A high-performance vowel spotting system based on a multi-stage architecture

J. Sirigos, N. Fakotakis and G. Kokkinakis
E-mail: john@pat.forthnet.gr
Wire Communications Laboratory, University of Patras, 26110 Greece

ABSTRACT

In this paper we present a novel multi-level vowel detection system with improved accuracy. Multi layer perceptrons (MLP), Discrete Hidden Markov Models (DSHMM) and heuristic rules are combined in three different levels to reduce the probability of false acceptance and rejection of vowel sounds. The TIMIT database was used to train and test this system. The rules are variable and are automatically customized by statistics extracted from the database, which concern the duration, the energy and the distance between vowels. The proposed method can easily be extended to languages other than English as long as a proper database exists for training the system. Its accuracy was measured to 99.22% using all the test data sets of the TIMIT database. Thus, the proposed vowel detection process can be reliably used for speech processing applications (speaker or speech recognition) where accurate vowel spotting algorithms are necessary.

1. INTRODUCTION

The vowel sounds are perhaps the most important class of sounds for speech processing applications. Many developed speech and speaker recognition systems rely heavily on vowel recognition to achieve high performance.

Several methods have been proposed for vowel detection. RMS, energy, zero-crossing rate [4], spectral information [8][9], formant frequencies [1][7] are some of the parameters used. Also, modeling the vowel sounds with Hidden Markov Models [6], Gaussian mixtures or neural networks [5] can lead to vowel identification systems. Nevertheless, each single method cannot achieve the demanded high accuracy (above 99%) based only on the speech signal, i.e. without the need of a grammar structure which makes the system language dependent.

In this paper we describe the successful combination of three different techniques in building a multi-level high accuracy vowel detection system. Multi layer perceptrons, discrete hidden markov models and heuristic rules are used. A sophisticated control unit decides on the successive use of each technique and provides the necessary information to each stage of the system.

The paper is organized as follows: In Section 2 we discuss the proposed system describing briefly each element. In Section 3 we describe the database and we present the experimental results. Finally, we summarize the major results and outline our future work.

2. SYSTEM ARCHITECTURE

The block diagram of the overall system is presented in figure 1 and each element is described below.

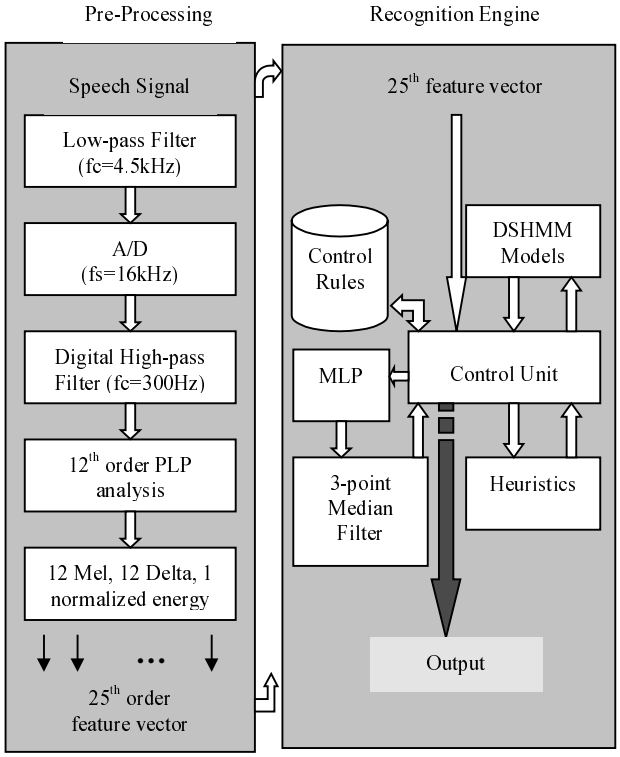

Figure 1. Block diagram of the vowel detection system.

2.1 Pre-processing

The pre-processing module of the system transforms the speech signal into proper feature vectors to be used as input to the recognition engine.

