VARYING TIME CONSTANTS AND GAIN ADAPTATION IN FEATURE EXTRACTION FOR SPEECH PROCESSING

David V. Anderson, Sourabh Ravindran

Center for Signal and Image Processing
Georgia Institute of Technology
Atlanta, GA 30332

Malcolm Slaney

Yahoo! Research
701 First Avenue
Sunnyvale, CA 94089

ABSTRACT

Previously we showed that band-pass filtered MFCC-like features are useful for noise robust speech discrimination and recognition. In this paper we aim to improve the previously presented features by incorporating varying time constants and gain adaptation in each frequency channel. We show that varying the time constants leads to a representation that is less prone to the effects of noise. Further, we show that gain adaptation can not only provide better performance in clean condition but can also be used to improve the noise robustness of the features. These improvements come at a very small increase in computational cost. Speech discrimination and recognition results are presented.

Index Terms— Noise robust, speech processing, adaptive gain, varying time constants.

1. INTRODUCTION

Mel-frequency cepstral coefficients (MFCCs), although very useful for speech and audio processing in clean conditions, are not very robust to noise. We take a brief look at some possible reasons of this behavior. In most audio feature extraction processes the number of samples used to represent each frame is small compared to the original sampled waveform. Given that there will be loss of information in building a compact representation of the audio signal, the key to generating better representations is to discard information that is least significant. In case of MFCCs, the FFT followed by grouping into critical bands using triangular filters lead to discarding of information that is not easily quantifiable. The temporal information in the signal is distributed in the magnitude and phase of the multiple frequency bins and combining them could lead to the masking of pertinent information.

The MFCC front-end, due to its dependence on block processing and combination of frequency bins, gives a representation that has low time and frequency resolution. In the human auditory system the asymmetrical shape of the cochlear filters allow for good time resolution (due to its gradual roll-off on the low frequency side) and good frequency resolution (due to the sharp cutoff on the high frequency side) [1]. But even without the asymmetrical shape, band-pass filtering is desirable since it avoids the windowing effects due to block processing and provides better temporal resolution compared to the short-time Fourier transform (wherein temporal resolution is restricted by the size of the analysis window and frame rate). Also, the use of triangular filters for critical band filtering leads to large changes in gain for small changes in the frequency [2]. Thus the energy estimation in each channel is smoother if frequency decomposition is performed using exponentially spaced band-pass filters and the signal strength in each channel is estimated using an envelope detector (implemented using a rectifier and a low-pass filter). Low-pass filtering before downsampling ensures that there is no temporal aliasing. The low-pass filter does not discard perceptually relevent information since we know that the central auditory neurons cannot respond to very fast temporal modulations [3]. The fast temporal variations that are smoothed out are most likely perceptually insignificant. Further, envelope extraction following band-pass filtering allows us the opportunity to filter out the noise modulations in each channel to some extent.

Noise robust auditory features (NRAF) [4], which are based on a model of the human auditory system [3], introduce improvements to MFCC that address the above mentioned issues without adding significant computational costs [4]. In this paper we propose further modifications to the feature extraction process that make it more robust for speech processing. It is shown that varying the time constants in each channel and adaptive gain compression of the envelope in each channel lead to improved noise robustness. The new feature extraction process is shown in Fig. 1. The use of a half wave rectifier (HWR) for envelope detection has a physiological grounding (see [3]) but it also justified from a signal processing perspective since the use of a HWR avoids pitch doubling in lower frequency channels [1].

2. EXPERIMENTAL SETUP

The improvements afforded by the modifications are evaluated for a speech versus non-speech classification task at various signal-to-noise ratios (SNRs) and a connected digit recog-

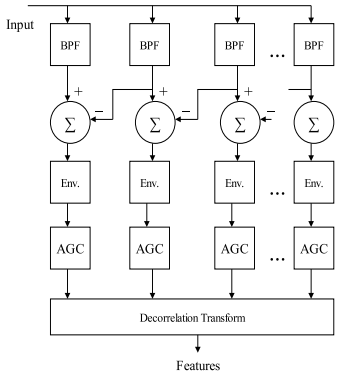

Fig. 1. Block diagram showing the feature extraction process. Band-pass filtering is followed by spatial derivative, an envelope detector (implemented as a half-wave rectifier and low-pass filter), an automatic gain control stage and a decorrelation transform.

