International Journal of Signal Processing, Image Processing and Pattern Recognition
Vol.8, No.3 (2015), pp.131-140
http://dx.doi.org/10.14257/ijsip.2015.8.3.12

Phoneme Classification Using New Feature Extraction Techniques based on Mellin Transform

Rimah Amami* and Noureddine Ellouze

National School of Engineering, Tunis, Tunisia
rimah.amami@yahoo.fr

Abstract

This paper presents a new hierarchical phoneme recognition system using the SVM classifier and different feature representations based on mellin transform.

The proposed architecture uses different representations with each group of phonemes of the speech database TIMIT which are distributed in a way to reduce the confusions between phonemes having similar articulatory strcuture. The main idea of this new architecture is based on the principle that each group of phonemes has his own characteristics which requires us to choose the adequate representation for the given group.

Experiments have proven the robustness of our new hierarchical phoneme recognition system (called MMP) and the use of conventional feature representations based on mellin transform explains its superior recognition performance.

Keywords: *Phoneme recognition, MFCC, mellin transform, Phonemes confusions, SVM*

1. Introduction

Automatic speech recognition filed has been considered and explored over the years and the phoneme recognition is one of the most important tasks reported in the state-of-the-art speech recognition systems. Its importance resides in the fact that phonemes maps, meaningful characteristics of the speech signal as they are the smallest speech sound units in a language.

In automatic speech recognition systems, phonemes provide the advantage that their number is limited in any language.

Therefore, phonemes recognition is considered the first step in speech recognition processing. However, the reliable performance and fully robustness of those systems are still not within reach.

The fundamental difficulty of phoneme recognition lies in the speech variability and perception, the acoustic-phonetic characteristics of speech, the classifier and feature extraction technique used for the recognition task, the acoustic conditions, speaker condition, etc.

In the literature, phoneme recognition systems rely on the importance of the choice of the classifier which maps the training model without taking into account the impact of the acoustic-phonetic characteristics. Hence, the purpose of this paper is to present our findings during the implementation and evaluation of our hierarchical phoneme recognition system. The main objective is to use a suitable feature extraction technique at each level of the hierarchical phoneme system. It must be pointed out that the feature analysis step plays a determinant role in the overall performance of the speech recognition system because it extract the relevant and pertinent information of the speech signal.

In practice, the information in speech signal is represented by short term amplitude spectrum of the speech wave form. In other words, the features extraction of phonemes data are based on

ISSN: 2005-4254 IJSIP

the short term amplitude spectrum from speech which helps to reduce the large variability of the speech signal and removing irrelevant aspects of the speech.

Thus, in this paper we try to choose the feature technique which best describe each group of phonemes (*i.e.* vowels or consonants) of the hierarchical system. It is shown that using feature extraction methods based on scale transformation improves the speech recognition accuracy. Thus, in this study, we propose a phoneme recognition model using mellin transform combining to different feature extraction techniques. Therefore, the new extraction method has the advantage of both scale transform and the given conventional feature extraction method.

Moreover, the aim of applying the mellin transform to the spectral envelope ofthe signal is to achieve some kind of pitch, gender, age normalization of the pronounced phoneme and all this in efficient way. On the other hand, the classifier SVM is used to build the training model to consider in the recognition task [1-3].

The rest of this paper is organized as follows: in Section 2, we present an overview of support vector method used for phoneme classification. In section 3,we discuss the different feature extraction techniques used based on mellin transform. In Section 4, we discuss the architecture of the hierarchical phoneme recognition system. In Section 5, we present the experimental condition and discuss recognition experiments. We end with conclusions and future work.

2. Support Vector Machine

Support Vector Machine is a learning machine which was developed by Vladimir Vapnik to construct decision functions in the input space based on the theory of Structural Risk Minimization [1]. This theory aims to find the function f(x,a) which minimizes the risk functional:

$$R(\alpha) = \int L(y, f(x,\alpha))dP(x,y) \tag{1}$$

Furthermore, SVM consists of constructing one or several hyperplanes in order to separate the different classes. Nevertheless, an optimal hyperplane must be found. Vapnik and Cortes [1] defined an optimal hyperplane as the linear decision function with maximal margin between the vectors of the two classes. We consider the optimal hyperplane if it is separated the examples without error and if the distance between the closest example and the hyperplane is maximal. The hyperplane can be described as:

$$W^T x + b = 0, x \in R^d \tag{2}$$

In a binary task, the distance from each example to hyperplane is: $\text{margin} = 2/\|W\|$. The best hyperplane will find by making the margin largest. Hence, the optimal hyperplane is the one that minimizes functional: $\|W\|/2$. The solution to this optimization problem can be cast into the Lagrange function:

$$L(w,b,\alpha) = \tfrac{1}{2} W^T W - \sum_{i=1}^{l} \alpha_i \left[y_i (W^T x_i + b) - 1 \right] \tag{3}$$

