Binary Phoneme Classification Using Fixed and Adaptive Segment-Based Neural Network Approach

Lotfi Messikh[1] and Mouldi Bedda[2]
[1]Electronic Department Annaba University Algeria, Algeria
[2]College of Engineering, ALJouf University, KSA

Abstract: *This paper addresses the problem of binary phoneme classification via a neural net segment-based approach. Phoneme groups are categorized based on articulatory information. For an efficient segmental acoustic properties capture, the phoneme associated with a speech segment is represented using MFCC's features extracted from different portions of that segment as well as its duration. These portions are obtained with fixed or variable size analysis. The classification is done with a Multi-Layer Perceptron trained using the Mackay's Bayesian approach. Experimental results obtained from the Otago speech corpus favourites the use of fixed segmentation strategies over adaptive ones for resolving consonants/vowels, Fricatives/non fricatives, nasals/non nasals and stops/non stops binary classification problems.*

Keywords: *Signal segmentation, binary phoneme classification, segment-based processing, and neural network.*

Received November 11, 2008; accepted May 17, 2009

1. Introduction

Hidden Markov Models (HMMs) and Artificial Neural Network Models (ANNs) have been the most dominant frame-based acoustic modelling techniques for automatic speech recognition to date. To address the limitations of theses models in constructing the acoustic representation of the words in the dictionary, in incorporating acoustic information about speech, such as phone duration and intonation, and capturing efficiently the spectral/temporal relationships over the whole phone, segment-based models, have been developed [1, 2, 3, 4, 5, 6]. Typically such approach contains three parts: segmentation, feature extraction and classification. In the first part, the signal is divided into segments, each segment corresponding to one phone. While the frame duration is fixed, the segment duration varies since it depends on the phone length. Feature extraction can be done on each segment using fixed or variable size analysis windows. The coefficients from several windows are combined into a vector of features, which is input to the classification unit. The classification part assigns preferences to the relevant phones. The classification can be hierarchical or non hierarchical. For hierarchical classification, the sound is firstly classified into a phonetic group, and then, it is classified within the group. The classification can be done for example with a Multi-Layer Perceptron (MLP), which can performs notable nonlinear feature discrimination.

In this work, we conduct a MLP-based phoneme classification experiments based on MFCC segmental feature vector, and perform comparison between fixed and adaptive size analysis windows. Before doing the feature extraction phase, the phonemes are divided into a given number of portions by one of the segmentation procedures given in section two. In the case of an adaptive analysis, the variable size windows are obtained via optimising the global squared error criterion of a composite linear prediction model, and is achieved by means of dynamic programming.

2. Segmental Spectral Representation

2.1. Fixed Segmentation Approach

In the commonly used fixed segment-Based Spectral Representation (FSR), a speech segment is divided into three consecutive regions having fixed relative durations in comparison with the length of the interested segment. According to [1], the relative window sizes are set respectively to 25%, 50% and 25%. A 13-dimentional MFCC-based feature vector is extracted from the signal in each of the three region of the interested segment. The three feature vectors are then concatenated together, and also with the duration of the segment. Therefore, a feature vector of 40 dimensions is used to represent the signal in one speech segment.

In this study we shall divide a speech segment into M consecutive regions. The relative window sizes

are set respectively to $(2M-2)^{-1}$ % $(M-1)^{-1}$%,…, $(M-1)^{-1}$%, $(2M-2)^{-1}$ %

A 13-Dimentional MFCC-based feature vector is extracted from the signal in each of the M regions of the interested segment. The M feature vectors are then concatenated together, and also with the duration of the segment.

2.2. Adaptive Segmentation

In the context of speech modelling, the major difficulty of this fixed segment-based approach is the possible presence of abrupt signal transitions within a windowed segment leading to poor obtained estimations. Instead of using the above fixed segment-based Spectral Representation, one can use an Adaptive segment-based Spectral Representation (ASR) [5, 7]. One of the possible solutions consider that the observations are generated by switching among M different $AR(p)$ models of coefficients $a_j = (a_{1,j},...,a_{p,j})$; i.e.,

$$x(n) = \sum_{m=1}^{M} w_m(n) x_m(n) \tag{1}$$

where $w_m(n)$ selects the samples generated by the mth AR model:

$$w_m(n) = \begin{cases} 1, if\ \ x(n)\ is\ \ generated \quad by\ \ AR_m \\ 0, otherwise \end{cases} \tag{2}$$

The output at time instant n for model AR_m of p-order is given by:

$$x_m(n) = \sum_{i=1}^{p} a_{i,m} x(n-i) + e_{p,m}(n) \tag{3}$$

where $e_{p,m}(n)$ is a zero mean uncorrelated Gaussian noise with variance $\sigma^2_{p,m}$.

