Proceedings of the CAINE 2005 Conference. Honolulu, Hawaii. November 2005.

Lebanese Colloquial Arabic Speech Recognition

Ramzi A. Haraty and Omar el Ariss
Lebanese American University
Beirut, Lebanon
Email: rharaty@lau.edu.lb

Abstract

Although there was, and still continues, extensive research and advancements in speech recognition on English language, there has been little research done on Arabic language. In addition to that, most of the research done is either for the standard Arabic language or the Egyptian colloquial language. Commercial applications related to this field are mostly based on telephony technology. In this paper, the implementation of a Lebanese colloquial Arabic discrete speech recognition is described.

1 Introduction

The natural form of communication among humans through speech is to be sought into computer technology, and if it is successfully imitated then the human-computer interaction will be more transparent. The advancement of computer technology seen through the growing usage of personal digital assistant (PDA) and tablet PC, makes speech a central, if not the only, means of communication between the human and the machine [2, 3].

Speech recognition through computer software encounters diverse types of difficulties due to the enormous information that is carried with the speech signal. Therefore, the need to apply constraints to simplify the difficulties are needed in order to make the recognition process possible. Some of the constraints could be the recognition of isolated words, limitation in the vocabulary size, or a limitation in the number of speakers. As the technology advances, the constraints become weaker and the process of recognition starts to behave as human like. Some of the difficulties that a speech recognition system encounters are [3, 5]:

1. The voluminous data in the speech sound wave: Although it may seem as if we speak using a single tone, the quantity of data in the sound wave is overwhelming.

2. Word knowledge: Speech is not just acoustic sound patterns, additional knowledge, as word meanings, is needed in order to recognize exactly the intended speech. Therefore, words with widely different meanings may share the same sequence of sound patterns. For example:

- The word'كَلَّ' that means exhausted, and the word 'كَلاَّ' that means no or never.
- The word 'جَرَّ' that means to drag, the word 'جَرَّى' that means to make something to stream, and the word 'جَرَّة' that means a jar.

3. The continuous flow of speech: Speech is uttered as a continuous flow of sounds and even when words are spoken distinctly there is no inherent separation between the words. To illustrate this idea, any unfamiliar foreign language can be heard as a continuous stream of sound without any distinction or identification of the word boundaries.

4. Variability: A person's voice and speech patterns can be entirely different from those of another person. The elements that cause the difference

are many: size and shape of the mouth, length and width of the neck, age, sex, regional dialect, health, and personal style of speech. An example of this variability is: speakers in Egypt pronounce the phoneme 'ج' in the word 'جمال' different than the speakers in Lebanon. Another variability is that some speakers talk more slowly or more nasally. Even a single speaker will exhibit variability. The sound pattern of a word changes when speakers whisper, shout, and become angry, sad, tired, or ill. Even when speaking normally, individual speakers rarely say a word the same way twice. Variability is a basic characteristic of speech.

6. Coarticulation effects: The acoustic realization of a phoneme may heavily depend on the acoustic context in which it occurs. This effect is usually called coarticulation. Thus, the acoustic feature of a phoneme is affected by the neighboring phonemes, the position of a phoneme in a word, and the position of this word in a sentence. Such acoustic features are very different from those of isolated phonemes, since the articulatory organs do not move as much in continuous speech as in isolated utterances. We can see the effect of coarticulation in the following phrase 'و في الأيام'. Here the phoneme 'ي' in the word 'في' is affected by the neighboring phoneme 'فِ' and by the phoneme 'لْ' in the word 'الأيام'. Therefore the acoustic realization is different from the stand alone phoneme 'يِ'.

7. Insufficient linguistic knowledge: With the help of linguistic knowledge, such as syntactic and semantic constraints, the listener can usually predict the next word. Unfortunately, this kind of knowledge is not applied in the field of speech recognition due to the difficulty to model this mechanism.

