HMM-Neural Network Monophone Models for Computer-Based Articulation Training for the Hearing Impaired

Mukund Devarajan, Fansheng Meng, Penny Hix, and Stephen A. Zahorian

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. In previous papers, the signal processing steps and display options have been described for giving real-time feedback about the quality of pronunciation for 10 steady-state American English monopthong vowels (/aa/, /iy/, /uw/, /ae/, /er/, /ih/, /eh/, /ao/, /ah/, and /uh/). This vowel training aid is thus referred to as a Vowel Articulation Training Aid (VATA). In the present paper, methods are described to develop a monophone-based Hidden Markov Model/Neural Network recognizer such that real time visual feedback can be given about the quality of pronunciation of short words and phrases. Experimental results are reported which indicate a high degree of accuracy for labeling and segmenting the CVC database developed for "training" the display.

1. BACKGROUND

Vowel Displays

The existing system [1][2][3][4] has two main displays. One is a bargraph display, which gives feedback about how well speech utterances match discrete vowel categories. The other is an "ellipse" display, which provides more continuous feedback about vowel pronunciation. The system is designed to provide feedback for the vowels /aa/, /iy/, /uw/, /ae/, /er/, /ih/, /eh/, /ao/, /ah/, and /uh/, which correspond to the vowel sounds found in the words "cot," "beet," "boot," "bag," "bird," "pig," "bed," "dog," "cup," and "book" respectively. In addition to the bargraph and ellipse display, three game displays have been developed. A speaker group selection option with "CHILD," "FEMALE" and "MALE" settings allows all displays 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" uses a classifier based on all speakers.

2. CVC DISPLAY PROCESSING STEPS

The objective of a CVC display is to visually present indicators of pronunciation "correctness" at the phone level in real time, with minimum time delays, in response to short words produced by a speaker. The operating system interacts with the sound card to continuously acquire contiguous sections of data (a "segment") from the audio data stream, using a double buffering approach for processing. The basic data acquisition was developed with operating system services, rather than directly with the hardware, and therefore the overall CVC system is able to work with most commonly available Windows-compatible sound cards.

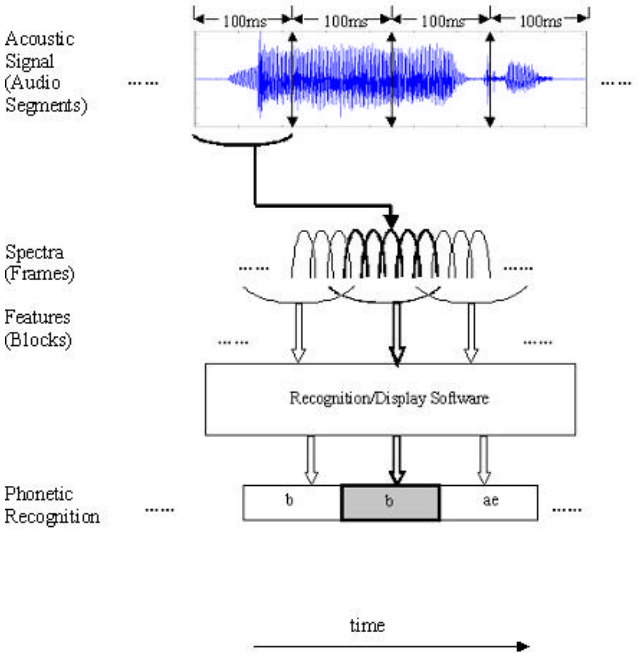

Figure 1. CVC display block diagram

A very important consideration in the overall implementation of the display is to minimize delays between input speech and display changes—to make the overall system as "real time" as possible. The data management for this signal processing can be explained in terms of three levels of buffering, as illustrated in Figure 1. First, as just mentioned above, the non-overlapping but continuous "segment" buffers form the interface between the data acquisition subsystem and the signal processing and recognition software. The first step of processing is based on overlapping frames of data--- thus frames form the second level of buffering.