Now, the problem can be stated as follows: given the number of models M, their order p, and the observation vector $x = (x(0), x(1),..., x(N-1))$, determine the boundaries $t = (t_1, t_2,..., t_{M-1})$ between segments and find the best model for each segment.

The goal of the general Rate/Distortion Algorithm (RDA) is to arrive at a minimization of the global squared error with respect to the local linear prediction orders and to the data segmentation using the global rate as a parameter controlling the number of segments and the distribution of linear prediction resources amongst the segments [7]. In the context of a fixed segments number and a fixed order for each AR model, the goal is simply to minimise the global squared error with respect to the local linear prediction coefficients and to the data segmentation. Formally, this amounts to solving the following problem:

$$\min_{0<Ct_1<...<Ct_{M-1}<N} \left\{ \sum_{m=1}^{M} \sum_{n=Ct_{m-1}}^{Ct_m - 1} \left(e_{p,m}(n)\right)^2 \right\} \tag{4}$$

where the integer number C is introduced to answer a sufficient window size to accurate AR parameter estimation.

In this paper, the AR parameters for a given dataset are determined by the autocorrelation method [6] with a rectangular windowing. The optimizations of equation 4 are performed by means of dynamic programming.

3. MLP-Based Phoneme Classification

Four binary classification problems are considered here: consonant/vowel, Fricative/non fricative, nasal/non nasal and stop/non stop. For each classification problem, the associated phone-level transcription of the phonemes Otago database [8] training set were converted into feature transcriptions. The resulting feature transcriptions and the parameterized speech segments constituted the training material for a MLP classifier. The network has one hidden layer and one output layer. All elements of the feature vector are scaled so that they will zero mean and unity standard deviation on the training set. They serves as input data for the MLP classifier. The output layer has two output nodes that decides whether the current segment is one of the two considered classes. In the hidden and output layer, we used hyperbolic tangent and logarithmic sigmoid respectively as activation functions. For improving network generalisation only one neuron is used in the hidden layer and the MLP was trained using the Bayesian framework of Mackay [8].

4. Experimentation

The realized phonemes used this experimentation are from the Otago speech corpora [8], they were sampled at a 22.05 kHz and 16 bit resolution. To reduce the computational complexity and to enhance the spectral discrimination, each signal was down sampled at 8 kHz, pre-emphasized ($z_0 = 0.995$), normalized to unit amplitude. In the case of variable scale analysis, the speech signals are also segmented into juxtaposed frames of 4ms in order to reduce the computational complexity required in optimising the speech partitions. A total of 4657 clean phonemes were used in our study. 3055 speech corpus phonemes, recorded from 8 speakers (5 males and 3 females), were used in our training set. The rest 1602 digit corpus phonemes, recorded from 20 speakers (11 males and 9 females) are used for classification evaluation as our test set.

Figure 1 shows the classification-rate on the training set versus the number of used phoneme portions for both fixed segment-based approach (solid line) and adaptive segment-based approach (dashed line), for Consonant (C), Fricative (F), Nasal (N) and

Stop (S) detection. The FSR and the ASR provides good results as the number of portions grows. In most cases the FSR is more robust then the ASR. Table 1 lists the classification rate for the training set. The classification rate is defined as the ratio of the number of correctly classified tokens to that of the total tokens of the considered set.

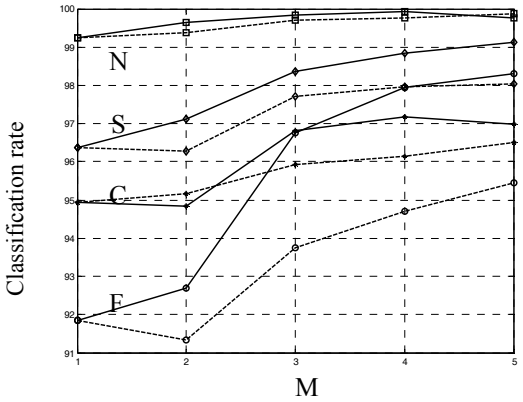

Figure 1. Classification results on the training set.

Table 1. Classification rate (%) for the training set.

