Study of acoustic analysis and modeling for Automatic Standard Arabic Speech Recognition

Rim Walha’, Fadoua Drira’, Haikal El-Abed”, Adel M. Alimi’

* National School of Engineers of Sfax, 3038 Sfax, Tunisia
** Institute for Communications Technologie, 38106 Braunschweig, Germany
E-mails : walha.rim@gmail.com, fadoua.drira@gmail.com, elabed@tu-bs.de, adel.alimi@ieee.org
Phones : +(216)25-25-29-21 ; +(216)21-41-31-36

Abstract:
In this paper, we are interested in the development of an HMM-based Automatic Speech Recognition (ASR) system dedicated to Standard Arabic (SA) language.

The acoustical analysis is one of the most delicate tasks in an ASR system development. In order to guarantee enough informative acoustic parameters describing each audio frame, we made a first study on frame windowing (size and period) and a second study on features extraction methods traditionally used in ASR such as MFCC (Mel-scale Frequency Cepstral Coefficients) and PLP (Perceptual Linear Prediction). This second study allowed us to select the most appropriate coefficients number describing each audio frame.

Modeling acoustic models is a key step in any ASR system. That’s why, we analyzed the effect of varying Gaussian number per HMM state and the number of embedded re-estimations of the Baum-Welch algorithm.

To evaluate the proposed ASR system, a multi-speaker SA connected-digits corpus is collected, transcribed and used throughout all experiments. This paper concludes by evaluating our ASR system on a speaker-independent continue SA speech corpus. The phonemes recognition rate is about 94% which is relatively high when comparing it with another ASR system evaluated on the same corpus.

Keywords: ASR, HMM, acoustical analysis, acoustic modeling, Standard Arabic language.

1. Introduction

The ASR is one of the challenging tasks in the domain of human-machine interaction. It is an information technology that allows computer software to interpret a human natural language.

Despite the important recent advances in ASR field, current systems have not yet the capacity to understand human speech fineness and they are still below human hearing system performance. That’s why ASR remains an active research topic.

In this work, we are interested in the development and the study of an HMM-based ASR system dedicated to Arabic language. In the beginning of this paper, we present a brief review of ASR system and we justify the choice of Arabic language. After that, we describe in section 3 the proposed ASR system. Section 4 presents the corpora used in this study and discusses experimental results. Section 5 concludes and gives some perspectives of this work.

2. Automatic Speech Recognition

2.1. Brief review

Many ASR systems dedicated mainly for European language are currently available on the market such as ViaVoice from IBM and Dragon Naturally Speaking from Nuance. These systems support the recognition of a large vocabulary. However, if new versions of these programs appear very frequently, it is probably due to their current lack of robustness.

An ASR system is generally intended for a given language. Unfortunately, and unlike other languages such as English, French and Japanese, Arabic language still remains very little approached in ASR field. During the past few years, some recent research works on Arabic ASR have been dedicated to single phonemes [Kabache et al., 2005; Selouani et al., 1999], and others to single words [Hazmoun et al., 2009; Alotaibi, 2008], Ejbali et al. [Ejbali et al., 2009] have worked on continue SA speech. However, recognition rates of these systems are still far from the perfection.

2.2. Standard Arabic versus other languages

The ASR applied to Arabic language is a challenging task. Despite it is the fourth most widely spoken language in the world nowadays, Arabic language has limited number of research efforts compared to other languages such as English, French and Japanese. This is mainly related to the language specificities and the limitation of useful tools which make the researchers facing multiple difficulties such as the insufficiency of linguistic resources as well as the very limited number of available transcribed Arabic speech corpora.

The researches on Arabic language are mainly concentrated on SA which is a formal linguistic standard used throughout the Arabic-speaking world, employed in the media, taught in schools, and spoken in the formal framework.

SA language is a Semitic language composed of 34 phonemes, of which 6 are basic vowels and 28 are consonants. Among these consonants, 3 (ل, ع, ط) are either consonants or long vowels according to their appearance context in the word. The Arabic phonetics originality is mainly based on the lengthening relevance in the vocalic system and on the presence of emphatic and geminated
consonants. The allowed syllable structures in Arabic are CV, CVC, and CVCC where V indicates a (long or short) vowel while C indicates a consonant. Arabic utterances can only start with a consonant [Alkhouri, 1990].