The third and final level of buffering is referred to as "blocks," with each block consisting of an integer number of frames, with block spacing also consisting of an integer number of frames. All parameters ("features") used for recognition are computed

from all the frames in a block with parameter updates based on the block spacing. Similarly recognizer decisions are made from all the block output parameters. Display updates are made once per each new segment, but using all the block-based recognizer decisions which occur in that segment. Each of the processing and recognition modules maintains internal delay memories that hold all previous data and intermediate calculations that are needed. Thus the "frame' and 'block' level buffers can be asynchronous with respect to the segment buffers. Typical values for the various buffers are 100 ms for the segment buffer, 30 ms frames with a frame spacing of 10 ms, and blocks consisting of 10 frames with a spacing of 5 frames between blocks. The time delay between input speech and display changes is on the order of one segment time, or 100 ms for the values given.

The actual signal processing consists of mid-frequency pre-emphasis, windowing, spectral magnitude calculations, DCTC computations for each spectral frame, DCS calculations over the DCTCs in each block, scaling, and finally neural network classifications and HMM phone log probability calculations. DCTCs, similar to cepstral coefficients, are computed using cosine basis vectors which are modified so that the frequency resolution approximates a mel frequency scale. In the final step, the DCTC's are block-encoded with a sliding overlapping block using another cosine transform over time that is used to compactly represent the trajectory of each DCTC. The cosine basis vectors in this second transform are also modified so that the temporal resolution is better near the middle portion of each block relative to the endpoints. The coefficients of this second transform are called Discrete Cosine Series Coefficients (DCSC). This method is very flexible with a small number of parameters that control the details of the analysis, particularly in terms of spectral/temporal frequency resolution tradeoffs. Typically 15 DCTC terms are computed, followed by 5 DCS terms for each DCTC (75 total parameters), but then as subset of 40 of these terms are used for recognition.

As for the actual phone recognition, the active dictionary varies from a single word, for the case when the system simply prompts the user to pronounce a certain word, up to the entire trained dictionary (at this point, consisting of 13 CVC words, containing 18 phones). In either case, the recognition module returns the identity of the top scoring phone models, log probability ratios for each model, neural network classification scores for each phone, time markers for the beginning and end of each phone, and the identity of the most likely word. The goal is to use this information to provide feedback about the produced word in a variety of formats including game displays.

In order to achieve these objectives for the display of phonetic information, it was first necessary to record and prepare a large database of CVC tokens. Preparation includes elimination of poorly pronounced tokens, determination of time markers for phonetic segments in the properly produced segments, and finally creation of HMM and neural network phone models for each phone of interest.

3. DATABASE OF CVCS

A database of CVC sounds was collected from adult males (120 speakers), adult females (142 speakers), and children between the ages of 6 and 13 (55 speakers). The tokens recorded are listed in Table 1. All tokens were automatically endpointed, using an endpoint routine similar to [5]. Listeners then evaluated tokens, and those that appeared to be incorrectly pronounced were eliminated from the database. This data was collected over a period of several years during which the sampling rate was increased from 11025Hz to 22050 Hz. For all experimental results reported in this paper, the 22 kHz data was smoothed with a 2-point smoothing window, and decimated to 11 kHz.

Table 1. List of CVC Tokens Recorded

CVC	Pronunciation	CVC	Pronunciation
Bag	b ae g	Boyd	b oy d
Bed	b eh d	Cake	k ey k
Beet	b iy t	Cot	k ao t
Bird	b er d	Cup	k ah p
Boat	b ow t	Dog	d ao g
Book	b uh k	Pig	p ih g
Boot	b uw t		

4. SEGMENTATION AND LABELING OF CVC DATABASE

After considerable experimentation, manual examination, and testing, the following multi-step approach was used for segmentation of the CVC database.

