SPEAKER IDENTIFICATION BASED ON HYBRID FEATURE EXTRACTION TECHNIQUES

BY

FERAS EID DHEIF ALLAH ABUALADAS

A thesis submitted in fulfillment of the requirement for the degree of Doctor of Philosophy in Computer Science

Kulliyyah of Information and Communication Technology
International Islamic University Malaysia

FEBRUARY 2020

ABSTRACT

Speech contains many features that can be used to determine gender and speaker identity; it is a natural form of communication between humans. One of the most exciting areas of signal processing is speech processing. Speech contains many features or characteristics that can discriminate the identity of a person. The human voice is considered one of the important biometric characteristic that can be used for person identification. The proposed speaker identification system (SIS) consists of four phases, namely, pre-processing phase (involves sample resizing to 40000 samples and normalization to ensure that the sound volume will modifying as a standard level), feature extraction phase (involves extracting a set of fundamental voice features that can represent or identify the entire signal of speech), feature selection phase (involves selecting the best features that describe the speaker, where dealing with hundreds number of features leads to increase the workload of recognition) and recognition phase (involves Backpropagation (BP) neural network in this research). In this work the effects of appropriate extracted voice features from various levels of discrete wavelet transformation (DWT) and the concatenation of DWT and curvelet transformation (DWT+Curvelet hereinafter) are studied. The effects of reducing the number of features via Principal component analysis (PCA) on speaker identification is also investigated, and the (BP) neural network was introduced as a classifier. The classifier is trained with a different set of features extracted from three different levels of DWT; these features are extracted one level at a time. The recognition capabilities of the classifier for all levels are compared to determine the best level. This research explores any positive or negative effects of DWT+Curvelet on the classification capability of the proposed system. in addition, this work investigates the effects of reducing the number of features via PCA with DWT and DWT+Curvelet In this research, different three datasets were used for speaker identification system, where these dataset used for train and testing the Feed-Forward Backpropagation (BP). In this approach it is clear that introducing PCA with BP networks improved the accuracy and is an effective method for speaker identification system, where it keeps the effective information and reduces the redundancy of characteristic parameters Four experiments are performed as follows using the three datasets: Experiment 1: only DWT features that extracted from each level of discrete wavelet transformation independently are used to train and test the Neural Network; Experiment 2: the features extracted from each level of (DWT+Curvelet) used to train and test the Neural Network; Experiment 3: With DWT features after utilized principal component analysis used to train and test the neural network; Experiment 4: With (DWT+Curvelet) features after utilized principal component analysis used to train and test the Neural Network. Practical results showed that, the accuracy is improved in level 1 and 2 with database 1 and increased by approximately 5% and 4%, respectively; whereas the accuracy was improved in all levels 1, 2 and 3 with Database 2 and 3 and increased by approximately 11%, 4% and 2% for database 2 and 9%, 11%, 5% for database 3 respectively, when applying (DWT+Curvelet). The system was trained and tested using (Cross-validation).

