International Conference on Control, Engineering & Information Technology (CEIT'13)
Proceedings Engineering & Technology - Vol.2, pp.90-94, 2013
Copyright - IPCO

Optimization of Features Parameters for HMM Phoneme Recognition of TIMIT Corpus

Ines BEN FREDJ[1], Kaïs OUNI[2]

Signals and Mechatronic Systems,
Higher School of Technology and Computer Science ESTI,
Carthage University,
45 Rue des Entrepreneurs, Charguia II, 2035, Tunis, Tunisia.
[1]ines_benfredj@yahoo.fr
[2]kais.ouni@esti.rnu.tn

***Abstract* —Phoneme is the smallest contrastive unit in the sound system of a language. Moreover, it has a meaningful role in speech recognition. In this study, we are interesting for phonemes recognition of Timit database using HTK toolkit for HMM. The main goal is to determine the optimal parameters for the recognizer. For this reason, different speech analysis techniques were operated such as Mel Frequency Cepstral Coefficient (MFCC), Linear Predictive Coding (LPC) and Perceptual Linear Prediction (PLP). These techniques were improved by adding temporal derivatives and energy to introduce temporal dynamic of parameters. Results revealed that MFCC and PLP techniques gave a reliable recognition rates using 39 coefficients.**

***Keywords*— Features extraction, HMM, HTK, LPC, MFCC, PLP, TIMIT**

I. INTRODUCTION

An Automatic speech recognition (ASR) system comprises two parts: an encoder and a decoder. Encoder analyses the signal to extract a number of relevant parameters. Decoder uses these parameters to reconstruct a synthetic speech signal [1] [2].

ASR presents two major problems [3]. The first one is the modelling problem. The basic question is “How to represent speech signal to simulate well the production and the perception of the speech by human?” The answer is related to the acoustic modelling. In acoustic modelling, the features of the speech are extracted in terms of vectors [4].

The second problem is decoding problem. The underlying question is “How will the system find the right word in the vocabulary in a good way?” To find the most likely word sequence, the ASR system searches a network of words. The size and the nature of the search space are primarily determined by the language model. Language model requires an artificial intelligence techniques such as fuzzy logic, support vector machines (SVM), hidden markov models (HMM), etc.

Many researchers have focused on different ways in modelling problem [5]. Others were interested in decoding problem to find the most appropriate language model for an effective speech recognition system.

For this purpose, the present study was prepared. We are interested to phoneme recognition of Timit database using HTK toolkit. We used different speech parameterization techniques such as MFCC, LPC and PLP. We evaluated these techniques with different coefficients to obtain better choice for the phoneme features recognized by HMM.

The rest of this paper is organized as follows: In the next two sections, MFCC, LPC and PLP are defined. Then, Hidden Markov Model Toolkit (HTK) is presented. After that, we explain our approach of recognition and we expose experimental results with comments. At last, comes conclusions and future works.

II. SPEECH PARAMETERIZATION TECHNIQUES

A. MFCC

The analysis MFCC consists of the evaluation of Cepstral Coefficients from a frequency distribution according to the Mel scale [9].

The algorithm of MFCC is as follows:

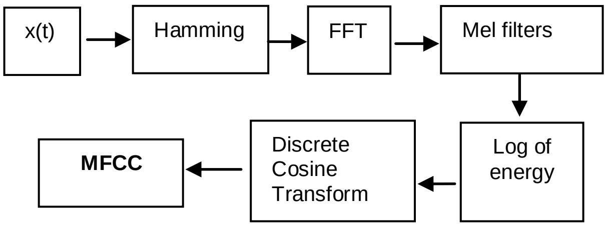

Fig. 1 MFCC algorithm

We take the Fourier transform of a signal windowed by the hamming window. We map the powers of the spectrum obtained above onto the Mel scale. We take the logs of the powers at each of the Mel frequencies [10].

We get the discrete cosine transform of the list of Mel log powers to obtain the MFCC coefficients.

B. LPC

Linear Predictive Coding (LPC) is a useful method for encoding good quality speech at a low bit rate. It provides extremely accurate estimates of speech parameters [8].

LPC algorithm is described in figure 2.