8. Arabic language difficulties: Arabic language presents problems that are not encountered in mainstream languages like English or Spanish. The cause to such difficulties is the extreme dialectical variation and non-standardized speech representations. Some of the dialectical variation for the word 'رحم' are: 'رَحِمَ', 'رُحِمَ', 'رَحَّمَ','رَحِمْ".

9. Noise: The speech signal carries with it different types of noise

- Background noise.
- Noise produced by the input device (telephone or microphone).
- Sounds made by the speaker; such as lip smack, nervous breathing.
- Non-communication vocalizations made by the speaker such as "uh", a cough.

The research proposed here is for an Arabic speech recognition application, concentrating on the Lebanese dialect for the six digits (1-6). The speech recognition system is a small vocabulary based system that is speaker dependent and accepts discrete speech, that is the user has to pause between words in order to identify the word boundaries. The system starts by sampling the speech, which is the process of transforming the sound from analog to digital, and then extracts the features by using the Mel-frequency cepstral coefficients (MFCC). The extracted feature is then compared with the system's stored model; in this case the stored model chosen is a word-based model. The reference model used is template matching. Dynamic Time Warping is applied in comparing the input sound with the stored templates to improve the difference in duration.

In the following section a detailed description of the Arabic language will be given. The description given will be related to the field of speech recognition. In section 3 the implementation stages of the proposed system will be described, followed by descriptions of the calculation of the delta

coefficients (section 4), Dynamic Time Warping (section 5), and performance evaluation (section 6).

2 The Arabic Language

Linguistically speaking, Arabic language does not have a normalized form that is used in all circumstances of speech and writing. Arabic used in daily informal communication is not the same form of Arabic that is used on TV to broadcast the news. The forms of Arabic are as follows:

- Classical or formal Arabic: is the old form of the language. It can be seen in the Jahelia poetry.
- Modern Standard Arabic (MSA): is a version of classical Arabic with modernized vocabulary. It is considered to be the formal language that is common in all Arabic speaking countries. Modern Standard Arabic is the form of Arabic used in all written texts.
- Colloquial or dialectical Arabic: there are many different dialects that differ considerably from each other and from the Modern Standard Arabic. According to the [4], colloquial Arabic can be divided into the following subgroups: Gulf Arabic, Egyptian Arabic, Levantine Arabic, and North African Arabic. This categorization is too general and wrong at the same time. Dialectical forms of Arabic can be many even in one country. For example Lebanon, dialects are different in the south, north, Beirut, and the mountains, further dialectical subdivisions can also be made. Another example, in Oman the dialect spoken is similar to the dialect spoken in Sudan and not to the other Gulf countries. The regional dialects of Arabic are spoken languages; very little written dialectical material exists.

Although some consider the alphabet to consist of twenty-eight letters (excluding the hamza) [4, 7], the Arabic alphabet consists of twenty-nine letters. Additional symbols or letters can be introduced for certain phones that are not present in the Arabic alphabet (like the English phonemes [p] and [v]).

Arabic doesn't have letters for vowels; all the alphabets are consonants. Diacritics play an important role in forming short vowels. The fatha, kasra, damma, and tanween all form different short vowels for the same letter. Long vowels can also be produced by adding an 'ا' after a short vowel. Also the madda diacritic form a long vowel for the letter 'ا'. The sokoon means that the letter is a consonant, while the shadda doubles the letter (the first is a consonant while the other letter is a vowel). The lack of diacritics in a word might cause considerable ambiguities, leading the vocabulary to be used in a speech recognition system to give wrong results. The word 'كتب' as an example has a possibility of 21 diacritizations. Therefore in order for a word-based speech recognition system to recognize those diacritization, the system must have at least one model for every diacritization form. Table 1 lists all the Arabic diacritics:

Table 1: Arabic Diacritics

Symbol	Name	Meaning	Example
◌ْ	Sokoon	Consonant letter	حبْس
◌َ	Fatha	Short vowel	كَذَبَ
◌ُ	Damma	Short vowel	كُل
◌ِ	Kasra	Short vowel	عِند
◌ّ	Shadda	Letter doubling	شدَّة
◌ً	Tanween el-fatha	Adds [an] to the letter	ايضاً

2.1 Letter production in the Arabic language

Arabic letters can be divided into subgroups depending on the place and manner of articulation. Figure 1 shows a detailed diagram for the articulation of the Arabic

letters. Some of the relevant categorizations are mentioned below:

- According to Ibn Sina [7], letters can be either single or composite. Letters that are produced by complete blockage of airflow are considered to be single letters. The single letters are: ‘ب’, ‘ت’, ‘ج’, ‘د’, ‘ط’, ‘ض’, ‘ق’, ‘ك’, ‘ل’, ‘م’, ‘ن’. The rest of the letters are considered to be composite.

- According to Yousof Bin Abi Beker El Sakaki [7], letters can also be loud or quiet. The loud letters are the product of the restriction of the airflow, while the quiet letters have no restriction of airflow during production of the letter. The loud letters are: 'ء', 'ا', 'ق', 'ك', 'ج', 'ي', 'ر', 'ن', 'ط', 'د', 'ت', 'ب', 'م', 'و'. The rest of the letters are considered to be quiet.

- According to Abdallah Bin Mohammed El Khafaji [7], the letters can be dense, intermediate, or loose. The letters are considered to be dense when the airflow is obstructed during production, and considered to be loose if there was no obstruction of the airflow. The dense letters are: 'ء', 'ق', 'ك', 'ج', 'ط', 'د', 'ت', 'ب'. The intermediate letters are: 'ا', 'ع', 'ر', 'ل', 'ي', 'ن', 'م', 'و'. The rest of the letters are considered to be loose.

- Other categorizations [7] are: Closed letters are the letters that are produced by closing the lips and raising the tongue, the letters are: 'ص', 'ض', 'ط', 'ظ'. The tip letters are the letters that are produced by using the tip of the tongue, and the letters are: 'ل', 'ر', 'ن', 'ف', 'ب', 'م'.

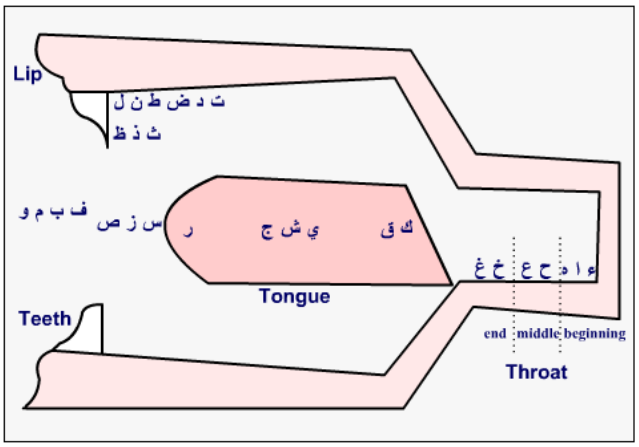

Figure 1: Places of articulation for the Arabic letters

3 Feature Extraction

Figure 2 shows the structure of a speech signal analysis component in an Automatic Speech Recognition system. The speech analysis, as shown below, can be summarized into three main stages, the first is done through hardware while the remaining two are implemented through software. The first stage can be shown as the movement of speech through the microphone, followed by the passage of the microphone output through the A/D converter. The microphone transforms the pressure wave into an electrical analog signal, while the A/D converter digitizes or transforms the analog signal into a digital signal. The second stage is the extraction of the features from a digitized speech signal.

The third stage recognizes the word uttered from the features extracted from the speech signal.

