SYLLABLE-BASED AUTOMATIC ARABIC SPEECH RECOGNITION

MOHAMED MOSTAFA AZMI 1, HESHAM TOLBA 2, SHERIF MAHDY 3, MERVAT FASHAL 4
1,2 Elect. Eng. Dept., 3 IT Dept., and 4 Phonetics Dept.
1 Alexandria Higher Institute of Engineering, 2 Faculty of Engineering 3 Faculty of Information Technology and 4 Faculty of Arts.
1, 2, 4 Alexandria University and 3 Cairo University,
Alexandria, EGYPT.

Abstract: - In this paper, we concentrate on the automatic recognition of Egyptian Arabic speech using syllables. Arabic spoken digits were described by showing their constructing phonemes, triphones, syllables and words. Speaker-independent hidden markov models (HMMs)-based speech recognition system was designed using Hidden markov model toolkit (HTK). The database used for both training and testing consists from forty-four Egyptian speakers. Experiments show that the recognition rate using syllables outperformed the rate obtained using monophones, triphones and words by 2.68%, 1.19% and 1.79% respectively. A syllable unit spans a longer time frame, typically three phones, thereby offering a more parsimonious framework for modeling pronunciation variation in spontaneous speech. Moreover, syllable-based recognition has relatively smaller number of used units and runs faster than word-based recognition.

Key-Words: - Speech recognition, syllables, Arabic language, HMMs.

1 Introduction

Automatic Speech Recognition (ASR) is a technology that allows a computer to identify the words that a person speaks into a microphone or telephone. It has a wide area of applications: command recognition (voice user interface with the computer), dictation and interactive voice response. It can be used to learn a foreign language. ASR can help handicapped people to interact with society. It is a technology which makes life easier and very promising [1]. Speech recognition task is split into two parts a front-end and an acoustic unit. A front-end transforms the speech signal into feature vectors containing spectral and/or temporal information using mel-frequency cepstral coefficients (MFCCs). Acoustic unit matches units of features. Units can be words or sub-words, such as phonemes, triphones or syllables. Based on the task (e.g. single digit or continuous speech recognition) the unit size is chosen. Triphones (a phoneme with a left and a right context) also can be used. Word-based recognition was used because the recognition structure is simple but its drawback needs large number of data for training. The recognizer which depends on the phoneme as a phonetic unit is easy to train. Also, it has context sensitive because each unit potentially affected by its predecessors and its followers. However, triphones are a relatively inefficient decompositional unit due to the large number of triphone patterns with a non-zero probability of occurrence, leading to systems that require vast amounts of memory for model storage. Otherwise, Syllables have long unit and they have the least context sensitive[2].The advantage of using syllables as a unit of training is that pronunciation variation is trained right into the acoustic model and does not need to be modeled separately in the

<table>
<thead>
<tr>
<th>Digit</th>
<th>Arabic Writing</th>
<th>Syllables</th>
<th>Representation of syllables</th>
<th>phonemes</th>
<th>Triphones</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>راوي</td>
<td>CV-CV</td>
<td>z-ru</td>
<td>z-t, i-ru</td>
<td>ru</td>
</tr>
<tr>
<td></td>
<td>مصطفى</td>
<td>CV-CV</td>
<td>ss, i-f</td>
<td></td>
<td>r</td>
</tr>
<tr>
<td>1</td>
<td>كتاب</td>
<td>CV-CV</td>
<td>wa-s-X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>حضور</td>
<td>CV-CV-CV</td>
<td>A-1-b-a</td>
<td>A-r, A-t-i-t</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>كلمة</td>
<td>CV-CV-CV</td>
<td>t-a-t-a</td>
<td>t-t,a-t-a</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>ألم</td>
<td>CV-CV-CV</td>
<td>A-a-b-a</td>
<td>A-a-r-t</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>مساحة</td>
<td>CV-CV</td>
<td>x-a-m-s-a</td>
<td>x-ta s-t-a</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>حجم</td>
<td>CV-CV</td>
<td>s-t-b</td>
<td>s-t-i-t-a</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>رمز</td>
<td>CV-CV</td>
<td>s-a</td>
<td>s-t-a</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>مياه</td>
<td>CV-CVC</td>
<td>t-a-n-a</td>
<td>t-1-t-a</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>حصة</td>
<td>CV-CVC</td>
<td>t-s-a</td>
<td>t-t, t-i-t</td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Examples of some Arabic digits presented using syllables, phonemes and triphones.

ISSN: 1790-5117

University of Cambridge, UK, February 20-22, 2008
dictionary. Syllable models also automatically capture co-articulation effects [3].

2 Automatic Recognition of Arabic Speech

Arabic is a Semitic language and it is one of the oldest languages in the world. It is the fifth widely used language nowadays [4].

