



Speaker Recognition with the Auditory Image Model and Self-Organizing Feature Maps: A comparison with traditional techniques

Timothy R. Anderson and Roy D. Patterson

Abstract— Speech and speaker recognition were examined using Self-Organizing Feature Maps (SOFMs) and three different representations of speech – traditional Mel-Cepstral Coefficients (MCC) and the integrated outputs of two different models of the auditory periphery: the Auditory Image Model (AIM) of Patterson and Payton's auditory model (PAM). AIM is a functional model of human hearing up to the level of our initial experience of a sound, that is, our 'auditory image' of the sound. PAM is a neurophysiologically based model of the auditory periphery. In the current experiments, the input vectors for the recognizer were based on the neural activity patterns flowing from the cochlear simulations of AIM and PAM. The phoneme recognition results are based on the 39 phoneme classes of K.F. Lee. The results showed that the auditory models supported better recognition accuracy than MCC using the training and test sets from dialect regions 1 and 2 of the TIMIT database (1140 sentences from 114 speakers for training, 370 sentences from 37 speakers for testing). The two representations made different types of phoneme recognition errors. Speaker recognition experiments using these same representations showed that AIM provided results comparable to that of MCC. PAM did not perform as well.

Keywords— Auditory Image Model, speaker identification, phoneme recognition, Self-Organizing Feature Maps

1. INTRODUCTION

Speaker recognition researchers have traditionally used Linear Predictive Coding (LPC) cepstral analysis as the representation of choice [11]. Vector Quantization (VQ) techniques using LPC cepstral features have been shown to be an effective combination for speaker recognition [12]. More recent research [2] has demonstrated success using an auditory model for feature representation and neural network based VQ for classification.

Research has been conducted to determine what classes of sounds contain the most effective features for speaker recognition. Sambur [10] has shown that, in general, voiced sounds contain the best features. Specifically, nasals and vowels are the best classes to use for speaker recognition.

This research is part of a continuing effort to examine the benefits of using auditory models and neural networks for speech and speaker recognition. This paper examines the use of auditory model representations and Kohonen self-organizing feature maps to perform both the classification of speech and the identification of speaker. Only vowels

T. R. Anderson is with the Armstrong Laboratory, Bioacoustics and Biocommunications Branch, Wright-Patterson AFB, OH USA. E-mail: tra@neural.aamrl.wpafb.af.mil.

R. D. Patterson is with the Medical Research Council, Applied Psychology Unit, 15 Chaucer Road, Cambridge, England CB2 2EF. E-mail: roy.patterson@mrc-applied-psychology.cambridge.ac.uk

were used to perform the speaker identification. The results are compared with those using LPC features.

2. AUDITORY MODEL DESCRIPTIONS

2.1 Payton

Payton [9] developed a model of peripheral auditory processing that incorporates processing steps describing the conversion from the acoustic pressure-wave signal at the eardrum to the time-course activity in auditory neurons. The model can process arbitrary time-domain waveforms and yield the probability of neural firing. The model contains separate modules for each anatomical section of the periphery. These modules are: a middle ear section, a model of the basilar membrane, and a hair-cell section that includes synaptic connections to auditory-nerve fibers. The predicted membrane displacements, for 20 locations along the basilar membrane, are used as input to subsequent stages of the model. The haircell/synapse section of Payton's composite model predicts the probability of firing for 20 auditory neurons covering the frequency range 440 Hz to 6600 Hz. The model converts the acoustic pressure wave into action-potential firing probability as a function of time through the stimulus.

2.2 Auditory Image Model

In Patterson's Auditory Image Model (AIM) [8], the spectral analysis is performed by a gammatone auditory filterbank which converts the incoming wave into a surface that provides a reasonable representation of the motion of the basilar membrane as a function of time. The operation of the inner hair cells is simulated by a module that includes a bank of logarithmic compressors and a bank of adaptive threshold generators. Together they convert the basilar membrane motion into a surface that represents the pattern of activity at the output of the cochlea. The adaptive thresholding mechanism removes the temporal and spectral smearing introduced by the filterbank and enhances features in the filterbank output.

3. SIGNAL REPRESENTATION

Historically, the short-time spectrum has played a major role in speech analysis. Specifically, Linear Predictive Coding (LPC) has been used extensively for speech coding and speech recognition because of its computational efficiency. Previous work by this author has compared the performance of LPC and auditory models for phoneme recognition using a small database of speakers [1].

For the present work the 20 channel output of the hair cell/synapse section of Payton's auditory model (PAM) was averaged over a 16 millisecond window with a 5 millisecond frame rate. Patterson's model was used to calculate hair cell outputs of 20 filters covering the same frequency range as that of Payton's model. The output of Patterson's model was also averaged over a 16 millisecond window with a 5 millisecond frame rate. A third representation, Linear Predictive Coding (LPC) derived mel-cepstral coefficients, was used as the baseline. A 16 millisecond Hamming window was used with a 5 millisecond frame rate. Twenty LPC coefficients for each frame were converted to mel-scale using the bilinear transformation method [7].

