Discriminative and Maximum Likelihood Classifiers for Computer-Based Visual Feedback for Speech Training for the Hearing Impaired

Stephen A. Zahorian and A. Matthew Zimmer
Department of Electrical and Computer Engineering, Old Dominion University
Norfolk, Virginia 23529, USA

ABSTRACT

A visual speech training aid for persons with hearing impairments has been developed using a Windows-based multimedia computer. The training aid provides real time visual feedback as to the quality of pronunciation for 10 steady-state American English monopthong vowel phonemes (/aa/, /iy/, /uw/, /ae/, /er/, /ih/, /eh/, /ao/, /ah/, and /uh/). This training aid is thus referred to as a Vowel Articulation Training Aid (VATA). Neural network (NN) classifiers are used to produce the two main displays: a 10-category "vowel bargraph" which provides "discrete" feedback, and an "ellipse display" which provides continuous feedback over a 2-D field, similar to a formant1-formant2 display often used by speech scientists to depict vowels. Continuous feedback such as this is desirable for speech training to help improve articulation. In this paper a method is described for combining the neural network classifiers used in the speech display with maximum likelihood classifiers which help to reject those sounds which were not used for training the neural network.

Keywords: Vowel classification, Classification decision, Maximum likelihood, Speech recognition, Speech training aid, Spoken language processing, Speech therapy

1. BACKGROUND

Displays Available

The system has two main displays. One is a bargraph display, which gives feedback about how well speech utterances fit into discrete categories. The other is an "ellipse" display, which provides continuously variable feedback about utterances. The system gives feedback for the sounds /ah/, /ee/, /ue/, /ae/, /ur/, /ih/, /eh/, /aw/, /uh/, and /oo/, which correspond to the vowel sounds found in the words "cot," "beet," "boot," "bag," "bird," "pig," "bed," "dog," "cup," and "book" respectively. The labels for this system (/ah/, etc.) were assigned by the ODU speech lab, and correspond to the ARPABET labels /aa/, /iy/, /uw/, /ae/, /er/, /ih/, /eh/, /ao/, /ah/, and /uh/ commonly used in speech processing literature.

The bargraph display (Figure 1) resembles a histogram, with one bar for each vowel sound of interest. The height of the vowel's bar varies in proportion to the accuracy of the speaker's pronunciation of that vowel. Correct pronunciation yields one steady, clearly defined bar, while the rest assume zero or small values. Incorrectly pronounced sounds may produce displays showing two or more partially activated category bars, no activated bars, or rapid fluctuations between bars.

