Submitted to ICASSP 2006.
Confidential until published.

IMPROVING THE NOISE-ROBUSTNESS OF MEL-FREQUENCY CEPSTRAL COEFFICIENTS FOR SPEECH DISCRIMINATION

Sourabh Ravindran, David V. Anderson

School of Electrical and Computer Engineering
Georgia Institute of Technology
Atlanta GA 30332
Email: stg@ece.gatech.edu, dva@ece.gatech.edu

Malcolm Slaney

Yahoo! Research
701 First Avenue
Sunnyvale, CA 94089
Email: malcolm@ieee.org

ABSTRACT

In this paper we study the noise-robustness of mel-frequency cepstral coefficients (MFCCs) and explore ways to improve their performance in noisy conditions. Improvements based on a more accurate model of the early auditory system are suggested to make the MFCC features more robust to noise while preserving their class discrimination ability. A speech discrimination task is chosen to evaluate the performance gains afforded by the modifications. This simple task allows us to study the features in isolation from complex systems required for more challenging tasks, such as speech recognition in noisy conditions.

1. INTRODUCTION

MFCC's are very useful features for audio processing in clean conditions. However, performance using MFCC features deteriorates in the presence of noise. There has been an increased effort in recent times to find new features that are more noise robust compared to MFCCs. Features such as, spectro-temporal modulation features [1] are more robust to noise but are computationally expensive. Skowronski and Harris [2] suggested modification of MFCC that uses the known relationship between center frequency and critical bandwidth from human psychoacoustics to decouple filter bandwidth from filter spacing. They also studied the effects of wider filter bandwidth on noise robustness. Herein, we suggest different modifications to MFCCs that make it more robust to noise without adding prohibitive computational costs.

MFCC features approximate the frequency decomposition along the basilar membrane by a short-time Fourier Transform. The auditory critical bands are modelled using triangular filters, compression is expressed as a log function and a discrete cosine transform (DCT) is used to decorrelate the features [3].

In this paper we cite two reasons for the poor noise performance of MFCCs. First, block processing with Fourier transform and the use of triangular filters for grouping of frequency bins into critical bands results in a signal in each channel that is not as smooth as that obtained with band-pass filters. Second, MFCC's poor noise performance can be attributed to log compression [4], [5], [6]. The large negative excursions of the log function for values close to zero leads to a splattering of energy after the DCT whereas root compression (expressed as $(\cdot)^{\alpha}$, with $0 < \alpha < 1$) followed by the DCT leads to better compaction of energy, as is shown later.

The rest of the paper is organized as follows, Section 2 explains the experimental setup. Sections 3, 4 and 5 talk about methods to improve noise-robustness of MFCCs. Section 6 uses an information theoretic measure of clustering to support the results obtained in earlier sections, followed by conclusions.

2. EXPERIMENTAL SETUP

In order to study the noise-robustness of MFCCs, we use a simple task that isolates the feature performance from the system performance (for instance, in a task like speech recognition there are too many parameters that might affect the outcome of the experiments). Hence, a speech discrimination task [1] at various signal-to-noise ratios (SNR) was chosen to evaluate the features. The audio database was built from five publicly available corpora. Speech samples were taken from TIMIT acoustic-phonetic continuous speech corpus. The training set consisted of 300 examples from TIMIT's training subset and test set consisted of 150 different sentences spoken by 50 different speakers (25 male, 25 female) from TIMIT's test subset. Sentences and speakers in the training and test sets are different. The non-speech class consisted of 450 examples (300 for training and 150 for testing) which included animal vocalization from BBC Sound Effects audio CD collection, music samples from RWC Genre Database [7] and environmental sounds from Noisex and Aurora databases. Segments of one second duration were selected for training and testing. The task consisted of predicting whether a given one second test segment belongs to the speech or the non-speech class. Pink noise was synthetically added to generate different SNRs.

Each audio segment was divided into frames of length 25.625 msec with a frame rate of 100 Hz. Forty 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 chosen to be 133.33 Hz, a linear spacing of 66.66 Hz and log spacing of 1.07 were used. In extracting the features we followed the Sphinx III specifications [8]. For the band-pass filter implementation of MFCC, forty fourth-order bandpass-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. In all experiments the first thirteen coefficients were used to perform the classification. A Gaussian mixture model 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 the two classes where added to make the final decision. Features from each one second segment were mean subtracted and variance normalized [9].