4. DATABASE DESCRIPTION

The speech data used for this work comprised a subset of the TIMIT acoustic-phonetic database [5]. The work reported here used all 10 sentences from 114 talkers from dialect regions 1 and 2 (37 female and 77 male) for training and a different 37 talkers from regions 1 and 2 (12 female and 25 male) for testing. During the development of the SPHINX [6] speech recognition system at Carnegie Mellon University (CMU), the TIMIT phoneme labels were slightly modified. This modified convention was adopted for the present research in order to provide a better means of comparing results with other established systems. This convention yields 39 phones in separate categories.

5. SELF-ORGANIZING FEATURE MAPS

The neural network selected for this investigation was the self-organizing feature map [4]. This neural network was selected because of its ability to learn a topological mapping of an input data space into a pattern space which defines discrimination or decision surfaces. This process has been used by Kohonen in a system for phonetic recognition of Finnish and Japanese [3]. The operation of this network resembles the classical vector-quantization method called k-means clustering. Self-organizing feature maps are more general because topologically close nodes are sensitive to inputs that are physically similar. Output nodes will be ordered in a natural manner.

Kohonen's algorithm adjusts weights from common input nodes to output nodes arranged in a two-dimensional grid. Each input node is connected to every output node. Real-valued input vectors are presented sequentially in time to the network without specifying the desired output. After enough input vectors have been presented, each node's weights will specify a cluster center. These cluster centers approximate the probability density function of the input vectors. The weight adjustment is based on a distortion measure. In this work, a mean squared error distortion measure was used, based on the input and stored weights.

Most speech recognition systems employing vector quantization use a codebook size of 256. This codebook size was used in this experiment as well, in the form of a 16x16 grid. An initial neighborhood of 4x4 nodes was linearly reduced to 1.5x1.5 during the first 30 speakers' data and held constant at 1.5x1.5 thereafter. The initial learning

rate of 0.9, was linearly reduced to 0.1 during the first 30 speakers' data and reduced linearly from 0.1 to 0.0 during the remaining speakers.

The calibration process was similar to that used by Kohonen. Once trained, learning was turned off (the weights were fixed) and the training data was presented to the feature map a second time. The node that responded to each training token was associated with that token label. The token label that had the largest number of responses for each node was assigned as the label for that node.

Learning Vector Quantization (LVQ) was used on the calibrated codebook to adjust the codewords for improved performance [4]. In this work LVQ3 was used with a window width of 0.2, an epsilon of 0.02, and an eta of 0.1.

6. PHONEME RECOGNITION PERFORMANCE

Experiments which compared the performance of the auditory models with mel-cepstral coefficients (MCC) showed that both auditory-model representations performed significantly better than MCC in terms of phoneme-recognition accuracy under the conditions tested (high signal-to-noise and a large database of speakers). The three representations made different types of broad class recognition errors. Results in the form of average recognition, substitution, deletion, and insertion rates are shown in Table I. Broad class recognition rates are shown in Table II.

TABLE I
SUMMARY RESULTS

	MCC	AIM	PAM
Correct	49.18	50.87	50.38
Deletions	11.67	7.50	7.38
Insertions	2.20	1.61	1.34
Substitutions	36.95	40.02	40.90

TABLE II
BROAD CLASS RESULTS

	MCC	AIM	PAM
Fricatives	62.42	72.57	72.73
Glides	42.13	34.13	36.49
Nasals	52.08	43.65	45.20
Silence	91.19	93.65	94.73
Stops	0.0	38.05	27.60
Vowels	82.52	87.01	86.07
Total	71.31	76.11	75.71

The AIM representation performed significantly better,¹ in terms of phoneme-recognition performance, than either of the other representations. AIM performed significantly better than MCC in the broad class categories of Fricatives, Silence, Stops and Total broad class performance. AIM performed significantly better than PAM in the broad class categories of Stops, Vowels and Total broad class performance.

¹Significance in this paper refers to statistically significant at the 95% confidence level with 36 degrees of freedom.

The PAM representation performed significantly better than AIM in the broad class categories of Glides and Silence. PAM performed significantly better than MCC in phoneme recognition performance and the broad class categories of Fricatives, Silence, Stops, Vowels, and Total broad class performance.

7. SPEAKER RECOGNITION

Sambur [10] has shown that, in general, vowels are the best broad phoneme class to extract features from for speaker recognition. Using this as a guideline, speaker recognition experiments were performed using only the vowels. Speaker dependent codebooks were created using Kohonen learning as described above. Each codebook contained 64 elements in an 8 x 8 grid. The codebooks were trained with 40 epochs of the data. The learning rate was a monotonically decreasing piece-wise linear sawtooth pattern. The vowel data used for training the speaker dependent codebooks were from 7 sentences (the si and sx sentences) from the test set above. This provided a 37 speaker data base for speaker identification. The test set consisted of vowel data from 1 sentence (the sa1 sentence) from each speaker.