Where $y_i(W^T x + b) \geq 1, i = 1,2,\ldots,l$ and the Lagrange multiplier α_i is corresponding to every training sample. Note that The Lagrangian has to be minimized with respect to W, b and maximized with respect to $\alpha_i \geq 0$.

Consider those conditions; the Lagrange functional can be substituted into the following equation in order to take into account the kühn–Tucker conditions:

$$L(W,b,a)=\sum_{i=1}^{l}\alpha_i - 1/2\sum y_i y_j \alpha_i \alpha_j (x_i.x_j) \tag{4}$$

Then, the classification consists of seeking the maximum of this function in the nonnegative quadrant with respect to $\sum_{i=1}^{l}\alpha_i y_i = 0$.

Note that the examples whose $\alpha_i \geq 0$ are called "support vectors". They are used to decide which hyperplane should be taken since this set of vectors is separated by the optimal hyperplane. As we said above, the SVM is basically used as a linear decision function when the data are separable, however, in this paper; we consider that the data are linearly nonseparable. Therefore, we should introduce a nonlinear function with a nonnegative variables ($\varphi(\xi_i)$) which can map the data in a high-dimensional feature space where they are linearly separable. The optimal hyperplane in a nonlinear space can be determined by the vector W, which minimises the functional:

$$\varphi(W,\xi)=1/2\|W\|^2 + C\sum_{i=1}^{l}\xi_i \tag{5}$$

Where ξ is a slack variable and C a pre-specified value which is used to control the amount of regularization. However, this solution subject to constraints:

$$y_i((W.x_i)-b)\geq 1-\xi_i \quad i=1,2,...,l \tag{6}$$

Using the same formalism with Lagrange multipliers in the linear space can get the optimal hyperplane in a nonlinearly space under some constraints:

$$\begin{aligned} &0\leq \alpha_i \leq C, \quad i=1,2,...l \\ &\sum_{i=1}^{l}\alpha_i y_i = 0 \end{aligned} \tag{7}$$

Then, we have the dual form of the functional:

$$\begin{aligned} &L(W,b,\xi,\alpha) \\ &=\sum_{i=1}^{l}\alpha_i - \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j y_i y_j \varphi(x_i)\varphi(x_j) \\ &=\sum_{i=1}^{l}\alpha_i - \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j y_i y_j K(x_i,x_j) \end{aligned} \tag{8}$$

Where α is the Lagrange multiplier. The Lagrangian has to be minimised with the respect to W, b, x and maximised with the respect to α. The $K(x_i,x_j)=\varphi(x_i)\bullet\varphi(x_j)$ is called kernel function which defines the dot product between two vectors in Z-space.

Furthermore, the main problem for SVM training is the density of the matrix $K(x_i,x_j)=\varphi(x_i)\bullet\varphi(x_j)$, this may lead to a memory problem as long as this matrix is too large to be stored. Since the traditional optimization methods cannot be directly applied to solve this problem, we can apply the decomposition methods for SVM which is considering as the one of the most known methods to train SVM. This method is an iterative procedure which

considers only a small subset of α per iteration, denoted as working set *B*. Thanks to this method, the memory problem is solved. A special decomposition method is the Sequential Minimal Optimization (SMO) which restricts *B* to only two elements.

This work was done using LIBSVM which consider the SMO algorithm as a decomposition method.

It must be pointed out that, a multi-class recognition problem is decoupling to a two-class problem. Therefore, we used the one-against-one approach. This approach consist of constructed k(k - 1)/2 classifiers where each one trains samples from two classes. For the recognition decision making, the majority voting strategy was applied.

