EURASIP Journal on Applied Signal Processing 2005:9, 1374–1381

A Physiologically Inspired Method for Audio Classification

Sourabh Ravindran
School of Electrical and Computer Engineering, College of Engineering, Georgia Institute of Technology, Atlanta, GA 30332-0250, USA
Email: stg@ece.gatech.edu

Kristopher Schlemmer
School of Electrical and Computer Engineering, College of Engineering, Georgia Institute of Technology, Atlanta, GA 30332-0250, USA
Email: kms@ece.gatech.edu

David V. Anderson
School of Electrical and Computer Engineering, College of Engineering, Georgia Institute of Technology, Atlanta, GA 30332-0250, USA
Email: dva@ece.gatech.edu

Received 2 November 2003; Revised 9 August 2004

We explore the use of physiologically inspired auditory features with both physiologically motivated and statistical audio classification methods. We use features derived from a biophysically defensible model of the early auditory system for audio classification using a neural network classifier. We also use a Gaussian-mixture-model (GMM)-based classifier for the purpose of comparison and show that the neural-network-based approach works better. Further, we use features from a more advanced model of the auditory system and show that the features extracted from this model of the primary auditory cortex perform better than the features from the early auditory stage. The features give good classification performance with only one-second data segments used for training and testing.

Keywords and phrases: auditory model, feature extraction, neural nets, audio classification, Gaussian mixture models.

1. INTRODUCTION

Human-like performance by machines in tasks of speech and audio processing has remained an elusive goal. In an attempt to bridge the gap in performance between humans and machines, there has been an increased effort to study and model physiological processes. However, the widespread use of biologically inspired features proposed in the past has been hampered mainly by either the lack of robustness to noise or the formidable computational costs.

In physiological systems, sensor processing occurs in several stages. It is likely the case that signal features and biological processing techniques evolved together and are complementary or well matched. It is precisely because of this reason that modeling the feature extraction processes should go hand in hand with the modeling of the processes that use these features.

We present features extracted from a model of the early auditory system that have been shown to be robust to noise [1, 2]. The feature extraction can be implemented in low-power analog VLSI circuitry which apart from providing substantial power gains also enables us to achieve feature extraction in real time. We specifically study a four-class audio classification problem and use a neural-network-based classifier for the classification. The method used herein is similar to that used by Teolis and Shamma [3] for classifying transient signals. The primary difference in our approach is in the additional processing of the auditory features before feeding them to the neural network. The rest of the paper is organized as follows. Section 2 introduces the early auditory system. Section 3 discusses models of the early auditory system and the primary auditory cortex. Section 4 explains the feature extraction process and Section 5 introduces the two methods used to evaluate the features in a four-case audio classification problem. Section 6 presents the experiments, followed by the results and the conclusion.

2. EARLY AUDITORY SYSTEM

As sounds enter the ear, a small amount of signal conditioning and spectral shaping occurs in the outer ear, but the

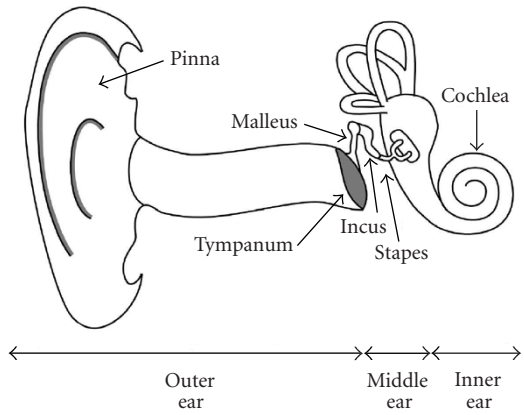

Figure 1: A cut-away view of the human ear. This shows the three stages of the ear. The outer ear includes the pinna, the ear canal, and the tympanum (ear drum). The middle ear is composed of three small bones, or ossicles. Simply put, these three bones work together for gain control and for impedance matching between the outer and the inner ear. The inner ear is the snail-shaped bone called the cochlea. This is where the incoming sounds are decomposed into the respective frequency components.

