TRANSFORMATIONS OF SPEECH SPECTRA TO A TWO-DIMENSIONAL CONTINUOUS-VALUED PHONETIC FEATURE SPACE FOR VOWEL TRAINING

Andrew E. Beck and Stephen A. Zahorian

Department of Electrical and Computer Engineering
Old Dominion University
Norfolk, VA 23529

ABSTRACT

A vowel articulation training aid for the deaf has been developed. This is an improved DSP version of the analog filter-bank system presented at **ICASSP90** [1]. The real-time signal processing strategy and the mapping of the continuous feature space to a two-dimensional display space are described. A combined linear/nonlinear transformation, in the form of an artificial neural network, is used to convert cepstral coefficients to a 2-D space suitable for computer displays. A multi-stage artificial neural network training algorithm has been developed that yields better performance and reduced training time relative to conventional training techniques.

INTRODUCTION

Most existing speech training aids for the hearing impaired suffer from deficiencies such as difficulty in interpreting the displays, lack of consistency between displayed parameters and variables required for speech production, and generally inadequate feedback to the user for speech correction. We are attempting to eliminate these weaknesses with a computer-based visual speech training system. This DSP version is more flexible than a previously-developed analog filter-bank system and appears to give better performance.

The initial signal processing has been optimized to provide reliable and easy to interpret continuous "phonetic" visual feedback to the deaf user as a substitute for the lost auditory feedback. That is, small changes in vowel production result in small changes in the display variables and large changes in pronunciation result in large changes in the display. This "acoustic-to-phonetic" transformation could be performed through a linear method such as discriminant analysis, or a nonlinear method such as Kohonen's self organizing feature map [2]. Our method utilizes a neural network algorithm with two nonlinear hidden layers and a linear output layer to convert acoustic speech features to a continuous-valued phonetic feature space. A multi-stage training process has been developed for this algorithm, resulting in a better performing network and a decrease in training time compared to conventional training methods.

The focus of this paper is to describe the real-time signal processing for the training aid, and also on the transformation of speech variables to a continuous phonetic feature space. Vowels were chosen for display in this aid because it has been found that good competence in vowel production is essential before most other sounds can be articulated clearly [3]. Additionally, vowels are primarily characterized by their steady-state short-time amplitude spectrum and therefore vowel displays are relatively easy to implement in a real-time system.

REAL-TIME SIGNAL PROCESSING

The vowel training aid is implemented on an IBM-compatible 386 PC equipped with a 40 mHz TMS320C25 DSP board for speech processing. The basic signal processing algorithm is outlined in Fig. 1.

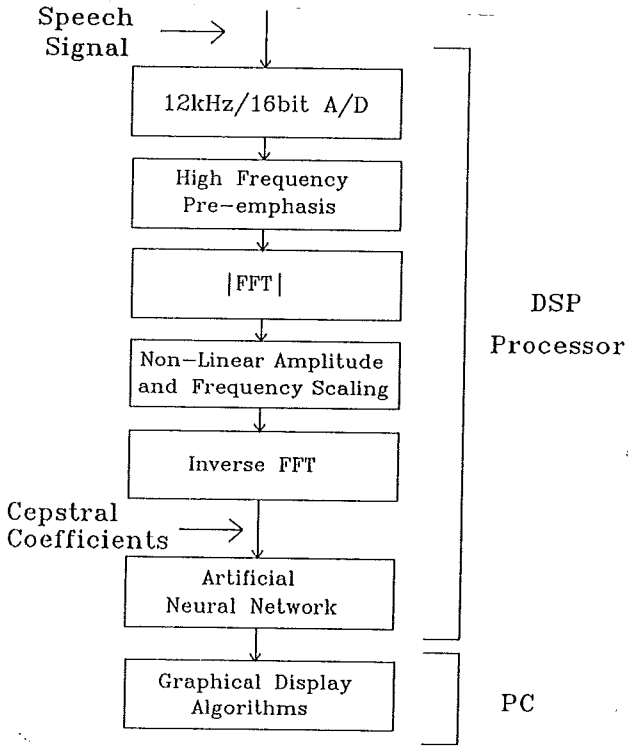

Figure 1. Basic signal processing algorithm.

Summarizing briefly, amplified speech is sampled at 12 kHz, high frequency pre-emphasized ($1-.95z^{-1}$) and stored in a 512-point frame. It is then Hamming windowed, analyzed with a 512-point FFT, and converted to a log magnitude scale. The resulting spectrum is nonlinearly scaled in frequency using a Bark warping function. Cepstral coefficients are computed using a 256-point inverse FFT. The cepstral coefficients are then smoothed using a single pole IIR filter and passed through a neural network and linear transformation. The result is a x-y coordinate pair in the 2-D phonetic space which is transferred to the PC for graphical display. Total processing time per frame is 23.75 ms out of an available 42.67 ms.

