Syllable-based automatic arabic speech recognition

Mohamed Mostafa Azmi[a] and Hesham tolba[b]
[a] *Alexandria Higher Institute of Engg. And Technology, Alexandria University, Alexandria, Egypt*
[b] *Electronic Engg. Dept., Faculty of Engg., Alexandria University, Alexandria, Egypt*
engazm@yahoo.com, htol@link.net

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 is obtained via seventy-five Egyptian speakers. Experiments show that the recognition rate using syllables outperformed the rate obtained using monophones, triphones and words by 7.68%, 1.79% and 12.5% 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.

في هذا البحث سنركز على التعرف الآلى على الكلام العربي المصري باستخدام المقاطع. سوف يتم دراسة تركيب الأرقام المنطوقة العربية من أحادي النغمة و ثلاثي النغمة والمقاطع الصوتية والكلمات الصوتية. وباستخدام نماذج ماركوف المختفية (HMMs) تم تصميم نظام تعرف الكلام من الأرقام المنطوقة باستقلالية . وتتكون قاعدة البيانات من خمس وسبعين متحدث مصري. وتشير التجارب التي أجريت أن معدل تعرف الكلام باستخدام المقاطع الصوتية يفوق معدل تعرف الكلام باستخدام أحادي النغمة و ثلاثي النغمة والكلمات الصوتية بمعدل ٧,٦٨% و ١,٧٩% و ١٢,٥% على الترتيب. والملاحظ أن وحدة المقطع الصوتي من الكلمة تتسع لإطار زمني طويل –تقريبا ثلاثة أضعاف أحادي النغمة - وبذلك تقدم نطاق أكثر حرصاً لعمل نماذج النطق المتغيرة في الكلام التلقائي. علاوة على ذلك، فإن نظام التعرف على الكلام بإستخدام المقاطع الصوتية يملك عدد أقل من الوحدات المستخدمة وبالتالي يعمل أسرع من نظام التعرف على الكلام باستخدام الكلمات الصوتية.

Keywords: 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 a small number of phonemes. But, phonemes are context sensitive because each unit is 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

dictionary. Syllables 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. [5] investigated the use of morphology-based language model at different stages in a speech recognition system for conversational Arabic. In 2002, K. Kirchhoff et al. [6] 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. In 2004, D. Vergyri et al. [7] 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%. In 2006, Markus [8] obtained recognition rate 60.08% using triphone-based recognition of arabic. In 2007, H. Satori et al. [9] obtained recognition rate 86.66% using moroccan arabic digits monophone-based recognition.

3. Syllable-based ASR

Standard arabic has 34 basic phonemes, of which six are vowels, and 28 are consonants [10]. Arabic language has fewer vowels than english language. It has three long and three short vowels, while american english language consists of 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 language. One of these differences is how to pronounce the 10 digits, zero through nine. It is clear from the database that “zero” is repeated two times because usually it is repeated 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 [16] conducted tests on a medium-sized corpus of spontaneous speech (german language) in comparison with a triphone-based recognition revealed a superior performance of the syllable-based recognition for the present data set. 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. [15] 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%. In 2002, Sethy et al.[14] obtained 80% of syllable-based recognition. 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 language using syllables to improve the performance of recognition of arabic speech.

4 Experiments and results

4.1. Database and platform

In order to evaluate the performance of syllable-based recognition, we performed some experiments on different individuals (fifty-seven men and eighteen female). The trained data consists of thirty-eight Egyptian speakers. The tested data consists of thirty-seven Egyptian speakers. Speakers were asked to utter different digits as a telephone number. All our experiments were conducted using Egyptian Arabic speech. The speech data is recorded at 16 KHz using a microphone connected to a laptop PC. 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 through out 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 training of the acoustic models is based on HTK ver 3.3 training tools. We have trained different sets of acoustic models. They use Mel scale cepstral coefficients with their first and second derivatives (MFCC_E_D_A). For example, in syllable-based recognizer, the acoustic modeling used syllables HMMs trained using conventional maximum likelihood estimation. First, initial syllable prototypes are bootstrapped and trained using the speech-training database. HTK tool's compute the global mean and variance (HCOMPV) to flat start bootstrapping and used in this phase. The advantage of using flat start bootstrapping is that it overrides the limitation of having segmented trained data. The strategy used in HCOMPV tool is to make all models equal initially and move directly to embedded training using HTK tool's embedded re-estimation (HEREST). After the flat start initialization we use three embedded training iterations using HEREST for training the syllable models. A silence fix technique is applied where a one state short tee model is created and this single state is tied to the center state of the silence model, so both models have the same silence training data. After doing the silence fix step another several iterations of embedded training cycle are done. HTK tool's HTK tool which implements Viterbi algorithm directly (HVite) is used to recognize the speech-testing database. The recognition rate is obtained by comparing the output of HVite tool and the labeled text of the speech-testing database.

