Audio Search Based on Keyword Spotting in Arabic Language

Mostafa Awaid
Biomedical Engineering Department
Higher Technological Institute
10th of Ramadan City,
Egypt

Sahar A. Fawzi
Systems and Biomedical
Engineering Department
Faculty of Engineering, Cairo
University Giza, Egypt

Ahmed H. Kandil
Systems and Biomedical
Engineering Department
Faculty of Engineering, Cairo
University Giza, Egypt Systems and
Biomedical Engineering Department
Higher Institute of Engineering
El-Shorouk, Egypt

Abstract— **Keyword spotting is an important application of speech recognition. This research introduces a keyword spotting approach to perform audio searching of uttered words in Arabic speech. The matching process depends on the utterance nucleus which is insensitive to its context. For spotting the targeted utterances, the matched nuclei are expanded to cover the whole utterances. Applying this approach to Quran and standard Arabic has promising results. To improve this spotting approach, it is combined with a text search in case of the existence of a transcript. This can be applied on Quran as there is exact correspondence between the audio and text files of each verse. The developed approach starts by text search to identify the verses that include the target utterance(s). For each allocated verse, the occurrence(s) of the target utterance is determined. The targeted utterance (the reference) is manually segmented from an allocated verse. Then Keyword spotting is performed for the extracted reference to the corresponding audio file. The accuracy of the spotted utterances achieved 97%. The experiments showed that the use of the combined text and audio search has reduced the search time by 90% when compared with audio search only tested on the same content. The developed approach has been applied to non-transcribed audio files (preaches and News) for searching chosen utterances. The results are promising. The accuracy of spotting was around 84% in case of preaches and 88% in case of the news.**

Keywords—Speech Recognition; Keyword Spotting; Template Matching

I. Introduction

Keyword spotting (KWS) is a technique used to allocate and identify target words/utterances in continuous speech. Keyword spotting systems can be classified into two categories: speaker dependent and speaker independent. For speaker dependent systems, models are developed for a specific speaker. While speaker independent systems need to be more generic and hence need more complex design.

Arabic language is the official language in more than twenty countries with population of more than one billion persons. Since it is the language of Islam religion, more people need use it and to learn its proper pronunciation. Arabic syllables begins with a consonant (c) followed by a vowel (v) or long vowel (v:) and may include one or two extra consonants. Syllables are classified according to the length of the vowels, which also known as Harakatt [1]. The five basic syllable structures in classical Arabic are: CV, CV: , CVC , CV: C , and CVCC.

Audio keyword spotting systems are difficult due to the huge variability of pronunciations between different speakers or even between repetitions of the same word by the same speaker. There exist different approaches to implement audio keyword spotting systems. Template matching approaches are used in small-scale systems and may result in accurate results when exact matching is needed [2, 3].

For audio files with corresponding text available, as in the case of the holly Quran, a text search can be used to help allocate the sentences (verses) containing the requested utterance.

This paper is organized as follows. Section two includes a description of the system. The results and discussion are presented in section Three. The conclusions are given in section four.

II. Description Of The System

The proposed audio keyword spotting system is supposed to search and allocate requested speech segments (word, connected words, sentences) within continuous speech using a Template Matching based approach with the help of text search. There is no restriction concerning the number of occurrences of the word.

The system is divided into two successive phases. The first phase is the text search in which the target utterance is given as a text for the system. The text search results in a set of sentences that include the target text. A target utterance is segmented from one of allocated audio files. Then, all allocated audio files is searched for the targeted utterance (Keyword spotting phase), as shown in Fig. 1.

The system can be used to extract the matched utterances of a given word or phrase independent of their contextual sensitivity. The developed approach has overcome this difficulty by extracting the nucleus of the utterance defined as the reference speech segment (excluding the peripheral syllables). After allocating the targeted utterances, a reconstruction procedure is applied to accomplish full matching with the original utterance. The allocation process was implemented through Pre-processing, Features Extraction, and Classification (Cross Correlation or Minimum Mean Square Error) [4].

