Arabic in my Hand: Small-footprint Synthesis of Egyptian Arabic

Laura Mayfield Tomokiyo, Alan W Black, and Kevin A. Lenzo

Cepstral LLC
1801 E. Carson St.
Pittsburgh, PA 15203 USA
laura@cepstral.com

Abstract

The research described in this paper addresses the dual concerns of synthesis of Arabic, a language that has shot to prominence in the past few years, and synthesis on a handheld device, realization of which presents difficult software engineering problems. Our system was developed in conjunction with the DARPA BABYLON project, and has been integrated with English synthesis, English and Arabic ASR, and machine translation on a single off-the-shelf PDA.

We present a concatenative, general-domain Arabic synthesizer that runs 7 times faster than real time with a 9MB footprint. The voice itself was developed over only a few months, without access to costly prepared databases. It has been evaluated using standard test protocols with results comparable to those achieved by English voices of the same size with the same level of development.

1. Introduction

The value of electronic language support in geopolitically sensitive environments has been proven through DARPA projects such as DIPLOMAT [7] and TONGUES [3]. In these programs, rapid deployment of languages such as Serbo-Croatian, Korean, and Haitian Creole was investigated, with an eye toward supporting peacekeeping efforts in Haiti, the former Yugoslavia, and the Korean Peninsula.

As it has become apparent that insertion of language technologies in the field is desirable, practical issues such as size, weight and durability of the device have become a priority. World events have also brought into focus a completely new set of target languages.

The research described in this paper addresses the dual concerns of synthesis of Arabic, a language that has shot to prominence in the past few years, and synthesis on a handheld device, realization of which presents difficult software engineering problems. Our system was developed in conjunction with the DARPA BABYLON project, and has been integrated with English synthesis, English and Arabic ASR, and machine translation on a single off-the-shelf PDA.

The Arabic language offers a number of challenges for speech synthesis. In the written language, vowels are represented partially at best, and must be inferred. Naturally, this is a problem in the generation phase, when one must know what vowel is to be synthesized. It is also a problem in training. In a concatenative synthesis system such as ours, a database is ordinarily annotated at the phoneme level; one must choose between working from a traditional text and labeling only consonants, and phonetically transcribing the text in order to include vowel labels.

Speech synthesis is the "face" of a speech-to-speech translation system. Not only must it speak the text given to it intelligibly, but it must also say it in the right register, with the right gender, in the right dialect – the design of a speech synthesis system must anticipate the reaction of the target listener and adjust its output accordingly. This is *particularly* true for sensitive contexts like those faced by BABYLON.

In this paper we describe rapid development of a small-footprint Arabic voice, focusing on the challenges encountered.

2. Special Challenges of Arabic

2.1. Dialects

The variety of Arabic dialects reflect the ethnic and social diversity of its speakers. There are two main classes of Arabic dialects: the Eastern dialects of Egypt, Sudan, and the middle East, and the Western dialects of North Africa. These dialect classes are distinguished by the reduction of the vowel system in the Eastern dialects and a contrast in the stress system [10]. The dialect classes can be further broken down into Gulf, Levantine, Egyptian/Sudanese, and Maghrebi dialect groups. Between these groups we see major phonological differences in realization of specific phonemes, such as the uvular stop *qaf*, the palatal fricative *jim*, and the interdental fricatives *tha* and *dha*. Morphological differences are also significant enough that spoken Moroccan and Yemeni are mutually unintelligible.

Dialects are a concern for speech synthesis for several reasons. What dialect is to be generated? One must decide between generating MSA, which is a mother tongue for no one, and one of the dialects, risking markedness and possibly a narrower potential listener base. MSA is widely understood only by communities with formal education in it, however, so its listener base is also limited.

Our system generates Egyptian Arabic, specifically Cairene, which enjoys broad exposure because of the prominence of Egyptian entertainment media across the Arab world.

2.2. Contrast between Spoken and Written Language

Arabic speakers rely on Modern Standard Arabic (MSA) to communicate across dialect groups. MSA is not spoken as a mother tongue, but rather as a lingua franca in situations where one's native dialect will not be understood. Spoken Arabic has very little occasion to be written down. Spoken events that might ordinarily be transcribed, such as the evening news, are delivered in MSA. Newspapers, literature, and electronic texts are almost exclusively found in MSA; beyond very recent efforts to build electronic corpora in Arabic, orthographic renditions of the spoken dialect are mainly limited to comic books and other such unusual forms.

