AUTOMATIC SEGMENTATION OF ARABIC SPEECH

Moustafa Elshafei Mohammad Ali, Husni Al-Muhtaseb, and Mansour Al-Ghamdi*

King Fahd University of Petroleum and Minerals

*King Abdulaziz City of Science and Technology

ABSTRACT

This paper presents the results and conclusions of a study on speech segmentation system for Arabic speech. Automatic phonetic segmentation is the problem of automatically locating the boundaries between the sounds corresponding to the phones that make up a fragment of speech. The developed system accepts a speech utterance and its orthographic transcription, and generates its phonemic transcription and the segmentation information of the utterance. The system was trained using a corpus of Arabic TV news. A phonetic dictionary was also developed using orthographic-to-phonemic transcription rules. The corpus is used to develop context-independent hidden Markov models for the Arabic phonemes. The system was validated against manually segmented speech utterances.

1. INTRODUCTION

Speech technology development strongly relies on corpus-based methodologies and, therefore, on the availability of good speech corpora. In order for a corpus to be really useful for the development of speech recognizers or speech synthesizers, apart from the speech itself, it should contain information about its contents (labeling) and about the time alignment between labeling and speech (segmentation). Phones are usually considered the smallest units of speech, by the concatenation of which any other speech unit (syllable, word, phrase, etc.) can be built [1,2]

This paper investigates the development of a speech segmentation system for Arabic speech. Speech segmentations, at the phonemic level such as TIMIT or the word level such as Switchboard, CGN, have become a standard annotation in speech corpora for training speech recognition systems. The segmentations link the orthographic/phonemic transcription of the speech to time stamps in the speech signals. Such segmentations are indispensable for the initial training of acoustic automatic speech recognition (ASR) models, the development of text-to-speech (TTS) systems and speech research in general. In standard Corpus such as TMIT the segmentation was performed manually by experts. As manual phonetic transcriptions require availability of expertise and a lot of time and money, they are not always available for speech corpora.

The method proposed in this paper uses general features and acoustic modeling which are common in ASR [3,4]. The proposed segmentation system starts from a phonetic transcription that is automatically generated on the basis of its orthography. The complete segmentation process is composed of two subtasks. First, a number of alternative phonetic transcriptions is produced on the basis of a given orthographic transcription. Then, this single phonetic transcription serves as input to a segmentation system based on either the Viterbi or the Forward-Backward algorithm to select the acoustically best matching phonetic representation..

During the last few years, however, the need to develop new voices and languages quickly and with the maximum quality (which frequently implies large inventories) has raised the interest in automatic segmentation techniques to partially automate the development of

synthesis inventories and models [4,5]. It is worth mentioning that some researchers believe that speech recognition can benefit from more precise segmentation in training or decoding [6]. The most frequent approach for automatic phonetic segmentation is to modify an HMM based phonetic recognizer to adapt it to the task of automatic phonetic segmentation [3-8]. The main modification needed consists in letting the recognizer know the phonetic transcription of the sentence to segment (which is considered known) by building a recognizer's grammar for that transcription and performing a forced alignment. rule based detection of phonetic boundaries, neural networks as phonetic boundary detectors, alignment of the utterance to the same utterance produced by a speech synthesizer, and dynamic time warping (DTW) [7,8,12]. Finally, some researchers have tried to combine HMMs with other features and techniques. In [11] a pre-segmentation technique is used followed by HMMs and cepstral coefficients to align the spectrally stable segments to phones.

2. FROM ORTHOGRAPHY TO PHONETICS

In this section, we report briefly a developed algorithm for automatic diacritization of the Arabic text. The detailed algorithm is published in [9]. We formulated the problem of generating Arabic diacritized text from undiacritized text using a Hidden Markov Model (HMM) approach. The word sequence of undiacritized Arabic text is considered an observation sequence from an HMM, where the hidden states are the possible diacritized expressions of the words. The optimal sequence of diacritized words (or states) is then obtained efficiently using Viterbi Algorithm.