The kernel functions are one of the major tricks of SVM. Those functions are used when the samples are linearly nonseparable. Thus, the kernel tricks extends the class of decision functions to the nonlinear case by mapping the samples from the input space X into a high-dimensional feature R without ever having to compute the mapping explicitly, in the hope that the samples will gain meaningful linear structure in R by the function :

$$\varphi : X \rightarrow R \tag{9}$$

The function φ does need to be known, the kernel function K calculate the inner product in the feature space:

$$K(x_i, x_j) = \varphi(x_i) \bullet \varphi(x_j) \tag{10}$$

Furthermore, the kernel function can be interpreted as a measure of similarity between the samples x_i and x_j which it allows SVM classifiers to perform separations even with very complex boundaries.

Moreover, the kernel K must satisfy Mercer's condition in order to be chosen. This theorem, which avoids an explicit formulation of this nonlinear mapping, states that the kernel function K must be continuous, symmetric, and have a positive definite gram matrix. Kernels which satisfy the Mercer's theorem are positive semi-definite, it means that their kernel matrices have no nonnegative Eigen values.

If a kernel does not satisfy these Mercer's conditions, then the Quadratic programming (QP) may have no solution.

There are several possibilities for the choice of this kernel function, including linear, polynomial, sigmoid and RBF. In the sequel of this paper, we will try to find the best choice of the kernels function [3].

RBF (Gaussian) kernels are a family of kernels where a distance measure is smoothed by a radial function (exponential function). This kernel nonlinearly maps samples into a higher dimensional space so it, unlike the linear kernel, can handle the case when the relation between class labels and attributes is nonlinear.

Furthermore, the linear kernel is a special case of RBF since the linear kernel with a penalty parameter C has the same performance as the RBF kernel with some parameters (C, Gamma).

$$K(x_i, x_j) = \exp(-\sigma \left\| x_i - x_j \right\|^2), \sigma > 0 \tag{11}$$

The adjustable parameter σ plays a major role in the performance of the kernel, and should be carefully tuned. If overestimated, the exponential will behave almost linearly and the higher-dimensional projection will start to lose its nonlinear power. On the other hand, if underestimated, the function will lack regularization and the decision boundary will be highly sensitive to noise in training data. Thus, the behavior of SVM depends on the choice of the width parameter σ [3].

3. Feature Extraction Techniques based Mellin Transform

The choice of the feature technique to be used is the one of the first decisions in the speech recognition system. Indeed, the manner in which the basic signal which will be classified is represented, plays a crucial role in the development of an efficient and robust speech recognition system. Through more than three decades of research on the speech, in particular phoneme, recognition task, many feature extraction techniques of the speech signal were proposed and tried in the state-of-the-art. Mel-Frequency Cepstral Coefficients (MFCC) and Perceptual Linear Prediction (PLP) are currently the most known and used feature representations.

Despite this important literature dedicated to the speech recognition area, current phoneme recognition system still incapable of performing efficiently. The main reason of this unsatisfactory performance lies in the characteristics of phoneme data which must be known such as time, frequency. Thus, to guarantee a robust phoneme recognition, the feature extraction technique used must take into account the different characteristics inherent in the phoneme sets.

Moreover, in this work, we combined the mellin transform with the feature extraction techniques based. The key property of mellin transform is the scale invariance which makes the features insensitive to different scale in the signal of phonemes.

In this paper, we propose to use the combined mellin-MFCC features for the recognition of vowels phonemes and the combined mellin-PLP features for the recognition of consonants phonemes.

3.1 The Mellin Transform

The various differences among speakers affect considerably the variability in the conventional features techniques used in phoneme recognition which leads to reduce the rates of phoneme recognition. In this paper, we proposed a new feature extraction methods based on scale transform and conventional feature extraction approach such as MFCC and PLP. The mellin transform is an integral transform introduced first by Robert H. Mellin [4, 5]. This transform is closely similair to the Fourier transform and the Laplace transform. The mellin transform is used in electrical engineering to represent a signal in term of scale.

