Diphone Speech Synthesis System for Arabic Using MARY TTS

M. Z. Rashad [1], Hazem M. El-Bakry [2], Islam R. Isma'il [2]
[1] Department of Computer Science, [2] Department of Information Systems
Faculty of Computer and Information Sciences, Mansoura University, Egypt
magdi_12003@yahoo.com, helbakry20@yahoo.com, islamcis@yahoo.com

Abstract

Concatenative speech synthesis systems generate speech by concatenating small prerecorded speech units which are stored in the speech unit inventory. The most commonly used type of these units is the diphone which is a unit that starts at the middle of one phone and extends to the middle of the following one. Diphones have the advantage of modeling coarticulation by including the transition to the next phone inside the diphone itself. In this paper, a diphone speech synthesis system for the Arabic language using MARY TTS has been developed and evaluated by two types of tests which are the Diagnostic Rhyme Test (DRT) that measures the intelligibility of the synthesized speech and the Categorical Estimation (CE) test that measures the overall quality of the synthesized speech. The results of these tests are illustrated in the experiments and results section.

Keywords: speech synthesis**,** concatenative synthesis**,** diphone inventory, natural language processing, Markup language, digital signal processing.

1. Introduction

Speech synthesis is the automatic generation of speech (acoustic waveforms) from text [1]. The fundamental difference between text-to-speech synthesizer and any other talking machine (e.g., a cassette-player) is that we are interested in the automatic production of new sentences [2]. This mapping of text into speech is performed in two phases usually called as high- and low-level synthesis. The first one is the high-level synthesis also called text analysis, where the input text is transcribed into a phonetic representation, and the second one is the generation of speech waveforms, where the acoustic output is produced from this phonetic and prosodic information.

There are three main approaches to speech synthesis: articulatory synthesis, formant synthesis, and concatenative synthesis. Articulatory synthesis generates speech by direct modeling of human articulator behavior. Formant synthesis models the pole frequencies of speech signal. Formants are the resonance frequencies of the vocal tract. Since the formants constitute the main frequencies that make sounds distinct, speech is synthesized using these estimated frequencies. On the other hand, concatenative speech synthesis produces speech by concatenating small, prerecorded units of speech, such as phonemes, diphones, and triphones to construct the utterance. The unit length affects the quality of the synthesized speech. With longer units, the naturalness increases, less concatenation points are needed, but more memory is needed and the number of units stored in the database becomes very numerous. With shorter units, less memory is needed, but the sample collecting and labeling techniques become more complex.

The most widely used units in concatenative synthesis are diphones. A diphone is a unit that starts at the middle of one phone and extends to the middle of the following one. Diphones have the advantage of modeling coarticulation by including the transition to the next phone inside the diphone itself. The full list of diphones is called diphone inventory, and once determined, they need to be found in real speech. To build the diphone inventory, natural speech must be recorded such that all phonemes within all possible contexts (allophones) are included, then diphones must be labeled and segmented. Once the diphone inventory is built, the pitch and duration of each diphone need to be modified to match the prosodic part of the specification.

2. Architecture of MARY TTS

MARY stands for Modular Architecture for Research on speech sYnthesis and it is a tool used for research, development and teaching in the field of text-to-speech [3]. The modular design of the MARY system has many advantages: it is easy to modify a certain module without affecting other modules; any module with a similar interface can be used instead or in addition to an employed module. To integrate two modules, the only requirement is that the output data types of the first module must match the input data types of the second module.

Figure 3 shows the processing modules of the MARY system and the output of each module. As shown in the figure, the MARY TTS accepts two types of input text plain text and markup text (such as SABLE-annotated text or SSML-annotated text). The use of SABLE-annotated text for example as input gives the users the ability to add information to the text that improve the way it is spoken such as pausing at the right places or emphasizing on certain words [4]. This input text is embedded into MaryXML document for the following processing steps. MaryXML is an internal, low-level markup that reflects the information provided and required by the modules of the MARY system. In other words, MaryXML markup enables the user not only to display intermediate processing results but also to modify this results so that the user can know the influence of a specific piece of information on the output of a given processing step.

The processing modules of the MARY TTS can be grouped into four parts: the preprocessing or text normalization, the natural language processing, the calculation of acoustic parameters and the synthesis.

