Research Journal of Applied Sciences, Engineering and Technology 7(24): 5082-5097, 2014
DOI:10.19026/rjaset.7.903
ISSN: 2040-7459; e-ISSN: 2040-7467
© 2014 Maxwell Scientific Publication Corp.
Submitted: May 31, 2013 Accepted: April 09, 2014 Published: June 25, 2014

Research Article

Arabic Audio News Retrieval System Using Dependent Speaker Mode, Mel Frequency Cepstral Coefficient and Dynamic Time Warping Techniques

[1]Hasan Muaidi, [2]Ayat Al-Ahmad, [1]Thaer Khdoor, [1]Shihadeh Alqrainy and [1]Mahmud Alkoffash
[1]Prince Abdullah Bin Ghazi Faculty of Information Technology,
Al-Balqa' Applied University, Salt, Jordan
[2]The Faculty of Prince Al-Hussein Bin Abdullah II of Information Technology,
Hashemite University, Zarqa, Jordan

Abstract: Recently, audio data has increasingly becomes one of the prevalent source of information, especially after the exponential growth of using Internet, digital libraries systems and digital mobile devices. The currently massive amount of audio data stimulates working on developing custom audio retrieval tools to facilitate the audio retrieval tasks. The most familiar audio retrieval systems are based on searching using keyword, title or authors. This study presents the feasibility of using MEL Frequency Cepstral Coefficients (MFCCs) to extract features and Dynamic Time Warping (DTW) to compare the test patterns for Arabic audio news. The study proposes and implements architecture for content based audio retrieval system that is dedicated for the Arabic Audio News. The proposed architecture (ARANEWS) utilizes automatic speech recognition for isolated Arabic keyword speech mode; template based automatic speech recognition approach, MFCCs and DTW. ARANEWS presents a style of retrieval system that based on modeling signal waves and measuring the similarity between features that are extracted from spoken queries and spoken keywords. One of the major components that compose ARANEWS system is feature Database (ARANEWSDB). ARANEWSDB stores the extracted features (MFCCs) from the spoken keywords that are prepared to retrieve Arabic audio news. ARANEWS supports using Query by Humming (QBH) and Query by Example (QBE) instead of using query by text.

Keywords: Arabic information retrieval, audio news retrieval system, dynamic time warping, frequency cepstral coefficient

INTRODUCTION

There are many ideas may come to mind during designing of audio information retrieval system for audio news. One of these ideas is using the lyrics of news. The lyrics are extracted from each audio news using specific techniques. The keywords that reflect the nature of news are chosen from extracted lyrics and finally each audio news file is linked with its lyrics and with its extracted keywords. This kind of retrieval system supports the query by text; the type of data that are extracted and stored inside database is text data type. The retrieval approaches that are used in such kind of systems are similar to those which used in traditional search engine such as Google. This method is called spoken document retrieval; in which written queries are used to search speech archives for relevant speech information (Fujii *et al*., 2002).

Another idea is using speech to text technique with extracted lyrics from audio news. The lyrics are extracted from each audio news file, the keywords for each audio news are chosen and stored within text database and the type of data that are extracted and stored inside database is text data type. The user of audio retrieval system utters keyword and the speech to text application converts it to text. All next steps are the traditional text retrieval process. This speech-based methods have been explored in the information retrieval is called speech-driven (spoken query). Speech-driven (spoken query) retrieval in which spoken queries are used to retrieve relevant textual information, this kind of retrieval systems supports the (QBH).

In this study aims to go out from the traditional style of retrieving that depends on text and transferring to audio retrieving that depends on calculating the degree of similarity between sound waves. The speech recognition component ties with information retrieval system to compose one effective component (Ratanamahatana and Tohlong, 2006).

It becomes obvious that the users of Arab audio need Arabic audio systems such as Arabic news information retrieval system, Arabic interviews

Corresponding Author: Hasan Muaidi, Prince Abdullah Bin Ghazi Faculty of Information Technology, Al-Balqa' Applied University, Salt, Jordan information retrieval system and Arabic music information retrieval system. We believe that more work needs to be done in this study domain to enable searching a database of Arabic audio database. Hence, the main aim of this study is to build an Arabic audio news search engine that can satisfy the search requirements of the Arabic users.

The proposed architecture based on two bases: the first one is the interaction between retrieval system and human will be by using human voice directly or file contains recorded human voice. The second base is the retrieved ranked list will be audio files not text files. This kind of retrieval systems which based on the idea of interacting between human and retrieval system by using voice represents a hope for many persons who are suffering from visual disabilities and hand mobility disabilities.

A brief overview on content based audio:

The concept of content based audio retrieval: A content based audio retrieval system is the system that processes information contained in audio data and creates an abstraction of its content in term of audio features. Human brain has very powerful mechanisms, so it has the ability to distinguish between the different kinds of sound and assign them to correct semantic categories; this task is often simple for human, but it is very difficult for computer system, because of audio signals are represented by numeric series without any semantic meaning. The major goal of content based audio retrieval system is the determination of perceptually similar audio content (Mitrovic *et al*., 2010). ARANEWS is content based audio retrieval system; it creates an abstraction of audio Arabic queries and Arabic audio keywords in term of features. The extracted features are examined to determine the perceptually similar audio content.

METHODOLOGY

The fields of study in content based audio retrieval: Many fields are classified under the umbrella of content based audio retrieval such as segmentation, Automatic Speech Recognition (ASR), Music Information Retrieval (MIR) and environmental sound retrieval (Mitrovic *et al*., 2010).

The segmentation aims to distinguish between different types of sound such as speech, music, silence and environmental sound. Speech recognition concentrates on recognition of spoken word, recognition of speakers, recognition of spoken language and recognition of emotion. Music information retrieval works on analyzing the structure of music and retrieving similar pieces of music, instruments and musical genres. Environmental sound retrieval includes types of sounds neither music nor speech. ARANEWS

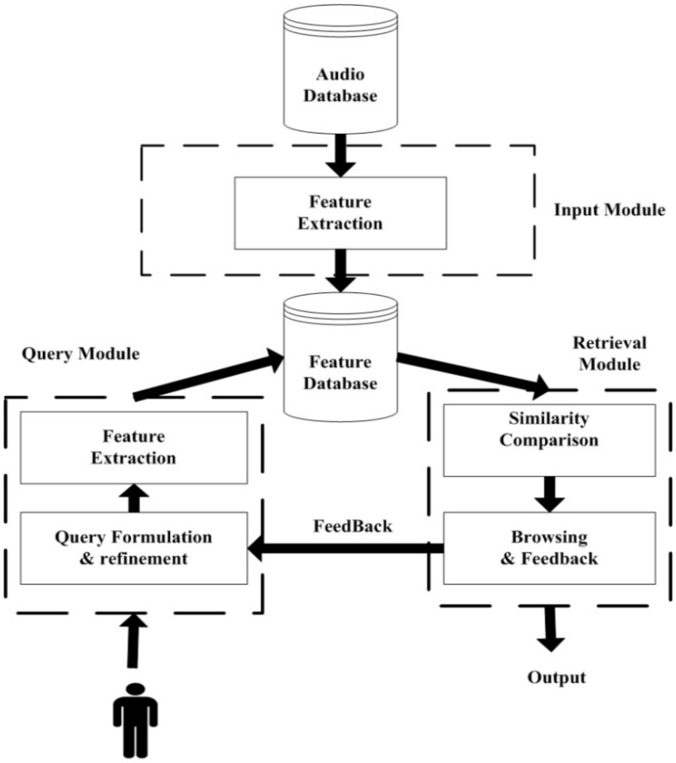


Fig. 1: Architecture of typical audio retrieval system

employs speech recognition field, it implements speech recognition component to recognize Arabic queries and Arabic keywords (Mitrovic *et al*., 2010).

The architecture of typical audio retrieval system: The architecture of typical audio retrieval system consists of three major components: input module, query module and the retrieval module. The task of input module is to extract features from audio objects that have been stored within database (audio database). The task of query module is to formulate the queries and to extract features from it using the same procedure as in the input module. The task of retrieval module is to estimate the similarity between the extracted features that are stored within feature database and the extracted features that are produced from queries.

The proposed Arabic Audio News retrieval system (ARANEWS) reflects the same architecture of the typical audio retrieval system. It consists of input module, query module and retrieval module. The Input module stores the extracted features from Arabic audio news within designed feature database. Query module extracts features from formulated and refined Arabic queries that are submitted from users. Retrieval modules measure the similarity between the extracted features from queries and extracted features from Arabic news. Figure 1 represents the architecture of typical audio retrieval system.

