20

Analysis and Implementation of an Automated Delimiter of "Quranic" Verses in Audio Files using Speech Recognition Techniques

Tabbal Hassan, Al-Falou Wassim and Monla Bassem
Lebanese University
Lebanon

1. Introduction

With more than 300 audio recitations of the Holy Quran available free on the internet, there is an increasing need to synchronize textual information with the audio recitation. The synchronized text can contain simply the textual representation of the recited verses or more information like the translation to a foreign language and the meaning of difficult verses. The general approach used nowadays consists of manually marking the beginning and ending of every verse. Taking into account the special way to recite Quran : "The art of tajweed", the manual method required a lot of work and proved to be unable to adapt to new recitors .

The goal of this chapter is to try the applicability and effectiveness of a new approach that uses common speech recognition techniques to automatically find and delimit verses in audio recitations regardless of the recitor. "The art of tajweed" defines some flexible yet well-defined rules to recite the Quran creating a big difference between normal Arabic speech and recited Quranic verses, thus it is interesting to analyze the impact of this "art" on the automatic recognition process and especially on the acoustic model. The study uses the Sphinx Framework (Carnegie Mellon University) as a research environment.

2. The sphinx IV framework

Sphinx is an open source (since version 2) project, financed by DARPA and developed at the Carnegie Melon University CMU in Pittsburgh with the contribution of Sun Microsystems, Mitsubishi Electric Research Lab, Hewlett Packard, University of California at Santa Cruz and the Massachusetts Institute of Technology. The sphinx-4 framework is a rewrite in java of the original sphinx engine. It follows a modular and extensible architecture to suit the needs of the researchers (Sphinx Group,2004).

Source: Robust Speech Recognition and Understanding, Book edited by: Michael Grimm and Kristian Kroschel,
ISBN 987-3-90213-08-0, pp.460, I-Tech, Vienna, Austria, June 2007