The developed corpus focuses on recognition of radio and TV news transcription in Modern Standard Arabic (MSA). The MSA is widely used and accepted over the entire region and contains a reasonable set of vocabulary for development and testing the continuous speech recognition system. The audio files were recorded from several TV news channels at a sampling rate of 22kHz. A total of 161 news stories, summing up to 4.5 hours of speech, were recorded and split up to 3627 files with an average file length of 4.5 seconds. The length of wave files range from 0.8 seconds to 15.6 seconds.

For the sake of training, the audio files were resampled at 16kHz. Additionally, a 0.1 second silence period is added to the beginning and end of each file. Some of the files have background noise that are of the following types:

1-Background music that accompanies the news headlines: although this kind of music was deliberately avoided while recording, some files might have fainting music at the beginning.

2-Some background noise might occur when the reporter is in an open location such as a stadium or a stock market.

Secondly, the orthographic transcription formed the basis for all other transcriptions and annotations. The orthographic transcription should require a minimum of interpretation. Thus grammatical 'errors' were not to be corrected and broken-off words were written down as such (they remained incomplete). In line with the recommendations made in e.g. the documentation with the Switchboard and SpeechDat corpora, it was decided to adopt normal common spelling conventions.

All the 3627 files were completely transcribed with fully diacritized text. The transcription is meant to reflect the way the speaker has uttered the words, even if it is grammatically wrong. It is a common practice in Modern Standard Arabic and most Arabic dialect to drop the vowels at the end of words; this situation is represented in the transcription by either using a silence mark (Sukun) or dropping the vowel, which is considered the same as a Sukun in later training stages.

The transcription file contains 31,055 words. The vocabulary list contains 12604 words.

All the recorded material was transcribed orthographically. The orthographic transcription is a verbatim record of what was actually said. In the transcription process repetitions, hesitations, false starts and such were transcribed. Background noise, on the other hand, was seldom represented in the transcriptions. Moreover, the transcription has been checked manually.

The automatic conversion from an orthographic to a phonetic transcription takes two steps. First, several techniques are applied to produce a network of plausible pronunciation variants. In a second step, the single best matching phonetic string is selected by means of an ASR system. Grapheme-to-phoneme conversion is a central task in any text-to-speech system [1,2]. Given an alphabet of spelling symbols (graphemes) and an alphabet of phonetic symbols, a mapping should be achieved transliterating strings of graphemes into strings of phonetic symbols. It is well known that this mapping is difficult because in general, not all graphemes are realized in the phonemic transcription, and the same grapheme may correspond to different phonetic symbol, depending on context.

A full network of alternative phonetic transcriptions is generated on the basis of orthographic information. Lexicon lookup is a simple but efficient way to acquire phonetic word transcriptions. Yet, not every orthographic unit is a plain word. Some speech fragments contain sloppy speaking styles including broken-off words, mispronunciations and other spontaneous speech effects.

Arabic provides (multiple) phonetic transcriptions for most of the standard words. Rules were developed to cover non-listed compounds, derivations and in ctions formed on the basis of Arabic entries. Lexicon lookup is also the first option for foreign words. If a foreign word is part of more than one of these lexica, the different phonetic realizations are put in parallel since the orthography does not specify which foreign language was used.

The outcome of the above techniques is a compact pronunciation network. To select the transcription matching best with the speech signal, all phonetic alternatives are acoustically scored (maximum likelihood) in a single pass (Viterbi) through our speech recognition system and the most probable one is retained. The phoneme models are statistically represented as three-state left-to-right Hidden Markov Models (HMMs).

3. ARABIC PHONEME SET

Table 1 shows the listing of the phoneme set used in training and the corresponding symbol:

Table 1. The complete phoneme list used in training

الحرف	الرمز الصوتي		الحرف	الرمز الصوتي
/AE/	بَ		/KH/	خ
/AE:/	بَـاب		/D/	د
/AA/	خَ		/DH/	ذ
/AH/	قَـد		/R/	ر
/UH/	بُ		/Z/	ز
/UW/	دُون		/S/	س
/UX/	غُصن		/SS/	ص
/IH/	بِ		/DD/	ض
/IY/	فِـيل		/TT/	ط
/IX/	صِنف		/DH2/	ظ
/AW/	لَـوم		/AI/	ع
/AY/	ضَيف		/GH/	غ