nition task. Mel-frequency cepstral coefficients (MFCCs) are used as the baseline. For the speech versus non-speech classification, each audio segment was divided into frames of length 25.625 msec with a frame rate of 100 Hz. Twelve MFCC's were extracted from each frame. Thirteen linearly spaced and twenty-seven log spaced triangular filters were used to group the FFT bins. The lowest frequency was 133.33 Hz, the linear spacing was 66.66 Hz and the log spacing was 1.07. In extracting the features we followed the Sphinx III specifications [5]. For the NRAF features, forty fourth-order band-pass filters (spanning the same frequency range as MFCCs) were used. The BPFs are approximately one-seventh octave, constant *Q* filters. The filters had to be chosen to be approximately one-seventh octave to match the number of triangular filters used for the standard MFCC features. The first thirteen coefficients (of the DCT) were used to perform the classification. A Gaussian mixture model (GMM) based classifier was used to predict the log-likelihood of each frame belonging to a particular class. The log-likelihoods of all frames in a segment belonging to each class were added to make the final decision. Post-processing consisted of mean subtracting and variance normalizing the features from each one second segment [6].

For the speech recognition task, MFCCs were extracted using the HTK toolkit front-end [7], 23 channels were used. Thirteen MFCC coefficients (including the zeroth coefficient) were mean and variance normalized (MVN) [6] and delta and acceleration features were computed to form a 39-dimensional feature vector. The NRAF features were also extracted in a similar way. Thirty-two one-sixth octave filters were used for the BPF implementation. Delta and acceleration coefficients were extracted from the MVN processed static features. The zeroth coefficient was used since it is shown to respond better to MVN than using the log energy. Logarithmic compression was used for both feature sets.

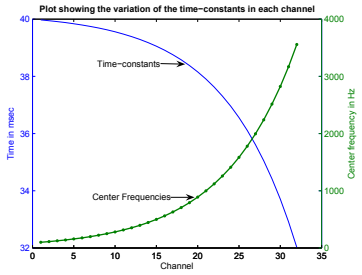

Fig. 2. Figure depicting the change in the time constants of the LPF in each frequency channel. The center frequency of each channel is also shown.

3. VARYING TIME CONSTANTS

In the auditory model proposed by Yang et. al [3] the low-pass filters in each channel model the inability of neurons in the central auditory system to respond to fast temporal fluctuations. The time constants for all the low-pass filters were the same. However, from a signal processing perspective the combination of the rectifier and the low-pass filter is an envelope detector and it is desirable to have envelope detectors with varying time constants in different frequency channels. Moreover, we know that the human auditory system has better temporal resolution at higher frequency and higher frequency resolution at lower frequencies. In order to mimic this behavior the time constants (in msec) are set according to,

$$tc(i) = \frac{k_1}{f_s}\left(\frac{f_s}{2} - f_c(i)\right) + k_2$$

where $f_c(i)$ is the center frequency of the i^{th} channel and k_1 and k_2 are parameters used to set the range of time constants. For the speech recognition tests we set k_1 = 18.4 and k_2 = 31. Figure 2 shows the variation of the time constants with center frequency. The range of the time constants was chosen empirically and is not the optimal operating point. There is a tradeoff between a shorter time constant, which gives better temporal resolution and better performance at low noise levels, and longer time constants that give better noise robustness. Choosing the time-constants based on the ambient SNR would provide the best overall improvement (i.e. in all SNR conditions). But this requires the use of a noise estimation algorithm. Further, it should be noted that the time constants also depend on the type of audio. In a physiological system the time constants for speech are restricted both by the production and the hearing mechanism, while for music or noise this may not be the case.

The improved noise robustness can be explained as follows. Let the noisy speech signal be represented as,

$$x(t) = s(t) + n(t)$$

where $s(t)$ is the speech signal and $n(t)$ is the additive noise. Assuming an acoustic signal can be expressed as,