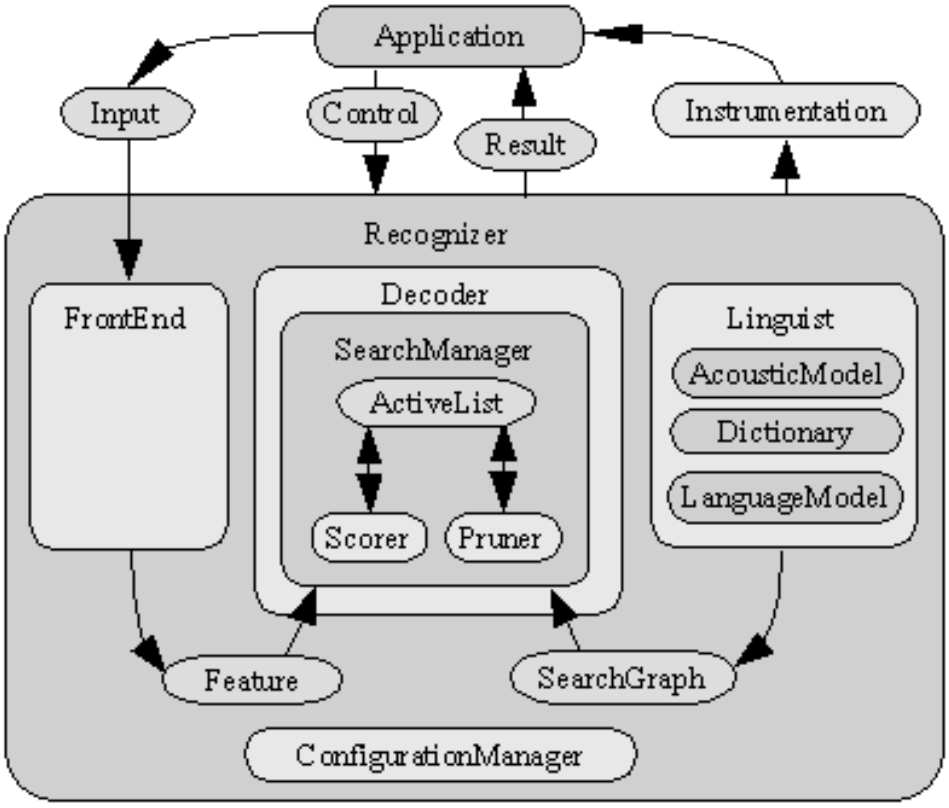


Figure 1. The functional bloc diagram of Sphinx-4

Sphinx-4 (Sphinx Group,2004) is based on HMM (Hidden Markov Model), and provides the researcher with a tool: SphinxTrain for the development of Acoustic models. The "Configuration Manager" uses an external configuration file to bind, at run-time, each part of the system with the corresponding algorithm. This flexibility makes it a practical choice for researchers because of the ability to test different algorithms or even develop new ones without the need of recompilation of the application. Once properly configured, the "FrontEnd" must either receive its input from audio files (batch mode) or directly from the microphone (live mode). Although not implemented yet, but the sphinx team envisaged the ability to switch between live mode and batch mode at run-time. The "frontend" generates the feature vectors (Cepstrum, delta cepstrum and delta delta cepstrum).The generated vectors will then be processed by the decoder that uses information from both the language model and the acoustic model to generate the search space for the HMM nodes. A number of search algorithms can be specified (Breadth-First, Beam Search, Viterbi, A*,…). The result of the decoding phase will then be returned to the calling application.

3. Holy Quran recitations and acoustic model

3.1 The "Art of Tajweed"

The recitation of the holy Quran differs from the normal reading of Arabic text due to a special art: "Fan al tajweed". "Tajweed" is considered as an art because not all recitors will perform the same verses in the same way. Furthermore, the same recitor may perform the same verses differently due to the flexibility of the laws of tajweed. Another word to describe the art of reciting Quran is "Tarteel". While obeying to the same laws of "tajweed", "tarteel" is generally identified by a faster reading pace in contrast with "tajweed" where reading is slower and where there is more focus on the artistic (musical) aspect of reading the holy Quran. "Tarteel" is the preferred reading style of recitors from Gulf countries while Syrian, Egyptian, Lebanese and other Middle Eastern recitors, prefers "tajweed".

There is 10 different law sets according to the 10 certified scholars[Hafs, Kaloun, Warsh,….] who teached the recitation of the Holy Quran (M.Habash,1998). Furthermore, recitors tend to vary the tone of their recitations according to musical "maqams" (The basic number of maqams is seven but there are other variations resulted from the combination of different maqams).

3.2 Impact of the "Art of Tajweed" on the acoustic model

The laws of "tajweed" introduce additional difficulties to the inherently difficult Arabic Speech Recognition problem. The most important part in our project was to identify what aspects of the laws of tajweed will affect the recognition phase and for which factors. After that analysis phase, we assumed that the seven "maqams" does not require any special treatment because of the statistical nature of the Hidden Markov Model (HMM). We considered only the laws of the art of Tajweed according to Hafs (used in 98% of the recitations) and found the following laws to have the most influence on the recognition of a specific recitation:

- Necessary prolongation of 6 vowels
- Obligatory prolongation of 4 or 5 vowels
- Permissible prolongation of 2,4 or 6 vowels
- Normal prolongation of 2 vowels
- Nasalization (ghunnah) of 2 vowels
- Silent unannounced letters
- Emphatic pronunciation of the letter R

Note that there is also the echoing sound that is produced with some unrest letters but we found that it has no effect on the recognition because the echo will be considered as noise and thus the noise-canceling filter will eliminate it.

In order to deal with these rules, we considered the prolongation as the repetition of the vowel n-corresponding times. The same consideration was used for the nasalization. The emphatic pronunciation of R led us to introduce another phoneme, or voice, that will also be used with other emphatic letters such as Kaf, Khaa when they are voweled by a fatha.

This conclusion can be verified by examining thoroughly the spectrogram of the Quranic recitations.