2.1 Text normalization

The processing modules that perform text normalization include the tokenizer, abbreviation expansion, and numeral expansion. The tokenizer divides the input texts into tokens. Each token is enclosed by a <t>….</t> MaryXML tag and subsequent processing modules add additional information to that tag as attribute-value pairs. A group of tokens that makes a sentence are enclosed by the <s>….</s> tag. The tokens may include non-standard words such as numbers and abbreviations. Determining the correct pronunciation of these types of tokens is not straightforward as the type of the non-standard word must be determined first [5]. For example, a number may be an ordinal number or a cardinal number. It may be a measure or a part of a date. An abbreviation may be pronounced as a letter sequence, as a whole word or it may need expansion.

2.2 The natural language processing

The purpose of this phase is to transcribe the input text into phonetic representation and to find prosodic information that make the pronunciation of the sentences flow naturally. The modules that perform natural language processing include part-of-speech tagger and chunker, lexicon, letter-to-sound, prosody, and postlexical phonological rules. The part-of-speech tagger helps to disambiguate many of the homographs which are words that have the same spelling but different pronunciation. To find the pronunciation of each token, a large pronunciation lexicon is used. For tokens that are not found in that lexicon, a set of letter-to-sound rules can be used to find their pronunciations. The output of these modules is a phonemic transcription added to each token as well as the source of this transcription.

2.3 The calculation of acoustic parameters

This module generates an acoustic parameter file by applying duration rules (Klatt rules) and intonation realization rules (ToBI based approach). This parameter file is used as input to the synthesizer so that its format must be compatible to the type of the synthesizer used. MARY uses MBROLA diphone synthesizer so that every phone symbol is assigned a duration in milliseconds and some phones are assigned a time and frequency where time is in percent of the phone duration and frequency is in Hertz.

2.4 The synthesizer

The synthesizer creates a sound file based on the output of the preceding module. MBROLA is used for diphone synthesis and the synthesizer also contains basic unit selection code, based on the cluster unit selection code, derived from FreeTTS [6].

3. Special challenges of the Arabic language

The Arabic phoneme set consists of 28 consonants, 3 short vowels, and three long vowels [7]. Arabic short vowels are written with diacritics placed above or below the consonant that precedes them. The Arabic letters are written from right to left and most of them are attached to one another. Most Arabic words can be reduced to a root which often consists of three letters. Modifying this root by adding prefixes and/or suffixes and changing the vowels results in many word patterns. For example, modifying the vowels inside the verb is used to convert it from the active form into the passive form [8].

3.1 Diacritization

Diacritization is the process of adding vowels to an unmarked text. The Arabic text written in newspapers, scientific or literature books does not contain vowels and other markings needed to pronounce the text correctly. Vowels are added to the text only in the cases where ambiguity appears and cannot be resolved from the context otherwise writers assume that the reader has enough knowledge of the language that enables him to infer the correct vowels.
One approach to solve the diacritization problem is to implement a module for automatic vowelization. Al-Ghamdi et al. followed that approach [9]. Since the accuracy of automatic vowelization is not high and speech synthesis requires a higher accuracy than

speech recognition, some authors such as El-Imam require that the input text be fully diacritized [10].

3.2 Dialects

Arabic is spoken in more than 20 countries by more than 300 million people so that there are different dialects of the Arabic language that reflect the social diversity of its speakers. The Arabic dialects include the Egyptian/Sudanese dialect, the Gulf dialect, the Levantine dialect, and the western dialect of North Africa. This diversity of the Arabic dialects is considered a problem for speech synthesis form many reasons. First, what dialect is to be generated? A choice must be done between generating Modern Standard Arabic (MSA) and one of the dialects. Second, MSA is understood by people with a high level of education so that its listener base is limited [11].

3.3 Differences in gender

The Arabic speech is influenced by the gender of the person the speech is directed to or is about. For example, the imperative form of the word (تَبَسَّمَ) "ta-bas-sa-ma" which means he smiled depends on the gender of the listener. If the speech is directed to a male, the masculine form (تَبَسَّمْ) " ta-bas-sam " is used on the other hand, if the speech is directed to a female, the feminine form (تَبَسَّمِي) " ta-bas-sa-mi " is used. As a consequence to that when the speech is the final product of a system such as a translation system or a synthesis system, inappropriate gender marking is more obvious and unsatisfactory than it is when the system generates only text.

