SYLLABLE-BASED AUTOMATIC ARABIC SPEECH RECOGNITION IN NOISY-TELEPHONE CHANNEL

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.
engazm@yahoo.com,htol@link.net,sabdou@jaguar.it.miami.edu & mervat_fashal@yahoo.com

Abstract: - The performance of well-trained speech recognizers using high quality full bandwidth speech data is usually degraded when used in real world environments. In particular, telephone speech recognition is extremely difficult due to the limited bandwidth of transmission channels. In this paper, we concentrate on the telephone 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. In clean environment, 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. Also in noisy telephone channel, syllables outperformed the rate obtained using monophones, triphones and words by 2.09%, 1.5% and 0.9% respectively. Comparative experiments have indicated that the use of syllables as acoustic units leads to an improvement in the recognition performance of HMM-based ASR systems in noisy environments. 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, Noisy-channel.

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 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. Syllable models also automatically capture co-articulation effects [3].

Digit	Arabic Writing	Syllables	Representation of syllables & phonemes	Triphones
0	صفر	CVCC	(ss.i.f.r)	ss+i ss-i+f i-f+r f-r
0	زيرو	CV-CV	(z.ii-r.uu)	z+ii z-ii +r ii-r+uu r-uu
1	واحد	CV-CVC	(w.aa)-(X.id)	w+aa w-aa+X aa-X+i X-i+d i-d
2		CVC-CVC	(A.i.t)-(n.ii.n)	A+i A-i+t i-t+n t-n+ii n-ii+n ii-n
3		CV-CV-CVC	(t.a)-(l.aa)-(t.a.h)	t+a t-a+l a-l+aa l-aa+t aa-t+a t-a+h a-h
4		CVC-CV-CVC	(A.a.r)-(b.a)-(E.a.h)	A+a A-a+r a-r+b r-b+a b-a+E a-E+a E-a+h
5		CVC-CVC	(x.a.m)-(s.a.h)	x+a x-a+m a-m+s m-s+a s-a+h
6		CVC-CVC	(s.i.t)-(t.a.h)	s+i s-i+t i-t+t t-t+a t-a+h a-h

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

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 Acoustic Units

The most natural unit of speech, the *word*, is able to capture within word contextual effects [10]. Word models had been used as the basic speech unit for both isolated word recognition and for connected word recognition systems because words have the property that their acoustic representation is well defined, and the acoustic variability occurs mainly at the beginning and the end of the word. Also, a word lexicon is unnecessary in the recognizer when word models are used. Despite the advantages of the use of the word as an acoustic unit in isolated word and connected word recognizers, it is not a practical choice for large vocabulary continuous speech recognition because the amount of the training data and the storage required are enormous. In addition, the phonetic content of the individual words overlaps in a large vocabulary [11]. Hence some more efficient representation, (i.e., subword) is required for large vocabulary systems. Subword speech units are more efficient to be used for large vocabulary CSR systems. There are several possible choices for subword units that can be used to describe a language. These include: phones (which are the basis for writing down a language and the smallest segments of sounds that can be distinguished within words), multi-phones, e.g., syllables, demisyllables, and diphones [12].

2.1 Context-Independent Phones

Phones are the smallest units that can be used to represent the speech. Since the number of phones representing a language is small (in English there are approximately 50 phones), then phones can be sufficiently trained with a small number of sentences. Moreover, they are also vocabulary independent. However, a phone is strongly affected by its neighboring phones, i.e., by the context. Hence, it is not advantageous to use such a unit, because the phone model assumes that a phone in any context is equivalent to the same phone in any other context. In addition, phonetic models are inconsistent across different vocabularies. Therefore some more efficient representation is required.

A multiphone unit overcomes the disadvantages of the phone unit by taking into account the neighbors' phones. The multiphone units are able to model the co-articulatory effects, hence they are more consistent. These units include syllables, demisyllables, and diphones. A very significant reduction in the number of units can be achieved by employing demisyllables instead of syllables (in English, the number of syllables is about 10,000 while the number of demisyllables is about 2,000. Two problems were found when such units were used for speech recognition. The first problem is due to the large number of multiphones needed, and the second one is the lack of generalizability of such units when used in vocabulary-independent systems. The number of diphones required to present a language is similar to that of demisyllables (in English there are approximately 1,000-2,000). The problem of segmentation, deciding where one ends and the next begins, however, is much more difficult with diphones than demisyllables.