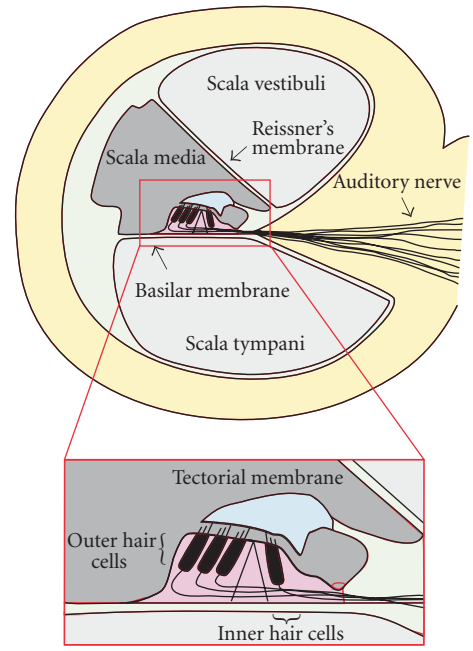

Figure 2: A cross-section of the human cochlea. Within the bone are three fluid-filled chambers that are separated by two membranes. The input to the cochlea is in the scala vestibuli, which is connected at the apical end to the scala tympani. Pressure differences between these two chambers lead to movement in the basilar membrane. The scala media is isolated from the other two chambers.

signals remain relatively unscathed until they contact the ear drum, which is the pathway to the middle ear. The middle ear is composed of three small bones, or ossicles. Simply put, these three bones work together for gain control and for impedance matching between the outer and the inner ear (matching the low impedance of the auditory canal with the high impedance of the cochlear fluid). The middle ear couples the sound energy in the auditory canal to the inner ear or the cochlea which is a snail-shaped bone. The placement of the cochlea with respect to the rest of the ear is shown in Figure 1.

Figure 2 shows a cross-sectional view of the cochlea. The input to the cochlea is through the oval window, and barring a scale factor resulting from the gain control, the signal that enters the oval window of the cochlea is largely the same as that which enters the ear. The oval window leads to one of three fluid-filled compartments within the cochlea. These chambers called scala vestibuli, scala media, and scala tympani are separated by flexible membranes. Reissner's membrane separates the scala vestibuli from the scala media, and the basilar membrane separates the scala tympani from the scala media [4, 5].

As the oval window is pushed in and out as a result of incident sound waves, pressure waves enter the cochlea in the scala vestibuli and then propagate down the length of the cochlea. Since the scala vestibuli and the scala tympani are connected, the increased pressure propagates back down the length of the cochlea through the scala tympani to the front end, also called the basal end. When the pressure wave hits the basal end, it causes a small window, called the round window, that is similar in composition to the oval window, to bow outwards to absorb the increased pressure. During this process, the two membrane dividers bend and bow in response to the changes in pressure [6] giving rise to a traveling wave in the basilar membrane.

At the basal end, the basilar membrane is very narrow but gets wider towards the apical end. Further, the stiffness of the basilar membrane decreases down its length from the base to the apex. Due to these variations along its length, different parts of the basilar membrane resonate at different frequencies, and the frequencies at which they resonate are highly dependent upon the location within the cochlea. The travelling wave that develops inside the cochlea propagates down the length of the cochlea until it reaches the point where the basilar membrane resonates with the same frequency as the input signal. The wave will essentially die out after the point where resonance occurs because the basilar membrane will no longer support the propagation. It has been observed that the lower frequencies travel further than the higher frequencies. Also the basilar membrane has exponential changes in the resonant frequency for linear distances down the length of the cochlea.

The basilar membrane is also attached to what is known as the organ of Corti. One important feature of the organ of Corti is that it has sensory cells called inner hair cells (IHC) that sense the motion of the basilar membrane. As the basilar membrane moves up and down in response to the pressure waves, it causes the local movement of the cochlear fluid. The viscous drag of the fluid bends the cilia attached to the IHC. The bending of the cilia controls the ionic flow into the hair cells through a nonlinear channel. Due to this ionic current flow, charge builds up across the hair-cell membrane.

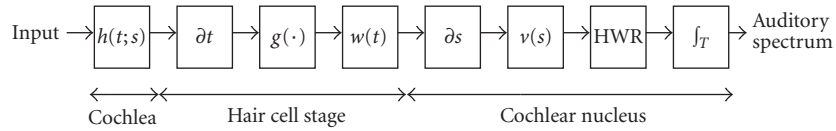

Figure 3: Mathematical model of the early auditory system consisting of filtering in the cochlea (analysis stage), conversion of mechanical displacement into electrical activity in the IHC (transduction stage), and the lateral inhibitory network in the cochlear nucleus (reduction stage) [1].

This mechanism converts the mechanical displacement of the basilar membrane into electrical activity. Once the potential builds up above a certain threshold, the hair cell fires. This neural spike is carried to the cochlear nucleus by the auditory nerve fibre. The neurons in the cochlear nucleus (CN) exhibit inhibition characteristics and it is believed that lateral inhibition exists in the cochlear nucleus. The lateral interaction of the neurons is spatially limited, that is, as the distance between the neurons increases, the interaction decreases [7].

3. MATHEMATICAL MODEL OF THE AUDITORY SYSTEM

3.1. Model of the early auditory system

Yang et al. [8] have presented a biophysically defensible mathematical model of the early auditory system. The model is shown in Figure 3 and described below.

