Performance of Isolated Digit Speech Recognition in Crowded Environment

This thesis is presented to the Centre for Graduate Studies

in fulfillment of the requirements for

Masters of Science (Information Technology)

Universiti Utara Malaysia

By

Muhamad Arif bin Hashim

JABATAN HAL EHWAL AKADEMIK
(DEPARTMENT OF ACADEMIC AFFAIRS)
UNIVERSITI UTARA MALAYSIA

PERAKUAN KERJA/TESIS
(Certification of Thesis Work)

Kami, yang bertandatangan, memperakukan bahawa
(We, the undersigned, certify that)

MUHAMAD ARIF BIN HASHIM

calon untuk Ijazah
(candidate for the degree of) *SARJANA SAINS (TEKNOLOGI MAKLUMAT)*

telah mengemukakan tesis/disertasinya yang bertajuk
(has presented his/her thesis work of the following title)

PERFORMANCE OF ISOLATED DIGIT SPEECH RECOGNITION IN CROWDED ENVIRONMENT

seperti yang tercatat di muka surat tajuk dan kulit tesis/disertasi
(as it appears on the title page and front cover of thesis work)

bahawa tesis/disertasi tersebut boleh diterima dari segi bentuk serta kandungan, dan liputan bidang ilmu yang memuaskan, sebagaimana yang ditunjukkan oleh calon dalam ujian lisan yang diadakan pada : **05 Ogos 2007**

(that the thesis/dissertation is acceptable in form and content, and that a satisfactory knowledge of the field covered by the thesis was demonstrated by the candidate through an oral examination held on

Pengerusi Viva *(Chairman for Viva)*	:	Prof. Madya Dr. Norshuhada binti Shiratuddin	Tandatangan: *(Signature)*
Pemeriksa Luar *(External Examiner)*	:	Prof. Madya Ir. Dr. Riza bin Sulaiman	Tandatangan: *(Signature)*
Pemeriksa Dalaman *(Internal Examiner)*	:	Prof. Madya Dr. Norita binti Md. Norwawi	Tandatangan: *(Signature)*
Penyelia Utama *(Principal Supervisor)*	:	Dr. Azman bin Yasin	Tandatangan: *(Signature)*
Dekan, Fakulti Teknologi Maklumat *(Dean, Faculty of Information Technology)*	:	Prof. Madya Dr. Suhaidi bin Hassan	Tandatangan: *(Signature)*

Tarikh
(Date) : **05 OGOS 2007**

PERMISSION TO USE

In presenting this thesis in fulfillment of the requirements for a Master of Science (Information Technology) degree from Universiti Utara Malaysia, I agree that the University Library may make it freely available for inspection. I further agree that permission for copying of this thesis in any manner, in whole or in part, for scholarly purposes may be granted by my supervisor(s), or, in their absence, by the Dean of Graduate School. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to Universiti Utara Malaysia for any scholarly use which may be made of any material from my thesis.

Requests for permission to copy or make other use of materials in the thesis, in whole or in part, should be addressed to:

Director
Centre of Graduate Studies
Universiti Utara Malaysia
06010 Sintok
Kedah Darul Aman

ABSTRAK

Sistem pengecaman ucapan ialah sistem yang mengesah apa yang dikatakan oleh seseorang. Objektifnya adalah mengestrak, mengkategori dan mengenalpasti maklumat yang terkandung di dalam ucapan tersebut. Salah satu masalah terbesar di dalam domain pengecaman ucapan adalah gangguan oleh persekitaran yang hingar. Gangguan ini mengurangkan tahap kebolehpercayaan dan ketepatan sistem tersebut. Objektif kajian ini ialah untuk mengenalpasti kekuatan dan kelemahan setiap ucapan digit (dari sifar ke sembilan) di dalam persekitaran orang ramai yang hingar berdasarkan pengukuran pretasi prototaip pengecaman ucapan *Vector Quantization* (VQSR). Prototaip VQSR menggunakan dua ukuran jarak: ukuran jarak *Euclidean* dan ukuran jarak *city block*. Ucapan digit hingar, yang dicipta dengan menggabungkan ucapan digit dari pangkalan data TIDigit dengan bunyi hingar kafeteria dari pangkalan data CLSU, digunakan untuk melatih dan menguji prototaip ini. Prototaip ini juga diuji menggunakan data yang diambil di cafeteria and dipenuhi orang ramai dan hingar. Keputusan dari fasa latihan dan pengujian direkod dan dibandingkan menggunakan satu set analisis pengukuran prestasi. Set ini mengandungi analisis *Sensitivity*, *Specificity*, *Total Accuracy*, *False Acceptance Rate*, *False Rejection Rate* dan *Half Total Error Rate*. Berdasarkan kepada penilaian yang dibuat, ucapan digit yang mempunyai ketahanan dan kebolehpercayaan yang tinggi terhadap bunyi hingar dapat dikenalpasti dan diaplikasikan untuk kegunaan sistem pengecaman ucapan. Akhir sekali, model yang dihasilkan dan cara penilaian prestasi bunyi digit ini boleh digunakan di lain-lain domain yang berkaitan dengan penggunaan ucapan.

