

TAPI IN TELEPHONE QUALITY SPEECH DATABASE

MOHD SHUKRI BONGSU

A project report submitted in partial fulfillment
of the requirements for the award of the degree of
Master of Engineering
(Electrical – Electronics & Telecommunication)

Faculty of Electrical Engineering
Universiti Teknologi Malaysia

MAY 2007

To beloved wife and sons

ACKNOWLEDGEMENT

First of all, I would like to take this opportunity to thank my supervisor, Associate Professor Ir. Dr. Sheikh Hussein bin Shaikh Salleh. It is him who had made this project possible. He has shown me guidance, important advices, and inspiration throughout my project. Thank you, Sir, I will never forget your kindness and most of all, your dedication which I really felt.

Furthermore, to my beloved family, friends and fellow course mates must not be left out. I want to thank them for sharing and discussing knowledge with me and always give suggestions and opinion on my project, which I find very precious. I really appreciate the helping hands and encouragements given when I really need it.

Finally, I want to record my special thanks to Mr Amar for his full assistance and guidance along the way to successfully complete the project.

ABSTRACT

It is always been the speech recognition team's vision to be able to apply the developed speech technology in real world so that more people can benefit from it. One of the targets is towards telephony. People will be able to talk comfortably to computer through telephone to obtain certain information. Studies and effort have been carried out to improve the accuracy and efficiency of telephone speech recognition. This project aims to build a computer telephony using Telephony Application Programming Interface (TAPI) to collect telephone quality speech database, which are very useful in testing and improving certain speech recognition system's ability in recognizing telephone speech. First, discussion on TAPI itself will be presented. Then the whole system will be designed using TAPI. The implementation of the design will be done by using Visual Basic 6.0 as programming language. When the computer telephony is completed, speech samples will be collected to keep as database. The speech samples will be collected through various type of Public Services Telephone Network. Parts of this database are the will be used for experiments to compare the performance of a certain speech recognition system that are trained in two ways: using soundcard quality speech (clean speech) as training samples and other one using telephone quality speech samples.

ABSTRAK

Ahli-ahli kumpulan teknologi pengecaman suara telah lama berhasrat untuk mengaplikasi system pengecaman suara pada system telefon. Justeru, usaha dan kajian telah dijalankan bagi meningkatkan prestasi suatu pengecaman suara supaya mempamerkan ketepatan yang lebih tinggi dan efisien dalam mengecam suara melalui wacana telefon. Projek ini bertujuan untuk membangunkan sebuah sistem yang beroperasi sebagai telefoni komputer menggunakan TAPI. Fungsi sistem ini ialah untuk mengumpul pangkalan data suara berkualiti telefon. Pangkalan ini amat berguna dan penting untuk meningkatkan dan menguji kebolehan suatu sistem pengecaman suara. Implementasi rekabentuk tersebut dilakukan dengan menggunakan bahasa pengaturcaraan Visual Basic 6.0. Setelah sistem selesai dibangunkan, kerja mengumpul pangkalan data dijalankan dengan menggunakan sistem tersebut. Kerja pengumpulan pangkalan data bagi sampel suara telefon melibatkan kepelbagaian rangkaian infa telefon yang sediada di Telekom Malaysia Bhd. Sebahagian daripada pangkalan data yang dikumpulkan akan digunakan dalam eksperimen bagi membandingkan prestasi suatu sistem pengecaman suara yang dilatih menggunakan suara terus. Manakala sebahagian daripadanya akan digunakan untuk melatih sistem pengecaman suara menggunakan kualiti suara telefon. Ini bertujuan untuk mendapatkan perbezaan prestasi kepada sesuatu sistem pengecaman suara yang dilatih menggunakan kualiti suara yang berbeza iaitu menggunakan kualiti suara terus dan juga suara telefon.

TABLE OF CONTENTS

CHAPTER	TITLE	PAGE
	TITLE PAGE	i
	ADMISSION PAGE	ii
	DEDICATION	iii
	ACKNOWLEDGEMENT	iv
	ABSTRACT	v
	ABSTRAK	vi
	TABLE OF CONTENTS	vii
	LIST OF TABLES	xii
	LIST OF FIGURES	xiv
	LIST OF GRAPHS	xvi
	LIST OF ABBREVIATIONS	xvii
	LIST OF APPENDICES	xix
1	INTRODUCTION	1
	1.1 Background of the problem	1
	1.2 Project objective	2
	1.3 Scope of project	2
2	WHAT IS TAPI	3
	2.1 Introduction	3
	2.2 The Telephony API Model	4
	2.3 Typical Configuration	4
	2.3.1 Phone-Based Configuration	5
	2.3.2 Personal Computer Based Configuration	5