It should be noted that phone-based models of words are essential for large-vocabulary systems for which training of complete word models is not feasible. On the other hand, word-dependent phone models take into consideration the phonological variations of the different phonemes for each word in the vocabulary. The parameters of a word-dependent phone depend on the word in which the phone occurs. Word-dependent phone models require considerable training and storage.

2.2 Context-Dependent Phones

Context-dependent modeling is necessary for large vocabulary continuous speech recognition. A triphone model takes into account both the left and the right neighboring phones. Triphone models capture the most important co-articulatory effects, so they are much more powerful and consistent than phone models. The disadvantages of the triphone models are the need of a large amount of memory, and the poor training due to the triphone models' large number. Several solutions in [10] had been proposed to solve these problems.

Generalized triphones are created from triphone models using a clustering procedure that combines triphone HMMs according to an information-theoretic distance measure. This would lead to a much more manageable number of models that can be better trained, and hence overcome the training's and storage's problems for the triphone models. It was found that generalized models work better than traditional triphone models [10].

There are a great number of factors which cause variability in speech. Articulation variabilities, such as left-context, right-context, and linguistic and speaker variabilities affect the acoustic behavior of phones, and consequently generate different realizations for each phone, which called *allophones*. By taking into account important acoustic-phonetic variabilities, more detailed and consistent models, allophonic models, can be

created to be used in a large vocabulary recognition system. The left- and right-contexts used in triphones models are two of many variabilities that affect the realization of a phone. Hence, triphones can be considered as a special case of allophones. The disadvantages of these models are the need of a large amount of memory, and the poor training due to their huge number. Two clustering methods had been proposed in [10] to solve these problems. The first method, agglomerative clustering, is completely data-driven and finds clusters without external guidance. The second one uses a decision tree clustering algorithm which integrates both linguistic knowledge and data-driven statistical modeling. It was found in [10] that both models are equally consistent, but the latter overcomes the former's problems. It was found also that the decision-tree-based allophonic models improved the performance of a vocabulary independent continuous speech recognition system to the level of vocabulary-dependent system, and even better.

An acoustic unit named the fenone [13] is a very small unit of speech corresponding to a small number of frames in the acoustic waveform. Fenone-based models can be applied when the speech waveform for any word is reduced to a vector-quantized observation string. Variations of the observation string for any other utterance of the same word can be modeled by replacing each observation in the original training sequence by a small HMM capable of learning and generating the variability surrounding the original single-frame observation. The benefit of using these models is that they are more robust to SI training than phonetic models [13].

2.3 Criteria for desirable Subword Units

Desirable subword units must satisfy three criteria. First, subword units must be consistent, i.e., the variabilities within a model are minimized. Second, they must be trainable, i.e., have sufficient training data for each model. Third, they must be generalizable, i.e., reasonable models for the subword units must be found in spite of the lack of precise coverage in training.

3 Syllable-Based ASR

3.1 Syllable Structure

A syllable is typically composed of more than one phoneme. It is phonologically known that syllable is a complex unit made up of nuclear and marginal elements. Nuclear elements are the vowels or syllabic segments, and marginal elements are the consonants or non-syllabic segments. Standard dictionaries provide syllabification that is influenced by the morphological structure of words [14].

Moreover, a syllable can be described by a series of grammars. The simplest grammar is the phoneme grammar, where a syllable is tagged with the corresponding phoneme sequence. The consonant-vowel grammar describes a syllable as a consonant-vowel-consonant (CVC) sequence. The syllable structure grammar divides a syllable into onset, nucleus and coda (ONC) [14]. The nucleus is obligatory which can be either a vowel or a diphtong. An onset is the first part of a syllable consisting of consonants and ending to the nucleus of the syllable. A nucleus is the vowel part of a syllable. A coda is the part of a syllable that follows the nucleus. A coda is constructed of consonants. The nucleus and coda are combined to form the rhyme of a syllable. A syllable has a rhyme, even if it doesn't have a coda. In the syllable structure grammar, the consonants are assigned as onset or coda. It contains more information than the CVC structure for multi-syllable words [14].

Our aim in this paper is to study the effect of using syllables for the ASR process in noisy environments. Syllables have long unit and they have the least context sensitive. The advantage of using syllables as an acoustic unit for the ASR process is that pronunciation variation is trained right into the acoustic model and does not need to be modeled separately in the dictionary. Syllable models also automatically capture co-articulation effects [15, 16].