When viewing the way the cochlea acts on signals of different frequencies from an engineering perspective, it can be seen that the cochlea has bandpass frequency responses for each location. An accurate but computationally prohibitive model would have a bank of bandpass filters with center frequencies corresponding to the resonant frequency of every point along the cochlea—the cochlea has about 3000 inner hair cells acting as transduction points. In practice 10–20 filters per octave are considered an adequate approximation. The cochlear filters $h(t;s)$ typically have 20 dB/decade rolloffs on the low-frequency side and a very sharp rolloff on the high-frequency side.

The coupling of the cochlear fluid and the inner hair cells is modeled by a time derivative (∂t). This can be justified since the extent of IHC cilia deflection depends on the viscous drag of the cochlear fluid and the drag is directly dependent on the velocity of motion. The nonlinearity of the ionic channel is modeled by a sigmoid-like function $g(\cdot)$ and the leakiness of the cell membrane is modeled by a lowpass filter $w(t)$.

Lateral inhibition in the cochlear nucleus is modeled by a spatial derivative (∂s). The spatial derivative is leaky in the sense that it is accompanied by a local smoothing that reflects the limited spatial extent of the interactions of the CN neurons. Thus, the spatial derivative is often modeled along with a spatial lowpass filter $v(s)$. The nonlinearity of the CN neurons is modeled by a half-wave rectifier (HWR) and the inability of the central auditory neurons to react to rapid temporal changes is modeled by temporal integration ($\int_T$). The output of this model is referred to as the auditory spectrum and it has been shown that this representation is more robust to noise as compared to the normal power spectrum [1].

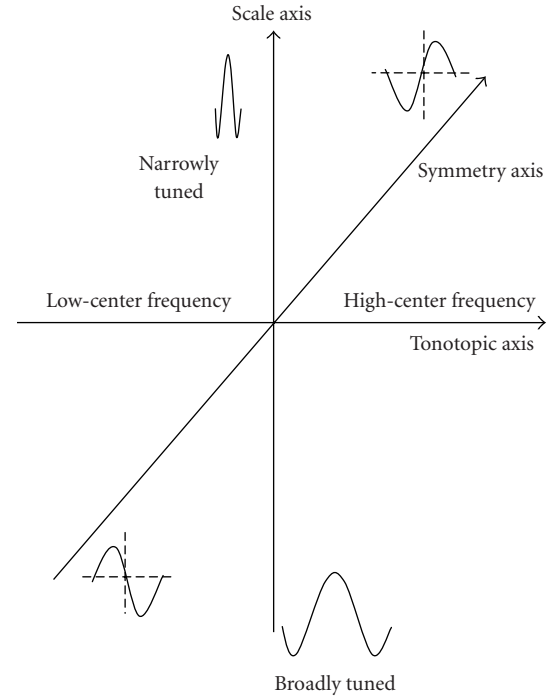

Figure 4: Schematic of the cortical model. It is proposed in [9] that the response fields of neurons in the primary auditory cortex are arranged along three mutually perpendicular axes: the tonotopic axis, the bandwidth or scale axis, and the symmetry or phase axis.

3.2. Cortical model

Wang and Shamma [9] have proposed a model of the spectral shape analysis in the primary auditory cortex. The schematic of the model is shown in Figure 4. According to this model, neurons in the primary auditory cortex (A1) are organized along three mutually perpendicular axes. The response fields of neurons lined along the tonotopic axis are tuned to different center frequencies. The bandwidth of the response field of neurons lined along the scale axis monotonically decreases along that axis. Along the symmetry axis, the response field of the neurons displays a systematic change in symmetry. At the center of A1, the response field has an excitatory center, surrounded by inhibitory sidebands. The response field tends to be more asymmetrical with increasing distance from the center of A1. It has been argued that the tonotopic axis is akin to a Fourier transform, and the presence of different scales over which this transform is performed leads to a multiscale Fourier transform. It has been shown that performing such

an operation on the auditory spectrum leads to the extraction of spatial and temporal modulation information [10]. This model is used to extract the cortical features explained in Section 4.

4. FEATURE EXTRACTION

4.1. Simplified model of the early auditory system

The cochlear filters $h(t;s)$ are implemented through a bandpass filter bank (BPF), with 40 dB/decade rolloff on the low frequency. This models the 20 dB/decade cochlear filter rolloff and also provides a time differentiation of the input signal. The nonlinearity of the ionic channel $g(\cdot)$ is implemented by a sigmoid-like function. The temporal lowpass filter $w(t)$ is ignored and it has been shown that at moderate sound intensities, this is a valid approximation [1]. The spatial derivative ∂s is approximated by a difference operation between the adjacent frequency channels and the spatial low-pass filter $v(s)$ is ignored (this corresponds to limiting the spatial extent of lateral inhibition of the CN neurons to adjacent channels). The half-wave rectification stage is retained and the temporal averaging is implemented by a lowpass filter (LPF).