The speech signal is low-pass filtered at 4.5kHz, sampled at 16kHz and quantized with 16-bit accuracy. The digitized signal is further high-pass filtered at 300Hz by a fourth-order Butterworth digital filter to eliminate low-frequency hum or noise. A feature vector is obtained for each 25ms segment of speech with a 15ms frame step in an overlapping mode.

After applying a Hamming window each frame is analyzed using the PLP (Perceptual Linear Predicative) speech analysis technique [2] to obtain the characteristic parameters of the signal. In the PLP technique several well-known properties of hearing are simulated and an autoregressive 12th order all-pole model approximates the resulting spectrum of speech.

The 12 Mel cepstral coefficients and the corresponding delta coefficients, along with a normalized energy factor result to a 25^{th} order parameter vector.

2.2 Recognition Engine

For detecting the vowels, we use a multi-level combination of three different classification stages: multi layer perceptrons, hidden markov models and heuristics rules, along with three sets of control rules.

2.2.1 Classification stages

Multi layer perceptrons (MLP) are employed in the first stage of the recognition module. In particular a 3-layer (2 hidden layers) feedforward artificial neural network is used. The overall architecture of the network, i.e., the number of hidden layers and the number of nodes per hidden layer, is determined experimentally by training the network. Our experiments resulted to a network size of 25x12x8x1. A fast version of the back propagation algorithm was used for training [3]. The classification output was then passed through a three-point median filter to eliminate isolated "impulse" noise.

Discrete Hidden Markov Models (HMM) are used in the second-stage. Specifically a standard 4-state left-to-right model is used and the codebook size is set to 512. Three different models are established for each vowel: one model for the middle part of the vowel, one for the left and one for the right. So three models are available after training the discrete HMMs.

The third stage consists of a set of *heuristics rules* (HRULES) based on previous work [4]. The spotting process employing the heuristics consists of the following steps:

Step 1. *Energy contour estimation.* The short time energy function is calculated using 40ms frames with 5ms steps in an overlapping mode. The energy function is then smoothed using a 3-point Hanning window.

Step 2. *Vowel candidates location.* From the smooth energy function, the extremes (peaks and dips) are located by applying a usual peak-picking procedure. Peaks with energy value below a threshold T_L are rejected.

Step 3. *Ripple rejection.* For every candidate a relative measure of its peak and dip energy level is estimated.

Step 4. *Strong consonant rejection.* The parameter back-to-total cavity volume ratio (BTR parameter) is used to reject any remaining strong consonants, which behave as vowels.

Step 5. All remaining candidates are accepted as being the speech events (vowel phones).

The parameters used at the previous steps are calculated automatically from the specific database employed for training the system. The duration of each vowel is calculated as a function of peaks, dips, BTR parameters and the distances between successive peaks.

2.2.2 Control Unit

The control unit, based on three different set of rules (SET1, SET2, SET3) described below, selects the appropriate classification stage. Figure 2 shows the block diagram of the control unit whose functions are described in the following steps (see also Fig 1):

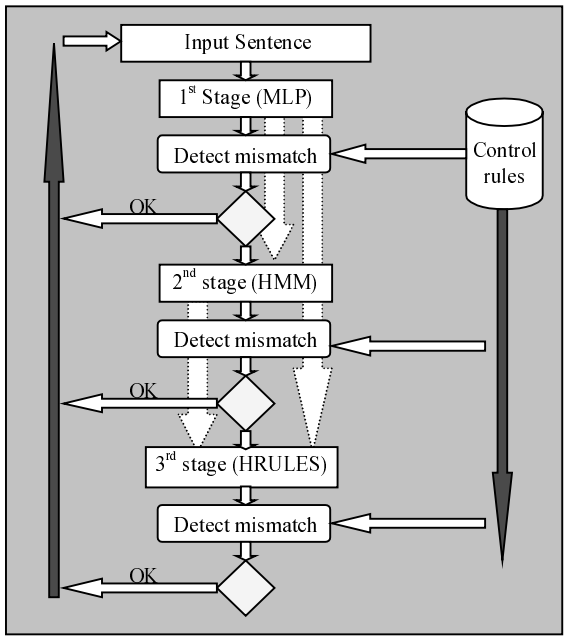

Figure 2. Block diagram of the Control unit.

Step 1: A sentence is put in the pre-processing unit. The control unit activates the 1st stage (MLP).