Feature extraction: Audio features represent specific properties of audio signals. Feature extraction is the process that captures audio properties such as the fundamental frequency and the loudness of a signal; the outputs that produced from feature extraction process are parametric numerical descriptions of signals (Mitrovic *et al*., 2010).

The amount of raw data of the audio is too big, so the audio retrieval system cannot process the raw data of audio directly. Feature extraction process aims to reduce data by extracting the most meaningful information from signal which lead to reduces the dimensionality of the input vectors while maintaining the discriminating power of the signals (Mitrovic *et al*., 2010; Gaikwad *et al*., 2010). The features are extracted once from audio files and then stored within feature database. ARANEWS system constructs ARANEWSDB (feature database) component that is designed to store extracted features from Arabic audio within it.

Query types: The retrieval systems can support using many types of query such as query by text, Query by Humming (QBH) and Query by Example (QBE). This study focuses on using (QBH) and (QBE).

QBH is search mechanism that allows the user to find audio by humming part of it into a microphone connected to computer (Mitrovic *et al*., 2010). QBE is search mechanism that allows user to search for audio "object" using an existing audio "existing object". The user has to prepare or find example of his/her favorite audio, usually the example query is audio file that is prepared or is found by the user (Helén and Lahti, 2006). When the user submits (QBE) or (QBH) to audio retrieval system, the system extracts the features from (QBE) or (QBH) and compares them with the stored extracted features within database. ARANEWS supports using (QBH) and using (QBE). The user utters Arabic keywords or submits the pre-recorded audio file that contains the Arabic keywords to ARANEWS, ARANEWS in its turn refine query and extract features from it. Figure 2 represents (QBE) works inside audio retrieval system.

A brief overview on speech recognition:

The definition of speech recognition: The technology and communications revolution opens new horizons to develop new generation of applications that characterize by exploiting the idea of interaction between the applications and human voice. If machines work on voice commands, human life will appear much more comfortable. Speech recognition process aims to give a machine the ability to "hear", "understand" and "act" upon spoken information.

Speech Recognition is defined as the process of converting speech signal to a sequence of words by means algorithm implemented as a computer program (Gaikwad *et al*., 2010). The speech recognition process is used within ARANEWS to recognize the QBH and QBE that are submitted by the user.

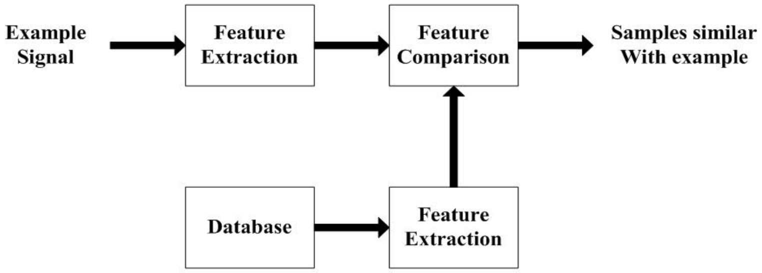


Fig. 2: (QBE) works inside audio retrieval system

Dimensions to classify the task of speech recognition: The dimensions of speech recognition help to classify the task of speech recognition component within any system. The first dimension is speech type "isolated word versus continuous speech". Some speech recognition systems are designed to recognize single word at a time "isolated word", while there are other speech recognition systems are designed to recognize sequence of words "sentence" at a time. Isolated word tends to be pronounced more clearly than continuous speech. ARANEWS works with isolated word parameter which generally more robust and easier.

The second dimension is speaker dependency "speaker dependent versus speaker independent systems". Speaker dependent speech recognition system is the system where signal patterns adapted and constructed to single speaker, whereas the speaker independent speech recognition system is the system where signal patterns adapted and constructed to multiple speakers (Reddy, 2005).

Speaker dependent speech recognition system is more accurate than speaker independent speech system, but it requires training phase which is not feasible for many applications. Speaker independent speech recognition system is more flexible and more difficult than speaker dependent speech recognition system. It requires global representation for signals to cover all types of voices, all possible ways of pronouncing words and the independent speech system must discriminate between all the various words of the vocabulary. ARANEWS system is speaker dependent speech recognition system. The third dimension is vocabulary size "small versus large". ARANEWS uses small vocabulary size for its dictionary.

Speech recognition such as pattern recognition has two phases training and testing phases. Extracting features from signals is a common process in both phases. Training phase is characterized by extracting features using large number of speech examples "training data", whereas testing phase is characterized by extracting features from testing data "data speech". Testing data are matched with model that is constructed from training data (Tiwari, 2005). Figure 3 represents the block diagram of training and testing phases in a speech recognition system.

Template based approach: There exist many approaches of automatic speech recognition such as: acoustic phonetic approach, artificial intelligence approach and template based approach. ARANEWS system has used template based approach for speech recognition.

Template based approaches is based on collection of prototypical speech patterns are stored as reference patterns that represent the dictionary of candidates words. A collection of prototypical speech is a set of prerecorded words (templates) in order to find the best match. This approach is modeled by constructing many templates per word (Gaikwad *et al*., 2010; Reddy, 2005).

Recognition is carried by matching an unknown spoken utterance with each of these reference templates and selecting the category of the best matching pattern, this approach is practical with limited size of dictionary (Gaikwad *et al*., 2010; Reddy, 2005).

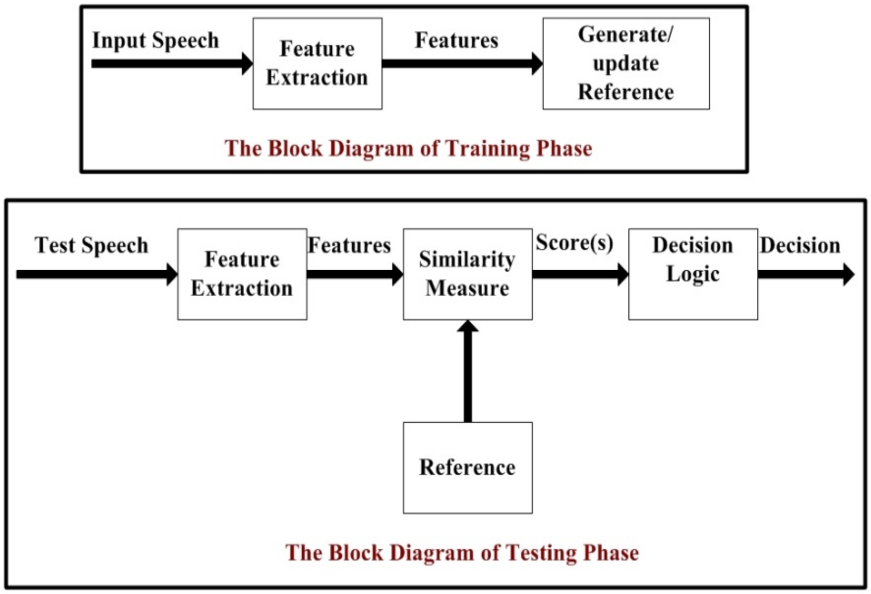


Fig. 3: Block diagram of training and testing phases in speech recognition system

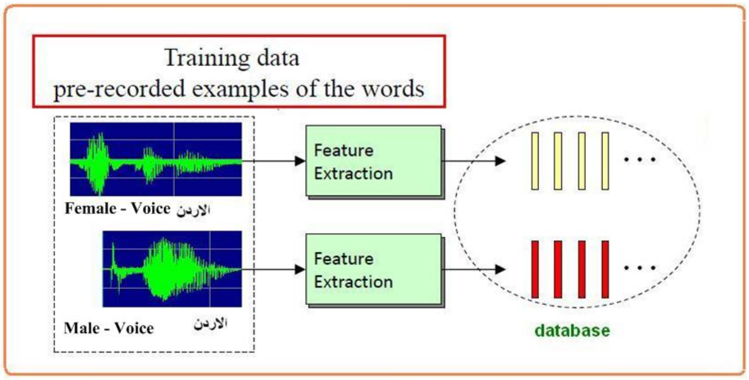


Fig. 4: The typical procedure for template based approach that is used in ARANEWS

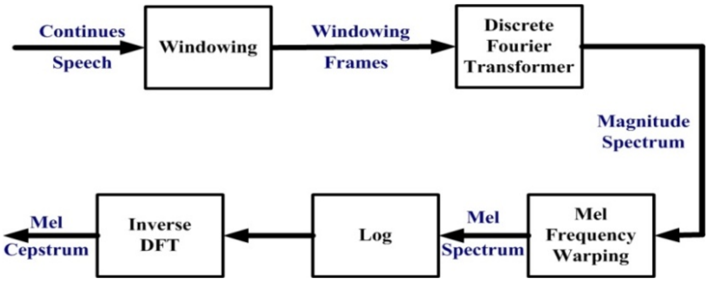


Fig. 5: The main steps for calculating MFCC

Each Arabic audio keyword that used within ARANEWS system has to be recorded many times by the dependent user who will use the system. Figure 4 represents the typical procedure for template based approach that is used within ARANEWS system.

