LOW-POWER AUDIO CLASSIFICATION FOR UBIQUITOUS SENSOR NETWORKS

Sourabh Ravindran, David Anderson

Georgia Institute of Technology
School of Electrical and Computer Engineering
Atlanta, GA 30332

Malcolm Slaney

IBM Almaden Research Center
650 Harry Road
San Jose, CA 95120

ABSTRACT

In the past researchers have proposed a variety of features that are based on the human auditory system. However none of these features have been able to replace mel-frequency cepstral coefficients (MFCCs) as the preferred feature for audio classification problems, either because of computational costs involved or because of their poor performance in the presence of noise. In this paper we present new features derived from a model of the early auditory system. We compare the performance of the new features with MFCC in a four-class audio classification problem and show that they perform better. We also test the noise robustness of the new features in a two-way audio classification problem and show that it outperforms the MFCCs. Further, these new features can be implemented in low-power analog VLSI circuitry making them ideal for low-power sensor networks.

1. INTRODUCTION

In order to implement signal processing algorithms for audio classification and auditory scene analysis in hand-held and remote-sensing devices it is important that a power efficient method be deviced to extract features that could be used for further signal processing. Researchers [1], [2] have shown the feasibility of implementing the MFCCs [3] in low-power analog VLSI circuitry. However MFCCs do not perform as well in the presence of noise. In this paper we present *noise-robust auditory features* (NRAF) as a viable alternative to MFCCs. NRAF can be implemented in low-power analog VLSI circuitry with a cost comparable to that of implementing MFCCs and provide better performance in the presence of noise.

Of late there has been a lot of interest in fabricating and utilizing miniature, low-power, and intelligent sensor elements and arrays. A low-power analog VLSI front-end for audio classification can be interfaced with such systems to provide end-to-end acoustic surveillance. The low-power feature of the front-end would allow it to be integrated with a networked array of autonomous sensors that can then be deployed in the field [4]. Such a low-power audio classifier is especially important as a front-end to an autonomous sensor that uses a small battery for power and is expected to monitor its environment for months at a time. The rest of the sensor's processing circuits need only be powered when the right kinds of sounds are present. Another important use for a low-power audio classification is in hearing aids where auditory scene analysis can be performed to automatically switch between different hearing aid algorithms based on the current environment. Such a front-end would also be very useful in the design of smart microphones that have sophisticated capabilities beyond that of passive sound reception.

The rest of the paper is organized as follows, section 2 explains the features used for the audio classification task. Section 3 talks about the experimental setup, followed by results and conclusions.

2. FEATURES

The NRAF features are derived from a model of the early auditory system [5]. The input signal is passed through a bandpass filter bank. The signal is then non-linearly compressed followed by a difference with the adjacent channels. This is followed by a half-wave rectification and smoothing filter. The half-wave rectification followed by the smoothing is in some sense a peak detector. The output at this stage is referred to as the auditory spectrum [5]. Figures 3 and 4 show the auditory spectrum and the spectrogram for a noisy speech input. It can be seen that the auditory spectrogram filters out some of the noise that appears in the normal spectrogram. The auditory spectrum is robust to noise for two reasons: the difference operation between adjacent channels reduces the effects of stationary noise, and a non-linear property known as phase locking emphasizes the signal. We compute a discrete cosine transform (DCT) of the logarithm of the output of the smoothing filter to obtain the NRAFs. The filter bank consists of 128 filters tuned from 180 Hz to 7246 Hz. The smoothing (temporal integration) is done over 8 msec. Thus we have 128 features for every 8 msec "frame." Principal component analysis is performed to re-

duce the dimension of the features to 64. Figure 2 shows the block diagram of the NRAF feature extraction. Building blocks of the feature extraction circuit have been built using a CMOS 0.5μm process. Each stage in the processing pathway such as the BPF filter bank consumes less than 20μW.

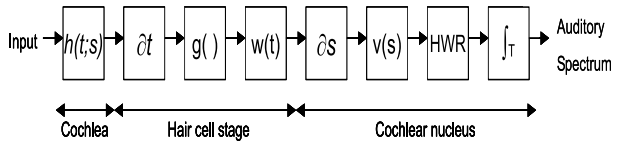

Fig. 1. Mathematical model of the early auditory system consisting of filtering in the cochlea (analysis stage), conversion of mechanical displacement into electrical activity in the IHC (transduction stage) and the lateral inhibitory network in the cochlear nucleus(reduction stage) [5].

For the purpose of comparison we also extracted the MFCCs. Each one second training sample is divided into 32 msec frames with 50% overlap and 13 MFCC coefficients are computed from each frame.