/UN/	نُـنجي		/F/	ف
/AN/	نَم		/V/	فـيـتـامـين
/IN/	مِـمـا		/Q/	ق
/E/	ء		/K/	ك
/B/	ب		/L/	ل
/T/	ت		/M/	م
/TH/	ث		/N/	ن
/JH/	جيم فصيحة		/H/	هـــ
/G/	جيم مصريـة		/W/	و
/ZH/	جيم مـعطشة		/Y/	ي
/HH/	ح			

4. ARABIC PHONETIC DICTIONARY

Using the selected phoneme set, we developed a Java tool (***ArabicPhoneticDicitonary***) that automatically generates a dictionary for a given transcription.

The tool takes care of the following issues:

1- Choosing the correct phoneme combination based on the location of the letters and their neighbors.
2- Providing multiple pronunciations for words that might be pronounced in different way according to:
    a. The context in witch the words is uttered, which might change the way the beginning and the end of the word is pronounced. For example, Hamzat al-wasl (ٱ) at the beginning of the word and the Ta' al marbouta (ـة) at the end of the word.
    b. Words that have multiple readings due to dialect issues.
    c. Foreign names, such as Lagrange, Vector, and Surgery, where the translation might not reflect the exact pronunciation.

We defined a set of rules based on regular expressions to define the phonemic definition of words. The tools scans the word letter by letter, and if the conditions of a rule for a specific letter are satisfied, then the replacement for that letter is added to a tree structure that represents all the possible pronunciations for that words.

Each rule has the following structure:

LETTER:

(pre_condition) . (post_condition) -> replacement

Where LETTER represents the current letter in the word, pre_condition and post_conditon are regular expressions that represent other letters surrounding the current letter, and replacement is the replacement phoneme or phonemes. The number of pronunciations in the developed phonetic dictionary came to 21,525 entries. A sample from he developed phoneme dictionary is listed below.

آبَارِ E AE: B AE: R IX N
آخَرَ E AE: KH AA R
آخَرَ E AE: KH AA R AA
آخَرُوْنَ E AE: KH AA R UW N AE
آخَرِيْنَ E AE: KH AA R IX: N AE
آخَرِيْنْ E AE: KH AA R IX: N
آخَرْ E AE: KH AA R
آخِذَةٌ E AE: KH IX DH AE T UH N
آخِرَ E AE: KH IX R AA
آخِرْ E AE: KH IX R

آذارْ E AE: DH AE: R
آرْ E AE: R
آسْبَا E AE: S Y AE:
آسْيَانْ E AE: S Y AE: N
آسْيَوِيَّةْ E AE: S Y AE W IH Y AE H
آسْيَوِيَّةْ(٢) E AE: S Y AE W IH Y AE T
آفَاق E AE: F AE: Q IX
آفَاقْ E AE: F AE: Q
آل E AE: L
آلافٍ E AE: L F IH N
آلافْ E AE: L F
آلافِ E AE: L AE F IH

5. HMMS FOR PHONETIC SEGMENTATION

The second step for refining the phonetic segmentation is to use a modified HMM phonetic recognizer. The objective here is to build on the extensive knowledge and infrastructure available in the speech recognition field to discover alternative phoneme pronunciations for words.

The sampling rate is 16 ksps, and analysis window is 25.6 msec (about 410 samples), with consecutive frames overlap by 10 msec. Each window is pre-emphasized and is multiplied by a Hamming window [10]. The basic feature vector uses the Mel Frequency Cepstrum Coefficients MFCC. The mel-frequency scale is linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000 Hz. The MFCCs are obtained by taking the Discrete Cosine Transform (DCT) of the log power spectrum from Mel spaced filter banks. The system uses a 12-coefficients basic feature vector. The basic feature vector is usually normalized by subtracting the mean over the sentence utterance. $x(0)$ represents the log mel spectrum energy, and is used to derive other feature parameters. The basic feature vector is highly localized. To account for the temporal properties, 3 other derived vectors are constructed from the basic MFCC coefficients: a 40-ms and 80-ms differenced MFCCs (24 parameters), a 12-coefficient second order differenced MFCCs, and 3–dimensional vector representing the normalized power (log energy), differenced power, and second-order differenced power The HMM shown in Figure 1 to represent the speech phonemes. The model, known as Bakis model, has a fixed topology consisting of 3 emitting sates and one output state.