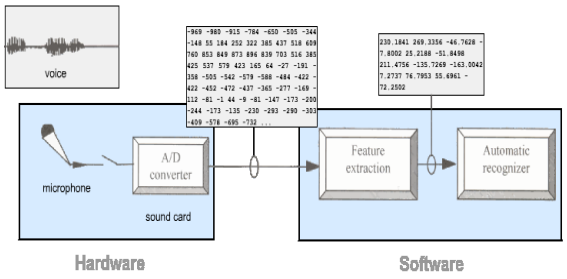

Figure 2: ***Structure of a speech signal analysis***

The extraction of reliable features is one of the most important issues in speech recognition. The Mel-Frequency Cepstrum Coefficient (MFCC) is chosen to be the feature extraction method due to the better performance, and the ability of the frequency domain to model adequately the sound. Figure 3 shows the components of an MFCC process with the number of input values for every component.

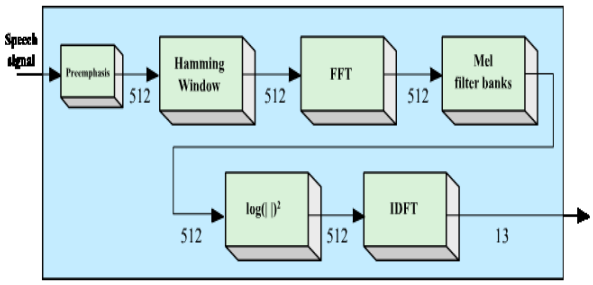

Figure 3: ***Components of MFCC***

3.1 Preemphasis

Formants, which are the peaks that result from the resonance of the vocal tract, usually define the structure of a phoneme. The high frequency formants carry with them relevant information, but they have smaller amplitude with respect to law frequency formants. Therefore, an amplitude that is the same for all formants should be attained. This can be done through the use of a Preemphasis filter, which flattens the spectral tilt. Preemphasis can be accomplished after the digitization of a speech signal through the application of the first-order Finite Impulse Response (FIR) filter [2, 3]

$$H(z) = 1 - \alpha z\text{-}1$$

where α is the Preemphasis parameter set to a value close to 1, in this case 0.95. Applying the FIR filter to the speech signal, the preempahsized signal is related to the input signal by the relation:

$$x'(n) = x(n) - \alpha x(n-1)$$

3.2 Windowing

Fourier transform, which will be discussed in the next section, is reliable only when the signal is in a stationary position. For voice, this holds only within a short time interval usually less than 100 milliseconds. Therefore, the speech signal is decomposed into a series of short segments, called analysis frames, then each frame will be analyzed and useful features will be extracted from it. A 512 points frame is chosen in this research, this frame segmentation can be seen in figure 4 [2, 3].

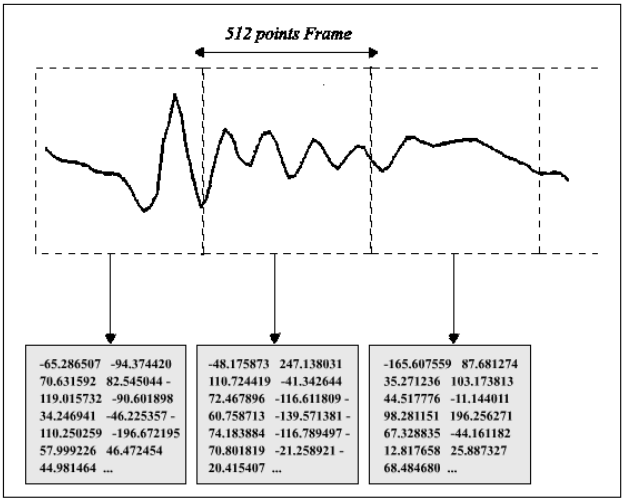

Figure 4: ***Frame segmentation of a speech signal***

To minimize the discontinuity and

therefore preventing spectral leakage of a signal at the beginning and end of each frame, every frame is multiplied by a window function. Window functions are signals that are concentrated in time, often limited in duration, that consist of a central lobe which contains most of the energy of the window and side lobes which decay rapidly. There are many different window functions, like rectangular, hanning, hamming, triangular, Kaiser, and many others, that can be applied to a speech signal. Here, the hamming window will be used. The characteristics and the application of this window to the speech signal can be seen in figure 5 [1, 2, 3, 6].