خلاصة البحث

يحتوي الكلام على العديد من الميزات التي يمكن استخدامها لتحديد هوية كلٍّ من الجنس وهوية المتحدث (وجنس المتكلم) وهي شكل طبيعي من أشكال التواصل بين البشر. ويعتبر مجال معالجة الكلام من اكثر مجالات معالجة الاشارة اعجابا. يحتوي الكلام على العديد من الميزات أو الخصائص التي يمكن أن تميز هوية الشخص. يعتبر صوت الإنسان أحد الخصائص الحيوية الهامة التي يمكن استخدامها لتحديد هوية الشخص. يتكون نظام تحديد هوية المتكلم المقترح من أربع مراحل (مرحلة ما قبل المعالجة وتتضمن تقليل عدد العينات المكونة للصوت حيث تم تقليل عدد العينات الى 40000 وتم تطبيق التطبيع (الاعتماد الوظيفي للبيانات) لضمان تعديل الصوت بمستوى معياري) , مرحلة استخراج الميزات (لاستخراج مجموعة من الميزات الاساسية التي تمثل اشارة الكلام بشكل كامل), مرحلة تحديد الميزات (حيث يتم اختيار الميزات الافضل التي تصف المتحدث حيث ان التعامل مع مئات الميزات يؤدي الى زيادة العبأ في عملية التمييز) ومرحلة التمييز حيث تم استخدام الشبكات العصبية ذات التغذية المرتدة. يهتم هذا البحث بدراسة تأثير الميزات المستخلصة من مستويات مختلفة من المتحول الموجي المنفصل بالاضافة الى الميزات المستخلصة من دمج المحول الموجي المنفصل مع التحويل القوسي ودراسة تأثير تقليل عدد الميزات المستخلصة بإستخدام تحليل المكون الاساسي على تمييز المتحدث حيث تم اعتماد الشبكات العصبية كمصنف.حيث تم تدريب الشبكات العصبية باستخدام مجموعة مختلفة من الميزات المختلفة المستخلصة من ثلاثة مستويات مختلفة من المتحول الموجي المنفصل (مستوى واحد في الوقت الواحد) ومقارنة قدرتها على التمييز وتحديد اي مستوى افضل من المستويات الثلاثة. ويوضح هذا البحث فعالية دمج المحول الموجي المنفصل مع التحويل القوبي لمعرفة ما اذا كان هناك اي تأثير على قدرة المصنف حيث بالامكان ان يكون ايجابي او سلبي بالاضافة الى دراسة تأثير تقليل عدد الميزات مع المتحول الموجي المنفصل و مع عملية الدمج بين المتحول الموجي المنفصل و التحويل القوسي. في هذا البحث تم استخدام ثلاثة مجموعات مختلفة من البيانات لنظام تحديد هوية المتحدث حيث استخدمت هذه المجموعات لتدريب الشبكات العصبية ذات التغذية المرتدة. تم اجراء اربعة تجارب مختلفة مع كل واحد من مجموعة البيانات, التجربة 1: استخدام الميزات المستخلصة من المحول الموجي المنفصل بشكل مستقل لتدريب الشبكة العصبية, التجربة 2 : استخراج الميزات من المستويات المختلفة للمحول الموجي المنفصل ودمجها مع التحويل القوسى,التجربة 3:تدريب الشبكة العصبية بالميزات المستخلصة من المحول الموجي المنفصل بعد استخدام تحليل المكون الرئيسي , التجربة 4: تدريب الشبكة العصبية بالميزات المستخلصة من المحول الموجي المنفصل و التحويل القوسي بعد استخدام تحليل المكون الرئيسي. في هذا النهج اتضح ان استخدام تحليل المكون الرئيسي مع الشبكات العصبية قد ادى الى تحسين الدقة ويعتبر هذا النهج هو وسيلة فعالة لنظام تحديد المتحدث حيث يحتفظ بالمعلومات الفعالة ويقلل من تكرار الميزات .وأظهرت النتائج العملية أنه تم تحسين الدقة في المستويين 1 و2 مع قاعدة البيانات 1 وزيادة بنسبة 5 ٪ و 4 ٪ على التوالي؛ في حين تم تحسين الدقة في جميع المستويات 1 ، 2 و 3 باستخدام قاعدة البيانات 2 و 3 وزادت بنسبة حوالي 11 ٪ , 4 ٪ و 2 ٪ لقاعدة البيانات 2 و 9 ٪ , 11 ٪ و 5 ٪ لقاعدة البيانات 3 على التوالي ، عند تطبيق المحول الموجي المنفصل بدقة مع التحويل القوسي. كما تم اســخدام نظرية التحقق من الصحة (Cross Validation) لاغراض التدريب.

.

APPROVAL PAGE

The thesis of Feras Eid Dheif Allah Abualadas has been approved by the following:

Akram M. Zeki
Supervisor

Muzhir Shaban Al-Ani
Co-Supervisor

Az-Eddine Messikh
Co-Supervisor

Asadullah Shah
Internal Examiner

Hamid Abdullah Jalab
External Examiner

Adel Al-Jumaily
External Examiner

Saim Kayadibi
Chairman

DECLARATION

I hereby declare that this thesis is the result of my own investigations, except where otherwise stated. I also declare that it has not been previously or concurrently submitted as a whole for any other degrees at IIUM or other institutions.