1. A heuristic rule-based segmentation, based primarily on pitch, spectral band energies and endpoint detection was used as a first pass for segmentation.
2. An HMM triphone-based word recognizer was trained as a recognizer for the entire database, using the segmented data from step 1 to initialize HMM models.
3. Using the models trained in step 2, recognition was performed on all the training data. All CVCs that were not "correctly" recognized at this step were eliminated from the database.
4. HMM models were retrained using the data after step 3.
5. The HMM models obtained in step 4 were used for forced alignment of CVC tokens.

Each of these steps is now described in somewhat more detail below and illustrated with experimental data.

Step 1 Heuristic Rule-Based Segmentation

The onset and offset of speech (i.e., the endpoints) are first determined using an endpointing routine similar to [5]. Next a pitch track is computed for each utterance, using the pitch routine of [6]. Additionally three normalized spectral band energies were computed—low-frequency energy band (300-800 Hz), mid-frequency energy band (800-2000 Hz) and high-frequency energy band (2000-4000Hz). The vowel region of each syllable is

located using a combination of the voiced portion of the pitch track, the normalized low-frequency band energy, and heuristic rules (including maximum and minimum durations for each initial and final stop). Next, the spectral band with the maximum value between the vowel end and the final endpoint of the CVC from all the 3 bands is selected as the band to be used to determine the start point for the final consonant. The first point in time at which this normalized band energy exceeds an empirically determined threshold (typically .3) is selected as the beginning of the final consonant. In addition, if this energy band falls below another threshold (typically .1) for at least a certain time duration, in the region between the end of vowel and the beginning of the final consonant, then that low-energy region is labeled as closure. Figures 2 and 3 illustrate this heuristic labeling method for two CVCs. Note that the method appears to give correct segmentation for Figure 2 whereas the initial stop and final stop appear to be in error in Figure 3. Visual inspection of several hundred tokens indicated that this heuristic segmentation appeared to be generally correct (no really large errors) for about 80% of tokens, but did have some really large errors for the remaining 20% of tokens. Thus the heuristic method was not considered accurate enough for a "final" pass of segmentation, but was valuable for bootstrapping HMM models. The HMM models, followed by forced alignment, did appear to provide more accurate segmentation.

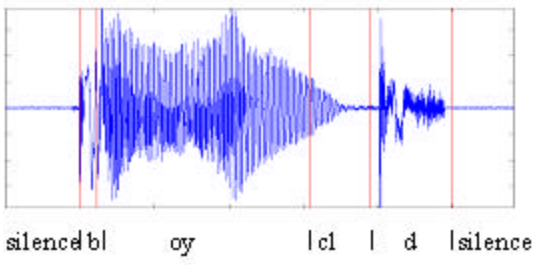

Figure 2. Example of correct heuristic segmentation for “boyd”

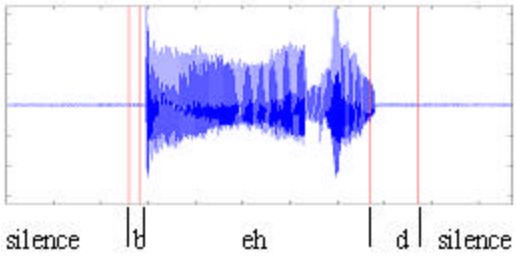

Figure 3. Example of incorrect heuristic segmentation for "bed."

Step 2 Bootstrap triphone models

The heuristic labeling process just explained was used for initializing HMM triphone models for the CVCs. (Initialization based on “flat-start” data (no segmentation used) resulted in lower recognition performance than initialization based on the heuristically labeled data.) The HTK (Hidden Markov model Tool Kit ver 3.1) [7] was used in this process to build a recognizer, which was used to determine the degree to which the words present in the training database could be automatically recognized. The HTK was configured with a dictionary which allowed the presence or absence of closure in each CVC. Mel frequency cepstral coefficients (MFCCs) derived from the FFT-based log spectra were used as the features, augmented by delta and acceleration coefficients (totally 39 terms).