Although Arabic is currently one of the most widely spoken languages in the world. There has been relatively few speech recognition researches on Arabic compared to other languages. Moreover, most previous works have concentrated on the recognition of formal rather than dialectal Arabic. The first work on Arabic ASR concentrated on developing recognizers for modern standard Arabic (MSA). The most difficult problems in developing highly accurate ASRs for Arabic are the predominance of non diacritic text material, the enormous dialectal variety and the morphological complexity. D. Vergyri et al. investigated the use of morphology-based language model at different stages in a speech recognition system for conversational Arabic [5]. In 2002, K. Kirchhoff et al. investigated novel approaches to automatic vowel restoration, morphology-based language modeling and the integration of out of corpus language model data and got significant word error rate improvements [6]. In 2004, D. Vergyri et al suggested that it is possible to use automatically diacritized training data for acoustic modeling, even if the data has a comparatively high diacritization error rate 23% [7]. In 2006, Markus obtained recognition rate 60.08% using triphone-based recognition of Arabic [8]. In 2007, H. Satori et al. obtained recognition rate 86.66% using Moroccan Arabic digits monophone-based recognition [9].

3 Syllable-Based ASR

Standard Arabic has 34 basic phonemes, of which six are vowels, and 28 are consonants [10]. Arabic has fewer vowels than English. It has three long and three short vowels, while American English consists from at least 12 vowels [11]. Arabic phonemes contain two distinctive classes, which are named pharyngeal and emphatic phonemes. These two classes can be found only in Semitic languages [10-12]. The allowed syllables in Arabic language 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 [10]. All Arabic syllables must contain at least one vowel.

Also Arabic vowels cannot be initials and can occur either between two consonants or final in a word. Arabic syllables can be classified as short or long syllables. The CV type is a short one while all others are long syllables. Syllables can also be classified as open or closed. An open syllable ends with a vowel, while a closed syllable ends with a consonant. In Arabic, a vowel always forms a syllable nucleus and there are as many syllables in a word as vowels in it [13]. Arabic language is a Semitic language that has many differences when compared to European languages such as English. One of these differences is how to pronounce the 11 digits, zero through nine. In Table 1, examples of some Arabic digits using syllables, phonemes and triphones are shown. It is clear from Table 1 that “zero” is repeated two times because it is usually uttered as “zero” or as “sifr”. Except for (sifr), all Arabic digits are polysyllabic words. The motivation behind using syllables comes from recent research on syllable-based recognition [14-15] as well as studies of human perception [16-17] which demonstrate the central role of the syllable played in human perception and generation of speech.

One important factor that supports the use of syllables as the acoustic unit for recognition is the relative insulation of syllable from pronunciation variations arising from addition and deletion of phonemes as well as co-articulation. For example, in 1996, K. Kirchhoff conducted tests on a medium-sized corpus of spontaneous speech (German) in comparison with a triphone-based recognition revealed a superior performance of the syllable-based recognition for the present data set [16]. In 1998, S. L. Wu et al. compared between syllable-based recognition and monophone-based recognition. They discovered that the recognition rate using syllable is higher than phoneme. In 2001, A. Ganapathiraju et al. conducted experiments on large vocabulary continuous English speech recognition; they found that the syllable-based recognition exceeds the recognition of the triphone-based system by 20% [15]. In 2002, Sethy et al. obtained 80% of syllable-based recognition [14]. According to the previous researches, high performance rate of syllable-based recognition is obtained. So, in this paper, we concentrate on the recognition of Egyptian Arabic using syllables to improve the performance of recognition of Arabic speech.
4 Experiments & Results

4.1 Database & Platform
In order to evaluate the performance of syllable-based system, we performed some experiments on different individuals (forty four men) each one of them was asked to utter different Arabic digits. The trained data was created by of twenty-two Egyptian speakers. The tested data was created by twenty-two Egyptian speakers. Speakers were asked to utter different digits as a telephone number. All our experiments were conducted using Egyptian Arabic speech. Four separate recognizers are built corresponding to the different acoustic units of interest i.e. phonemes, triphones, syllables and words.

The recognition platform that we used throughout all our experiments is based on HMMs using HTK. HTK is a portable toolkit for building and manipulating HMMs [18]. HTK is primarily used for research in speech recognition. HTK was used for the back end processing. The data was sampled at 16 kHz. Frame features were extracted to reduce the amount of the information in the input signal. Thirty-nine MFCCs were extracted at frame rate of 10 ms using a 25 ms Hamming window. First-order and second-order differentials were used. Twenty-four channels were calculated and added to the static coefficients to form a vector of 39. Then, the training process starts with HMMs modeling units. First, global means and variances are computed from the audio features. Second, HMMs for each unit initialized with global means and variances are created. Then, forward/backward algorithm is performed. In the testing process, most probable pronunciation to each word in the transcript is assigned.

4.2 Experiments

4.2.1 Monophone-based recognition
The number of phonemes used in our database is twenty-five. Fig.1 (a) shows the effect of increasing the number of states per model on the recognition rate and accuracy of monophone-based recognition. The recognition rate for 3-states, 5-states, 7-states and 9-states were found to be 66.27%, 90.75%, 85.37% and 79.1% respectively. The accuracy rate for 3-states, 5-states, 7-states and 9-states were found to be 59.4%, 78.5%, 75.52% and 72.84% respectively.

