International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:11, No:10, 2017

The Capacity of Mel Frequency Cepstral Coefficients for Speech Recognition

Fawaz S. Al-Anzi, Dia AbuZeina

***Abstract*—**Speech recognition is of an important contribution in promoting new technologies in human computer interaction. Today, there is a growing need to employ speech technology in daily life and business activities. However, speech recognition is a challenging task that requires different stages before obtaining the desired output. Among automatic speech recognition (ASR) components is the feature extraction process, which parameterizes the speech signal to produce the corresponding feature vectors. Feature extraction process aims at approximating the linguistic content that is conveyed by the input speech signal. In speech processing field, there are several methods to extract speech features, however, Mel Frequency Cepstral Coefficients (MFCC) is the popular technique. It has been long observed that the MFCC is dominantly used in the well-known recognizers such as the Carnegie Mellon University (CMU) Sphinx and the Markov Model Toolkit (HTK). Hence, this paper focuses on the MFCC method as the standard choice to identify the different speech segments in order to obtain the language phonemes for further training and decoding steps. Due to MFCC good performance, the previous studies show that the MFCC dominates the Arabic ASR research. In this paper, we demonstrate MFCC as well as the intermediate steps that are performed to get these coefficients using the HTK toolkit.

***Keywords*—**Speech recognition, acoustic features, Mel Frequency Cepstral Coefficients.

I. Introduction

ASR is an attractive user-friendly technology to felicitate human computer interface (HCI) in different domains. In the last years, there has been a growing interest to reinforce natural man-machine communication through speech technology. In this regard, much research has been devoted to introduce innovative ideas in the industry for automation purpose (e.g. banking services, cars, control machines, etc.). In general, sound is made out of vibrations of an object to generate a type of energy. The energy causes a movement in the air particles that propagate as audible waves. The air particles movement keeps going until they run out of energy. Humans can hear sound waves with frequencies between about 20 Hz (cycles per second) and 20 kHz. However, the most sensitive limit of human hearing is in the 2000 - 5000 Hz frequency range. In general, machine-learning systems perform feature extraction process at the first place in order to produce the feature values based on the input patterns, these speech features are then pass to an ASR system.

MFCC is the classical front-end analysis in speech recognition to produce the sequence of real-valued numbers that represent feature vectors based on the input signal. Since 1980, it has dominated the ASR feature extraction methods due to its good performance. The success of MFCC makes it the standard choice in the state-of-the-art speech recognizers such as the CMU Sphinx [1], the HTK [2], and the Kaldi speech recognizer [3]. The literature shows that there is a variety of feature extraction methods; however, it is clearly observed that MFCC is extensively used in the most speech classification tasks. An example of another feature extraction method is Perceptual Linear Prediction (PLP) [4]. In fact, previous studies show that MFCC is an appropriate choice to maximize the recognition performance as reported by [5]. It indicates that the MFCC is characterized by better performance and ability of the frequency domain to model adequately the sound. Reference [6] indicated the MFCC and the relative spectral analysis PLP are the most commonly used due to their ability to provide more robust features in adverse conditions. Similarly, Reference [7] demonstrated that the most of today's ASR systems are based on some types of MFCC, which have proven to be effective and robust under various conditions.

The rest of this paper is organized as follows. In the next section, we present some of the challenges of speech features. In Section III, we present the background of MFCC technique followed by the literate review in Section IV. Finally, we conclude in Section V.

II. Speech Features Challenges

Due to the difficulties of handling speech features, it has been long observed that ASR researches employ off-the-shelf toolboxes for features extraction. It is clear that employing MFFC, or even other speech features, for speech applications is not a straightforward task since some of the intermediate functions are difficult for non-specialist researchers. For instance, writing a program for fast Fourier transform (FFT), which is the heart of computing MFCC, requires highly qualified scientists or engineers who have a solid background in complex mathematics, and then, can understand and write FFT program from scratch. No doubt, conducting valuable research that includes speech processing (e.g. speech recognition or speech synthesis) requires deep understanding of signal processing. Speech features pose some challenges in terms of the nature of the data. For instance, textual data or even images features are constant, which remain fixed wherever they appear. To clarify, the features of an article (i.e. the words or the roots are always the same for a particular text; however, speech features are not constant as they are continuously changed according to different aspects such as

Fawaz S. Al-Anzi and Dia AbuZeina are with the Department of Computer Engineering, Kuwait University, Kuwait City, Kuwait (e-mail: fawaz.alanzi@ku.edu.kw, dia.abuzeina@ku.edu.kw).