The principal difficulty in simply transcribing spoken Arabic is that the dialects lack a fixed orthography. Individual words consist of a root, a sequence of 3-4 consonants that represent a broad concept; vowel diacritics and other phonological annotations, which are usually omitted in the written form;

and morphological components. For example, the consonant sequence *ktb* represents the concept of writing, and has standalone readings such as "kutub" or "kattaba," or can add morphological components to become "aktib," "maktaba," and so forth. Because only the voweling for MSA is learned in school, speakers of the same dialect can differ significantly in their sense of which vowel is being used in the spoken language, and there is a strong tendency to write the standard orthography as prescribed by MSA even when the morphology is not the same.

The Egyptian Arabic database that we have collected (described in Section 6.2) has been spoken and transcribed by native speakers of Cairene, led by an experienced Arabic linguist. Care has been taken to remove influence from written language in both the transcription and the elicitation. The proper phonological description of individual words, however, remains an open question.

2.3. Gender Differences in Speech

Male and female speakers use many different word forms in Arabic. Certain inflectional components reflect the gender of the speaker. Arabic syntax is also affected by the gender of the listener.

Generating the appropriate grammar for a given situation is, of course, the task of the language generation module and not the synthesizer, and we have not addressed this in our system. It should be noted, however, that when speech output is the final product of a translation system, inappropriate gender marking is perhaps more obvious and unsatisfactory than it is when the system generates only text.

2.4. Voweling

As has been mentioned, normal Arabic text written for adults does not contain vowels and other phonological markings necessary to expand the orthography to a reasonably phonetic form. This is in some sense analogous to the grapheme-to-phoneme problem for English; the correct pronunciation of an English word is not often obvious from its spelling, and there are many words for which multiple pronunciations are possible. For English, however, we can rely on electronic lexicons that provide the correct pronunciation for an orthographic string. A comparable body of work does not exist for Arabic.

For synthesis, we must know what the correct vowel is. Diacritics indicating the correct MSA vowel are shown in religious texts and literature for children, and are known as the *vocalization* or the *voweling*. The process of adding all of the diacritics to an unmarked text is called *diacritization.*

There are two obvious approaches to solving the voweling problem for spoken language: inferring the vowels and enumerating the lexicon. The former has been applied with some success in recognition; vowels were guessed with 80% accuracy [6]. Synthesis requires a much higher level of accuracy than recognition, however, and we have selected the enumerative approach to voweling.

3. Related work

Clearly, determining the correct voweling is a major consideration for Arabic TTS systems. Kirchhoff et al. [6] describe an approach to automatic romanization for spontaneous speech recognition that achieves 80% token accuracy in generating the correct diacritization as estimated by comparison with manual diacritization. This is an enormous improvement over the 50% accuracy measured for commercially-available diacritizers, which are targeted toward MSA. For TTS, however, a much higher level of accuracy is required; this state-of-the-art result emphasizes the need in synthesis for manual diacritization. Even manual diacritization of dialects, however, is not unambiguous.

Although context-dependent units are generally thought to provide the most natural synthesized sound, a large number of them is needed to accurately cover the phonological and prosodic space of a language. Context-independent diphone units can provide broad coverage with relatively small storage requirements. Elshafei, Al-Muhtaseb, and Al-Ghamdi [5] argue that the degradation in coarticulatory modeling is not as severe for Arabic as for other languages partially due to its consonant-heavy structure. They describe a synthesis system for classical Arabic that uses diphones and a few other specific subsyllable units. They generate vowels automatically, but require a morpho-syntactic analyzer because the correct phonetic realization (with vowelization and consonant doubling) can only be inferred with information about word classes and dependencies. The vowel inference task is easier for classical Arabic, which is well-described and for which the effects of dialect and spontaneous speech can be ignored.

