S10.12

SPEAKER-INDEPENDENT VOWEL RECOGNITION: SPECTROGRAMS VERSUS COCHLEAGRAMS

Yeshwant K. Muthusamy, Ronald A. Cole

Department of Computer Science and Engineering
Oregon Graduate Institute of Science and Technology
19600 NW Von Neumann Dr., Beaverton, OR 97006

Malcolm Slaney

Advanced Technology Group
Apple Computer, Inc.
20525 Mariani Avenue, Cupertino, CA 95014

ABSTRACT

We examined the ability of multi-layer perceptrons (MLPs) trained with backpropagation to classify vowels excised from natural continuous speech. Two spectral representations were compared: spectrograms and cochleagrams. The features used to train the MLPs included DFT or cochleagram coefficients from a single frame in the middle of the vowel, or coefficients from each third of the vowel. We also investigated the effect of three additional features – estimates of pitch, duration and the relative amplitude of the vowel. Our experiments showed that *with coefficients alone, the cochleagram was* superior to the spectrogram in classification performance for all experimental conditions. With the three additional features, however, the results were comparable. Perceptual experiments with trained human listeners on the same data set revealed that MLPs perform much better than humans on vowels excised from context.

1. INTRODUCTION

A common aspect of vowel classification experiments with artificial neural networks (ANNs) has been the use of some type of spectral representation of the speech signal as input to the network. Networks trained on spectral coefficients extracted from multiple frames in the vowel have displayed performance comparable to that of human listeners [1,2].

Leung [3] used an auditory-based spectral representation [4] and MLPs to classify the 16 vowels and diphthongs of American English in natural continuous speech. The vowel tokens were excised from all phonetic contexts in sentences of the TIMIT database, a standardized acoustic phonetic corpus of continuous speech, displaying a wide range of American dialectical variation [5,6]. Recognition accuracy of 60% was obtained with spectral information alone, and 77% with the addition of duration and phonetic context.

Recent experiments have shown the superiority of auditory models over conventional representations (e.g. DFT, filterbank mel-cepstrum) for recognition of words and digits under noisy conditions [7,8]. However, there have been no direct comparisons of auditory models and conventional representations for classification of vowels in natural continuous speech. In this paper, we report results of experiments comparing two spectral representations: the constant-increment DFT and the cochleagram, a computational model of the peripheral auditory system [9,10].

We address the following questions about the ability of MLPs to perform speaker-independent vowel classification using these two representations:

- How much information does a *single* spectral slice convey about the identity of a vowel?
- How does performance improve with the use of spectra from each third of the vowel?
- What is the effect of adding three additional features – median pitch, duration, and relative amplitude – to the spectral information?
- How does classification performance compare to that of trained human listeners on the same task?

2. THE COCHLEAGRAM

We use a cochlear model designed by Lyon [9] and described by Slaney [10] to convert a sound waveform into a multidimensional vector that represents the information sent from the ear to the brain. This system is diagrammed in Figure 1. It is important to remember that the cochlear model used here does not try to accurately model the internal structure of the ear but only to approximate the information contained in the auditory nerve.

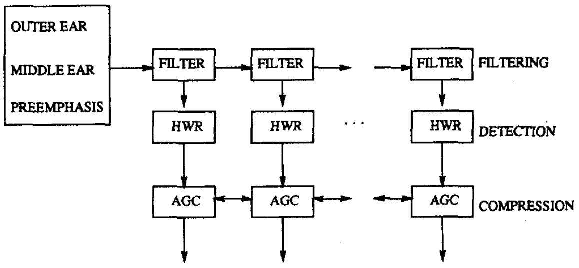

Figure 1. Lyon's Cochlear Model

The cochlear model in this study uses a relatively simple filter to simulate the response of the outer and middle ears. A cascade of second order filters is used to model the propagation of sound along the *basilar membrane.* At each point along the cochlea, the basilar membrane responds best to a broad range of frequencies and it is this movement that is sensed by the *inner hair cells.* The "best" frequency of the cochlea varies smoothly from high frequencies at the base to low frequencies at the apex. Inner hair cells only respond to movement of the basilar membrane in one direction. This is simulated in Lyon's model with an array of Half Wave Rectifiers (HWRs) that detect the output of each second order filter. The HWR nonlinearity serves to convert the motion of the basilar membrane at each point along the cochlea into a signal that represents the energy of the acoustic input and retains the fine time structure.