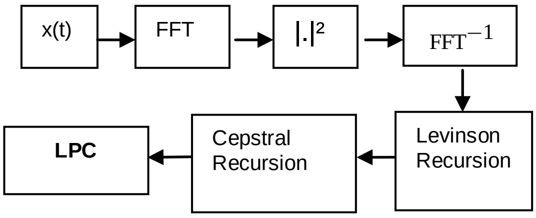

Fig. 2 LPC algorithm

First, Fourier transform of the signal is applied. Then, calculating the inverse Fourier transform of its module squared. Finally, we pass to Levinson and cepstral recursion for getting LPC coefficients.

C. PLP

PLP was studied by Hermansky in 1990 [13]. This technique is based on concepts from the psychophysics of hearing to derive an estimate of the auditory spectrum: the critical-band spectral resolution, the equal-loudness curve and the intensity-loudness power law.

The power spectrum is obtained with a Bark filter bank with a subsequent equal loudness pre-emphasis and a compression based on cube-root.

The auditory spectrum is then approximated by an auto-regressive all-pole model.

PLP algorithm is presented as follows:

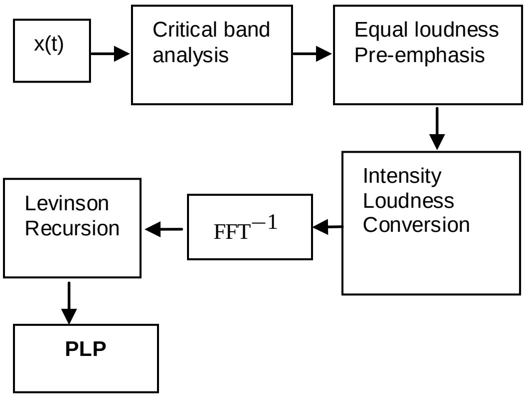

Fig. 3 PLP algorithm

III. Hidden Markov Models: HMM

Hidden Markov Models (introduced in the late 1960s and early 1970s) are currently the most used in speech recognition [14]. HMM are probabilistic models useful for modelling stochastic sequence with underlying finite state structure. Indeed, these models are an intense mathematical structure which explains the remarkable results they give.

An HMM is characterized by the number of states, the functions of observation and the transition probability between states. In fact, the main goal is to determine the probability of a sequence of observations $O = o_1, o_2, ...o_N$ where N is the length of the sequence. An HMM with "n" states $S = s_1, s_2, ..., s_n$ can be presented by a set of parameters λ = { π , A, B} where:

- π represent the initial distribution probability that describes the probability division of the observation symbol in the initial moment noting $\sum_{i=1}^{n} \pi_i = 1$ and $\pi_i >= 0$.
- A is the transition probability matrix { $a_{i,j}$ | i,j=1,2,...,n} where $a_{i,j}$ is the probability of transition from state "i" to state "j" noting $\sum_{j=1}^{n} a_{i,j} = 1$ and $a_{i,j} >= 0$.
- B is the observation matrix { $b_{i,k}$ | i=1,2,...,n , k=1,2,...,m} where $b_{i,k}$ is the probability of observation symbol with index "k" emitted by the current state "i", "m" is the number of observation symbols, $\sum_{k=1}^{m} b_{i,k} = 1$, $b_{i,k} >= 0$ and "n" as noted is the number of states.

A prominent toolkit based on Hmms was used in this work: HTK (Hidden Markov Model Toolkit). It is a portable toolkit for building and manipulating hmms [14].

The first version of HTK was developed by the Cambridge University Engineering Department (CUED) in 1989 [15].

HTK is principally used for speech recognition purpose other than HMMs have a lot of other possible applications.

HTK consists of a set of library modules and tools available in C source form. It is available on free download, beside with a good and complete documentation [15].

HTK offers a refined solutions for the vocal analysis, the training HMM and the test results.

IV. Recognition Approach

A. Database

TIMIT database [16] is used to train and evaluate speaker-independent phoneme recognizers. It consists of 630 speakers from 8 major dialect regions of the United States; each saying 10 sentences which gives 6300 sentences.

Table I describes the structure of Timit corpus.

TABLE I
TIMIT CORPUS

Dialect	Designation	Speakers number	
		Male	Female
DR1	New England	31	18
DR2	Northern	71	31
DR3	North Midland	79	23
DR4	South Midland	69	31
DR5	Southern	62	36
DR6	NewYork City	30	16
DR7	Western	74	26
DR8	Army Brat (moved round)	22	11

All dialects of TIMIT speech corpus sampled in 16 kHz were used [16].

