AUTOMATIC SPEECH SYNTHESIS FOR ARABIC LANGUAGE USING THE GENERATIVES SCHEMES METHOD

Chegrani Lamari, Guerti Mhania, Boudraa Bachir

Algeria

ABSTRACT

This study aims to create language units for an Arabic speaker synthesizer based on the predefined schemes for the syllabic structures of Arabic. The focus of this research is to design a system for spoken language assistance for the blind people in Arab countries. Simple verbs, names and particles are within our reach. Moreover, with only 84 sub-syllables, we can produce speech at different levels of complexity, including syllable, word, sentence, or text level based on the different schemes devised.

KEYWORDS

text-to-speech; Arabic scheme; speech synthesis; concatenative synthesis; generated scheme; generation of Sequence.

1. INTRODUCTION

Humans primarily communicate through speech, which is the production of word sounds. In the last few decades, research on speech synthesis focused on making people with disabilities understand recognizable synthesizers and automation has made it possible to produce mechanized computers capable of producing understandable synthetic speech. The quest for a talking machine started in the eighteen century, and although modern speech synthesizers have made significant progress in their quality of articulation, they still lack the issues pertaining to the quality and naturalness of sound produced. Such research makes speech synthesis a crucial area of development in primary languages including Arabic. Because of various applications, the speech synthesizer or Text-To-Speech (TTS) technology is one of the most revolutionary inventions today. It can be found in human-computer interfaces in multi media softwares and apparatuses for device usage, is used for reading text of emails or sms on mobile phones, in computers for reading emails and text documents, etc. It is handy for most blind people as a basic reading machine, gives opportunities for communication to deaf and speech disabled people who are not familiar with sign language, and can be used in many educational activities such as speech therapy and teaching pronunciation for foreign languages. With over 437 million speakers around the globe, Arabic ranks as the most spoken language in the world, belonging to 25 countries [1]. Alongside this, the language also holds religious significance for more than 1.6 billion Muslims [2].

Moreover, the Arab world has an excess of 5 million visually impaired people [3]. It is important that an automated accurate and simple TTS system is developed for the Arabic language so that these people can communicate using SMS, email, webpages, etc. A speech synthesizer consists of a Digital Signal Processing (DSP) module and a text processing unit, which are its main two portions. The former does two critical functions. First, it converts "unformatted" text containing

Bibhu Dash et al: IOTBC, NLPAI, BDML, EDUPAN, CITE - 2025
pp. 41-49, 2025. IJCI – 2025

DOI:10.5121/ijci.2025.140203

symbols such as numbers and acronyms to its more readable form. This procedure is also referred to as text normalization. Then, it provides the text's data in another symbolic form to the DSP or synthesizer, which changes the symbols into speech. Different means and various languages have been synthesized by many researchers to produce speech. In 1987 NETtalk, Sejnowsky and Rosenberg built a neural network learning to speak English text. This system was developed using numerous parallel network systems capable of capturing various regularities and exceptions of English pronunciation and so facilitated transforming English text strings into phonetic representations [4]. Karaali et al. [5] developed a rule-based system that consists of two neural networks. The time-delay neural network converts the phonetic representation of speech into the acoustic representation and generates speech. The second neural network is applied for output timing control. Concatenation-based speech synthesis uses the speech inventory by choosing units and algorithms for joining them along with some kind of processing, such as smoothing the borders between the concatenated parts. [6].

Telecommunications serviced at present time is the primary market for such techniques. These types of services exemplify scenarios that use speech synthesis as an important means for a computer system to transmit information to the user.Our work centers on the development of speech synthesis system based on symmetry technique in generating formant transition patterns using a dictionary of 84 different acoustic units consisting only of [CV]-type sound units (Consonant-Vowel).Based on mathematical models, speech sequences are generated into small, medium or large unit sequence varieties such as syllable sequence, word sequence and sentence sequence.The work of this article reports the process of text-to-sound sequence conversion, first from small to exponentially large sequences of sounds.Thus, for example in creating a sentence sequence, we first construct a syllable sequence, followed by a word sequence, and lastly a sentence sequence; with smooth increments in quality at every transition upward from syllable to sentence.

2. METHOD OF GENERATING SOUNDS FROM GIVEN SEQUENCES

We put forth the method of construction schemes and carrier envelopes of formant transitions. We present definitions concerning this method and the generative forms of the morphological units of the Arabic language. Then, we expose the improved method, which includes the formant transitions that ensure continuity and intelligibility of the synthetic speech as well as naturalness. In this appendix, two methods for generating schemes are presented: the superposition technique and the [CVCC]-enhanced generation method.