ABSTRACT

Speech recognition is a process that recognizes what the speaker says. Its objective is to extract, characterize and recognize the information in the speech signal conveying what the speaker says. One of major problems in speech recognition domain is disturbance caused by background noise. This disturbance can decrease the effectiveness and reliability of the system and its accuracy. This research objective is to measure the performance of isolated digit speech recognition in crowded environment. VQSR prototype uses two kinds of distance measure: Euclidean distance and city block distance. Noisy digit speech, which is constructed from TIDigit speech database and cafeteria noise from CLSU database, is used to train and test the prototype. The prototype is also tested using real data that been recorded in a crowded and noisy cafeteria. Results of training and testing phases are recorded and compared between these two distance measures using a set of performance measurement analysis. This set includes Sensitivity, Specificity, Total Accuracy, False Acceptance Rate, False Rejection Rate and Half Total Error Rate analysis. Based on the performance measurement, a robust and reliable digit speech can be used by user that has high possibility of success and low probability in making errors. Finally, the proposed model and guideline in evaluating the digit speech performance can be use in other speech domain.

ACKNOWLEDGEMENTS

In the name of Allah, the Most Gracious and the Most Merciful.

I would like to extend my thanks and gratitude to:

Allah the Almighty for giving me the good health and mind in doing the research;

The Ministry of Science and Technology for the financial support;

The Universiti Utara Malaysia for the facilities and resources provided;

My supervisor, Dr. Azman Yasin and ex-supervisor Mr. Roshidi Din for their support and knowledge in guiding and assisting me throughout the research;

My beloved father and mother, Hj. Hashim Ishak and Mrs Jamaliah Ilyas and also my siblings, Haslayati, Hafiz, Hanif, Haziman and Hazran for the endless love and patience;

My colleagues, Mr. Azizi, Mr. Hafiz, Mr. Amirulikhzan, Mr. Hanif , Mr.Imam Hazwam and others for the helps, kindness and motivations;

Lastly to my late mother, Mrs. Latipah Ab. Rahim, I dedicate this to you with word

"I've change my star afterall"

Table of contents

List of Tables

List of Figures

List of Abbreviations

Acc	-	Total Accuracy
ANN	-	Artificial Neural Network
ASR	-	Automatic Speech Recognition
ATM	-	Auto teller Machine
CPS	-	Consumer Perception Survey
DFT	-	Discrete Fourier Transform
DTW	-	Dynamic Time Warping
FAR	-	False Acceptance Rate
FFT	-	Fast Fourier Transform
FN	-	False Negative
FP	-	False Positive
FRR	-	False Rejection Rate
HMM	-	Hidden Markov Model
HTER	-	Half Total Error Rate
LBGVQ	-	Linde, Buzo and Gray – Vector Quantization
LPC	-	Linear Predictive Coding
LPCC	-	Linear Prediction Cepstral Coefficients
MFCC	-	Mel-frequency Cepstrum Coefficients
PIN	-	Personal Identification Number
PLP	-	Perceptual Linear Predictive
Sen	-	Sensitivity
Spec	-	Specificity
TN	-	True Negative
TP	-	True Positive
UUM	-	Universiti Utara Malaysia
VQ	-	Vector Quantization
VQSR	-	Vector Quantization Speech Recognition