In addition, we have organized the database into six homogenous groups which represent vowels, semivowels, affricates, fricatives, stops and nasals classes as illustrated table II.

TABLE II
DISTIBUTION CLASSES OF TIMIT CORPUS

Class	Label
Affricates	/jh/ /ch/
Fricatives	/s/ /sh/ /z/ /zh/ /f/ /th/ /v/ /dh/
Nasals	/m/ /n/ /ng/ /em/ /en/ /eng/ /nx/
Semi-vowels	/l/ /r/ /w/ /y/ /hh/ /hv/ /el/
Stops	/b/ /d/ /g/ /p/ /t/ /k/ /dx/ /q/ /bcl/ /dcl/ /gcl/ /pcl/ /tcl/ /kcl/
Vowels	/iy/ /ih/ /eh/ /ey/ /ae/ /aa/ /aw/ /ay/ /ah/ /ao/ /oy/ /ow/ /uh/ /uw/ /ux/ /er/ /ax/ /ix/ /axr/ /ax-h/
Others	/pau/ /epi/ /h#/ /1/ /2/

We apply MFCC, LPC and PLP to obtain a database of cepstral parameters. They was extracted from the speech signal with 256 sample frames and was Hamming windowed in segments of 25 ms length every 10 ms with a sampling frequency equal to 16000 KHz. Coefficients number varies from 12 to 39 including first and second derivatives and energy.

B. Training

Training is described as follow:

We start by preparing the dictionary that contains a list of all the phonemes. After that, we label the wav files to mark the beginning and the end of each phoneme and to get a database of labels relative to each sentence.

Then, we extract the coefficients MFCC, LPC and PLP to obtain a database of features and we describe a prototype HMM for each phoneme. A prototype is characterized by the number of states, the functions of observation and the transition probability between states. We have used a prototype of five states defined by the following matrix of probability:

$$A=\begin{pmatrix} \mathbf{0} & \mathbf{0,6} & \mathbf{0,4} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0,6} & \mathbf{0,4} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0,7} & \mathbf{0,3} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \end{pmatrix}$$

Fig. 4 Probability matrix of HMMs states

Each HMM is initialized and trained with the corresponding training set to get a model set [17].

Training is summarized in Fig. 5.

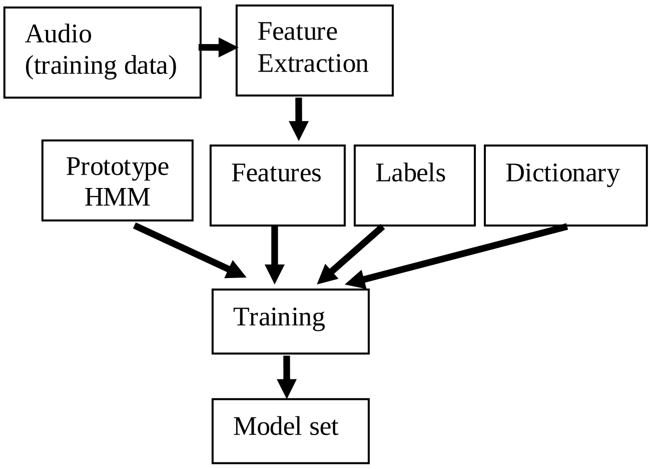

Fig. 5 Training schema

C. Recognition

Recognition is described as follow (Fig. 6):

Before using our phoneme models, we have to define the basic architecture of our recognizer (the task grammar). It's illustrated by a start silence, followed by a single phoneme, followed by an end silence. The task grammar has to be compiled to obtain the task network.

At this stage, our speech recognition task completely defined by its network, its dictionary, and its HMM Model set, is ready for use.

Evaluation and recognition should be done on the test data which should be labelled as for the training data.

An input speech signal is first transformed into a series of acoustical vectors, in the same way as what was done with the training. The input features are then process by a Viterbi algorithm, which matches them against the Markov models recognizer.

The output is stored in a file which contains the transcription of the input.

The performance measures will just result from the comparison between the reference transcription and the recognition hypothesis of each data.