3. TECHNIQUE OF SUPERPOSITION

For the determination of the number of null points down and the perfect formation for the speech sounds synthesis having adequate number of variables corresponding to the written expressions of sequences of the Arabic language, we are trying here to apply a new idea of creating a generative form having no discontinuity points. Such an approach being based upon the principle of superposition would come up with a new technique for automatic synthesis of Arabic speech called-devising the best way of scheming. The principle of superposition applied here serves to limit the number of successive variables that are assumed to represent a quantity equal to the number of phonemes denoting a given sequence, which has created a number of null points reaching sometimes 3 in the process of concatenation. This approach, however, has the advantage of producing very specific combinations of variables equated to the scenario we are planning. For the algebraic sum of the two vectors to be equivalent to a third, the first and second vectors need to have equal number of components such that each of the components of an arbitrary placement in the first vector is summed together with a corresponding component of the other vector. The property of a vector that illustrates the algebric addition is known as longitudinal concatenation.

In this method we will make the superposition of the several vectors that contain values concerning a specific sound. The results obtained are the following:

- In the case of two consonants betweenV_n and V_{n+1}, we make a superposition of the variable C_{nnn} with C_{n+1} of the generative form.
- In the case of a single consonant between Vn and Vn+1, we make the superposition of the variable Cnnn with Vn+1 and Cnn with Cn+1 of the generative form.
- In the case of no consonant between V_n and V_{n+1}, we make the superposition of the variables V_n, C_{nn} and C_{nnn} with C_{n+1}, V_{n+1} and $C_{n+1,n+1}$ of the generative form.

The sum of the two successive vectors is made depending on the following possible cases:

For generating the sequence S = [V**C**+**C**V]:

$$f_S(C)=\frac{C_1}{}\left(\frac{V_1|C_{1,1}}{}\left|\frac{C_{1,1,1}}{C_2}\right|\frac{}{V_2}\right)\frac{}{C_{2,2}|C_{2,2,2}}$$

For generating the sequence S = [V+**C**V]:

$$f_S(C)=\frac{C_1}{}\left(\frac{V_1}{}\left|\frac{C_{1,1}}{C_2}\right|\frac{C_{1,1,1}}{V_2}\right)\frac{}{C_{2,2}|C_{2,2,2}}$$

For generating the sequence S = [**V**+**V**]:

$$f_S(C)=\frac{C_1}{}\left(\frac{V_1}{C_2}\middle|\frac{C_{1,1}}{V_2}\right)\left.\frac{C_{1,1,1}}{C_{2,2}}\right|\frac{}{C_{2,2,2}}$$

From these examples showing how to use superimposed variables to form the identical shape of the same written sequence, we can use these produced sequences to concatenate all the cases resulted from the Arabic language.

4. THE IMPROVED METHOD OF GENERATION BY [CVCC]

The generation of schemes is done according to generation rules, just as before, but we will find the classification it relates to the cases studied, corresponding exactly to the number of constructive variables of the generative form; that is, the variables (Cn, Cnn, and Cnnn) represented respectively by (F, ʕ, and L). We distinguish three cases as follows:

1. The case of two consonants between V_n et V_{n+1}: in this case we have two probabilities:

   a. $C_{n+1} = 0$, The resulting scheme is of the following form:

$$f(C)=C_1V_1C_{11}C_{111}V_2C_{22}C_{222}$$

In the case of Fat’ha [V=a]:

$$\begin{array}{rl|c|c|c|c|c|c|}
f(C) = & C_1 & \mathrm{a} & C_{11} & C_{111} & \mathrm{a} & C_{22} & C_{222} \\
Sch(f) = & \mathrm{F} & \mathrm{a} & \varsigma & \mathrm{L} & \mathrm{a} & \varsigma & \mathrm{L}
\end{array}$$

Fat’ha	Dhamma	Kasra
FaʕLaʕL	**FuʕLuʕL**	**FiʕLiʕL**

b. $C_{nnn} = 0$, The resulting scheme is of the following form:

$$f(C) = C_1V_1C_{11}C_2V_2C_{22}C_{222}$$

In the case of Fat’ha [V=a]

$$\begin{array}{rl|c|c|c|c|c|c|}
f(C) = & C_1 & \mathrm{a} & C_{11} & C_2 & \mathrm{a} & C_{22} & C_{222} \\
Sch(f) = & \mathrm{F} & \mathrm{a} & \varsigma & \mathrm{F} & \mathrm{a} & \varsigma & \mathrm{L}
\end{array}$$