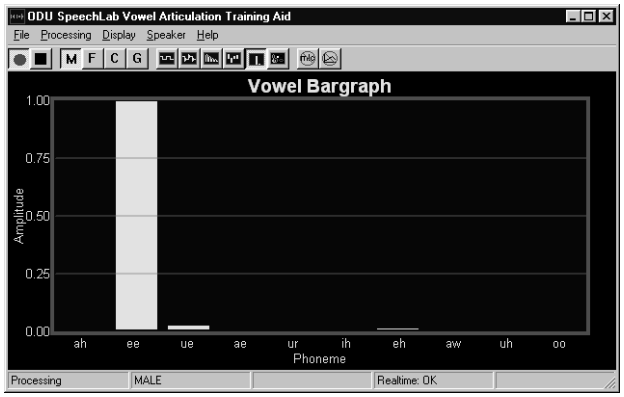

Figure 1. Bargraph display showing response for correct pronunciation of /ee/

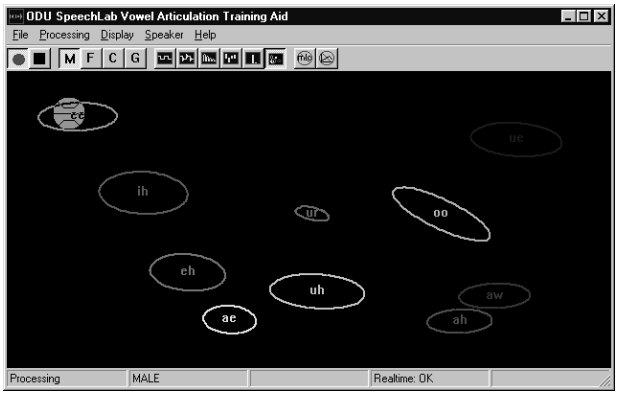

Figure 2. Ellipse Display showing response for correct pronunciation of /ee/

The ellipse display (Figure 2) divides the screen into several elliptical regions, similar to a standard F1/F2 type display. Unlike the F1/F2 display, this 'ellipse' display bases its output on Neural Network Classifier output values obtained using Discrete Cosine Transform Coefficients (DCTCs) of the log magnitude spectrum rather than the formant frequencies of the utterance. Correct pronunciation of a particular sound places a basketball icon within the corresponding ellipse and causes the icon's color to match that of the ellipse. Incorrect or unclear pronunciation results in the ball icon 'wandering' about the screen or coming to rest in an area not enclosed by an ellipse. By observing the continuous motion of the ball, a speaker gains information about how to adjust his or her pronunciation in order to produce the desired vowel sound.

A speaker group selection option with "CHILD," "FEMALE" and "MALE" settings allows both the ellipse and bar graph display modes to be fine-tuned for better classification of sounds produced by child, adult female, or adult male speakers respectively. A fourth speaker group option, "GENERAL" causes the display to attempt to classify sounds produced by a speaker from any of the three other categories.

2. PROCESSING STEPS

A block diagram for the VATA system is shown in Figure 3. The operating system interacts with the sound card to acquire a section of data (a "segment") from the from the continuous audio data stream. Special driver software written for the VSD program interacts with the Windows multimedia services to handle double buffering and ensures that no samples are lost. Since the driver software communicates with the operating system services and not directly the hardware, the VATA system should work with most commonly available Windows-compatible sound cards.

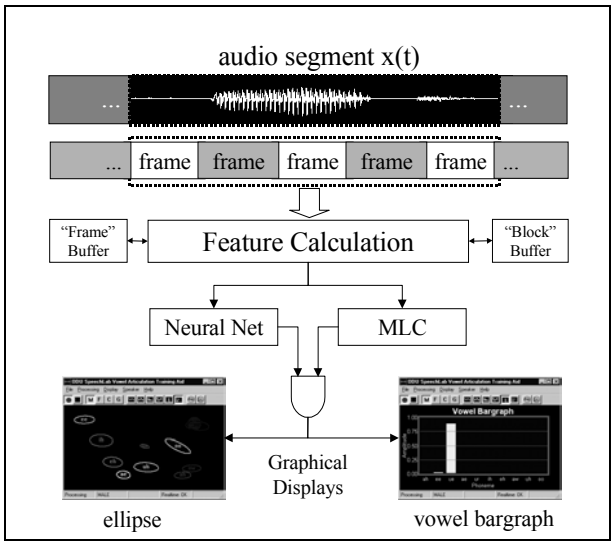

Figure 3. VATA Block Diagram

The first step in signal processing is to divide the audio segment into sub-segments called "frames." The frame is the basic unit on which signal processing is performed. Following division of the signal into frames, signal processing is performed as shown in figure 4. Each frame is first passed through a high-frequency pre-emphasis filter, typically centered at 3.2kHz. Next a Fast-Fourier Transform (of typically 512 points) is performed on each frame and the base-10 logarithm of the magnitude of the resulting coefficients is taken. This "log-magnitude spectrum" is then averaged over several (usually about 5 to 10) frames, and a Discrete Cosine Transform (DCT)

expansion using (1) is performed to yield the Discrete Cosine Transform ("Cepstral") coefficients, which are taken as the "features" of the signal.

$$DCTC(i) = \sum_{k=0}^{N-1} X(k)\phi(i,k) \tag{1}$$

where

$$\phi(i,k) = \cos\left(\frac{\pi i(k+0.5)}{N}\right) \tag{2}$$

Other work has shown that modification of the basis vectors defined by (2) via a bilinear frequency warping function,

$$f' = f + \frac{1}{\pi}\tan^{-1}\left\{\frac{\alpha \sin(2\pi f)}{1-\alpha\cos(2\pi f)}\right\} \tag{3}$$

can substantially increase [2]. A typical value for the bilinear warping coefficient, α is 0.45.

