PERSONAL COMPUTER SOFTWARE VOWEL TRAINING AID FOR THE HEARING IMPAIRED

A. Matthew Zimmer, Bingjun Dai, Stephen A. Zahorian

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

ABSTRACT

A vowel training aid system for hearing impaired persons which uses a Windows-based multimedia computer has been developed. The system provides two main displays which give visual feedback for vowels spoken in isolation and short word contexts. Feature extraction methods and neural network processing techniques provide a high degree of accuracy for speaker independent vowel training. The system typically provides correct classification of over 85% of steady state vowels spoken by adult male, adult female and child (both genders combined) speakers. Similar classification accuracy is also observed for vowels spoken in short words. Low cost and good performance make this system potentially useful for speech training at home.

1. INTRODUCTION

A semi-speaker independent visual speech training aid for persons with hearing impairments has been developed using a standard Windows 95 or NT based multimedia computer. The training aid provides visual feedback about the quality of pronunciation for 10 steady state American English monopthong vowel phonemes (list these in terms of darpabet codes) . Neural network 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 an F1-F2 display. Four additional displays provide access to the results of the signal processing steps used to produce the main displays.

A previous system for steady-state vowels [1] required specialized hardware and a custom-programmed user interface. A major difficulty with the previous system was software upgrades, since the programming was accomplished with a combination of PC C, signal processing card C, signal processing card assembly language, with a requirement that all software components be compatible. The newer display system reduces cost and maintenance difficulty by using a standard Windows graphical user interface, which permits easy operation and without specialized hardware and software. Unfortunately, the less expensive multimedia soundcards have poorer noise performance and require additional signal processing measures to ensure spectral stability and display accuracy.

Work is underway on a second set of displays to provide feedback for vowels spoken in CVC (consonant, vowel, consonant) contexts. A large database (~160 speakers) of steady-state vowel and CVC recordings has been collected for use in training and testing the neural networks for both the steady-state and CVC-context vowel displays. Results from neural network training experiments indicate that the larger recording database will help the system achieve greater speaker independence. Previous testing of a similar system for steady state vowel training [citation...] showed that hearing-impaired schoolchildren who used the system exhibited improved vowel articulation, but did require training for vowels produced in word contexts.

2. PROCESSING STEPS

Six steps are used to implement the two main displays: preemphasis, log-magnitude spectrum calculation, morphological time smoothing, calculation of discrete cosine transform coefficients, final smoothing, and classification. The system acquires a continuous speech signal from the multimedia sound card using the standard Windows multimedia services. A custom waveform audio API has been developed to provide smooth double buffering and automatic signal threshold detection. The API sends the continuous signal to the main signal processing routines in 90ms (typical) segments. The display output is updated once for each new segment acquired.

Pre-emphasis is applied to each segment using a 2nd order IIR filter, with a peak frequency response at approximately 3 kHz. The filtered acoustic segment is then subdivided into 30-ms frames overlapping by 15ms. A 512 point FFT is computed for each frame. The log magnitude of each frequency sample is calculated, and a 40dB noise floor is applied. Using a time window of 10 frames, the peak value at each frequency over the

selected frames replaces the original spectral values. A discrete cosine transform (DCT) is performed on the resulting "time-smoothed" spectrum. These 12 DCTC's are further time smoothed by block averaging to compute "features" of the speech signal for the neural network classifier.

The two main displays are derived from neural networks trained with "backpropagation," but with slightly different architectures. The bargraph display uses a network with 12 inputs, 25 hidden layer nodes, and 10 outputs. Each output corresponds to one of the 10 vowel phonemes, and is displayed directly as a bar height corresponding to one of the neural network outputs. A correct vowel utterance results in only one bar with high amplitude, and all other bars with low amplitude. The ellipse display network has two hidden layers with an additional linear output layer used to map the vowels to a continuous 2-dimensional space. In the actual display, ten elliptical regions are outlined in different colors and correspond to the target vowel sounds. In operation, a correctly pronounced vowel guides a basketball icon into the ellipse region for that vowel and changes the ball's color to match the ellipse color. Incorrect vowel pronunciation causes the basketball icon to wander or appear in locations outside of the ellipses, or the ellipses of alternate vowels.

