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...
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.

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.

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
The MFCC algorithm is a speech processing technique that is widely used in automatic speech recognition systems. It is based on the short-time Fourier transform (STFT) of a speech signal. The STFT converts the speech signal into a series of complex exponentials, which are then analyzed to extract features that are invariant to variations in the speech signal. The features are then used to train a classifier for speech recognition.

The MFCC feature extraction process consists of several steps:

1. **Signal preprocessing**: The input speech signal is first preprocessed to remove any unwanted noise or artifacts. This step may include filtering, normalization, and other transformations.
2. **Frame division**: The preprocessed signal is divided into short-time frames of a fixed length, typically 20-30 milliseconds. Each frame contains a segment of the speech signal.
3. **Spectral analysis**: The short-time Fourier transform (STFT) is applied to each frame. The STFT converts the time-domain signal into a frequency-domain representation.
4. **Filter-bank**: The frequency-domain representation is then passed through a set of band-pass filters, known as the Mel-frequency filter bank. These filters are designed to approximate the human auditory system's sensitivity to different frequencies. The output of the filter bank is known as the Mel-frequency spectrum.
5. **Logarithmic compression**: The Mel-frequency spectrum is then compressed using a logarithmic scale. This step is important because the human ear is more sensitive to higher frequencies at low amplitudes and to lower frequencies at high amplitudes.
6. **Deltas and energy**: The logarithmic Mel-frequency spectrum is then used to calculate several higher-order features, known as deltas and energy. These features capture information about the rate of change in the spectrum and the energy distribution.
7. **Feature vector**: The final MFCC feature vector is obtained by concatenating the energy, first-order, second-order, and third-order delta features.

The MFCC feature vector is typically a 12-dimensional vector, although other configurations are also used. The MFCC features are used to represent the speech signal in a way that is robust to variations in the input signal. The features are then used to train a classifier for speech recognition.

The extracted MFCC features can be used in a variety of applications, including automatic speech recognition, speaker identification, and emotion recognition. The MFCC features are also used in combination with other features, such as pitch and formant frequencies, to improve the performance of speech recognition systems.
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>
<thead>
<tr>
<th>TABLE III</th>
<th>PREVIOUS STUDIES EMPLOYING MFCC</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Isolated speech (digits or control command)</strong></td>
<td></td>
</tr>
<tr>
<td>Reference</td>
<td>Year</td>
</tr>
<tr>
<td>----------</td>
<td>------</td>
</tr>
<tr>
<td>[16]</td>
<td>2003</td>
</tr>
<tr>
<td>[17]</td>
<td>2003</td>
</tr>
<tr>
<td>[18]</td>
<td>2006</td>
</tr>
<tr>
<td>[19]</td>
<td>2007</td>
</tr>
<tr>
<td>[20]</td>
<td>2008</td>
</tr>
<tr>
<td>[21]</td>
<td>2008</td>
</tr>
<tr>
<td>[22]</td>
<td>2009</td>
</tr>
</tbody>
</table>

<p>| <strong>Continuous speech</strong> | | |</p>
<table>
<thead>
<tr>
<th>Reference</th>
<th>Year</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>[23]</td>
<td>2007</td>
<td>MFCC</td>
</tr>
<tr>
<td>[25]</td>
<td>2010</td>
<td>MFCC</td>
</tr>
<tr>
<td>[26]</td>
<td>2011</td>
<td>MFCC</td>
</tr>
<tr>
<td>[27]</td>
<td>2011</td>
<td>MFCC</td>
</tr>
<tr>
<td>[28]</td>
<td>2012</td>
<td>MFCC</td>
</tr>
<tr>
<td>[29]</td>
<td>2012</td>
<td>MFCC</td>
</tr>
<tr>
<td>[30]</td>
<td>2017</td>
<td>MFCC</td>
</tr>
</tbody>
</table>

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