2. Proposition of an Arabic ASR system

The proposed ASR system is based on a statistical approach introduced by F. Jelinek [Jelinek, 1976]. It includes five modules: acoustical analysis module, modeling module, transcription module, training module and decoding module. Figure 1 illustrates an overview of the proposed HMM-based system.

![Fig. 1 Overview of the proposed ASR system](image)

Besides their popularity and success, HMMs offer powerful algorithms such as the Baum-Welch algorithm which makes faster and easier the HMMs learning and the Viterbi algorithm which is useful in the research of the best solution. In addition, HMMs are very flexible and have a great ability to process events in time (they cover much of intra-and inter-speaker variation).

3.1. Acoustical analysis module

Acoustical analysis module includes pre-treatments such as recording, digitalization, pre-emphasis, and frame windowing by using Hamming window. It includes also the extraction of features performed to give an observation vector of the acoustic parameters for each frame.

This module is one of the most complex steps in the development of an ASR system. Thus, the acoustic parameters choice conditions the system performances. In order to guarantee enough informative observation vectors, we made experiments related to frame windowing (size and period) and feature extraction methods traditionally used in ASR such as MFCC and PLP. A detailed discussion of these experiments is given later in section 4.

3.2. Modeling module

This module includes both of linguistic and acoustic modeling modules.

For the first modeling process, we used a simple word grammar to describe the sequence of words successfully recognized by the system. This grammar can be depicted through network transitions as it is illustrated in Figure 2.

![Fig. 2 Description of the grammar by network transitions](image)

Concerning the acoustic modeling module, the choice of the speech recognition unit is very important. We used the phoneme as an acoustic unit (34 phonemes allow to describe a standard spoken Arabic). According to their performances and popularity, acoustic units are modeled by continuous-density HMM. To model phoneme, we choose a simple topology ‘left-right’ having three active states authorizing the looping to the current state and the passage to the following state. Indeed, the proposed topology is well-adapted in automatic continuous speech recognition [Lefevre, 2000].

A silence model was also used to model non-speech acoustic artefacts.

3.3. Transcription module

For the transcription module, the speech recognition word vocabulary and the audio corpus are specified in terms of the basic recognition units. The first output of this module is an audio corpus which is orthographically and phonetically transcribed. The second output is a pronunciation dictionary containing phonetic models.

The phonetic transcription is a work of interpretation which requires a scrupulous attention. As an Arabic word may be pronounced by various manners, according to its position in the sentence, its morphological variability, or simply according to the habits of speakers, we can integrate phonetic variants to relax the pronunciation and take into account the speech variations. Thus, every Arabic word could have several phonetic transcriptions in the pronunciation dictionary.

3.4. Training module

The training of acoustic models is realized under HTK toolkit by using embedded training method based on the Baum-Welch algorithm [Young et al., 2009]. Several experiments were designed to evaluate the effect of varying the number of embedded re-estimations of the Baum-Welch algorithm and the effect of varying the number of Gaussian Mixtures.

3.5. Decoding module

Decoding module is also realized under HTK toolkit [Young et al., 2009].
4. Experimental results and discussion

The proposed system performances are evaluated by the phoneme recognition percentage defined by this formula:

\[ \% \text{Phoneme recognition} = \frac{(N - O - S - I)}{N} \times 100 \]

where O, S, I, N are respectively deletions, insertions, substitutions and the total number of speech units of the reference transcription.

4.1. Corpora

We have evaluated the performances of the proposed ASR system on two SA corpora. The first corpus, collected by ourselves, is a multi-speaker connected-digits database. It comprised a small vocabulary of ten digits (from 0 to 9). The training data, spoken by 41 speakers (18 males and 23 females), contains 513 connected-digits utterances. The test data was spoken by 24 speakers (8 males and 9 females) including 17 speakers having participated in the training data construction and 7 speakers not involved in the training data construction. The 105 connected-digits utterances formed the test data. Our corpus was recorded in a normal office environment and sampled at 16 kHz sampling rate. The phonetic transcriptions associated to the audio data were realized on the basis of Arabic phonemes and well checked to be reflected in best the acoustic context.