4.2. Experiments

4.2.1. Monophone-based recognition

The number of phonemes used in our database is twenty-five phonemes. Fig. 1-a shows the effect of increasing the number of states per model on the recognition rate of monophone-based recognition. The recognition rate for 3-states, 5-states, 7-states and 9-states were found to be 67.5%, 86.6%, 84.82% and 79.82% respectively.

4.2.2. Triphone-based recognition

The number of triphones used in our database is sixty-five triphones. Fig. 1-b shows effect of increasing the number of states per model on the recognition rate of triphone-based recognition. The recognition rate for 3-states, 5-states, 7-states and 9-states were found to be 80.36%, 92.5%, 87.14% and 79.82% respectively.

4.2.3 Syllable-based recognition

The number of syllables used in our database is twenty-two syllables. Fig. 1-c shows the effect of increasing the number of states per model on the recognition rate of syllable-based recognition. The recognition rate for 3-states, 5-states, 7-states, 9-states and 11-states were found to be 59.64%, 89.11%, 94.29%, 92.5% and 92.32% respectively.

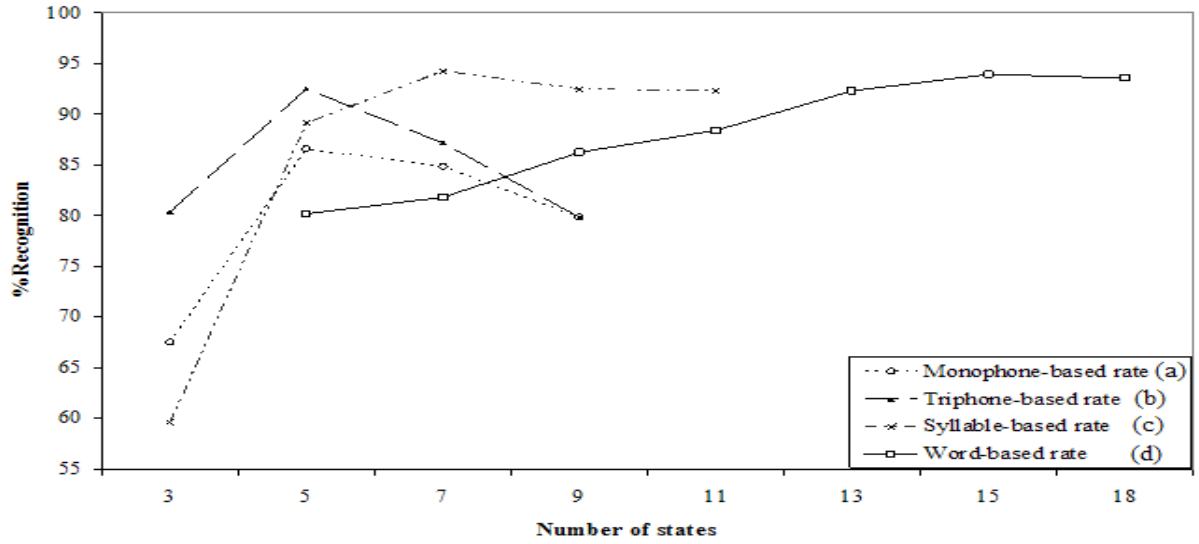

Fig. 1. The main dimensions of the proposed wraparound mode converter.

4.2.4. Word-based recognition

The number of words used in this recognizer is thirteen words. Fig. 1-d shows the effect of increasing the number of states per model on the recognition rate of word-based recognition. The recognition rate for 5-states, 7-states, 9-states, 11-states, 13-states, 15-states and 18-states were found to be 80.18%, 81.79%, 86.25%, 88.39%, 92.32%, 93.93% and 93.57% respectively.

As shown in table 1, H represents the number of correct words. D represents number of deleted words. S is the rate of number of substituted words, and I is the rate of number of inserted words. Also from table 1, we can conclude the highest rate of recognition. The selected monophone-based recognition rate is 86.61%. The selected triphone-based recognition rate is 92.5%. The selected syllable-based recognition rate is 94.29%. The selected word-based recognition rate is 81.79% using 7-states of HMM-based but at 15-states of HMM-based, the recognition rate is 93.94%. The syllable-based system is the highest recognition rate using 7-states of HMM-based. But, monophone-based recognition and triphone-based recognition are used at 5-states of HMM-based. 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.

