AUDITORY MODEL REPRESENTATION FOR SPEAKER RECOGNITION

John M. Colombi[†], *Timothy R. Anderson*[‡], *Steven K. Rogers*[†], *Dennis W. Ruck*[†], *Gregory T. Warhola*[†]

[†]AFIT/EN, Wright-Patterson AFB, OH, 45433
[‡]AL/CFBA, Wright-Patterson AFB, OH, 45433

ABSTRACT

This examination of the KING database compares proven spectral processing techniques to an auditory model representation for speaker recognition. The feature sets compared were LPC cepstral coefficients and auditory nerve firing rates provided by the Payton model. The two feature sets were quantized by two clustering algorithms, a Linde-Buzo-Gray (LBG) algorithm and a Kohonen self-organizing feature map. The resulting Vector Quantized (VQ) distortion based classification indicates the auditory model provides comparable accuracies to LPC cepstral in non-studio quality environments and over multiple sessions. For a 10 speaker subset using only voiced frames of 15 second segments, both achieve over 80% identification rate. Cepstral performs better on verification tasks measured with receiver operating characteristics (ROC) curves.

1. INTRODUCTION

Currently, much of speaker recognition research uses proven LPC cepstral and various weighted and transitional derivatives of the voice production model [1]. These have been shown to be better feature sets than other spectral representations. Success has been reported on VQ techniques as an effective classification for speaker recognition [2]. More recent research demonstrates successful classification with Hidden Markov Models, Gaussian mixture methods and artificial neural classifiers. Concurrent to this effort, it was shown that an auditory model by Seneff can be used effectively for speaker identification using several types of classifiers [3] on the TIMIT database.

Using a subset of the KING database, this paper examines the performance of auditory mean firing rate responses compared to various cepstral representations using VQ distortion-based recognition methods. Codebooks are created for each speaker and minimum MSE is used for classification.

2. EXPERIMENTAL METHODS

2.1. Databases

The King Database consists of 51 speakers collected in 10 sessions, speaking on several tasks for approximately one minute each session. The narrowband speech is recorded over long distance telephone lines and sampled at 8 kHz. Typical evaluation on this database consists of training on the first 3 sessions, and testing on sessions 4 and 5.

The first 10 speakers of the KING database (sessions 1-5) were used in the comparison (all male). The data was framed using a 32 msec Hamming window, stepped every 10.6 msecs. Tenth order cepstral coefficients were derived from 10th order LPC coefficients, calculated using the autocorrelation method. Each frame was tagged with a probability of voicing using an algorithm similar to the pitch tracking algorithm of Secrest and Doddington [4], and a segment of 15 seconds was chosen per utterance which contained the maximum consecutive voicing.

2.2. Payton Auditory Model

The Payton auditory model is a composite collection of stages based on physiological data [5]. The model accepts sampled data and provides predicted neural firing responses for 20 points along the basilar membrane, corresponding to center frequencies of 440Hz to 6600Hz. See Figure 1. This model is unique in that the displacement of the basilar membrane with respect to time and location is modeled, processing biologically plausible variables of fluid dynamics, damping, stiffness, size and shape of the membrane. The output of this basilar membrane section is sharpened and the displacement stimuli is input to the non-linear transduction process of the inner hair cells/synapse. Other auditory models only approximate this displacement and transduction through a series of filterbanks [6]. Comparison of Payton's representation to other models for phoneme recognition is demonstrated by Anderson [7].

It was necessary to scale each of the utterances to drive the auditory model to approximate conversational levels, insuring not to saturate the neural responses. The Payton model references 0 dB with respect to a 1 kHz sine wave of certain energy, enough to drive firing of the 1 kHz Center Frequency synapse to threshold, a firing level equal to 10% of its dynamic range. A scaling value of 8000 was experimentally determined to provide adequate firing responses, corresponding to 47 dB model reference. The responses of the model were averaged over 32 msec, calculated every 10.6 msec; thus, corresponding frames were identical to the cepstral representation.