Output probability distributions in HMM states are modeled with mixtures of 8 diagonal covariance Gaussians. First, context-independent HMMs of one Gaussian were trained. Then these HMMs were successively extended to one more Gaussian and re-estimated to
up to 8 Gaussians. Baum-Welch re-estimation was used during the whole process. The phonetic labels used were generated automatically from the orthographic transcription using a set of rules. Alternative pronunciations were not taken into account, but we do not expect this to constitute a major problem because the training material was recorded and manually verified to avoid important dialectal and pronunciation variations.

It is a common practice to use context-independent HMMs for speech segmentation [3,4]. This contrasts with the generally extended practice of using context- dependent HMMs for speech recognition. Context-dependent HMMs can better model the spectral movements in phonetic transitions. However, the segmentations they produce tend to be less precise than the ones produced by context-independent HMMs. A theoretical explanation for this behavior was presented in [5], where it was argued that the cause is the loss of alignment, during the training process, between the context-dependent HMMs and the phones. Context-dependent HMMs are always trained with realizations of phones in the same context. For that reason,

the HMMs do not have any information to discriminate between the phone and its context. As a result the HMM (particularly the lateral states) can end up modeling part of other phones or not all the phone. Context-independent HMMs, on the other hand, are trained with realizations of phones in different contexts. For that reason they should be able to discriminate between the phone to model (invariable in all the training examples) and its context (which varies).

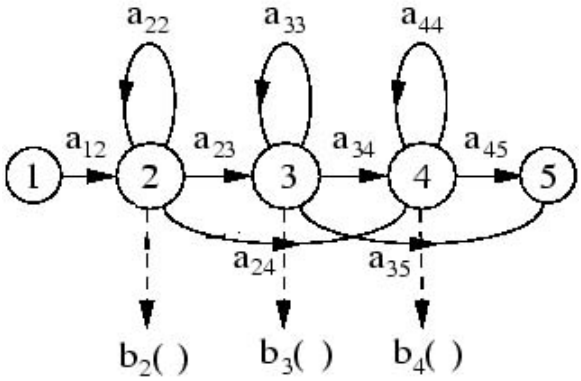

Figure 1. The 5-states HMM phoneme model.

Once a phonetic transcription has been selected, automatic segmentation can proceed in the following way. Sentence models are rst generated by simply concatenating all relevant phoneme models. Next, the speech data are assigned by respectively Viterbi to the acoustic model of the complete phoneme sequence.

The Viterbi algorithm returns the single best path through the model given the observed speech signal x_i , $i = 1,2,...T$, where T is the number of frames in the utterance.

$$s_i = \arg\{ \max_{s_i \subset S} \prod_{i=1}^{T} f(x_i \mid s_i) p(s_i \mid s_{i-1}) , \quad (1)$$

With si a sequence of HMM states (one state for each time frame) which is consistent with the sequence model S,T being the number of time frames. Thus, the Viterbi algorithm results in the segmentation which reaches maximum likelihood for the given feature vectors.

6. EVALUATION

It is possible to evaluate segmentation performance using indirect figures of merit, for example measuring the word error rate of a recognizer that uses a segmentation stage or measuring the subjective quality of a speech synthesizer obtained using automatic segmentation. However, the most common and direct form of evaluation is comparing the segmentation to a manual segmentation and computing some figures of merit. Among the different figures of merit used we may mention: mean error , RMS error, and percentage of errors smaller than a tolerance value [3,4,7]. The most commonly reported figure of merit and perhaps the most useful one for comparison is the percentage of boundaries with errors smaller than 20 ms.