The Mellin transform is defined as follows:

$$M_f(p) = \int_0^{\infty} f(t) t^{p-1} dt \tag{12}$$

Where p is a mellin parameter. The scale transform is specific restriction of the mellin transform when s = jc-1/2 and it can be defined as follows:

$$D_f(c) = \int_0^{\infty} f(t) t^{jc-\frac{1}{2}} dt \tag{13}$$

As we said above, the most important property of the scale transform is the scale invariant unlike the Fourier transform which is shift invariant. Indeed, the scale-invariance property of scale transform have the capability to improve the performance of phoneme recognition systems. In fact, the scale-invariance property means that the signal differing just by scale transformation have the same transform magnitude distribution. Thus, a scale modification is an expansion or a compression with an energy preservation along the time axis of the original

function. Therefore, the function g which is the scaled version of the original function f have both the same magnitude transform. The scale transform magnitude of g et f is given by:

$$\left|\mathbf{D_g(c)}\right| = \left|\mathbf{D_f(c)}\right| \tag{14}$$

In this study, we applied a fast mellin transform (FMT) since only an efficient and fast discrete implementation of the mellin transform can be used to achieve effective modifications of signals [6]. In other words, the fast mellin transform is realized by using the similarity between the mellin and Fourier transforms. The robustness of the fast mellin transform lies in its capability to compute rapidly the scale magnitude.

3.2 Mellin-MFCC based Features

Mel frequency Cepstral coefficient is the most known feature extraction method for the speech recognition. This method was first proposed by Davis and Mermelstein [7, 8]. The main idea of this algorithm consider that the MFCC are the cepstral coefficients calculated from the mel-frequency warped Fourier transform representation of the log magnitude spectrum. The Delta and the Delta-Delta cepstral coefficients are an estimate of the time derivative of the

MFCCs. Those coefficients have shown a determinant capability to capture the transitional characteristics of the speech signal that can contribute to ameliorate the recognition task. As already told; The proposed idea is to use scale transform and MFCC feature extraction technique which both has the advantages of the scale invariant property of scale transform and the spectral representation of MFCC (see Figure 1).

Figrue 1. Mellin-MFCC based Features

MFCC features are commonly used for the phoneme recognition task and in particular for the vowel recognition [9].

3.3 Mellin-PLP based Features

As we already said, MFCC and PLP feature representations are the most used in the speech recognition task. They have closely similar performance since they both consider the nature of the human auditory system during the features extraction [10-11] and both based on short-time magnitude spectra.

The Perceptual Linear Prediction (PLP) was firstly introduced by Hynek Hermansky [12]. This technique is viewed as a hybrid of DFT and LP (linear predictive) approaches and it is based on the short-term spectrum of speech. The PLP algorithm modifies the short-term spectrum of the speech by several psychophysically based transformations. Later researches [19-20] have shown that the PLP features outperform MFCC in similar conditions, and generally no large difference in performance was observed between them while tested on cleaned phonemes datasets. Several studies have demonstrated the efficiency of applying PLP features for consonants phonemes [10, 13-15]. The third figure shows how the mellin-PLP based features are computed (see Figure 2).

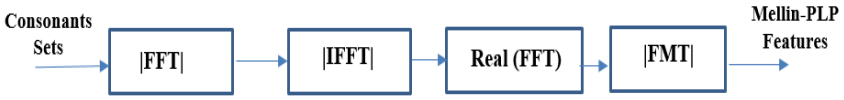

Figure 2. Mellin-PLP based Features

4. Hierarchical Recognition Systems Architecture

The phoneme can be classified into two main classes, vowels and consonants and both of which are further divided into subclasses. This traditional classification is based on the individual and common characteristics of the phoneme such as the place of the articulation, etc. In previous study [14, 10], we find that the conventional phoneme recognition system is vulnerable face to the phonemes confusions using the traditional classification of phonemes. This phenomena can be explain by many factors, such as the variability inter and intra-speaker and the environment. The variability of speech which characterizes phonemes by an inevitable way since it is impossible to speak systematically twice with identical sound manner. For example, if the same speaker produces twice the phoneme / iy / under the same conditions, these productions will be physically different since each pronunciation signal amplitude will be different, the two sounds do not have the same duration, etc. Hence, the variability of phonemes is the main cause of the confusions between phonemes having similar articulatory strcuture which is inherent into our phoneme recognition system.

On the other hand, the TIMIT corpus comprises a huge confusion for the majority of phonemes detected with pronunciation. This fact leads to have for one phoneme a variety of possible pronunciation.

