VOWEL-NON VOWEL CLASSIFICATION OF SPEECH USING AN MLP AND RULES

John Sirigos, john@wcl.ee.upatras.gr
Vassilis Darsinos, darsinos@wcl.ee.upatras.gr
Nikos Fakotakis, fakotaki@wcl.ee.upatras.gr
George Kokkinakis, gkokkin@wcl.ee.upatras.gr

Wire Communications Laboratory, University of Patras, 26500 Patras, Greece

ABSTRACT

In this paper we present a high precision speaker independent vowel/non vowel classifier based on a simple feed forward MLP (Multi Layer Perceptron) and several rules. RASTA-PLP analysis of the speech signal resulting to mel-cepstral coefficients and a formant tracking method are used in order to provide the feature vectors for the MLP. To train and test the system we used a part of the TIMIT database. The results indicate that the performance of this classifier for speaker independent vowel classification is approximately 97.25% so it can be favorably used for speaker recognition or speech labeling purposes.

1 INTRODUCTION

The classification of the speech signal into vowel and non vowel segments provides a preliminary acoustic labeling of speech, which can be very important for both speech and speaker recognition procedures.

The vowel/non vowel classification could be made using a single parameter derived from the speech signal such as rms energy or zero-crossing rate. Such a method can only achieve limited accuracy because the value of any single parameter usually overlaps between categories, particularly when the speech is not recorded in a high fidelity environment. There is also the possibility that such a parameter could also describe other phoneme categories as well as vowels.

In this paper we describe a vowel/non vowel classification system which achieves high accuracy rates.

The proposed classifier is based on formant tracking, RASTA-PLP analysis for feature extraction and on an MLP network and rules for the classification and decision procedure.

2 GENERAL DESCRIPTION

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

2.1 Pre-processing

The sampled speech data are segmented into 30ms frames with an overlap of 20ms. After application of a hamming window, each of these frames is analyzed using the RASTA-PLP speech analysis technique, in order to make speech analysis more robust to spectral distortions [1][2]. A 15th order all-pole model is fitted to the spectrum computed with the RASTA-PLP method.

2.1.1 Rasta PLP Analysis

The Rasta-PLP analysis to estimate the 15 mel cepstral coefficients, includes the following steps:

a. Fast Fourier Transform: The windowed speech segment is transformed into the frequency domain.

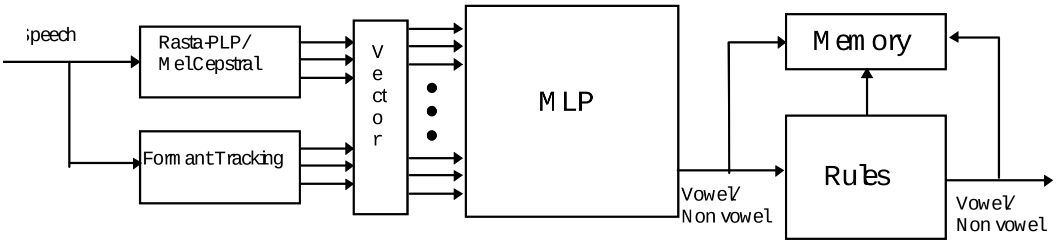

Figure 1. A block diagram of the vowel/non-vowel classifier.

b. Critical band integration and re-sampling: The spectrum is warped along its frequency axis into the Bark frequency scale and then a down-sampling is performed.

c. Logarithm: The logarithm of the critical-band spectrum, is calculated.

d. Equal-Loudness Curve: The spectrum is preemphasize using the simulated equal-loudness curve.

e. Power-Law of Hearing: Cubic-root amplitude compression is carried out.

f. Inverse Logarithm: The inverse logarithm of this relative log spectrum is calculated, yielding a relative auditory spectrum.

g. Inverse Discrete Fourier transform: The inverse discrete fourier transform is performed.

h. Solving the set of linear equations (Durbin): An all pole model of the spectrum is computed following the conventional PLP technique.

i. Cepstral recursion: The 15 mel cepstral coefficients are computed.

2.1.2 Formant extraction and tracking

The speech signal passes through the formant tracking module where the first four formants and the corresponding bandwidths are computed for each frame. For this purpose a N-Best tracking algorithm is used.

The usefulness of formant information has been recognized in the past and put to use in speech recognition systems. However, the accuracy and reliability of the formant trackers turned to be too low for the demands of speech recognition systems. Hence the current focus on using spectral representation in conjunction with dynamic programming tools.

Estimating the formants based on short term spectral analysis can be done successfully in the case where the spectral information is clearly pronounced. However the more interesting case is when the short time spectrum is relatively flat or ambiguous. In that case the location of the formants can only be determined by tracking the formants. Taking in mind the global information, the algorithm is allowed to compensate for incomplete local information.

It has been observed [7] that for most speech segments, only a few consistent interpretations of formants can be made. Thus, designing a formant tracker that finds the N-best tracks seems to be the right solution, instead of a single best tracker.

This approach has several advantages over more conventional systems. Noise in a particular frequency band influences all spectral coefficients. On the other hand, a formant representation is more robust to such noise. Thus if the estimation of the formant is obscured, it will be recovered by using consistency constraints with respect to the adjacent frames to estimate the formant location. The same process will reduce the problems for heavily glottalized speech as well as strong extraneous noise events as clicks [9].

2.1.3 Elementary tracks

In the first step the identification of individual formants is accomplished. There are three basic mechanisms for estimating formant candidates: a) computing the complex roots of a linear predictor polynomial [10] b) peak picking of a short-time spectral representation [11] and c) analysis by synthesis [12]. In this work the first solution is adopted using a 12-th order LPC polynomial and solving for the formants and bandwidths. The elementary tracks are computed connecting the formant candidates and using frequency constraints for each formant and constraints of the types F1-F2 and F2-F3.