3. NOISE ISSUES

The predecessor of the current Windows NT/'95-based system was implemented with TMS320C25 DSP board from Texas Instruments. The conversion from a dedicated DSP board to a standard multimedia sound card brought low cost, but also degradation in noise performance, since most sound cards do not have high signal to noise ratios in typical operation.. When we tested the system with steady-state vowels, the spectral display had considerable jitter in the envelope. Based on approximate tests, the effective signal to noise ratio of typical speech signals and average noise levels was only 30 dB. As a consequence, the DCTC feature display was not nearly as stable (for steady state vowel sounds) as for the DSP board based system. Presumably this additional noise degrades the recognition performance, particularly for "close" vowels.

In our initial comparison examinations of the feature displays for the PC only system and the older DSP system the additional jitter in the display was so dramatic that a software implementation bug (related to the double buffering scheme) was suspected. However, careful checking and experimental testing of the new code indicated that the software was not causing the inconsistency in the feature displays.

Therefore, it was concluded that the problem was due to front end noise, and, as discussed in the next few paragraphs, several algorithmic refinements were investigated to ameliorate the effects of this noise.

First, a second order pre-emphasis filter centered at 3kHz was added before the remainder of the signal processing in an attempt to suppress low-frequency and high-frequency noises. This filter can also be viewed as an approximate matched filter to the average speech spectrum. Although the recognition performance was not significantly enhanced by this filter (using tests similar to those described later in this paper), this refinement was included in the processing since in other tests we have found this step to result in a small but measurable improvement (mention last years ICASSP paper).

Time smoothing of the spectrum was the second method investigated to reduce noise and increase feature stability for steady state vowels Three types of smoothing methods were tried: peak-value smoothing, median-value smoothing and average-value smoothing. Experiments with our pre-recorded database of steady state vowels showed that peak-value smoothing resulted in the best recognition performance. Moreover, longer smoothing windows gave the largest improvement. For example, the overall vowel recognition rate for test vowels increased by 3% when the smoothing window length was increased from 0 frames to 65 frames. However, a smoothing window of 10 frames or less seemed to be appropriate, since smoothing with too long a window introduces intolerable response latency in the real time program. With a smoothing window length of 10, there is about a 1% increase in recognition rate as compared to the case without smoothing.

In addition, we also tried adjusting other processing parameters such as *FFT length, Frame length* and *Block length*. We found that increasing the *FFT length* brought slight performance improvements, but at the expense of increased computational complexity. As a compromise, an *FFT length* of 512 with an associated *Frame length* of 30 milliseconds (sampling rate of 11.025 kHz) was selected. The *Block length* variable, which is used to determine a final number of frames as a last averaging step to compute features from DCTCs, was selected as 5 frames, again as a compromise between performance and response latency and computational complexity

In summary, since the signal to noise ratio of a sound card is less than that of a dedicated DSP board, some performance degradation must be expected. The improvement from optimization of the processing steps

described in this section is limited. Better performance would be expected if a high-quality sound card were used, but at a higher hardware cost.

4. NEURAL NETWORK TRAINING ISSUES

Neural Networks require training to function properly. If a sufficient amount of training is performed, the neural network can generalize to properly classify data that it was not specifically trained to recognize. For the visual speech display system, when the system performs satisfactorily with speakers who are not part of the training data set, it can be said that the system is "speaker independent."

The overall system performance can be reasonably predicted from the performance of the neural network classifiers--training recognition rate and the test recognition rate. The test recognition rate, which is obtained from speakers not used for neural network training, is a more accurate predictor of how well the neural network will generalize and can be interpreted as a measure of speaker independence. As the size of the training data set increases, it is expected that the training recognition rate will decrease as the variability within the training data set increases. Conversely, the test recognition rate should increase as the increased amount of training data improves the networks ability to generalize . These points are illustrated in stylized from in Figure 1.