CH2847-2/90/0000-0533 $1.00 © 1990 IEEE

Finally, four stages of Automatic Gain Control (AGC) allow the cochlear model to compress the dynamic range of the input to a level that can be carried on the auditory nerve. The AGC used here also serves to simulate the ear's adaptation to loud sounds. Sixty-four of the lowest frequency output channels from this cochlear model, spanning the range 0.1–4 kHz, are used in this experiment.

3. EXPERIMENTS

3.1. Stimuli

The stimuli for the experiments consisted of featural descriptions of the 12 monophthongal vowels of English, shown in Table 1. The vowels were excised from all phonetic contexts in utterances of the TIMIT database. The diphthongs /oy/, /ay/, /ey/, /aw/ were excluded because they are characterized by spectral change, and are therefore inappropriate for experiments using information from a single spectral slice.

Table 1. The 12 Vowel Classes

Phone	Example	Phone	Example
/iy/	b*ea*t	/ah/	b*u*tt
/ih/	b*i*t	/uw/	b*oo*t
/eh/	b*e*t	/uh/	b*oo*k
/ae/	b*a*t	/ao/	b*ough*t
/ix/	ros*e*s	/aa/	c*o*t
/ax/	th*e*	/er/	b*ir*d

In all experiments, the training set consisted of 342 exemplars of each vowel provided by 320 speakers, for a total of 4104 vowel spectra. The test set consisted of 137 exemplars of each vowel, provided by a different set of 100 speakers, for a total of 1644 vowel spectra.[1]

3.2. Experimental Design

In the first set of experiments, we compared the performance of artificial neural networks trained on spectrogram or cochleagram coefficients, under the following conditions: i) 64 coefficients from the center frame of each vowel token (explained in the next section), ii) 64 coefficients taken from the center frame of each third of the vowel ($3 X 64 = 192$ coefficients), and iii) the averaged coefficients from each third of the vowel ($3 X 64 = 192$ coefficients).

In a second set of experiments, the coefficients in each of the three conditions above were augmented by three additional features – estimates of the fundamental frequency (median pitch), duration and relative amplitude of the vowel. In addition, we examined the effects of the three features, added individually and in pairs, to the cochleagram.

Finally, perceptual experiments were conducted on five trained human listeners to compare their performance with that of the MLPs.

3.3. Spectral Representations

Two spectral representations were compared: (a) Constant-increment discrete Fourier transform (DFT), and (b) the cochleagram.

DFT. A 256-point real DFT was computed on each utterance, with a 10 ms Hanning window and 3 ms increment, yielding 128 spectral coefficients (spanning the range 0–8 kHz). Since the important information about vowel identity is found below 4 kHz, only the first 64 spectral coefficients (0–4 kHz) from each frame were used. The center frame of each vowel token was located using the hand-segmented phonetic transcriptions provided in the TIMIT database.

Cochleagram. Unlike the DFT which uses a linear frequency scale from 0 to 8 kHz, the cochleagram uses a Bark scale. The range 0–8kHz is encompassed by 84 spectral coefficients. However, the first 64 spectral coefficients span the range 0.1–4 kHz. Thus the number of spectral coefficients used was the same for the two representations.

The coefficient values were normalized to lie between 0 and 1 in order to train the neural networks. Normalization was done by computing the "relative value" of each coefficient with respect to the maxima and minima of all 64 coefficients in each frame:

$$\textit{normalized value} = \frac{(X - min)}{(max - min)} \qquad (1)$$

where X is the value of any spectral coefficient, max is the value of the largest of the 64 spectral coefficients, and min is the value of the smallest of the 64 spectral coefficients.

3.4. Additional Features: Pitch, Duration and Relative Amplitude

Three additional features were computed for both the training and test utterances and appended to the corresponding feature vectors in the DFT and cochleagram representations.

Median Pitch. Pitch peaks were located automatically using a neural network classifier [11]. The median pitch was calculated based on the 10 pitch peaks closest to the center of the vowel.