Fat’ha	Dhamma	Kasra
FaʕFaʕL	**FuʕFuʕL**	**FiʕFiʕL**

2. The case of a single consonant between Vn and Vn+1: in this case we have three probabilities:

a. $(C_{nnn}, C_{n+1})=(0,0)$, the resulting form is as follows:

$$f(C) = C_1V_1C_{11}V_2C_{22}C_{222}$$

In the case of Fat’ha, the scheme is of the following form:

$$\begin{array}{rl|c|c|c|c|c|}
f(C) = & C_1 & \mathrm{a} & C_{11} & \mathrm{a} & C_{22} & C_{222} \\
Sch(f) = & \mathrm{F} & \mathrm{a} & \varsigma & \mathrm{a} & \varsigma & \mathrm{L}
\end{array}$$

Fat’ha	Dhamma	Kasra
FaʕaʕL	**FuʕuʕL**	**FiʕiʕL**

b. $(C_{nn}, C_{n+1})=(0,0)$, the resulting form is as follows:

$$f(C) = C_1V_1C_{111}V_2C_{22}C_{222}$$

In the case of Fat’ha, the scheme is of the following form:

$$\begin{array}{rl|c|c|c|c|c|}
f(C) = & C_1 & \mathrm{a} & C_{111} & \mathrm{a} & C_{22} & C_{222} \\
Sch(f) = & \mathrm{F} & \mathrm{a} & \mathrm{L} & \mathrm{a} & \varsigma & \mathrm{L}
\end{array}$$

Fat’ha	Dhamma	Kasra
FaLaʕL	**FuLuʕL**	**FiLiʕL**

c. $(C_{nn}, C_{nnn})=(0,0)$, the resulting form is as follows:

$$f(C) = C_1V_1C_2V_2C_{22}C_{222}$$

In the case of Fat’ha:

$$\begin{array}{r|c|c|c|c|c|c|}
f(C) = & C_1 & \mathrm{a} & C_2 & \mathrm{a} & C_{22} & C_{222} \\
Sch(f) = & \mathrm{F} & \mathrm{a} & \mathrm{F} & \mathrm{a} & \varsigma & \mathrm{L}
\end{array}$$

Fat'ha	Dhamma	Kasra
FaFaʕL	**FuFuʕL**	**FiFiʕL**

3. The case of no consonant between V_n and V_{n+1}: In this case we have only one probability:

   a. $(C_{nn}, C_{nnn}, C_{n+1}) = (0, 0, 0)$, the resulting form is as follows:

$$f(C) = C_1 V_1 V_2 C_{22} C_{222}$$

In the case of Fat'ha, the scheme is of the following form:

$$\begin{array}{r|c|c|c|c|c|}
f(C) = & C_1 & \mathrm{a} & \mathrm{a} & C_{22} & C_{222} \\
Sch(f) = & \mathrm{F} & \mathrm{a} & \mathrm{a} & \varsigma & \mathrm{L}
\end{array}$$

Fat'ha	Dhamma	Kasra
FaaʕL	**FuuʕL**	**FiiʕL**

In the case of long vowels, we can also cite the generation of diphthongs, which have two forms in the Arabic language: [au] and [ai], these generated schemes are:

Fat'ha +Fat'ha	Fat'ha +Dhamma	Fat'ha +Kasra
FaaʕL	**FauʕL**	**FaiʕL**

The trouble with generating a scheme of a sentence lies at the very end of the sentence, where we will generatively either end with (CnCnnn) or, for generated schemes, with (ʕL), where the insertion of pauses at the end of that sentence is very necessaryˆso that those we insert correspond whichever way to the generative orders.

4.1. Speech Generation Results

Hereafter, we will take a look at whether a certain method of generating a syllable, but that does not exist in the list of syllables whilst earlier segmentation controls indicated its existence from the corpus studied: both have inherent formant transitions. The interest in this study has to do with the types of [CV] syllables-perhaps, while few phonemes exist that would constitute these syllables and not appear on the list, they continue to pose a challenge to generating all possible speech combinations according to the standards of Arabic. Below in table 1 are cited results from 10 listeners having heard 13 sentences created with the approach synthesis by produced schemes. Table 1 sets forth listeners A1 through A10 by symbolic representation.