4.2. Features

4.2.1. Auditory model features

The auditory spectrum derived from the simplified model of the early auditory system is a two-dimensional time-frequency representation. The filter bank consists of 128 channels tuned from 180 Hz to 7246 Hz and the temporal averaging (lowpass filtering) is done over 8-milliseconds "frames", thus the auditory spectrum for one second of data is a 128×125 two-dimensional matrix. The neural response over time is modeled by a mean activity level (temporal average) and by the variation of activity over time (temporal variance). Thus taking the temporal average and temporal variance of the auditory spectrum, we end up with a 256-dimensional feature vector for each one-second segment. We refer to these as the AM short features.

4.2.2. Noise-robust auditory features

We modified the early auditory model to incorporate the log compression due to the outer hair cells [11] and also introduced a decorrelation stage. The decorrelation stage is important for practical reasons; while the neural networks naturally perform this operation, doing so explicitly reduces the training requirements for the networks. We refer to this representation as noise-robust auditory features (NRAF). The noise robustness of these features is shown elsewhere [2]. The NRAF extraction can be implemented in low-power analog VLSI circuitry as shown in Figure 5. The auditory spectrum is log compressed and transformed using a discrete cosine transform (DCT) which effectively decorrelates the channels. The temporal average and temporal variance of this representation yield a 256-dimensional feature vector for every one-second segment. We refer to these as the NRAF short features. The NRAF feature extraction is similar to extracting the continuous-time mel-frequency cepstral coefficients (MFCCs) [12], however the NRAF features are more faithful to the processes in the auditory system.

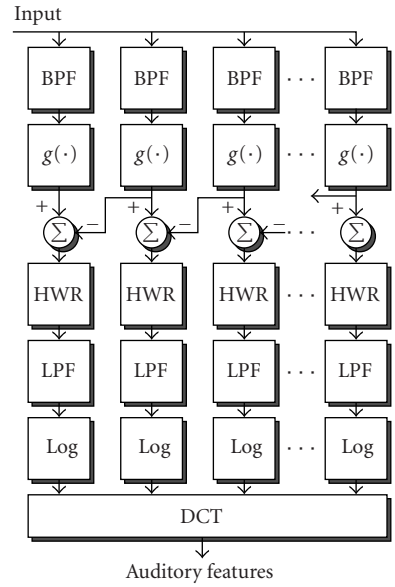

Figure 5: The bandpass-filtered version of the input is nonlinearly compressed and fed back to the input. The difference operation between lower and higher channels approximates a spatial derivative. The half-wave rectification followed by the smoothing filter picks out the peak. Log compression is performed followed by the DCT to decorrelate the signal.

4.2.3. Rate-scale-frequency features

A multiscale transform (performed using the cortical model) on the auditory spectrum leads to a four-dimensional representation referred to as rate-scale-frequency-time (RSFT) [9]. The processing done by the cortical model on the auditory spectrum is similar to a two-dimensional wavelet transform. Frequency represents the tonotopic axis in the basilar membrane and in our implementation is the center frequency of the bandpass filters of the early auditory model. Each unit along the time axis corresponds to 8 milliseconds. This is the duration over which the temporal integration is performed in the early auditory model. Rate corresponds to the center frequency of the temporal filters used in the transform and yields temporal modulation information. Scale corresponds to the center frequency of the spatial (frequency) filters used in the transform and yields spatial modulation information. The RSFT representation is collapsed across the time dimension to obtain the RSF features. Principal component analysis is performed to reduce the RSF features to a dimension of 256. These features are referred to as the RSF short features.

4.2.4. Mel-frequency cepstral coefficients

For the purpose of comparison, we also extracted the MFCCs. Each one-second training sample is divided into 32-millisecond frames with 50% overlap and 13 MFCC