Step 2: The output of the MLP, after passing through the median filter, is fed back to the control unit.

Step 3: The control unit, by using the SET1 rules, decides whether there is a falsely rejected vowel or a falsely accepted non-vowel phoneme. When a mismatch of this kind is detected, it proceeds to the next step, otherwise it goes to step 1.

Step 4: The second stage (HMM) is activated. The control unit passes the mismatched part of this sentence to the HMM stage.

Step 5: Both the MLP and HMM outputs along with the SET2 rules are used to calculate a "success" factor. If this factor is greater than 1, then the process goes to step 1, otherwise it goes to step 6.

Step 6: The third stage (HRULES) is activated, which takes as input the mismatch region (calculated by the two previous stages).

Step 7: All the different outputs of the three stages along with the SET3 rules are used to calculate a new "success factor". If it exceeds 1, the process goes to step 1 and handles the next sentence. Otherwise it goes to step 8.

Step 8: A fatal error is raised which means that there is a doubtful region in the sentence. Go to step 1.

When a fatal error is raised, the system can not correct it. A possible solution is to increase the frame rate of the pre-processing and repeat the above procedure.

We can see from the above procedure that the rules used by the control unit are of great importance for the performance of the whole system because the flow decision is mainly based on them.

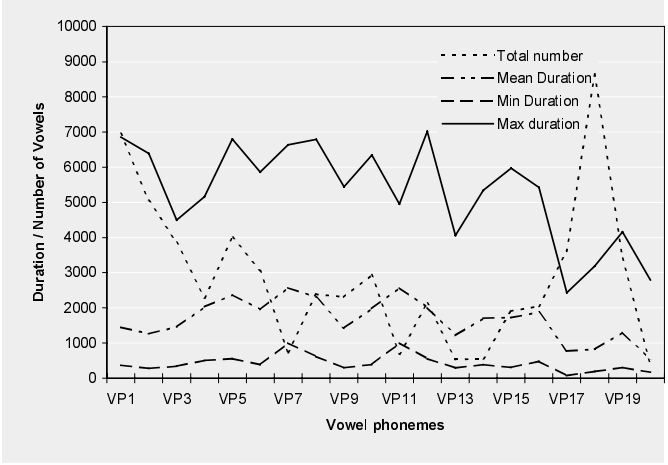

Figure 3. Duration statistics of the training set for the total numbers of each vowel phoneme (VP) category. The minimum, maximum and mean duration of each phoneme category are shown.

All the parameters of the rules are calculated once by statistics extracted from the database used for training the system.

The three sets of rules (SET1, SET2 and SET3) are estimated as a function of the following three factors: distance (FI), duration (FU) and amplitude (FA), derived as follows:

$$FI = \left| I_x \sum_{i=1}^{3} a_i - \left(a_1 \cdot SI_{Min} + a_2 \cdot SI_{Max} + a_3 \cdot SI_{Mean} \right) \right|,$$

$$FU = \left| U_x \sum_{i=1}^{3} b_i - \left(b_1 \cdot SU_{Min} + b_2 \cdot SU_{Max} + b_3 \cdot SU_{Mean} \right) \right|,$$

$$FA = \left| A_x \sum_{i=1}^{3} c_i - \left(c_1 \cdot SA_{Min} + c_2 \cdot SA_{Max} + c_3 \cdot SA_{Mean} \right) \right|,$$

where SI, SU and SA are the parameters (min, max and mean value) of distance between successive vowels, duration and amplitude of the vowel, respectively. The a_i, b_i and c_i (i=1,2,3) are parameters calculated as a function of the statistics of the database and the outputs of the classification stages for the specific region. I_x is the distance between two vowels, U_x is the duration of the x^{th} vowel and A_x the amplitude of the x^{th} vowel.