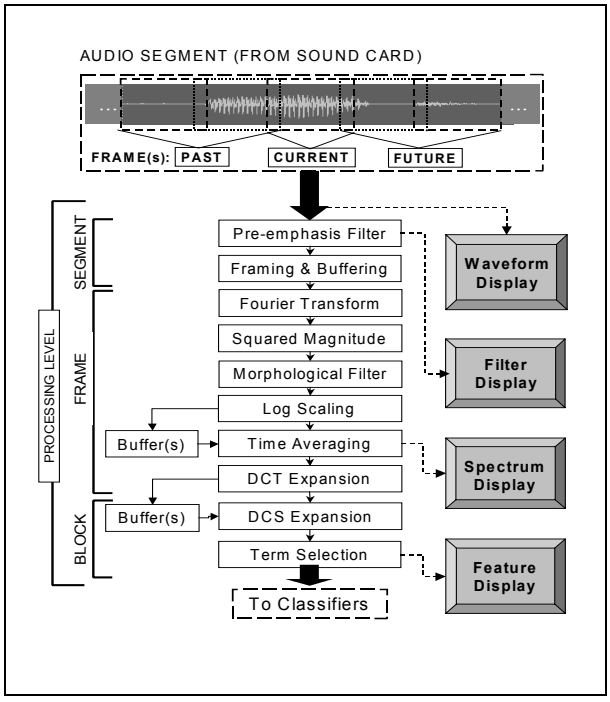

Figure 4. Signal processing and Feature Calculation Detail

Several frames' DCTC values can be combined to form a "block" of feature data over which a Discrete Cosine Series expansion can be performed to encode feature variation over time. This step is useful for encoding feature changes useful in the analysis of non-steady-state sounds such as dipthongs and some consonants, but is not typically performed in the analysis of monopthong vowel sounds. For such cases, the DCTC features are simply averaged over the frames in the block.

The first several (typically 12) cepstral coefficients (DCTCs) are then normalized (zero mean and standard deviation of ± .2, or typically a range of ±1) and passed to a neural network (NN) classifier. The maximum likelihood classifier (MLC), the focus of this paper, is described in more detail below. The NN, a feed-forward multi-layer perceptron, attempts to classify the speaker's utterance into one of the ten phoneme categories. The neural networks used in VATA have one input node per feature, 15 to 25 nodes in the hidden layer, and ten output nodes (one per vowel class). The NN's are trained using error backpropagation for (typically) 250,000 iterations.

3. THE MAXIMUM LIKELIHOOD CLASSIFIER

One of the major problems with a neural network approach for any type of pattern classification is that “out-of-category” exemplars are also classified, usually according to whichever trained class is “closest” to the exemplar. During testing of the VATA system, it became apparent that occasionally sounds which were not "correct" examples of any of the 10 steady-state vowels for which the system was designed caused "false" correct displays. This stems directly from the implicit assumptions underlying a neural network classifier. In particular, a neural network must choose from among only the categories for which it was trained; it cannot choose an "other" category. This has the unfortunate consequence of producing feedback corresponding to a "correct" pronunciation when in fact no valid phoneme sound is in the audio stream. For example, ambient room noise could often trigger the display for a correct /ee/ phoneme. This unintended behavior could, for a user relying solely on VATA for pronunciation feedback, cause confusion and frustration.

To reduce the number of "false correct" displays, a modified Euclidean-Distance measure, which is a form of a Maximum Likelihood Classifier (MLC) has been incorporated with the neural network classifier as a secondary verification system. The focus of this paper is to describe the incorporation of the MLC distance with the neural network to provide better feedback in the VATA system.

In its most basic form, the MLC serves as a distance measure for a multivariate Gaussian feature value from the mean value for each class. The general equation requires the use of the mean vector and covariance matrix for each class (computed from a database of training data) and is given by:

$$D_i(\mathbf{f}) = (\mathbf{f} - \bar{\mathbf{f}}_i)^T \mathbf{R}_i^{-1} (\mathbf{f} - \bar{\mathbf{f}}_i) + \ln|P_i| \tag{4}$$

This general equation has the disadvantage of requiring many parameters (and thus requiring lots of training data) and also being computationally complex. For the case of the speech display, the features are approximately uncorrelated, and hence its inverse is approximately diagonal. Also the (apriori probability) term remains constant and can be ignored. Taking advantage of these two considerations, Equation 2 reduces to the much simpler form shown in Equation 3, which we refer to as the modified Euclidean distance:

$$D_i(\mathbf{f}) = \sqrt{\sum_{j=1}^{m} \left(\frac{\mathrm{f}_j - \bar{\mathrm{f}}_{ij}}{\sigma_{ij}} \right)^2} \tag{5}$$