2.3.3	Shared or Unified Line Configuration	6
2.3.4	Multiline Configuration	7
2.4	TAPI Architecture	8
2.4.1	Assisted Telephony Services	9
2.4.2	Basic Telephony Services	9
2.4.3	Supplemental Telephony Services	9
2.4.4	Extended Telephony Services	10
2.5	TAPI Hardware Consideration	10
2.6	A Quick Review of How Modems Work	11
3	TELEPHONE NETWORK	12
3.1	Introduction	12
3.2	Local Access Network	13
3.2.1	Introduction	13
3.2.2	Function of Local Line Network	14
3.2.3	Local Loop	14
3.2.4	Local Line Cable Network (Copper)	15
3.2.5	Primary Cable Network	16
3.2.6	Secondary Cable Network	16
3.2.7	Optical Fibers in the Local Network	17
3.2.8	Radio in the Local Loop	18
3.3	Central Office Switching Systems	20
3.3.1	Step by Step Switching	20
3.3.2	Crossbar Switch	21
3.3.3	Electronic Switching System	23
3.4	Multiplexing	27
3.5	Overview Of TM Public Service Telephone Network (PSTN)	28
3.5.1	Public Switched Telephone Network (PSTN)	28
4	AUTOMATIC SPEECH RECOGNITION SYSTEM	33

4.1	Overview of Speech Recognizer	33
4.2	Feature Extraction Techniques	35
4.2.1	Linear Predictive Coding	36
4.2.2	LPC-derived Cepstrum	39
4.2.3	Mel-Frequency Cepstral Coefficients	40
4.3	Dynamic Time Warping, Hidden Markov Models and Vector Quantization	42
4.4	Telephone Speech Database	44
4.4.1	TIMIT and Derivatives (LDC)	44
4.4.2	SIVA (ELRA)	46
4.4.3	PolyVar (ELRA)	48
4.4.4	POLYCOST (ELRA)	49
4.4.5	KING (LDC)	50
4.4.6	YOHO (LDC)	51
4.4.7	Switchboard I-II Including NIST Evaluation Subsets (LDC)	53
4.4.8	Speaker Recognition Corpus (OGI)	56
4.5	Telephone Speech Recognition Previous Research	57
4.5.1	Continuous Recognition Of Large- Vocabulary Telephone-Quality Speech - By Pedro J. Moreno, Matthew A. Siegler, Uday Jain, Richard M. Stern	57
4.5.2	Constructing Telephone Acoustic Models From A High-Quality Speech Corpus – By Mitchel Weintraub And Leonardo Neumeyer	60
4.5.3	Lexical Stress Modeling for Improved Speech Recognition of Spontaneous Telephone Speech in the JUPITER Domain – by Chao Wang and Stephanie Seneff	63
4.5.4	Telephone Speech Recognition Using Neural Networks And Hidden Markov Models – By Dongsuk Yukyz And James Flanagan	66
4.5.5	Estimation Of Channel Bias For Telephone	69

Speech Recognition – By Jen-Tzung Chien

Hsiao-Chuan Wang And Lee-Min Lee

5	METHODOLOGY	70
	5.1 Introduction	70
	5.2 Software/Program Design Specification	71
	5.3 Field Testing	72
	5.3.1 Local Loop Access Network	72
	5.3.2 Transmission Network System	73
	5.3.3 Through Different Switching System/Brand	74
	5.4 Experiment	74
6	RESULTAND ANALYSIS	78
	6.1 Introduction	78
	6.2 Speaker dependent testing result	78
	6.3 Speech Samples recorded through Copper Cable Access testing result	80
	6.4 Speech Samples recorded through Fiber Access Network testing result	82
	6.5 Speech Samples recorded through Transmission Line testing result	84
	6.6 Speech Samples recorded through Same Switching Network testing result	86
	6.7 Speech Samples recorded through Different Switching Network testing result.	87
	6.8 Analysis and discussion	88
	6.9 Summary	104
7	CONCLUSION	105
	7.1 General Summary	105
	7.2 Future Improvement	106