Three different methods were used to test the discriminating abilities of MFCCs and NRAFs. In the first method the mean and variance of the features over all frames are computed and used to train the classifier. We refer to this as the "mean–variance" method. In the second method feature vectors from each frame are used as training vectors to train the classifier. Each frame is treated as a new input to the classifier. We refer to this as the "all-frame" method. In the third method we implemented the "stacking" method described by Slaney [6] to enhance the performance of MFCCs. Three frames, the frame before the current frame, the current frame and the one following the current frame are stacked together to form a 39-dimensional feature vector. An optimal dimensionality reduction transform [7] is then used to reduce the dimension to 8. Thus each frame is represented by a 8-dimensional feature vector which is used to train the classifier. The stacking method was not be implemented for the NRAFs due to the computational costs involved in performing the optimal dimensionality reduction transform on stacked frames that are represented by 128-dimensional feature vectors.

3. EXPERIMENT

The database consisted of four classes: noise, animal sounds, music and speech. Each of the sound samples was one second long. The noise class comprised of nine different types of noises from the NOISEX database including babble noise. The animal class comprised of a random selection of animal sounds from the BBC Sound Effects au-

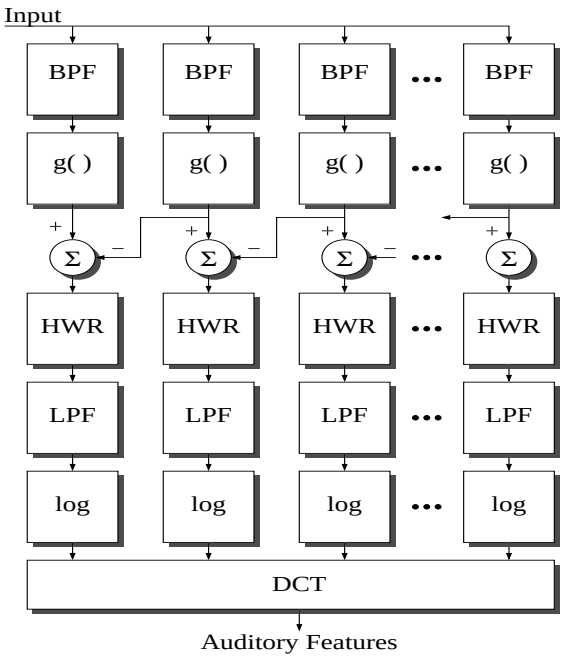

Fig. 2. The bandpass filtered version of the input is nonlinearly compressed. The difference operation between lower and higher channels approximates a spatial derivative. The half-wave rectification followed by the smoothing filter picks out the peak. The DCT is performed to decorrelate the signal.

dio CD collection. The music class was formulated using the RWC music database and included different genres of music [8]. The speech class was made up of spoken digits from the TIDIGITS and AURORA database. The training set consisted of a total of 4325 samples with 1144 noise, 732 animal, 1460 music and 989 speech samples and the test set consisted of 1124 samples with 344 noise, 180 animal, 354 music and 246 speech samples. The sounds in the database are publicly available and the sample file name and file offsets are available from the authors.

A Gaussian Mixture model (GMM) was used to model each class of data and the feature vectors from each class were used to train the GMM. During testing, the likelihood that a test sample belongs to each model is computed and the sample is assigned to the class whose model produces the highest likelihood. During testing in the all-frames method the likelihood of each frame belonging to the four different classes is computed and a majority voting is performed to determine the class of the sample.

To test the noise robustness of the MFCCs and NRAFs, a two class (speech and music) audio classification problem was chosen. White noise was added to produce samples with various signal-to-noise ratios. The GMM model was

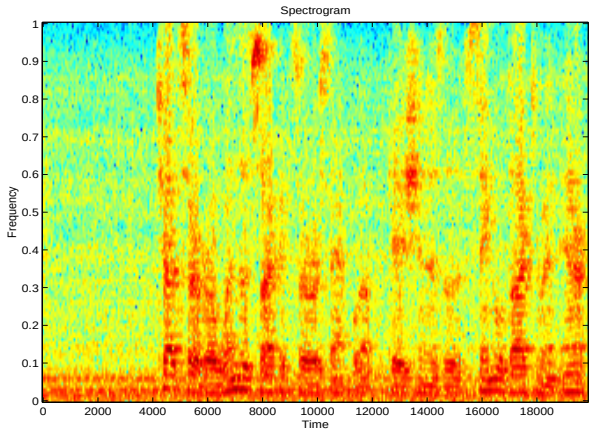

Fig. 3. Figure showing the spectrogram for a noisy speech input.

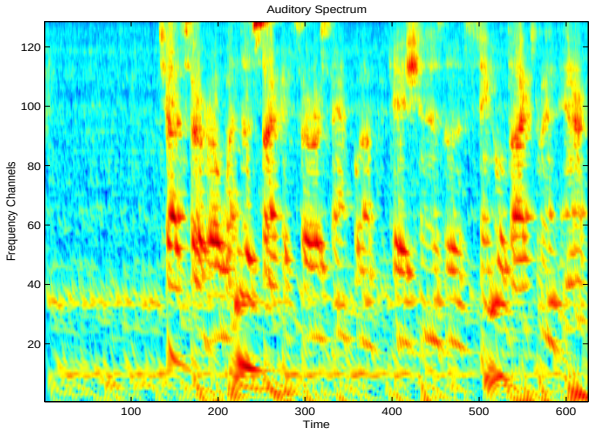

Fig. 4. Figure showing an auditory spectrum for a noisy speech input [5].

trained using the clean samples and tested with the noisy samples.