El-Imam [4] addresses the problem of vowel generation by requiring fully diacritized input text. El-Imam's system has a fixed unit inventory of 452 subphonetic units; 400 representing the basic steady-state and transition units, and 52 representing allophonic variation. Letter-to-sound rules are manually enumerated. Ben Sassi, Braham, and Balghith [9] also specify letter-to-sound rules manually in a neural network based diphone system. Each phoneme is represented by a feature vector, for which length of both vowels and consonants is an element. Diphone feature vectors are compositions of their constituent phonemes. A fully diacritized set of phonetically balanced sentences was used for training in this system.

4. BABYLON

The present work was carried out under the umbrella of BABYLON, a DARPA-driven collaboration between multiple sites to explore deployment of speech and translation technologies on portable platforms in a military/diplomatic context. Languages under investigation include Arabic (Lebanese and Egyptian), Pashto, Dari, Farsi, and Chinese.

Each participant is tasked with developing a different combination of technologies, platform, and language. The role of our team is to develop full speech-to-speech translation on a handheld device for Egyptian Arabic. That is, our device (a Compaq iPaq) hosts unrestricted ASR for both English and Arabic, a translation module, and synthesis for English and Arabic, all tuned for the medical triage domain (the domain is discussed in more detail in Section 6.1).

The eventual objective of the program is to determine which technologies and platforms show the most promise for insertion where language support is needed in sensitive environments.

5. The System

Cepstral's Theta$^{\text{TM}}$synthesizer is a unit selection synthesizer specifically developed to be optimal on a low resource device. As the BABYLON system supports two synthesizers, two recognizers and two translation engines at the same time on a device with a comparatively slow processor and only about 40M total available to our system, a small, fast engine is important.

The engine itself is designed to work on processors without floating point support; all scores and measures are done in fixed point. The core architecture of Theta$^{\text{TM}}$is based on CMU Flite [2], but the unit selection algorithm has been optimized for space and speed.

The second important aspect of the system is that of database compression. Unit selection synthesizers require large databases to provide appropriate variation. Thus we ensure that the units within the database are the most useful ones for synthesis, and we take advantage of the fact that the voice is from a single speaker and use speech compression techniques to reduce the data representation significantly, but still ensure we can unpack efficiently at run time.

6. The Voice

6.1. Domain

The target domain can be characterized as medical interviewing. Users of the system are doctors and patients in a clinic situation, with patients coming to the clinic with routine complaints such as toothaches and stomachaches, as well as more urgent problems like broken bones and wounds. The domain does not extend to full medical triage, due to the ASR-side difficulties with recognition of excited speech. In principle, though, the synthesizer we describe could be used under such conditions as well.

Our system supports both Arabic-speaking doctor / English-speaking patient and English-speaking doctor / Arabic-speaking patient situations, although realistically the latter is of greater concern.

This scenario presents an interesting challenge to our sociolinguistic model. In some Arabic-speaking communities, male doctors would not be able to treat female patients, and female doctors would not treat male patients. We have for the time being ignored this problem, and the system always speaks with a male voice, although the patient can be male or female if gender differences are supported by the translation model. In real-world use, however, we would want the user to have the option of providing a female synthesized voice when appropriate.

6.2. Corpus

Because no corpus of spoken Arabic data existed for this domain, and the cost of the commercially-prepared CALLHOME corpus was prohibitive for us, we collected a corpus from scratch.

6.2.1. Elicitation

An English corpus of data for the medical domain had already been elicited as part of our team's BABYLON effort. A selection of the English sentences were first translated by native Arabic speakers. This Arabic data was then expanded. Without looking at the source English sentences, the Arabic speakers were asked to provide up to ten possible rephrasings of each Arabic sentence, in the target Egyptian dialect. The rephrasings were generated in a verbal brainstorming session, with one speaker transcribing the sentences that were spoken in order to capture the naturalness of spoken language.

This process yielded approximately 5000 sentences. A subset was selected to maximize phonetic and prosodic coverage, and these sentences were recorded by our model speaker for the unit database.

6.2.2. Transcription

The challenges of transcribing spoken Arabic have already been described. Namely, without official voweling to fall back on, speakers must rely on their own intuitions of what vowels are being pronounced, and this intuition varies from speaker to speaker.

In order to remove some of the influence of the written language, transcribers worked with a roman alphabet. Transcription conventions were based on the LDC conventions for CALLHOME with some extensions.