The second corpus, already defined by R. Ejbali et al. [Ejbali et al., 2009], is a speaker-independent (i.e., speakers used for the training phase are different from those used for the testing phase) continue speech database. It contains the pronunciation of 20 lists. Each list consists of 10 phonetically balanced Arabic sentences. Training data which is about 1 hour and 10 minutes was spoken by 13 speakers (7 males and 6 females). The test data which is about 7 minutes was spoken by 2 speakers (1 male and 1 female) not involved in the training data construction. Each audio file is associated with a transcription text file.

4.2. Parameter estimation

In these first series of experiments, we performed a training of the continuous-density single Gaussian Mixture (GM) phoneme models using the connected-digits corpus collected for SA language. Results were observed by making 10 embedded re-estimations of the Baum-Welch Algorithm used for training phoneme models.

4.2.1. Effect of varying frame windowing

First experiments have been conducted to examine the effect of varying the size and period of frame windowing on phoneme recognition performance (Table I). Each frame was represented by 12 acoustic parameters augmented by the corresponding delta (Δ') and delta-delta (Δ") coefficients. We used, in a first time, MFCC and, in a second time, PLP as an acoustic analysis method. Using either MFCC or PLP, Table I shows that the best phonemes recognition rate is reached by applying 38 ms window size every 18 ms.

4.2.2. Effect of varying the number of acoustic parameters

Second series of tests have been conducted to examine the effect of varying the number of acoustic parameters on phoneme recognition performance. Based on the previous experimental results, frame windowing was characterized by 38 ms window size every 18 ms in this experiment and each frame was represented by acoustic parameters augmented by the corresponding delta and delta-delta coefficients. Table II shows the phoneme recognition rate according to the number of acoustic parameters based on MFCC coefficients in a first time and on PLP coefficients in a second time. Results illustrate that the maximal phoneme recognition rate is reached by using 16 acoustic parameters augmented by the corresponding delta and delta-delta coefficients when we used MFCC or PLP.

4.2.3. Combination of acoustic parameters with energy

In this experiment, we examined the effect of combining the MFCC and the PLP coefficients with the normalized energy (Figure 3). The combination was made by a simple concatenation of acoustic parameters. We analyzed also the effect of varying the number of embedded re-estimations of the Baum-Welch algorithm used for training phoneme models (up to 20 embedded re-estimations).

<table>
<thead>
<tr>
<th>Frame Windowing</th>
<th>MFCC</th>
<th>PLP</th>
</tr>
</thead>
<tbody>
<tr>
<td>20 ms every 10 ms</td>
<td>90.23%</td>
<td>87.83%</td>
</tr>
<tr>
<td>25 ms every 10 ms</td>
<td>91.51%</td>
<td>88.66%</td>
</tr>
<tr>
<td>30 ms every 15 ms</td>
<td>92.61%</td>
<td>93.01%</td>
</tr>
<tr>
<td>34 ms every 16 ms</td>
<td>93.01%</td>
<td>93.53%</td>
</tr>
<tr>
<td>38 ms every 18 ms</td>
<td>93.68%</td>
<td>96.34%</td>
</tr>
<tr>
<td>42 ms every 20 ms</td>
<td>93.32%</td>
<td>96.22%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Acoustic parameters</th>
<th>MFCC</th>
<th>PLP</th>
</tr>
</thead>
<tbody>
<tr>
<td>10 acoustic parameters+ Δ'+ Δ&quot;</td>
<td>90.99%</td>
<td>95.46%</td>
</tr>
<tr>
<td>12 acoustic parameters+ Δ'+ Δ&quot;</td>
<td>93.06%</td>
<td>96.34%</td>
</tr>
<tr>
<td>14 acoustic parameters+ Δ'+ Δ&quot;</td>
<td>93.27%</td>
<td>96.22%</td>
</tr>
<tr>
<td>16 acoustic parameters+ Δ'+ Δ&quot;</td>
<td>96.48%</td>
<td>96.4%</td>
</tr>
<tr>
<td>18 acoustic parameters+ Δ'+ Δ&quot;</td>
<td>95.72%</td>
<td>96.05%</td>
</tr>
<tr>
<td>20 acoustic parameters+ Δ'+ Δ&quot;</td>
<td>94.84%</td>
<td>94.62%</td>
</tr>
</tbody>
</table>