Mel frequency cepstral coefficients "MFCC feature": To develop applications that use signals of speech, the concept of signal modeling imposes itself strongly. Signal modeling is the process of converting speech signal into a set of parameters (parametric representation). Many discriminative features are extracted from speech either to identify speakers or to recognize the speech. One of the most important extracted features from speech is Mel Frequency Cepstral Coefficients (MFCCs) feature (Picone, 1993; Ali *et al*., 2013; Thakur and Sahayam, 2013).

MFCC feature is acoustic feature that utilizes the idea of modeling human auditory system. It tries to mimic the way of our ears work, the ears analyzes speech waves linearly at low frequencies and logarithmically at high frequencies (Tiwari, 2005; Shaneh and Taheri, 2009; Muda *et al*., 2010; Dhingra *et al*., 2013; Ali *et al*., 2013).

MFCC plays on five facts to mimic the human hearing perception; the first fact is the human hearing perception does not follow a linear scale, the second fact is each tone has an actual frequency measured by ’hertz’, the third fact is each tone has subjective frequency "pitch" is measured by a scale called the mel scale, the fourth fact is the main purpose from subjective frequency is to capture the important characteristic of phonetic and the final fact is mel frequency scale is a linear below 1000 Hz and logarithmic above 1000 Hz (Tiwari, 2005; Shaneh and Taheri, 2009; Muda *et al*., 2010; Thakur and Sahayam, 2013). This study aims to investigate in viability of using MFCC as extracted feature to recognize Arabic speech. MFCC feature is extracted and derived as follow (Tiwari, 2005):

- Framing and windowing of signal
- Taking fourier transform of a signal
- Mapping the power of the spectrum above onto the mel scale
- Taking the logs of powers at each of the mel frequencies
- Taking the discrete cosine transform of the list of mel log powers
- The MFCC feature is the amplitudes of the resulting spectrum

Figure 5 represents the main steps that calculate MFCC features.

MFCC algorithm: step1: framing and windowing of signal: Framing is the process of blocking the speech signal into frames of n samples in the time domain (Shaneh and Taheri, 2009; Muda *et al*., 2010;

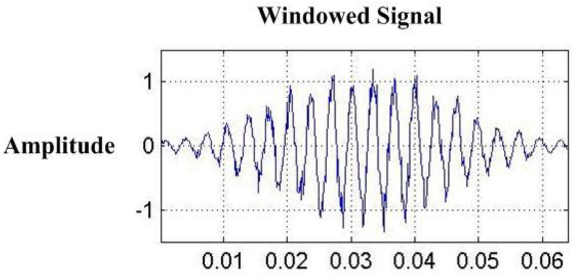


Fig. 6: The frame after applying the window function

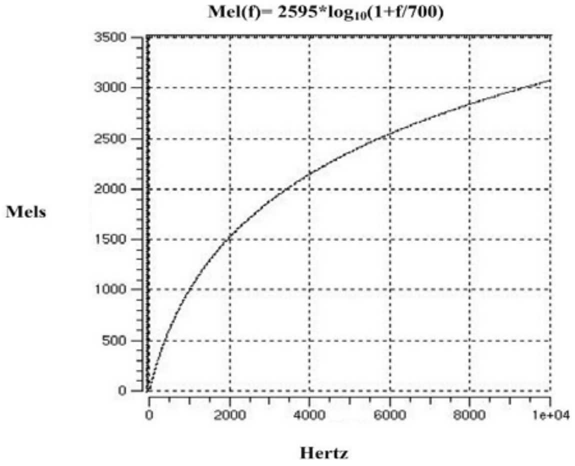


Fig. 7: The formula that is used to convert from actual frequency that is measured by Hz into subjective pitch that is measured by mel scale

Gawali *et al*., 2011). After framing step, each individual frame is windowed using window function. The window function is a mathematically function that is used to minimize signal discontinuities at the beginning and at the end of each frame by taking the block of next frame in consider and integrates all closet frequency lines. This step makes the end of each frame connects smoothly with the beginning of the next (Shaneh and Taheri, 2009; Muda *et al*., 2010; Gawali *et al*., 2011). Figure 6 represents the frame after the window function is applied.

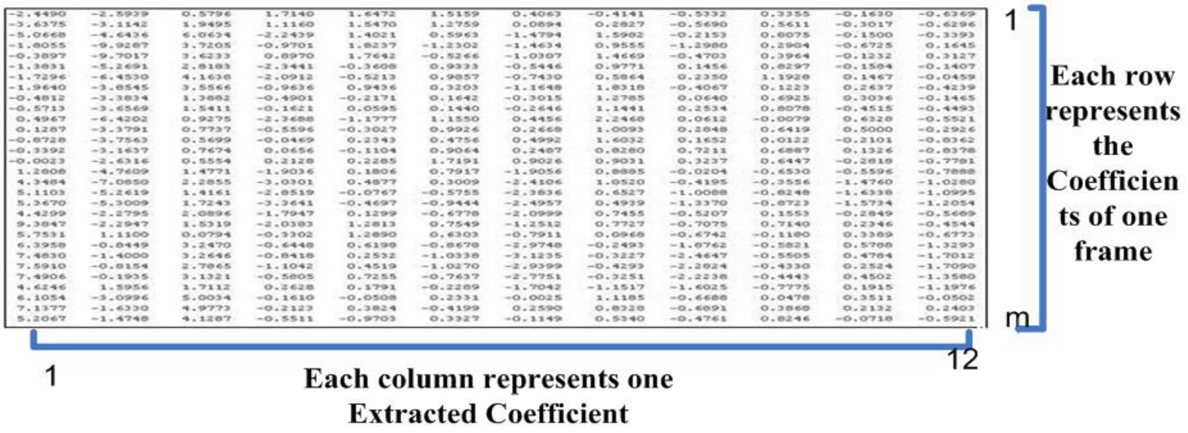
MFCC algorithm: step2: taking fourier transform of a signal: In this step each tone with actual frequency that is measured by hertz is converted into subjective pitch that is measured by scale called "mel" scale.

The main purpose from the process of converting is to mimic the behavior of ear, human ear act as filter, it concentrates on only certain frequency (Shaneh and Taheri, 2009). Figure 7 represents the formula that is used to convert from actual frequency that is measured by Hz into subjective pitch that is measured by "mel" scale.

MFCC algorithm: step3: mapping of powers of spectrum onto mel: In digital signal processing, signals are studied in three domains frequency domain, time domain and wavelet domain. Time domain is a term that is used to describe the domain for analysis of signals with respect to time rather than frequency, when an audio signal is examined in the time domain, the X-axis is time, so the value of the Y-axis depends on the changing of the signal with respect to time (Pope *et al*., 2004).

Frequency domain is a term that is used to describe the domain for analysis of signals with respect to frequency, rather than time. When an audio signal is examined in the frequency domain, the X-axis is frequency, so the value of the Y-axis depends on the changing of the signal with respect to frequency (Pope *et al*., 2004). The Fourier Transform is used to convert each frame from the time domain into the frequency domain (Shaneh and Taheri, 2009; Muda *et al*., 2010).

MFCC algorithm: step 4: taking the logs of powers at each of the mel frequencies: The logarithm of powers is taken at each of the Mel frequencies (Tiwari, 2005; Shaneh and Taheri, 2009).



