy. For best recognition performance, the FI should be monotonically related to the SNR in each frame. If the background noise is slowly varying, an estimate of the noise power can be computed during non-speech frames, and used during speech frames to compute local SNR. In mobile user environments, however, the background noise level may vary as rapidly as the speech signal, so that quasi-stationary assumptions are not appropriate.

Here, the roles of F0 and the summed ACF are again exploited. Of the several methods explored, the FI computed with Equation (9) seems to work well in white and babble noise. For each frame, FI is taken to be the ratio of R at the pitch delay and R at the zeroth delay:

$$FI = R[M]/R[0]. \quad (9)$$

TEST SYSTEM

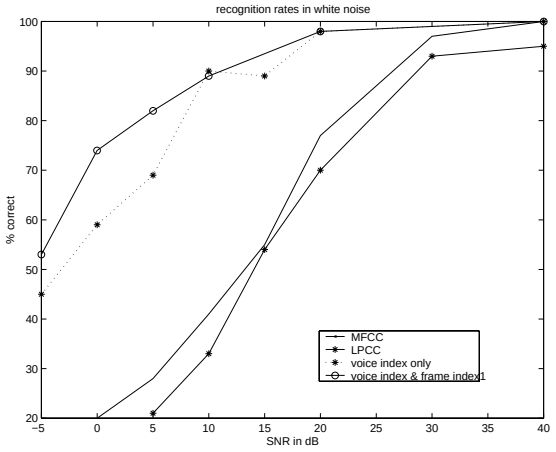
A simple recognition system is implemented to test the merit of the proposed methods. The system is a speaker independent, small vocabulary, word recognition system based on linear time warping (LTW). The results obtained through LTW are almost identical to those obtained from dynamic time warping (DTW). LTW is used because of its simplicity and computational efficiency. The distance measure used is the Euclidean distance between cepstral coefficient vectors. Since LPCC, MFCC, and the proposed methods all compute cepstral coefficients as the feature vectors, the comparison is done in a straightforward manner.

The test cases and the templates are taken from the TIDIGIT database. There are 11 different words, digits 0-9 and "o". The utterances are manually cut. The templates are created from clean speech, by speakers who are different from the speakers of the test utterances. .

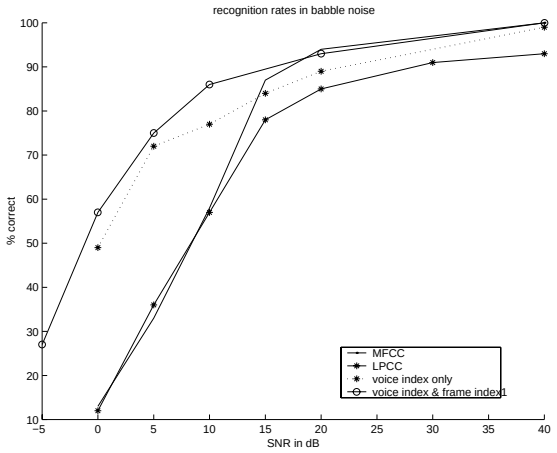
RESULTS

Figure 3 demonstrates the performance of the VI method versus that of MFCC and LPCC. Improvement in recognition rate is particularly striking in white noise. With other types of background noise, the VI method generally shows an advantage over MFCC as well, though usually in a lesser degree.

A related result is shown in Figure 4, where the VI method is compared to a "hard voice index". The "hard" VI method differs from the regular VI method in that only binary values are possible for the VI, namely, 0 and 1. The VI is 1 if it is above the average voice index, and 0 otherwise. Many previously published computational auditory scene analysis algorithms perform a binary allocation of each band to either signal or background [8]. Figure 4 suggests that a probabilistic or fuzzy auditory scene analysis algorithm may be more useful for ASR than a binary "hard-decision" method.



(A)



(B)

Figure 3. Recognition scores using the VI and the FI, in varies types of noise: (a) white noise, (b) babble noise.

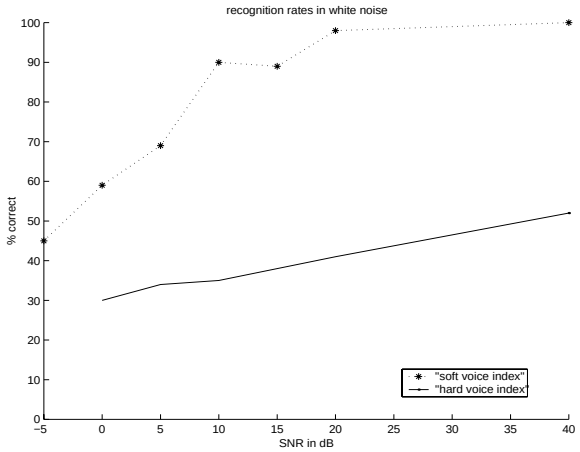


Figure 4. Recognition scores with the VI method and the "hard" VI method.

CONCLUSIONS

The VI method improves isolated digit recognition accuracy, especially under very nosy conditions. At 0dB SNR in white noise backgrounds, the FI and VI methods combined with linear time warping yield 73% correct digit recognition accuracy. The FI and VI methods do not require the noise to be either slowly varying or known in advance. The only assumption is that the noise does not have a prominent speech-like F0, or that if it does, F0 of the speech signal is more prominent than F0 of the noise in every frame. The VI method may be combined with other techniques at other stages of the recognizer to further improve the overall performance. In fact, it might be an interesting experiment to combine the VI method with RASTA processing in the log domain. The resulting method can expect to achieve robustness to both additive and convolutional noises.

REFERENCES

[1] S. Davis and P Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences," *IEEE Trans. Acoust. Speech and Audio Processing*, vol. 28, no 4, pp. 357-366, 1980.

[2] H. Hermansky and N. Morgan, "RASTA processing of Speech," *IEEE Trans. Speech Audio Processing*, vol. 2, no 4, pp. 587-589, 1994.

[3] O. Ghitza, "Auditory Models and Human Performance in Tasks Related to Speech," *IEEE Trans. Speech Audio Processing*, vol. 2, no 1, pp. 115-132, January 1994.

[4] R.F. Lyon, "The all-pole gammatone filter and auditory model," *Forum Acusticum*, Antwerp, Belgium, April 1996.

[5] A. Robert and J.L. Eriksson, "A Composite Model of the Auditory Periphery for Simulating Responses to Complex Sound," *J. Acoust. Soc. Am.*, vol. 106, pp.1852-1864, 1999.

[6] R. Meddis and M. J. Hewitt, "Virtual pitch and phase sensitivity of a computer model of the auditory periphery. I: Pitch identification," *J. Acoust. Soc. Am.*, vol. 89, no. 6, pp. 2866-2882, June 1991.

[7] R. Meddis and M. J. Hewitt, "Modeling the identification of concurrent vowels with different fundamental frequencies," *J. Acoust. Soc. Am.*, vol. 91, no. 1, pp. 233-245, January 1992.

[8] Y. Gong, "Speech recognition in noisy environments: A survey," *Speech Communication*, vol. 16, pp. 261-291, 1995.

[9] R. Meddis and M. J. Hewitt, "Modeling the identification of concurrent vowels with different fundamental frequencies," *J. Acoust. Soc. Am.*, vol. 91, no. 1, pp. 233-245, January 1992.