<table>
<thead>
<tr>
<th>%H</th>
<th>%D</th>
<th>%S</th>
<th>%I</th>
</tr>
</thead>
<tbody>
<tr>
<td>90.75</td>
<td>3.58</td>
<td>5.67</td>
<td>12.24</td>
</tr>
</tbody>
</table>

4.2.2 Triphone-based recognition
The number of triphones used in our database is sixty-five. Fig.1 (b) shows effect of increasing the number of states per model on the recognition rate and accuracy of triphone-based recognition. The recognition rate for 3-states, 5-states, 7-states and 9-states were found to be 90.75%, 92.24%, 85.37% and 79.1% respectively. The accuracy rate for 3-states, 5-states, 7-states and 9-states were found to be 74.33%, 86.57%, 75.82% and 73.73% respectively.

<table>
<thead>
<tr>
<th>%H</th>
<th>%D</th>
<th>%S</th>
<th>%I</th>
</tr>
</thead>
<tbody>
<tr>
<td>92.24</td>
<td>4.18</td>
<td>3.58</td>
<td>5.67</td>
</tr>
</tbody>
</table>

4.2.3 Syllable-based recognition
The number of syllables used in our database is twenty-two. Fig.1 (c) shows the effect of increasing the number of states per model on the recognition rate and accuracy of syllable-based recognition. The recognition rate for 3-states, 5-states, 7-states, 9-states, 11-states and 13-states were found to be 53.43%, 93.43%, 92.84%, 93.13%, 92.84% and 89.25% respectively. The accuracy rate for 3-states, 5-states, 7-states, 9-states, 11-states and 13-states were found to be 45.67%, 79.1%, 77.61%, 80.3%, 80.9% and 76.42% respectively.

<table>
<thead>
<tr>
<th>%H</th>
<th>%D</th>
<th>%S</th>
<th>%I</th>
</tr>
</thead>
<tbody>
<tr>
<td>93.43</td>
<td>2.09</td>
<td>4.48</td>
<td>14.33</td>
</tr>
</tbody>
</table>

4.2.4 Word-based recognition
The number of words used in this recognizer is thirteen. Fig. 1 (d) shows the effect of increasing the number of states per model on the recognition rate and accuracy of word-based recognition. The recognition rate for 5-states, 7-states, 9-states, 11-states, 13-states and 15-states were found to be 91.64%, 95.22%, 94.93%, 89.85%, 97.01% and 96.42% respectively. The accuracy rate for 5-states, 7-states, 9-states, 11-states, 13-states and 15-states were found to be 78.51%, 78.21%, 97.7%, 74.93%, 85.37% and 86.8% respectively.

<table>
<thead>
<tr>
<th>%H</th>
<th>%D</th>
<th>%S</th>
<th>%I</th>
</tr>
</thead>
<tbody>
<tr>
<td>91.64</td>
<td>4.48</td>
<td>3.88</td>
<td>13.13</td>
</tr>
</tbody>
</table>

Table 2: A comparison between the recognition rates for the performance of our proposed recognizer using the different units.
As shown in Table 2: H represents the number of correct words. D represents number of deleted words. S is the rate of number of substituted words. I is the rate of number of inserted words. Several experiments were done as shown in Fig.1. As shown in Table 2, we can conclude the highest rate of recognition. The selected monophone-based recognition rate is 90.75%. The selected triphone-based recognition rate is 92.24%. The selected syllable-based recognition rate is 93.43%. The selected word-based recognition rate is 91.64% using 5-states of HMM-based but at 13-states of HMM-based, the recognition rate is 97.01%. The syllable-based system is the highest recognition rate using 5-states of HMM-based. Although word-based recognition rate in 13-states is higher than syllable-based recognition rate in 5-states, but syllable-based recognition is preferred because it has relatively smaller number of used units (syllables) and runs faster than word-based recognition. In fact, the performance of the proposed approach could be enhanced by increasing the amount of training data by increasing the number of speakers used to obtain our database.

5 Conclusion & Future Work
Several experiments were conducted on automatic recognition of Egyptian Arabic speech recognition based on HMMs using HTK. These experiments showed that the best recognition performance is obtained when we use syllables to recognize Egyptian Arabic speech compared to the rates obtained for recognition using monophones, triphones and words. Motivated by the obtained results, we are currently preparing our database in order to use syllables to recognize LVCSR. Also, we are studying the effects of wireless channels on the recognition of Arabic speech using syllables.

References:


![Graphs showing the effect of increasing the number of states per model on the recognition rate and accuracy of different recognition methods.](image-url)

Fig. 1: A bar graph showing the effect of increasing the number of states per model on the recognition rate and accuracy of (a) monophone-based recognition, (b) triphone-based recognition, (c) syllable-based recognition, and (d) word-based recognition.