-2.4490	-2.5939	0.5796	1.7140	1.6472	1.5159	0.4063	-0.4141	-0.5332	0.3355	-0.1630	-0.6369
-3.6375	-3.1142	1.9495	1.1160	1.5470	1.2759	0.0894	0.2827	-0.5690	0.5611	-0.3017	-0.6296
-5.0668	-4.6436	6.0634	-2.2439	1.4021	0.5963	-1.4794	1.5982	-0.2153	0.8075	-0.1500	-0.3393
-1.8055	-9.9287	3.7205	-0.9701	1.8237	-1.2302	-1.4634	0.9555	-1.2980	0.2904	-0.6725	0.1645
-0.3897	-9.7017	3.6233	0.8970	1.7642	-0.5266	-1.0307	1.4669	-0.4703	0.3964	-0.1232	0.3127
-1.3831	-5.2691	2.8183	-2.3441	-0.3608	0.9333	-0.5446	0.9771	0.1456	0.8297	-0.1504	-0.1407
-1.7296	-6.4530	4.1638	-2.0912	-0.5213	0.9857	-0.7430	0.5864	0.2350	1.1928	0.1467	-0.0459
-1.9640	-3.8545	3.5566	-0.9636	0.9436	0.3203	-1.1648	1.8318	-0.4067	0.1223	0.2637	-0.4239
-0.4812	-3.3834	1.3882	-0.4901	-0.2171	0.1642	-0.3015	1.2785	0.0640	0.6925	0.3036	-0.1465
-0.5713	-3.6569	1.5411	-0.1621	0.0595	0.1440	-0.2646	1.1441	0.2534	0.8078	-0.4515	-0.4493
0.4967	-6.4202	0.9275	-2.3688	-1.1777	1.1550	0.4456	2.2468	0.0612	-0.0079	0.6328	-0.5521
0.1287	-3.3791	0.7737	-0.5596	-0.3027	0.9926	0.2668	1.0093	0.2848	0.6419	0.5000	-0.2926
-0.8728	-3.7563	0.5699	-0.0469	0.2343	0.4756	0.4992	1.6032	0.1652	0.0122	-0.2101	-0.8362
-0.3392	-3.1637	0.7674	0.0656	-0.1104	0.9064	0.2487	0.8280	0.7211	0.6887	0.1326	-0.8378
-0.0023	-2.6316	0.5554	0.2128	0.2285	1.7191	0.9026	0.9031	0.3237	0.6447	-0.2818	-0.7781
1.2808	-4.7609	1.4771	-1.9036	0.1806	0.7917	-1.9056	0.8885	-0.0204	-0.6530	-0.5596	-0.7888
4.3484	-7.0850	2.2855	-3.0301	0.4877	0.3009	-2.4106	1.0520	-0.4195	-0.3556	-1.4760	-1.0280
5.1103	-5.2619	1.4161	-2.8519	-0.0767	-0.5755	-2.3836	0.6527	-1.0088	-0.8248	-1.6338	-1.0995
5.3670	-5.3009	1.7243	-3.3641	-0.4697	-0.9444	-2.4957	0.4939	-1.3370	-0.8723	-1.5734	-1.2054
4.4299	-2.2795	2.0896	-1.7947	0.1299	-0.6778	-2.0999	0.7455	-0.5207	0.1553	-0.2849	-0.5689
9.3847	-2.2947	1.5319	-2.0383	1.2813	0.7549	-1.2512	0.7727	-0.7075	0.7140	0.2346	-0.4544
5.7531	1.1100	0.0794	-0.3302	1.2890	0.6303	-0.7911	0.0968	-0.6742	-0.1180	0.3389	-0.6773
6.3958	-0.8449	3.2470	-0.6448	0.6198	-0.8678	-2.9748	-0.2493	-1.8762	-0.5821	0.5788	-1.3293
7.4830	-1.4000	3.2646	-0.8418	0.2532	-1.0338	-3.1235	-0.3227	-2.4647	-0.5505	0.4784	-1.7012
7.5910	-0.8154	2.7865	-1.1042	0.4519	-1.0270	-2.9399	-0.4293	-2.2824	-0.4330	0.2524	-1.7090
7.4906	-0.1935	3.1321	-0.5805	0.7255	-0.7637	-2.7751	-0.3251	-2.2238	-0.4443	0.4502	-1.3580
4.6246	1.5956	1.7112	0.2628	0.1791	-0.2289	-1.7042	-1.1517	-1.6025	-0.7775	0.1915	-1.1976
6.1054	-3.0996	5.0034	-0.1610	-0.0508	0.2331	-0.0025	1.1185	-0.6688	0.0478	0.3511	-0.0502
7.1377	-1.6330	4.9773	-0.2123	0.3824	-0.4199	0.2590	0.8328	-0.6891	0.3868	0.2132	0.2403
5.2067	-1.4748	4.1287	-0.5511	-0.9703	0.3327	-0.1149	0.5340	-0.4761	0.8246	-0.0718	-0.5921

1
m
Each row represents the Coefficients of one frame
1
12
Each column represents one Extracted Coefficient

Fig. 8: The numeric representation of MFCC matrix

MFCC algorithm: step 5: taking the discrete cosine transform of the list of mel log powers: This step is to convert the log mel spectrum into the time domain using discrete cosine transform. The MFCC feature is the amplitudes of the resulting spectrum, this representation of speech spectrum provides a good representation of the local spectral properties. The results are produced from this step is called mel frequency cepstrum coefficients (Tiwari, 2005; Shaneh and Taheri, 2009; Muda *et al*., 2010; Salvador and Chan, 2007). Figure 8 represents the numeric representation of MFCC matrix. Each row represents the coefficients of one frame and each column represents one extracted coefficient. ARANEWS system uses twelve coefficients for each frame.

Dynamic time warping "DTW": Any information retrieval system needs techniques to calculate distance between training data and testing data. ARANEWS has used DTW algorithm as a technique for this purpose.

DTW is an algorithm that based on dynamic programming. It is used to measure similarity between two time series that may vary on time or speed. DTW is used in speech recognition to determine if the two waveforms represent the same spoken word (Muda *et al*., 2010; Gawali *et al*., 2011; Salvador and Chan, 2007).

DTW algorithm is used to compute the best possible alignment between test data (submitted query) and training data (stored keywords, references). DTW finds optimal alignment between two time series, one time series may be "warped" non-linearly by stretching or by shrinking it along time axis (Muda *et al*., 2010). Figure 9 represents how one time series warped to another time series. Each vertical line connects a point in one time series to its correspondingly similar point in other time series, as shown in Fig. 9 the two blue, red and green points are similar. If two time series are identical, no warping would be necessary and all lines would be straight vertical.

DTW is used dynamic programming approach to find minimum distance warp path. Figure 10 represents time to time matrix. The input pattern (query) goes along button. The input "SsPEEhH" is a noisy version from the keyword "SPEECH". The template "stored keyword" goes along up side, the idea is letter ’h’ is closer to match ’H’ compared with any letter in the template (http://haroon.site11.com/dtw.html).

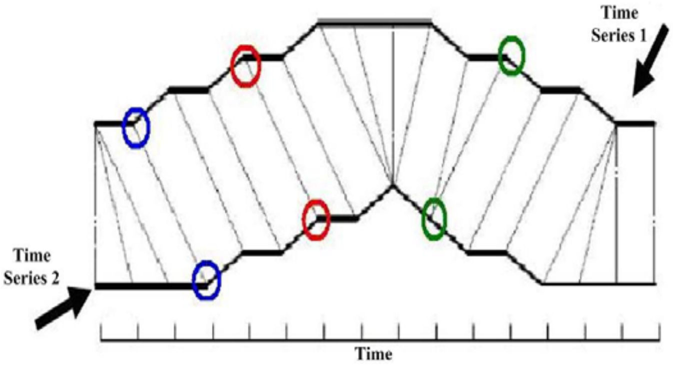


Fig. 9: One time series warped to another time series

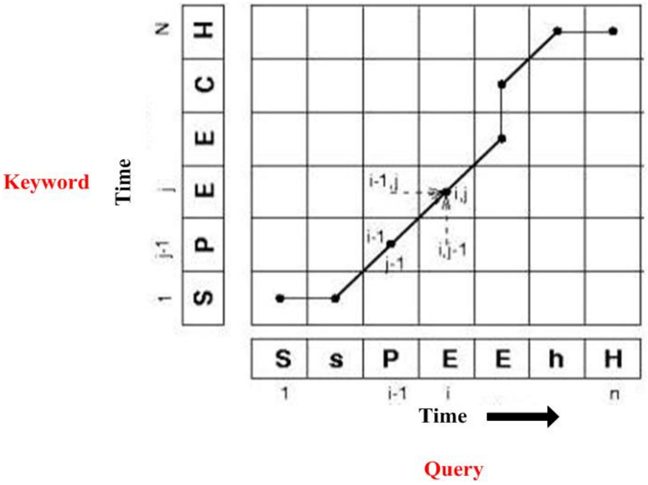


Fig. 10: Time to time matrix

DTW computes an n-by-m matrix, where the (ith, jth) element of the matrix contains the constructed distance d (q_i, c_j) between the two points q_i and c_j. Then, the absolute distance between the values of two sequences is calculated using the Euclidean distances shown in Eq. (1) (Muda *et al*., 2010; Thakur *et al*., 2011; Bala *et al*., 2010).