Duration. The duration estimate was taken from the phonetic transcriptions provided in the TIMIT database.

Relative Amplitude. The amplitude estimate was based on the peak-to-peak amplitude computed in a 10 msec window in the filtered waveform between 0 and 700 Hz. The relative amplitude of the vowel was the maximum peak-to-peak amplitude in a 30 ms window around the center of the vowel, divided by the maximum peak-to-peak amplitude in a larger window, extending 300 ms behind and 250 ms ahead of the vowel.

3.5. Procedûre

The neural network classifiers were fully connected feed-forward networks (no recurrent links). The number of input units of the network was determined by the number of features used in the experiment (e.g., 64 or 192). All of the networks had 12 output units, corresponding to the 12 vowel categories. The number of hidden layers and the number of units in each hidden layer (one or two) was determined experimentally. We parametrically investigated network configurations with one and two hidden layers and different numbers of hidden units in each layer.

The networks were trained using backpropagation with conjugate gradient optimization [12]. The procedure for training and testing a network proceeded as follows: The network was trained on 100 iterations through the 4104 training vectors. The trained network was then evaluated on the training set and

[1] The number of tokens of each vowel class was determined by the number of tokens in the least frequent class. It was found that the vowel */uh/* had the least number of tokens, 342 in the training set, and 137 in the test set. (By comparison, the corresponding figures for the most frequent vowel class */iz/* were 5798 and 1809 respectively). Thus, for each of the remaining 11 vowel classes, 342 tokens were selected by iterating through all the 320 speakers, picking one token at random from each speaker, until 342 different tokens were obtained. This procedure ensured that there was wide across-speaker variation in the tokens selected. A similar procedure was followed in creating the test set (137 tokens per class).

the 1644 test vectors. This process was continued and the performance of the network on the training and test vectors was recorded after every 100 iterations through the training set. The training was stopped when the network had converged; convergence was observed as a consistent decrease or leveling off of the classification percentage on the test data over successive sets of 100 iterations. Typically, the networks converged after 1100–1200 iterations and took 35–40 hours on a Sun 4/60.

4. RESULTS

Table 2 shows the classification performance of the networks on the DFT and cochleagram coefficients, for the three experimental conditions. It can be seen that (a) with spectral coefficients alone, the networks performed consistently better on the cochleagram than on the DFT in all three experimental conditions, and (b) with the cochleagram, the networks showed comparable performance with 3-slice data and the 3 averaged spectra. With the DFT, the 3 averaged spectra produced better performance than 3 slices. Both these results suggest that a single time frame of the cochleagram provides more information about a vowel's identity than a corresponding slice of the DFT.

Table 2. DFT vs. Cochleagram (Spectral Coefficients)

Hidden Units	DFT			Cochleagrams		
	1 slice	3 slices	Averaged Thirds	1 slice	3 slices	Averaged Thirds
16	46.72	48.36	50.91	47.93	53.09	54.68
32	45.80	48.97	52.74	50.79	53.80	55.29
40	47.38	49.39	52.62	50.06	54.70	55.05
32-16	48.42	48.05	52.25	51.40	55.66	55.66

We believe two features of the cochlear model are responsible for the superior performance of the cochleagram. First, the HWR nonlinearity implements a simple form of phase locking. Phase locking, or the tendency for nerve cells to lock to the dominant frequency, has been shown to be important in representing formants [13]. Secondly, the information content of a message is independent of its loudness. The AGC inherent in a cochlear model effectively removes the amplitude information from the signal while it emphasizes critical sound onsets.

Table 3 shows the classification performance of the networks on the DFT and cochleagram coefficients augmented with pitch, duration and amplitude, for the three experimental conditions. It can be seen that the addition of median pitch, duration and relative amplitude substantially improved classification performance for the DFT, but resulted in comparable or even worse performance for the cochleagrams.

Analysis of confusion matrices revealed that the main benefit of adding PDA estimates to the DFT was a reduction in the number of confusions between /ix/ – /ih/, /ao/ – /aa/, and /ax/ – /ah/. No such pattern was discernible with the cochleagram.

