Speech Recognition System Based on OLLO French Corpus by Using MFCCs

Braham Chaouche Youcef[1](✉), Yessaad Mohamed Elemine[1],
Benmaiza Islam[2], and Bouttout Farid[3]

[1] LMSE Laboratory, Department of Electronics, University of Mohamed
El Bachir El Ibrahimi, 34265 Bordj Bou Arréridj, Algeria
bcyoucef@gmail.com, yessaad.amine@gmail.com
[2] Laboratory of Spoken Communication and Signal Processing,
Faculty of Electronics and Computer Sciences, USTHB, 16000 Algiers, Algeria
islam.bm34@gmail.com
[3] Laboratory of Signal Processing, Department of Electronics,
University of Constantine, 25000 Constantine, Algeria
f.bouttout@gmail.com

Abstract. The automatic speech recognition is an area of active study since the early 1950s, and the latest technologies in the field of stochastic processes and the discovery of Hidden Markov Models have given a new direction for this area.

This paper describes an approach of speech recognition by using the Mel-Scale Frequency Cepstral Coefficients (MFCC) from speech recognition experiments done on OLLO French corpus by different features. Our work consists in finding the most appropriate choice for this task using the Mel-Scale Frequency Cepstral Coefficients (MFCC) extracted from speech signal.

To evaluate this analysis, we built an ASR reference system based on the modeling of phonemes by the HMM (Hidden Markov Models) associated with the GMM models (Gaussian Mixture Model) using the HTK tool. The implementation of this system was made using several experiments in order to choose the best parameters used in two main steps to build an ASR system, acoustic analysis and decoding. The experiments show that the choice of 25 Gaussian components provides a good compromise between recognition accuracy and computation time, and we found also that the best parameters leading to good recognition accuracy are MFCC_E_D_A coefficients with 92.5%.

In this paper the quality and testing of speaker recognition and gender recognition system is completed and analysed.

Keywords: ASR system · HMM · MFCC · GMM · OLLO · HTK

1 Introduction

Speech is the most natural means of communication between humans. With the development of information technology and the massive use of computer. Man-Machine Dialogue (MMD) using the word as a means of communication has been an increased interest from both the scientific and the industrial community. Automatic speech recognition (ASR), the main component of the MMD system, is a central topic

M. Chadli et al. (eds.), *Recent Advances in Electrical Engineering and Control Applications*,
Lecture Notes in Electrical Engineering 411, DOI 10.1007/978-3-319-48929-2_25

in the broader one of Natural Language Processing (NLP) domain. The general structure of HMM-based speech recognition system [1] consists of two phases: a learning phase whose goal is the construction of acoustic models (HMM models) and recognition phases which the most likely word being imposed. Generally, ASR systems use cepstral parameters called standard parameters as acoustic representation of the speech signal. Cepstral parameters currently the most successful are the MFCCs coefficients (Mel Frequency Cepstral Coefficients). The procedure of calculation of the coefficients is performed on several stages.

The rest of this paper is organized as follows. Section 2 gives a description of Mel-Frequency Cepestral Coefficients (MFCCs). Section 3 introduces a description about OLLO French corpus. The experiments and the results obtained are given in Sect. 4. Concluding remarks are given in Sect. 5.

2 The Mel-Frequency Cepstrum Coefficient (MFCC)

The Mel-frequency Cepstrum Coefficient (MFCC) technique is often used to create the fingerprint of the sound files. The MFCC are based on the known variation of the human ear's critical bandwidth frequencies with filters spaced linearly at low frequencies and logarithmically at high frequencies used to capture the important characteristics of speech. Studies have shown that human perception of the frequency contents of sounds for speech signals does not follow a linear scale [2]. Thus for each tone with an actual frequency, f, measured in Hz, a subjective pitch is measured on a scale called the Mel scale. The following formula is used to compute the Mels for a particular frequency:

$$mel(f) = 2595 \times \log_{10}(1 + \frac{f}{700}) \quad (1)$$

A block diagram of the MFCC processes is shown in Fig. 1. The windowing block minimizes the discontinuities of the signal by narrowing the beginning and end of each frame to zero. The following step consists of applying DFT in order to convert each frame from the time domain to the frequency domain. Then, the signal is passed through the Mel filter bank spectrum to mimic human hearing. In the final step, the Cepstrum, the Mel-spectrum scale is converted back to standard frequency scale. This spectrum provides a good representation of the spectral properties of the signal which is key for representing and recognizing characteristics of the speaker.

3 Coprus

The OLLO French part of the database is a corpus spoken by 10 native French individuals who lived in Belgium. There are six men and four women in different ages [3]. Each person speaks 150 utterances, every utterance is spoken in 6 different style. Each utterance is spoken three times per person.