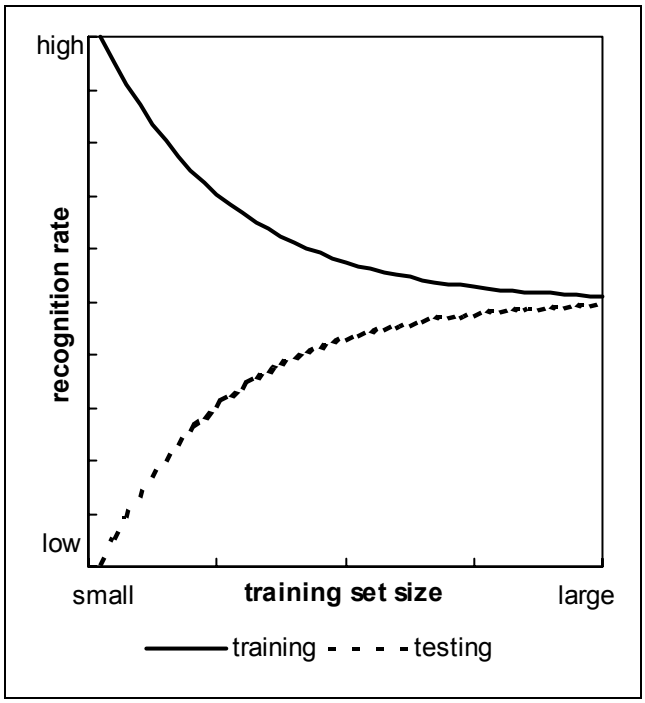

Figure 1. Expected recognition rate behavior as a function of training set size

To achieve speaker independence for the visual speech display, a large amount of speech data is required to train the neural networks. Speakers were divided into three primary groups: "male" (males over 14 years old), "female" (females over 14 years old), and "child" (males and females under 14 years old). By partitioning all speakers into these three groups, better within-group speaker independence would be expected with less training data than if all speakers were considered as one group. A fourth group ("general") which encompasses all speakers was also defined and provides a qualitative measure of speaker independence without group

partitions. For the two displays, one neural network was trained per group, for a total of 8 neural networks in the complete system.

A database of speech recordings was collected from 56 male, 58 female, and 46 child speakers using specially-developed recording software which provides automatic recording, endpoint detection, rough segmentation and database management. Speech files were stored in the TIMIT format and were automatically organized into a systematic directory structure. A second program performed more accurate segmentation of CVC recordings and phoneme labeling based on energy measures.

5. TRAINING EXPERIMENT RESULTS

Figure 2 depicts the trend of training experiment results for bargraph-display networks trained with steady state vowels for three data set sizes: 12 training speakers (4 test speakers) 24 training speakers (8 test speakers) and 33 speakers (11 test speakers). Note that the ratio of training to set speaker data set sizes is 3:1 in each case. Results for the ellipse display network are similar and are typically between 5% and 10% lower than the bargraph rates (Table 1). The general trend of the training and test results follows the stylized "expected" trend of figure 1. The male speaker category shows the greatest rise in test recognition rate as training set size increases. The female case shoes the least change as the number of training speakers rises. The child speaker case exhibits the lowest recognition rates, indicating the greatest amount of variability in the recorded speech data.