The developed phoneme HMM models are used to perform automatic phoneme and state alignment of some speech utterances. The automated alignment was then verified manually

by inspection of the waveform using standard speech editing tools. The resulted automated alignment coincide with the manual alignment to within one frame (10 msec) error.
The example below shows the transcript of an utterance from the corpus, the resulting phoneme alignment and the state alignment are shown in Table. 2.

The transcription for the file is:
<s> اِرْتَفَعَتْ الأَسْهُمْ الْكُوَيْتِيَّة اِرْتِفَاعًا طَفِيفًا الْيَوْمْ السَّبْتْ **</s>**

Table 2. Forced alignment of an utterance using the CI models.
The table showes frame number, phoneme name, and the emitting state number.

No	Ph.	St.		No	Ph	St		No	Ph	St		No	Ph	St		No	Ph	St
1	SIL	0		51	E	0		101	L	2		151	IH	0		201	AE:	2
2	SIL	0		52	E	1		102	L	2		152	IH	1		202	AE:	2
3	SIL	1		53	E	1		103	L	2		153	IH	2		203	TT	0
4	SIL	1		54	E	2		104	L	2		154	R	0		204	TT	1
5	SIL	1		55	L	0		105	K	0		155	R	1		205	TT	2
6	SIL	2		56	L	0		106	K	1		156	R	2		206	TT	2
7	E	0		57	L	1		107	K	2		157	R	2		207	TT	2
8	E	1		58	L	2		108	K	2		158	T	0		208	TT	2
9	E	2		59	E	0		109	K	2		159	T	0		209	AH	0
10	IH	0		60	E	1		110	UH	0		160	T	1		210	AH	0
11	IH	1		61	E	2		111	UH	1		161	T	2		211	AH	0
12	IH	2		62	AE	0		112	UH	1		162	IH	0		212	AH	0
13	R	0		63	AE	0		113	UH	1		163	IH	0		213	AH	0
14	R	1		64	AE	0		114	UH	2		164	IH	0		214	AH	1
15	R	2		65	AE	0		115	W	0		165	IH	0		215	AH	1
16	T	0		66	AE	1		116	W	1		166	IH	0		216	AH	2
17	T	1		67	AE	1		117	W	1		167	IH	1		217	AH	2
18	T	2		68	AE	1		118	W	2		168	IH	1		218	F	0
19	AE	0		69	AE	1		119	AY	0		169	IH	1		219	F	0
20	AE	1		70	AE	2		120	AY	0		170	IH	2		220	F	1
21	AE	2		71	S	0		121	AY	0		171	IH	2		221	F	1
22	F	0		72	S	0		122	AY	0		172	F	0		222	F	1
23	F	1		73	S	1		123	AY	1		173	F	0		223	F	1
24	F	2		74	S	2		124	AY	1		174	F	0		224	F	1
25	AE	0		75	S	2		125	AY	1		175	F	1		225	F	1
26	AE	1		76	H	0		126	AY	1		176	F	1		226	F	2
27	AE	1		77	H	1		127	AY	2		177	F	2		227	F	2
28	AE	1		78	H	2		128	AY	2		178	AE:	0		228	IY	0
29	AE	2		79	H	2		129	AY	2		179	AE:	0		229	IY	0
30	AI	0		80	H	2		130	T	0		180	AE:	0		230	IY	0
31	AI	0		81	UH	0		131	T	1		181	AE:	1		231	IY	1
32	AI	0		82	UH	0		132	T	1		182	AE:	2		232	IY	1
33	AI	1		83	UH	1		133	T	2		183	AE:	2		233	IY	1
34	AI	2		84	UH	2		134	IH	0		184	AE:	2		234	IY	1
35	AI	2		85	UH	2		135	IH	1		185	AE:	2		235	IY	1
36	AI	2		86	UH	2		136	IH	1		186	AE:	2		236	IY	2
37	AE	0		87	M	0		137	IH	1		187	AE:	2		237	F	0
38	AE	0		88	M	0		138	IH	1		188	AI	0		238	F	1
39	AE	1		89	M	1		139	IH	2		189	AI	1		239	F	1
40	AE	1		90	M	2		140	IH	2		190	AI	2		240	F	1
41	AE	1		91	E	0		141	Y	0		191	AE:	0		241	F	1
42	AE	1		92	E	1		142	Y	0		192	AE:	0		242	F	2
43	AE	1		93	E	2		143	Y	1		193	AE:	0		243	AE	0