This equation can be further modified to allow for unequal weighting of the various features:

$$D_i(\mathbf{f}) = \sqrt{\sum_{j=1}^{m} w_j \left(\frac{\mathrm{f}_j - \bar{\mathrm{f}}_{ij}}{\sigma_{ij}} \right)^2} \tag{6}$$

These weights, w_j are important to use if the “information” contained in the underlying features is not proportional to the feature variances. For the case of the DCTC features for vowel recognition, our previous work has shown that the DCTCs do not uniformly contribute to vowel recognition [2]. Based on this earlier work, relative weights of (.82, 1.65, 2.47, 2.47, 2.06, 1.65, 1.24, .83, .41, .41, .41, .21, .21, .08, .08) were used. The actual weights were the relative weights given, but normalized such that the sum of the weights was 1.0.

To provide verification that the vowel display is producing accurate feedback, the MLC calculates the distance of the average features for the NN's choice from the features of the token under evaluation by the NN. If the feature distance is within the threshold criterion,

$$D_i(\mathbf{f}) < \alpha\sqrt{m} \tag{7}$$

where m is the number of features (typically 10 to 15 for VATA), and α is an arbitrary scale factor used for performance tuning, the neural network's decision is accepted, otherwise it is discarded. If α is too small, the MLC will reject many correct tokens; if α is too large, the out-of-category tokens will still not be rejected. Tests (described below) have shown that with the properly determined threshold ($\alpha \approx 1.2$), the MLC does reject unwanted sounds, and makes very few false rejections.

4. EXPERIMENTAL TESTS

To experimentally verify the operation of the MLC method described above, experiments were done as follows: Features were computed using a database of 10 vowel sounds obtained from three speaker panels, as described below, with each speaker producing each vowel sound three times. Vowels were pronounced as “isolated” words, in response to a computer prompt. The central 200 ms interval section of each vowel was used for processing. A two layer feedforward fully interconnected neural network with sigmoidal nonlinearities was trained as a classifier, using nine vowels (those listed above except for /ur/). The means and standard deviations for the features for these

nine vowels were also computed, on a vowel by vowel basis, since these are the features needed for the MLC. The neural network/MLC classifier system was then evaluated using test data from 24 different speakers, and using data for 9 vowels, consisting of all the training vowels, except with /ur/ substituted for /oo/. Thus, ideally the classifier should have correctly recognized 8 vowels, but rejected /ur/ as being out of category.

The experiment was repeated separately for male speakers, female speakers, and male/female speakers combined. 50 training speakers were used for the male speaker case, 50 for the female speaker case, and 80 speakers (40 male, 40 female) for the combined case. There were 24 test speakers for the male case, 24 for the female case, and 48 test speakers for the combined case. In each experiment the false acceptance rate, and false rejection rates were obtained as a function of the parameter α above. False acceptances were considered as instances of the /ur/ being accepted as any one of the vowels. False rejections were considered as instances of any vowels classified correctly by the neural network, but incorrectly rejected by the MLC.

Results are shown in figures 5, 6, and 7, as false acceptance and false rejections as a function of α. As expected for low values of α, the false rejection rate is very high. For high values of α, the false acceptance is high. However, for the case of the male and female speakers considered individually, value of α of approximately 1.5 results in very few false rejections, but still rejects most of the out of category tokens

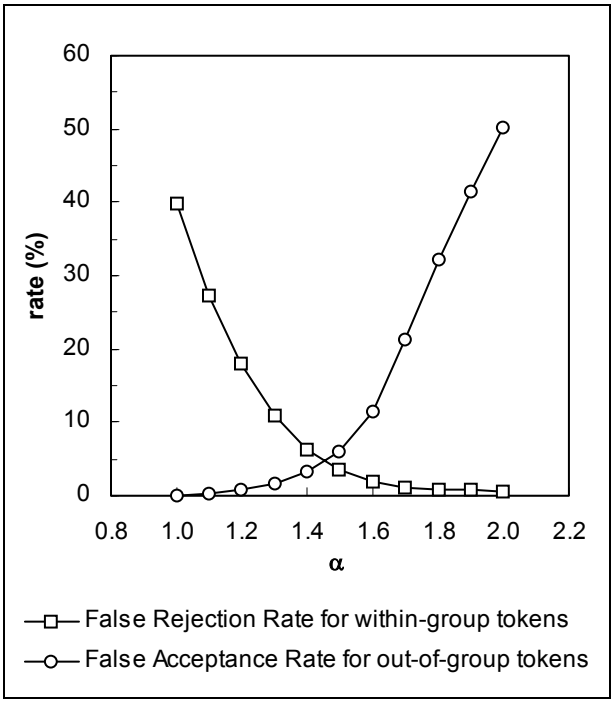

Figure 5. False acceptance/rejection rates as a function of decision threshold value α, male speaker case.