3. AMPLITUDE COMPRESSION

Previous work [4], [5], [6] shows that root compression is better than logarithmic compression for noise robustness. In this section we revalidate this result and try to explain the effect of different amplitude compressions in terms of a discrimination measure. Figure 1 shows the performance of the classifier using MFCC features with root and log compression. We see that root compression is much more robust to noise as compared to log compression. Root function is better behaved as opposed to the log function. The log function gives large negative values for inputs close to zero and this leads to a spreading of the energy in the transform domain (after DCT).

A simple experiment was devised to show that root compression followed by DCT leads to better compaction of energy. The envelope of a speech segment was amplitude compressed and transformed using DCT. Varying number of transformed coefficients were used to reconstruct the amplitude compressed signal and the reconstruction error was calculated. Figure 2 shows the plot of reconstruction error versus the number of coefficients used for the reconstruction, for both log and root compression. It is clear that root compression followed by DCT leads to better compaction of energy since the reconstruction error using fewer coefficients is much lower as compared to the log case.

Another interesting aspect of amplitude compression is the trade-off between performance in clean conditions and robustness to noise. The trade-off can be better understood in terms of between-class and within-class scatter. We define between-class scatter as the distance between the mean of the two clusters and within-class scatter as the mean of the distances of each data point from the mean of its cluster. The ratio of between-class scatter and within-class scatter is used as a measure of discrimination ability. More compression leads to lower within-class scatter, the between-class scatter is also lower but since there is no noise to confuse the classifier the accuracy is higher. In noisy conditions more compression leads to more errors due to the reduced between-class scatter (the reduction in within-class scatter is not able to offset the reduction in between-class scatter). Table 1 shows the effect of log and different degrees of root compression on the discrimination ability of the MFCC features. As is clear from the table, more compression leads to better performance in clean but this adversely affects the performance in noisy conditions. In the sections that follow, root compression refers to the use of a compression factor, $\alpha = 0.3$.

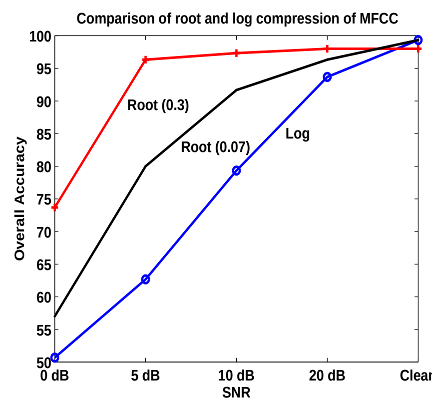

Fig. 1. Figure showing the performance of the classifier using MFCC features with log and root compression for a speech discrimination task. Different SNRs were generated by adding pink noise synthetically.

Measure of discrimination ability for log and various degrees of root compression		
Compression	Clean	Noisy
Log	0.4364	0.2372
Root ($\alpha = 0.07$)	0.4182	0.2387
Root ($\alpha = 0.3$)	0.3645	0.2473
Root ($\alpha = 0.7$)	0.3196	0.2719

Table 1. Table showing that more compression yields greater discrimination in clean conditions. However, in noisy conditions less compression yields better class discrimination.

4. ALIASING AND SMOOTHING

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 some 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

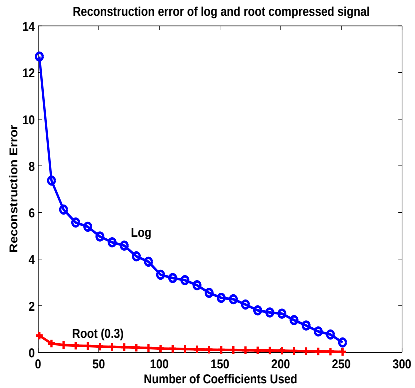

Fig. 2. Figure showing that root compression followed by DCT leads to better compaction of energy. Reconstruction error is plotted as a function of number of coefficients used for the reconstruction.

