Information Systems Department
Faculty of Computer and Information Sciences
Ain Shams University

Content Based Search Engine for Arabic Audio Files

Thesis submitted as a partial fulfillment of the requirements for the degree of Master of Science in Computer and Information Sciences

By

Mona A.Azim A.Gawad

Teaching Assistant at Information Systems Department,
Faculty of Computer and Information Sciences,
Ain Shams University

Under Supervision of

Prof. Mohamed Fahmy Tolba
Scientific Computing Department,
Faculty of Computer and Information Sciences,
Ain Shams University

Prof. Tarek Fouad Gharib
Information Systems Department
and Head of Information Systems Department,
Faculty of Computer and Information Sciences,
Ain Shams University

Prof. Nagwa Lotfy Badr
Information Systems Department
and Vice Dean of education and student affairs,
Faculty of Computer and Information Sciences,
Ain Shams University

April – 2017
Cairo

Acknowledgment

First and foremost, my deep gratitude to Allah, the most beneficial, the most graceful, for helping me completing this work to the fullest.

Second, I would like to express my veracious and deep respect and gratitude to all my supervisors*; Prof. Mohamed Fahmy Tolba*, *Prof. Tarek Fouad Gharib* and *Prof. Nagwa Lotfy Badr* at the Faculty of Computer and Information Sciences, Ain Shams University, who dedicated much of their precious time to assist me in elaborating this thesis, guide me in understanding and analyzing problems, and continuously provide support valuable comments that influenced my career and will be my success recipe throughout my entire life.

Finally, I would like to acknowledge Dr. Abd El Aziz Abd El Hamid at the Faculty of Computer and Information Sciences, Ain Shams University for his continuous support and efforts throughout this work until it reached the desired form. In addition, I would like to acknowledge my dearly loved parents who were burdened with even more roles and responsibilities in order to give me the chance to accomplish this work and now I hope that they are proud to see this work has been finished successfully.

Abstract

Information systems were subject to considerable changes. These changes were arisen due to the fact that information representation underwent massive evolution. This evolution from heavyweight information representation (i.e. papers, books, music sheets) towards lightweight information (i.e. multimedia and digital forms) is driven by spectacular changes to the constrains of storing, managing and retrieving such types of information.

Multimedia Information Retrieval (MIR) is about obtaining semantic information from different multimedia data sources (i.e. audio, video or images). The research in this field faces many challenges due to the variety of data types, methodologies and research problems. One of the challenging sub-field of MIR is audio retrieval.

Audio retrieval aims to extract certain piece of information from audio signals contents based on a submitted query. Audio data types may include pure music, environmental sounds, songs and speech utterances.

In this work, we have implemented a content based retrieval system for Arabic speech signals to retrieve the speech audios relevant to the user text query. A Large Vocabulary Arabic Continuous Speech Recognition (LVACSR) System was built as a preceding step towards defining speech transcriptions used in the retrieval system. Hidden Markov Models (HMMs) based LVACSR system has adopted a novel approach in state tying that uses a new developed phonetic decision tree for the Arabic language for building acoustic models of the LVACSR system. Building such models was accomplished by firstly building phoneme-based HMMs then building triphones based HMMs. The development of the triphones models process includes using both Data driven based and phonetic tree based state tying techniques. After testing the built models, the tree-based

tied states models have shown maximum word correctness and accuracy. The text transcriptions generated from the recognition system that used the tree-based tied states models as acoustic models were used as an input to the text- based retrieval system. In the text-based retrieval system, the text transcriptions were stemmed then indexed in the indexer database then the most relevant audios to the user query were retrieved. The text-based retrieval system used three different character based similarity distance to obtain the desired audios list. Levenshtein distance, Jaro-Wrinkel distance and N-gram distance were calculated to measure the similarity between the user query and the indexed text transcriptions of the audios. After obtaining the most relevant audios list, the precision and recall were calculated to evaluate the text-based retrieval system. The results showed that the Levenshtein distance and Jaro-Wrinkel distance were more accurate than the N-gram distance when used as a similarity measure.

List of Publications

Azim, M. A., Hamid, A. A. A., Badr, N. L., & Tolba, M. F. (2016, December). Large Vocabulary Arabic Continuous Speech Recognition Using Tied States Acoustic Models. *Asian Journal of Information Technology.(Accepted on December 10,2016)*

Azim, M. A., Hamid, A. A. A., Badr, N. L., & Tolba, M. F. (2016, October). Tree-Based HMM State Tying for Arabic Continuous Speech Recognition. In *International Conference on Advanced Intelligent Systems and Informatics* (pp. 96-103). Springer International Publishing.