CHAPTER 1

INTRODUCTION

Speech recognition technology have been a very popular technology that been applied into human daily life activities. From telephone network services (Zissman, 1996) to medical applications (Grasso, 2003), speech recognition was applied to ease up and simplified various works. These achievements were due to improved speech technology, falling costs of speech hardware and the changing customer perceptions which turned speech recognition technology no longer a dependant on large-scale deployments but a cost saving. In order to gain a good ROI (return on investment), a Consumer Perception Survey has been done by Fluency Voice experts (2007) revealed that almost three-quarters (74 percent) of the survey respondents reported that they had used speech technology in the past. Over half (58 percent) have used a speech system recently, and almost all (88 percent) of those reported that they found it easy-to-use. For the trials themselves, the vast majority of users were able to successfully complete all the transaction using speech recognition system.

Meanwhile, around a half (58 percent) of respondents endorse using speech as an alternative to the internet for making purchases. Confirming that the use of this technology is on the increase, a two-thirds (66 percent) of respondents prefer to use speech technology over the internet if a better purchasing deal is in the offering. In the

The contents of the thesis is for internal user only

References

Adda-Decker, M., Antoine, F., Boula de Mareuil, P., Vasilescu, I., Lamel, L., Vaissiere, J., Geoffrois, J. & Liénard,. J. S. (2003). Phonetic knowledge, phonotactics and perceptual validation for automatic language identification. *Proc. 15th ICPhS*, Barcelona.

Aida–Zade, K. R., Ardil, C. & Rustamov, S. S. (2006). Investigation of Combined use of MFCC and LPC Features in Speech Recognition Systems. *International Journal of Signal Processing, 3* (2). 105-111.

Bengio, S., Keller, M. & Mari´ethoz J. (2005). The Expected Performance Curve. *International Conference on Machine Learning, ICML, Workshop on ROC Analysis in Machine Learning*, Bonn, Germany.

Bishop, C. (1995). Neural network for pattern recognition. Oxford University Press.

Bourouba, E. H., Bedda, M. & Djemili, R. (2006). Isolated Words Recognition System Based on Hybrid Approach DTW/GHMM. *Informatica*. 373-384.

Campbell, J. P. (1997). Speaker Recognition: A Tutorial. *Proceedings of the IEEE, 85* (9). 1437-1462.

Chien, J. & Furui, S. (2005). Predictive Hidden Markov Model Selection for Speech Recognition. *IEEE Transactions on Speech and Audio Processing, 13*(3). 377-387.

Common Criteria Biometric Evaluation Methodology Working Group. (2002). Common Methodology for Information Technology Security Evaluation (version 1.0). Retrieved May 13, 2006, from http://www.cesg.gov.uk/site/ast/biometrics/media/BEM_10.pdf

Cook, S. (2002). Speech Recognition HOWTO. Retrieved Retrieved October 3, 2006, from http://www.ibiblio.org/pub/linux/docs/HOWTO/other-formats/pdf/Speech-Recognition-HOWTO.pdf

Cosi, et al. (1998). Connected Digit Recognition Experiments with the OGI Toolkit's Neural Network and HMM-Based Recognizer. *Interactive Voice for Telecommunication Applications, IVTTA'-98*, 135-140.

Chitu, A. G. et al. (2007). Comparison between Different Feature Extraction Techniques for Audio-Visual Speech Recognition. *Journal on Multimodal Interfaces*, pp. 16, Springer.

Ephraim, Y., Lev-Ari, H. & Roberts, W. J. J. (2005). A brief survey of speech enhancement. *CRC Electronic Handbook*, 2nd edition, CRC Press.

Fluency Voice. (2007). Consumers give their voice to speech recognition. *FST*. Retrieved March 1, 2007, from http://www.gdsinternational.com/infocentre/artsum.asp?mag=187&iss=183&art=268962&lang=en

Forsberg, M. (2003). Why is Speech Recognition Difficult?. Retrieved March 1, 2007, from http://www.speech.kth.se/~rolf/gslt_papers/MarkusForsberg.pdf

Furui, S. (2005). 50 years of progress in speech and speaker recognition. *Proc. SPECOM 2005*. 1-9.