This work is supported by Kuwait University Research Administration Research Project Number EO06/12.

gender, accent, and age, etc. Simply, it is hard to directly compare speech features due to the (small) differences in vibrations that lead to completely different sounds. The speech-recording environment might have noise such as background music, a second speaker, unwanted breathing, and be affected by the quality of the microphone, or the health and psychological state of person. Reference [8] has a thorough study of the pronunciation variations sources that degrade the performance of ASR systems. In fact, humans can easily interpret signals by extracting relevant information; however, this task is more complex when performed using signal processing and machine learning algorithms. More problems can be observed regarding the speech context. Sounds are quite substantially changed by the surrounding context. The vocal tract goes through different stages getting from 't' to 'a' and getting from 'r' to 'a', and the parameters during the transition will be different as indicated in [9]. Moreover, sounds can last different amounts of time. Deciding where one ends and the next one starts is hard. Moreover, the speech extraction process is a tricky task that requires care and skill. The input waveform is sliced up into frames (usually of 20~30 milliseconds) to generate speech spectrum, which is the distribution of energy as a function of frequency for a particular sound source [10]. Therefore, the waveform is transformed into spectral features (i.e. acoustic feature vectors), as shown in Fig. 1. The figure is obtained from reference [11], which has more details of speech and language processing. For general overview of the difficulty to handle speech recognition, reference [12] elaborates on some of the difficulties with ASR.

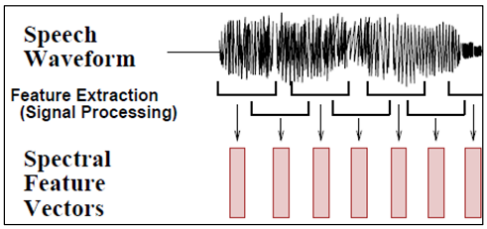

Fig. 1 Extracting features by dividing the signal into frames

III. MFCC Background

To compute the MFCC, the time domain representation of the input speech signal is used to produce the spectral properties, as the patterns are more evident in the frequency domain. The MFCC consists of a set (39 coefficients) that represents the speech signal by dividing it to a set of overlapping short segments called frames. In particular, MFCC coefficients represent the spectral envelope of the speech signal on the Mel-frequency scale. Fig. 2 shows the steps to extract the MFCC of a speech signal. For better performance, the temporal properties might be considered to obtain the first and the second derivative (named respectively ΔMFCC and ΔΔMFCC) of the first order 13 coefficients. We emphasize that the first step, which is sampling and quantization, is performed by the sound card (i.e. a hardware related issue) and is not a part of the MFCC process. However, it is shown in the figure as an indication of the nature of the input data for the Pre-emphasis stage. The goal of the sampling and quantization (also called digitization) step is to convert the analog signal to digital forms for further processing. The sampling rate is the number of samples taken per second, while quantization is the process of representing real-valued numbers as integers. It is worthy to indicate that the MFCC process is not invertible; it is impossible to get the signal back from the set of MFCCs.

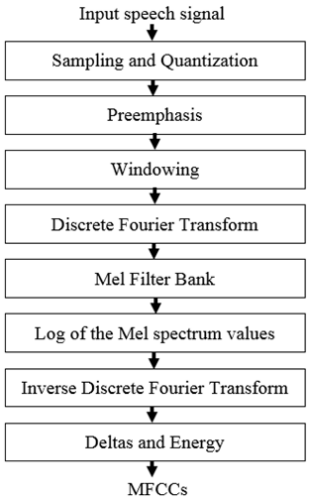

Fig. 2 Extracting features using MFCC algorithm

Reference [13] highlighted some reasons of MFCC popularity in parametric representation of the spectrum as follows. First, the calculation of these parameters leads to a source-filter separation. Second, the parameters have an analytically tractable model. Third, experience proves that these parameters work well in recognition applications. The following is a brief description of the tasks to extract the speech features:

Pre-emphasis: Pre-emphasis is performed after the digitization step. It aims at increasing the amplitude of high frequency bands and decreases the amplitudes of lower bands. That is, this stage is to attain the high frequency formants that carry the relevant information. Without Pre-emphasis, it might be difficult for the receiver to interpret the signal due to the suppression during the sound production mechanism. Hence, the purpose of Pre-emphasis is to apply to the signal with the proper weight sometimes called alpha. The Pre-emphasis is also considered as noise reduction module as it leaves the desired signal untouched, but reduces the noise power considerably.