The Payton model is extremely computationally complex, which necessitated the choice of 15 second segments. Due to numerical considerations in solving the basilar membrane equations, the model is currently required to run at 160 kHz. This overall impact is that processing takes approximately 1000 times real time. Other auditory periphery models, based on filterbank designs, are computationally more effi-

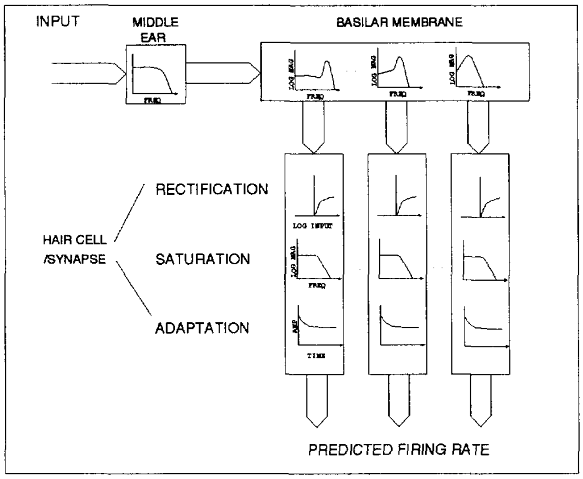

Figure 1: Payton Composite Model [5]

cient, yet only model the general characteristics of basilar membrane displacement and neural responses.

2.3. Recognition

Prior experimentation has shown voiced speech carries more speaker dependent information than unvoiced speech; all high probability of voicing frames within the 15 second segments were selected for quantization. Speaker dependent codebooks were created, using the iterative LBG splitting algorithm with convergence threshold .05 and 64 codewords. Kohonen self-organizing feature maps were also tested, running approximately 40 epochs, using linear decreasing learning rate and neighborhood radius, and map size of 8 x 8 nodes [8]. Many parameters of Kohonen learning must be experimentally determined [9] such as learning schedule (linear, exponential, hyperbolic), neighborhood function(linear, Gaussian), output neighborhood structure (2D, 3D, etc.) and conscience implementation. Since neighborhood preservation was not required for classification, effects of Kohonen learning without neighborhood were examined. The learning schedule chosen was a monotonically decreasing piece-wise linear "saw-tooth" pattern, which had previously demonstrated improved phoneme recognition.

Identification is based on minimum average distortion defined over all speaker codebooks and N frames. For speaker s, the distortion, D_s, is,

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

where the index over m_{sj} codewords is $k = 1, \ldots, 64$. Other distortion metrics for cepstral vectors include weighted cepstral (diagonal covariance), Mahalanobis (full covariance) and Root Power Sums (index^2), examined in [9]. For verification, the classic method iteratively thresholds distortions calculated using speaker dependent codebooks and examines probability of detection (P_d) of targets and probability of false alarm (P_{fa}) of imposters.

Transitional characteristics of the processed speech signal also contain speaker dependent information using polynomial expansions [10], linear regression [1] or differenced coefficients [11]. Lee [11], in preliminary tests for the SPHINX system, settled on only differenced coefficients using a 40 msec window, symmetric about the current frame. The m differenced (or delta) coefficient at time t is simply,

$$d_m(t) = c_m(t+\delta) - c_m(t-\delta) \qquad (1)$$

Others have used slightly larger windows, such as 90 msec [1].

3. EXPERIMENTS AND RESULTS

3.1. Identification

Results for baseline 10 class identification are provided in Tables 1 and 2 using quantizer designs of LBG, Kohonen and Kohonen without neighborhood (Comp). These results for LBG are in agreement with 15 seconds of testing utterances in published data [12]. The Payton representation was first normalized such that each vector had zero mean [3]. Also, the five higher frequency Payton channels were removed, since these model basilar membrane locations having characteristics frequencies greater than 4 kHz. These 15 coefficients provide better recognition than the baseline cepstral representation.

Table 1: Speaker Identification using 15 second training per session and 15 seconds testing (before voiced segmentation) per session of 10th order LPC cepstral coefficients.

Train	Test	LBG	Kohonen	Comp
Sessions 1,2,3	4,5	80.0%	55.0%	45.0%
Sessions 2,3,4	1,5	75.0%	55.0%	45.0%
Sessions 3,4,5	1,2	70.0%	55.0%	60.0%
Average		75.0%	55.0%	50.0%

Table 2: Speaker Identification using 15 Payton coefficients and zero-mean normalization.

Train	Test	LBG	Kohonen	Comp
Sessions 1,2,3	4,5	85.0%	55.0%	45.0%
Sessions 2,3,4	1,5	75.0%	80.0%	60.0%
Sessions 3,4,5	1,2	80.0%	75.0%	80.0%
Average		80.0%	70.0%	61.7%

A recent article [13] reported increases with liftering techniques on cepstral coefficients, both on individual vectors using bandpass liftering, and temporally over sequences of vectors using RASTA liftering. Bandpass liftering [14] deemphasizes the low and high order cepstral components using a raised sinusoid window.

$$w(k) = 1 + L/2\sin(\pi k/L)$$

where $1 \leq k \leq L$ are the cepstral coefficient index. Removal of cepstral time averages may reduce transmission and recording characteristic effects [15]. This normalization and liftering techniques were applied to the 10 speaker tests and shown in Table 3 using LBG design.