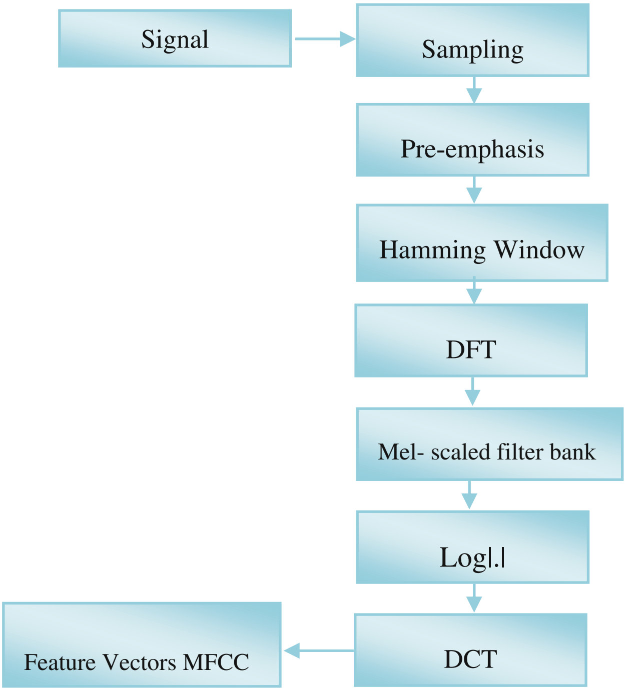

Fig. 1. Block diagram of MFCC

So each person contains 2700 (150*6*3) files. As we are performing dependant speaker recognition tasks, the set of the learning (L1) contains the first of three each logatome's repetitions in the corpus. The remaining two repetitions are contained within the set of test (L2).

The allocation of the corpus is used to build our ASR system. In our experiments, the sampling rate is down to 8 kHz.

To validate our reference system for the phonetic transcription of the speech signal, we took the ACC recognition accuracy as an evaluation criterion, calculated by the formula:

$$Acc = \left(\frac{N - D - S - I}{N}\right) \times 100\% \qquad (2)$$

Where:

H: Number of recognized words;
D: Number of deleted words;
S: Number of substituted words;
I: Number of inserted words;
N: Total number of words.

4 Experimental Results and Analyse

Experiments included comparative evaluations of the recognition results using the mel–cepstral using OLLO databases (Fig. 2).

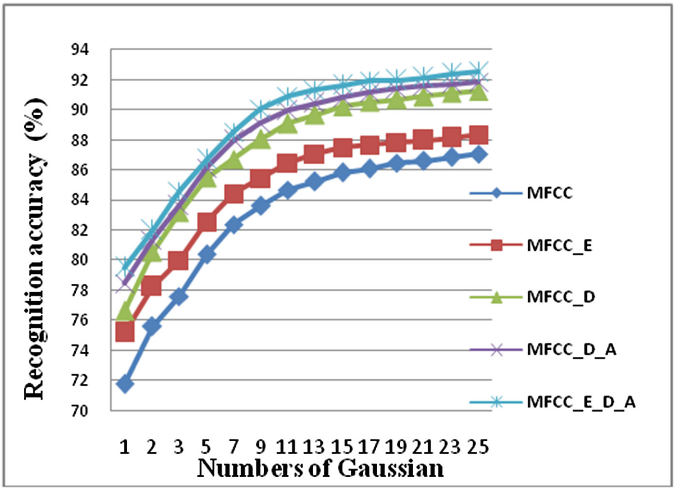

Fig. 2. Comparative accuracy of recognition between static/dynamic coefficients depending on the number of Gaussians.

We used 25 ms speech window with mel–cepstral features [4], due to specific decomposition structure. We also used the same overlapping rate of speech window with the value of 10 ms. Speech feature vector computation included calculation of log energies in the mel–scaled filterbanks [5].

In order to have an effective system of reference, we conducted an experiment to determine the best number of Gaussian mixture components per state and the best type of acoustic parameters MFCC. In this experiment, we set the number of states to three active states [6], then we took the following types of acoustic parameters: MFCC, MFCC_E, MFCC_D, MFCC_D_A and MFCC_E_D_A. In each type of parameters, we varied the number of Gaussian 1 to 25 to determine the best number of Gaussians to have the best accuracy. The basic system uses MFCC coefficients and energy, and the differential coefficients of the parameters.

The logarithm of the energy of the frame is added to the 12 cepstral coefficients to form a vector of 13 coefficients. Differential coefficients of the first and second order, calculated automatically by the tools of HTK [7], are optionally used with the static coefficients.