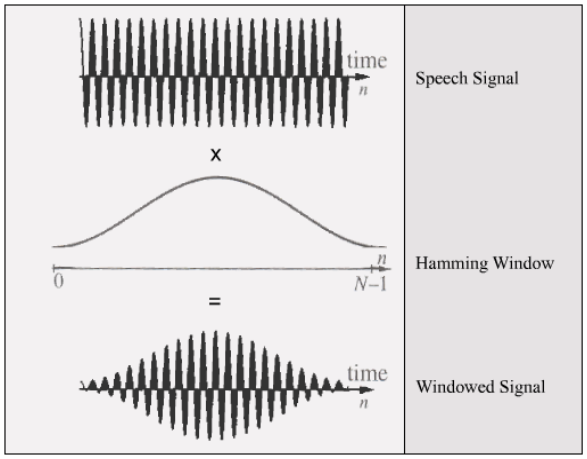

Figure 5: ***Characteristics of a Hamming Window***

The hamming window is defined as

$$W_H(n) = 0.54 - 0.46 \cos(2n\pi/N-1)$$

and the application of this window function to the speech signal is

$$xt(n) = W_H(n).x'(n)$$

3.3 Fast Fourier Transform

Discrete Fourier Transform (DFT) is considered to be the basis of spectral analysis, and spectral analysis reveals speech features that are due to the shape of the vocal tract. The Discrete Fourier Transform of a finite duration sequence $\{x(n)\}$ where $0 \leq n \leq N - 1$ is defined as:

$$X(k) = \sum x(n)e^{-j(2\pi/N)nk} = \sum x(n)W^{nk}$$
where $(0 \leq n, k \leq N - 1)$

It can be easily seen that W^{nk} is periodic of period N, and this periodicity is the key to the Fast Fourier Transform. The Fast Fourier Transform (FFT) is an algorithm that consists of variety of trick for reducing the computation time required to compute a DFT. Although FFT algorithms are well known and widely used, they are rather intricate and often difficult to grasp due to the great variety of different FFT algorithms such as radix-4, split-radix, radix-8, radix-16, and decimation-in-time (DIF) algorithms [1, 6].

This research implements the radix-2 algorithm. The idea behind this algorithm is to break the original N point sequence into two shorter sequences. This process continues by iterating, as long as N is an integer power of 2, until two point DFT's are left to be evaluated. The algorithm described here has been called the decimation-in-time (DIT) algorithm, since at each stage of the process, the input sequence is divided into smaller sequences; that is the input sequence is decimated at each stage [3, 6].

3.4 Mel Filter bank processing

This procedure has the role of smoothing the spectrum, closely modeling the sensitivity of the human ear. The Mel frequency scale is composed of a set of band-pass filters, generally 24 filters are used. The part of the spectrum which is below 1 kHz is usually processed by more filter banks since it contains more relevant information. Mel filters are linear below 1 kHz, and logarithmic above, with equal numbers of samples taken below and above [1, 3].

3.5 Log energy and IDFT

After smoothing the spectrum, the logarithm of the square magnitude of the coefficients are computed. The final step in MFCC consists of performing the Inverse Discrete

Fourier Transform (IDFT) on the logarithm coefficients. The IDFT can be calculated using the FFT procedure.