Table 3: Speaker Identification using bandpass liftering cepstral time average removal.

Train	Test	Lifter	Remove Mean
Sessions 1,2,3	4,5	75.0%	65.0%
Sessions 2,3,4	1,5	70.0%	70.0%
Sessions 3,4,5	1,2	75.0%	70.0%
Average		73.3%	68.3%

A series of delta coefficients were extracted from the KING instantaneous cepstral vectors. No documentation has been reported on successful use of temporal characteristics on the KING narrowband corpus. Various codebooks were created for increasing delta windows of ± 1 frame to ±6 frames. This corresponds to window times of 21.2 msec to 106 msec. Performance increases over instantaneous cepstral were demonstrated by up to 13.3%. Shown in 2 are various window sizes and the effects of bandpass liftering and time average removal for the cepstral representation.

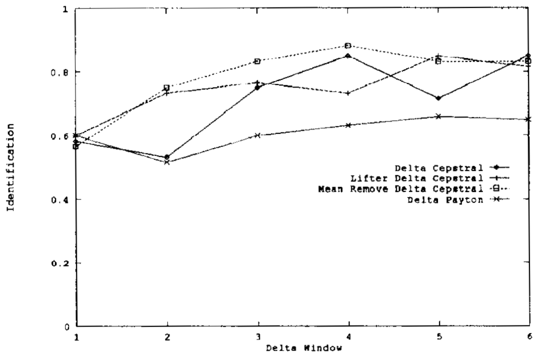

Figure 2: Speaker Identification using Δ cepstral and Δ Payton coefficients using LBG quantizer design. Each point is the average of the three training and test method.

A series of delta windows were examined for the Payton model, also shown in Figure 2. This technique for the auditory model attempted to capture temporal firing information without specifically estimating neural pulse trains. Since it has been shown that delta cepstral contains uncorrelated information to that of instantaneous cepstral, this technique was applied to Payton. Improvements were not demonstrated over instantaneous coefficients.

3.2. Verification

Verification is performed using a technique recently documented by Kao et al [12]. LBG Codebooks are designed using speech from a set of targets, and testing using both targets and imposters, again using the 15 second segments per session. For KING, speakers 1 - 10 were used as targets, and speakers 14 - 26 were used as imposters, also The receiver operating characteristics for both LPC cepstral and Payton are shown in Figure 3. Training and testing for verification uses data from session 1 - 3 and sessions 4 - 5 respectively. Verification is significantly worse using the delta representations, shown in Figure 4.

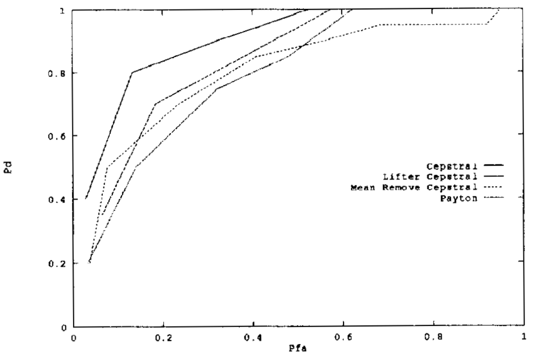

Figure 3: Speaker Verification ROC using cepstral and Payton coefficients. Cepstral coefficients are also liftered and mean removed. Payton 15 coefficients are zero mean normalized.

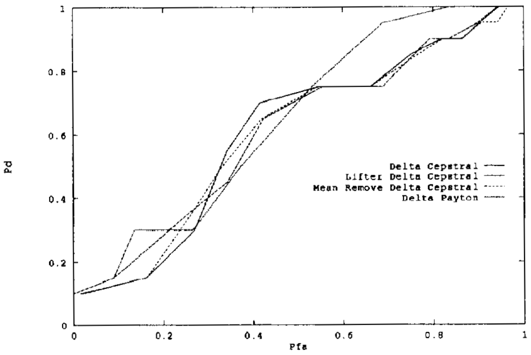

Figure 4: Speaker Verification ROC of Δ cepstral and Δ Payton distortions derived from codebooks using a δ of ±4.