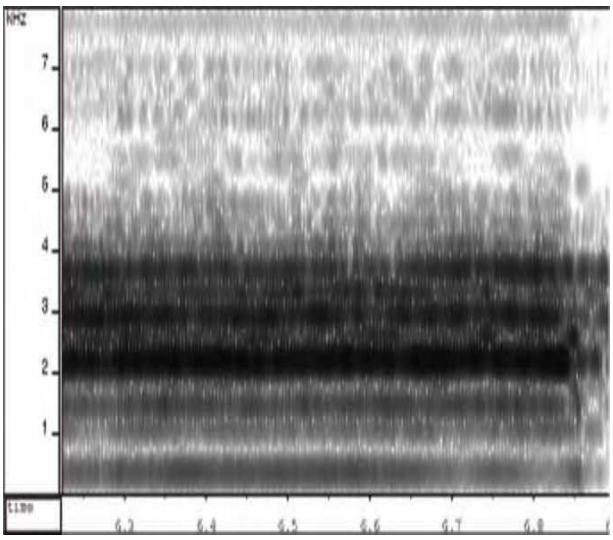


Figure 2. Spectrogram of a 6 vowels prolongation in "Bismillahi Al-Rahmani Al-Rahim"

The above spectrogram shows us that the six vowels prolongation of the “I” in “Al-Rahim” does not present so much variation, so considering it as the repetition of six “I” is a correct assumption.

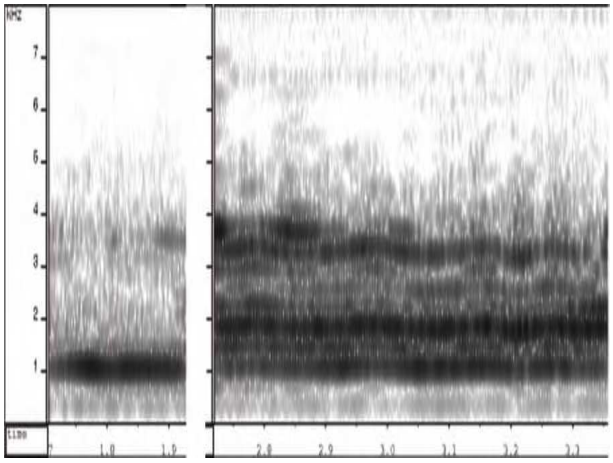


Figure 3. Comparaison between 2 "fatha" one emphatic with the letter R (left) and a normal with 2 vowels prolongation

If we compare the emphatic “fatha” and the normal “fatha”, the differences between their spectrogram justifies the need to have two different phonemes representing each one of them. (All spectrograms were generated by the Open source tool WaveSurfer)
We can deduce from this study the following set of phonemes:

Symbol	Alphabet	Transliteration
AA	ء	Alef
B	ب	Ba′
T	ت	Ta′
TO	ط	Emph. Ta′
TH	ث	Tha′
J	ج	Jim
H	ح	Ha′
KH	خ	Emph. Kha′
D	د	Dal
DO	ض	Emph. Dad
DZ	ذ	Thal
DZO	ظ	Emph. Tha′
Z	ز	Zay
R	ر	Ra′
S	س	Sin
SH	ش	Shin
SO	ص	Emph. Sad
AIN	ع	Ayn
GAIN	غ	Ghayn
F	ف	Fa′
KO	ق	Emph. Kaf
K	ك	Kef
L	ل	Lam
M	م	Mim
N	ن	Noun
HI	ه	Ha′
W	و	Waw
Y	ي	Ye′
A	◌َ	Fatha
I	◌ِ	Kasra
E	◌َ مع مد أو ◌ِ	Between kasra and fatha
AO	◌َ (مفخمة)	Emph. fatha
OU	◌ُ	Damma

Table 1. List of selected phonemes

The preceding set was then used to train the states of the HMM that corresponds to the acoustic model. About 1 hour of audio recitations of Sourate Al-Ikhlass for different recitors including normal (with no tajweed) and women performed recitations were used alongside with the corresponding dictionary mapping each word to the corresponding symbolic representation, to feed the sphinxTrain application that generated the corresponding Acoustic model for the application. It is recommended however to have a minimum of 8 hours of recorded audio in order to get efficient recognition, but we estimated that for the scope of our research, the use of only 1 hour is sufficient especially that we will be testing the system on a limited vocabulary(only sourate "Al-Ikhlass"). The following table shows an excerpt from the dictionary used for the training and recognition phases.

