A Comparative Study of Feature Extraction Methods Applied to Continuous Speech Recognition in Romanian Language

Corneliu Octavian DUMITRU [1], Inge GAVAT [1]
[1] Politehnica University Bucharest, Faculty of Electronics Telecommunication and Information Technology, Splaiul Independentei 313, Bucharest, Romania
E-mail: *odumitru@alpha.imag.pub.ro, igavat@alpha.imag.pub.ro*

Abstract - *This paper describes continuous speech recognition experiments on a Romanian language speech database,by using Hidden Markov Models (HMM). We compare the recognition rates obtained in our ASR system realising front-ends based on features extracted by perceptual variants of cepstral analysis and linear prediction and by simple linear prediction. The best results obtained with 36 coefficients mel-frecquency cepstral coefficients (MFCC) are used as basis to rank the front-ends based on LPC. The second rank is very promising for the performance obtained with 5 perceptual linear prediction (PLP) coefficients, obviously better at the last ranked performance of the simple linear prediction coefficients (LPC). We reorganized the database as follows: one database for male speakers, one database for female speakers and one database for both male and female speakers.*

Keywords – *PLP, MFCC, LPC, Hidden Markov Models (HMM), speech recognition*

1. INTRODUCTION

Speech recognition is potentially very useful in many domains, like: telephone communication, dictation, translation and applications for physically handicapped, but also for man-machine spoken dialog in robotics and in information technology.

In this moment, in the field of automatic speech recognition (ASR) there are many challenging aspects, but one of the most important concerns the solution adopted for the classifier of feature vectors sequences in order to decode the corresponding spoken string.

The classifier is a parametric structure and the parameters of this are determined in a process of learning, by examples in the so-called *training phase*. The possible solution for the classifier is based on statistical methods like Hidden Markov Models (HMMs).

The paper is structured as follows: chapter 2 is dedicated to the speech signal analysis in order to realise parameterization by cepstral analysis and perceptual linear prediction. In chapter 3 we present the HMMs. Databases and experimental results are exposed in chapter 4 and 5. Conclusions and references are presented in the end.

2. SPEECH PROCESSING

First step in all recognition tasks is speech analysis, where the speech signal is processed in order to obtain important characteristics, further called features or parameters [1]. By using only the important characteristics of the signal, the amount of data used for comparisons is greatly reduced and thus, less computation and less time is needed for comparisons.

Our feature extraction is based on perceptual linear predictive coding and perceptual cepstral coding, methods that will be presented further.

The block scheme of the processor is shown in Fig. 1. Few blocks are common in both linear prediction and cepstral coding [1], [2]. The first block in the scheme is the frame blocking, used because audio signals are fundamentally a non-stationary signal, so we cut short fragments of the speech signal, which are called frames and the speech is approximated as a quasi-stationary random process during a frame. Then we passed each frame through a Hamming window. We can compute at this time the energy of each frame and we can use the energy set of coefficients in the recognition process for more accuracy.

The classical method in obtaining process of the LPC coefficients requires computation of the autocorrelation coefficients; but according to the Wiener-Hinchin theorem, the autocorrelation function is the original of the power spectrum and therefore an alternative representation for the LPC analysis becomes possible, like is shown in Fig. 1.

In Fig. 1, if the spectral manipulation blocks are missing, it will be performed by cepstral analysis giving the cepstral coefficients (c_p) or linear predictive analysis giving the LPC coefficients; through the conversion block the lasts can be converted in other types of coefficients like the reflection coefficients or the LPC cepstral coefficients. By using the spectral manipulation blocks, we can obtain the perceptive parameters.

