Effects of dynamic derivatives of speech signals on fuzzy phoneme recognition

I. Ben Fredj, K. Ouni

Research Unit Signals and Mechatronic Systems,

Higher School of Technology and Computer Science ESTI,

Carthage University, Tunis, Tunisia

ines_benfredj@yahoo.fr , kais.ouni@esti.rnu.tn

Abstract

For many years, fuzzy logic methods have become mostly applied in different areas such as speech recognition. In this work, a fuzzy phoneme recognizer is presented and evaluated with different speech parameterization techniques such as LPC, MFCC and PLP. The recognizer consists on the extraction of a fuzzy reference vectors for classification and recognition. Performances are studied by improving the input data and varying features coefficients by adding temporal dynamic derivatives of coefficients.
The experimental results show that temporal information of different features can provide a relevant issue for the task of recognition such as significant rates attained 97, 31 %, 61, 91 % and 73, 04 % for respectively 36 MFCC, 39 LPCC and 39 PLP.

Index Terms

Fuzzy Logic, LPCC, MFCC, PLP, Timit.

1 Introduction

Automatic speech recognition (ASR) systems consider that speech signal is the result of the encoding of some message as a sequence of symbols [1]. The basic purpose is to decode this message and then convert it either into writing or into commands to be processed. Two main problems should be studied: modelling and decoding [2] [3].
Speech signal is transformed into a set of vectors of discrete parameters and then recognized using a language model. Modeling problem consists on to find a number of relevant parameters to represent speech signal. A variety of selection for this task can be useful. Some frequently applied method for speech recognition is linear prediction [4] and mel-cepstrum [5].
Besides, decoding problem is to use a language model for retrieving the appropriate word in the most effective way. Language models are based on several techniques of artificial intelligence as fuzzy logic [6], support vector machines (SVM) [7], hidden Markov models (HMM) [8] [9], etc.
In this work, we aim to evaluate the fuzzy recognizer by introducing different features. This evaluation is illustrated by improving speech parameters using temporal derivatives and energy which have proven a great way to get optimal features for reliable accuracy [10] [11] [12].
In the next section, we provide a summary of the signal parameterization techniques. Then, we present in brief our recognition approach. After that, we discuss experimental results and we finish with conclusions and perspectives.

2 Speech parameterization

We present in the next section, the speech parameterization techniques implemented for the fuzzy recognizer such as the Mel Frequency Cepstral Coding (MFCC), the Linear Predictive Cepstral Coefficients (LPCC) and the Perceptual Linear Prediction (PLP).

2.1 MFCC

MFCC is a technique widely used in speech processing.
It is based on the variation of the critical bands of the human ear with frequency; filters are spaced linearly at low frequencies and logarithmically at high frequencies [13].
These filters are modeled by a non-linear scale based on knowledge of human perception: the Mel scale.
To calculate MFCC coefficients, the Hamming window is applied into the transformation from the time domain to the frequency domain. This transformation is made using the Fourier transform.
A filter is applied subsequently by triangular filters spaced according to the Mel scale.
This scale reproduces the selectivity of the human hearing.
The log is calculated and a discrete cosine transform is applied to return to the time domain.
Figure 1 illustrates the algorithm for calculating the MFCC coefficients.

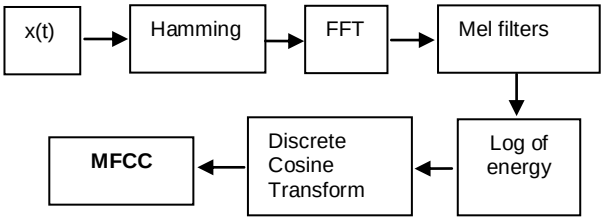

Figure 1 – *MFCC algorithm*

2.2 LPCC

The LPCC can be calculated from the LPC signal [14] analysis by a recursive procedure. In other words, they are converted to LPC cepstrum coefficients.