Each element (i, j) in matrix corresponds to the alignment between the points q_i and c_j. Then, accumulated distance is measured by Eq. (2) (Muda *et al*., 2010; Thakur *et al*., 2011; Bala *et al*., 2010):

$$d(q_i, c_j) = d(q_i, c_j) \quad (1)$$

$$D(i, j) = \min [D(i-1, j-1), D(i-1, j), D(i, j-1)] + d(i, j) \quad (2)$$

ARANEWS architecture: ARANEWS architecture is composed of three modules. The first module is an input module "Administrator side". The second module is the query module "client side" and the third module is the retrieval module. Each module works on achieving specific tasks to retrieve Arabic audio news. The idea of the proposed system is based on dependency between the input module and query module. The person who submits the queries is the same person who will record the Arabic keywords that control the retrieval operation. Figure 11 represents the architecture of ARANEWS system.

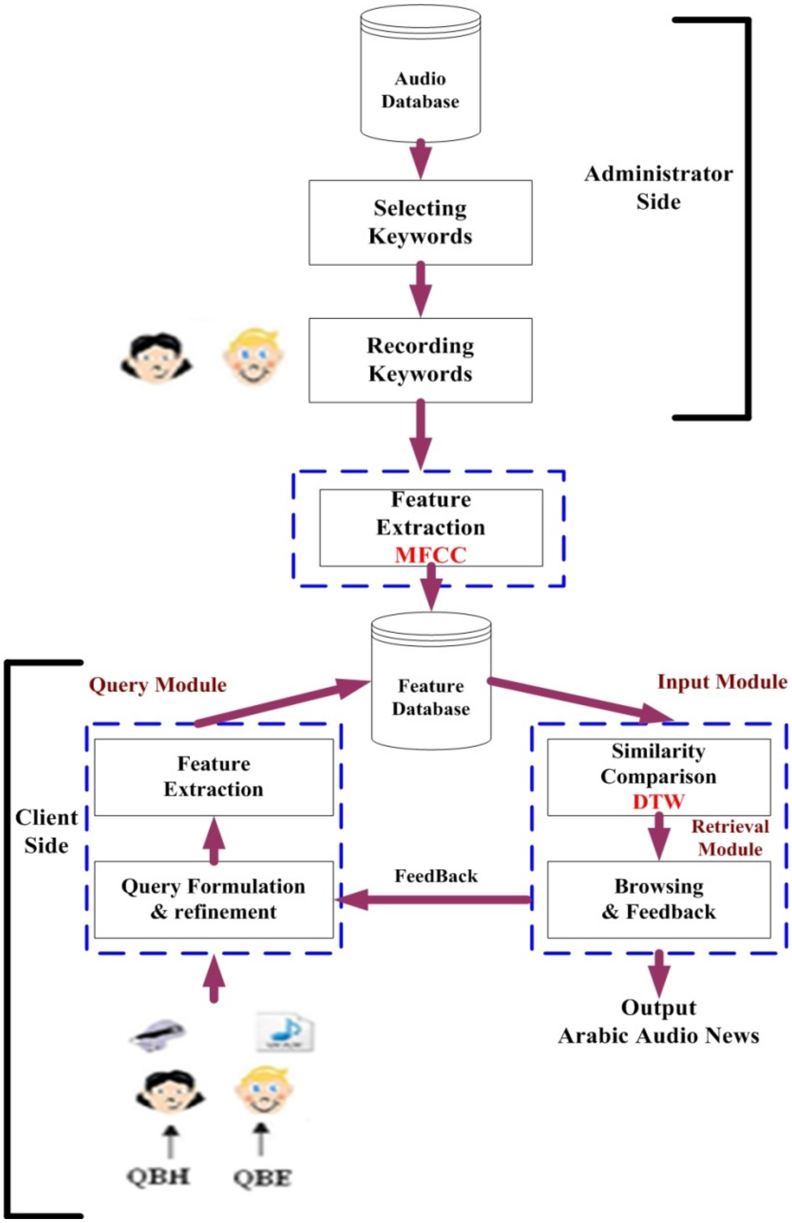


Fig. 11: The architecture of ARANEWS system

Input module "administrator side": The main task of the input module "Administrator Side" is to prepare the collections of Arabic audio news (Audio Database) which will be retrieved from the query module "client side". Each collection is concerned with a common subject and it is indexed by using a group of audio keywords. For example the administrators collect a group of Arabic audio news that concerned with the football. This group is indexed by using a group of Arabic audio keywords such as "Messi", "Barcelona" and "Real Madrid". Each keyword has to be recorded many times by the user of the system to constitute templates for it. All recorded audio keywords are refined using specific algorithms to remove silence and noise from them. The approach of selecting audio keywords that are indexed Audio news and the enhanced speech techniques are discussed later. Preparing training dataset for the input module is shown in Fig. 12.

The MFCC feature is extracted from all recorded refined keywords to constitute the training data "references" of the system. Training data are stored within feature database "ARANEWSDB" to be used in the matching process with testing data that submitted from query module.

Query module "client side": The main task of query module is formulating and refining queries that are submitted by user. The user can submit QBH or QBE in ARANEWS system. Both types of query are refined using algorithms that remove silence and noise from them. The algorithms that are used to refine query in the Query module is the same as algorithms that are used to refine training data in the input module. MFCC feature is extracted from submitted query to constitute testing data which will be matched with training data in retrieval module.

If the user submits query using QBH, he must determine the length of query before recording it via microphone. ARANEWS system allows four lengths for QBH which are two seconds, three seconds, four seconds and five seconds. After the recording process, the user must press search button in the client screen to retrieve audio news based on the submitted recorded keyword. If the user submits query using QBE, he must click on the browse button to select file that contains audio keyword and then he can press search button in the client screen to retrieve audio news.

Retrieval module: The main task of retrieval module is measuring similarity between training data and testing data to retrieve Arabic audio news. DTW algorithm is