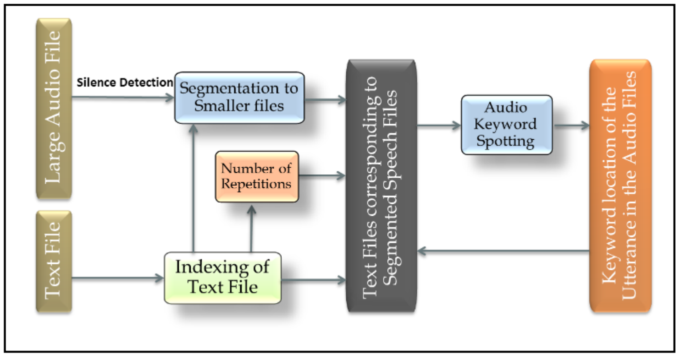

Large Audio File
Text File
Silence Detection
Segmentation to Smaller files
Number of Repetitions
Indexing of Text File
Text Files corresponding to Segmented Speech Files
Audio Keyword Spotting
Keyword location of the Utterance in the Audio Files

Fig. 1. General Block Diagram of the system [4]

A. *Text Search:*

The text file is searched for the target text ignoring the vowelization differences between the targeted word(s) and the matched one(s) in the sentence. The audio file is segmented into shorter audio files knowing that each silence corresponds to a sentence separator. The allocated sentences and their corresponding audio files are the new search domain.

B. *Targeted utterance preparation:*

The first step to create feature vectors representing the acoustic signal is pre-processing. A high pass filter is used to decrease the noise and to flatten the speech signal spectrum (Pre-emphasis), using (1).

$$H(z) = 1 - 0.95z^{-1} \qquad (1)$$

Since the vocal tract changes relatively slow, speech is considered a random process with slowly varying properties [5]. So, the speech utterance is divided into a number of overlapping frames having durations of around 10 msec. A Hamming window is applied to minimize the discontinuities at the beginning and end of each frame. The Hamming window is given by (2).

$$W(n) = 0.54 - 0.46 \cos\left(\frac{2\pi n}{N-1}\right) \qquad (2)$$

In order to omit the co-articulation effect, frames corresponding to the first and last syllables of the utterance are ignored. The remaining frames represent the utterance nucleus.

C. *Features Extraction*

The speech features techniques used are Mel-Frequency Cepstral Coefficient (MFCC) and Linear Predictive Cepstral Coefficient (LPCC).

1) *Mel-Frequency Cepstral Coefficient (MFCC)*

MFCC is a popular feature set, used in speech recognition, and based on the frequency domain of Mel scale for the human ear scale [3]. Mel-scale is based on filter bank processing. The Mel-frequency scale formula is based on mathematical equation given by (3).

$$f_{Mel} = 1127.01 \ln\left(\frac{f}{700} + 1\right) \qquad (3)$$

Steps to derive MFCC*:*

- Fourier transform of each frame of the signal is obtained.
- Mel scaling is applied using triangular overlapping windows.
- Calculate the log of the power spectrum at each of the Mel frequencies.
- Compute the discrete cosine transform (DCT).
- The MFCCs are the amplitudes of the resulting spectrum.

2) *Linear Predictive Cepstral Coefficient (LPCC)*

LPCC is one of the most powerful speech analysis techniques for extracting good quality features [6]. The process for obtaining the LPCC features vectors is shown in Fig. 2

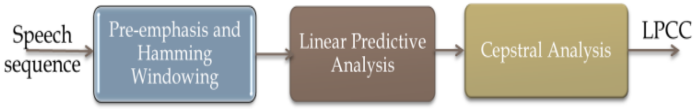

Speech sequence
Pre-emphasis and Hamming Windowing
Linear Predictive Analysis
Cepstral Analysis
LPCC

Fig. 2. Block diagram of the computation steps of LPCC.

The notion behind LPCC is to model the human vocal tract by an all-pole filter. The LPC Coefficients $a_{i,}$ are the coefficients of the all pass transfer function *H (z)* modeling the vocal tract [7], as shown by (4):

$$H(z) = \frac{1}{1-\sum_{i=1}^{p} a_i z^{-i}} \qquad (4)$$

D. *Classification:*

Two classification methods are used.

1) *Mean Square Error (MSE)*

Mean Square Error (MSE) is a signal fidelity measure used to compare two signals. Such a measure provides a quantitative score describing the degree of fidelity/ similarity [8]. Since **x** = {*xi*|*i* = 1, 2, … , *N*} and **y** = {*yi*|*i* = 1, 2, …, *N*} are two finite-length, discrete signals representing two distinct utterances, where *N* is the number of signal frames and *xi*, *yi* are the features vectors of frames constituting ***x*** and ***y***, respectively. The MSE between the two signals is determined be (5).

$$MSE(x, y) = \frac{1}{N} \sum_{i=1}^{N} (x_i - y_i)^2 \qquad (5)$$

Finally, the minimum square error computed is compared with the predetermined threshold, to accept or reject the tested pattern.