Feras Eid Dheif Allah Abualadas

Signature .... Date ....

INTERNATIONAL ISLAMIC UNIVERSITY MALAYSIA

DECLARATION OF COPYRIGHT AND AFFIRMATION OF FAIR USE OF UNPUBLISHED RESEARCH

SPEAKER IDENTIFICATION BASED ON HYBRID FEATURE EXTRACTION TECHNIQUES

I declare that the copyright holders of this thesis are jointly owned by the student and IIUM.

1. Any material contained in or derived from this unpublished research may be used by others in their writing with due acknowledgement.

2. IIUM or its library will have the right to make and transmit copies (print or electronic) for institutional and academic purposes.

3. The IIUM library will have the right to make, store in a retrieved system and supply copies of this unpublished research if requested by other universities and research libraries.

By signing this form, I acknowledged that I have read and understand the IIUM Intellectual Property Right and Commercialization policy.

Affirmed by Feras Eid Dheif Allah Abualadas

……..…………………….. Signature

……………………….. Date

DEDICATION

To The Soul Of My Mother, Which I Wish To Be Beside Me Now, To Her Ambitious And Her Wishes …

To My Father For His Endless Love And Support, The Origin Of My Success …

To My Brother, My Sisters For Their Supporting Words...

To My Daughter (shams), Which Will Come To The World Soon

To My Wife….

To My Supervisors And Teachers With Respect...

To My Friends Who Always Behind Me

TO My Ambitiousness And My Dreams

Which ,I Wish To Be Real…

Feras E. Abualadas

ACKNOWLEDGEMENTS

Firstly, it is my utmost pleasure to dedicate this work to Allah, to my dear father, my brother and sisters, my wife, daughter, the soul of my mother, who are granted me the gift of their unwavering belief in my ability to accomplish this goal: thank you for your support and patience.

I wish to express my appreciation and thanks to those who provided their time, effort and support for this thesis.

Finally, a very special thanks to Professor Dr. Akram Zeki, Professor Dr. Muzhir Shaban Al-Ani and Professor Dr. Az-Eddine Messikh for their continuous support, encouragement and leadership, and for that, I will be forever grateful.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ADC	Analog To Digital Converter
ANN	Artificial Neural Network
ASR	Automatic Speaker Recognition
BP	Backpropagation
BPNN	Back-Propagation Neural Network
BLSTM	Bidirectional Long Short-Term Memory
CWT	Continuous Wavelet Transform
DSP	Digital Signal Process
DCT	Discrete Cosine Transform
DCT	Discrete Cosine Transform
FDCT	Discrete Curvelet Transform
DWT	Discrete Wavelet Transformation
FFT	Fast Fourier Transform
FFNN	Feed Forward Neural Network
LFCC	Frequency Cepstral Coefficients
GMM	Gaussian Mixture Model
HMM	Hidden Markov Model
ILPR	Integrated Linear Prediction Residual
Knn	K-Nearest Neighbours
KNN	K-Nearest-Neighbors
LPCC	Linear Predictive Cepstral Coefficient
LSH	Locality Sensitive Hashing
MFCC	Mel-Frequency Cepstrum Coefficients
MLP	Multilayer Perceptrons

NN	Neural Networks
PLPCC	Perceptual Linear Prediction Cepstral Coefficient
PNCC	Power Normalized Cepstral Coefficients
PCA	Principal Component Analysis
PNN	Probabilistic Neural Network
RF	Random Forest
RASTA PLP	Relative Spectral Perceptual Linear Prediction
SOM	Self Organizing Map
STFT	Short Time Fourier Transform
SVS	Speaker Verification System
SWT	Stationary Wavelet Transform
SVM	Support Vector Machine
SIS	Speaker Identification System
USFFT	Unequally Spaced Fast Fourier Transform
VQLBG	Vector Quantisation-Linde, Buzo, And Gray
VQ	Vector Quantization
WLPC	Wavelet Based Linear Predictive Coding

CHAPTER ONE

INTRODUCTION

1.1 BACKGROUND OF THE STUDY

Speech processing is one of the most exciting areas of signal processing given that speech is the most natural form of human communication. The human voice is considered as one of the main biometrics that could be used to identify a person since person voice contains information that could help in specifying his identity (Al-Hassani & Kadhim, 2012a).