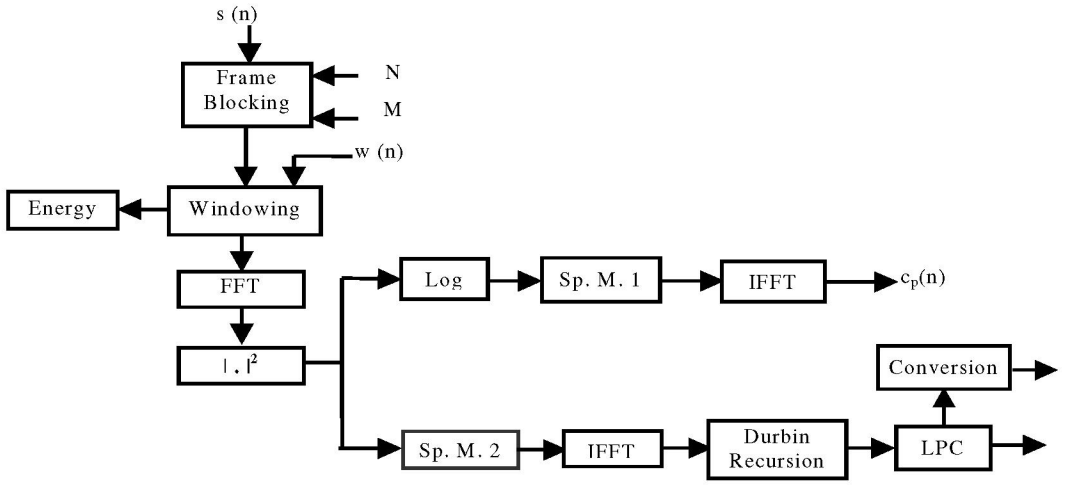

Fig. 1. The block scheme of the processor.

2.1. Cepstral analysis

Cepstral analysis is a very reliable method to speech parameterization and it can be realized applying the blocked and windowed time discrete signal s(n) to the processing chain depicted in Fig. 1.

After the DFT, the modulus of the signal is calculated and the logarithm is taken, the result being proportional in fact to the power spectrum of the speech signal.

Through IDFT, the real cepstrum is obtained, the "filter" characterization being comprised near the cepstrum origin. The re-sampling of the real cepstrum leads to the cepstral coefficients, which alone or in addition with the energy E, and/or the first and second order differences constitute a feature vector successfully applied in speech recognition.

2.2. Spectral manipulation 1 (Sp. M.1)

Spectral manipulation 1 (Sp. M.1) for mel-cepstral coding is represented in Fig. 2.

In order to obtain a parametric representation with the mel-cepstral coefficients and their first and second order variations, the power spectrum is processed by using a set of filters with the transfer functions represented in the Fig. 2. That filter bank is a model for the critical band perception of the human cochlea.

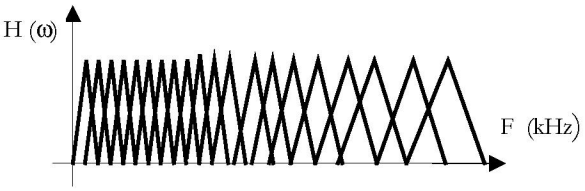

Fig. 2. The transfer functions of filters.

The first and second order variations of the mel-cepstral coefficients are used for speech recorded in noisy environments or under the influence of stress or emotional factors.

2.3. Spectral manipulation 2 (Sp. M.2)

Spectral manipulation 2 (Sp. M. 2) for perceptual linear prediction is represented in Fig. 3

The PLP audio analysis method is more adapted to human hearing, in comparison to the classic Linear Prediction Coding (LPC).

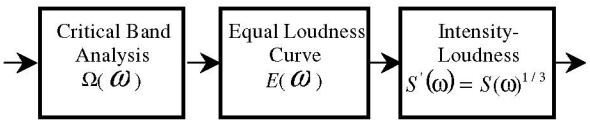

Fig. 3. Block representation for Sp. M. 2.

The power spectrum is computed as follows:

$$P(\omega) = Re(S(\omega))^2 + Im(S(\omega))^2 \qquad (1)$$

The first step is a conversion from frequency to bark, which is a better representation of the human hearing resolution in frequency. The bark frequency corresponding to an audio frequency is:

$$\Omega(\omega) = 6\ln\left(\frac{\omega}{1200\ \pi} + \left(\left(\frac{\omega}{1200\ \pi}\right)^2 + 1\right)^{0.5}\right) \qquad (2)$$

The resulting warped spectrum is convoluted with the power spectrum of the critical band-masking curve, which act like a bank of filters centered on Ω_i .