ACOUSTIC TO PHONETIC TRANSFORMATION

The cepstral coefficients, as mentioned above, were chosen to represent the acoustic speech signal. Although vowels are generally characterized by the first two or three formants, formants are difficult to track in real-time. Previous experiments have also shown that vowel classification based on cepstral coefficients, which encode global spectral shape, yields slightly higher recognition rates (2-5%) than classification from formants [1].

In order to allow a convenient and easy-to-interpret visual display of vowel information, we transform the cepstral coefficients to a reduced dimensionality "display" space with "target" display positions specified for each vowel. This procedure is a combination nonlinear/linear transformation based on a multi-layer feedforward artificial neural network. A nonlinear transformation occurs from the inputs to the hidden layers of the network followed by a linear transformation to the final output space. The training algorithm developed consists of two "phases." First a classifier network with a single hidden layer and sigmoid nonlinearities at both the hidden and output nodes is trained using backpropagation (a single output node per category). Next, minimum mean-square error techniques [4] are used to map the outputs of the classifier network to specified target positions in the continuous-valued 2-D display space. The neural network and linear transformation are then combined to form a composite neural network with two nonlinear hidden layers and a linear output layer. The final network is then "fine-tuned" with further backpropagation training. Experimental results show that a network with nine to fourteen cepstral coefficients as input features, and fifteen hidden nodes at the first layer, exhibits optimal performance. Note that the second hidden layer, originally trained as a classifier, must have the same number of nodes as categories. The overall network could be trained using backpropagation only, but we have found the multi-stage approach reduces the training time needed and results in a network with a more

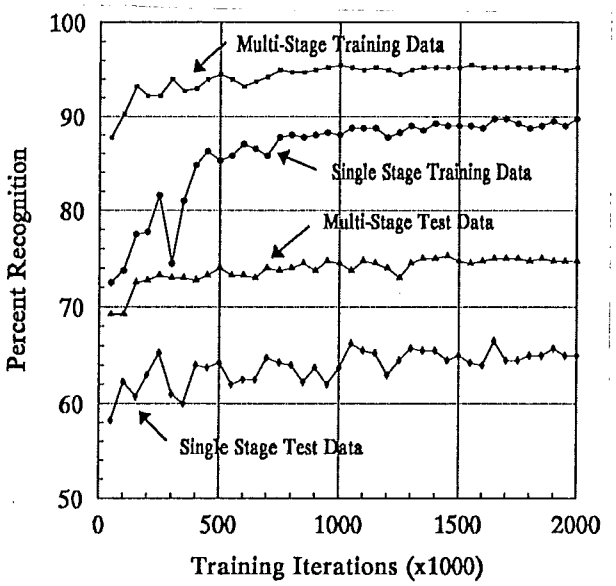

Figure 2. Performance comparison for the two methods.

accurate mapping to the 2-D output space. The results of an experiment illustrating this point are illustrated in Fig. 2. Vowel data from eight adult female speakers were used to train two networks. The first network used only backpropagation training. The second network was trained with the two-stage approach. Ten tokens of each of ten vowel sounds were used for each of the eight speakers. One-half of the speakers were used to train each network and the other half used only for testing. Nine features were used for inputs. Networks were evaluated by classifying vowel data according to minimum Euclidean distance from specified target positions. As seen in Fig. 2, the single stage network trained to a maximum of 89.8% on the training data and reached a maximum of 66.5% on the test data. This compares to the multi-stage method which achieved 95.5% recognition on the training data and tested to 75.3%. Thus the multi-stage approach exceeded the single-stage by 5.7% on training data and 8.8% on test data. The figure also shows that the multi-stage approach trains more rapidly than the single-stage method.

The same experiment was conducted using fourteen input features, with similar results. The fourteen feature single-stage network reached a maximum of 93.8% on training data and 71.8% on test data. The multi-stage version trained to 95.8% and tested to 76.8%. Thus the fourteen feature multi-stage network out-performed the single stage by 2% on training data and 5% on test. It is hypothesized that this increase in performance of the multi-stage method is due to its ability to avoid "bad" local minima encountered by the single-stage approach.

The cluster plots depicted in Figures 3 and 4 also illustrate the accuracy with which vowel data can be

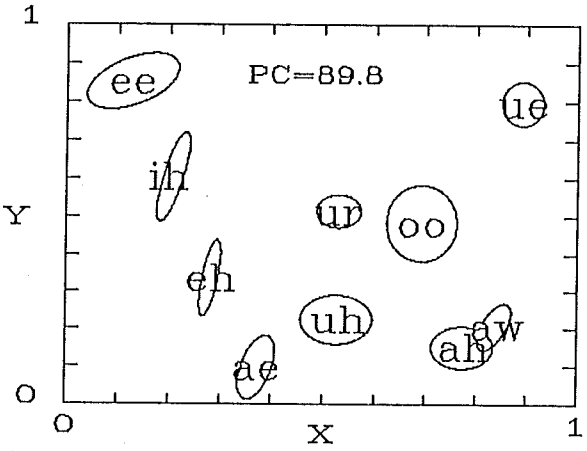

Figure 3a. 2-D vowel cluster plot using the single-stage training method (training data).