Maintaining inter-coder consistency in transcription was very difficult, and required multiple iterations of hand-checking and converting to the Arabic script for verification.

6.3. Unit considerations

6.3.1. Vowel length

Arabic is generally regarded as having 28 consonants, 3 long vowels (/a:/,/i:/,/u:/), and 3 short vowels (/a/,/i/,/u/). The opposition between the three long vowels is present in all modern Arabic dialects. In Cairene, the distinction in the short vowels between /i/ and /u/ has been greatly reduced, although it still exists. Some analyses suggest that this reduction has led to the development of an enriched long vowel system, with /e:/ and /o:/ added to the inventory. We chose to limit our voweling to long and short /a/, /i/, and /u/. We found it quite difficult to achieve inter-coder agreement even with this vowel set, and the presence of an /e/ or /o/ vowel was not something that our speakers were able to identify with a great deal of consistency. This may be due to influence from MSA, where formal methods for indicating voweling exist only for /a/, /i/, and /u/. Although empirical evidence does exist of productive minimal pairs involving /e/ and /u/ for some speakers, it did not appear that lack of this opposition affected perception of synthesized Arabic.

6.3.2. Word-initial glottal stops

The glottal stop is part of the consonant inventory of Arabic, and is particularly common in Cairene because the historical uvular stop *qaf* is realized as a glottal stop. These glottal stops are, naturally, explicitly represented in the orthography. Arabic exhibits pre-vocalic glottal stops at the beginning of words, also explicitly shown in text.

Word-initial glottal stops are predictable in Arabic; they occur, just as they do in English and many other languages, when the word starts with a vowel, except in "weak" environments that are also predictable. Although word-initial glottal stops were transcribed, they were collapsed with the following phoneme in the unit database. Because only a word-initial unit will be selected for a word-initial context in synthesis, a glottal stop is never inappropriately generated in a prevocalic context, and this binding reduces the size of the unit database.

6.3.3. Epenthetic vowels

Arabic is consonant-rich, and there are many complex consonant clusters. These clusters are often simplified with an epenthetic vowel. The realization of the epenthesis varies greatly from speaker to speaker. Phonetically, it is almost always an /ɪ/, at least in Cairo. For example, the sentence "candy faqr dam gAmid" is pronounced /c a n d i f a q r d a m i g A m i d/ by all three of our speakers.

We chose to label these insertions explicitly in the recorded database, but not generate epenthetic units. The insertions are not required, and it was our observation that the residual effect of the epenthesis was often enough to give the illusion of simplification in the synthesized utterance.

6.3.4. Metathesis

The consonant cluster simplification process can also appear to take on the form of metathesis, or the transposition of segments. For example, in the sentence "baHis bi+alam bi+il+zAt sAciB(t) maSHa" the first two words are pronounced /b a H i s i b a l a m/ by all three speakers that we recorded. Another interesting feature of this example is that the apparent transposition triggers a cross-word resyllabification.

After close examination and consultation with an Arabic linguist, we have concluded that these events are not metathesis but rather a combination of epenthesis and reduction or elision.

This phenomenon occurs mainly with an /i/ following a (possibly cross-word) consonant cluster, frequently in the words “bi” and “li.” Because vowel epenthesis in Cairene is primarily realized as [ɪ]/, it *sounds* as if the vowel and its preceding consonant are simply reversed, but when the same phenomenon is examined in other dialects the simplifying vowel varies in realization, indicating that it is not a post-cluster /i/ that is pulled back, but rather an inserted simplifying vowel and an elided /i/.

We have labeled the preceding vowel as an epenthetic vowel, excluding it from the unit database as described above. The elided vowel is removed from the labels.

6.4. Voice-building Process

The voice development process for Cepstral’s Theta$^{\text{TM}}$synthesizer is based on Festival, which is well documented through the Festvox project [1]. A set of sentences which cover the phonetic, prosodic, and lexical space of the language is selected from a corpus of naturally-occurring text or transcribed spoken language. For a task such as this in which something is known of the target domain, the sentences are selected from domain-relevant text, but general-domain language such as greetings must also be included.