3.2 Syllable-Based ASR

The first step in designing a syllable-based recognition system is the preparation of the syllabic lexicon. Syllables are represented in terms of the underlying phone sequence. Hence, given the phonetic transcription of the speech in a standardized format (IPA for example), a syllable representation could be written using a set of syllable symbols from the phonemes comprising the syllable. The second step in designing a syllable lexicon is to identify the phone clusters, which correspond to the correct syllabic representation. The process of clustering phones to get a syllable representation is called syllabification. This process is described in [16] as a set of rules which define permitted syllable-initial consonant clusters, syllable-final consonant clusters and prohibited onsets. NIST Syllabification software available [17]

implements the syllabification rules and offers a set of alternative possible syllable clusters given a phoneme sequence which are used to generate the syllabic lexicon.

3.3 Word-Based ASR

The implementation of the word recognizer is similar to the syllable recognizer. The only difference being that the pronunciation phone sequence for every word in the lexicon is used as a separate word level unit. Thus we have acoustic units corresponding to all the different words in the lexicon. Homophones were given the same lexical representation. Model topology and initialization strategies are identical to the syllable-based recognizer. Initialization from phoneme level models in this manner ensures that the syllable and word level models have performance identical to or only slightly lesser than the corresponding phoneme recognizer even without further acoustic training. Training on acoustic data leads to substantial improvement in accuracy as the temporal and spectral correlation information gets embedded in the longer length units. However the achievable gain results depends on the coverage of the unit in the training data as well as the linguistic nature of the unit. Thus a word or syllable unit with no training data will not lead to improvement in accuracy as compared to the corresponding phonemic representation. Thus we need to identify the proper lexical representation or choice of units to represent.