4. Adding support for a new language to MARY TTS

Two tasks are required to add support for a new language to MARY TTS: the first task is constructing a minimal set of natural language processing (NLP) components for the new language. In this task some kind of script is applied on a voluminous body of encoded text in the target language, such as an XML dump of the Wikipedia in the target language to extract the actual text without markup and the most frequent words, and then a pronunciation lexicon has to be built up. Using MARY transcription tool, which supports a semi-automatic procedure for transcribing new language text and automatic training of letter-to-sound rules for that language, many of the most frequent words has to be manually transcribed then a ' trainpredict ' button in the GUI is used to automatically train a simple letter-to-sound algorithm and predict pronunciations for the untranscribed words in the list. To be able to use the MARY transcription tool, an XML file describing the allophones that can be used for transcription and providing for each allophone the phonetic features that are to be used for characterizing the phone later is needed. The second task is the creation of a voice in the target language and the ' redstart' voice recording tool can be used for this purpose [12].

5. Experiments and Results

Speech quality can be measured by many tests each of which focuses on a certain aspect of the speech. There is no single test that is said to provide the correct results. The two most important criteria that are measured when evaluating a synthesized speech are the intelligibility and the naturalness of the speech. The intelligibility of the speech means whether or not the synthesizer's output could be understood by a human listener while the naturalness of the speech means whether or no the synthesized speech sounds like the human speech. The feeling of naturalness about speech is based on a complex set of features. For example, the naturalness scoring introduced by Sluijter et al. enumerated eleven parameters which listeners are asked to consider on five-point scales [13].
Two types of tests were applied two evaluate the speech of the developed system regarding the intelligibility and the naturalness aspects. The first test which measures the intelligibility is the Diagnostic Rhyme Test (DRT). In this test, twenty pairs of words that differ only in a single consonant are uttered and the listeners are asked to mark on an answer sheet which word of each pair of the words they think is correct [14]. In the second evaluation test, which is Categorical Estimation (CE), the listeners were asked a few questions about several attributes such as the speed, the pronunciation, the stress, etc. [15] of the speech and they were asked to rank the voice quality using a five level scale.
The test group consisted of sixteen persons and the previously mentioned two tests were repeated twice to see whether or not the test results will increase by the learning effect which means that the listeners may become accustomed to the synthesized speech they hear and they understand it better after every listening session [16]. The following tables and charts illustrate the results of these tests.

Table 1: Perception of the test words

Word no.	1	2	3	4	5	6	7	8	9	10
No. of respondents	14	14	13	15	16	15	12	14	13	16
Percent of respondents	87.5	87.5	81.2	93.7	100	93.7	75	87.5	81.2	100
No. of respondents	15	14	13	15	16	15	13	14	14	16
Percent of respondents	93.7	87.5	81.2	93.7	100	93.7	81.2	87.5	87.5	100

Word no.	11	12	13	14	15	16	17	18	19	20
No. of respondents	15	15	15	16	13	12	14	14	15	15
Percent of respondents	93.7	93.7	93.7	100	81.2	75	87.5	87.5	93.7	93.7
No. of respondents	16	15	14	16	13	12	14	15	16	15
Percent of respondents	100	93.7	87.5	100	81.2	75	87.5	93.7	100	93.7

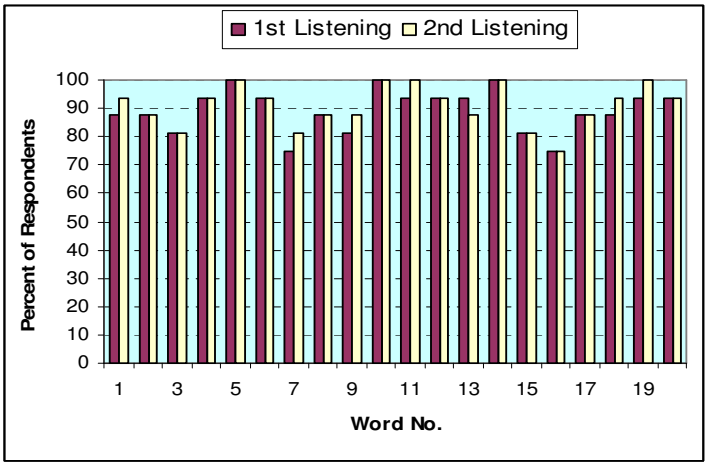

Figure 1: Perception of the test words

As shown in the previous table and diagram, the listener identifies 89.3% of the test words in the 1st listening session and this percent increases slightly and becomes 90.9% in the 2nd listening session.

Table 2: the overall quality assessment

Scale	1	2	3	4	5
Naturalness	0.19	0.50	0.25	0.06	0.00
Pronunciation	0.13	0.31	0.31	0.19	0.06
Speed	0.00	0.19	0.62	0.13	0.06
Clarity	0.00	0.13	0.50	0.31	0.06

The five-level scale shown in the previous table is enumerated from 1 to 5 where 1 represents the most negative indicator, 2 is less negative and so on until 5 which represents the most positive indicator. For example, the explanation of the five-level scale for the speed attribute is 1 means too slow, 2 means slow, 3 means normal speed, 4 means fast, and 5 means too fast. The results of the 1st and the 2nd listening session are the same so that only on set of the results are shown in the table and the diagram.