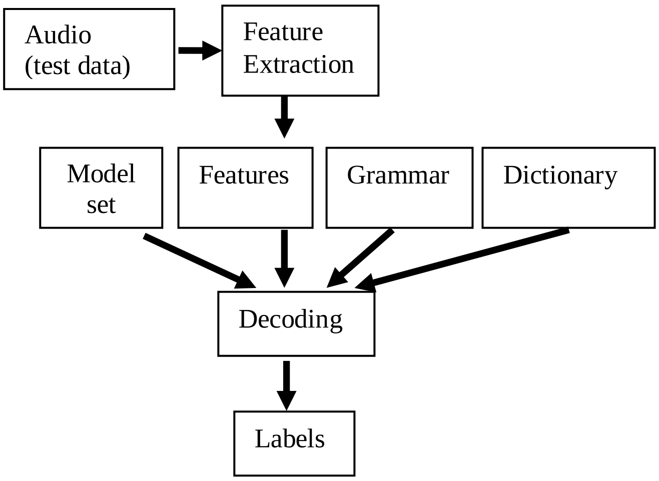

Fig. 6 Recognition schema

V. Experimental Results

Fig.7 and Fig.8 presents classification and recognition rates obtained using MFCC, LPC and PLP coefficients.

We varied the number of these coefficients from 12 to 39 by adding dynamic features: delta (first derivative), delta2 (second derivative) and energy.

We notice that MFCC and PLP gave very similar results using different number of coefficients. As well, LPC coefficients have yielded modest results. Also, we see that increasing number of features affects positively classification and recognition rates. Using 12 coefficients, the recognition rates are 48.43%, 32.85% and 47.76% for MFCC, LPC and PLP respectively. These rates have slightly increased by including first derivative such as 59.39% for MFCC, 35.56% for LPC and 59.45% for PLP. Second derivative was also useful since it have ameliorated the accuracy of the recognizer for MFCC by 2.77%, PLP by 2.66% and LPC by 1.67%.

Performances are better when first and second derivatives and energy are included for all features.

Overall, the recognizer can run well and the best rate that could be achieved is 67.57% for training and 65.96% for recognition. This result is achieved by using 39 coefficients of PLP.

However, we got some low rates for some features and coefficients such as for MFCC, LPC and PLP using 12 coefficients and all the rates obtained with LPC.

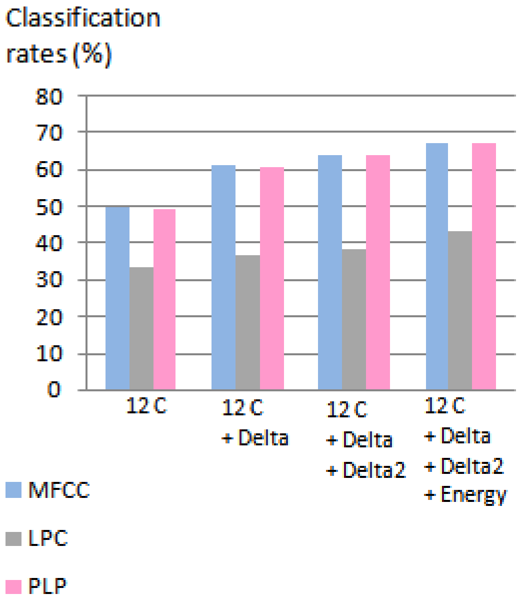

Fig. 7 Classification results

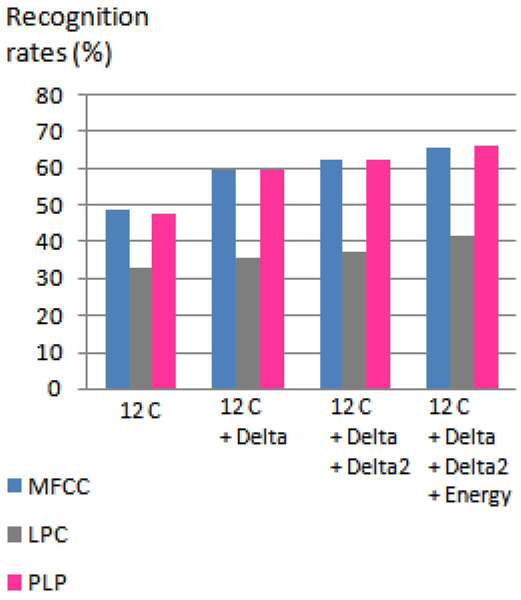

Fig. 8 Recognition results

VI. Conclusions and Future Works

In this paper, we presented an approach of phonemes recognition of Timit database using HTK toolkit.

We evaluated the recognizer with different techniques of features extraction such as MFCC, LPC and PLP.