Gavat, I., Zirra, M. & Enescu, V. (1996). A Hybrid NN-HMM System for Connected Digit Recognition over Telephone in Romanian Language. *Third IEEE Workshop on Interactive Voice Technology for Telecommunications Applications*. 37-40.

Grasso, M. A. (2003). The Long-Term Adoption of Speech Recognition in Medical Applications. *Proceedings of the 16th IEEE Symposium on Computer-Based Medical Systems, CBMS'03*. 257 – 262.

Hedberg, S.R. (1997). Dictating this article to my computer: automatic speech recognitionis coming of age. *IEEE Intelligent Systems and Their Applications, 12*(6). 9-11.

Hennebert, J., Hasler, M. & Dedieu, H. (1993). Neural Networks in Speech Recognition. Retireved February 11, 2005, from http://citeseer.ist.psu.edu/cache/papers/cs/3630/http:zSzzSzcircwww.epfl.chzSzstaffzSzhenneberzSzjeanzSzpapierszSzMicrocompzSzneuralnet.pdf/neural-networks-in-speech.pdf

Hunt, M. J. (1999). Spectral signal processing for ASR. *Proc. ASRU'99*.

Juang, B.H., D. Childers, R.V. Cox, R. De Mori, S. Furui, J. Mariani, P. Price, S. Sagayama, M.M. Sondhi, & R. Weishedel. (1998). Speech Processing: Past, Present and Outlook. *IEEE Signal Processing Magazine,* May 1998.

Kasper, K. et al. (1995). A fully recurrent neural network for recognition of noisy telephone speech. ***International Conference on Acoustics, Speech, and Signal Processing, 1995. ICASSP-95.*** vol. 5. 3331 – 3334.

Kinnunen, T., Karpov, E. & Fränti, P. (2006). Real-Time Speaker Identification and Verification. *IEEE Transactions on Audio, Speech, and Language Processing. 14*(1). 277-288.

Kirchhoff, K. & Parandekar, S. (2001). Multi-stream statistical language modeling with application to automatic language identification. *Proceedings of Eurospeech 01*. 803-806.

Kirchhoff, K., Parandekar, S. & Bilmes, J. (2002). Mixed-Memory Markov Models for Automatic Language Identification. *IEEE International. Conference on Acoustics, Speech, and Signal Processing*, Proceedings of ICASSP'02, 1. 761-764.

Klusacek, D., Navratil, J., Reynolds, D. & Campbell, J. (2003). Conditional Pronunciation Modeling In Speaker Detection. *IEEE International. Conference on Acoustics, Speech, and Signal Processing, Proceedings of ICASSP'03*, 4. 804-807.

Levy, C., Linares, G. & Nocera, P. (2003). Comparison of Several Acoustic Modeling Techniques and Decoding Algorithms for Embedded Speech Recognition Systems. *2003 Workshop on DSP in Mobile and Vehicular Systems*. Nagoya, Japan.

Low, R. & Togreri. (1998). Speech Recognition Using the Probabilistic Neural Network. *Proceedings of ICSLP98 (SST Student Day)*, Sydney, Australia.

Lockwood, P. & Boudy, J. (1991). Experiments with a Non-linear Spectral Subtractor (NSS), Hidden Markov Models and the Projection, for Robust Speech Recognition in Cars. *Proceedings European Conference on Speech Communication and Technology, Proc. EUROSPEECH*..79-82.

Md. Rashidul Hasan et.al. (2004). Speaker Identification using Mel Frequency Cepstral Coefficient. *3rd International Conference on Electrical & Computer Engineering, ICECE 2004*. 565-568.

Miller, D. R. H., Leek, T. & Schwartz, R. M. (1999). A Hidden Markov Model Information Retrieval System. *22nd ACM International Conference on Research and Development in Information Retrieva, Proceedings of SIGIR-99*. 214-221.

Mohri, M. & Riley, M. (1997). Weighted Determination and Minimization for Large Vocabulary Speech Recognition. *Proceedings European Conference on Speech Communication and Technology, Proc. Eurospeech'97*, 1. 131-134.

Mut, O. & Göktürk, M. (2005). Improved Weighted Matching for Speaker Recognition. Transactions on Engineering, Computing and Technology. 5. 229-231.