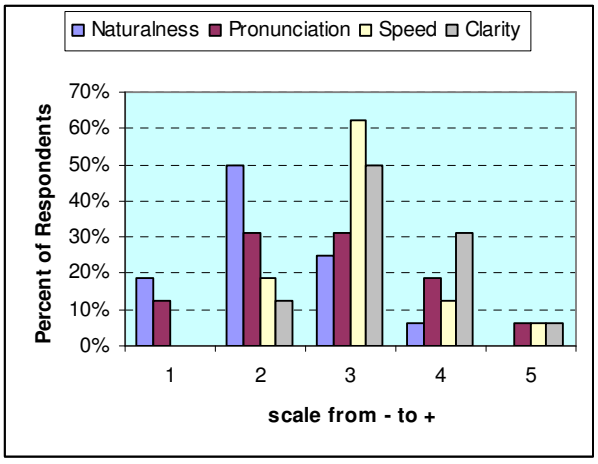

Figure 2: the overall quality assessment

6. Conclusion and Future Work

A modified MARY TTS for Arabic language has been presented. An XML file that contains the consonants and vowels of the Arabic phonemes has been constructed. Furthermore, the characteristics of these phonemes have been described. The pronunciation of the most frequent Arabic words has been provided. In order to conclude the relation between the words and their pronunciations, the modified MARY TTS has been trained for these words. Two tests have been implemented. It has been found that the result of the DRT is approximately 91% which can be considered a satisfactory percentage. On the other hand, the results of the CE need more improvements. Better results can be achieved if multiple instances of each speech unit are stored in the unit inventory and the instance that best match the context according to some type of cost is chosen. The focus of the future work will be on that technique which is the unit-selection concatenative speech synthesis as it produces speech that is closest to nature.

7. References

[1] Klatt D. H., "Review of text-to-speech conversion for English", Journal of the Acoustical Society of America, vol. 82(3), 1987.

[2] Thierry Dutoit, "High-Quality Text-to-Speech Synthesis: an Overview", Journal of Electrical & Electronics Engineering, Australia: Special Issue on Speech Recognition and Synthesis, vol. 17, pp. 25-37, 1999.

[3] M. Schroder and J. Trouvain, “The German text-to-speech synthesis system MARY: A tool for research, development and teaching”, International Journal of Speech Technology, vol.6, pp.365–377, 2003.

[4] Sproat R., Hunt A., Ostendorf M., Taylor P., Black A., Lenzo K. and Edgington M., "SABLE: A Standard for TTS Markup", Proc. ICSLP Sydney, pp. 1719-1724, 1998.

[5] Sproat R. et al., "Normalization of non-standard words", Computer Speech & Language, vol. 15(3), pp. 287-333, 2001.

[6] http://www.freetts.sourceforge.net/

[7] Ibraheem Anees, "Al-Aswat Al-Arabia", Anglo-Egyptian Publisher, Egypt, 1990.

[8] J. Haywood and H. Ahmad "A new Arabic grammar", Lund Humphries, London, 2003.

[9] Husni Al-Muhtaseb, Moustafa Elshafei and M. Al-Gamdi "Techniques for high quality Arabic speech synthesis", Information Sciences, vol. 140,pp. 255–267, 2002.

[10] Yousif A. El-Imam, "An unrestricted vocabulary Arabic speech synthesis system", IEEE Transactions on Acoustics, Speech and Signal Processing, vol. 37(12), pp. 1829–1845, 1989

[11] Laura Mayfield Tomokiyo, Alan W Black and Kevin A. Lenzo, "Arabic in my Hand: Small-footprint Synthesis of Egyptian Arabic", Proceedings of the Eurospeech'03, pp. 2049-2052, 2003.

[12] M. Schrder et al., "Multilingual MARY TTS participation in the Blizzard Challenge 2009", In Blizzard Challenge, Edinburgh, UK, 2009.

[13] Sluijter A., Bosgoed E., Kerkhoff J., Meier E., Rietveld T., Swerts M. and Terken J., "Evaluation of speech synthesis systems for Dutch in telecommunication applications", Proceedings of the 3rd ESCA/COCOSDA Workshop of Speech Synthesis, 1998.