The SET1 is estimated as a function of the distance factor (FI), duration factor (FU) and the MLP outputs. The SET2 is estimated as a function of the distance factor (FI), duration factor (FU), HMM outputs and MLP outputs. The SET3 is estimated as a function of the distance factor (FI), duration factor (FU), amplitude factor (FA), and HRULES, HMM and MLP outputs. Both SET2 and SET3 use a "success" factor(S_F) given from the following equation:

$$S_F = \frac{\left(F_D + \sum F_{DU} \right)_{NEW}}{\left(F_D + \sum F_{DU} \right)_{OLD}}$$

For SET2 and SET3 the outputs of the MLP, HMM and HRULES are weighted by a factor which depends on the digital error rates of the training procedures.

3. EXPERIMENTAL RESULTS

3.1 Database description

For the training and testing procedures the Texas Instruments/Massachusetts Institute of Technology (TIMIT) acoustic-phonetic corpus of read speech was used [10]. This database contains a total of 6,300 utterances, 10 sentences spoken by 630 speakers from 8 major dialect regions of the United States. The database is separated in two portions, one for training and one for testing. We used the training set of the database (462 speakers) for training our system and the test set (168 speakers – 1344 sentences) for testing. The vowel categories used for the transcription files of the TIMIT database were 20.

3.2 Training

In the following we describe the training procedure of the system using the TIMIT database. We describe the procedures for estimating the statistics and training the MLP and the HMM stages. The heuristic rules are estimated as a function of the statistics.

3.2.1 Statistics

The first step in training the system is to calculate the statistics based on the training database. The training set of the TIMIT database consists of 462 speakers (4,620 sentences – 57,463 vowels). The statistics are vowel duration parameters (min, max and mean duration for each vowel category), distance between vowels (min, max and mean distance between the successive vowels of a sentence), peaks and dips of the speech data files and the corresponding energy function.

The statistics for the duration of the vowel phonemes of the training set are shown in figure 3. Similar statistics were obtained for the distances between different vowel phoneme categories in the sentences of the training set.

3.2.2 Training the MLP

In order to train the MLP we used vowels and non-vowels from the training set. For each phoneme category we applied a standard K-means algorithm with a different number of centroids depending on the number of vowels found.

- <500 Vowels: Number of centers: 128
- 500-1,500 Vowels: Number of centers: 256
- 1,500-3,000 Vowels: Number of centers: 512
- >3,000 Vowels: Number of centers: 1,024

The K-means was applied three times: once for the middle part of the vowel, once for the left part and once for the right part. For each part it is obvious that there were more than one parameter vectors and all those vectors were used in the K-means.

The results of the multiple K-means along with non-vowel parameter vectors, are used for training the MLP network.

3.2.3 Training the DSHMMs

For the 20 vowel phoneme categories of the database we created 3 discrete HMM models. One model for the middle part of the vowel, one for the left and one for the right. The VQ size was set to 512 and the standard K-means algorithm was used to create the codebooks. We didn't use different HMMs for each category

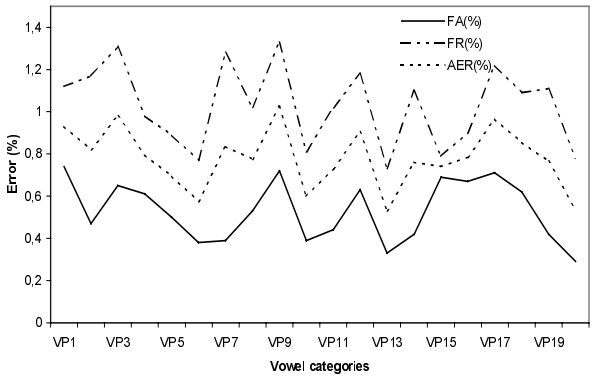

Figure 4. False acceptance (FA), false rejection (FR) and average error rate (AER) for each of the vowel phoneme categories of the TIMIT database.

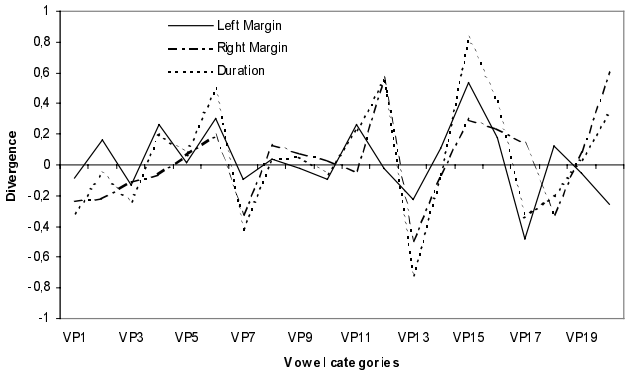

Figure 5. Divergence of the located vowels due to the left margin, right margin and duration of the transcription files of the database. The 1 and –1 are for one vector left to a one vector right miss-positioning.

as our objective is vowel spotting and not the distinction between vowels.