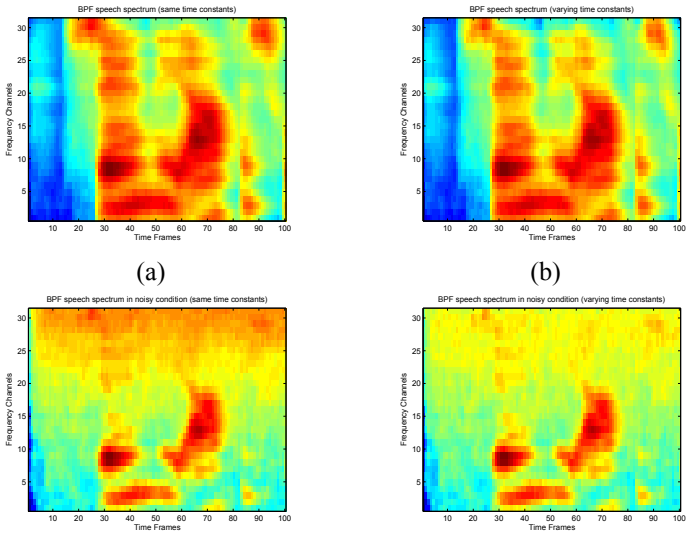

Fig. 3. Speech spectrum in clean condition with a) same time constant in each channel b) varying time constants in each channel. Spectrum in noisy condition with c) same time constant in each channel d) varying time constants in each channel. As is clear, varying time constants help reduce the effect on noise on the speech spectrum, especially in high frequency channels.

$$s(t) = \sum_i e_{s_i}(t) v_i(t)$$

where $v_i(t)$ is the carrier signal and $e_{s_i}(t)$ is the modulating signal in the i^{th} channel. From Fig. 1 we see that for the i^{th} channel, output after the spatial derivative is,

$$(s_i(t) + n_i(t)) - (s_{i+1}(t) + n_{i+1}(t))$$

or,

$$(e_{s_i}(t) * v_i(t) - e_{s_{i+1}(t)} * v_{i+1}(t)) + (n_i(t) - n_{i+1}(t))$$

The output after the peak detector is

$$(e_{s_i}(t) - e_{s_{i+1}}(t)) + (e_{n_i}(t) - e_{n_{i+1}}(t))$$

where $e_{n_i}(t)$ is the noise modulation in the i^{th} channel. If we assume that the noise spectrum is approximately flat, the noise term is dominated by the signal term. But in the general, it is possible to adjust time constants in each channel to selectively extract the speech modulation and hence weed out the noise component, making the representation more robust to noise. Figure 3 shows the advantage of using varying time constants. As can be seen from Figs. 3(a) and 3(b), in clean conditions the two spectrums look the same. However in the presence of noise, varying the time constants to suit the speech modulation masks the noise to some extent leading to a better representation (see Figs. 3(c) and 3(d)). The features extracted with varying time constants are referred to as NRAF-TC.

4. ADAPTIVE GAIN COMPRESSION (AGC)

A further advantage of processing the speech signal using BPF and envelope detectors is that since we already have the envelope in each channel, adaptive gain compression can be performed with very little additional computation. Anderson et al. [8], [9] showed the usefulness of gain adaptation in hearing aids. We follow the same approach in non-linearly compressing the envelope in each channel. The relationship between the non-linearly compressed envelope and the original envelope can be expressed as,

$$\hat{e}_{s_i}(t) = \beta e^{\alpha}_{s_i}(t)$$

or,

$$log\hat{e}_{s_i}(t) = \alpha log e_{s_i}(t) + log\beta$$

where $e_{s_i}(t)$ is the original envelope and $\hat{e}_{s_i}(t)$ is the compressed envelope. α and β are computed based on the desired range of compressed envelope. Features with adaptive gain compression are referred to as NRAF-AGC. The improvement afforded by AGC was evaluated for the speech versus non-speech classification task and speech recognition on a subset of the TIDIGITS database. White noise was synthetically added to generate the various SNRs. It should be noted that NRAF-AGC was obtained by incorporating AGC in the NRAF-TC implementation.

5. RESULTS

5.1. Speech Versus Non-speech Classification

Figure 4 shows the relative performance of the NRAF, NRAF-TC and NRAF-AGC features with respect to the baseline feature (MFCC). As expected, NRAF-TC improves the noise-robustness of NRAF. Incorporating gain adaptation in the NRAF-TC feature further boosts the performance.

5.2. Speech Recognition

The speech recognition results for the Aurora 2 task in clean training condition are presented in Fig. 5. We used 12 components per mixture for the silence model and 6 components for every other state. NRAFs demonstrate a clear advantage over MFCCs in noisy conditions. Varying time constants further improves the performance. Gain adaptation gives an improvement in low SNR conditions at the expense of slight deterioration in the high SNR case. On the other hand by choosing the parameters differently the performance in clean and high SNR conditions can be improved at the cost of slight degradation in low SNR conditions (see Table 1).