Bismi	B I S M I
Lahi	L L A HI I
Rahmani	R R AO HI M A N I
Rahim	R R AO HI I I M
Rahim(2)	R R AO HI I I I I I I M
Rahim(3)	R R AO HI I I I I M
Koul	KO OU L
Houwa	HI OU W A

Table 1. Excerpt from the dictionnary of Sourate Al-Ikhlass

4. Language Model

As with the majority of speech recognition solutions, a language model is used to increase the accuracy of the recognition process. The most important choices is either to create a statistical model of the words in a given language (or linguistic context) or to create a grammar file. The first approach is most suitable for large vocabulary applications while the latter is very well adapted to small ones. In the case of the holy Quran, Creating a statistical language model is the best approach but, in our research, we chose to use a grammar file based on the Java Speech Grammar Format JSGF specification that is well supported by Sphinx-4 because of the relatively small vocabulary that we chose for our test. Furthermore, It is imperative to have a high accuracy ratio based on an "all or none" paradigm meaning that if an "aya" could not be 100% recognized, it is better to drop it rather than filling it with garbage words because of the holiness of the Quran. These JSGF rules are similar to those used for conversational systems and are, actually, not suitable for large vocabulary continuous speech recognition but we generated them as such to reflect the structure of the "Sourat".

```
grammar Quran;

public <Ikhlass> = (Bismi  Lahi  Rahmani  Rahim |
         (Koul houwa llahou Ahad | Koul houwa llahou Ahadounil |Allahou
Samad |Lam Yaled wa Lam youlad |            wa Lam yakoun lahou koufouan
Ahad)
```

Listing 1. Excerpt from the grammar file

5. System Design

The core recognition process is provided automatically by the sphinx engine using the appropriate language and acoustic models. The sphinx framework must be configured using an xml based configuration file.

5.1. Data preparation

We configured our system with the following pipeline before being processed by the recognizer:

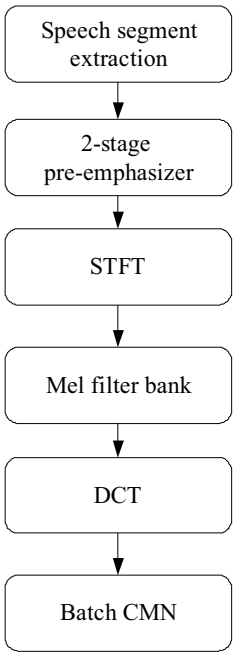


Figure 4. frontend pipeline of the system

The **Speech Segment Extraction** bloc labels speech frames as "speech" and "no speech" using a threshold that we set to -10db. The **pre-emphasis filter** consists of a digital network that flattens the signal:

$$\tilde{s}(n) = s(n) - as(n-1) \tag{1}$$

Where a is the pre-emphasis factor 0<a<1

Although it is often enough to use one pre-emphasis filter, we have found that for some audio files the recognition ratio could be increased with the use of a 2-stage pre-emphasis filter with different factor values (0.92 and 0.97).

The **Short Time Fourier Transform STFT** bloc uses a raised cosine windower to apply the Fourier transform on detected speech blocs. We specified the number of point of FFT points to 512 points.

The **Mel Filter Bank** bloc transforms the frequency domain to the Mel frequency domain that mimics the sensitivity and perception of the human ear by transforming the frequency domain from a linear to a non-linear one. This goal is achieved by using a set of 30 triangular Mel filters where each filter is given by:

$$H_m(k)=\begin{cases}0 & k<f(m)\\ \dfrac{k-f(m-1)}{f(m)-f(m-1)} & f(m-1)\le k\le f(m)\\ \dfrac{f(m+1)-k}{f(m+1)-f(m)} & f(m)\le k\le f(m+1)\\ 0 & k>f(m)\end{cases}$$ With m = 1,2,…30