Initial phoneme labels are generated automatically by building database-specific acoustic models using the CMU SphinxTrain package and then forced aligning with the CMU Sphinx recognition system. The labels are then manually corrected. The manual correction in this voice was not done by native speakers of Arabic. The labeling team, while highly experienced in phonetic annotation, had no knowledge of Arabic beyond a basic introduction to the writing system and phoneme inventory. They did have access to native speakers for questions, but in most cases had very little difficulty defining boundaries and identifying speech errors and errors in the autolabeling. Most of the problems referred to native speakers involved labeling of the uvular fricative *‘ayn* and incorrect transcription of doubled consonants and vowel length.

After labels have been hand-corrected, the voice can be built and evaluated.

7. Evaluation

The Arabic voice was evaluated with Diagnostic Rhyme (DRT)[8] and Modified Rhyme (MRT) tests, and with a sentence-level test in which listeners were asked to mark any words that sounded bad. The DRT asks listeners to choose which of a pair of monosyllabic words that differ only in one feature in the word-initial phoneme is being synthesized. For example, a test for voicing might include the words “tart” and “dart.” The MRT presents the listener with a list of monosyllabic words that differ in either the first or last phoneme, asking them to choose which word is being synthesized.

Standard DRT distinctive feature categories are defined for English and include voicing, nasality, sustenation, sibilation, graveness and compactness. The DRT and MRT categories were adapted to include important distinctions in Arabic; emphaticness was added as a diagnostic class. Gemination was *not* tested; it was our experience that speakers varied considerably in their perception of geminate consonants.

The score of a test was defined to be

$$\frac{\text{number of items guessed correctly}}{\text{total number of items}}$$

Results are shown in Table 1, and are comparable to English voices of the same size and degree of development.

DRT	MRT	Sentence
78.3	72.0	84.7

Table 1: Evaluation results for the Arabic voice

8. Conclusion

We have presented a small-footprint speech synthesizer for Egyptian Arabic that runs on a PDA with a size of 9MB. The synthesizer has been evaluated with standard word-level and sentence-level tests with results comparable to those achieved for English voices of similar size and level of development (a few months to build a new-language voice from scratch).

We have discussed some of the challenges encountered, describing our solutions to them. Challenges include vowelization of undiacritized text, synthesis of spoken dialect, and engine optimization.

Promising future work includes experimenting with vowel processing, possibly treating short vowels as unit features, or automatic diacritization optimized for speech synthesis.

9. References

[1] A. Black and K. Lenzo. Building voices in the Festival speech synthesis system. http://festvox.org/bsv/, 2000.

[2] A. Black and K. Lenzo. Flite: a small fast run-time synthesis engine. In *4th ESCA Workshop on Speech Synthesis*, Scotland., 2001.

[3] Alan W Black, Ralf D. Brown, Robert Frederking, Rita Singh, John Moody, and Eric Steinbrecher. TONGUES: Rapid Development of a Speech-to-Speech Translation System. In *Proceedings of HLT-2002: Second International Conference on Human Language Technology Research*, March 2002.

[4] Yousif A. El-Imam. An unrestricted vocabulary arabic speech synthesis system. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 37(12):1829–1845, 1989.

[5] Moustafa Elshafei, Husni Al-Muhtaseb, and Mansour Al-Ghamdi. Techniques for high quality arabic speech synthesis. *Information Sciences*, 140:255–267, 2002.

[6] Katrin Kirchhoff et al. Novel speech recognition models for arabic. Technical report, Johns Hopkins University, 2003.

[7] R. Frederking, A. Rudnicky, and C. Hogan. Interactive speech translation in the diplomat project. In *Proceedings of the Spoken Language Translation Workshop at ACL*, 1997.

[8] J. Logan, B. Greene, and D. Pisoni. Segmental intelligibility of synthetic speech produced by rule. *Journal of the Acoustical Society of America*, 86(2):566–581, 1989.

[9] Sihem Ben Sassi, Rafik Braham, and Abdelfattah Belgith. Neural speech synthesis system for arabic language using celp algorithm. In *Proc. IEEE Conference on Computer Systems and Applications*, pages 119–121, 2001.

[10] Janet C.E. Watson. *The Phonology and Morphology of Arabic*. The Phonlogy of the World’s Languages. Oxford University Press, 2002.