A second set of experiments was conducted to compare recognition rates for different pairings of CVC and steady-state vowel data using training databases with 72 (24 each from Male, Female, and Child categories) speakers and test databases of 24 (8 from each category). Each data type was used once as training data and once as testing data, resulting in four total pairings. Results from

Speaker Case:	Male	Female	Child	General
Train Rate for Bargraph (%)	96.2	95.5	93.2	90.0
Test Rate for Bargraph (%)	87.1	88.2	76.3	83.7
Train Rate for Ellipse (%)	90.5	91.1	82.1	79.7
Test Rate for Ellipse (%)	86.9	84.1	70.7	74.4

Table 1. Comparison of Training and Test Recognition Rates

Training Set: 24 Speakers from each gender, Steady Vowel
Test Set: 8 Speakers from each gender, Steady Vowel

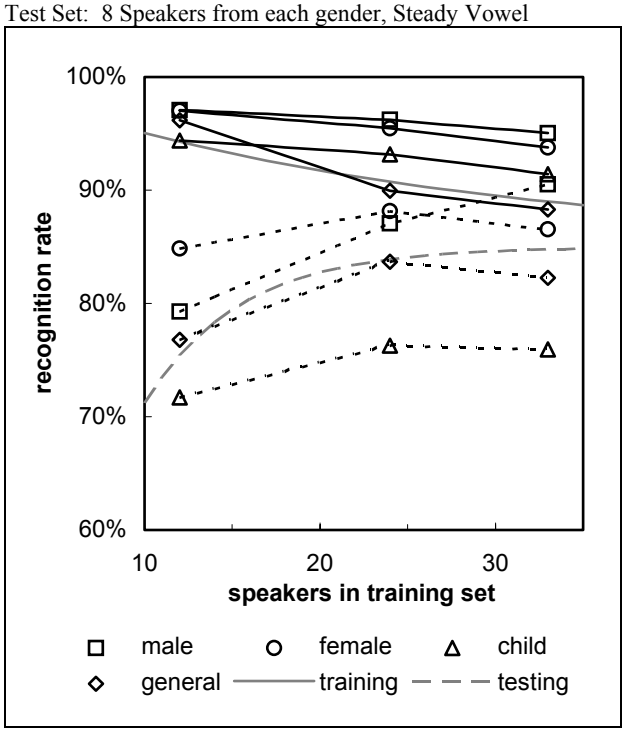

Figure 2. Observed recognition rate behavior as a function of training set size for steady-state vowels.

these experiments are shown in Table 2. While better test recognition rates are exhibited in the "homogeneous" pairings of CVC/CVC and SV/SV, consistent performance is shown in the "heterogeneous" pairings with about a 5 to 6% decrease in recognition rate, indicating that this system performs approximately equally as well for either type of vowel data.

Trained with: / Tested with:	SV / SV	SV / CVC	CVC / CVC	CVC / SV
Train Rate (%)	90.0	90.0	89.4	89.4
Test Rate (%)	83.7	75.9	84.6	77.4

Table 2. Comparison of Training and Test Recognition Rates for different pairings of vowel data.
Training Set: 24 Speakers from each gender
Test Set: 8 Speakers from each gender
SV: Steady vowels spoken in isolation
CVC: Vowels extracted from CVC words

6. CONCLUSION

The low cost and high performance of this system indicate that it has potential for speech training at home for the hearing impaired. Although the use of common multimedia sound card adds a significant amount of noise to the input signal when compared to high quality dedicated signal processing cards, additional signal processing and careful parameter selection lessen the impact on the system's performance. A large speaker database provides sufficient training data to provide high neural network classifier performance—typically over 85%. Since the highest test results obtained in this study are still substantially below the training results (typically at least twice as many errors for the test data as for the training data), continued collection of training data should cause the real time system performance to improve, and increase the attractiveness of this system for widespread use for speech training . The results for the vowels for CVCs also indicate that a display for CVC vowels can be implemented with accuracy comparable to that obtained for steady vowels, provided the neural networks are trained with data obtained form CVC tokens.

7. REFERENCES

[1] Beck A., and Zahorian S. "Transformations of Speech Spectra to a Two-dimensional Continuous Valued Phonetic Feature Space for Vowel Training." *IEEE International Conference on Acoustics, Speech, and Signal Processing*, San Francisco, CA, pages 241-244. March, 1992.