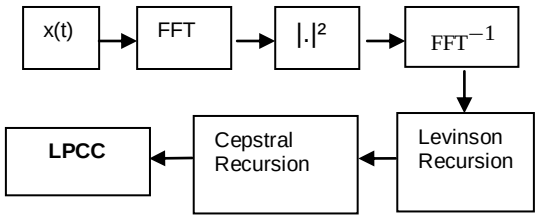

Figure 2 – *LPCC algorithm*

LPC analysis is shown in [15]. 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.

2.3 PLP

The method of the PLP, introduced in 1991 by Hermansky [16] tries to simulate the human auditory system by introducing mechanisms psycho acoustics of the human hearing.

It is based on the LPC analysis; indeed, PLP model the auditory spectrum by an all-pole model of order reduced by using the auto-correlation of the linear prediction technique.

PLP consists on a filter into critical bands of the signal spectrum in the short term, followed by an adjustment of the intensity.

The amplitude of the signal is then compressed and finally the linear predictive analysis is implicated.

This last step is actually a spectral compression technique that modifies the spectrum (set of frequencies constituting a signal) of short-term power before its approximation by an autoregressive model.

PLP algorithm is as shown by figure 3.

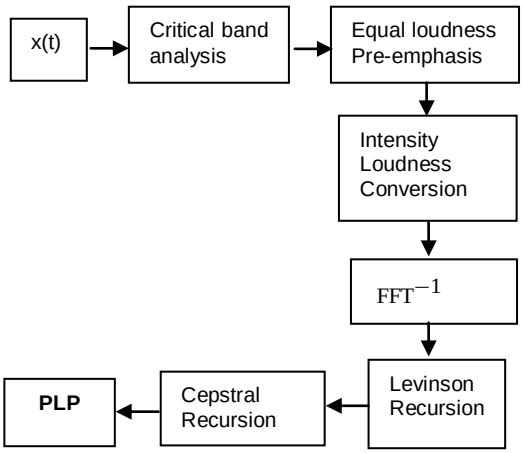

Figure 3 – *PLP algorithm*

2.4 Dynamic derivatives, energy

Static parameters which represent the speech signal are sometimes insufficient for the task of recognition; for this reason, we use always to extract the dynamic parameters. These parameters are the first and second derivatives (Δ, $\Delta\,\Delta$) of the cepstral coefficients of the statics features (see figure 4) [17].

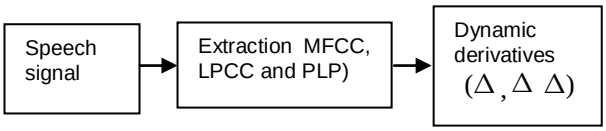

Figure 4 – *Extraction of dynamic features*

These parameters play a meaningful role in human perception as they represent temporal variations in the spectrum of the signal [18].

These coefficients showed an advantageous ability to capture the transient characteristics of the speech signal that can help to improve the recognition task [19].

The coefficients Δ (first derivatives) are often estimated using the second order of the Taylor's expansion:

$$c_i'(t) = \frac{c_i(t+1) - c_i(t-1) + 2(c_i(t+2) - c_i(t-2))}{10} \quad (1)$$

Where $\mathbf{c_i(t)}$ is the $\mathbf{i^{th}}$ coefficient for the frame "t"; $\mathbf{c_i'(t)}$ is its derivative.

Also, second derivatives $(\Delta\,\Delta)$ are estimated in the same way from Δ coefficients.

In addition, we can use also another intuitive parameter: the signal energy.

This is the power of signal measured in successive frames in order to illustrate its changes over time. It is a good feature to distinguish difference between phonemes.

The energy of the signal s_n is calculated as follow:

$$E_n = \sum_{n=0}^{N-1} s_n^2 \qquad (2)$$

Generally, the log normalized of the energy is used as the energy coefficient.