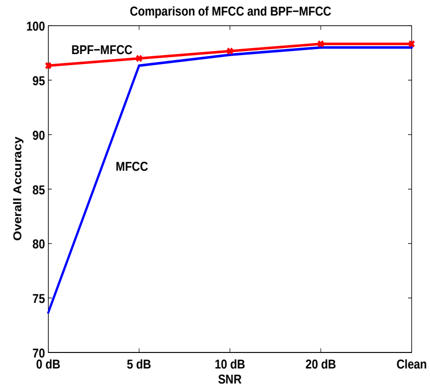

Fig. 3. Figure showing the comparative performance of MFCC and BPF-MFCC for the speech discrimination task. Different SNRs were obtained by synthetically adding pink noise. Root compression was used for both features.

MFCCs, the FFT and triangular filters lead to discarding of information that is not exactly quantifiable. We also know that due to the sharp peaks of the triangular filters MFCCs are sensitive to small changes in the frequency [10]. 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 a peak detector (implemented using a rectifier and a low-pass filter). Features extracted from a smooth representation are not perturbed by perceptually irrelevant variations in the signal. We know that central auditory neurons cannot respond to very fast temporal modulations [11] and hence smoothing over 10–20ms does not discard perceptually relevant features. The high frequency components that are smoothed out are most likely perceptually insignificant. By low-pass filtering the signal and then down sampling, we are discarding information in a more intelligent way. We refer to features obtained in this manner as BPF-MFCC [12]. Figure 3 shows the improvement in performance of MFCC for the speech discrimination task by removal of FFT and triangular filters. The BPF-MFCC representation degrades more gracefully with falling SNR.

5. SPATIAL DERIVATIVE

In this section we show that the performance of BPF-MFCC in clean conditions can be further improved by introducing yet another processing stage which is directly motivated by physiological processing. The fourth-order BPFs used for the frequency decomposition are not very sharp and result in some amount of frequency spreading across the channels. This spreading can be limited by the use of a spatial derivative (implemented as a difference operation between adjacent channels). The spatial derivative is used to model the lateral inhibitory network in the cochlear nucleus [11]. We refer to BPF-MFCC with spatial derivative as noise-robust auditory features (NRAF). Apart from limiting the frequency spreading by sharpening the filter response, the spatial derivative stage, in clean and high SNR conditions, enhances the contrast across the spectral profile. This is a form of edge detection operation common in image processing, although the effect in audio is less dramatic due to lack of abrupt changes across frequency channels.

The comparison of MFCC, BPF-MFCC and NRAF is shown in Fig. 4. As predicted the performance of NRAF in clean and moderate SNR cases is better than that of BPF-MFCC, but the performance in high noise case (0 dB SNR) is lower. This can be explained as follows, in very low SNR cases where the noise variance is equal to or greater than the signal variance, the spatial derivative results in some loss of signal component due to subtraction, i.e. the difference operation removes some signal component from channels whose adjacent higher channels are noisy.

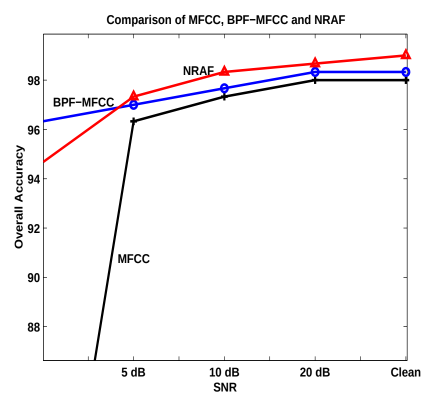

Fig. 4. Figure showing the comparative performance of MFCC, BPF-MFCC and NRAF for the speech discrimination task. Different SNRs were obtained by synthetically adding pink noise. Root compression was used for all the features.

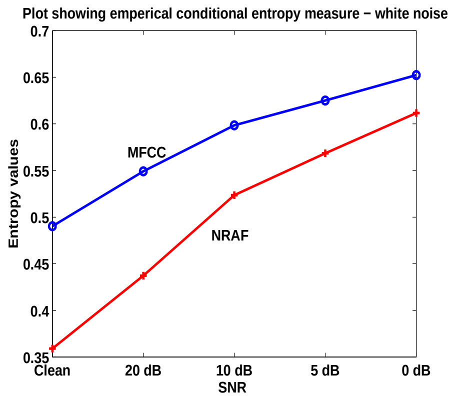

Fig. 5. Figure showing the empirical conditional entropy measures for MFCC and NRAF for a two-class, two-cluster case. It is seen that NRAFs are better features than MFCCs for speech discrimination. White noise was synthetically added to generate different SNRs.