2) *Cross-Correlation*

Cross-correlation is a measure of similarity of two waveforms. This process is considered as a Template Matching (TM) based approach. The cross-correlation coefficient should be high between the target and the reference utterances.

3) *Recovery of the full utterance:*

In this phase, the frames corresponding to the first and last syllables are concatenated to the nucleus in order to restore the complete target utterance. This is achieved through applying the template matching explained above.

The hybrid technique described in this paper was fully implemented in MATLAB 7.12.

III. Results and Discussion

The system performance is tested in two tracks. The first track is by applying it on Quranic words uttered by professionals readers, in this case a text search is accompanying the audio search. The rules of recitation of the Quran lead to consistent pronunciations. The system accuracy is evaluated by detecting correct and complete target words. In the second track, only audio search is applied on standard Arabic audio files representing lecture of Quran explanation, BBC Arabic news and a BBC Arabic interview.

A. *Experimental setup*

For the first track: More than three hundred utterances were used as reference patterns. Selected words or syllables were used as keywords and 15 hours' of audio files of Quran data were used for evaluation. Feature parameters used were 8 and 12 MFCCs (Mel-Frequency Cepstral Coefficients) and 12 LPCCs (Linear predictive Cepstral Coefficients). The algorithm was applied on words/ phrases from Quran context.

For the second track: Fifty utterances were used as reference patterns. Selected words were used as keywords and 60 minutes of audio files of Quran explanation "Tafseer", BBC Arabic news and a BBC Arabic interview were used for evaluation. Feature parameter used was 8 MFCCs (Mel-Frequency Cepstral Coefficients) because; it's the best feature according to the previous experiments. The algorithm was applied on words/ phrases from episodes of "Tafseet" and BBC News.

B. *Experiments results of the first track*

- In order to measure the value added by using text search before Audio Keyword Spotting, a Quranic audio file representing the first 42 verses from "*EL-Rahman*" Surah was searched to allocate the repeated utterance "آلاء ربكما". The Audio Keyword Spotting allocated the utterances in time duration of 300 seconds. When text search was added, a search time of 30 seconds has been achieved with the same accuracy. So a reduction of 90% of the search time was achieved.
- Silence detection was performed in order to divide large audio files into smaller ones which make the audio search more efficient, as shown in Fig. 3.

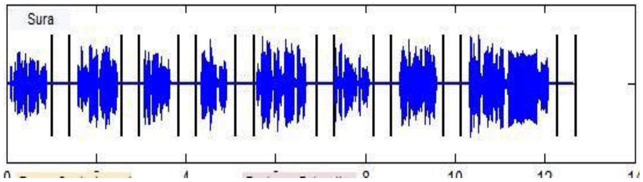


Fig. 3. Detecting the duration of silences Inter-Verse in "سورة العصر"

- In another experiment the Keyword ["الحمد لله"; Al-hamdulellah] was allocated in a long utterance as shown in Fig. 4.

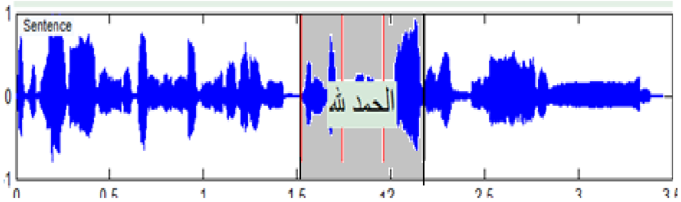
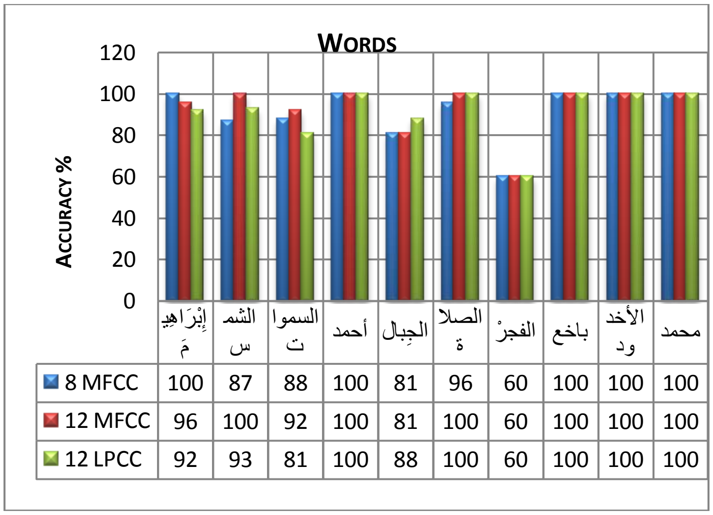