REFERENCES

109

APPENDIX A

112

LIST OF TABLES

TABLE NO.	TITLE	PAGE
4.1	TIMIT Corpus Description	45
4.2	SIVA Corpus Description	47
4.3	PolyVar Corpus description	48
4.4	POLYCOST Corpus Description	49
4.5	KING Corpus Description	50
4.6	YOHO Corpus Description	52
4.7	Switchboard I-II Corpus Description	54
4.8	OGI Speaker Recognition Corpus Description	56
4.9	Word error rate and out-of-vocabulary rate for speakers reported in the official CSR 1994 H2 evaluation set. Boldface entries indicate significantly high values. The overall error rate was 23.5% and the OOV rate was 0.9%.	58
4.10	Word error rates and trigram-hit ratios for speakers in the 1994 H2 evaluation set. Boldface entries indicate significant values. The overall error rate is 23.5% and the trigram hit ratio was 66%.	59
4.11	Effect of Different Training and Front-End Bandwidth on Test Set Performance. Results are Word Error Rate on the 400 Sentence Simultaneous Test Set	61
4.12	Performance on ATIS Telephone Test Data using Wide-Bandwidth HMM Acoustic Models and Different Signal Processing Estimators	61
4.13	Performance on the Aug 1993 WSJ Spoke S6 Development Test Set for Simultaneous	62

Sennheiser/Telephone Recordings

4.14	Word Error for Spontaneous Conversational Speech over Long Distance Telephone Lines	62
4.15	Summary of data sets in the JUPITER corpus	63
4.16	Speech recognition error rates (in percentage) on the development data and test data. “WER” is the word error rate, which is the sum of the substitution, insertion, and deletion error rates. “SER” is the sentence error rate	64
4.17	Table 4.17: Classification accuracy (in percentage) of each individual prosodic feature on the development data	64
4.18	Classification accuracy (in percentage) of various combinations of features on the development data. The combinations of features are described by feature indices as defined in Table 4.17 and this table	65
4.19	Speech recognition error rates (in percentage) on the development data. “WER” is the word error rate, which is the sum of the substitution, insertion, and deletion error rates. “SER” is the sentence error rate	65
4.20	Speech recognition error rates (in percentage) on the test data. The significance level between the baseline performance and the improved performance is also listed	66
4.21	TIMIT Database Result	67
4.22	Table 4.22: Comparison of recognition rates and computational speed of baseline, CMN and three kinds of MAP channel estimator	69
6.1	Speaker Dependent Testing Result	78-79
6.2	Copper Cable Access Speech Testing	80
6.3	Fiber Network Access Speech Testing	82
6.4	Transmission/Junction Speech Testing	84
6.5	Same Exchange Speech Testing	86
6.6	Different Exchange Speech Testing	87

LIST OF FIGURES

FIGURE NO.	TITLE	PAGE
2.1	A typical phone-based TAPI configuration	5
2.2	Typical PC based TAPI configuration	6
2.3	Typical shared line TAPI configuration	7
2.4	Typical unified line TAPI configuration	8
3.1	Local Loop Access Network	15
3.2	Copper Cable Access Network	16
3.3	Fiber Loop Access Network	17
3.4	Radio In Local Loop Access Network	19
3.5	Step-by-Step Switching	21
3.6	Crossbar Switching Technique	23
3.7	Telekom Malaysia Existing Network – 2 Layer Hierarchy	29
3.8	TM Junction Telephony Network DLS – DLS and DLS – DLS	32
4.1	Overview of the ASR	34
4.2	Linear prediction model of speech	35
5.1	Incoming calls made using copper cable	72
5.2	Incoming call made using fiber network system	73
5.3	Incoming call made from different exchange through Tx System	73

5.4	Incoming call made from same/different exchange type	74
6.1	Training Sample of number “Empat” for dependent speaker Azizah	88
6.2	Testing Sample of number “Empat” for dependent speaker Azizah	89
6.3	Testing Sample of number “Empat” for dependent speaker Azizah	90
6.4	Training Sample of number “Kosong” for dependent speaker Shukri	91
6.5	Testing Sample of number “Kosong” for dependent speaker Shukri	92
6.6	Testing Sample of number “Dua” for dependent speaker Malik	96
6.7	Clear & Smooth signal sample of Clear Speech	97
6.8	Sample Signal for number or “Enam” on Copper Access Network	99
6.9	Sample Signal for number or “Enam” on Fiber Access Network	100
6.10	Sample Signal for number or “Enam” on Noisy Environment	102
7.1	Bay Packets’ Agility Voicemail Advance Intelligent Network Concept	107