3 Recognition Approach

3.1 Fuzzy Logic

Fuzzy logic is an artificial intelligence technique that was introduced by Lotfi Zadeh in 1965 [20] [21] [22].
It is a method of resolving for problems setting and decision-making based on the mathematical theory of fuzzy sets.
This theory is an extension of the classical set theory to identify sets of imprecisely defined.
Fuzzy logic describes the uncertain and imprecise and it is based on the concept of the membership of an element to one or more classes.
Unlike classical logic, according fuzzy logic, each element belong partially or gradually defined sets.
For example, the statement "Today, it is a nice day!", according fuzzy logic, is 100% true if there are no clouds, 80% true if there are a few clouds, 50% true if there are a lots of clouds and 0% true if it rains all day.
To conclude, in fuzzy logic an element belongs to a "fuzzy" set (not strictly to one set).

3.2 Database

TIMIT database [23] is used for classification and recognition.
It is composed by 630 speakers from 8 different dialects of the United States. Each speaker saying 10 sentences which gives 6300 sentences.
Table 1 describes the structure of the corpus Timit.

Table 1 – Timit Corpus

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

We have organized Timit corpus into seven homogenous phonemes classes which represent affricates, fricatives, nasals, semi-vowels, stops, vowels and others as illustrates table 2.

Table 2 – Classes distribution of TIMIT corpus

Class	Label (phoneme)
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 applied MFCC, LPCC and PLP to obtain a database of cepstral parameters.
Each phoneme is represented by a vector of 12 coefficients which characterize three middle windows of a phoneme.
Features were extracted from the speech signal with 256 sample frames and were Hamming windowed in segments of 25 ms length every 10 ms with a sampling frequency equal to 16000 KHz.
In addition, Timit database is divided into training and recognition data.
The number of samples is presented later with results.

3.3 Fuzzy recognition algorithm

The fuzzy recognition algorithm adopted is based on the extraction of reference vectors: minimal, mean and maximal vectors. Each class is characterized by a set a reference vectors [24] [25].
The fuzzy algorithm id used for classification (train data) and recognition (test data).
After extracting MFCC, LPCC and PLP parameters, we obtain for each phoneme a matrix of coefficients.

Noting v_{max} , v_{mean} and v_{min} the maximal, mean and minimal vectors;

The degree of membership $D_{r,c}$ of a vector « V_r » to a class « c » is given by:

$$D_{r,c} = \frac{V_{max\_c} - V_{mean\_c}}{V_r - V_{mean\_c}} \quad \text{if} \quad V_{mean\_c} \le V_r \le V_{max\_c}$$

$$D_{r,c} = \frac{V_r - V_{min\_c}}{V_{mean\_c} - V_r} \quad \text{if} \quad V_{min\_c} \le V_r < V_{mean\_c}$$

$$D_{r,C} = 0 \quad \text{Otherwise}$$

Indeed, we use the norms of vectors to compare the input vector with references vectors of each class.
So, the term " $v_{moy} \le V_r \le v_{max}$ " means that " $norm(v_{moy}) \le norm(V_r) \le norm(v_{max})$ ".
Thus, we obtain for each sample a degree of membership to each class. We choose then the class relative for the highest degree of membership (figure 4).