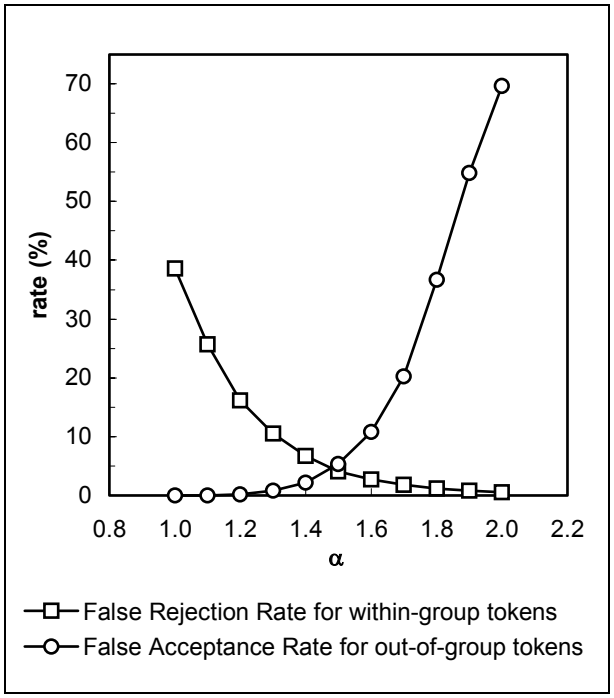

Figure 6. False acceptance/rejection rates as a function of decision threshold value α, female speaker case.

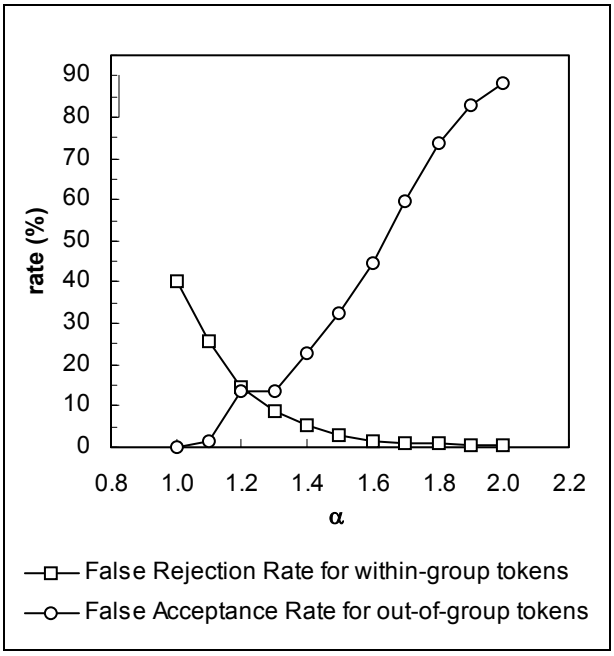

Figure 7. False acceptance/rejection rates as a function of decision threshold value α, combined speaker case.

5. CONCLUSIONS

A method has been described for modifying a neural network classifier to include a maximum Likelihood classifier term. This method has been implemented in a real time vowel articulation training aid for the hearing impaired. This additional term allows the overall classifier to reject "out of category" sounds. In real-time operation, the overall system functions as intended, with most consonant sounds no longer accepted by the classifier.

6. REFERENCES

[1] Zahorian S., and Jagharghi, A., (1993) "Spectral-shape features versus formants as acoustic correlates for vowels", J. Acoust. Soc. Amer. Vol.94, No.4, pp. 1966-1982.

[2] Zahorian S., and Nossair, Z B., (1999) "A Partitioned neural network approach for vowel classification using smoothed time/frequency features", IEEE Trans. on Speech and Audio Processing, vol. 7, no. 4, pp. 414-425.

[3] Zimmer A., Dai, B., and Zahorian, S, (1998) "Personal Computer Software Vowel Training Aid for the Hearing Impaired", International Conference on Acoustics, Speech, and Signal Processing, Vol 6, pp. 3625-3628

7. ACKNOWLEDGEMENT

This work was partially supported by NSF grant BES-9977260.