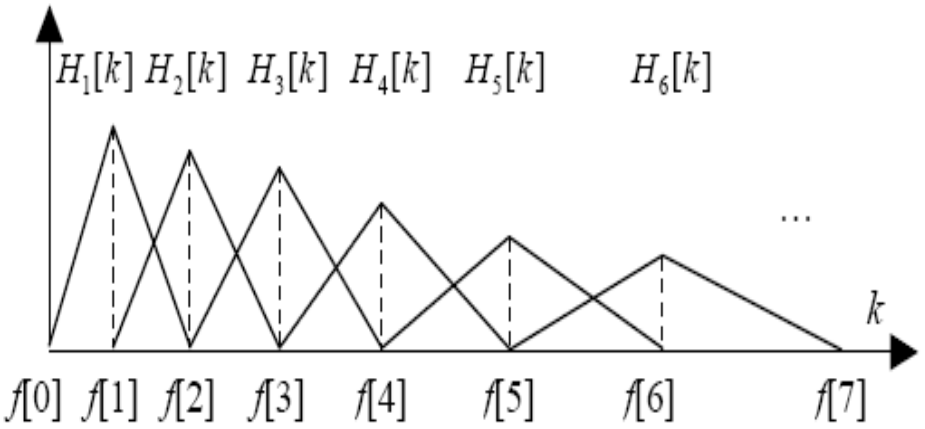


Figure 5. Representation of the Mel triangular filters

The **DCT** bloc applies the Discrete Cosine Transform on the result to extract the Mel Frequency Cepstral Coefficients MFCC (When all of the FFT coefficients are real values, DCT is often used to calculate the cepstras instead of the inverse FFT). The use of the MFCC has proven remarkable results in the field of speech recognition.

The final Cepstral Mean Normalization **CMN** operation is used to reduce the distorsion effect introduced by the transmission medium (microphone). It consists of subtracting the mean vector $\bar{\boldsymbol{x}}$ from each vector $\boldsymbol{x_t}$ to obtain the normalized cepstrum vector. This is justified by recalling that the cepstral transformation transforms the convolution to addition due to the use of the logarithm, thus the mean of the cepstral holds the characteristic of the transmission medium (X. Huang, et al.,2001).

5.2. Application Design

The output of the front-end was then used to feed the sphinx core recognizer, which uses the Hidden Markov Models HMM as the recognition tool.

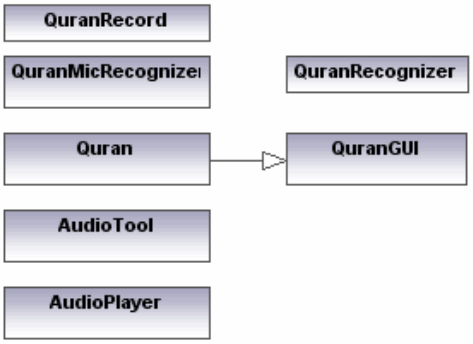


Figure 6. Class Diagram of the Quranic Recognizer application

The application uses a multithreaded architecture for better performance. **QuranRecognizer** takes audio files as an input and launches the recognition process. The result of the recognition is recorded in a list of type **QuranRecord** where each entry specifies the recognized word and its starting and ending time in the audio file. **AudioTool** and **AudioPlayer** are used for the playback of the audio files after the recognition process.
The internal application uses the **Observer** design pattern in order to notify the main application of each result of the detection process.