4. SUMMARY

Initial results using a biologically motivated model to represent speaker spectral content were demonstrated and compared to cepstral, using the same sampled data for both representations. Zero-mean normalized, mean-rate response Payton outputs using LBG quantization provided better performance than cepstral with and without liftering or mean removal. Differenced cepstral coefficients provided better identification than instantaneous cepstral; delta payton provided no improvements. Verification, as measured by ROC curves show comparable results. Improvements would naturally be seen, as documented in [12, 13], using greater amounts of training and testing data. Classification was provided by overall LMSE distortion using 64 codeword codebooks. The only other result reported on an auditory model representation for speaker identification used the TIMIT corpus, which is studio-quality, single session data; KING provides a more realistic corpus. Past KING research has developed particular train and test scenarios. This article presents results for a hold-two-out training pattern, showing how results do vary by as much as 35% between training sessions, depending on codebook design.

These cepstral results differ from past research in the amount of data used for quantizer design and test, constricted by Payton processing. The 15 second windows typically held 300 to 500 frames per session or 3 to 5 seconds of high voiced speech, as tagged by the pitch tracking probability of voicing mechanism. Choosing high ($\geq .9$) probability frames gave improved performance over low ($\geq .1$) ones. These latter frames contain transitional areas into and out of high voiced (vowel) areas. This leads us to believe high voiced areas may be better for speaker recognition than using a speech/ non-speech segmentation before the quantization, as is often performed.

This initial examination of an auditory representation for speaker identification shows promise. Whereas distortion metrics and signal processing methods have been extensively developed for LPC and cepstral representations, these currently do no exist for neural data. Improvements in auditory modeling, often used for physiological understanding, should continued to be exploited for speech and speaker recognition. Future research will examine other temporal aspects of the neural signals of the auditory periphery, such as some form of general phase synchrony (Seneff's GSD) or localized synchrony responses (Sach and Young's ALSR).

5. ACKNOWLEDGEMENT

The authors would like to thank Janet Slifka, Systems Research Laboratories, Inc., for development of several Entropic ESPS utilities used throughout this work.

REFERENCES

[1] F. K. Soong and A. E. Rosenburg, "On the use of instantaneous and transitional spectral information in speaker recognition," *IEEE Trans. ASSP*, vol. 36, pp. 871–79, June 1988.

[2] Soong, Rosenburg, Rabiner, and Juang, "A vector quantization approach to speaker recognition," in *ICASSP 85*, vol. 1, pp. 387–390, 1985.

[3] H. Hattori, "Text-independent speaker recognition using neural networks," in *ICASSP 92*, vol. 2, pp. 153–156, 1992.

[4] B. Secrest and G. Doddington, "An integrated pitch tracking algorithm for speech systems," *ICASSP 83*, pp. 1352 – 1355, 1983.

[5] K. L. Payton, "Vowel processing by a model of the auditory periphery: A comparison to eighth-nerve responses," *J. Acoust. Soc. Amer.*, vol. 83, no. 1, pp. 145–162, 1988.

[6] S. Seneff, "A joint synchrony/mean-rate model of auditory speech processing," *Journal of Phonetics*, vol. 16, pp. 55 – 75, 1988.

[7] T. R. Anderson, "A comparison of auditory models for speaker independent phoneme recognition." ICASSP 93, 1993.

[8] T. Kohonen, "Tutorial/ self-organizing maps," *Proceedings of the 1988 International Conference of Neural Networks*, 1988.

[9] J. Colombi, "Comparison of cepstral and auditory model representation for speaker recognition," Master's thesis, School of Engineering, Air Force Institute of Technology (AU), Wright-Patterson AFB OH, Dec. 1992.

[10] S. Furui, "Cepstral analysis techniques for automatic speaker verification," *IEEE Trans. ASSP*, vol. 29, pp. 254–72, Apr. 1981.

[11] K.-F. Lee, "An overview of the sphinx speech recognition system," *IEEE Trans. ASSP*, vol. 38, January 1990.

[12] Y.-H. Kao, P. K. Rajasekaran, and J. S. Baras, "Free-text identification over long distance telephone channel using hyposthesized phonetic segmentation," in *ICASSP 92*, vol. 2, pp. 177–180, 1992.

[13] Y.-H. Kao, P. K. Rajasekaran, and J. S. Baras, "Robust free-text speaker identification over long distance telephone channels." ICASSP 93, 1993.

[14] B.-H. Juang, L. R. Rabiner, and J. G. Wilpon, "On the use of bandpass liftering in speech recognition," *I.E.E.E Trans of ASSP*, vol. 35, July 1987.

[15] B. S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *J. Acoust. Soc. Amer.*, pp. 1304–12, June 1974.