Table 3. DFT vs. Cochleagram (Spectral coefficients with PDA)

Hidden Units	DFT			Cochleagrams		
	1 slice	3 slices	Averaged Thirds	1 slice	3 slices	Averaged Thirds
16	55.35	54.74	57.00	53.65	54.05	54.70
32	56.02	55.47	58.58	53.10	57.27	55.73
40	55.66	55.35	58.39	54.44	56.31	56.69
32-16	56.27	55.35	57.79	54.38	56.24	55.08

To examine the individual effects of the three features on the cochleagram, additional experiments were conducted. In these experiments, MLPs (each with 40 hidden units) were trained and tested on the 192 cochleagram coefficients (representing the averaged thirds of each vowel) augmented with all combinations of the three features – pitch alone, duration alone, amplitude alone, pitch and duration, and so on.

The results are tabulated in Table 4. The result with all 3 features added (PDA) is included for comparison. It can be seen that there are small differences between the seven feature conditions, but no evidence that the individual features improve performance beyond that obtained with the cochleagram coefficients alone.

5. LISTENING EXPERIMENTS

In order to better interpret these results, we conducted listening experiments using the vowel tokens in the training and test sets. These experiments allowed us to compare human vowel identification performance with that of the neural network classifiers. Although many vowel recognition experiments have been performed, none have used a large number of monophthongal vowels excised from continuous speech from a variety of contexts, with sufficient training on this type of classification task. We believe that a lengthy training procedure is necessary, since subjects are not used to hearing segments removed from fluent speech. Training on a large set of tokens, with feedback on each trial, provides a fair estimate of listeners' subsequent classification performance (without feedback) on the test set.

5.1. Stimuli

The stimuli for the listening experiments consisted of a subset of the vowels used in the training and test sets described above, excised from utterances in the TIMIT database. The training set for these experiments consisted of 900 vowels drawn at random from the larger set of 4104 vowel tokens. The test set had 600 vowel tokens (50 of each class), drawn from the larger test set of 1644 vowel tokens.

The boundaries of each segment were located based on the phonetic labels. Using these boundaries, the actual vowel onset and offset points were selected so as to minimize the transients in the excised segment. The vowel segment was then converted to sound using the digital to analog converter (0–4 kHz) in a Sun 4/60, and presented to the subject over a loudspeaker.

Table 4. Averaged Thirds of the Cochleagram with combinations of P, D and A (40 hidden-unit network)

Feature(s)	P	D	A	PD	PA	DA	PDA
% correct	56.56	55.47	56.18	55.98	56.82	56.95	56.69

5.2. Procedure

Training. Five male subjects from the OGI Speech Group served as subjects. In the training phase, the subjects were asked to identify the sounds chosen at random from the training set. Words containing the 12 vowel sounds (e.g., b*ee*t, b*i*t, b*e*t) were displayed on the console. The subjects could listen to each vowel sound as many times as needed. They indicated their response by clicking the mouse on the appropriate word on the display. Feedback was given on each trial. Each subject went through 9 sessions listening to 100 vowels in each session.

Testing. In the test phase, the subjects were presented with the 600 vowel sounds from the test set in sessions of 100, *with no feedback*. All the listeners were tested on the same 600 tokens. As in the training phase, they could listen to each vowel sound as many times as needed.

5.3. Results

On the training set, the average performance of the 5 listeners was 45.5%. To track improvement in performance over the 900 trials, we computed the average performance over all listeners for 3 blocks of 300 trials each. The numbers were 42%, 46% and 48%, showing an increase in performance with increased training.

On the test set, the average performance (listener-to-label agreement) of the 5 listeners was 49.2%. Individual listener performances ranged from 46.1% to 51.1%. The average listener-to-listener agreement was 53.1%. The pairwise listener-to-listener agreement ranged from 47.7% to 59.3%. Thus, the listeners seemed to agree with each other more often than they did with the phonetic labels.

6. DISCUSSION

The information in a single spectral slice enables neural networks to distinguish between 12 spectrally confusable vowel classes with an accuracy of about 48% (DFT) and 51% (cochleagram), as opposed to 8.33% by chance. When three slices or averaged spectra are used, performance improves to about 53% (averaged thirds of the DFT) and 56% (averaged thirds of the cochleagram). This performance can be further improved to about 59% (averaged thirds of the DFT with PDA) by the addition of features that capture important information in the vowel segment. However, these features do not produce any improvement when added to the cochleagram.