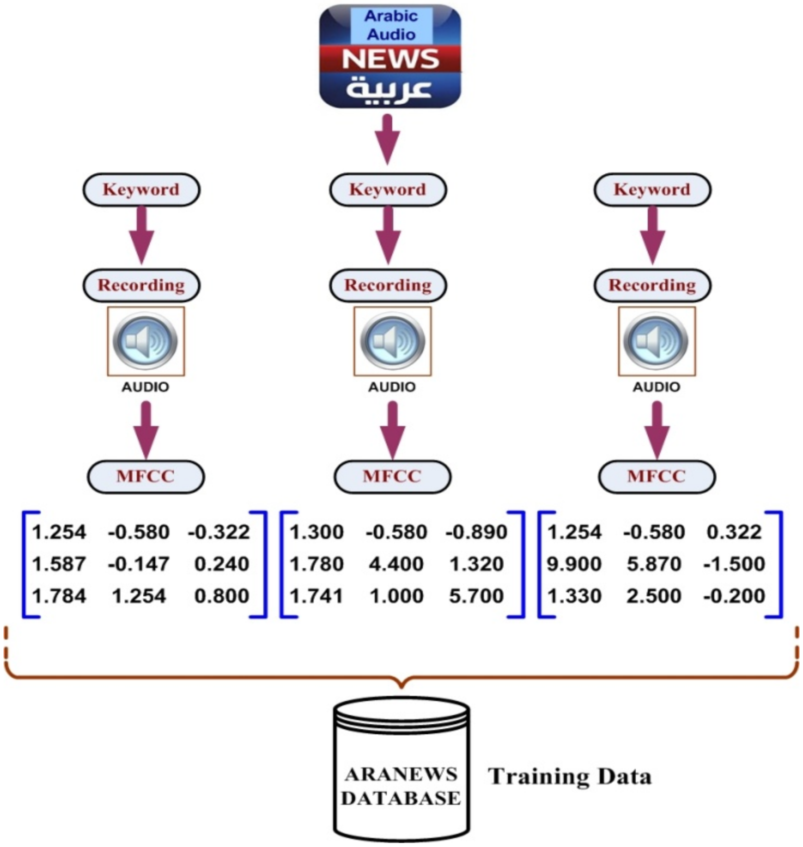


Fig. 12: Preparing training dataset for the input module

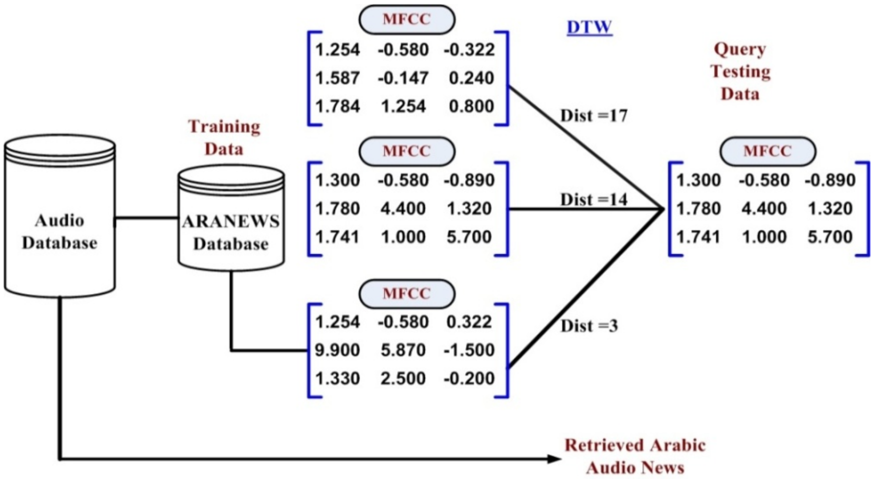


Fig. 13: Calculating distances between training data and testing data using DTW algorithm

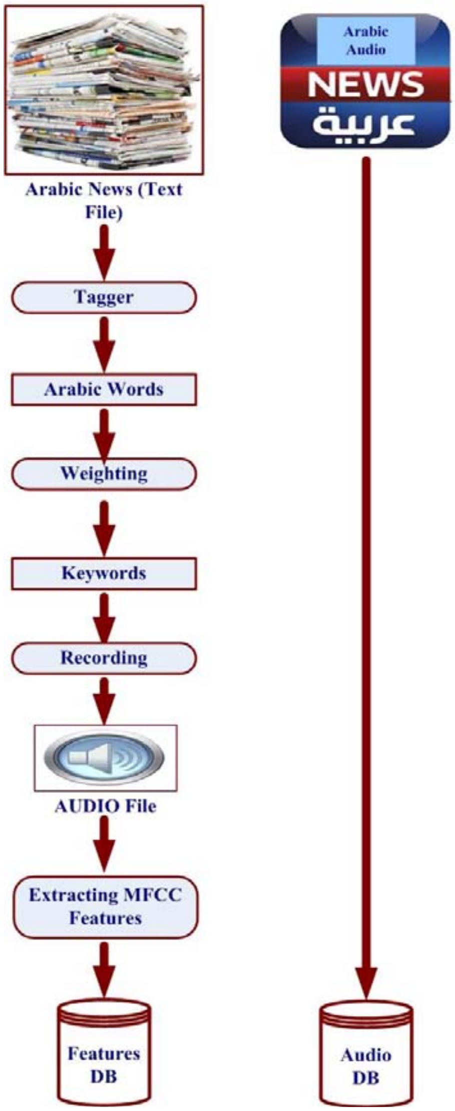


Fig. 14: The proposed technique to select keywords for ARANEWS

used to measure distances between extracted MFCC feature from query and extracted MFCC features from training data. The calculated distances is sorted in ascending order and all Arabic audio news that are indexed by the keyword which gets the minimum distance will be retrieved to the user. Figure 13 clears this process.

A proposed technique for selecting keywords: ARANEWS system selects keywords based on news text files, each selected keyword will be recorded many times by the user of system to prepare the training data at the input module. The main reason of using text files in selecting keywords process instead of analyzing the audio news and extracting the audio keywords directly is the lacking of researches that work on analyzing the long Arabic audio clip.

Each Arabic news in ARANEWS system is gathered as text file and audio file. Both files retain the same news; the test file is passed through Arabic tagger. The Arabic tagger returns separated words after removing stop words and particles. Each word comes from tagger will be weighted. The words with the highest weight will be selected keywords that adopted to control the retrieval process. Figure 14 represents the proposed technique to select keywords for ARANEWS system.

Speech enhancement techniques: ARANEWS system supports using QBH and QBE, in both cases the recording process may be affected by the nature of environment "clean vs. noisy" and the quality of microphone. The noise effects on the accuracy of original signals which necessarily affect on the accuracy of extracted MFCC features, the accuracy of extracted MFCC will affect by its role on the calculated distances and the retrieved audio news.

Training data and testing data passes through two gates before extracting MFCC features, removing

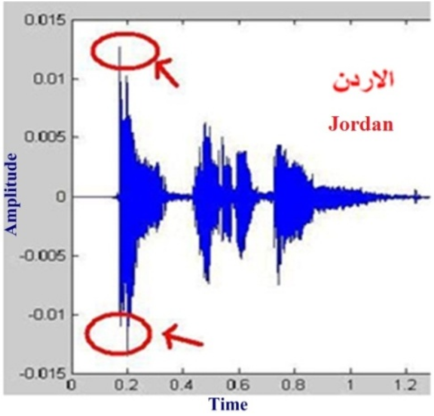


Fig. 15: Signals of Arabic keyword JORDAN before using WIENER filter

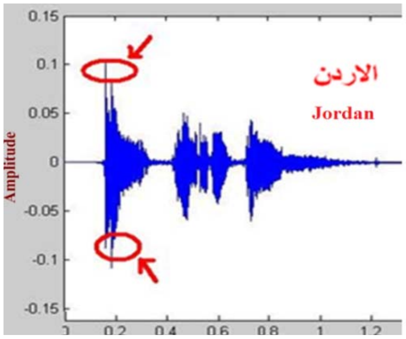


Fig. 16: Signals of Arabic keyword JORDAN after using WIENER filter

silence gate and removing noise gate. Removing silence gate works on removing silence from the beginning and the end of recorded data. There are no meaningful data can be extracted from silence parts in speech signals, so processing these parts is considered as overhead working on the system.

Removing noise gate works on reducing the percentage of noise that destroys original signals. WIENER filter has been exploited as removing silence gate within ARANEWS system, it is used in different sectors such as communication, signal processing and voice recognition.

WIENER filter was introduced by Robert Wiener and it was published in 1949. WIENER filter has been designed on the assumption that we have knowledge about the properties of the original signals and noise, it extracts signal from noise and reproduce signals as accurately as possible with noise reduction. Figure 15 and 16 represent the signals of recorded Arabic keywords before and after using WIENER filter.

EXPERIMENTAL SETUP AND RESULTS

The accuracy of speech recognition component: The performance of ARANEWS system is based on the accuracy of speech recognition component that has been used. MFCC feature and DTW are used in speech recognition component. If speech recognition component recognizes the query "keyword" correctly, then all Arabic audio news that has been indexed by this audio keyword will be retrieved.

Three factors have been chosen to be examined in the experiments. The first factor is the size of dictionary "number of audio keywords" that indexed Arabic Audio news. The second factor is the noise and the third one is the number of templates "how many times each keyword will be recorded". The length of query in all experiments is two seconds.

All experiments have been designed as follow. The user at client will examine each keyword three times with ARANEWS by submitting three queries for each keyword to ARANEWS. If ARANEWS retrieves related audio news, then this attempt will be considered successful attempt, otherwise it will be considered as failed attempt. The overall accuracy for ARANEWS system is calculated using Eq. (3):

$$\text{Accuracy (\%)} = \frac{\text{Total of successful attempts}}{\text{Total number of all attempts}} \qquad (3)$$

The impact of dictionary size on the accuracy: This experiment is reiterated four times with four different sizes of dictionary which are 5, 10, 15 and 20 keywords, respectively. The experiment has not been carried out in standard environments (noise free environment). The experiment has been carried out in environment with a low level of noise. The following list depicts the training requirements for the current experiment:

- **Speaker:** One female "Ayat" at administrator side and the same female at client side
- **Tools:** Microphone with normal specification at both sides
- **Environment:** Low level of noise
- **Sample rate for recording process:** Twenty two thousand and twenty five Hz at both sides
- **Bit depth for recording process:** Sixteen bits at both sides
- **The number of channels for recording process:** Mono
- **The number of templates for each keyword:** One template

Table 1 shows the relationship between the various dictionary sizes with respect to the corresponding accuracies. The results show that when the size of dictionary increases the accuracy of system decreases.