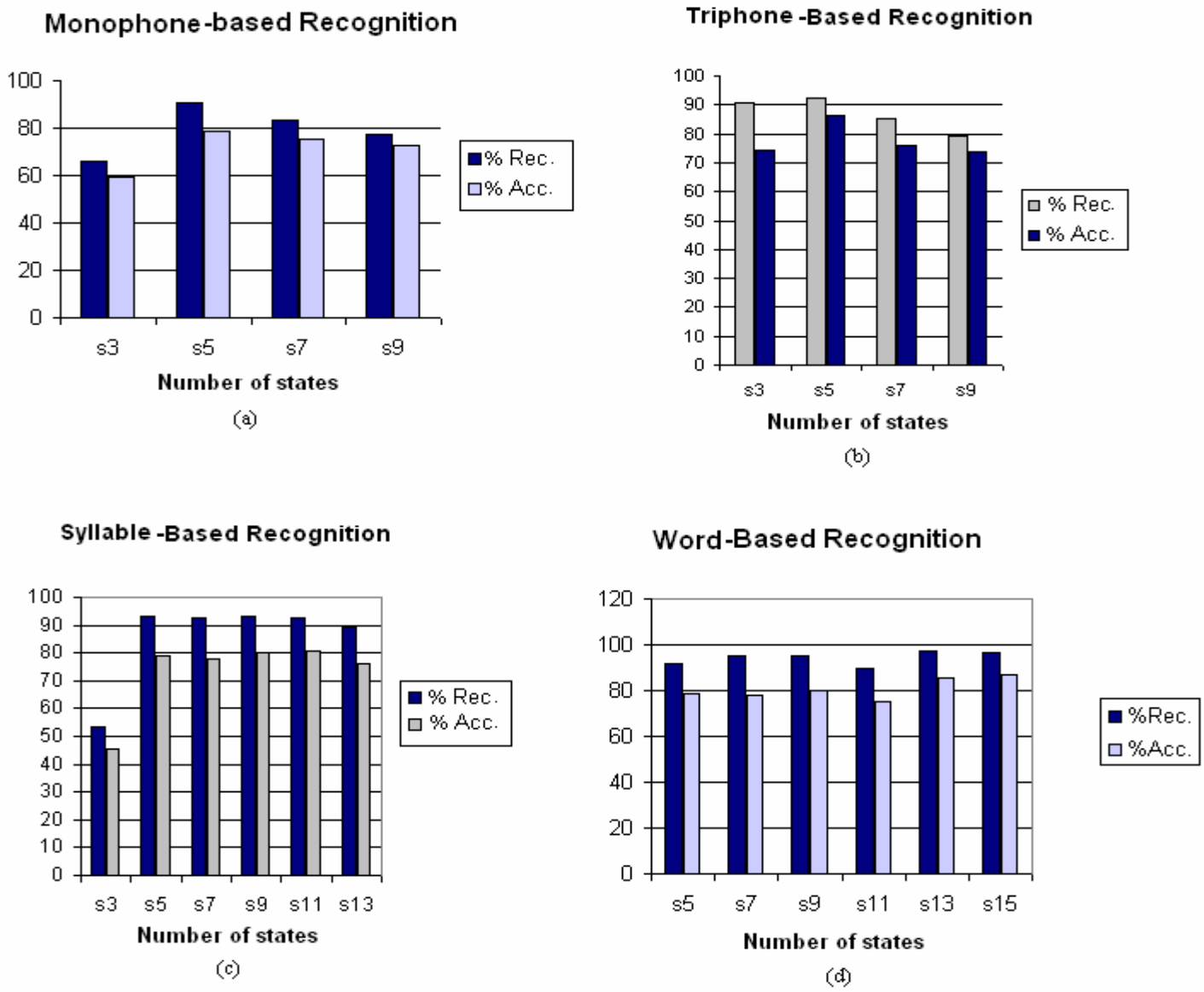

Fig. 1: A bar graph showing the effect of increasing the number of states per model on the recognition rate and accuracy of (a)monophone-based recognition (b) triphone-based recognition (c)syllable-based recognition (d) word-based recognition.

3.4 Syllable-Based ASR of Arabic Speech

Standard Arabic has 34 basic phonemes, of which six are vowels, and 28 are consonants [18]. Arabic has fewer vowels than English. It has three long and three short vowels, while American English consists from at least 12 vowels [13]. Arabic phonemes contain two distinctive classes, which are named pharyngeal and emphatic phonemes. These two classes can be found only in Semitic languages [13][18][19]. 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 [18]. 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 [20]. 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 [15-16] as well as studies of human perception [21] 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 [17]. 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% [16]. In 2002, Sethy et al. obtained 80% of syllable-based recognition [15]. 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 Proposed ASR Engine

In most ASR systems the speech signal is segmented into consecutive frames and despite clear correlations between successive frames, each frame is parameterized separately. The parameterization process serves to maintain the relevant part of the information within a speech signal while eliminating the irrelevant part for the ASR process. A wide range of possibilities exists for parametrically representing the speech signal such as: short-time spectral envelope, LPC coefficients, MFCCs, short-time energy, zero crossing rates and other related parameters [22]. Among all the parameterization methods, the cepstrum has shown to be favorable for ASR and is widely used in many ASR systems [22].
To better represent temporal variations in the speech signal, higher-order time derivatives (or simply, *delta* parameters for first derivatives, *delta-delta* parameters for the second derivatives) of signal measurements are added to the set of static parameters. The combination of dynamic and static features had shown additional discriminability for speech pattern comparison and consequently improved the accuracy of the speech recognition process. Moreover, temporal variations in the speech signal, obtained by applying time derivatives to the speech signal, when combined with the static features mentioned above, had shown additional discriminability for speech pattern comparison.