4. RESULTS

Among the three methods used to test the performance of MFCCs the mean–variance method performed best. Although we believe that with sufficient number of frames stacked together the stacked frames method should perform at least as well as the mean–variance method. For the NRAF, using all the frames to train the GMM performs slightly better than the mean–variance method. Overall, NRAF does better than MFCCs. The results are tabulated in Tables 1. The confusion matrix of MFCC and NRAF experiments (Tables 2–6.) show that NRAFs learn the speech and noise class very well. MFCCs do well on the speech class but their performance on the other classes is not as good. Both features do equally bad on the animal class, which is the hardest to classify due to the variety of sounds present and also due to the close proximity of some of the sounds to the noise class.

Performance of MFCCs and NRAFs		
Method	MFCC	NRAF
Mean–variance	85.85 %	90.22 %
All-frames	81.05 %	92.97 %
Stacked	82.38 %	-

Table 1. Table showing performance of MFCCs and NRAFs for a four-class audio classification problem using different methods.

	Noise	Animal	Music	Speech
Noise	310	18	30	0
Animal	0	140	55	0
Music	34	22	269	0
Speech	0	0	0	246

Table 2. Table showing Confusion matrix for MFCC (mean-variance). This method gave an accuracy of 85.85 %

	Noise	Animal	Music	Speech
Noise	253	9	31	0
Animal	47	137	39	0
Music	0	34	275	0
Speech	44	0	9	246

Table 3. Table showing Confusion matrix for MFCC (all-frames). This method gave an accuracy of 81.05 %

The noise robustness results are as shown in Table 7. We see that the NRAF features outperform MFCCs. An error of 41% percent corresponds to the case were all the samples of class two (speech) are misclassified as that of class one (music). We used the mean–variance method for comparison because this did best for MFCCs, but the results for NRAF might improve by using the all-frame method.

5. CONCLUSION

In this paper we presented noise-robust features that perform better than the standard MFCCs. These features can be implemented in low-power analog VLSI circuits making them attractive to ubiquitous sensor applications. Future

work would involve developing features that provide better discrimination and also designing a classification structure that can be implemented in low-power analog VLSI circuitry. This would enable us to have a low-power autonomous classification system.

6. ACKNOWLEDGEMENTS

We are grateful to Matasaka Goto for making the RWC music database available. We would like to thank the Telluride Neuromorphic Engineering Workshop for motivating this research. We are also thankful to the reviewers for their feedback and comments.

	Noise	Animal	Music	Speech
Noise	343	56	55	1
Animal	0	115	42	0
Music	1	8	226	3
Speech	0	1	31	242

Table 4. Table showing Confusion matrix for MFCC (stacked). This gave an accuracy of 82.38 %

	Noise	Animal	Music	Speech
Noise	294	19	1	20
Animal	50	140	12	3
Music	0	9	339	2
Speech	0	12	2	241

Table 5. Table showing Confusion matrix for NRAF (mean-variance). This method gave an accuracy of 90.22 %

	Noise	Animal	Music	Speech
Noise	344	31	0	0
Animal	0	148	49	0
Music	0	1	305	0
Speech	0	0	0	246

Table 6. Table showing confusion matrix for NRAF (all-frames). This method gave an accuracy of 92.79 %

Percentage error of MFCC and NRAF for different SNRs		
SNR	MFCC	NRAF
Clean	0.00 %	0.33 %
15 dB	40.33 %	17.66 %
10 dB	41.00 %	26.40 %
5 dB	41.00 %	38.16 %

Table 7. Table showing performance of MFCC and NRAF features in a two-class (speech and music) classification problem at different SNRs. The mean–variance method is used

7. REFERENCES

[1] P. Smith, M. Kucic, R. Ellis, D. Graham, and P. Hasler, "Mel-frequency cepstrum encoding in analog floating-gate circuitry," in *IEEE International Symposium on Circuits and Systems*, Phoenix, AZ, May 2002, vol. 4, pp. 671–674.

[2] Sourabh Ravindran, Cenk Demiroglu, and David Anderson, "Speech recognition using filter-bank features," in *Asilomar Conference on Signals and Systems*, Pacific Grove, CA, Nov. 2003.

[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] M. Clapp and Ralph Etienne-Cummings, "Ultrasonic bearing estimation using a mems microphone array and spatiotemporal filters," in *IEEE International Symposium on Circuits and Systems*, Scottsdale, AZ, May 2002.

[5] Kuansan Wang and Shihab Shamma, "Self-normalization and noise-robustness in early auditory representations," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 3, pp. 421–435, July 1994.

[6] Malcolm Slaney, "Semantic-audio retrieval," in *IEEE International Conference on Accoustics, Speech, and Signal Processing*, 2002.

[7] J. Duchene and S. Leclercq, "An optimal transformation for discriminant and principal component analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 10, no. 6, pp. 978–983, Nov. 1988.

[8] 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.