Table 1: hearing result for each listener and for each sentence [7]

Phrase	**A1**	**A2**	**A3**	**A4**	**A5**	**A6**	**A7**	**A8**	**A9**	**A10**	**Taux**
Ph. 01	1	1	1	1	1	1	1	1	1	1	100%
Ph. 02	0.5	1	1	1	0.5	0.5	0.5	0.5	0.5	1	75%
Ph. 03	1	1	1	1	1	1	1	1	1	1	100%
Ph. 04	0.65	0.65	0.65	0.65	1	0.65	0.65	0.65	0.65	1	72%
Ph. 05	1	1	1	1	1	1	1	1	1	0.7	97%
Ph. 06	0.8	0.8	1	1	1	0.8	1	1	0.8	1	92%
Ph. 07	0.5	0.5	1	0.5	0.5	0.5	0.5	0.5	0.5	1	60%
Ph. 08	1	1	1	1	1	1	1	1	0.75	0.75	95%
Ph. O9	1	1	1	1	1	1	1	1	1	1	100%
Ph. 10	1	1	1	1	1	1	1	1	1	1	100%
Ph. 11	1	1	0.75	1	1	1	1	1	0.75	1	95%
Ph. 12	1	1	1	1	1	1	1	1	1	1	100%
Ph. 13	1	1	1	1	1	1	1	1	0.75	1	97.5%
Taux	0,88	0,92	0,95	0,93	0,92	0,88	0,90	0,90	0,82	0,96	91%

With the exception of the second sentence, which has a rate of 75% compared to the content of the written sentence sequence, we can see in Table 1 that sentences 4 and 7 are misidentified with a rate lower than 75%. This error was caused by the use of a sound segment of the type [ha[reversed of bad segmentation, and of a nature close to ^{2}[Ha[, which creates a bad generation for the word [ˀaðhabu] and to hear it as [ˀaðHabu]. With the help of the repair, a new [ha] segment is found that is adequate to hear the created sound match the specified sequence. The identification rate of all sentences for each listener is shown in the following curve. In this instance, the lowest rate is 82% for the ninth listener (see table 1), an average third-year student with a very low level. This indicates that the rate does not accurately reflect the generation's output, necessitating a correction at the listener level to accurately identify the content of this generation.

5. Correction of Misidentified Sequences

The erroneous identification of the two word sequences, [sentence 4] and [sentence 7] is associated with the bad quality of the sound units [wɑ] and [ruu] that should not have taken concatenated positions due to the fact that the sound [wɑ] is emphatic; it is similar to the sound [oi] in French and requires emphasizing the entire concatenated sequence in a stretch that disturbs the real sound of the concatenated unit, which leads listeners to hear something like [ʕ], which does not exist in the unit itself [qudwatan]. Others heard [n] instead of [d], and so forth, consequently resulting in the results that were inserted in Table 2. In order to provide satisfying revision to this concatenation result and generate good sound units that correspond to the written ones, a sound unit of type [wa] was verified experimentally; a suitable unit [kaw] was found with it being reflected across the line of symmetry indicating that the reverse sound unit is [wak], by performing a deletion of a sound of [k] of this syllable, [wa] is left at the end of this partitioning. This operation was repeated for the syllabic unit [ru], like in its generation in a way that would be easier to be understood by the listeners of [yuqaamiruun]. The results are listed in the following table:

Table 2: Test results and identification rate of word sequence [yuqaamiruuna] and [qudwatan] after correction [7]

listeners	**Yuqaamiruuna**	**qudwatan**
A1	100%	100%
A2	83%	87%
A3	83%	100
A4	100%	100%
A5	100%	100%
A6	100%	100%
A7	91%	100%
A8	100%	100%
A9	100%	100%
A10	91%	100%

Following the correction, the second and third listeners' identification rate of the generated sequences of [yuqaamiruuna] rises to over 83%, the seventh and tenth listeners' to 91%, and for the same sequence of the seventh sentence, all other listeners' to 100%. The identification rate of the whole group reaches 94.8% which is a very favorable result for this type of speech synthesizer of the Arabic language [7].

After the correction, the identification rate of the sequence [qudwatan] of the 4th sentence for the second listener increases to more than 87% and to 100% for all the other listeners. This gives an average of 98.7% for this group of listeners.

The results inserted in this table are very satisfactory to judge this synthesis method, and to say that we have obtained good concatenation results.

For the new sequences, the audition result rate is 98% for all generated sequences, we made a small correction which consists of adding pauses between syllables for the first and fourth sentence, this correction is made as follows:

[Ph1] = [li#yaʤ**#ðib**#nal#Hab#l]
[Ph4] = [yar#ta**#**Ti**#**mu#bi#lʤi#daar]

These corrections are to add linear smoothing to the sound sequences, its role is very beneficial in these places, in the first place before [ðib] is to remove the overlap between [ʤ] and [ð], and gives [ð] to appear really and entirely as natural. In the second place the space between [ta] and [Ti] serves to differentiate between an ordinary sound and an emphatic sound, it is to make a distinction between [t] and [T], and finally, the space added between [Ti] and [mu] which is for the appearance of [mu] clearly without ambiguity. The results of the listeners are very perfect, they have a rate of 100% of identification of all the four sequences of the sentences generated with good intelligibility and extreme continuity.