The impact of noise on the accuracy: Noise is one of the most dangerous enemies that affect on the performance of speech recognition component. This experiment has been reiterated twelve times with three environments which are environment with low level

Table 1: The impact of dictionary size on accuracy

Exp. No.	Dictionary size	Accuracy (%)	Environment	Query length (sec)	Avg.
1	5	100	Low noise	2	85.32
2	10	83.30	Low noise	2	
3	15	82.20	Low noise	2	
4	20	76.60	Low noise	2	

Table 2: The impact of noise level on accuracy

Exp. No.	Dictionary size	Accuracy (%)	Environment	Query length (sec)	Avg. (%)
1	5	100.00	Low level noise	2	85.32
2	10	83.30		2	
3	15	82.20		2	
4	20	76.60		2	
5	5	93.30	Medium level noise	2	76.00
6	10	70.00		2	
7	15	71.10		2	
8	20	70.00		2	
9	5	80.00	High level noise	2	53.50
10	10	56.50		2	
11	15	37.70		2	
12	20	40.00		2	

noise, environment with medium level noise and environment with high level noise. The following list depicts the training requirements for Experiment 2:

- **Speaker:** One female "Ayat" at administrator side and the same female at client side
- **Tools:** Microphone with normal specification at both sides
- **Environment:** Three levels of noises: low, medium and high
- **Sample rate for recording process:** Twenty two thousand and twenty five Hz at both sides
- **Bit depth for recording process:** Sixteen bits at both sides
- **The number of channels for recording process:** Mono
- **The number of templates for each keyword:** One template

Table 2 shows the relationship between the various noise levels with respect to the corresponding accuracies. The results show that the accuracy of speech recognition component decreases when the level of noise increases.

The impact of number of templates on the accuracy: Each keyword can be pronounced in different speed from the same user. Logically if the number of recorded copies for each keyword increases, the number of extracted MFCC copies for each keyword "templates" will increase and that will cause increasing the percentage of accuracy for speech recognition component. Figure 17 represents three templates for Arabic Keyword JORDAN that has been uttered by the same speakers with different speeds of pronunciation. This experiment has been reiterated twelve times with three different numbers of templates which are: 1, 3 and 6 templates, respectively. The following list depicts the training requirements for Experiment 3:

- **Speaker:** One female "Ayat" at administrator side and the same female at client side
- **Tools:** Microphone with normal specification at both sides
- **Environment:** Low level of noise
- **Sample rate for recording process:** Twenty two thousand and twenty five Hz at both sides
- **Bit depth for recording process:** Sixteen bits at both sides
- **The number of channels for recording process:** Mono
- **The number of templates for each keyword:** One, three and six templates, respectively

Table 3 shows the relationship between the various numbers of templates with respect to the corresponding accuracies. The results show that when the number of templates increases to six, the accuracy of speech recognition component increases explicitly.

The overall accuracy of ARANEWS audio retrieval system:

Preparing datasets: Two thousands Arabic news have been collected to evaluate the performance of ARANEWS system. Each one hundred Arabic news is classified under one category and each category is indexed by one Arabic audio keyword. Each news is collected as audio news and as text news. Sound Forge tool has been used to extract audio tracks from some video files that contain the Arabic audio news. Figure 18 depicts the preprocessing steps of collecting audio Arabic news.

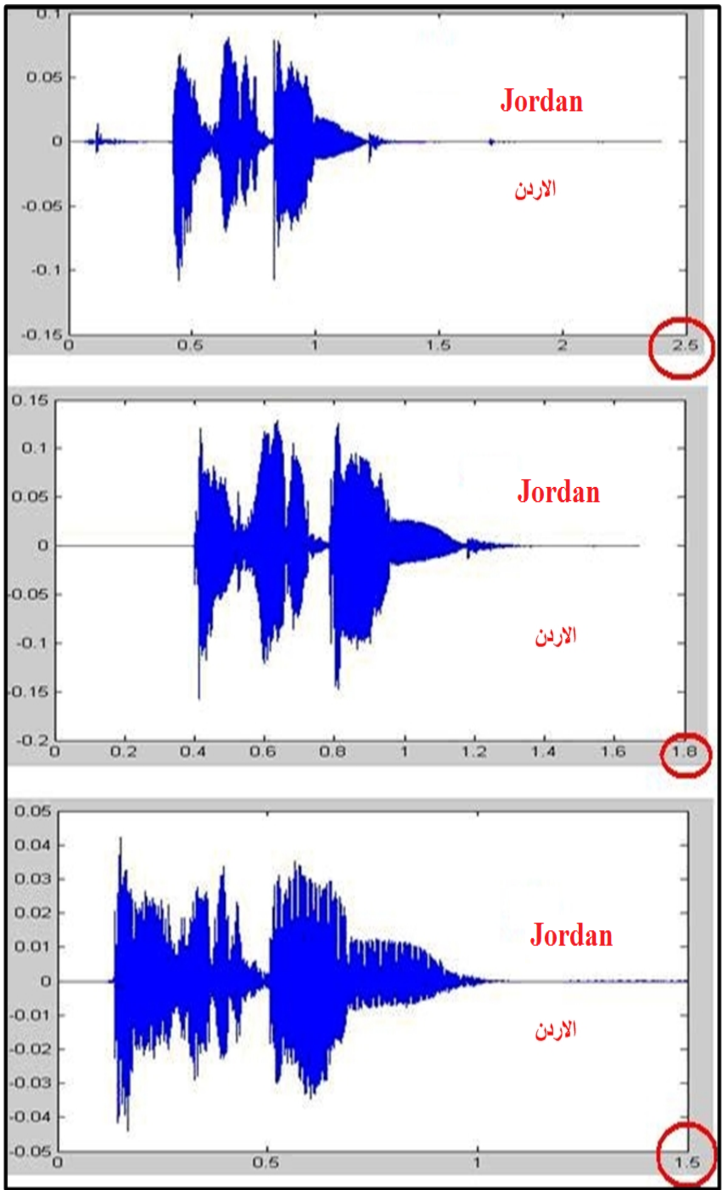


Fig. 17: Three templates for Arabic keyword JORDAN that has been uttered by the same speakers with different speeds of pronunciation

Table 3: The impact of number of templates on the accuracy

Exp. No.	Dictionary size	Accuracy (%)	No. of templates	Query length (sec)	Avg. (%)
1	5	100.00	One template	2	81.25
2	10	80.00		2	
3	15	80.00		2	
4	20	66.60		2	
5	5	100.00	Three templates	2	81.25
6	10	80.00		2	
7	15	80.00		2	
8	20	66.60		2	
9	5	100.00	Six templates	2	89.25
10	10	90.00		2	
11	15	91.10		2	
12	20	76.60		2	

Standard measures: Recall, precision and F-measure are standard metrics expressing the quality of information retrieval methods (Lu and Sajjanhar, 1998; Van Rijsbergen, 1979). Recall measures the ability of

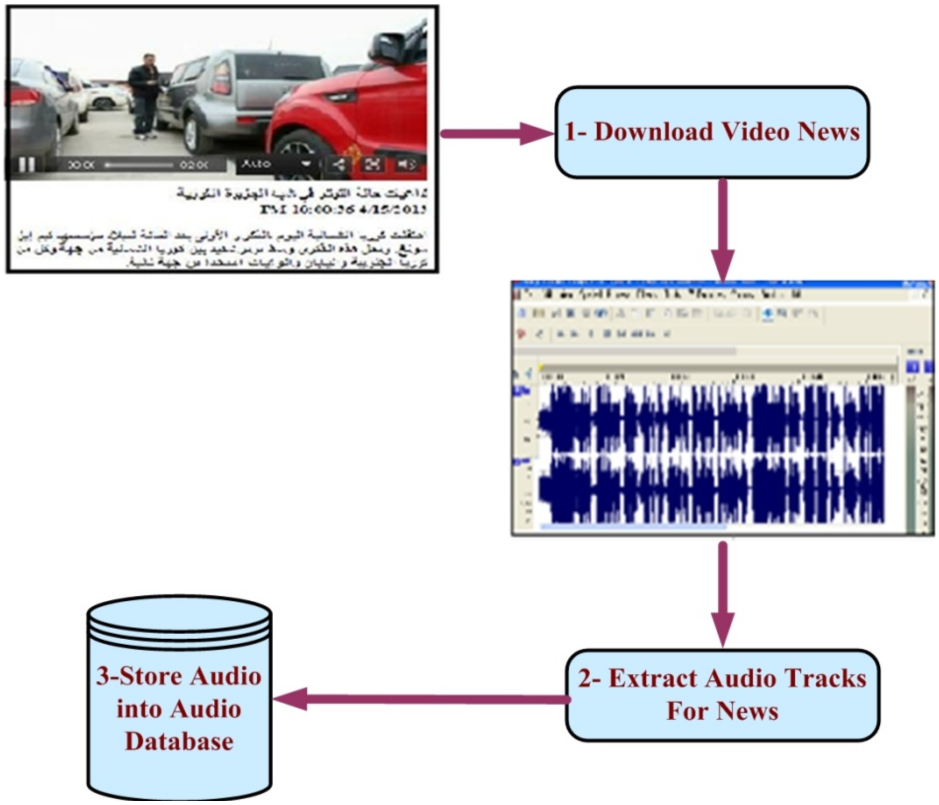

1- Download Video News
2- Extract Audio Tracks For News
3-Store Audio into Audio Database