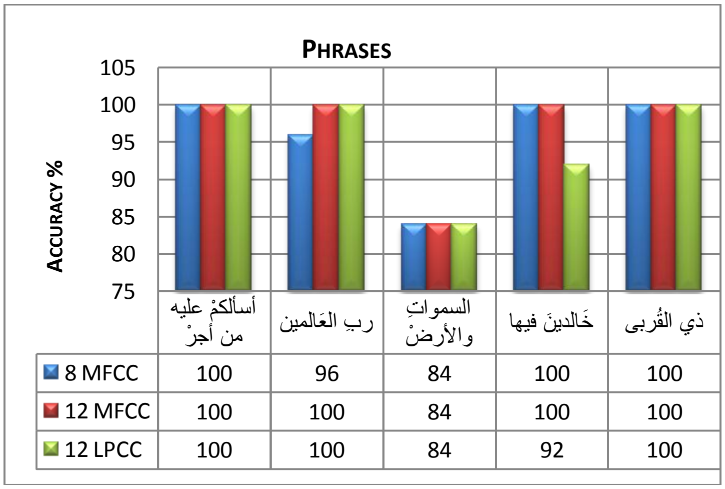
Fig. 4. Detect a Keyword in utterance.

The results obtained by applying the hybrid text/audio search on different Quranic utterances, using different sets of parameters, are summarized in the charts represented in Fig.5 and Fig.6.



	إبْرَاهِيمَ	الشمس	السموات	أحمد	الجِبال	الصلاة	الفجرْ	باخع	الأخدود	محمد
🟦 8 MFCC	100	87	88	100	81	96	60	100	100	100
🟥 12 MFCC	96	100	92	100	81	100	60	100	100	100
🟩 12 LPCC	92	93	81	100	88	100	60	100	100	100

Fig. 5. Percentage of Accuracy for Words.



	أسألكمْ عليه من أجرْ	ربِ العَالمين	السمواتِ والأرضْ	خَالدينَ فيها	ذي القُربى
🟦 8 MFCC	100	96	84	100	100
🟥 12 MFCC	100	100	84	100	100
🟩 12 LPCC	100	100	84	92	100

Fig. 6. Percentage of Accuracy for Phrases.

Results obtained from the first set of experiments are presented in tables I, II, III. The uttered word search was performed using audio reference uttered by the same reader (*El-Hossary*) with different features and different coefficients.

TABLE I. FIRST TRACK RESULTS OF (8 MFCC) FEATURE VECTOR FOR QURAN

KEYWORD	TRANSCRIPT	TEMPLATES	ACCURACY%
إِبْرَاهِيمَ	Ebrahiem	26	100
ربِ العَالمين	Rab-Elalameen	25	96
الشمس	A-Shams	30	87
السموات	A-Samawat	26	88
السمواتِ والأرضْ	A-Samawat w-Al-Ard	25	84
الجِبال	Al-jebal	26	81
الصلاة	Al-Salah	25	96
خَالدينَ فيها	Khaldeena-Feha	25	100
أسألكمْ عليه من أجرْ	Asalokom Alyeh men-Agr	7	100
ذي القُربى	Ze-lqurba	5	100
الفجرْ	Al-Fajr	5	60
أحمد	Ahmad	1	67
باخع	Bakhea	2	100
الأخدود	Al-Okhdod	1	100
محمد	Muhammad	4	100
TOTAL ACCURACY			90.6%

TABLE II. FIRST TRACK RESULTS OF (12 MFCC) FEATURE VECTOR FOR QURAN

KEYWORD	TRANSCRIPT	TEMPLATES	ACCURACY%
إِبْرَاهِيمَ	Ebrahiem	26	96
ربِ العَالمين	Rab-Elalameen	25	100
الشمس	A-Shams	30	100
السموات	A-Samawat	26	92
السمواتِ والأرضْ	A-Samawat w-Al-Ard	25	84
الجِبال	Al-jebal	26	81
الصلاة	Al-Salah	25	100
خَالدينَ فيها	Khaldeena-Feha	25	100
أسألكمْ عليه من أجرْ	Asalokom Alyeh men-agr	7	100
ذي القُربى	Ze-lqurba	5	100
الفجرْ	Al-Fajr	5	60
أحمد	Ahmad	1	100
باخع	Bakhea	2	100
الأخدود	Al-Okhdod	1	100
محمد	Muhammad	4	100
TOTAL ACCURACY			94.2%