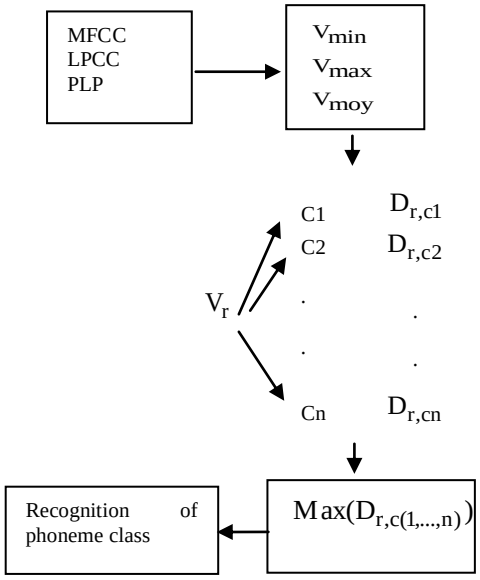

Figure 5 –*Fuzzy recognition algorithm*

To develop parameterization techniques (MFCC, LPCC and PLP), we have used the implementation of Dan Elis described in [26]. In addition, the simulation of the classifier was performed under Matlab environment.
Noting that the recognition of phoneme is its identification into macro classes; for example, to classify the phoneme / aa /, it consist to mention that it is a simple vowel. This is established if the degree of membership to the class of vowels is higher than degrees of membership to other classes. Besides, this phoneme is assigned to the class of the highest degree of membership.

4 Experimental results

Since the main goal is to determine how many coefficients enumerated in the features afford the best recognition accuracy for the fuzzy recognizer; a number of experiments were performed in which dynamic derivatives and energy were introduced such as the number of coefficients varied from 12 to 39.

Tables 3 and 4 show respectively classification and recognition rates.

Table 3 – Classification rates (%)

Features \ Number of coefficients	**12**	**24**	**36**	**39**
MFCC	88.60	91.27	**97.32**	83.36
LPCC	37.60	45.44	26.22	**72.74**
PLP	25.60	62.67	32.73	**78.32**

Table 4 – Recognition rates (%)

Features \ Number of coefficient	**12**	**24**	**36**	**39**
MFCC	84.34	91.79	**97.31**	83.47
LPCC	23.97	53.81	23.84	**61.91**
PLP	17.12	63.22	37.55	**73.04**

Experimental results indicate that selected recognition rates attained 97, 31%, 61, 91% and 73, 04 % using respectively 36 MFCC, 39 LPCC and 39 PLP.
This selection validates that MFCC coefficients were more reliable than LPCC and PLP.
In addition, we observe that first and second derivatives have begun to increase considerably classification and recognition rates for LPCC and PLP only by a combination with the energy coefficient (39 coefficients).
This conclusion confirms that the recognizer can run well using dynamic features and energy; this is also assured for MFCC coefficients (noting that best rates were obtained with 36 features).
On the other hand, we observe a slight difference between classification and recognition rates that is means that both are comparables which indicates a significant flexibility of our recognizer.

5 Conclusion

A phoneme recognition system based on fuzzy logic has been examined in this work. The main goal is to find the optimal speech parameters for this recognizer.
For this reason, we used MFCC, LPCC and PLP features.
We varied also the number of coefficients from 12 to 39 by introducing dynamic derivatives and energy.
Results show that MFCC coefficients were more efficient that LPCC and PLP coefficients since optimal rate attained 97, 31 % by 36 MFCC.
This study shows that our method is a promising approach for the construction of a phoneme recognizer.

Soon, we will focus on further study to optimize results obtained with LPCC and PLP techniques.

References

[1] B. H. Juang, L. R. Rabiner, Automatic Speech Recognition – A Brief History of the Technology Development, *Elsevier Encyclopedia of Language and Linguistics, Second Edition*, 2005.

[2] A. Sadiqui, N. Chenfour, Réalisation d'un système de reconnaissance automatique de la parole arabe base sur CMU sphinx, *Annals. Computer Science Series*, 8(1), pg. 27, 2010.

[3] J.P. Barker, M.P. Cooke and D.P.W. Ellis, Decoding speech in the presence of other sources, *Speech Communication*, pg. 5-25, 2005.

[4] P. P. Vaidyanathan, The theory of linear prediction, Morgan and Claypool Publishers, 2008.

[5] S. Molau, M. Pitz, R. Schl¨uter and H. Ney, Computing Mel-Frequency Cespstral Coefficients on the Power Spectrum. In *International Conference on Computing & Informatics*, pages 1-5, June 2006.