Therefore, all the 57,463 vowels found on the training set of the TIMIT database were used to train our HMM.

3.3 Testing

For testing the system we used the test set of the TIMIT database, which consists of 1,344 sentences (20,911 vowels) in total. Three different tests were performed.

First, we tested the improvement to the performance of the system that each stage achieves. Table 1 shows the error rates (false acceptance FA, false rejection FR and average error rate AER) for the 20,911 vowels by using one, two or all stages of the system.

	FA (%)	FR (%)	AER (%)
Stage 1	6.31	7.22	6.77
+ Stage 2	2.14	2.01	2.08
+ Stage 3	0.53	1.02	0.78

Table 1. False acceptance (FA), false rejection (FR) and average error rate (AER) of the system using the TIMIT test data sets.

The second test was for the performance of the detection system for each vowel category. In figure 4 we show the FA, FR and AER for each of the vowel categories.

The third test was for the position and duration of the detected vowels. Supposing that the correct vowel durations are those from the transcription files of TIMIT, the results show the percentage of the detected vowels and the divergence to the transcribted margins. The results are shown in figure 5.

4. CONCLUSIONS

We presented a vowel detection system, which uses neural networks, discrete hidden markov models and heuristics rules controlled by a central processing unit. This unit with the help of a sophisticated set of rules, decides which step is needed in order to avoid a false acceptance of a non-vowel or a false rejection of a vowel phoneme. The purpose of the overall system is to detect with a high precision the vowels in a spoken sentence. The only disadvantage of the proposed system is the need of a well-defined training database with transcription files of the speech data. Its performance is superior in comparison to other systems and apart from the long training time, the response of the system is in real time.

5. REFERENCES

[1] Laurence Rabiner, Bing-Hwang Juang. *Fundamentals of speech recognition.* Prentice Hall, 1993.

[2] H. Hermansky. *Perceptual linear predictive (PLP) analysis for speech.* J. Acoust. Soc. Am., pp. 1738-1752, 1990.

[3] D. Anguita, M. Pampolini, G. Parodi , R. Zunino. *YPROP: Yet Another Accelerating Technique for the Back Propagation.* ICANN '93, September 13-16 1993, Amsterdam, The Netherlands, pp. 500-503.

[4] N. Fakotakis, A. Tsopanoglou, G. Kokkinakis. *A text-independent speaker recognition system based on vowel spotting*. Speech Communication 12 (1993), pp. 57-68

[5] J. Sirigos, V. Darsinos, N. Fakotakis, G. Kokkinakis. *Vowel – non vowel decision using neural networks and rules*. Eusipco 96, Trieste, Italy.

[6] N. Fakotakis, K. Georgila, A. Tsopanoglou. *A continuous HMM Text-independent speaker recognition system based on vowel-spotting*. Eurospeech 97, Rhodes, Greece, pp. 2347-2350.

[7] Kewley-Port, D., & Watson, C.S. *Formant- frequency discrimination for isolated English vowels*. Journal of the Acoustical Society of America, 1995, 485- 496.

[8] G.F. Meyer, R.W.L. Kortelkaas and D.J. Hermes. *Vowel-Onset Detection with Models of the Auditory Periphery and Cochlear Nucleus.* Proc. Inst. of Acoustics V16(5) 231-238 (1994)

[9] Richard, G., Lin, Q., Zussa, F., Sinder, D., Che, C., and Flanagan, J. *Vowel recognition using an articulatory representation.* J. Acous. Soc. Am. Vol 98 (5S), p. 2931.

[10] W. Fisher, V. Zue, J. Bernstein and D. Pallet. *An Acoustic-Phonetic Database.* JASA, Suppl. A. Vol. 81 (592) 1986.