TABLE III. FIRST TRACK RESULTS OF (12 LPCC) FEATURE VECTOR FOR QURAN

KEYWORD	TRANSCRIPT	TEMPLATES	ACCURACY%
إِبْرَاهِيمَ	Ebrahiem	26	92
ربِ العَالمين	Rab-Elalameen	25	100
الشمس	A-Shams	30	93
السموات	A-Samawat	26	81
السمواتِ والأرضْ	A-Samawat w-Al-Ard	25	84
الجِبال	Al-jebal	26	88
الصلاة	Al-Salah	25	100
خَالدينَ فيها	Khaldeena-Feha	25	92
أسألكمْ عليه من أجرْ	Asalokom Alyeh men-agr	7	100
ذي القُربى	Ze-lqurba	5	100
الفجرْ	Al-Fajr	5	60
أحمد	Ahmad	1	100
باخع	Bakhea	2	100
الأخدود	Al-Okhdod	1	100
محمد	Muhammad	4	100
TOTAL ACCURACY			92.7%

Table IV shows the best feature vector obtained from the first set of experiments presented in tables I, II, III.

TABLE IV. FIRST TRACK RESULTS OF BEST FEATURE VECTOR FOR QURAN

KEYWORD	TRANSCRIPT	TEMPLATES	BEST FEATURE	ACCURACY%
إِبْرَاهِيمَ	Ebrahiem	26	8-MFCC	100
ربِ العَالمين	Rab-Elalameen	25	12-MFCC	100
الشمس	A-Shams	30	12-MFCC	100
السموات	A-Samawat	26	12-MFCC	92
السمواتِ والأرضْ	A-Samawat w-Al-Ard	25	8-MFCC	84
الجِبال	Al-jebal	26	12-LPCC	88
الصلاة	Al-Salah	25	12-LPCC	100
خَالدينَ فيها	Khaldeena-Feha	25	8-MFCC	100
أسألكمْ عليه من أجرْ	Asalokom Alyeh men-agr	7	12-LPCC	100
ذي القُربى	Ze-lqurba	5	8-MFCC	100
الفجرْ	Al-Fajr	5	8-MFCC	80
أحمد	Ahmad	1	8-MFCC	100
باخع	Bakhea	2	8-MFCC	100
الأخدود	Al-Okhdod	1	12-LPCC	100
محمد	Muhammad	4	12-LPCC	100
TOTAL ACCURACY				97%

Another set of experiments were performed to evaluate the effect of changing the reference reader, for the same utterance, which is referred to as cross-reader. Results obtained from the first set of experiments are presented in table V. The utterance to be allocated is pronounced by one reader (in this case *El-Hossary*) and the reference was recorded by another reader (in this case *El-Menshawy*) and vice versa. Promising results reached 72%.

TABLE V. FIRST TRACK CROSS-READER BETWEEN UTTERANCE (EL-MENSHAWY) AND REFERENCE (EL-HOSSARY)

KEYWORD	TRANSCRIPT	TEMPLATES	BEST FEATURE	ACCURACY %
ذي القُربى	Ze-lqurba	5	8-MFCC	80
أسألكمْ عليه من أجرْ	Asalokom alyh men-agr	7	12-LPC	57
الفجرْ	Al-fajr	6	12-LPCC	50
الأخدود	Al-Okhdod	1	12-MFCC	100
البروج	Al-Brouj	1	8-MFCC	100
باخع	Bakhea	2	12-LPCC	100
أحمد	Ahmad	1	12-MFCC	100
محمد	Muhammad	4	12-LPCC	25
الطامة	Al-Tamaa	1	12-LPCC	100
TOTAL ACCURACY				71.2 %

C. Experiments results of the second track

In this track, experiments were performed on general Arabic episodes such as lecture of Quran explanation "Tafseer", BBC Arabic news and a BBC Arabic interview. In this case the audio search was conducted over the whole record, since there were no text scripts available. Results obtained from the first set of experiments are presented in tables VI, VII.