Table 1
A comparison between the recognition rates for the performance of the proposed recognizer using different units

Unit-based recognition	%H	%D	%S	%I
Monophone	86.61	1.43	11.96	18.21
Triphone	92.5	2.14	5.36	14.29
Syllable	94.29	1.25	4.46	26.61
Word	81.79	10.5	7.68	18.39

5. Conclusions and future work

Several experiments were conducted on automatic 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 Large Vocabulary Continuous Speech Recognition (LVCSR). Also, we are studying the effects of wireless channels on the recognition of Arabic speech using syllables.

References

[1] A. Yousfi, "Introduction De La Vitesse D'élocution et de L'énergie dans un

Modèle de Reconnaissance Automatique De La Parole", Thèse De Doctorat, Faculté Des Sciences Oujda (2002).

[2] L. Rabiner and B.H. Juang, Fundamentals of Speech Recognition. New Jersey, Prentice Hall (1993).

[3] M. Larson, "Sub-Word-Based Language Models for Speech Recognition: Implications for Spoken Document Retrieval", GMD German National Research Center for Information Technology Institute for Media Communication.

[4] M. Al-Zabibi, "An Acoustic–Phonetic Approach in Automatic Arabic Speech Recognition", the British Library in Association with UMI (1990).

[5] D. Vergyri, K. Kirchhoff, K. Duh and A. Stolcke, "Morphology-based Language Modeling for Arabic Speech Recognition", In INTERSPEECH-2004, pp. 2245-2248 (2004).

[6] K. Kirchho, J. Bilmes, J. Henderson, R. Schwartz, M. Noamany, P. Schone, G. Ji, S. Das, M. Egan, F. He, D. Vergyri, D. Liu and N. Duta. Novel Approaches to Arabic Speech Recognition (2002).

[7] D. Vergyri, K. Kirchhoff, "Automatic Diacritization of Arabic for Acoustic Modeling in Speech Recognition". In Ali Farghaly and Karine Megerdoomian, editors, COLING 2004 Computational Approaches to Arabic Script-based Languages, pp. 66–73, Geneva, Switzerland (2004).

[8] Markus Cozowicz, "Large Vocabulary Continuous Speech Recognition Systems and Maximum Mutual Information Estimation", Diploma, Vienna University of Technology, August 23 (2006).

[9] H. Satori M. Harti and N. Chenfour, "Introduction to Arabic Speech Recognition Using CMU Sphinx System", Submitted to Int. Journal of Computer Science Application (2007).

[10] A. Muhammad, "Alaswaat Alaghawaiyah," Daar Alfalah, Jordan, (in Arabic) (1990).

[11] J. Deller, J. Proakis, J.H. Hansen, "Discrete-Time Processing of Speech Signal", Macmillan (1993).

[12] M. Elshafei, "Toward an Arabic Text-to-Speech System", the Arabian J. Science and Engineering, Vol. 4B (16), pp. 565–583 (1991).

[13] Y.A. El-Imam, "An Unrestricted Vocabulary Arabic Speech Synthesis System", IEEE Transactions on Acoustic, Speech and Signal Processing, Vol. 37, (12), pp. 1829–1845 (1989).

[14] Abhinav Sethy, Shrikanth Narayanan and S. Parthasarthy, "A syllable-based Approach for Improved Recognition of Spoken Names", Proceedings of the ISCA Pronunciation Modeling Workshop, Estes Park, Colorado, September (2002).

[15] Ganapathiraju, J. Hamaker, M. Ordowski, G. Doddington and J. Picone, "Syllable-Based Large Vocabulary Continuous Speech Recognition", IEEE Transactions on Speech and Audio Processing, Vol. 9 (4), pp. 358-366 (2001).

[16] K. Kirchhoff, "Syllable-Level Desynchronisation of Phonetic Features for Speech Recognition", International Conference of Spoken Language Processing, pp. 2274-2276 (1996).

[17] Su-Lin Wu, Brian Kingsbury, Nelson Morgan, and Steven Greenberg, "Incorporating Information from Syllable-length Time Scales into Automatic Speech Recognition", ICASSP-98, Seattle, pp. 721-724.

[18] S. Young, G. Evermann, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey,V. Valtchev and P. Woodland. The HTK Book. Revised for HTK Version 3.2 (2002).

Received April 4, 2006
Accepted May 31 2008