LIST OF GRAPHS

GRAPH NO.	TITLE	PAGE
6.1	Speaker Dependent Performance Accuracy Result	80
6.2	Copper Access Speech Sample Performance Accuracy Result	81
6.3	Fiber Access Speech Sample Performance Accuracy Result	83
6.4	Junction/Transmission Line Speech Sample Performance Accuracy Result	85
6.5	Same Exchange Type Speech Sample Performance Accuracy Result	86
6.6	Different Exchange Type Speech Sample Performance Accuracy Result	87

LIST OF ABBREVIATIONS

ASR	-	Automatic Speech Recognition
PSTN	-	Public Services Telephone Network
TAPI	-	Telephone Application Programming Interface
WOSA	-	Windows Open Services Architecture
API	-	Application Programming Interface
PC	-	Personal Computer
FDM	-	Frequency Division Multiplexing
TDM	-	Time Division Multiplexing
DEL	-	Direct Exchange Line
MDF	-	Main Distribution Frame
CPE	-	Customer Premises Equipment
DP	-	Distribution Point
D' side	-	Distribution Side
RILL	-	Radio In Local Loop
PBX	-	Private Branch Exchange
ESS	-	Electronic Switch System
SPC	-	Stored Program Control
ROM	-	Read Only Memory
SDM	-	Space Division Multiplexing
PCM	-	Pulse Code Modulation
DNHR	-	Dynamic Non-Hierarchical Routing
INT	-	International
DTS	-	Digital Trunk Switch
DLS	-	Digital Local Switch
DRS	-	Digital Remote Switch

LPC	-	Linear Predictive Coding
MFCC	-	Mel-Frequency Cepstrum Coefficients
VQ	-	Vector Quantization
DTW	-	Dynamic Time Wrapping
HMM	-	Hidden Markov Model
DFT	-	Discrete Fourier Transform
LPCC	-	Linear Prediction Cepstrum Coefficients
ISDN	-	Integrated Service Digital Network
Tx	-	Transmission line

LIST OF APPENDICIES

APPENDIX NO.	TITLE	PAGE
A	Application User Manual	109

CHAPTER 1

INTRODUCTION

1.1 Background of the problem

It is very common nowadays for people, especially university students to make a phone call to get some information like result or application status. Computer telephone usually answers the call automatically. All the caller need to do is just provide some information, for example identity card numbers by pressing telephone buttons and then wait for the response. The computer system at the other end of the line will search for desire information in database based on the callers' input.

This kind of system is indeed useful and convenient, as it can operate with consistent performance and longer time compared to a human telephony. However, the system can only handle inputs formed up by numbers or digits. It cannot handle alphabets, for instance, names. What shall we do if we want to know some information pertaining to persons name such as their extension telephone numbers and etc? The implementation of speech recognition technology through telephone should be best solution for this problem.

1.2 Project objective

The main objective of this project is to analyze the performance accuracy for Automatic Speech Recognition (ASR) Engine in recognizing Telephone Quality Speech for various type of PSTN Network Architecture comparing with Clear Quality Speech.

1.3 Scope of project

This project consists of 3 modules:

1. Design & Develop TAPI Application.
2. Database collection: Collective of speech samples through various type of Telephone Network.
 - I. Access network
 - Copper Access Network
 - Fiber Access Network.
 - II. Switching Network
 - Connected to same Switching Network Element (Brand/Type).
 - Connected to different Switching Network Element (Brand/Type).
 - III. Trunk or Junction Network
3. Performing evaluation of telephone speech quality using Automatic Speech Recognition (ASR) Engine.

CHAPTER 2

WHAT IS TAPI

2.1 Introduction

The telephony Application Interface (TAPI) is one of the most significant API sets to be released by Microsoft. The telephony API is a single set of function calls that allows programmers to manage and manipulate any type of communications link between the personal computer and the telephone line(s). This chapter provides a general overview of the Telephony API and how it fits into the WOSA (Windows Open Services Architecture) model.

2.2 The Telephony API Model

The TAPI design model is divided into two areas, each with its own set of API calls. Each API set focuses on what TAPI refers to as a “device”. The two TAPI devices are:

- i. “Line devices” to model the physical telephony lines used to send and receive voice and data between locations.
- ii. “Phone devices” to model the desktop handset used to place and receive calls.

2.3 Typical Configuration

The TAPI model is designed to function in several different physical configurations. There are four general physical configurations:

- i. *Phone-based*
- ii. *PC-based*
- iii. *Shared or unified line*
- iv. *Multiline*

2.3.1 Phone-Based Configuration

This configuration is best for voice-oriented call processing where the standard handset (or some variation) is used most frequently.