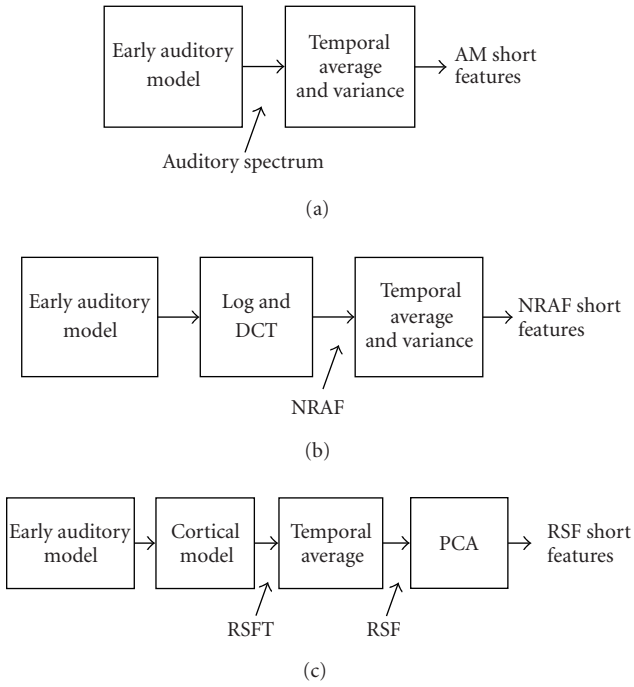

Figure 6: Extraction of the different features used in the classification.

coefficients are computed from each frame. The mean and variance of the 13 features over one-second segment were calculated to give a 26-dimensional feature vector.

Figure 6 gives a graphic representation of how different feature sets are obtained. The NRAF short features differ from the AM short features in that they incorporate an additional log compression and decorrelation stage. The RSF short features are obtained by a multiresolution processing on the auditory spectrum, followed by dimensionality reduction.

5. CLASSIFICATION METHODS

We used two different methods for classification, a Gaussian-mixture-model (GMM) -based classifier and a neural-net (NN) -based classifier. The GMM-based classifier is used as the nonanthropomorphic control case and was chosen because of its successful use in audio classification and speech recognition. It is easily trained and can match feature space morphologies.

5.1. GMM-based classifier

The feature vectors from each class were used to train the GMM models for those classes. During testing, the likelihood of a test sample belonging to each model is computed and the sample is assigned to the class whose model produces the highest likelihood. Diagonal covariance was assumed with a separate covariance matrix for each of the mixtures. The priors are set based on the number of data samples in each mixture. To implement the GMM, we used the Netlab software provided by Nabney and Bishop [13].

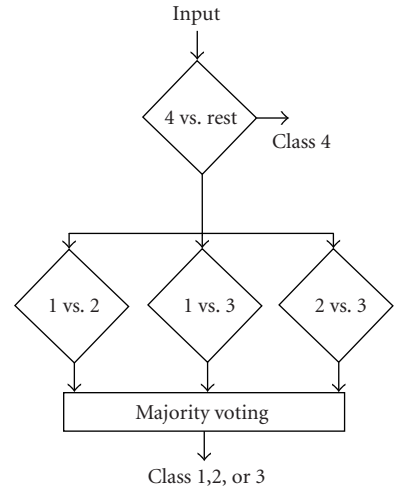

Figure 7: Flow chart showing the decision rights of the collaborative committee for the NN classifier.

5.2. NN-based classifier

Through iterative development, the optimal classification system was found to be a collaborative committee of four independently trained binary perceptrons. Three of the perceptrons were trained solely on two of the four classes (class 1 (noise) versus class 2 (animal sounds), class 1 versus class 3 (music), and class 2 versus class 3). Because of the linear separability of class 4 (speech) within the feature space, the fourth perceptron was trained on all four classes, learning to distinguish speech from the other three classes. All four perceptrons employed the $\tanh(\cdot)$ decision function and the conjugate gradient descent training with momentum learning. All training converged within 500 epochs of training, and consistent performance and generalization results were realized. With the four perceptrons independently trained, a collaborative committee was established with decision rights shown in Figure 7. The binary classifier for classifying speech versus the rest was allowed to make the decisions on the speech class due to its ability to learn this class extremely well. A collaborative committee using majority voting was instituted to arbitrate amongst the other three binary classifiers. The performance curve for training and testing for the three different features is as shown in Figures 8, 9, and 10. We see that while NRAF short and RSF short features generalize well, AM short features do not perform as well.

6. EXPERIMENTS

MFCCs are the most popular features in state-of-the-art audio classification and speech recognition. Peltonen et al. [14] showed that MFCCs used in conjunction with GMM-based classifiers performed very well for an auditory scene recognition experiment involving identifying 17 different auditory scenes from amongst 26 scenes. They reported near-human-like performance when using 30 seconds of data to perform the scene recognition. We use a similar approach using MFCCs and a GMM-based classifier as the baseline system.