4 Delta Coefficients

Dynamic spectral transitions or features are believed to play an important role in human perception. Based on this idea, the first- and second-delta coefficients are calculated. Although we can compute the MFCC in real-time, the 1st order delta coefficients and the 2nd order delta coefficients (also known as acceleration coefficients) cannot be calculated directly. This is due to the fact that delta coefficients depend on previous and future coefficients. Figure 6 shows the dependency of the previous and future frames. As can be seen, the ith 1st order delta coefficient depends on the i-2, i-1, i+1, and i+2 MFCC. Similarly, the ith 2nd order delta coefficient depends on the i-2, i-1, i+1, and i+2 1st order delta coefficients. The 1st and 2nd order delta computation of MFCC coefficients are given by [2, 8]:

$$\Delta c_k = c_{k+2} - c_{k-2}$$
$$\Delta\Delta c_k \text{ or } \Delta^2 c_k = \Delta c_{k+2} - \Delta c_{k-2}$$

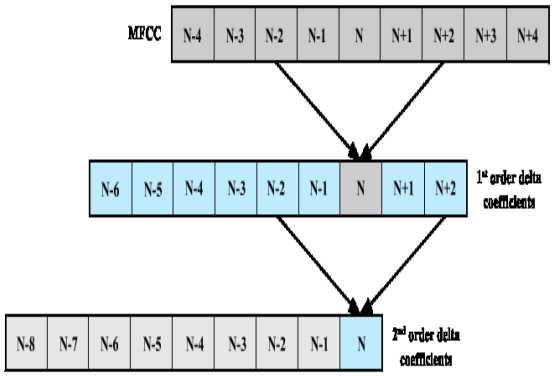

Figure 6: ***Delta coefficients computation***

The feature vector used for speech recognition is the combination of these features

$$x_k = \begin{pmatrix} c_k \\ \Delta c_k \\ \Delta\Delta c_k \end{pmatrix}$$

5 Dynamic Time Warping

Using word as a unit of recognition adds more complication to the recognition process. The speech signal is a time dependent process, therefore several utterances of the same word are likely to have different durations. Even the same word with the same duration might differ in the middle. The difference in duration is due to the different rates used while uttering the different parts of a word. For example, the word 'كيفك' has a different duration when the 'ي' is emphasized in the utterance 'كيييفك', and is also different in the utterance 'كيفاك' when 'فَ' is emphasized.

This problem can be solved through the use of Dynamic Time Warping (DTW). DTW nonlinearly expands or contracts the time axis to match the same phoneme position between the input speech, or the word uttered, and the reference template. The DTW process can be efficiently accomplished by using the Dynamic Programming (DP) technique. The DTW process can be represented mathematically as follows [2]:

$$g(i,j) = \min \begin{pmatrix} g(i, j-1) + d(i,j) \\ g(i-1, j-1) + 2d(i,j) \\ g(i-1, j) + d(i, j) \end{pmatrix}$$

Here, d(i,j) is the distance between i and j. An appropriate distance measure is the Euclidean distance. An illustration of how DTW works is shown in figure 7. The x-axis represents the input word, while the y-axis represents the template reference.

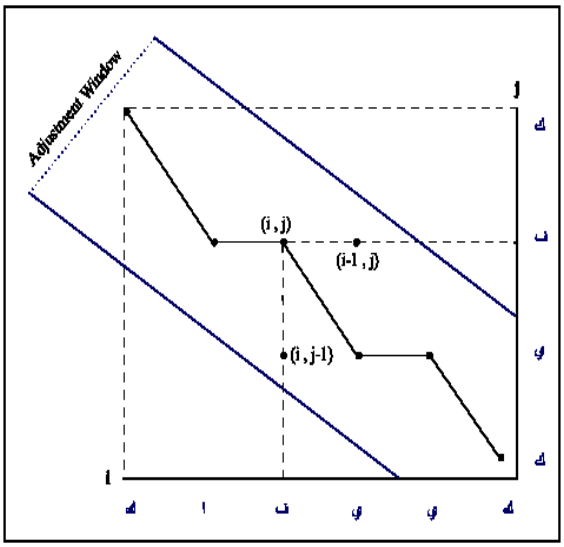

Figure 7: ***Application of DTW***

6 Performance Evaluation

In order to evaluate the speech recognition system, two testing sets were applied. In the first test, the speaker uttered each digit ten times. In the second test, the speaker uttered each digit fifteen times. The vocabulary contained two utterances for every digit. The speech input and the vocabulary templates are form the same speaker, this is because the recognition system is a speaker dependent system. Tables 2 and 3 show the results of the two testing set respectively:

where:

- Deletion (D): The speaker utters a word and the system omits it. For example, the system hears 'و لقد غادرت ' when the speaker says 'ولقد غادرت '.
- Substitution (S): A spoken word is replaced with another, usually similar, word. For example, the system hears 'الثامنة ' when the speaker says 'الثانية '.
- Insertion (I): The recognizer adds a word although the speaker didn't say it. For example, the system hears ' الى الساعة الخامس هي' when the speaker says ' الى الساعة الخامسة'.

% accuracy = (N - S - D - I)/N * 100

% error rate = 100 - % accuracy

Where N stand for the number of words of the correct speech that is for evaluation

7 Conclusion

The research represented here, to our knowledge, is the first attempt to implement a speech recognition system for the Lebanese colloquial Arabic language. Preliminary testing showed acceptable results. Future work will focus on making the system a speaker independent one, and changing the word-based model into a phoneme model.

8 References

[1] Becchetti, Claudio, and Lucio Prina Ricotti, *Speech Recognition: Theory and C++ Implementation*, 1999, Chichester, John Wiley & Sons.

[2] Furui, Sadaoki, *Digital Speech Processing, Synthesis, and Recognition*, 2001, New York, Marcel Dekker, Inc..

[3] Huang, Xuedong, Acero, Alec, and Hon Hsiao-Wuen, *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*, 2001, Upper Saddle River, Prentice Hall.

[4] Kirchhoff, Katrin, et al., *Novel Approaches to Arabic Speech Recognition: Final Report from the JHU summer workshop 2002*, 2002, Tech. Rep., John Hopkins University.

[5] Markowitz, Judith A., *Using Speech Recognition*, 1996, Upper Saddle River, Prentice Hall.

[6] Rabiner, Lawrence R., and Bernard Gold, *Theory and Application of Digital Signal Processing*, 1975, Englewood Cliffs, Prentice Hall.

[7] Tarazy, Fouad Hanna, *Al Aswat wa Makharej Al Hrouf Al Arabiet*, 1962, Beirut, Matbaet Dar Al Kotob.

[8] You, Kisun, Kim, Hoyoun, and Wonyong Sung, *Implementation of an Intonational Quality Assessment System for a Handheld Device*, October 2004, INTERSPEECH-ICSLP-2004, pp. 1857-1860.

Table 2: Testing result for 10 utterances for every word

Digit	Articulation	Correct	Substitution	Deletion	Insertion	%accuracy	%error rate
واحد	واحَدْ	8	2	0	0	80	20
إثنين	تِتِينْ	5	4	1	0	50	50
ثلاثة	تِلِيتِي	6	4	0	0	60	40
أربعة	أَرْبَعَ	7	3	0	0	70	30
خمسة	خَمْسِ	9	1	0	0	90	10
ستة	سِتّ	8	2	0	0	80	20
Total		**43**	**16**	**1**	**0**	**71.66**	**28.34**

Table 3: Testing result for 15 utterances for every word

Digit	Articulation	Correct	Substitution	Deletion	Insertion	%accuracy	%error rate
واحد	واحَدْ	10	5	0	0	66.67	33.33
إثنين	تِتِينْ	9	5	1	0	60.00	40
ثلاثة	تِلِيتِي	8	7	0	0	53.33	46.67
أربعة	أَرْبَعَ	11	4	0	0	73.33	26.67
خمسة	خَمْسِ	14	1	0	0	93.33	6.67
ستة	سِتّ	13	2	0	0	86.67	13.33
Total		**65**	**24**	**1**	**0**	**72.22**	**27.78**