A pattern is a structural characterization of objects or other some entity of interest such as data system, therefore; when talking about data we not just refer to the visible objects. Pattern recognition is a task of categorize measured sample or observed data as a member of one of various category (Bow, 2002).

Authentication of personal identity through biometrics is one of the most important applications that fall under pattern recognition. Biological characteristics used by biometric technology should be unique for each person. Other individuals cannot easily replace, steal or forge these features. Human voice is one of the biometrics where an individual can be distinguished by his/her voice, thus we refer to speaker recognition systems as those technologies which utilize human speech to recognize each individual from others (Togneri & Pullella, 2011b). The method of biometric identification is preferred over traditional identification methods involving passwords for various reasons. For example, biometric identification requires the person being identified to be physically present at the point of identification and is thus more secure relative to other methods. Moreover, the identification based on biometric techniques does not require passwords, tokens or smartcards (von Graevenitz, 2000).

Automatic speaker recognition (ASR) mention as voice biometric or voice recognition in some studies, when the identity of speakers analysis via machine without any help from human, it is named automatic speaker recognition, where the machine should extract particular information (features) from speaker voice (Kinnunen & Li, 2010). Speaker recognition system, identification and verification, should never be confused with speech recognition technologies. The purpose of Speaker recognition, where it is various from speech recognition is to identify or verify the speakers, on other hand speech recognition is the process that concerned in recognize the word or understand what is being said. Many applications used in automatic speech recognition such as correction/detection of specific pronunciation errors (Kalaivani & Thakur, 2017; Yousfi & Zeki, 2016).

Automatic speaker recognition can be divided in two different types, namely speaker identification and speaker verification. ASR is used in extensive application areas such as authentication and surveillance (N. Singh, Khan, & Shree, 2012). Some examples of the applications of speaker recognition systems are as follows:

1. **Authentication**: Different characteristics are used to identify person, such as voice, fingerprint, facial feature and signature. This method of authentication is classified as biometric authentication and involves fewer problems than the traditional method that uses passwords, PINs and credit cards that can be stolen or forgotten.

2. Surveillance: the technology of speaker recognition has an ability to use in detect, tract the persons by security agencies over radio/telephone conversation.

3. Forensics: It is significant application of speaker recognition, where this application can verify the suspect's during the crime if there is a sample of speech for comparison between the voice of suspect's and the sample of speech.

4. Security: Speaker recognition can be used in transactions such as credit card purchases, the authentication of which is integrated with methods such as facial recognition. The technology of speaker recognition can help authenticate access to computers, banks transactions or telephone conversations.

Human speech is produced by different acoustical excitations of the vocal tract. It comes in two forms: voiced and unvoiced; voiced speech is the result of the vibration of the vocal fold, whereas unvoiced speech is the result of turbulence in the vocal tract (Campbell & Tremain, 1986).

Three fundamental classes of voiced (female, male, child), according to the frequency of vocal cords vibration. These measures called pitch or fundamental frequency, where it is different for the three classes. the frequency of man speech ranges between 80 HZ to 150 HZ, that of female speech ranges between 150 HZ to 300 HZ and that of child ranges between 200 HZ to 500 HZ (Djeraba & Saadane, 2001).

Voice is produced by humans by using the vocal folds in reading, singing, talking, screaming, laughing and crying; the vocal folds are considered a basic sound source. The methods for generating voice involve the vocal folds within the larynx, the articulators and the lungs (Gbadamosi, 2013).

The vocal folds or vocal cords of males and females are different in size. Those of adult males measure between 1.75 and 2.5 cm, whereas those of females measure between 1.25 and 1.75 cm. As the vocal folds of males are thicker than those of females, adult males have lower pitched voices than adult females. Children have much shorter vocal folds than males and females (Dhoke, March 2016).