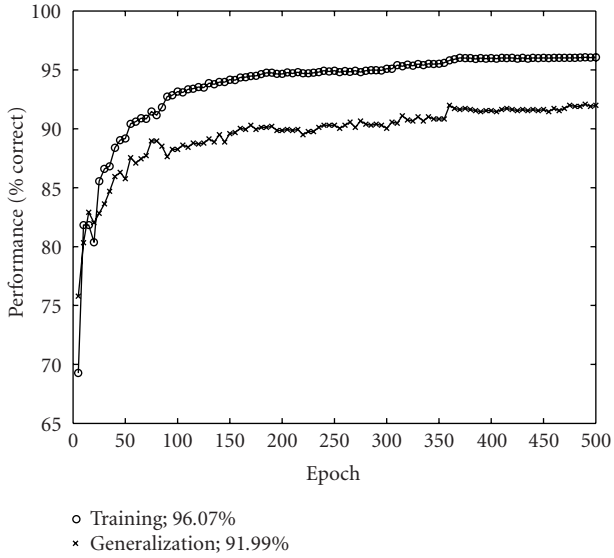

Figure 8: Performance curves during training and testing of the NN classifier with the NRAF short features.

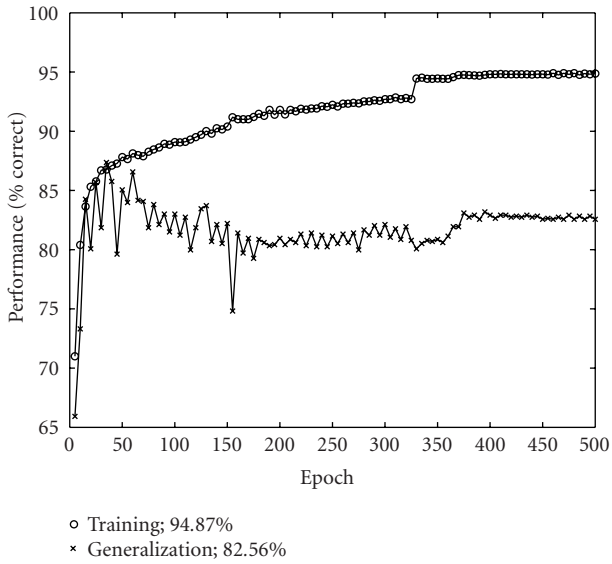

Figure 9: Performance curves during training and testing of the NN classifier with the AM short features.

The database consisted of four classes; noise, animal sounds, music, and speech. Each of the sound samples was a second long. The noise class was comprised of nine different types of noises from the NOISEX database which included babble noise. The animal class was comprised of a random selection of animal sounds from the BBC Sound Effects audio CD collection. The music class was formulated using the RWC music database [15] and included different genres of music. 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

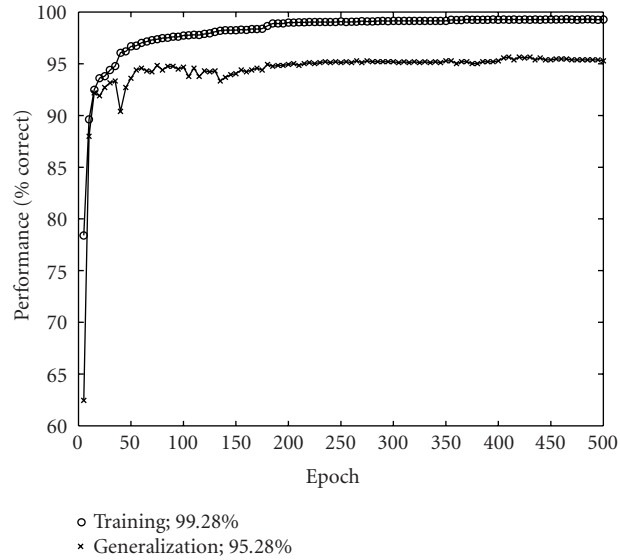

Figure 10: Performance curves during training and testing of the NN classifier with the RSF short features.

of 1124 samples with 344 noise, 180 animal, 354, music, and 246 speech samples.

Dimensionality reduction was necessitated by the inability of the GMM to handle large-dimensional feature vectors. For the AM short features, it was empirically found that reducing to a 64-dimensional vector by using principal component analysis (PCA) provided the best result. Since the PCA helps decorrelate the features, a diagonal covariance matrix was used in the GMMs. Performing linear discriminant analysis [16] for dimensionality reduction and decorrelation did not provide better results as compared to PCA. The NRAF short features were also reduced to 64 dimensions similarly. For these two feature sets, a 4-mixture GMM was used to perform the classification. The RSF short features were further reduced to a dimension of 55 using PCA, and a 6-mixture GMM was used to perform the classification. For MFCCs, a 4-mixture GMM was used. The GMMs were optimized to give the best results on the database. However, the improvement in accuracy comes at a cost of reduced generalization ability.