Speaker recognition was based on minimum average distortion defined over all speaker codebooks and N frames. For each speaker s in the database, the distortion, D_s , was calculated by,

$$D_s = \frac{1}{N} \sum_{i=1}^N \min_{j \in k} \|x_i - m_{sj}\|^2$$

where the index over m_{sj} codewords was $k = 1, \dots, 64$ and the x_i 's were feature vectors (frames) from an unknown speaker u . Speaker u was recognized as the speaker with the minimum distortion, \hat{D}_s .

Table III shows the speaker recognition results for the 37 speakers. The performance of AIM compares well with that of MCC. PAM does not do nearly as well. Previous experiments using PAM for speaker recognition have shown that its' performance is very sensitive to input gain.

TABLE III
SPEAKER RECOGNITION RESULTS

	MCC	AIM	PAM
# Correct	35	34	25
% Correct	94	91	67

8. CONCLUSIONS

This paper examined different auditory model representations to perform speaker-independent phoneme classification and speaker recognition. The study examined the phoneme classification performance using the self-organizing feature map and LVQ. Speaker recognition performance was examined using self-organizing feature maps. The results indicated that it was possible to assign an acoustic segment to one of 39 phoneme categories with at

least 50% recognition accuracy using either auditory model representation. The auditory models provided a broad class recognition rate of greater than 86% for vowels. The resulting context-independent phoneme-recognition performance was better than that of the SPHINX System [7] with the same feature set (i.e. cepstral coefficients only, no delta or power features).

This work demonstrates the potential of using a two-stage approach to speaker recognition. The first stage performs a classification into broad-class categories and the second stage uses one or more of those categories for speaker recognition. The use of this two-stage technique using specific phoneme class recognition is also possible.

The results presented are the first comparing the performance of AIM to MCC and PAM for speech and speaker recognition. These results demonstrate that AIM provides a significant improvement over MCC for speech recognition and comparable performance for speaker recognition. The improved speech recognition performance could translate into improved speaker recognition performance in a two stage speaker recognition system as described above.

This initial examination of AIM for speaker recognition shows promise. Whereas distortion metrics and signal processing methods have been extensively developed for LPC and cepstral representations, these currently do not exist for auditory model representations. Improvements in auditory modeling should continue to be exploited for speech and speaker recognition. Future research will examine temporal aspects of the auditory periphery models, such as the strobed triggered-temporal integration of AIM, and its use in speech and speaker recognition.

9. ACKNOWLEDGEMENTS

The author would like to express his thanks to Janet Slifka, Systems Research Laboratories, Inc, who developed the Kohonen code and many other utilities that greatly aided in the research and for reading and commenting on an earlier version of this paper. This work was supported in part by the Air Force Office of Scientific Research (AFOSR) through its Window-on-Science Program and through AFOSR Task 2313V3.

REFERENCES

- [1] T. R. Anderson. A comparison of auditory models for speaker-independent phoneme recognition. In *Int. Conf. on Acoustics, Speech, and Signal Processing*, volume IV, pages 231-234, 1993.
- [2] J. M. Colombi, T. R. Anderson, S.K. Rogers, D.W. Ruck, and G. T. Warhola. Auditory model representation for speaker recognition. In *Int. Conf. on Acoustics, Speech, and Signal Processing*, volume II, pages 700-703, 1993.
- [3] T. Kohonen. The neural phonetic typewriter. *IEEE Computer Magazine*, 21:11-22, 1988.
- [4] T. Kohonen. *Self-Organization and Associative Memory*. Springer-Verlag, Berlin, third edition, 1989.
- [5] L. Lamel, R. Kassel, and S. Seneff. Speech database development: Design and analysis of the acoustic-phonetic corpus. In *DARPA Speech Understanding Workshop*, pages 100-109, 1986.
- [6] K. F. Lee. *Automatic Speech Recognition: The Development of the SPHINX System*. Kluwer Academic, Boston, 1989.
- [7] K.F. Lee and H.W. Hon. Speaker-independent phone recognition using hidden markov models. *IEEE Trans. ASSP*, 37:1621-1648, November 1989.

- [8] R. D. Patterson and J. Holdsworth. A functional model of neural activity patterns and auditory images. In W. A. Ainsworth, editor, *Advances in Speech, Hearing and Language Processing*. JAI Press, 1994.
- [9] K. L. Payton. Vowel processing by a model of the auditory periphery: A comparison to eighth-nerve responses. *J. Acoust. Soc. Amer.*, 83:145-162, 1988.
- [10] M. R. Sambur. Selection of acoustic features for speaker identification. *IEEE Trans. of ASSP*, ASSP-23(2):176-182, April 1975.
- [11] F. K. Soong and A. E. Rosenburg. On the use of instantaneous and transitional spectral information in speaker recognition. *IEEE Trans. ASSP*, 36:871-879, June 1988.
- [12] F. K. Soong, A. E. Rosenburg, L. R. Rabiner, and B.-H. Juang. A vector quantization approach to speaker recognition. In *Int. Conf. on Acoustics, Speech, and Signal Processing*, volume 1, pages 387-390, 1985.