Azim, M. A., Badr, N. L., & Tolba, M. F. (2016, May). An Enhanced Arabic Phonemes Classification Approach. In *Proceedings of the 10th International Conference on Informatics and Systems* (pp. 210-214). ACM.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

CBASR **C**ontent **b**ased **A**rabic **S**peech **R**etrieval

ASR **A**utomatic **S**peech **R**ecognition.

CL **Cl**one Command

HERest **H**TK **E**mbedded **re-est**imation

HHEd **H**MM **H**Tk **Ed**itor

HLEd **H**MM **L**abel **Ed**itor

HMMs **H**idden **M**arkov **M**odels

HTK **H**MMs **T**ool **K**it

IVR **I**nteractive **V**oice **R**esponse

LVACSR System **L**arge **V**ocabulary **A**rabic **C**ontinuous **S**peech **R**ecognition System

QS **Q**uestion Command

SASSC **S**tandard **A**rabic **S**ingle **S**peaker **C**orpus

TC **T**riphones **C**onversion

TI **Ti**e Command

WER **W**ord **E**rror **R**ate

LM **L**anguage **M**odel

MFCC **M**el **F**requency **C**epstral **C**oefficients

Chapter 1

Introduction

Chapter 1. Introduction

1.1 Overview

Nowadays, digital archives are increasing rapidly and the need for effective retrieval systems is increasing correspondingly. In case of audio, traditional audio retrieval systems rely mainly on text-based approaches. These approaches depend on gathering and indexing audio files meta data then applying a text based retrieval technique on the indexed data. However, this approach is simple and straightforward and reuse the already refined text based search techniques, the misplaced audio meta data may result in poor retrieval quality.

Recently, content-based approaches were developed to improve the audio retrieval systems performance. These approaches aim to fulfill the user audio information needs by searching the entire audio library and retrieve the most relevant ones to the user query.

1.2 Motivation

The challenge of audio information retrieval systems is familiar to anyone who has been in a vacation and then after returning home he found his answering machine was full of messages and he has to be patient enough to listen to the entire tape till he found the urgent message from his boss.

Content-based audio retrieval has not been explored in the context of content based multimedia retrieval systems. However, if fully studied it can also be more useful in many applications such as audio data set storing and managing systems and digital libraries manipulation.

Searching and archiving digital libraries is an extensive and exhausting task to be done manually. Thus automating such task is

mandatory for certain beneficial organizations such as Radio stations and recordings incorporations.

1.3 Objectives

This work aims to build a content based Arabic speech retrieval system. The text transcriptions of the Arabic speech audio files were generated using a Large Vocabulary Arabic Continuous Speech Recognition (LVACSR) system. Phonemic based Hidden Markov Models (HMMs) where trained and tested to serve as the LVACSR system acoustic models. In order to improve the system word correctness, triphones based HMMs where built as well. Firstly, the triphones models applied the data driven based state tying. After that, a novel Arabic phonetic decision tree was used to apply the decision tree based state tying on the triphone models. Finally, for each of the trained models the retrieval results were evaluated.

1.4 Thesis Organization

The thesis is organized as follows. Chapter 2 reviews the recent advances in the audio retrieval systems and gives a brief overview of the latest trends of Arabic speech recognition systems. Chapter 3 introduces the proposed content based Arabic audio retrieval system and the typical architecture of the LVACSR with a detailed explanation of each system module and the text based retrieval system. Experimental Results and system evaluation are discussed in chapter 4. Conclusions and future work are presented in chapter 5.

Chapter 2

Literature Review

Chapter 2. Literature Review

2.1 Content based audio retrieval systems

Most of the information retrieval systems are used to focus on the textual type of information and text based information retrieval techniques have been developed and achieved considerable results.

Due to the increase of multimedia databases, the need of developing efficient retrieval techniques has arisen. Recently, the multimedia information retrieval has gained much interest by various researchers. The multimedia information retrieval is the research field that aims at extracting semantic information from multimedia data sources; such as images, audios and videos. Most of researchers in multimedia information retrieval systems focus on searching the visual media (i.e. image and videos).

However, audio is an important medium and an information carrier from the human auditory view point, so it is needed to search and retrieve the audio collections. In fact, audio can save large amount of rich information in the form of speech, musical and sound effects within a sort of small space. So it should be searched based on its aural content; such as speech uttered within audio, musical melodies and acoustic features.

The development of content based audio retrieval systems depends on the type of the audio either it is a piece of music, an uttered speech or a sound effect or the query type either it is query by example or text query.

The research in this field was initially proposed in [1] by implementing a general audio classification and retrieval system. In that system, audio is represented by perceptual and acoustical features, in which users can search or retrieve sounds by several types of query. In [2] a newly developed