Further experiments were performed to incorporate temporal information into the baseline classifier (GMM-based classifier using MFCCs). The MFCCs per frame were used to generate models for each class. It was determined that using 9 mixtures gave the best result. In the test phase the log likelihood of each frame in the one second segment belonging to each of the classes was computed and summed over one second. The class with the highest summed log likelihood was declared the winner.

7. RESULTS

As can be seen from Table 1, NRAF short features outperform MFCCs using GMM-based classifier. The NRAF short features also outperform the AM short features using both

Table 1: Performance (% correct) for different features and different classifiers.

	MFCC	NRAF short	AM short	RSF short
GMM	85.85%	90.21%	71.79%	70.99%
NN	—	91.99%	82.56%	95.28%

Table 2: Confusion matrix (rows give the decision and columns give the true class) for MFCC with GMM. This method gave an accuracy of 85.85%.

	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 3: Confusion matrix (rows give the decision and columns give the true class) for NRAF short with GMM. This gave an accuracy of 90.21%.

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

Table 4: Confusion matrix (rows give the decision and columns give the true class) for NRAF short with NN. This gave an accuracy of 91.99%.

	Noise	Animal	Music	Speech
Noise	340	34	6	0
Animal	0	133	31	2
Music	4	10	317	0
Speech	0	3	0	244

Table 5: Confusion matrix (rows give the decision and columns give the true class) for AM short with GMM. This gave an accuracy of 71.79%.

	Noise	Animal	Music	Speech
Noise	181	23	10	2
Animal	19	121	33	10
Music	52	36	285	14
Speech	92	0	26	220

Table 6: Confusion matrix (rows give the decision and columns give the true class) for AM short with NN. This gave an accuracy of 82.56%.

	Noise	Animal	Music	Speech
Noise	297	33	27	0
Animal	6	106	34	4
Music	39	40	284	1
Speech	2	1	9	241

Table 7: Confusion matrix (rows give the decision and columns give the true class) for RSF short with GMM. This gave an accuracy of 70.99%.

	Noise	Animal	Music	Speech
Noise	136	6	6	0
Animal	4	142	48	1
Music	54	30	280	5
Speech	150	2	20	240

Table 8: Confusion matrix (rows give the decision and columns give the true class) for RSF short with NN. This gave an accuracy of 95.28%.

	Noise	Animal	Music	Speech
Noise	337	28	2	0
Animal	0	143	5	0
Music	7	9	347	2
Speech	0	0	0	244

the GMM- and the NN-based classifiers. RSF short features outperform NRAF short features while using the NN classifier. The results for all the features (except MFCCs, which were not tested with the NN classifier) were better with an NN classifier as compared to a GMM classifier. Tables 2, 3, 4, 5, 6, 7, and 8 give the confusion matrices of the various features for the two different classifiers.

It is seen from the confusion matrices that MFCCs do a very good job of learning the speech class. All the other features used are also able to separate out the speech class with reasonable accuracy indicating the separability of the speech class in the feature space. It is interesting to note that most of the mistakes by MFCCs in the noise class are misclassifications as music (Table 2) but NRAF makes most of its mistakes in this class as misclassifications into the animal class (Table 6), which is more acceptable as some of the animal sounds are very close to noise. The animal class seems to be the most difficult to learn but RSF short features in conjunction with the NN classifier do a good job of learning this class. Most of the mistakes in this case are misclassifications as noise.

Table 9: Performance (% correct) for MFCCs. The one-second segment was divided into 30-millisecond frames, and the final decision was made by combining the frame decisions.

Classifier	Performance
GMM + majority rule	81.49%

The result of incorporating temporal information into the GMM-based classifier is shown in Table 9. It is seen that the performance decreases in comparison with using mean and variance of the MFCCs over one second. This could be attributed to the fact that there is too much variability in each of the classes. Performing temporal smoothing over one second makes the features more robust.

8. CONCLUSIONS

We have shown that for the given four classes, audio classification problem features derived from a model of the auditory system combine better with an NN classifier as compared to a GMM-based classifier. The GMM-based classifier was optimized to give the best results for the database while the NN classifier was trained with generalization in mind. The accuracy of the NN classifier can be increased but at the cost of reducing the generalization ability of the classifier. It could be argued that the few number of classes considered combined with the high dimensionality of the feature space might render the classes linearly separable and hence aid the NN approach. The performance of GMM- and neural-network-based classifiers was not tested for large number of classes and the scalability of NN classifier to large number of classes is an open question. Neural networks however provide an efficient and natural way of handling large-dimensional feature vectors as obtained from models of the human auditory system.