Using coefficients alone, we observed a consistent superiority of the cochleagram over the DFT in all experimental conditions. Although the difference was small (3% to 5%), it was observed over all network configurations in 12 experimental conditions. This indicates that the cochleagram captures more of the phonetic information in the vowel than the DFT. However, the relatively poor performance of the cochleagram with the additional features defies explanation. We are currently conducting more experiments to analyze these results.

An interesting result was the average performance of 49% by listeners, compared to 59% by the best network configuration. One explanation for the relatively poor performance of human listeners on these vowel sounds is the lack of contextual information in the stimuli. Since vowel sounds undergo considerable restructuring due to coarticulation effects, context is essential for accurate human recognition of these vowel sounds. Phillips [14] presented listeners with segments excised from continuous speech from a set of 19 vowel sounds, including diphthongs, with right and left phonetic context. The average listener-to-listener agreement on the labels was about 65%. The discrepancy between the two results can be attributed to the lack of contextual information in our stimuli.

7. ACKNOWLEDGEMENTS

The authors would like to thank Nicholas Horton, Richard Lyon and Mark Fanty for their insightful comments and discussions; Steven Neighorn and Jon Inouye for their immense help in setting up the listening experiments; and Vincent Weatherill and Murali Gopalakrishnan for participating in the listening experiments.

References

1. H. C. Leung and Victor W. Zue, "Some phonetic recognition experiments using artificial neural nets," pp. 422-426 in *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing*, (1988).
2. H.C. Leung and V.W. Zue, "Application of Error-backpropagation to Phonetic Classification," *IEEE '88 Neural Information Processing Systems - Natural and Synthetic*, (Nov. 28 - Dec. 1, 1988.).
3. H.C. Leung, *The Use of Artificial Neural Networks for Phonetic Recognition*, PhD Thesis, Massachusetts Institute of Technology (May 1989).
4. S. Seneff, "A computational model for the peripheral auditory system: Application to speech recognition research," pp. 37.8.1-37.8.4 in *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing*, (April, 1986).
5. W. Fisher, G. Doddington, and K. Goudie-Marshall, "The DARPA speech recognition research database: specification and status," pp. 93-100 in *Proceedings of the DARPA Speech Recognition Workshop*, (February, 1986).
6. L. Lamel, R. Kassel, and S. Seneff, "Speech database development: design and analysis of the acoustic-phonetic corpus," pp. 100-110 in *Proceedings of the DARPA Speech Recognition Workshop*, (February, 1986).
7. O. Ghitza, "Temporal non-place information in the auditory-nerve firing patterns as a front-end for speech recognition in a noisy environment," *Journal of Phonetics* **16** pp. 109-123 Academic Press Limited, (1988).
8. M. J. Hunt and C. Lefebvre, "Speaker dependent and independent speech recognition experiments with an auditory model," pp. 215-218 in *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing*, (April, 1988).
9. R. F. Lyon, "A Computational Model of Filtering, Detection, and Compression in the Cochlea," in *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing*, (May 1982).
10. M. Slaney, "Lyon's Cochlear Model," Apple Technical Report #13, Apple Computer Inc. (1988).
11. R. A. Cole, E. Barnard, M. Vea, and F. Alleva, "Classification of pitch periods using expert knowledge and neural net classifiers," *Journal of the Acoustical Society of America* **84** p. S60 (A) (1988).
12. E. Barnard and D. Casasent, "Image processing for image understanding with neural nets," in *International Joint Conference on Neural Nets*, (1989). (Submitted for publication.)
13. S.A. Shamma, "Speech processing in the auditory system I: The representation of speech sounds in the responses of the auditory nerve," *Journal of the Acoustical Society of America* **78**(5) pp. 1612-1621 (November 1985).
14. M. Phillips, "Speaker-independent classification of vowels and diphthongs in continuous speech," in *Proc. of the 11th International Congress of Phonetic Sciences*, , Estonia, USSR (1987).