In order to reduce this confusion, we proposed a solution for phoneme recognition based on hierarchical phoneme distribution [16]. The main idea is to organize a new group of phonemes in different classes in order to limit the confusion and isolate phonemes that cause problems to the recognition system (see Figure 3).

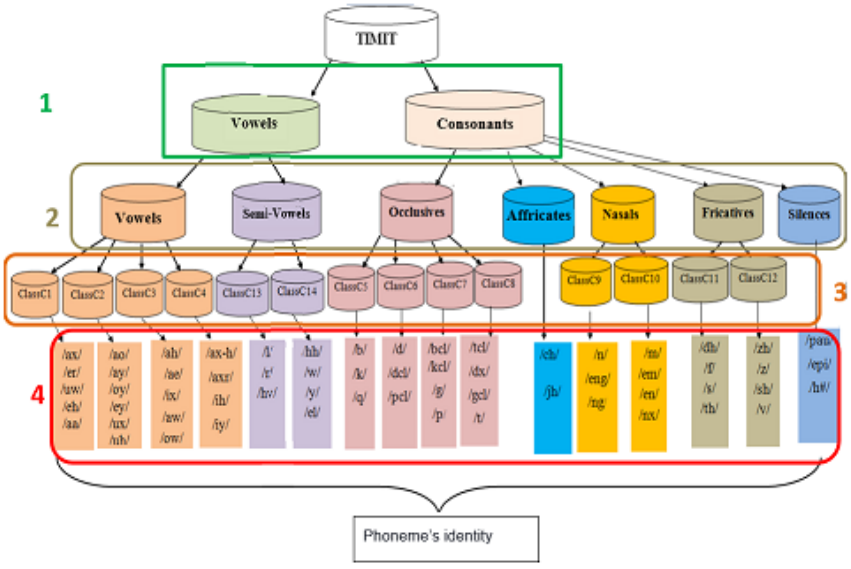

Figure 3. The Architecture of the Hierarchical Phoneme Recognition System with Four Recognitions Levels

It must be point out that the main idea of this hierarchical phoneme recognition system is to use feature extraction methods based Mellin Transfom.

However, for the vowels phonemes we used the Mellin-MFCC features and the Mellin-PLP features to recognize consonants phonemes. Generally, our previous empirical experiments show that recognition rates of consonants within PLP features outperforms slightly the recognition rates within MFCC features.

5. Experimental Conditions

This work was done using SVM classifier [17]. To evaluate the proposed techniques, we used the dialect region DR1 (New England) from TIMIT corpus [18]. Moreover, for the nonlinear SVM approach, we choose the RBF (Gaussian) Kernel trick, this choice was made after a previous study done on our datasets with different kernel tricks (Linear, Polynomial, Sigmoid). There are several ways to carry out a multiclass SVM classification. In the current work we use the "one-against-one" method [17] and the voting strategy. As the classification performance of SVMs is mainly affected by its model parameters particularly the Gaussian width Gamma and the regularization parameter C, we set, for all experiments, gamma as a value within 1/K where K is the number of features and C as value within 10 [14, 19].

Otherwise, for the recognition we used a training datasets to get the SVMs model and so the support vectors. A test datasets was used for classification. Hence, each phoneme was labeled by the number of class to which it belongs.

Moreover, each phoneme has a feature vector which contains 36 coefficients including first delta (Delta) and second delta (Delta-Delta). Indeed, the choice of the feature extractor was made in view of the fact that, those coefficients are the most known and used in pattern recognition researches [14].

5. Results and Ddiscussion

This section includes a comparison of the performance of five hierarchical phonemes recognition systems; The first one is using MFCC features for all the phoneme groups, the second is using PLP features, the third is using Mellin-MFCC features, the fourth is using Mellin-PLP features and the last recognition system is using Mellin-MFCC features to recognize vowels phonemes groups and Mellin-PLP features to recognize consonants phonemes groups. This proposed system is called MMP.