O'gorman, L. (2003). Comparing Passwords, Tokens and Biometrics for User Authentication. *Proceedings of the IEEE, 91*(12). 2019-2020.

Orman, O. D. & Arslan, L. M., (2001). Frequency Analysis of speaker Identification. *2001 Speaker Odyssey*.

Owens, F.J. (1993). Signal processing of speech, Macmillan.

Paliwal, K. & Atal, B. (1993). Efficient Vector Quantization of LPC Parameters at 24 Bits/Frame. *IEEE Transactions on Speech and Audio Processing, 1*(1). 3-14.

Pearce, D. & Hirsch, H. (2000). The Aurora Experimental Framework for the Performance Evaluation of Speech Recognition Systems under Noisy Conditions. *6th International Conference on Spoken Language Processing, ICSLP 2000.* 181-188.

Perez, E. & Rodriguez-Esteban, R. (2004). Oreja... for the design of psychoacoustic experiments (Manual User's version 1.0). Retrieved October 2, 2005, from http://www.ee.columbia.edu/~raul/oreja/manual.htm

Picone, J. (1996). Fundamentals of Speech Recognition: A Short Course. Retrieved October 2, 2005, from http://www.stillhq.com/diary/asr-short.pdf

Rabiner L.R. & Juang B.H. (1993). *Fundamentals of speech recognition*. Prentice-Hall, Englewood Cliffs, NJ.

Raiko, T. (2003). Speech recognition in noisy environments: A survey. Retrieved February 11, 2005, from http://www.cis.hut.fi/Opinnot/T-61.6020/2003/Kalvot2003/raiko1.pdf

Rodrigues, F. & Trancoso, I. (1999).Digit Recognition Using the SPEECHDAT Corpus. Conference on Telecommunications ,CONFTELE'99.

Saldana, I. & Ginberg, D. (2003). Remote Speaker and Speech Recognition : A senior design project. Retrieved August 11, 2005, from http://www.yov408.com/tutorials/speech_recog.pdf

Savage, J., Rivera, C. & Aguilar, V. (1998). Isolated Word Speech Recognition using Vector Quantization and Artificial Neural Network. *SPECOM 08, Russia.*

Sebastiani, F. (2002). Machine Learning in Automated Text Categorization. *ACM Computing Surveys*, *34*(1). 1–47.

Shannon, B. J. & Paliwal, K. (2005). Influences of Autocorrelation Lag Ranges on Robust Speech Recognition. *Proc. IEEE Intern. Conf. on Acoustics, Speech and Signal Processing*. vol. 1, 545-548

Siivola, V., Kurimo, M. & Lagus, K. (2001) Large Vocabulary Statistical Language Modeling for Continuous Speech Recognition in Finnish. *Proceedings of the 7th European Conference on Speech Communication and Technology, Proc. Eurospeech.*

Stokes-Rees, I. (2002). A Study of the Automatic Speech Recognition Process and Speaker Adaptation. Masters Thesis.

Vaishnavi, V. & Kuechler, B. (2004). Design Research in Information System. Retrieved March 26, 2005, from http://www.isworld.org/Researchdesign/drisISworld.htm

Verlinde, P., Chollet, G., & Acheroy, M. (2000). Multimodal identity verification using expert fusion. *InformationFusion, 1*. 17–33.

Xu, R. & Wunsch, D. (2005). Survey of Clustering Algorithms. *IEEE Transaction on Neural Network*. vol 16, no. 3.

Zhang, Z. (2002). A Study on Increasing Robustness against Speaker and Noise Variations in Speech Recognition. A Dissertation Submitted to Department of Computer Science, Graduate School of Information Science and Engineering, Tokyo Institute of Technology, Japan.

Zhao, J., Kuang, J. & Xie, X. (2003). The Formant Structure based Feature Parameter for Speech Recognition. *IEEE Workshop on Statistical Signal Processing 2003*. 605-608.

Zissman, M. (1996). Comparison of four approaches to automatic language identification of telephone speech. *IEEE Transaction on Speech and Audio Processing, 4*(1). 31-44.

Zweig, & Campbell (1993). ROC plots: a fundamental evaluation tool in clinical medicine. *Clinical Chemistry*, *39* (8). 561–577.