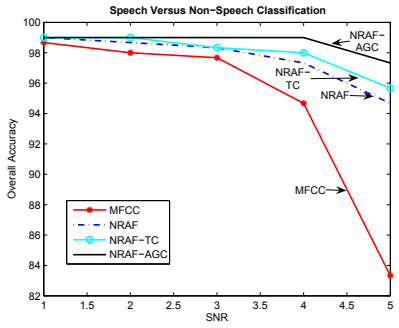

Fig. 4. Figure showing the comparative performance of MFCC, NRAF, NRAF-TC and NRAF-AGC for the speech versus non-speech classification task. Different SNRs were obtained by synthetically adding pink noise. Root compression was used for all the features.

Recognition results (training in clean condition)				
	NRAF (no AGC)	NRAF (with AGC, K=0.05)	NRAF (with AGC, K=0.01)	NRAF (with AGC, K=1.5)
Clean	99.51 %	99.48 %	99.42 %	99.54 %
20 dB	97.73 %	98.13 %	98.10 %	97.67 %
15 dB	95.73 %	96.50 %	96.56 %	95.61 %
10 dB	90.76 %	92.39 %	92.54 %	90.70 %
5 dB	79.71 %	83.02 %	83.79 %	79.09 %
0 dB	59.69 %	64.54 %	65.67 %	58.21 %
-5 dB	37.80 %	41.51 %	42.19 %	37.21 %

Table 1. Effect of AGC on the noise robustness of features. White noise was synthetically added to obtain different SNRs. K determines the range of the compressed envelope.

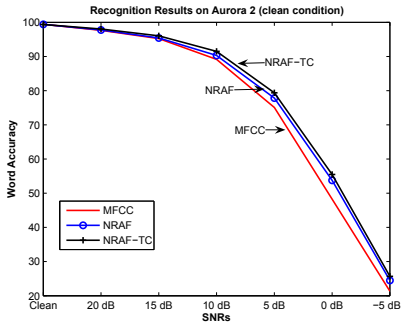

Fig. 5. Figure showing the performace of NRAF and MFCC on Aurora 2 task. A six component mixture was used for each state and silence was modeled using a 12 component mixture.

6. CONCLUSIONS

Noise-robust auditory features address some of the issues leading to the poor noise robustness of MFCCs. This paper presents further modifications to the NRAF features that improve its performance for speech processing tasks. Varying the time constant comes at no extra cost while gain adaptation adds a small overhead. It appears that adapting the time constants and gain based on the SNR would further improve the results. Initial results in this direction are encouraging.

7. REFERENCES

[1] Richard Lyon, "A computational model of filtering, detection, and compression in the cochlea," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Paris, France, May 1982.

[2] Steven Beet, "Email contribution to auditory mailing list, Nov, 2004. http://www.auditory.org/postings/2004 /833.html," .

[3] Xiaowei Yang, Kuansan Wang, and Shihab Shamma, "Auditory representations of acoustic signals," *IEEE Transactions on Information Theory*, vol. 38, no. 2, pp. 824–839, March 1992.

[4] Sourabh Ravindran, David V. Anderson, and Malcolm Slaney, "Improving the noise robustness of mel-frequency cepstral coefficients for speech processing," in *Proceedings of the ISCA Workshop on Statistical and Perceptual Audition*, Pittsburgh, PA, 2006.

[5] Michael Seltzer, "Sphinx III signal processing front end specification http://cmusphinx.sourceforge.net/ sphinx3/s3_fe_spec.pdf," .

[6] C.-P Chen, K. Filali, and J. A. Bilmes, "Frontend post-processing and backend model enhancement on the aurora 2.0/3.0 databases," in *International Conference on Speech and Language Processing*, 2002, pp. 241–244.

[7] ETSI ES 201 108 v1.1.3 (2003-09), "Speech Processing, Transmission and Quality Aspects (STQ); Distributed speech recognition; Front-end feature extraction algorithm; Compression algorithms," .

[8] David V. Anderson, "Model based development of a hearing aid," in *M.S. Thesis, Brigham Young University*, Provo, Utah, 1994.

[9] Douglas M. Chabries, David V. Anderson, Thomas G. Stockham Jr., and Richard W. Christiansen, "Application of a human auditory model to loudness perception and hearing," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, May 1995.