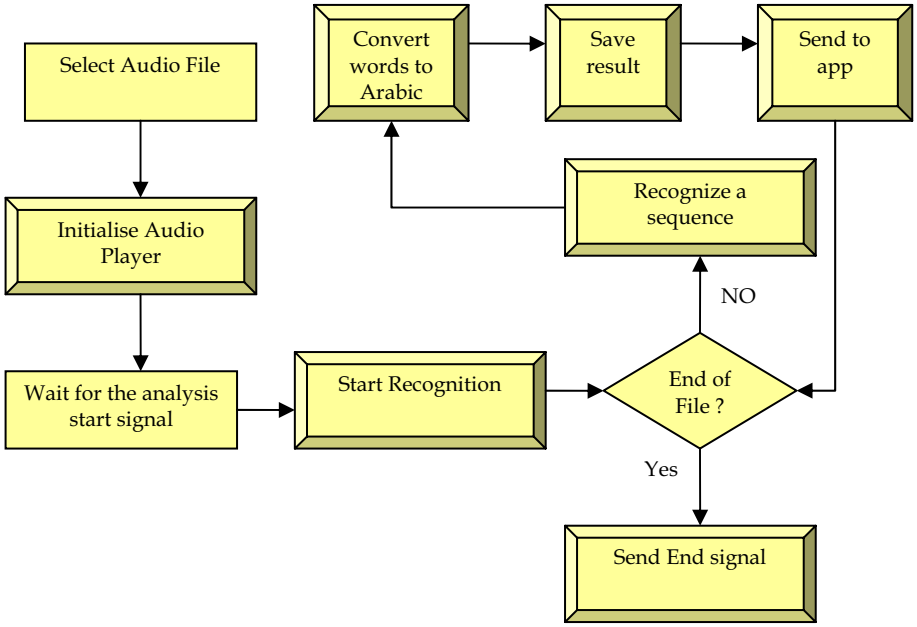


Figure 7. Program Flowchart

The version of SphinxTrain we used didn't support Unicode characters for text and thus we were forced to use transliteration for all the words in our dictionary. At the application level, we used a hash map to translate the results of the recognizer into common Arabic words. The search algorithm used in the decoder was the simple breadth first combined with the beam search. In the breadth first search algorithm, all the nodes on one level are examined before considering any node of the next level, it could thus take a longer time to reach the best solution if

6. Experiments and Results

We performed a large number of experiments on different individuals: each one was asked to recite sourat "Al-Ikhlass" several times and each time we recorded the number of ayates that were recognized correctly then a mean recognition ratio for each tester was calculated. The global mean is what we are showing in the following tables and the mean per individual represents the values shown in the graphs. The testers were chosen from different backgrounds without excluding women and children from them.

Type of Recitation	Number of Recitors	Mean Recognition Ratio
Tajweed	20	90%
Tarteel	20	92%

Table 2. Test results for professional recitors for sourate Al-Ikhlass

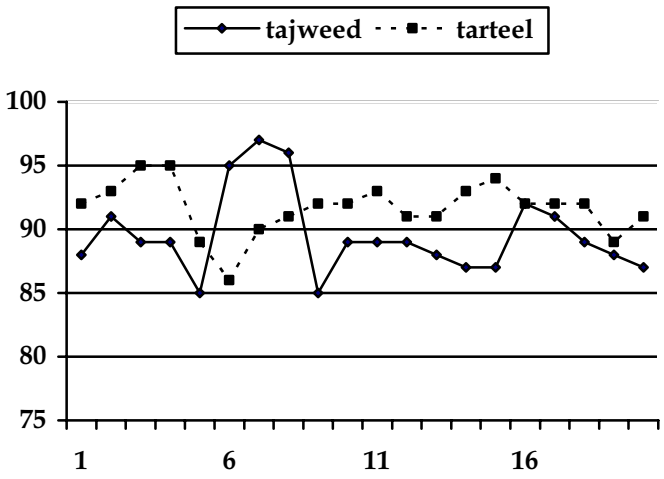


Figure 8. Chart illustrating the accuracy for both tarteel(dotted) and tajweed(plain) according to each recitor

The recognition ratio in the case of tarteel is slightly better than in the case of tajweed. One possible reason for this could be that the majority of the tarteel recitations available now follow the same monotony and the duration (in time) of each phoneme differ slightly from

one recitor to another. There is also the extra noise that is caused by the compression of the audio files and the law quality of the recordings. Although we have anticipated this by using noisy audio files during the training, but the differences in compression ratios between the files add a lot of variety for the added noise and thus causing extra errors.

Gender of recitor	Number of Recitors	Mean Recognition Ratio
male	20	90%
female	20	85%

Table 3. Test results for normal arabic speaking people for Sourate Al-Ikhlass

When unskilled persons tested the system (we even tested it on children), it behaved astonishingly well even when the recitor was a woman, a case that cannot be encountered in real life because it not common to have a woman reciting the Holy Quran. There is also an interesting observation drawn from these tests: It is always recommended [6] to train the system with more than 500 different voices in order to reach speaker independence. But we didn't train our system with this relatively large number and still we were able to have remarkable speaker independence results.