We tested the contribution of differential coefficients of the first and second order with 13 initial coefficients, working with vectors of dimensions d = 13 (12 MFCC; E), d = 26 (12 MFCC; E; 12 ΔMFCC; ΔE) and d = 39 (12 MFCC; E; 12 ΔMFCC; ΔE; 12 ΔΔMFCC; ΔΔE). The emission probability of each state of the HMM is represented by a linear combination of Gaussian G with diagonal covariance matrix. All other experimental conditions are those of the already described basic system.

Table 1 shows the accuracy for the five sets of experiments and the number of Gaussian probability density.

Table 1. Comparative accuracy of recognition between static/dynamic coefficients depending on the number of Gaussians.

Numbers of Gaussian	Recognition accuracy (%)				
	MFCC(12)	MFCC(13)	MFCC(24)	MFCC(26)	MFCC(39)
1	71.78	75.17	76.64	78.44	79.50
2	75.59	78.23	80.55	81.36	82.00
3	77.62	79.98	83.18	83.63	84.59
5	80.39	82.55	85.48	86.11	86.68
7	82.38	84.41	86.73	87.97	88.52
9	83.61	85.43	88.06	89.10	90.06
11	84.69	86.47	89.09	89.94	90.89
13	85.22	87.06	89.67	90.44	91.31
15	85.82	87.48	90.23	90.83	91.63
17	86.09	87.66	90.52	91.15	91.92
19	86.46	87.84	90.69	91.41	91.96
21	86.58	87.97	90.88	91.58	92.15
23	86.85	88.12	91.11	91.72	92.40
25	87.05	88.34	91.25	91.85	**92.50**

From Table 1, based on the parameterization MFCC_E_D_A typical system gives very good results. The good performance of the system based on MFCC parameters are due to the nature of their products based on human perception models. The results also show that increasing the number of Gaussian enables better modeling of the acoustic space. In the case of a word recognition system, we achieved the best number of Gaussians that allows stability accuracies of recognition (number of Gaussian equal to 25). It is therefore necessary to find a compromise between the recognition accuracy and the number of parameters. We note that beyond the 13th Gaussian, the system performance is not significantly improved (91.31 to 92.50).

Based on these results, we note the following points:

- The combination of MFCC_E_D_A type gives the best recognition accuracy with 92.50 %. However, this combination of 39 parameters requires more resources and computing time. By combining WCC_D_A against type 21 parameters (or 18 parameters in the case of level 5) provides a more compact representation relative to that MFCC_E_D_A or MFCC_D_A (36 parameters), requiring less time and computing resources. Thus, the combination WCC_D_A provides a good compromise in terms of accuracy and computational resources.

5 Conclusion

Our study is to apply the Mel-Scale Frequency Cepstral Coefficients (MFCC) in the acoustic analysis of speech signal to an ASR task.

To accomplish this task and to evaluate the acoustic analysis, we built a reference ASR system based on HMM models. Acoustic analysis of this reference system is based on the extraction of MFCC parameters. This system is built under the platform HTK and evaluated on the basis of OLLO database.

The construction of the reference system calls for the type of acoustic parameters and the number of Gaussian components for each active state HMM. To this end, we conducted various experiments to determine these two parameters. The results showed us that the most relevant acoustic parameters are MFCC_E_D_A coefficients.

The results also showed that the choice of 25 Gaussian components provides a good compromise between recognition accuracy and computation time. In the present work, After a comparative study between the acoustic analysis based on the MFCC parameters, we found that the best parameters leading to good recognition accuracy are MFCC_E_D_A coefficients. However, the representation of the latter requires a dimension (39 parameters) larger than that based on the parameters.

References

1. Haton, J.-P.: Reconnaissance Automatique de la Parole. Paris (2006)
2. Patel, K., Prasad, R.K.: Speech recognition and verification using MFCC & VQ. Int. J. Emerg. Sci. Eng. (IJESE) **1**(7) (2013). ISSN: 2319–6378
3. Huang, L., Zhang, X.: Speaker independent recognition on OLLO French corpus by using different features. In: 2010 First International Conference on Pervasive Computing, Signal Processing and Applications
4. Modic, R.: Comparative wavelet and MFCC speech recognition experiments on the Slovenian and English SpeechDat2
5. Nguyen, Q.C.: Reconnaissance de la Parole en Langue Vietnamienne. thèse de doctorat, Institut national polytechnique de Grenoble, Juin 2002
6. Bakis, R.: Continuous speech recognition via centisecond acoustic states, 91th. Meeting of the Acoust.Soc, avril 1976
7. Young, S., et al.: The HTK Book (for HTK Version 3.4), p. 198 (2006)