[14] Maria M., "A Prototype of an Arabic Diphone Speech Synthesizer in Festival", Master Thesis in Computational Linguistics, Uppsala university, 2004.

[15] Kraft V., Portele T., "Quality Evaluation of Five German Speech Synthesis Systems" Acta Acustica 3, pp. 351-365, 1995.

[16] Karlsson I., Neovius L., "Speech Synthesis Experiments with the GLOVE Synthesizer", Proceedings of Eurospeech, vol. 2, pp. 925-928, 1993.

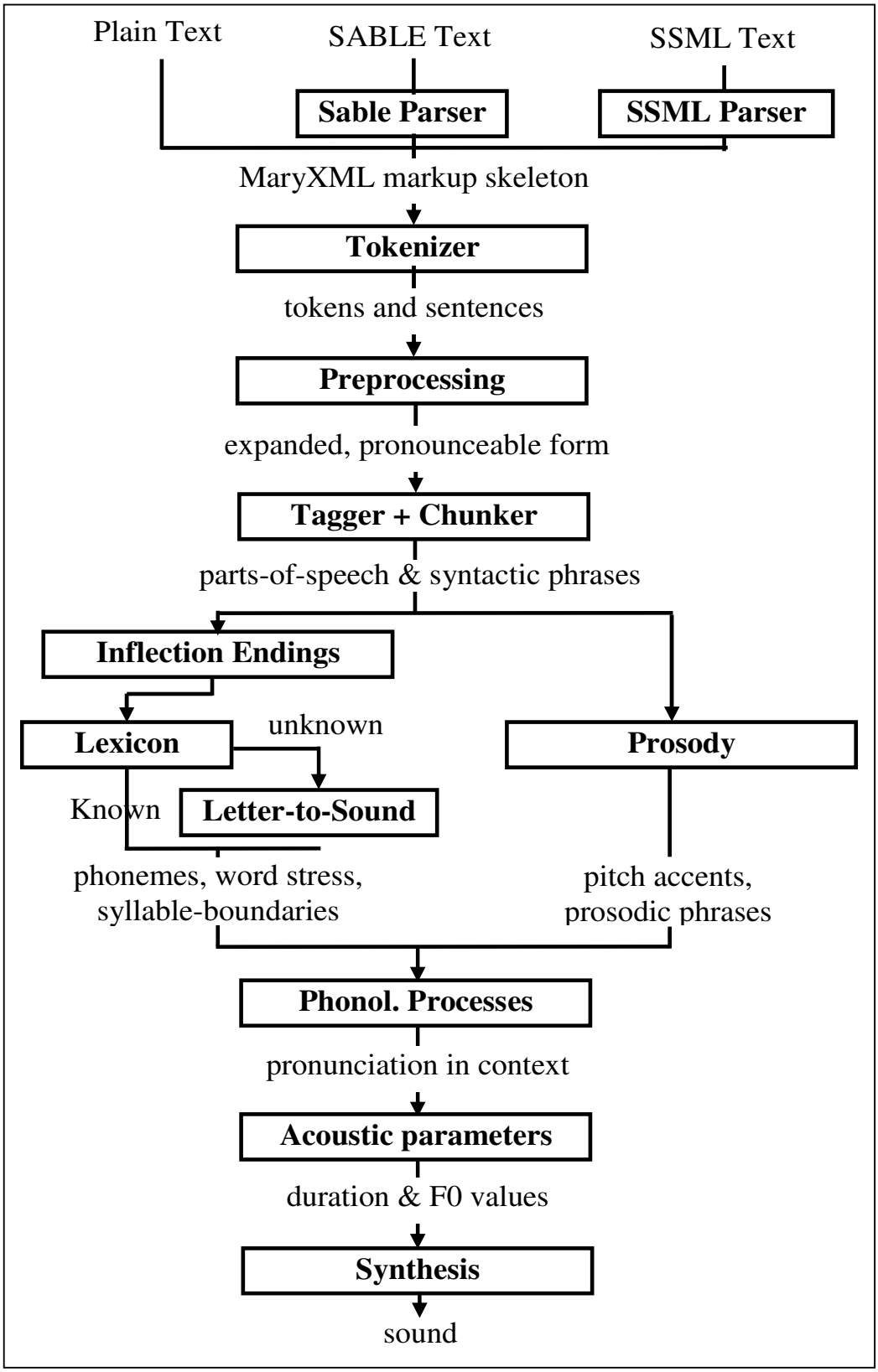

Figure 3. Architecture of MARY TTS