Windowing: The pre-emphasized speech signal is subjected to the short-time Fourier transform analysis with frame durations of 20-30 ms, frame shifts overlap of around 10 ms. In this stage, the speech signal is analyzed to extract the stationary portion of speech using a window function, which can be characterized by minimizing the discontinuities of the signal.

Discrete Fourier Transform: This stage is the basis of spectral analysis to extract the speech features based on magnitude spectrum computation. It is performed by decomposing an N point time domain signal to obtain the

magnitude frequency response of each frame. That is, it calculates the N frequency spectra corresponding to the N time domain signals

Mel Filter-bank: Computing the Mel frequency spectrum is performed after the discrete Fourier transform by passing the spectrum through Mel filters to obtain Mel spectrum. To produce the filter-bank energies, a number of triangular filters are used that are uniformly spaced on the Mel scale between the lower and upper frequency. It is used to approximate the frequency resolution of the human ear. That is, the Mel scale approximates the sensitivity of the human ear. The Mel frequency scale is linear up to 1000 Hz and logarithmic thereafter.

Log of the Mel spectrum values: Mel filter-bank is used to generate a range of natural logarithm values and replacing the original values by this range. It is an approximation of the spectrum to the Gaussian statistical distribution.

Inverse Discrete Fourier Transform: A set of low order coefficients is compressed in this transform and used to convert spectral information. This is called the Mel cepstrum representation.

Deltas and Energy: The previous step provides the 12 cepstral coefficient for each frame. This step is to add the 13th feature: the energy from the frame. It is useful to identify phone identity.

To explain the output of the MFCC algorithm, we used a small speech file that contains a single Arabic word "*as'hum*" that means "stocks". The speech waveform of this word is shown in Fig. 3. In addition, the spectrogram of this word is shown in Fig. 4. The spectrogram is a visualization tool that is used to understand the information in the signal using time and frequency. Acoustic phones and their properties are better observed in spectrogram. The spectrogram representation of the speech signal is based on short-time Fourier analysis. In the spectrogram, if gray scale is used, the higher the amplitude (the energy), the darker the corresponding region; however, if a color scale is used, the blue represents the low energy, while the red parts represent high energy [14].

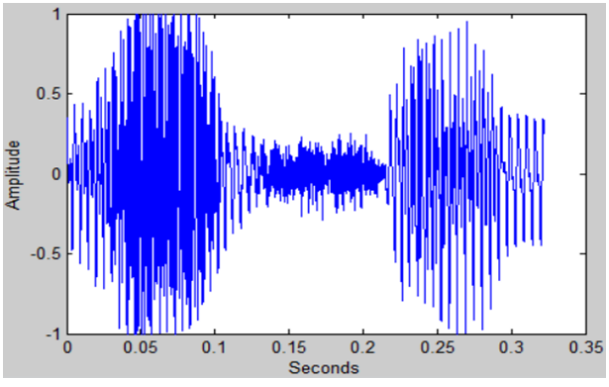

Fig. 3 A speech signal waveform of the Arabic single word "*as'hum*"

The HTK system was used to extract the MFCC speech features of the single word speech file that is represented in Fig. 3. The speech file is of length 0.323 seconds and uses a sampling rate of 16 kHz with 16-bit quantization for each sample. Table I shows the first 12-order of the MFCC coefficients after completing the feature extraction process. Each column represents the 13 features (the 13th feature is the energy from the frame) of a 25 milliseconds frame.

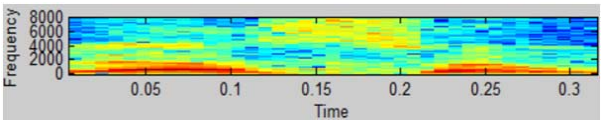

Fig. 4 The spectrogram of a single Arabic word speech file "*as'hum*"

TABLE I
MFCC OF A SINGLE WORD SPEECH FILE

	Frame				
Feat.	1	2	3	4➔29	30
1	-1.81965	-3.15548	-3.76447	…	1.40033
2	-3.10861	-7.92128	-9.68467	…	1.19619
3	1.95010	-2.75036	-4.08556	…	2.44150
4	-12.21996	-14.16043	-15.84370	…	-8.50239
5	-8.21085	-10.14035	-13.00282	…	-11.45462
6	-14.98533	-13.45490	-17.62610	…	-0.93472
7	-22.24395	-23.61402	-13.59459	…	-10.85606
8	2.53291	-1.30409	10.01444	…	-4.10245
9	-8.75291	-14.95345	-3.51132	…	-15.78435
10	-7.62615	-4.33472	-2.60953	…	-13.07826
11	-4.17761	-6.58369	-8.04277	…	3.56489
12	-8.05171	-7.96873	-10.30831	…	-7.61198
13	80.21140	83.59767	86.78981	…	76.75751