TABLE VI. SECOND TRACK RESULTS OF (8 MFCC) FEATURE VECTOR FOR TAFSEER KEYWORDS

KEYWORD	TRANSCRIPT	TEMPLATES	BEST FEATURE	ACCURACY%
الوسواس	Al-Waswas	3	8- MFCC	67
شهر	Shahr	4	8 -MFCC	100
الناس	A-Nas	8	8 -MPCC	100
القدر	Al-qdr	6	8 -MFCC	67
TOTAL ACCURACY				83.5%

TABLE VII. SECOND TRACK RESULTS OF (8 MFCC) FEATURE VECTOR FOR ARABIC NEWS KEYWORDS

KEYWORD	TRANSCRIPT	TEMPLATES	ACCURACY%
الأردن	Al-Ordon	4	100
اللاجئين السوريين	Al-ajean A-Soreen	2	100
المتظاهرين	Al-motazahreen	5	60
وزارة الداخلية	Wzart Al-dakhelea	4	50
بورسعيد	Por-Saeed	2	100
المجلس العسكري	Al-Magles Al-Askary	2	100
ميدان التحرير	Medan Al-Tahrer	2	100
الأخبار السعيدة	Al-Akhbar Al-Saeda	5	80
هولاندية	Holandya	3	100
TOTAL ACCURACY			87.8%

IV. CONCLUSIONS

In this work, a keyword spotting approach based on Template Matching was used to perform audio search for words/ phrases in audio files. The Audio Keyword Spotting is also used to allocate silence periods in audio files which results in dividing larger audio files into smaller ones. This division process improves the search process as it is performed in smaller audio files. Considering the audio files in Quran that have corresponding text files, a hybrid technique depending on both text search and audio keyword spotting is developed. This hybrid technique results is 97% accuracy when performing audio searching using audio reference of the same reader. 90% time reduction is achieved when compared with audio search only. This was accomplished using the sets of features (MFCCs and LPCCs). It was shown that the MFCC with an order 8 results in the best spotting accuracy. The accuracy of the developed spotting reached 72% when testing cross-readers utterances(the reference reader is tested against a different one).

Using the same audio keyword spotting technique to search in general audio files such as "News" and "preaches(Tafseer)" episodes, the recognition rates reached around 84% for preaches and around 88% for the same speaker in each test without the help of text search and with no recitation rules to control the speaker's pronunciation.

Despite the simplicity of the technique, it proves to be very efficient and shows high robustness to obtain high recognition rates under all circumstances.

V. FUTURE WORK

Record a larger evaluation standard database, for different speakers and different environments, to get more test cases.

The system can be expanded to cover the whole Quran by complete the implementation for acoustical database of the rest recitation rules. This can achieve by manually segmenting the phonetic units of each rule from various referenced readers sounds.

In the updates of this system, we may use the resulting automatically detected Keywords in online process such as News and Arabic dialog programs for different speakers.

REFERENCES

[1] A. Youssef, O. Emam.” An Arabic TTS based on the IBM Trainable Speech Sythesizer.” Department of Electronics & Communication Engineering, Cairo University, Giza, Egypt, 2004.

[2] Yung-Hwan Oh, Jeong-Sik Park and Kyung-Mi Park” Keyword Spotting in Broadcast News.” Department of Electrical Engineering & Computer Science Korea Advanced Institute of Science and Technology, Daejeon, Korea, 2007.

[3] J. Junkawitsch, L. Neubauer, H. Höge, and G. Ruske "A New Keyword Spotting Algorithm with Pre-Calculated Optimal Thresholds" Technical University of Munich, Germany, 1996.

[4] A. Kandil, A. Bialy, S. Fawzi, M. Awaid. “Towards Speech Corpus for the Quran Using Keyword Spotting.” M.S. thesis, Biomedical and systems Engineering Department, Cairo University, Giza, Egypt, 2013.

[5] Rabiner, L. and Juang, B. -H., Fundamentals of Speech Recognition, PTR Prentice Hall, San Francisco, NJ, 1993.

[6] Grangier D. and Bengio, S., “Learning the Inter-frame Distance for Discriminative Template-based Keyword”, International Conference on Speech Communication and Technology (INTERSPEECH), 2007.

[7] Octavian Cheng, Waleed Abdulla, Zoran Salcic "Performance Evaluation of Front-end Processing for Speech Recognition Systems.", Electrical and Computer Engineering Department, Auckland University, New Zealand, 2005.

[8] Zhou Wang and Alan C. Bovik. “Mean Squared Error: Love It or Leave It? [A new look at signal fidelity measures]” IEEE Signal Processing Magazine, pp. 98-117, January 2009.