1.2 STATEMENT OF THE PROBLEM

Due to the rapid development in various fields of information technology, data have become more vulnerable to theft and vandalism (Alarifi, Alkurtass, & Al-Salman, 2011). The level of security breaches and transaction fraud has increased recently, and the need for highly secure personal identification technologies has become apparent (Al-Hassani & Kadhim, 2012a). With the need to verify the identity and authority of users and customers, individuals face issues in remembering several PIN codes and carrying various cards and keys (Debnath, Soni, Baruah, & Sah, 2015). One of the systems that used to verify the user identity is the speaker recognition system, where each one of people have distinct vocal features, thus, it is beneficial to recognize between people based on their voice. The human voice carries two types of information, namely, low-level information and high-level information. In speaker voice recognition, the computer deals with low-level information, such as frequencies, tone, pitch period, spectral magnitude, bandwidths of the voice and rhythm. By contrast, humans recognize voice using high-level information, such as accent, dialect and style (El-Zaghmouri, 2015).

One of the most significant aspects of speaker identification is the feature extraction technique, which affects speaker identification performance. The selection of an appropriate approach for feature extraction is vital, and identification is carried out by comparing the characteristics that are unique for a given input (Tirumala, Shahamiri, Garhwal, & Wang, 2017). Another set of information attributes can be taken as features and they including the Mel-frequency Cepstrum Coefficients (MFCC) and Linear Predictive Cepstral Coefficient (LPCC) for voice recognition. For a robust voice recognition system, continuous wavelet transform coefficients (CWT) and discrete wavelet transform are used. Researchers used different techniques to extract features (single or combining) lead to extract effective features (Oscar Rangel, 2017). The output of feature extraction has many features of which none is important for speaker discrimination, and the number of features should be also relatively low were dealing with hundreds of features that leads to increasing the workload of recognition (Srinivasan, Ramalingam, & Sellam, 2012).

The current research is focused on combined two techniques for features extraction and selecting an effective features in order to generate effective minimum features for speaker identification approach where discrete wavelet transformation (DWT) and curvelet transform were used singly or combined with principal component analysis (PCA) to train the classifier. Two main factors, namely, number of levels and the minimum set of coefficients extracted from the level(s), should be selected to leads to better recognition. This topic still requires further research because maintaining the best discrimination capability with a minimum set of features is important to speed up the identification operation during the search for a large voice dataset that does not affect system accuracy.

1.3 RESEARCH OBJECTIVES

The main objective of this research is to build a speaker identification system (SIS) that automatically authenticates individuals on the basis of their voice by extracting and selecting a minimum number of features from one or more levels (hereinafter referred to as features). DWT and curvelet, along with PCA, are applied to speaker identification, and the best classification accuracy is achieved in the following ways.

1- To study the features extracted from different levels of discrete wavelet transformation.

2- To illustrate the effectiveness of using discrete wavelet transformation concatenated with curvelet transform (hereinafter referred to as DWT+Curvelet) for different levels.

3- To select the best features from each level by using a Principal Component Analysis (PCA) and suggesting minimum number of features that would not affect the system accuracy.

1.4 RESEARCH QUESTIONS

The problem of this study can be expressed by the following main question: How to develop SIS by using minimum set of features with high accuracy. This main question has the following sub-questions that must be answered in order to achieve the objectives of study:

1. What are the effects of each level of DWT on the speaker identification system?
2. What is the difference in the accuracy of the system by concatenation discrete wavelet and curvelet?
3. What are the effects of selected set of feature of DWT and curvelet by using Principal Component Analysis?

1.5 SIGNIFICANCE OF THE STUDY

This work concerned with studying the effect of appropriate extracted features from a different levels of DWT and the concatenation of two techniques (DWT+Curvelet). It also studies the effect of reducing the number of features by using principal component analysis (PCA) on speaker identification. Backpropagation (BP) neural network is also introduced as a classifier.