2.1.4 Formant selection and tracking

Using the elementary tracks as they are computed from the above procedure, a formant selection, correction and connection of individual tracks must be done. For this reason least-square polynomial tracks are computed for each formant, using the elementary tracks. The order of the polynomial is 4. After that each formant candidate is examined using distance criteria. The distances that are computed are: from the elementary track that it belongs, from the next elementary track and from the corresponding LS-polynomial. A set of rules based on these distances, decides if the formant belong to another track or if it must be discarded.

2.1.5 Interpolation

The final processing step reconstructs the missing regions of the formant tracks. This is accomplished using cubic spline interpolation. This procedure guarantees a formant smoothness and does not introduces any artifacts.

For each sentence the tracks of the first four formants are extracted in the three step procedure described above.

After that the resulting 23th order vectors are used as the feature sets for the MLP classifier.

2.2 Multi Layer Perceptron (MLP)

A simple feed forward MLP is used for the classification of each frame as vowel or non vowel. The size of the MLP was experimentally derived to 23x12x7x1 (23 for the input layer, 12 for the first hidden layer, 7 for the second hidden layer, and 1 for the output layer). To this end, we performed several tests with one and two hidden layers varying their units from 5 to 200.

The MLP classifier for the vowel/non-vowel model is trained using a fast version of the Back Propagation Algorithm as described in [4].

2.3 Rules - Memory

The output of the MLP net is fed into a set of heuristic rules and saved into the memory module. In this module all decisions of the MLP for the last 24 frames are held. The rules use this memory to decide whether a frame is vowel or non vowel considering 12 frames preceding and 12 frames following the current frame. That implies the use of the look-ahead memory.

Using simple distance and duration rules, the output of the MLP is rejected if the duration of a candidate vowel is less than 30ms and more than 300ms. Peaks present at the output exceeding prescribed values are also rejected along with low level outputs (less than 0.2).

3 DATABASE - TRAINING

For training and testing the classifier the Texas Instruments/Massachusetts Institute of Technology (TIMIT) acoustic-phonetic corpus of read speech was used. This database contains a total of 6,300 utterances, 10 sentences spoken by 630 speakers from 8 major dialect regions of the United States. 70% of the speakers are male and 30% female. Each speaker reads 2 dialect sentences, 5 phonemically-compact sentences and 3 phonetically-diverse sentences. The 10 sentences represent roughly 30 seconds of speech material per speaker. A dialect region (Northern) and the corresponding train set of 76 speakers (53 male and 23 female) was randomly chosen to be used for training and testing the system.

4 EXPERIMENTAL RESULTS

In order to train the speaker independent classifier, we used only 16 speakers from the above set (10 male and 6 female). For each speaker we used the five SX (phonetically-compact) sentences and their corresponding transcription file.

For testing the system we used the remaining 60 speakers (43 male and 17 female) and also North Midland and Western dialect train data sets. The false acceptance error rate achieved (FA) (falsely accepted vowels over total number of vowels) and the false rejection error rate (FR) (falsely rejected vowels over total number of vowels) are presented in the table below.

Data Set Region Dialect	***FA(%)***	***FR(%)***
Northern	2.31%	3.24%
North Midland	2.53%	3.12%
Western	2.27%	3.22%

5 CONCLUSIONS

The vowel/non-vowel classifier described in this paper features a high recognition accuracy and significant improvements, in comparison with similar reported methods [5][6]. The achieved results show that the system is speaker and dialect independent. Thus it can be useful for speech or speaker recognition purposes. More sophisticated rules can improve the overall system performance and reduce further the FA and FR error rates.

6 REFERENCES

[1] Hynek Hermansky, Nelson Morgan, Aruna Bayya, Phil Kohn, "*Compensation for the effect of the communication channel in auditory-like analysis of speech (RASTA-PLP)*", In. Proc. Eurospeech '91, pp. 1367-1371, Genove, Italy, 1991.

[2] J. Sirigos, N. Fakotakis, G. Kokkinakis, "*A comparison of several speech parameters for speaker independent speech recognition and speaker recognition*", In. Proc. Eurospeech '95, vol. II, pp. 1407-1409, Mandrid, Spain, 1995.

[3] Atal .B, Hanuer S. (1971) "Speech analysis and synthesis by linear prediction of the speech wave" JASA 50, 1971.

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

[5] Yingyong Qi, Bobby R. Hunt, "*Voiced-Unvoiced-Silence Classifications of Speech Using Hybrid Features and a Network Classifier*", IEEE Transactions on Speech and Audio Processing, Vol. 1, No. 2, pp. 250-255, April 1993.

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

[7] Laprie Y., Berger M. (1994) *"A new paradigm for reliable automatic formant tracking"* In. Proc. ICASSP, 1994.

[8] Allen J. (1994) *"How do humans process and recognize speech?", In. Proc.* ICASSP 1994.

[9] Miller J. (1989) "*Auditory-perceptual interpretation of the vowel"*. JASA 85, 1989

[10] Atatl .B, Hanauaer S. (1971*) "Speech analysis and synthesis by linear prediction of the speech wave"* JASA 50, 1971.

[11] Schafer R., Rabiner L. (1970*) "System for automatic formant analysis of voiced speech"* JASA 57 1970.

[12] Olive J. (1971) "*Automatic formant tracking by a Newton-Raphson Technique"* JASA 50, 1971.