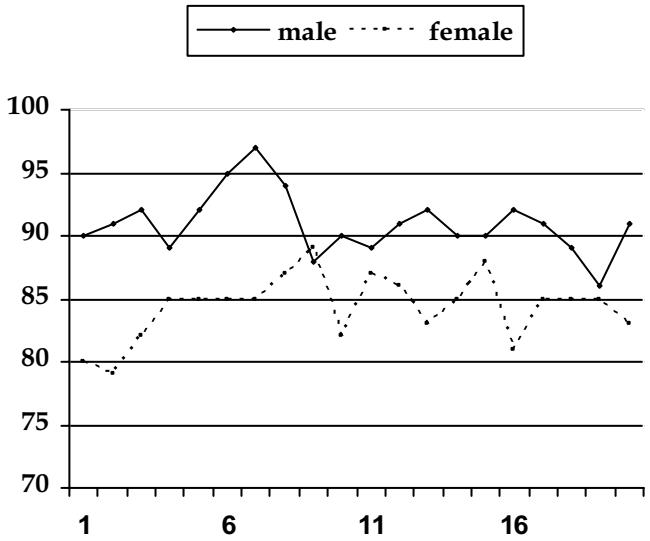


Figure 7. Chart illustrating the accuracy for male and female recitors (each tick on the x axis represents one recitor)

7. Conclusion

The system that we developed showed promising results although it was only tested against small Quran' chapters. We think that the incorporation of morphological knowledge of the Arabic language with a more sophisticated statistical model deduced from the full scope of

the Holy Quran can lead to a robust universal recognizer for the Arabic language. The results that we obtained were far beyond our expectations: It behaved as well accurately with children and female voices as well as with male voices. The HMM is know to be poor in its representation of the time that the phoneme takes, however, time in our application was a very important factor and to resolve this issue, we were forced to represent all the variations of each pronunciation, this could prove to be very painful if it is to be applied to the whole Quran. Another approach based on a smarter representation of the duration of each phoneme may represent a better solution to this problem. But, overall, the system proved that it is possible to construct an Automatic delimiter of the verses of the Holy Quran; May be a search inside the audio files will emerge one day as an alternative and more versatile way to search the Holy Quran.

8. References

The Holy Quran.

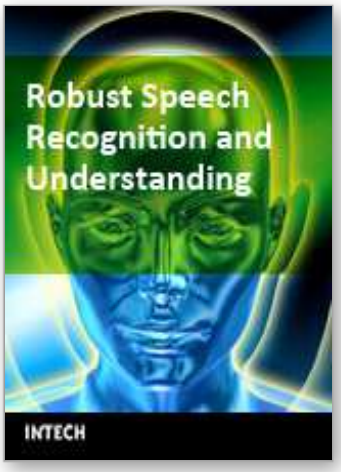
M. Habash, "How to memorize the Quran" , Dar al-Khayr, Beirut 1986.

Sphinx group, "Sphinx-4: A flexible Open Source Framework for Speech Recognition", Sun Microsystems, 2004.

O. Kimball, "Recognition of Conversational and Broadcast Arabic Speech", BBN technologies.

R. Descout, "Applied Arabic linguistics and signal and information processing" , Hemisphere, 1987.

X. Huang, A. Acero, H. Hon, "Spoken language processing a guide to theory, algorithm and system design",Prentice Hall 2001.



Robust Speech Recognition and Understanding

Edited by Michael Grimm and Kristian Kroschel

ISBN 978-3-902613-08-0
Hard cover, 460 pages
Publisher I-Tech Education and Publishing
Published online 01, June, 2007
Published in print edition June, 2007

This book on Robust Speech Recognition and Understanding brings together many different aspects of the current research on automatic speech recognition and language understanding. The first four chapters address the task of voice activity detection which is considered an important issue for all speech recognition systems. The next chapters give several extensions to state-of-the-art HMM methods. Furthermore, a number of chapters particularly address the task of robust ASR under noisy conditions. Two chapters on the automatic recognition of a speaker's emotional state highlight the importance of natural speech understanding and interpretation in voice-driven systems. The last chapters of the book address the application of conversational systems on robots, as well as the autonomous acquisition of vocalization skills.

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821