The HMM triphone prototype models were configured with 5 states, 39 features, and 1 Gaussian mixture. After the prototype models were established, each monophone HMM was initialized by the HTK tool (HINIT), which uses the Viterbi algorithm to find the most likely state sequence corresponding to each training sample. These isolated-unit models were refined using the Baum-Welch re-estimation procedure (HREST tool in HTK) for 3 passes. State tying was also used.

Total processing time needed for all the above processes including initialization and re-estimation was approximately 32 minutes on a 1GHz machine with 512 MB ram. Note that a total of 20 monophone models were created (12 vowels, 6 consonants, 1 silence and 1 closure model).

Step 3 Bad data removal

The HMM models trained in the previous step were used as a recognizer to eliminate all tokens not correctly recognized by the recognizer. Table 2 summarizes the database size before and after this step. Note the database includes over 300 speakers, including adult males, adult females, and children, thus making the recognition task somewhat difficult. Approximately 12% of the data was eliminated (recognition rate of 88.1%) by this step. The motivation for this step was that the majority of misrecognized tokens (entire database used for training) were not pronounced correctly or clearly. The goal was to be confident that nearly all tokens used for final training were clearly and correctly produced.

Table 2. CVC database before and after removal of tokens incorrectly recognized by HMM recognizer.

CVC	Male	Female	Child	Total
Original Database	4638	5514	2168	12320
After Removal	4172	4935	1747	10854

Step 4—Triphone model retraining

Triphone models were recreated using the identical processing methods as described above for step 2, except now, only data remaining after step 3 was used. Thus, it was expected that these HMM models would be better representations of the phones than those created in step 2, because of the elimination of the “bad” data. As a test, these triphone models were again used as a

recognizer, and approximately 98% accuracy was obtained on the training data.

Step 5—Final forced alignment

Finally the above retrained HMM models were used to obtain a forced Viterbi alignment, resulting in final labeled training data.

5. NEURAL NETWORK TRAINING

Three neural networks are used in the CVC display system--one each for the initial consonant, vowel and the final consonant. Each neural network has one input node per feature component and 25 nodes in the hidden layer. The initial consonant network has 4 output nodes, the vowel network has 12 output nodes and the final consonant network has 5 output nodes. The NN's are trained using error backpropagation for (typically) 250,000 iterations. These neural network classifiers resulted in the following recognition rates for the training data. As expected the recognition rates are generally higher if based on the data obtained from the final forced alignment, rather than the heuristic algorithm alignment.

Table 3. Percent correct results for neural network classification of the phones in CVC syllables based on either heuristic labeling of phonetic segments or HMM forced alignment.

	Initial Consonant	Vowel	Final Consonant
Heuristic Algorithm	95.3%	83.6%	82.8%
HMM / NN Based	94.0%	90.8%	85.2%

6. CONCLUSIONS

The extension of the Vowel Articulation Training Aid (VATA) for use with short words (CVCs) has been described. In response to audio input from a user, a visual display is generated to provide feedback about the quality of pronunciation of the phones in each CVC. A key aspect of the system is the short time delay between input speech and display changes. Experimental data gives indicates that speaker independent phonetic accuracy is on the order of 90%. Interested readers may obtain a copy of the run time program by emailing (szahoria@odu.edu)

7. REFERENCES

[1] Zahorian S.,Zimmer M., Meng F.,(2002) "Vowel Classification for computer-based visual feedback for speech training for the hearing impaired", International Conference on Spoken Language Processing.

[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

[4] 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.

[5] Evangelos et al., (1991) "Fast endpoint Detection Algorithm for Isolated word recognition in office environment", International Conference on Acoustics, Speech, and Signal Processing, pp. 733-736.

[6] Zahorian S., and Kasi, K., (2002) "Yet another Algorithm for Pitch Tracking", International Conference on Acoustics, Speech and Signal Processing.

[7] Young et al., (2001) Hidden Markov Model Toolkit v3.1 reference manual, Technical report, Speech group, Cambridge University Engineering Department, December 2001.

8. ACKNOWLEDGEMENT

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