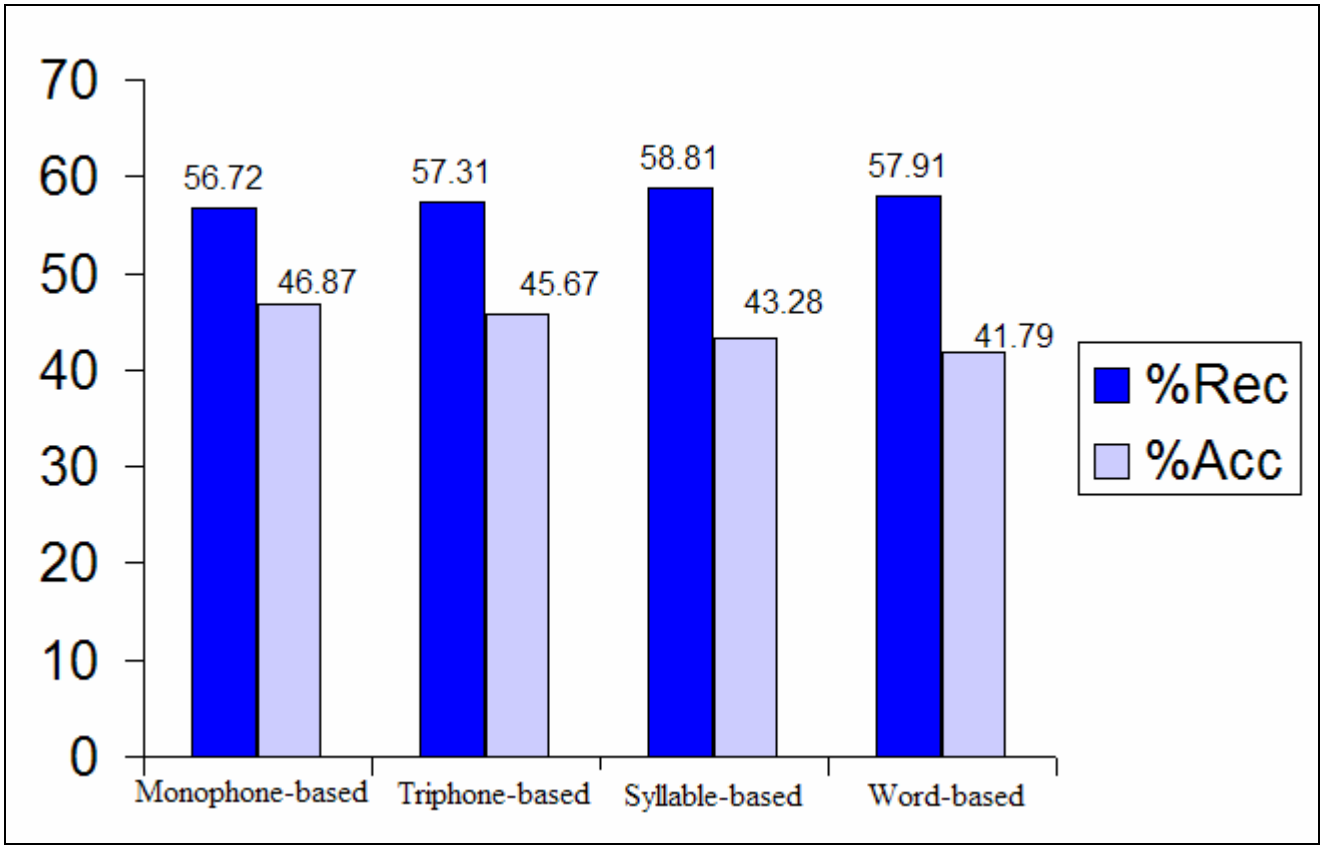

Fig. 2: The relation between different recognizers when the signal is passed through noisy-telephone channel

The primary ASR parameters are extracted from the spectral envelope, based on the assumption that enough information for ASR resides in the spectral envelope. About 10 coefficients derived from a Fourier transform, LPC analysis, or bank of bandpass filters are considered to provide sufficient and efficient information to model the short-time spectral envelope. MFCCs, LPC coefficients, LSFs, reduced forms of DFT, and zero crossing rates in bandpass channels are some of the known ASR parameters. However, over the past few decades MFCCs have been the most used parameters in the state-of-the-art ASR systems due to their good performance in clean speech recognition.

5 Experiments & Results

5.1 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. Then, the tested data is passed through noisy telephone channel.

In order to recognize the continuous speech data that has been enhanced as mentioned above, the HTK toolkit described in [23] has been used throughout all experiments. This toolkit is used to build an HMM-based speech recognition system. The HTK toolkit can be used for isolated or continuous whole-word-based speech recognition. The toolkit was designed to support continuous density HMMs with any numbers of state and mixture components. It also implements a general parameter-tying mechanism which allows the creation of complex model topologies to suit a variety of speech recognition applications. For more details see [23].

5.2 Experiments

5.2.1 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 in clean environment. The recognition rate for 3-

states, 5-states, 7-states and 9-states were found to be 66.27%, 90.75%, 83.58% and 77.01% 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. 5-states are chosen due to the highest recognition percent in monophone-based recognition. As shown in Fig.2, when the tested data is passed through noisy telephone channel, the recognition percent is 56.72% and the accuraccy percent is 46.87%.