TABLE II
TYPICAL HTK SETTINGS – CONFIGURATION FILE

Coding parameters	Comments
SOURCEFORMAT = WAV	The format of the source file
TARGETKIND = MFCC_0	Cepstral C_0 coefficient appended
TARGETRATE = 100000.0	10ms frame rate
SAVECOMPRESSED = T	Save the output file in compressed form
SAVEWITHCRC = T	Attach a checksum to output parameter file
WINDOWSIZE = 250000.0	25ms window
USEHAMMING = T	Use a Hamming window
PREEMCOEF = 0.97	Set pre-emphasis coefficient
NUMCHANS = 26	Number of filter-bank channels
CEPLIFTER = 22	Cepstral filtering coefficient
NUMCEPS = 12	Number of cepstral parameters

The extracted MFCC speech features shown in Table I were extracted using the HTK-HCopy command and the default parameters [2], as shown in Table II. A configuration file (generally called config) is needed which specifies all of the conversion parameters. The HCopy command is used as the following, supposing that the input speech file is "sample.wav":

```
HCopy -C config.txt sample.wav sample.mfc
```

However, the HCopy command creates a binary file (special format non-text file) that contains the MFCC data. Therefore, another option to obtain the MFCC data in textual form is by using HTK-HList command as the following:

```
HList -C config.txt -r sample.wav
```

IV. Literature Review

Based on a thorough review of Arabic speech recognition literature, it is observed that MFCC is extensively used in most studies of Arabic ASR. Table III shows some of the previous studies. However, some of the studies employ other feature extraction methods such as the first work in Table III, in which the LPCC is the shorthand of linear prediction spectrum coefficients, which is one of the famous speech features extraction method. As illustrated, the information in the table belongs to two main categories of speech recognition; isolated and continuous speech recognition. Table III also reveals that Arabic speech recognition is in row stages as most of works depend on off-the-shelf tools (MFCC-based tools), which reduce the opportunities to investigate different speech features as well as reduce the opportunity to present innovative ideas (i.e. featuring new methods).

TABLE III
Previous Studies Employing MFCC

Isolated speech (digits or control command)		
Reference	**Year**	**Features**
[15]	2001	LPCC
[16]	2003	MFCC
[17]	2003	MFCC
[18]	2006	MFCC
[19]	2007	MFCC
[5]	2007	MFCC
[20]	2008	MFCC
[21]	2008	MFCC
[22]	2009	MFCC
Continuous speech		
Reference	**Year**	**Features**
[23]	2007	MFCC
[24]	2008	MFCC
[25]	2010	MFCC
[26]	2011	MFCC
[27]	2011	MFCC
[28]	2012	MFCC
[29]	2012	MFCC
[30]	2017	MFCC

V. Conclusion

This paper demonstrates the MFCC speech features extraction method as one of the most commonly used in ASR systems. Compared to other speech features extraction methods, MFCC is the standard choice for front-end features in state-of-the-art ASR systems. According to our best knowledge and the review that we performed on the previous studies of Arabic ASR, we found that MFCC dominates the works in this field. We employed the HTK system to demonstrate the extraction process of MFCC speech feature vectors of a simple speech file. As a future work, it is worth to continue this work by conducting a practical research to compare MFCC with other methods such as LPCC and PLP.