	M=1	M=2	M=3	M=4	M=5
CFSR	94.93	94.83	96.82	97.18	96.99
CASR	94.93	95.16	95.94	96.14	96.50
FFSR	91.85	92.70	96.76	97.94	98.30
FASR	91.85	91.33	93.75	94.70	95.45
NFSR	99.25	99.64	99.84	99.93	99.77
NASR	99.25	99.38	99.71	99.77	99.87
SFSR	96.37	97.12	98.36	98.85	99.12
SASR	96.37	96.27	97.71	97.97	98.04

Table 2 shows the classification rate on the unseen test set versus the number of used phoneme portions for FSR and ASR, for C, F, N and S detection. On the unseen test the classification rate are less accurate and there are no obvious improvement of accuracy as M grows. However, in most studied cases, the FSR remains more robust then the ASR.

From the results in Tables 1 and 2, it appears that the use of the FSR leads to good results in comparison to the used ASR. It appears also that for detecting Consonants, Fricatives, stops and nasals, it is not necessary to detect abrupt changes which occurs inside the phoneme segments.

5. Conclusions

This paper presents a study of Consonant, Fricative, Nasal and Stop detection problems for English phonemes, where the segmental feature vectors are obtained after a fixed or an adaptive segmentation. Performance comparison of the two classification approaches indicates the superiority of FSR over ASR in terms of time consuming and accuracy performance.

Table 2. Classification rate (%) for unseen data.

	M=1	M=2	M=3	M=4	M=5
CFSR	83.02	82.96	83.90	84.02	81.15
CASR	83.02	82.65	82.27	82.65	82.46
FFSR	86.77	88.20	89.45	88.64	89.08
FASR	86.77	87.64	85.33	86.27	86.52
NFSR	94.57	93.32	94.32	92.82	93.13
NASR	94.57	92.26	92.76	92.51	91.45
SFSR	90.20	88.70	90.26	93.07	92.32
SASR	90.20	90.51	90.20	90.20	88.08

It has to be noted, however, that the proposed approach is not directly applicable to phoneme classification in continuous speech. While the speech signal must be adequately segmented into phoneme units, a fixed segmentation of these units seems to be sufficient to produce good classification results for the considered detection problems. Further studies are then needed to adapt our approach to such application.

References

[1] Ekkachaiworrasin N., Punyabukkana P., and Suchato A., "Phoneme Classification Study for Thai Segment-Based Acoustic Models," *in Proceedings of International Symposium on Communications and Information Technologies IEEE*, Bangkok, pp. 122-127, 2006.

[2] Glass J., "A Probabilistic Framework for Segment-Based Speech Recognition," *Computer Journal of Computer Speech and Language*, vol. 17, no. 3, pp. 137-152, 2003.

[3] Keiler F., Arfib D., and Zolzer U., "Efficient Linear Prediction for Digital Audio Effect," *in Proceedings of the COST G-6 Conference on Digital Audio Effect*, Italy, pp. 490-498, 2000.

[4] Mackay C., "A Practical Bayesian Framework for Back Propagation Networks," *Computer Journal of Neural Computation*, vol. 4, no. 6, pp. 448-472, 1992.

[5] Messikh L. and Bedda M., "Performance Comparison of Two Joint Three States Segmentation and AR Modelling Algorithms," *Computer Journal of Research Applied Sciences*, vol. 2, no. 6, pp. 704-707, 2007.

[6] Ostendorf M., Digalakis V., and Kimball O., "From HMMs to Segment Models: A Unified View of Stochastic Modeling for Speech Recognition," *Computer Journal of IEEE Transactions on Speech and Audio Processing*, vol. 4, no. 5, pp. 360-378, 1996.

[7] Prandoni P. and Vetterli M., "R/D Optimal Linear Prediction," *Computer Journal of IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 6, pp. 166-169, 2000.

[8] Sinclair S. and Watson C., "The Development of the Otago Speech Database," *in Proceedings of Artificial Neural Networks and Expert Systems*, CA, pp. 145-147, 1995.

Lotfi Messikh is an associate researcher at the center l'URER/MS, Adrar. He holds PhD in specch processing from the University of Badji Mokhtar Annaba, Algeria. His research interests focus on automatic speech processing and system design for photovoltaic applications.

Mouldi Bedda received his PhD degree in electrical engineering from the University Nancy 2, France in 1985. From 1985-2006, he worked with the University Badji Mokhtar Annaba, Algeria. He was the director of Automatic and Signals Laboratory from 2001-2006. From 2006 to date full professor at the college of engineering of aljouf university KSA. He research interest**:** DSP, speech processing , OCR, artificial intelligence, biomedical engineering and image processing