Table 2. (%)Recognition Rates of Five Hierarchical Phonemes Recognition Systems Using: (1) MFFC Features, (2) PLP Features, (3) Mellin-MFCC Features,(4) Mellin-PLP Features and (5) MMP

	MFCC	PLP	Mellin-MFCC	Mellin-PLP	MMP
Level 1	93.50	93.98	95.22	95.56	95.56
Level 2	69.10	70.78	71.13	71.61	**78.09**
Level 3	77.46	78.12	80.44	81.22	**85.31**
Level 4	83.28	85.31	87.90	87.68	**89.30**
Overall rates (%)	80.83	82.04	83.67	84.02	**87.51**

The Table 2 presents comparison of five hierarchical phonemes recognitions systems used different feature extraction methods including our new proposed method based on Mellin transform. Indeed, the results shows the robustness of the system MMP comparing the other

systems in all the levels and thus in the overall phonemes recognition rates within a correct phoneme recognition rate of 87.51%.

In the other hand, we have note that the proposed recognition system MMP permit to enhanced and improved performance the majority of phonemes.

6. Conclusion

In this paper, we proposed a new hierarchical phoneme recognition system using different feature extraction methods based in Mellin transform.

The experimental results demonstrated the capability of this new recognition system to improve the recognition rates. For further work, we propose to improve the robustness of our phoneme recognition system by introducing the Discrete Wavelet Transform (DWT).

References

[1] C. Cortes and V. Vapnik, "Support-vector networks", Machine Learning, Springer, vol. 20.
[2] M. Hofmann, "Support vector machines-kernels and the kernel trick", An elaboration for the Hauptseminar "Reading Club: Support Vector Machines.
[3] R. Amami, D. ben Ayed and N. Ellouze, "Practical Selection of SVM Supervised Parameters with Different Feature Representations for Vowel Recognition", International Journal of Digital Content Technology and its Applications (JDCTA), vol. 7, no. 9, (**2013**), pp. 418-424.
[4] J. M. Fallaha and A. A. Soleamanic, "A new online signature verification system based on combining mellin transform, mfcc and neural network", Digital Signal Processing, vol. 21, no. 2, (**2011**), pp. 404-416.
[5] D. S. A. and D. Rocchesso, "A fast mellin and scale transform", EURASIP J. Appl. Signal Process.
[6] A. D. Sena and D. Rocchesso, "A fast mellin transform with applications in dafx", (**2004**).
[7] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences", IEEE Transactions on Acoustics, Speech and signal Processing, vol. 28, no. 4, (**1980**), pp. 357–366.
[8] M. Sahidullah and S. Goutam, "Design, analysis and experimental evaluation of block based transformation in mfcc computation for speaker recognition", Speech Communication, vol. 54, no. 4, (**2012**), pp. 543-565.
[9] A. D. Sena and D. Rocchesso, "A study on using the mellin transform for vowel recognition", (**2005**).
[10] R. Amami, D. B. Ayed and N. Ellouze, "Practical selection of svm supervised parameters with different feature representations for vowel recognition", International Journal of Digital Content Technology and its Applications, vol. 7, (**2013**), pp. 418–424.
[11] F. Z. Chelali, A. Djeradi and R. Djeradi, "Speaker identification system based on plp coefficients and artificial neural network", The World Congress on Engineering, vol. 2.
[12] H. Hermansky, 'Perceptual linear predictive (plp) analysis of speech", Journal of the Acoustical Society of America, vol. 87, (**1990**), pp. 1738–1752.
[13] C. Er, "Speech recognition by clustering wavelet and plp coefficients", Master of Engineering in Electrical Engineering and Computer science.
[14] R. Amami, D. B. Ayed adn N. Ellouze, "Phoneme recognition using support vector machine and different features representations", 9th International Conference Distributed Computing and Artificial Intelligence, vol. 151, (**2012**), pp. 587–595.
[15] F. Mller, "Invariant features and enhanced speaker normalization for automatic speech recognition", PhD ThesisUniversity of Lbeck.
[16] A. D. Amami and R. N. Ellouze, "Incorporating belief function in svm for phoneme recognition", The 9th International Conference on Hybrid Artificial Intelligence Systems.
[17] C.-C. Chang and C.-J. Lin, "LIBSVM: a Library for Support Vector Machines", Department of Computer Science National Taiwan University, Taipei, Taiwan, (**2011)**.
[18] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue, "TIMIT Acoustic-Phonetic Continuous Speech Corpus ,USA, (**1993**).
[19] R. Jang, "Audio Signal Processing and Recognition", Chapter 12, Book online, Tsing Hua University, Taiwan, (**2009).**