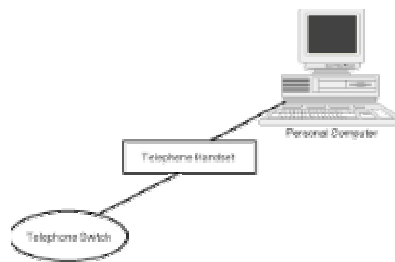


Figure 2.1: A typical phone-based TAPI configuration

This configuration is most useful when the telephone handset is the primary device for accessing the telephone line. Since the telephone rests between the PC and the switch, the PC may not be able to share in all the activity on the line.

2.3.2 Personal Computer Based Configuration

This configuration is best for data-oriented call processing where the PC is used most frequently for either voice or data processing.

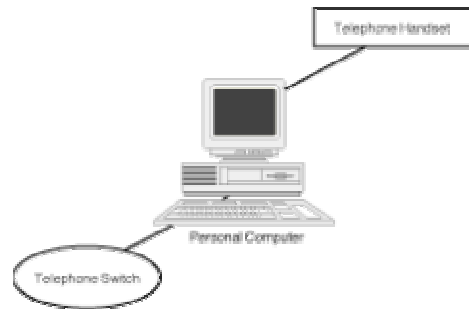


Figure 2.2: Typical PC based TAPI configuration

This configuration is most useful when the PC is the primary device for accessing the telephone line.

2.3.3 Shared or Unified Line Configuration

This is a compromise between phone-based and PC-based systems. It allows all devices to operate as equals along the service line.

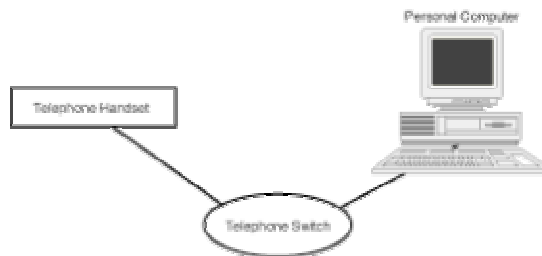


Figure 2.3: Typical shared line TAPI configuration

2.3.4 Multiline Configuration

The primary difference between this configuration and the others is that the PC acts as either a voice-server or a call switching center that connects the outside phone lines to one or more PCs and telephone handsets. The primary advantage of Multiline configurations is that we do not need a direct one-to-one relationship between phone lines and end devices (phones or PCs).



Figure 2.4: Typical unified line TAPI configuration

2.4 TAPI Architecture

The four different levels of TAPI services:

1. Assisted Telephony.
2. Basic Telephony
3. Supplemental
4. Extended Telephony

2.4.1 Assisted Telephony Services

The simplest form of TAPI service is Assisted Telephony. Under the Assisted Telephony interface, programmers can place outbound calls and check the current dialing location of the workstation.

2.4.2 Basic Telephony Services

Basic Telephony is the next level up in the TAPI service model. Basic Telephony function calls allow programmers to create applications that can provide basic in- and outbound voice and data calls over a single-line analog telephone.

2.4.3 Supplemental Telephony Services

The Supplemental Telephony functions provide advanced line device handling (conference, park, hold, forward, and so on). Access to these advanced services is dependent on the type of telephone line to which the workstation is connected. The Supplemental Telephony functions also allow programmers to handle service requests for multiple-line phones.

2.4.4 Extended Telephony Services

The last level of Telephony services is Extended Telephony. Extended Telephony service allows hardware vendors to define their own device-specific functions and services and still operate under the TAPI service model.

2.5 TAPI Hardware Consideration

The three primary types of telephony hardware for PCs:

- Basic data modems - can support Assisted Telephony services (outbound dialing) and usually are able to support only limited inbound call handling.
- Voice data Modems - capable of supporting the Basic Telephony services and many of the Supplemental services.
- Telephony cards - support all of the Basic Telephony and all of the Supplemental Telephony services, including phone device control.

2.6 A Quick Review of How Modems Work

Any information sent over the telephone line has to be in the form of sound waves. In order to accomplish this feat, hardware was invented to convert digital information into sound, and then back again from sound into digital information. Sending data over phone lines involves three main steps. First, a connection must be established between two modem devices over a telephone line. In the second step, the digital information is modulated into sound and then sent over the voice-grade telephone line to the modem. In the last step, the modem at the other end of the call converts (demodulates) the sound back into digital information and presents it to the computer for processing.