The spectrum is pre-emphasized by an equal loudness curve, which is an approximation to the non-equal sensitivity of human hearing at different frequencies, at about 40dB level. A filter having the following transfer function gives the curve:

$$E(\omega) = \frac{\left(\omega^2 + 56.8\times10^6\right)\omega^4}{\left(\omega^2 + 6.3\times10^6\right)^2 \times \left(\omega^2 + 0.38\times10^9\right)} \qquad (3)$$

The last operation prior to the all-pole modeling is the cubic-root amplitude compression (Intensity – Loudness Conversion), which simulates the non-linear relation between the intensity of sound and its perceived loudness. Together with the psychophysical equal-loudness pre-emphasis, this operation also reduces the spectral amplitude variation of the critical-band spectrum so that the following all-pole modeling could be done by a relatively low model order [3].

Autoregressive modeling represents the final stage of the PLP analysis, and consists of approximating the spectrum by an all-pole model, by using the autocorrelation method. An Inverse Discrete Fourier Transformation is applied to the spectrum samples, resulting the dual autocorrelation function. For a M-th order all-pole model, only the first M+1 autocorrelation values are needed. The Levinson - Durbin recursive algorithm is used to solve the Yule – Walker equations.

$$\begin{bmatrix} R(1) & R(2) & \dots & R(N) \\ R(2) & R(1) & \dots & R(N-1) \\ \dots & \dots & \dots & \dots \\ R(N) & R(N-1) & \dots & R(1) \end{bmatrix} \times \begin{bmatrix} A(2) \\ A(3) \\ \dots \\ A(N) \end{bmatrix} = \begin{bmatrix} -R(2) \\ -R(3) \\ \dots \\ -R(N+1) \end{bmatrix} \quad (4)$$

where $R(n)$ are the autocorrelation coefficients, and $A(n)$ are the all-pole model coefficients (the predictor), and $A(1)=1$.

The PLP speech analysis method is more adapted to human hearing, in comparison to the classic Linear Prediction Coding (LPC). The main difference between PLP and LPC analysis techniques is that the LP model assumes the all-pole transfer function of the vocal tract with a specified number of reasonances within the analysis band. The LP all-pole model approximates power distribution equally well at all frequencies of the analysis band.

This assumption is inconsistent with human hearing, because beyond 800 Hz, the spectral resolution of hearing decreases with frequency and hearing is also more sensitive in the middle frequency range of the audible spectrum.

3. HIDDEN MARKOV MODELS (HMM)

Our system was developed to recognize continuous speech in four important stages: the data preparation; the training of the models; the testing of the models; the final evaluation.

3.1. The data preparation

The data preparation represents the first stage in the recognizer development. Before it can be used in training, the database must be parameterized, by the described processing procedures [4], [5].

In order to build the HMMs, the set of corresponding speech data files and their associated phonetic transcriptions are required.

Typically the labels used in the original source transcriptions will not be exactly as required, because of the differences in the phone sets. For our experiments it was necessary that labels are context-dependent.

3.2. The training of the models

This is the second step of the system building; it is necessary to define the desired topology for each HMM by writing a prototype definition.

The purpose of the prototype definition is only to specify the overall characteristics and structure of the HMM. The actual parameters (e.g. the transition probabilities and the symbol emission distribution) will be computed later by training.

Initial values for the transition probabilities must be given. An acceptable and simple strategy for choosing these probabilities is to make all of the transitions out of any state equally likely. Then a training process takes place in the following steps:

- Allocate initial values and reset the accumulators for all parameters of all HMMs.
- Get the next training phrase. Construct a composite HMM by joining in sequence the HMMs corresponding to the symbol transcription of the training phrase.
- Calculate the "forward" and "backward" probabilities for the composite HMM.
- Use the "forward" and "backward" probabilities to compute the probabilities of state occupation at each time frame and update the accumulators.
- Repeat the step until all training data are processed.
- Use the accumulators to calculate the new parameter estimations for all of the HMMs.

These steps can then all be repeated as many times as necessary to achieve the required convergence.

3.3. The testing of models

The system provides a single recognition strategy by the use of the "Viterbi" algorithm. The recognition stage takes as input the spoken word sequences, a dictionary defining how each word is pronounced.

The corresponding set of HMMs is allocated.

Recognition can then be performed either on a list of stored speech files, either on a direct input.

3.4. The final evaluation

Once built the HMM-based recognizer, it is necessary to evaluate its performance. Using it to transcribe some pre-recorded test sentences and match the recognizer output with the correct reference transcriptions usually does this. This comparison is performed by the system with uses of the dynamic programming to align the two transcriptions and then count substitution, deletion

and insertion errors.