Fig. 3 Combination effect of acoustic parameters with energy on phonemes recognition
Results presented in Figure 3 illustrate a clear superiority of the curve representing the test using 16 PLP coefficients combined with the normalized energy and their corresponding delta and delta-delta coefficients and achieving 96.5% phonemes recognition for the 12 embedded re-estimations of the Baum-Welch Algorithm.

4.2.4. Effect of varying Gaussian number per HMM state

Based on all the previous experiments, the overall recognition performance does not exceed 96.5%. This is due to the fact that the single GM HMMs were not able to provide a good parametric modeling of the acoustic space. Therefore, these second series of experiments examined the effect of varying Gaussian number per HMM state. For each experiment, Gaussian number was split by a factor of 2. Based on the previous acoustic analysis experimental results, we used in these experiments, as the best features representing every audio frame, 16 PLP coefficients and even the normalized energy coefficient augmented by the corresponding delta and delta-delta coefficients.

The tests were performed by training acoustic models using up to 20 embedded re-estimations of the Baum-Welch Algorithm.

![Graph showing variation of phoneme recognition rate according to Gaussian number and embedded re-estimation number](image)

Fig. 4 Variation of phoneme recognition rate according to Gaussian number and embedded re-estimation number

Fig. 4 shows that the phoneme recognition performance increases with Gaussian number per HMM state. We observed that the computational complexity increase exponentially with the Gaussian number. Since the increase in performance from 8 to 32 Gaussians per HMM state did not compensate for the computational complexity, a decision was taken to use only 8 Gaussians. Hence, we concluded that by using 8 Gaussians and 10 embedded re-estimations of the Baum-Welch algorithm, satisfactory performance (98.62% phonemes recognition) can be achieved with reasonable computational complexity.

4.3. Performance comparison

In this section, we evaluated our phoneme-based ASR system on continue SA speech recognition. The test was performed by training acoustic models using the speaker-independent continue SA speech corpus collected by Ejbali et al. and described in section 4.1. As we noted in section 2.1, Ejbali et al. had developed a phoneme-based ASR system which was evaluated on the same corpus. Table III compares the characteristics and the performances of these two systems. It shows that the phoneme recognition rate reached by our system is about 94% which is higher than that obtained with the Ejbali et al. system.

<table>
<thead>
<tr>
<th>ASR system</th>
<th>Acoustic parameters</th>
<th>Number of Gaussian per HMM state</th>
<th>Number of embedded re-estimations</th>
<th>% Phoneme recognition</th>
</tr>
</thead>
<tbody>
<tr>
<td>System of Ejbali et al.</td>
<td>12 PLP+ Energie+ $\Delta^+$ $\Delta''$</td>
<td>64</td>
<td>10</td>
<td>80.36</td>
</tr>
<tr>
<td>Our system</td>
<td>16 PLP+ Energie+ $\Delta^+$ $\Delta''$</td>
<td>8</td>
<td>10</td>
<td>94.02</td>
</tr>
</tbody>
</table>

5. Conclusion and perspectives

The main contribution of this work is the proposition of a HMM-based ASR system suited for the SA language. The performance of this system has been evaluated using a speaker-dependent SA connected-digits corpus and a speaker-independent continue SA speech corpus. A well-established study was conducted to define the best acoustic parameters and models of a performant ASR system for SA language.

In a future work, we intend improving the proposed system by using a large vocabulary linguistic model which will be automatically generated from SA textual corpus. We will furthermore evaluate our system on a large vocabulary SA speech corpus.

6. References