5.2.2 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 in clean environment. 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. 5-states are chosen due to the highest recognition percent of triphone-based recognition. As shown in Fig.2, when the tested data is passed through noisy telephone channel, the recognition percent is 57.31% and the accuraccy percent is 45.67%.

5.2.3 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 in clean environment. 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. 5-states are chosen due to the highest recognition percent of syllable-based recognition. As shown in Fig.2, when the tested data is passed through noisy telephone channel, the recognition percent is 58.81% and the accuracy percent is 43.28%.

5.2.4 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 in clean environment. 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

	%H	%D	%S	%I
Monophone-based recognition	90.75	3.58	5.67	12.24
Triphone-based recognition	92.24	4.18	3.58	5.67
Syllable-based recognition	93.43	2.09	4.48	14.33
Word-based recognition	91.64	4.48	3.88	13.13

Table 2: A comparison between the recognition rates for the performance of our proposed recognizer using the different units.

	%H	%D	%S	%I
Monophone-based recognition	56.72	11.94	31.34	9.8
Triphone-based recognition	57.31	10.74	31.94	11.64
Syllable-based recognition	58.81	9.25	31.94	15.52
Word-based recognition	57.91	8.65	33.43	16.12

Table 3: A comparison between the recognition rates recognition rates for the performance of our proposed recognizer using the different units in noisy-telephone channel.

and 15-states were found to be 78.51%, 78.21%, 97.7%, 74.93%, 85.37% and 86.8% respectively. 5-states are chosen to be compared with monophones, triphones and syllables. As shown in Fig.2, when the tested data is passed through noisy telephone channel, the recognition percent is 57.91% and the accuraccy percent is 41.79%.
As shown in Table 2-3: 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.

6 Conclusions & Future Work

Several experiments were conducted on automatic recognition of Egyptian Arabic speech recognition based on HMMs using HTK in clean and noisy environment. 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 in clean environment. Also, in noisy-telephone channel, 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 vocabukary continous 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, A. Stolcke, "Morphology-based language modeling for Arabic speech recognition", In INTERSPEECH-2004, 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.2002 . Novel approaches to Arabic speech recognition.
[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] H. Hon, “Vocabulary Independent Speech Recognition: The VOCIND System”, PhD Thesis, CMU, Computer Science Department, 1992.
[11] L. Rabiner & B. H. Juang, “Fundamentals of Speech Recognition”, PTR Prentice-Hall Inc., 1993.
[12] P. Ladefoged, “A Course in Phonetics”, Harcourt Brace Jovanovich Inc, 1982.
[13] J. Deller, J. Proakis & J. Hansen, “Discrete-Time Processing of Speech Signals”, Macmillan Pub. Co., 1993.

[14] Jilei Tian, "Data-Driven Approaches for Automatic Detection of Syllable Boundaries", Proceeding of the ICSLP'04, 2004.
[15] Abhinav Sethy, Shrikanth Narayanan and S. Parthasarthy, "A syllable based approach for improved recognition of spoken names", Proceedings of the ISCA Pronunciation Modeling Workshop, Colorado, 2002.
[16] Ganapathiraju, J. Hamaker, M. Ordowski, G. Doddington & J. Picone, "Syllable-Based Large Vocabulary Continuous Speech Recognition", IEEE Transactions on Speech and Audio Processing, Vol. 9, No. 4, pp. 358-366, May 2001.
[17]K. Kirchhoff, "Syllable-level desynchronisation of phonetic features for speech recognition", Proceeding of the ICSLP'96, pp 2274-2276, 1996.
[18] A. Muhammad, "Alaswaat Alaghawaiyah," Daar Alfalah, Jordan, 1990 (in Arabic).
[19] M. Elshafei, "Toward an arabic text-to-speech system," The Arabian J.Science and Engineering vol. 4B no. 16, pp. 565–583, 1991.
[20] Y.A. El-Imam, "An unrestricted vocabulary arabic speech synthesis system", IEEE Transactions on Acoustic, Speech and Signal Processing vol. 37 , no. 12, pp.1829–1845, 1989.
[21] D. O'Shaughnessy, "Speech Communication: Human and Machine", IEEE Press, 2001.
[22] 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.
[23] S. Young, G. Evermann, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey,V. Valtchev, P. Woodland. The HTK Book. Revised for HTK Version 3.2 Dec. 2002.