References

[1] https://cmusphinx.github.io/wiki/faq/ Accessed on 31 March 2017.
[2] Young, Steve, et al. "The HTK book (for HTK version 3.4)." Cambridge university engineering department 2.2 (2006): 2-3.
[3] Povey, Daniel, et al. "The Kaldi speech recognition toolkit."IEEE 2011 workshop on automatic speech recognition and understanding. No. EPFL-CONF-192584. IEEE Signal Processing Society, 2011.
[4] Hermansky, Hynek. "Perceptual linear predictive (PLP) analysis of speech." The Journal of the Acoustical Society of America 87.4 (1990): 1738-1752.
[5] Haraty, Ramzi A., and Omar El Ariss. "CASRA+: a colloquial Arabic speech recognition application." American Journal of Applied Sciences 4.1 (2007): 23-32.
[6] Sharma, Davinder Pal, and Jamin Atkins. "Automatic speech recognition systems: challenges and recent implementation trends." International Journal of Signal and Imaging Systems Engineering 7.4 (2014): 220-234.
[7] Molau, Sirko, et al. "Computing mel-frequency cepstral coefficients on the power spectrum." Acoustics, Speech, and Signal Processing, 2001. Proceedings. (ICASSP'01). 2001 IEEE International Conference on. Vol. 1. IEEE, 2001.
[8] Benzeghiba, Mohamed, et al. "Automatic speech recognition and speech variability: A review." Speech communication 49.10 (2007): 763-786.
[9] Ramsay, Allan. "How Do Speech Recognisers Work?" A presentation. Kuwait Univeristy (2016).
[10] http://www.thefreedictionary.com/ Accessed on 31 March 2017.
[11] Jurafsky, Dan. Speech & language processing. Pearson Education India, 2000.
[12] Forsberg, Markus. "Why is speech recognition difficult." Chalmers University of Technology (2003).
[13] Alcaraz Meseguer, Noelia. Speech analysis for automatic speech recognition. MS thesis. Institutt for elektronikk og telekommunikasjon, 2009.
[14] Huang, Xuedong, et al. Spoken language processing: A guide to theory, algorithm, and system development. Prentice hall PTR, 2001.
[15] Bahi, Halima, and Mokhtar Sellami. "Combination of vector quantization and hidden Markov models for Arabic speech recognition." *Computer Systems and Applications, ACS/IEEE International Conference on. 2001*. IEEE, 2001.
[16] Elmisery, F. A., et al. "A FPGA-based HMM for a discrete Arabic speech recognition system." *Microelectronics, 2003. ICM 2003. Proceedings of the 15th International Conference on*. IEEE, 2003.
[17] Amrouche, Abderrahmane, and J. Michel Rouvaen. "Arabic isolated word recognition using general regression neural network." *Circuits and Systems, 2003 IEEE 46th Midwest Symposium on*. Vol. 2. IEEE, 2003.
[18] Bourouba, H., et al. "New hybrid system (supervised classifier/HMM) for isolated Arabic speech recognition." *Information and Communication Technologies, 2006. ICTTA'06. 2nd*. Vol. 1. IEEE, 2006.
[19] Satori, Hassan, Mostafa Harti, and Nouredine Chenfour. "Introduction to Arabic speech recognition using CMUSphinx system." *arXiv preprint arXiv:0704.2083* (2007).
[20] Essa, E. M., A. S. Tolba, and S. Elmougy. "A comparison of combined classifier architectures for Arabic Speech Recognition." *Computer Engineering & Systems, 2008. ICCES 2008. International Conference on*. IEEE, 2008.
[21] Azmi, M., et al. "Syllable-based automatic arabic speech recognition in noisy-telephone channel." *WSEAS Transactions on Signal Processing* 4.4 (2008): 211-220.
[22] Satori, Hassan, et al. "Investigation arabic speech recognition using CMU sphinx system." *Int. Arab J. Inf. Technol.* 6.2 (2009): 186-190.
[23] Alghamdi, Mansour, Moustafa Elshafei, and Husni Al-Muhtaseb. "Arabic broadcast news transcription system." *International Journal of Speech Technology* 10.4 (2007): 183-195.
[24] Alotaibi, YousefAjami, Sid-Ahmed Selouani, and Douglas O'shaughnessy. "Experiments on automatic recognition of nonnative Arabic speech." *EURASIP Journal on Audio, Speech, and Music Processing* 2008.1 (2008): 679831.
[25] Selouani, Sid Ahmed, and Malika Boudraa. "Algerian Arabic speech database (ALGASD): corpus design and automatic speech recognition application." Arabian Journal for Science and Engineering 35.2C (2010): 158.
[26] Abu Zeina, Dia, et al. "Toward enhanced Arabic speech recognition using part of speech tagging. "*International Journal of Speech Technology* 14.4 (2011): 419-426.
[27] Abu Zeina, Dia, et al. "Cross-word Arabic pronunciation variation modeling for speech recognition. "*International Journal of Speech Technology* 14.3 (2011): 227-236.

[28] Abu Zeina, Dia, et al. "Within-word pronunciation variation modeling for Arabic ASRs: a direct data-driven approach." *International Journal of Speech Technology* 15.2 (2012): 65-75.
[29] Abushariah, Mohammad Abd-Alrahman Mahmoud, et al. "Arabic speaker-independent continuous automatic speech recognition based on a phonetically rich and balanced speech corpus." *International Arab Journal of Information Technology (IAJIT)* 9.1 (2012): 84-93.
[30] Al-Anzi, Fawaz S., and Dia AbuZeina. "The impact of phonological rules on Arabic speech recognition." International Journal of Speech Technology (2017): 1-9.