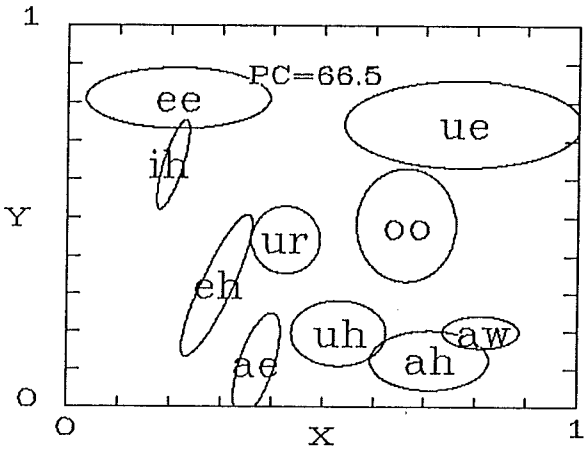

Figure 3b. 2-D vowel cluster plot using the single-stage training method (test data).

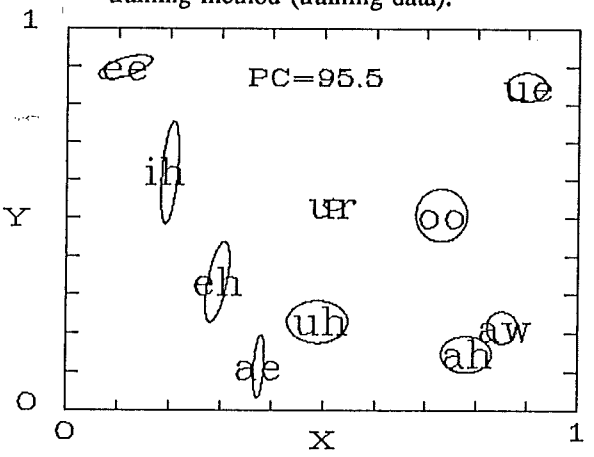

Figure 4a. 2-D vowel cluster plot using the multi-stage training method (training data).

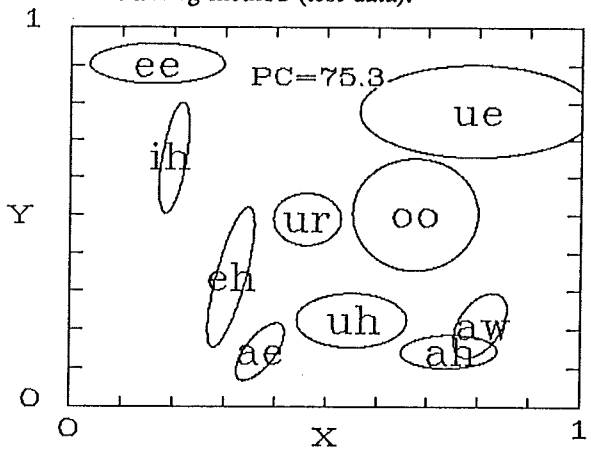

Figure 4b. 2-D vowel cluster plot using the multi-stage training method (test data).

projected to two-dimensional spaces. The plots were drawn using the networks obtained from the previous experiments with nine input features and show the training and test results for the single stage and multi-stage training processes. Target positions in the 2-D space were chosen to approximate vowel centroids in a log F1 versus log F2 plot. Thus the horizontal direction roughly corresponds to the front-back vowel dimension and the vertical direction to the low-high vowel dimension. The ellipses were drawn such that they enclose approximately 50% of the vowel data for each vowel in the 2-D space. The tilt of the ellipses are such that the major axis of each ellipse aligns with the direction of maximum data variation for the corresponding vowel. The percentage correct (PC) is given in terms of minimum Euclidean distance to the target specified for each vowel.

Inspection of the cluster plots shown in Figures 3 and 4 shows that the multi-stage training method is slightly better than the single stage approach in mapping data to a 2-D display. The smaller ellipses of Fig. 4, compared to Fig. 3, illustrate that the multi-stage mapping produces more tightly clustered data in the 2-D space. This is especially apparent for vowels such as "ee" (as in heed) and "ae" (as in had) for both the training and test data.