6. INFORMATION-THEORETIC CLUSTERING VALIDITY MEASURE

In this section we use an information theoretic measure of clustering to substantiate the fact that NRAFs are better than the original MFCCs not only in terms of noise-robustness but also in terms of class discrimination ability. Conditional entropy is used as a criterion for evaluating the clustering validity of clustering algorithms [13]. By using a very "naive" clustering algorithm the clustering properties of the underlying attributes can be studied. Mahalanobis distance from the mean of the two clusters is used as the clustering algorithm to study the effect of synthetically added noise on the clustering properties of MFCC and NRAF.

Given a set of class labels $c \epsilon C$ and clusters $k \epsilon K$, conditional entropy, $H(C|K)$ is approximated by empirical conditional entropy, $H^e(C|K)$ given by,

$$H^e(C|K) = -\sum_{c=1}^{|C|}\sum_{k=1}^{|K|}\frac{h(c,k)}{n}log\frac{h(c,k)}{h(k)}$$

where, $h(c,k)$ is the number of examples of class c assigned to cluster k, n is the total number of examples and $h(k)$ is the associated marginal. Conditional entropy gives the number of bits required to encode all the class information given we know the clusters $\{k_i\}$ and the model $\Pi = \{h(c,k)\}$. A lower value of conditional entropy indicates that the cluster labels are good indicators of the class labels. For the case of two classes and two clusters, the empirical conditional entropy measure is shown in Fig. 5. As is evident, NRAF has a lower value of conditional entropy, implying that it clusters better than MFCCs in clean and noisy conditions.

7. CONCLUSIONS AND FUTURE WORK

This paper justifies three improvements to MFCC features that are motivated by a more accurate auditory model. We tested these changes using a simple speech discrimination task. Replacing the log compression with a root compression improves the noise-robustness of MFCCs. Low-pass filtering the signal in each channel before decimation avoids aliasing and leads to a smoother signal envelope in each channel. Further, the benefit of spatial derivative in clean and high to moderate SNR cases is demonstrated. Future work would involve investigating the preliminary results that show that noise-robustness is also a function of the amount of smoothing performed. Further, the BPF-based implementation of MFCC allows the use of different time constants for the LPF in different channels, hopefully leading to an even better representation. Future work would also include using the modified MFCC features for a speech recognition task.

8. REFERENCES

[1] Nima Mesgarani, Shihab Shamma, and Malcolm Slaney, "Speech discrimination based on multiscale spectro-temporal modulations," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Montreal, Canada, May 2004.

[2] Mark D. Skowronski and John G. Harris, "Exploiting independent filter bandwidth of human factor cepstral coefficients in automatic speech recognition," *Journal of Acoustical Society of America*, vol. 116, no. 3, pp. 1774–1780, sept 2004.

[3] M.J. Hunt, M. Lenning, and P. Mermelstein, "Experiments in syllable-based recognition of continuous speech," in *IEEE International Conference on Accoustics, Speech, and Signal Processing*, Denver, CO, Apr. 1980.

[4] J.S. Lim, "Spectral root homomorphic deconvolution system," *IEEE Trans. ASSP*, vol. 27, no. 3, pp. 223–233, 1979.

[5] P. Alexandre and P. Lockwood, "Root cepstral analysis: A unified view," *Speech Communication*, vol. 3, pp. 277–288, 1993.

[6] Ruhi Sarikaya and John H.L. Hansen, "Analysis of the root-cepstrum for acoustic modeling and fast decoding in speech recognition," in *Eurospeech*, Aalborg, Denmark, 2001.

[7] Masataka Goto, Hiroki Hashiguchi, Takuichi Nishimura, and Ryuichi Oka, "RWC music database: Music genre database and musical instrument sound database," in *Proceedings of the 4th International Conference on Music Information Retrieval*, Oct. 2003, pp. 229–230.

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

[9] 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.

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

[11] 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.

[12] Paul Smith, Matt Kucic, Rich Ellis, Paul Hasler, and David V. Anderson, "Cepstrum frequency encoding in analog floating-gate circuitry," in *Proceedings of the IEEE International Symposium on Circuits and Systems*, Phoenix, AZ, May 2002, vol. IV, pp. 671–674.

[13] Byron E. Dom, "An information-theoretic external cluster-validity measure," in *IBM Research Report RJ 10219*, May 2001.