44	AE	2		94	L	0		144	Y	2		194	AE:	0		244	AE	0
45	AE	2		95	L	0		145	AE	0		195	AE:	0		245	AE	1
46	T	0		96	L	0		146	AE	1		196	AE:	1		246	AE	2
47	T	0		97	L	0		147	AE	2		197	AE:	1		247	AE	2
48	T	1		98	L	1		148	H	0		198	AE:	1		248	AE	2
49	T	2		99	L	1		149	H	1		199	AE:	1		249	AE	2
50	T	2		100	L	1		150	H	2		200	AE:	2		250	AE	2

7. CONCLUSION

The paper presents our initial results of an ongoing work for developing tools for automatic segmentation of Arabic speech. The development includes building a fully transcribed corpus of Arabic TV news, and an Arabic phonetic dictionary. The developed tools are also part of the effort for developing an Arabic speech recognition system for automatic TV news transcription.

8. ACKNOWLEDMENT

The authors would like to acknowledge King Abdulaziz City for Science and Technology (KACST) for its support of this work under grant AT-24-94, and King Fahd University of Petroleum and Minerals for its support through out the execution of this project.

REFERENCES

[1] Moustafa Elshafei, Husni Al-Muhtaseb, Mansour Al-Ghamdi, "Techniques for High Quality Arabic Speech Synthesis", Information Sciences 140(3-4): 255-267 (2002).

[2] M. Elshafei Ahmed, "Toward an Arabic Text-to-Speech System", in the special issue on Arabization, the Arabian Journal of Science and Engineering, Vol. 16, No. 4B, pp.565-583, October 1991.

[3] S. Cox, R. Brady, and P. Jackson, " Techniques for accurate automatic annotation of speech waveforms", in Proceedings of the International Conference on Spoken Language Processing, Vol V., Sydney, NSW, pp. 1947-1950, 1998.

[4] F. Malfrere, O. Deroo, and T. Dutoit, " Phonetic alignment: Speech synthesis based vs. hybrid HMM/ANN", in Proceedings of the International Conference on Spoken Language Processing, Vol IV., Sydney, NSW, pp. 1571-1574, 1998.

[5] A. Ljolje, J. Hirschberg, and J.P.H. Van Santen, "Automatic speech segmentation for concatenative inventory selection", in Progress in Speech Synthesis, J.P.H. Van Santen, Ed: Springer, 1997, pp. 305-311.

[6] C.D. Mitchel, M.P. Harper, and L.H. Jamieson, " Using explicit segmentation to improve HMM phone recognition", in Proceedings of the International Conference on Acoustics Speech and Signal Processing, Vol I, pp.229-232, 1995.

[7] A. Vorstermans, J.P. Martens, and B. Van Coile," Automatic segmentation and labeling of multi-lingual speech data", Speech Comm. Vol. 19, pp. 271-293, 1996.

[8] J. Van Santen and R. Sproat, " High-accuracy automatic segmentation", Proc. EUROSPEECH, Vol. VI, Budapest, Hungary, pp. 2809-2812, 1999.

[9] Moustafa Elshafei, Husni Al-Muhtaseb, and Mansour Alghamdi, "Machine Generation of Arabic Diacritical Marks", Proceedings of the 2006 International Conference on Machine Learning; Models, Technologies, and Applications (MLMTA'06), June 2006, USA.

[10]] X.Huang, A. Acero, and H. Hon, Spoken Language Processing, Prentice Hall PTR, 2001.

[11] B. Angelini, F. Brugnara, D. Falavigna, D. Giuliani, R. Gretter, and M. Omologo, " Automatic segmentation and labeling of English and Italian speech database", in Proceeding of EUROSPEECH, pp. 653-656, 1993.

[12] A. Farbat, G. Perennou, and R. Andre-Obrecht, " A segmentation approach versus a centisecond one for automatic phonetic time-alignment," in Proceeding of EUROSPEECH, pp. 657-660, 1993.