VOWEL TRAINING AID SYSTEM

Using the signal processing and acoustic to phonetic transformation to a 2-D "display" space discussed, a vowel articulation training aid has been developed. The system first draws ellipses on the computer monitor illustrating the target region for each vowel. In real-time, the user attempts to move a filled basketball to a desired ellipse by correctly pronouncing the corresponding vowel sound. Figure 5 shows a black and white reproduction of this display. In the real-time display, ball color, size and position are all dependent on the acoustic signal. The size of the ball is proportional to loudness and the color changes according to the color of the nearest ellipse. The target region for each vowel is centered at the "ideal" position for the vowel. The ellipses are drawn in the same manner as the cluster plots discussed earlier (Figures 3 and 4) for the training data case.

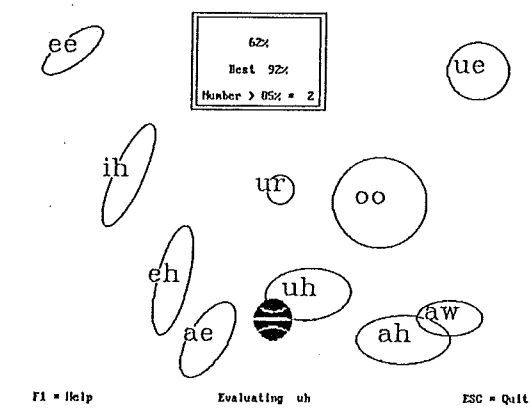

Figure 5. Real-time display screen of the vowel training aid system.

Several additional features are available to enhance the display and provide numerical evaluation. The ellipses can be expanded or reduced to adjust the sensitivity of the system in recognizing vowel sounds. Any combination of vowels can be selected for display allowing a user to focus on vowels of interest. Each vowel has a help display, which draws a face showing correct mouth and tongue position for that vowel. An evaluation option is also included which gives an indication of percent correct in terms of Euclidean distance to the desired vowels ellipse center. The user can define a percentage as the threshold for a correct pronunciation, and the system counts the number of utterances which exceed the threshold. The main novel feature of this display is, of course, the continuous feedback it offers. If the user slightly mispronounces a vowel sound, the ball moves in the direction of the desired sound but not quite inside the ellipse. Large mispronunciations not only result in large deviations from the desired position, but also give an indication as to what sound the mispronunciation is most similar. Thus the user can make changes in the vocal tract and see immediate, continuous feedback on the display.

Additional displays have been implemented that do not utilize the two-dimensional display mapping algorithm. A "bargraph" display has been developed which depicts each vowel sound as a vertical bar. Bar height is controlled by "correctness" of vowel pronunciation. Ideal pronunciation of a vowel sound results in the corresponding bar reaching maximum height and all others staying at zero. This version uses the same signal processing discussed above but uses a neural network with one hidden layer and nonlinear outputs nodes as a vowel classifier. A third display developed is a "pac-man" style game. The user moves a "pac-man" around the screen by correct pronunciation of the vowel sound corresponding to the desired direction.

EVALUATION RESULTS

The previous version of the display has been informally tested with hearing-impaired speakers [1]. The present version of the display has been tested with normally-hearing speakers who have vowel articulation problems. An adult French-native female worked with a speech therapist and the training aid concentrating on the vowel sounds "ae" (as in had) and "ur" (as in heard). Initially the user was unsuccessful in even imitative pronunciation. After several sessions on the system and help from the speech therapist, the user was not only able to pronounce both vowels imitatively, but spontaneously in words, with 85% accuracy as judged by the therapist. The therapist found the system "particularly beneficial with regard to the visual representation, which allows the user to discriminate subtle differences in the place and manner of production between phonemes." Further evaluations are now underway with normally hearing children with vowel articulation problems.

CONCLUSION

A method has been described for converting high-dimensionality spectral shape representations of speech spectra to low-dimensionality display representations for use as a vowel training aid. In addition a multi-stage training algorithm has been devised that results in faster more accurate training of a continuous input/output neural network. This algorithm has many potential applications outside of the one presented in this paper. The method could be used for any application that requires a nonlinear mapping between continuous-valued vector spaces. Experimental verification of the dimensionality transformation indicates that vowels can easily be discriminated in a two-dimensional space. Experiments utilizing the vowel training aid with hearing-impaired and normally-hearing speakers indicate that the displays developed are easy to interpret, and provide useful information for improving vowel articulation.

REFERENCES

[1] Zahorian, S.A. and Venkat, S. (1990). "Vowel articulation training aid for the deaf," ICASSP-90, 1121-1124.

[2] Kohonen, T. (1988). "The neural phonetic typewriter," Computer, (March, 1988), 11-22.

[3] Ling, D. (1976). Speech and the hearing-impaired child: Theory and practice, (The Alexander Graham Bell Association for the Deaf, Washington, D.C. ISBN 0-88200-074-8).

[4] Zahorian, S. A. and Jagharghi, A. J. (1992). "Minimum mean-square error transformations of categorical data to target positions," accepted for publication in the IEEE Trans. Acoust., Speech, Signal Process., scheduled for January, 1992.