ACKNOWLEDGMENTS

The authors are grateful to Shihab Shamma for the very useful insights he provided on the working of the auditory model. They would also like to thank Malcolm Slaney for all his help and the Telluride Neuromorphic Engineering Workshop for having motivated this work.

REFERENCES

[1] K. Wang and S. Shamma, "Self-normalization and noise-robustness in early auditory representations," *IEEE Trans. Speech Audio Processing*, vol. 2, no. 3, pp. 421–435, 1994 .

[2] S. Ravindran, D. Anderson, and M. Slaney, "Low-power audio classification for ubiquitous sensor networks," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP '04)*, Montreal, Canada, May 2004.

[3] A. Teolis and S. Shamma, "Classification of transient signals via auditory representations," Tech. Rep. TR 91-99, Systems Research Center, University of Maryland, College Park, Md, USA, 1991.

[4] E. Kandel, J. Schwartz, and T. Jessel, *Principles of Neural Science*, McGraw-Hill, New York, NY, USA, 4th edition, 2000.

[5] R. Berne and M. Levy, *Physiology*, Mosby, New York, NY, USA, 4th edition, 1998.

[6] P. Denes and E. Pinson, *The Speech Chain*, Freeman, New York, NY, USA, 2nd edition, 1993.

[7] C. Koch and I. Segev, Eds., *Methods in Neural Modelling*, MIT Press, Cambridge, Mass, USA, 1989.

[8] X. Yang, K. Wang, and S. Shamma, "Auditory representations of acoustic signals," *IEEE Trans. Inform. Theory*, vol. 38, no. 2, pp. 824–839, 1992.

[9] K. Wang and S. Shamma, "Spectral shape analysis in the central auditory system," *IEEE Trans. Speech Audio Processing*, vol. 3, no. 5, pp. 382–395, 1995.

[10] L. Atlas and S. Shamma, "Joint acoustic and modulation frequency," *EURASIP Journal on Applied Signal Processing*, vol. 2003, no. 7, pp. 668–675, 2003.

[11] J. O. Pickles, *An Introduction to the Physiology of Hearing*, Academic Press, New York, NY, USA, 2nd edition, 2000.

[12] P. D. Smith, M. Kucic, R. Ellis, P. Hasler, and D. V. Anderson, "Mel-frequency cepstrum encoding in analog floating-gate circuitry," in *Proc. IEEE Int. Symp. Circuits and Systems (ISCAS '02)*, vol. 4, pp. 671–674, Scottsdale, Ariz, USA, May 2002.

[13] I. Nabney and C. Bishop, "Netlab neural network software," http://www.ncrg.aston.ac.uk/netlab/.

[14] V. Peltonen, J. Tuomi, A. Klapuri, J. Huopaniemi, and T. Sorsa, "Computational auditory scene recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP '02)*, Orlando, Fla, USA, May 2002.

[15] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "RWC music database: Music genre database and musical instrument sound database," in *Proc. 4th International Conference on Music Information Retrieval (ISMIR '03)*, pp. 229–230, Baltimore, Md, USA, October 2003.

[16] J. Duchene and S. Leclercq, "An optimal transformation for discriminant and principal component analysis," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 10, no. 6, pp. 978–983, 1988.

Sourabh Ravindran received the B.E. degree in electronics and communication engineering from Bangalore University, Bangalore, India, in October 2000, and the M.S. degree in electrical and computer engineering from the Georgia Institute of Technology (Georgia Tech), Atlanta, Ga, in August 2003. He is currently pursuing the Ph.D. degree at Georgia Tech. His research interests include audio classification, auditory modeling, and speech recognition. He is a Student Member of IEEE.

Kristopher Schlemmer graduated summa cum laude as a Commonwealth Scholar from the University of Massachusetts Dartmouth in 2000, earning his B.S.E.E. degree, and from Georgia Institute of Technology in 2004, earning his M.S.E.C.E. degree. Mr. Schlemmer is currently employed at Raytheon Integrated Defense Systems in Portsmouth, Rhode Island, and his interests include analog design, digital signal processing, artificial intelligence, and neurocomputing.

David V. Anderson was born and raised in La Grande, Ore. He received the B.S. degree in electrical engineering (magna cum laude) and the M.S. degree from Brigham Young University in August 1993 and April 1994, respectively, where he worked on the development of a digital hearing aid. He received the Ph.D. degree from the Georgia Institute of Technology (Georgia Tech), Atlanta, in March 1999. He is currently on the faculty at Georgia Tech. His research interests include audition and psychoacoustics, signal processing in the context of human auditory characteristics, and the real-time application of such techniques. He is also actively involved in the development and promotion of computer-enhanced education. Dr. Anderson is a Member of the Acoustical Society of America, Tau Beta Pi, and the American Society for Engineering Education.