Fig. 18: The preprocessing steps of collecting audio Arabic news

Table 4: The performance of ARANEWS with four different sizes of dictionary

Exp. No.	Dictionary size	Noise level	Query length (sec)	Avg. of recall	Avg. of precision	Avg. of F-measure
1	5	Low level	2	1.000	1.000	1.000
2	10	Low level	2	0.833	0.833	0.833
3	15	Low level	2	0.832	0.832	0.832
4	20	Low level	2	0.766	0.766	0.766

retrieving relevant items from the database. It is the ratio of the number of relevant items retrieved to the total number of relevant items in the database:

$$\text{Recall} = \frac{\text{Number of relevant items retrieved}}{\text{Total number of relevant items}} \qquad (4)$$

Precision measures the retrieval accuracy and is defined as the ratio between the number of relevant items retrieved and the number of total items retrieved (Van Rijsbergen, 1979). While, F-measure is a measure that combines between precision and recall:

$$\text{Precision} = \frac{\text{Number of relevant items retrieved}}{\text{Total number of retrieved items}} \qquad (5)$$

$$\text{F} - \text{Measure} = 2 \times \frac{\text{Precision}\times\text{Recall}}{\text{Precision} + \text{Recall}} \qquad (6)$$

Three queries have been submitted to examine each keyword. Recall, precision and F-measure are calculated for each query. The average of recall, the average of precision and the average of F-measure are calculated for all queries that have been submitted to ARANEWS system.

ARANEWS evaluation: The performance of ARANEWS is examined with three factors which are the size of dictionary, noise and the number of templates. Table 4 represents the performance of ARANEWS with four different sizes of dictionary. The results show that the recall, precision and F-measure decrease when the size of dictionary increases.

Table 5 represents the performance of ARANEWS with three different environments which are low level noise environment, medium level noise environment and high level noise environment. The results show that the noise decreases the value of recall, precision and F-Measure.

Table 6 represents the performance of ARANEWS with three different numbers of templates. The results show that the values of recall, precision and F-Measure increases when the number of templates increases. The number of templates affects on the retrieval time. When

Table 5: The performance of ARANEWS with three different environments

Exp. No.	Dict. size	Env.	Query length (sec)	Accuracy (%)	Avg. of recall	Avg. of precision	Avg. of F-measure
1	5	Low level noise	2	85.32	1.000	1.000	1.000
2	10		2		0.833	0.833	0.833
3	15		2		0.833	0.833	0.833
4	20		2		0.766	0.766	0.766
5	5	Medium level noise	2	76.00	0.933	0.933	0.933
6	10		2		0.700	0.700	0.700
7	15		2		0.711	0.711	0.711
8	20		2		0.700	0.700	0.700
9	5	High level noise	2	53.50	0.800	0.800	0.800
10	10		2		0.566	0.566	0.566
11	15		2		0.377	0.377	0.377
12	20		2		0.400	0.400	0.400

Table 6: The performance of ARANEWS with 3 different numbers of templates reading

Exp. No.	Dict. size	Env. (%)	Query length	Accuracy (sec)	Avg. of recall	Avg. of precision	Avg. of F-measure
1	5	100.00	One template	2	1.000	1.000	1.000
2	10	80.00		2	0.800	0.800	0.800
3	15	80.00		2	0.800	0.800	0.800
4	20	66.60		2	0.650	0.650	0.650
5	5	100.00	Three templates	2	1.000	1.000	1.000
6	10	80.00		2	0.800	0.800	0.800
7	15	80.00		2	0.800	0.800	0.800
8	20	66.60		2	0.650	0.650	0.650
9	5	100.00	Six templates	2	1.000	1.000	1.000
10	10	90.00		2	0.900	0.900	0.900
11	15	91.10		2	0.911	0.911	0.911
12	20	76.60		2	0.766	0.766	0.766

the number of template increases, the retrieval time increases. Content based audio retrieval systems that use template based approach must achieve the trade off between the number of templates and the retrieval time.

CONCLUSION

ARANEWS system works efficiently with small dictionary size, reasonable number of templates and if it is used in environment with low level of noise. The approaches, algorithms and feature that have been invested to build ARANEWS can be used with all languages. ARANEWS system represents step toward controlling retrieving process by voice, such as this kind of retrieving system facilitates the life of many persons who are suffering from disabilities. Developing content based audio retrieval system is very rich field for research and it needs more investigation.

REFERENCES

Ali, M., M. Hossain and M. Bhuiyan, 2013. Automatic speech recognition technique for Bangla words. Int. J. Adv. Sci. Technol., 50: 51-60.

Bala, A., A. Kumar and N. Birla, 2010. Voice command recognition system based on MFCC and DTW. Int. J. Eng. Sci. Technol., 2(12): 7335-7342.

Dhingra, S., G. Nijhawan and P. Pandit, 2013. Isolated speech recognition using MFCC and DTW. Int. J. Adv. Res. Electr. Electron. Instrum. Eng., 2(8): 4085-4092.

Fujii, A., K. Itou and T. Ishikawa, 2002. Speech-driven Text Retrieval: Using Target IR Collections for Statistical Language Model Adaptation in Speech Recognition. In: Coden, A.R., E.W. Brown and S. Srinivasan (Eds.), IR Techniques. Springer-Verlag, Berlin, Heidelberg, LNCS 2273, pp: 94-104.

Gaikwad, S., B. Gawali and P. Yannawar, 2010. A review on speech recognition technique. Int. J. Comput. Appl., 10(3): 16-24.

Gawali, B.W., S. Gaikwad, P. Yannawar and S.C. Mehrotra, 2011. Marathi isolated word recognition system using MFCC and DTW features. Int. J. Inform. Technol., 1(1): 21-24.

Helén, M. and T. Lahti, 2006. Query by example methods for audio signals. Proceeding of the 7th Nordic Signal Processing Symposium. Reykjavik, pp: 302-305.

Lu, G. and A. Sajjanhar, 1998. On performance measurement of multimedia information retrieval systems. Proceeding of the International Conference on Computational Intelligence and Multimedia Applications. Monash University, pp: 781-787.

Mitrovic, D., M. Zeppelzauer and C. Breiteneder, 2010. Features for content-based audio retrieval. Adv. Comput., 78: 71-150.

Muda, L., M. Begam and I. Elamvazuthi, 2010. Voice recognition algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) techniques. J. Comput., 2(3): 138-143.

Picone, J., 1993. Signal modeling techniques in speech recognition. P. IEEE, 81(9): 1215-1247.

Pope, S., F. Holm and A. Kouznetsov, 2004. Feature extraction and database design for music software. Proceedings of the International Computer Music Conference, pp: 596-603.

Ratanamahatana, C. and P. Tohlong, 2006. Speech Audio Retrieval using Voice Query. In: Sugimoto, S. *et al*. (Eds.), ICADL 2006. Springer-Verlag Berline Heidelberg, LNCS 4312, pp: 494-497.

Reddy, D., 2005. Speech recognition by machine: A review. P. IEEE, 64(4): 501-531.

Salvador, S. and P. Chan, 2007. Toward accurate dynamic time warping in linear time and space. Intell. Data Anal., 11(5): 561-580.

Shaneh, M. and A. Taheri, 2009. Voice command recognition system based on MFCC and VQ algorithms. World Acad. Sci. Eng. Technol., 33: 534-538.

Thakur, A., N. Singla and V. Patil, 2011. Design of Hindi key word recognition system for home automation system using MFCC and DTW. Int. J. Adv. Eng. Sci. Technol., 11(1): 177-182.

Thakur, A. and N. Sahayam, 2013. Speech recognition using Euclidean distance. Int. J. Emerg. Technol. Adv. Eng., 3(3).

Tiwari, V., 2005. MFCC and its applications in speaker recognition. Int. J. Emerg. Technol., 1(1): 19-22.

Van Rijsbergen, C.J., 1979. Information Retrieval. 2nd Edn., Butterworths, London.