6. COMPARISON WITH EXIST METHOD

We compare the results of Arabic speech synthesis using the generative grapheme technique to produce [CVh] with the results of linear smoothing for a hybrid composition system using a polyphems method [8]. The important result that can be concluded from this comparison is the overall percentage of the regeneration method compared to the schemes method, which is approximately 97.9%, with the correction of only two ambiguous sounds in the fourth and seventh sentences, i.e., only 22 listening tests. In contrast, the results of the polyphems method use the correction and modification results for all sentences three times for each sentence, i.e., 60 listening tests. This yields the previous results [7]; the overall average is estimated at 90.9%, with a 7% difference from the overall speech recognition rate of the schemes method. Moreover, the values of this method are almost equal with values over 92%, which explains the effectiveness, efficiency, discrimination, preference and stability of the recognition rate of different sentences of the schemes method.

7. CONCLUSION

The identification rate of the constructed Arabic speech sequences largely depends on the quality of the sound snippets used. The generative form that carries formant transitions, a rather advantageous technique to automatic synthesis of speech within Arabic language, is based on the mirror technique that generates symmetric syllables of those opposites of type [CV]. These syllables are certainly carrying formant transitions to the left and right of a vowel which we have referred to as dependent on the left and right, information contained in the continuity and intelligibility section of the synthesized speech, that is the region between the consonant and the vowels carrying this information acoustically. It is the variation of the values of the frequencies F1 and F2, named formants. It exists in the sound units of type [CV] of various possible cases, (28x3), that is, 28 consonants combined with three different Arabic vowels (Fat'ha, Dhamma, and Kasra). So this is an extremely important stage. The corresponding vowels to the emphatic phonemes can be used in the case of realization of long vowels [CVV], we do segmentation in several qualities:

- of type]VV] to generate long vowels of closed syllables [CVV];
- of type]VV[to generate long vowels for open syllables, this is the case of generation of syllables of quality [CVVC and CVVCC];
- and of type]V] for the generation of closed syllables with short vowel [CV], this is the case of an emphatic consonant with short closed vowel (example: [bɑHr]).

As we have seen, these segments are identical in both the emphatic and ordinary cases. The possibility of generating all Arabic syllables symmetrically while maintaining formant transitions (mirror technique) and the total number of Arabic consonants to generate the cases of the two successive consonants (CVCC, CVVCC), or all Arabic sequences, is made possible by this technique, which enables us to synthesize speech using only [CV] type units.

REFERENCES

[1] "Liste des pays ayant l'arabe pour langue officielle," in *Langue internationale ou mondiale*, https://creativecommons.org/licenses/by-sa/3.0/deed.fr, Ed., ed. Etats-unis: organisation de bienfaisance régie par le paragraphe 501(c)(3) du code fiscal des États-Unis., 2022.

[2] B. J. Grim and B. Hsu, "Estimating the global Muslim population: Size and distribution of the world's Muslim population," *Interdisciplinary Journal of Research on Religion,* vol. 7, 2011.

[3] A. B. Kain and J. P. v. Santen, "A speech model of acoustic inventories based on asynchronous interpolation," in *Eighth European Conference on Speech Communication and Technology*, 2003.

[4] W. H. Organization, "Country focus: annual report 2008," World Health Organization2008.

[5] O. Karaali, G. Corrigan, and I. Gerson, "Speech synthesis with neural networks," in *World Congress on Neural Networks, San Diego,* 1996, pp. 45-50.

[6] A. Indumathi and E. Chandra, "Survey on speech synthesis," *Signal Processing: An International Journal (SPIJ),* vol. 6, no. 5, p. 140, 2012.

[7] L. Chegrani, G. Mhania, and B. Bachir, "The symmetric technique of formant transition generation for use in speech synthesis in Arabic," *International Journal of Information Technology,* vol. 17, no. 2, pp. 1235-1245, 2025/03/01 2025.

[8] T. Saidane, M. Zrigui, and M. B. Ahmed, "Un système de synthèse de la parole arabe par concaténation de polyphèmes: Les résultats de l’utilisation d’un lissage linéaire," in *3rd International Conference: Sciences of Electronic, Technologies of Information and Telecommunications, Tunis*, https://aclanthology.org/2005.jeptalnrecital-recitalcourt.11., 2005.