[6] L. Zadeh, Fuzzy logic, neural networks, and soft computing, *ACM*, 37 (3), pg. 77-84, March 1994.

[7] I. Tsochantaridis, T. Hofmann, T. Joachims and Y. Altun, Support vector machine learning for interdependent and structured output spaces. In *the twenty-first international conference on Machine learning ICML*, page 104, 2004.

[8] E.Gouws, K. Wolvaardt, N. Kleynhans and E. Barnard, Appropriate baseline values for HMM-based speech recognition. In *PRASA*, pages 169–172, 2004.

[9] L.R. Rabiner, A tutorial on hidden Markov models and selected applications in speech recognition. In Proceeding of the *IEEE*, 77(2), pages 257–286, 1989.

[10] H.V. Hamme, Handling time-derivative features in a missing data framework for robust automatic speech recognition. In *Proceedings of IEEE ICASSP, May* 2006.

[11] V. Tyagi, I. McCowan, H. Bourlard and H. Misra, On Factorizing spectral dynamics for robust speech recognition, In *Eurospeech*, 2003.

[12] I. Ben Fredj and K. Ouni, Optimization of features parameters for HMM phoneme recognition of TIMIT corpus. In Proceeding of the *International Conference on Control, Engineering & Information Technology (CEIT'13)*, 2013.

[13] C. Goh and K. Leon, Robust Computer Voice Recognition Using Improved. In *International Conference on New Trends in Information and Service Science*, 2009.

[14] C. Octavian, W. Abdulla and S. Zoran, Performance Evaluation of Front-end Processing for Speech Recognition Systems, *Electrical and Computer Engineering Department, School of Engineering*, the University of Auckland, Nouvelle Zélande, 2005.

[15] S. W. Thiang, Speech Recognition Using Linear Predictive Coding and Artificial Neural Network for Controlling Movement of Mobile Robot. In *International Conference on Information and Electronics Engineering*, pages 179-183, 2011.

[16] H. Hermansky, Perceptual linear predictive (PLP) analysis of speech, *Acoustical Society of America*, pg. 1738-1752, 1990.

[17] C. Lévy, Modèles acoustiques compacts pour les systèmes embarqués, *Académie D'aix-Marseille, Université D'Avignon et des pays de Vaucluse*, 2006.

[18] H. Xuedong and others, Improved Hidden Markov Modeling for Speaker-Independent Continuous Speech Recognition. In *Proceedings of a Workshop Held at Hidden Valley*, 1990.

[19] A. Deemagarn and A. Kawtrakul, Thai Connected Digit Speech Recognition Using Hidden Markov Models. In *the 9th Conference Speech and Computer, SPECOM*, 2004.

[20] L. Zadeh, Lotfi, Fuzzy sets, *Information and Control*, 8(3), 1965.

[21] M. Negnevitsky, Artificial Intelligence: A Guide to Intelligent Systems, *School of Electrical Engineering and Computer Science*, University of Tasmania, 2002.

[22] F. Chevrie and F. GUÉLY, La logique floue, *Groupe Schneider*, 1998.

[23] Linguistic Data Consortium, University of Pennsylvania, *http://www.ldc.upenn.edu/Catalog/readme_files/timit.readme.html*

[24] A. Sadiq, R.O.H. Thami, M. Daoudi and J.P. Vandeborre , Classification des Objets 3D Basée sur la Logique Floue. In *Compression et Représentation des Signaux Audiovisuels CORESA*, 2004.

[25] I. Ben Fredj and K.Ouni, Study of speech analysis techniques for the phonemes classification using fuzzy logic. In *Proceedings of the 8th International Multi-Conference on, Signals and Devices (SSD),* pages 1-5, 2011.

[26] D.P.W. Ellis, PLP and RASTA (and MFCC, and inversion) in Matlab using melfcc.m and invmelfcc.m, *http://www.ee.columbia.edu/~dpwe/resources/matlab/rastamat/*