Number of features varied from 12 to 39 by introducing first and second derivatives and energy to implementing temporal variation.

Results showed the relevance of MFCC and PLP coefficients.

Though, LPC technique was satisfying for the majority of classes.

Shortly, we plan to:

- Study the performance of the recognizer by changing the prototypes of the HMM (number of states, the probabilities of transitions…) or using multi-mixture models,
- Amplify the database and use a techniques of fusion of classes,
- Test HTK in a noisy environment with other database as Aurora database.

REFERENCES

[1] M.A. Anusuya and S.K. Katti, "Front end analysis of speech recognition: a review," *International Journal of Speech Technology*., pp. 99–145, 2011.

[2] B.H. Juang and L.R. Rabiner, "Automatic speech recognition - A brief history of the technology development," *Elsevier Encyclopedia of Language and Linguistics*, 2005.

[3] A.G. Veeravalli, W.D. Pan, R. Adhami and P.G. Cox, "A tutoriel on using hidden markov models for phoneme recognition," in *Proc. Thirty-Seventh Southeastern Symposium on System Theory (SSST05)*, pp. 154 – 157, 2005.

[4] O. Deroo, "Modèles dépendants du contexte et méthodes de fusion de données appliqués à la reconnaissance de la parole par modèles hybrides HMM/MLP," *PhD thesis*, Polytechnic Faculty of Mons, 1998.

[5] R.Gajsek and F. Mihelič, "Comparison of speech parameterization techniques for Slovenian language," in *Proc. 9th International PhD Workshop on Systems and Control*: Young Generation Viewpoint, Slovenia, 2008.

[6] N. Theera-Umpon, S. Chansareewittaya and S. Auephanwiriyakul, "Phoneme and tonal accent recognition for Thai speech," *Expert Systems with Applications*, pp. 13254–13259, 2011.

[7] J. Psutka, L.Müller, and J.V. Psutka, "Comparison of MFCC and PLP Parameterizations in the Speaker Independent Continuous Speech Recognition Task," in *Proc. European Conference on Speech Communication and Technology*, pp. 1813-1816, Scandinavia, 2001.

[8] L. Khoo, Z. Cvetković and P. Sollich, "Robustness of Phoneme Recognition Using Support Vector Machine," in *Proc. International Conference on Acoustics, Speech, and Signal Processing*, France, 2006.

[9] M. Kabache and M. Guerti, "Application des réseaux de neurones à la reconnaissance des phonèmes spécifiques à l'arabe standard," in *Proc. Sciences of Electronics, Technologies of Information and Telecommunications*, Tunisia, 2005.

[10] B.T. Meyer and B. Kollmeier, "Complementarity of MFCC, PLP and Gabor features in the presence of speech-intrinsic variabilities," in *Proc. Interspeech*, Brighton, 2009.

[11] Z, Hachkar, B. Mounir, A. Farchi and J. El Abbadi, "Comparison of MFCC and PLP Parameterization in pattern recognition of Arabic Alphabet Speech," *Canadian Journal on Artificial Intelligence, Machine Learning & Pattern Recognition*, 2011.

[12] Thiang and S. Wijoyo, "Speech Recognition Using Linear Predictive Coding and Artificial Neural Network for Controlling Movement of Mobile Robot," in *Proc. International Conference on Information and Electronics Engineering*, Singapore, pp. 179-183, 2011.

[13] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *The Journal of the Acoustical Society of America*, pp. 1738-1752, 1990.

[14] J. Picone, "Fundamentals of speech recognition," *Institute for Signal and Information Processing*, Mississippi State University, 1996.

[15] S.J. Young, G. Evermann, D. Kershaw, G. Moore, J. Odel, D. Ollason, D.Povey, V. Valtchev and P. Woodland, "*The HTK Book (for HTK Version 3.2)* ," Cambridge University, 2002.

[16] The Linguistic Data Consortium website. [Online]. Available: http://www.ldc.upenn.edu/Catalog/readme_files/timit.readme.html

[17] E.Gouws, K. Wolvaardt, N. Kleynhans and E. Barnard, "Appropriate baseline values for HMM-based speech recognition," in *Proc. PRASA*, pp. 169–172, 2004.

[18] M.Jamaati, H. Marvi and M. Lankarany, "Vowels recognition using mellin transform and PLP-based feature extraction," Proc. *Acoustics*, Paris, 2008.