4. DATABASE

For continuous speech recognition, usually our database is constituted for training by 3300 phrases, uttered by 11 speakers, 7 males and 4 females, each speaker reading 300 phrases, and for testing by 880 phrases uttered by the same speakers, each of them reading 80 phrases. The training database contains over 3200 distinct words, while the testing database contains 1500 distinct words [5].

The data are sampled by 16 kHz, quantified with 16 bits, and recorded in a laboratory environment.
In order to realize our experiments, the database was reorganized as follows: one database for male speakers (MS), one database for female speakers (FS) and one database for both male and female speakers [6].

In all cases we have excluded one MS and one FS from the training and we used them for testing.

5. EXPERIMENTAL RESULTS

The performance is expressed in the word recognition rate (WRR) [5].
The conditions for feature extraction are: perceptive cepstral analysis giving a 36-dimensional vector having as components 12 MFCCs with the corresponding first and second order derivatives, perceptual linear prediction giving a 5-dimensional feature vector having as components five PLP coefficients, and linear prediction, giving a 12-dimensional feature vector having as components the LP coefficients.

The results expressed in WRR obtained in the experiments realized under these conditions are summarized in Table 1, 2 and 3.

The results for WRR are:

- For LPC feature extraction the attained word recognition rates are low: 63,55% training and testing with FS.
- For PLP feature extraction, with 5 coefficients the obtained results are very promising with the PLP: giving word recognition rates about 75,78% training MS and FS and testing MS.
- For MFC feature extraction we obtained best results, as expected, considering that the MFCC are currently standard features in speech recognition: 90,41% training MS and testing with MS.

Table 1. WRR(%): training MS testing MS or FS

Training MS	MFCC_D_A	LPC	PLP
Testing MS	90,41	51,32	72,42
Testing FS	83,21	49,16	63,55

Table 2. WRR(%): training FS testing MS or FS

Training FS	MFCC_D_A	LPC	PLP
Testing MS	78,42	51,32	56,35
Testing FS	89,45	63,55	62,35

Table 3.WRR: training MS and FS testing MS or FS

Training MS and FS	MFCC_D_A	LPC	PLP
Testing MS	88,97	53,24	75,78
Testing FS	85,69	52,28	74,86

6. CONCLUSIONS

Evaluating the efficiency of feature extraction on WRR, one can say that the highest recognition rate was obtained by using cepstral analysis (90,41%), and the lowest recognition rates were obtained for LPC analysis (63,55%). Although in PLP analysis we have only used a very small number of parameters (5), the results that we obtained are promising (75,78%), the recognition rates being situated between the two cases mentioned above.

In the case of PLP coefficients, the best WRR are obtained on the database combined-trained with MS and FS for both cases of tests with MS or FS. In the case of LPC and MFC coefficients, the combined-trained database is not so efficient, the best results being obtained if the training and tests are made on the same type of database.

As concerns our future work, we intent to combine various feature extraction methods in order to improve the efficiency of this basic processes in the recognition framework.

REFERENCES

[1] B. Gold, N. Morgan, *Speech and audio signal processing,* John Wiley and Sons, N.Y., 2002.

[2] S. Furui, *Digital Speech Processing, Synthesis and Recognition*, 2-end, rev and expanded Marcel Dekker, N.Y., 2000.

[3] H. Hermansky, "Perceptual Linear Predictive Analysis of Speech", *J. Acoust. Soc. America*, Vol.87, No.4, pp. 1738-1752, April 1990.

[4] S. Goronzy, *Robust Adaptation to Non-Native Accents in Automatic Speech Recognition*, Springer–Verlag Berlin Heidelberg, Germany, 2002.

[5] C.O. Dumitru, I. Gavat, "Features Extraction, Modeling and Training Strategies in Continuous Speech Recognition for Romanian Language", *Proc. EUROCON 2005*, Serbia & Montenegro, 22-24 November, 1425-1428.

[6] C. Huang, T. Chen, E. Chang, "Speaker Selection Training For Large Vocabulary Continuous Speech Recognition", *Proc. ICLSP* Vol. 1, pp. 609-612, 2002.