MULTI-SPEAKER FREQUENCY WARPING VOCAL TRACT LENGTH NORMALIZATION FOR SPEAKER INDEPENDENT SPEECH RECOGNITION

JENSEN WONG JING LUNG

A thesis submitted in fulfillment of the
requirements for the award of the degree of
Master of Science (Computer Science)

Faculty of Computing
Universiti Teknologi Malaysia

JUNE 2014

This thesis is dedicated to my gracious God as well as my beloved father and mother for their endless support and encouragement.

ACKNOWLEDGEMENT

During the progress of my master research study, it's a blessing to meet many kinds of people who came across my path. First of all, my utmost appreciation, gratitude and respect to my main supervisor, Dr Md Sah Hj Salam, as well as my co-supervisor, Associate Professor Dr Mohd Shafry Mohd Rahim, who had patiently given me the guidance, comments and advice toward the completion of this research and Master thesis.

Special thanks go to my parents who have always morally supported me and encouraged me in my research. Same goes to my brothers who have never ceased to care for me and motivate me to keep on going. Not to forget my friends here who have always been with me throughout my difficult time as well as given me the courage and endurance to continue my Master degree.

Lastly, to those whom I have met, or those I didn't mention here, my biggest "Thank you" to all of you for the countless effort and support.

ABSTRACT

One of the important issues in speaker independent speech recognition system is to compensate speaker variability. Speaker variability is usually related to the physical difference in vocal tract length. Compensation on vocal tract length variation can be made using Vocal Tract Length Normalization (VTLN) method which is known to be able to normalize speech utterances via specific speaker frequency warping. However, this approach leads to repetition process in finding optimal value for warping per speakers, which increase computational cost. This work proposed an alternative approach in finding optimal warping factor in VTLN via multi-speaker frequency warping in which only one optimum warping factor value is used for all speakers. The proposed multi-speaker frequency warping VTLN is experimented using different experimental setup on language model, phoneme categorization and warping values through trial and error method. The data used in this work is large vocabulary TIMIT dataset and Hidden Markov Model Toolkit (HTK) is used for classification purpose. The obtained results show that the proposed approach has achieved improvement of up to 1.0% higher phoneme accuracy rate compared to the baseline result. The proposed approach performance is at par with speaker-specific warping approach but with added advantage of lesser computational cost.

ABSTRAK

Salah satu isu penting dalam sistem pengecaman ucapan bebas adalah mengimbangkan kepelbagaian pengucap. Kepelbagaian pengucap kebiasaannya adalah berkaitan dengan perbezaan fizikal dalam panjang saluran vokal. Pendekatan kaedah *Vocal Tract Length Normalization* (VTLN) diketahui dapat menormalkan ucapan-ucapan melalui ledingan frekuensi pengucap khusus. Ledingan ini boleh mengimbangi setiap saiz variasi saluran vocal dari setiap pengucap. Walaubagaimanapun, pendekatan ini melibatkan proses mencari nilai optima secara berulang dan ini menyebabkan kos pengiraan meningkat. Penyelidikan ini mencadangkan satu pendekatan alternatif dalam mencari faktor ledingan optimum VTLN melalui kaedah pelbagai pengucap frequensi ledingan, di mana hanya satu nilai faktor ledingan optimum diperlukan untuk semua pengucap. Cadangan VTLN ledingan frekuensi pelbagai pengucap diselidik secara ujikaji menggunakan persediaan yang berbeza pada model bahasa, pengkategorian *phoneme* dan nilai-nilai meleding melalui pendekatan cuba-jaya. Data yang digunakan dalam penyelidikan ini terdiri daripada set data vokabulari besar TIMIT dan *Hidden Markov Model Toolkit* (HTK) digunakan untuk tujuan klasifikasi. Hasil kajian menunjukkan bahawa pendekatan yang dicadangkan ini mencapai kemajuan kadar ketepatan sebanyak 1.0% lebih tinggi berbanding dengan kadar ketepatan persediaan asas. Walaupun cadangan pendekatan ini setanding dengan pendekatan ledingan pengucap khusus dari segi prestasi kadar ketepatan, tetapi ianya mempunyai kelebihan dari segi penggunaan kos pengiraan yang rendah.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATION

ASR	Automatic Speech Recognition
cMLLR	Constrained MLLR
CMLSN	Constrained MLLR Speaker Normalization
CMN	Cepstral Mean Normalization
DCT	Discrete Cosine Transform
FFT	Fast Fourier Transform
HMM	Hidden Markov Modeling
HTK	HMM ToolKit
HIFREQ	Highest Frequency Bandwidth
LOFREQ	Lowest Frequency Bandwidth
MAP	Maximum *a posteriori*
MFCC	Mel-Frequency Cepstral Coefficient
ML	Maximum Likelihood
MLLR	Maximum Likelihood Linear Regression
SAT	Speaker Adaptation Technique
SI-ASR	Speaker Independent Automatic Speech Recognition
TIMIT	DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus
VTL	Vocal Tract Length
VTLN	Vocal Tract Length Normalization
WER	Word Error Rate

LIST OF SYMBOLS

W	-	transcription of the input utterance
Y	-	acoustic observation sequence data
$P(Y\|W)$	-	acoustic model; probability of sequence of acoustic observations conditioned on the input utterance.
$P(Y)$	-	language model; Grammar
$\hat{W}$	-	estimated sequence of the transcription of the input utterance
$\hat{\mu}$	-	estimated mean for model mean vector μ
μ	-	model mean vector
G	-	transformation matrix
b	-	bias vector
f_N	-	Nyquist frequecy, upper frequency bandwidth
f_U	-	upper warp cutoff frequency
f_L	-	lower warp cutoff frequency
f_0	-	lower frequency bandwidth
α	-	warping factor
k	-	pre-emphasis coeffient value
S_n	-	samples
S'_n	-	pre-emphasized samples
As	-	sampling rate of the acoustic speech signal (Hertz)
f_b	-	time usage per frame (second)
B^{FFT}	-	total number of FFT bins
$f_{original}$	-	original maximum frequency bandwidth
f_{scaled}	-	scaled maximum frequency bandwidth
F	-	frequency warping range
$C^{\alpha F}$	-	VTLN-warped MFCC input
$M^{\alpha F}$	-	VTLN-warped HMM acoustic models

$Acc^{\alpha F}$ - accuracy rate for VTLN-warped HMM and MFCC

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

1.1 Introduction

Automatic speech recognition (ASR) is part of the speech technology that involved speech processing to simulate human intelligence. This technology simplifies the human-machine interaction through repetitive processes of sensing and converting spoken words into the machine-readable input to execute certain tasks or application functions. These certain tasks include turning the written text into speech, dictating speech into text, and executing commands through voice input.

There are many examples of speech technology that help simplify human daily tasks. In health care and medical field, ASR assists in gaining increased workflow efficiency and productivity in conjunction with electronic medical records for transcription, dictation as well as clinical decision support. In military operations, ASR is capable of handling high-performance fighter aircrafts (Englund, 2004) and helicopters, enabling immediate access and control of large, rapidly changing information databases in battle command center. In telephony field, ASR aids in processing the voice input through mobile apps in every mobile phone with high-speed processor and telephony server in handling either voice messages or step-by-step voice command interaction. ASR shows potential in implementing its technology on future applications and devices like automatic translation, vehicle

navigation systems, real-time voice writing, robotics, video game, and other domains related to the implementation of front-end and back-end documentation.

The revelation of ASR's potential for different fields in daily activities shows the importance of understanding the ASR's fundamental process in preparation, training and recognition. For the purpose of making ASR usable for all speakers in different fields, many ASR systems require vast amount of training speech data from wide variety of individuals without going through re-training process (Holmes and Holmes, 2001). Most training speech data consist of either in one single word or multiple words in one sentence. However, for handling large sized vocabulary, ASR focuses on using phonemes instead of using words to correctly recognize large vocabulary speech.

Even though large vocabulary ASR can be trained word-by-word, it is impractical as the total word vocabulary size can reach around thousands to hundreds of thousands of words (Flanagan, Allen and Hasegawa-Johnson, 2008; Rabiner and Juang, 1993; Rabiner and Juang, 2008). This large vocabulary size training and recognition process involves longer speech processing time as well as larger storage resources usage. Phoneme on the other hand has very small vocabulary size within the average range of 50 phonemes (Breen, Bowers and Welsh, 1996; Donovan, 1996; Flanagan, Allen and Hasegawa-Johnson, 2008; Rabiner and Juang, 1993; Rabiner and Juang, 2008; Garofolo *et al.*, 1993; Fernández, Graves and Schmidhuber, 2008), thus it gave the advantages of smaller storage usage and shorter overall speech processing time. Phoneme's small linguistic size also turns the phoneme recognition, which focuses on recognizing phonemes from every speech input, into a very delicate recognition task. This enables the phoneme recognition to observe actual recognition performance level on every phoneme in each word and sentence.

According to the definition given by John Holmes and Wendy Holmes (2001), phoneme is the smallest linguistic unit in a spoken language. Phoneme is the core fundamental unit in ASR system where ASR uses phonemes to match every spoken word based on word's pronunciation. A word pronunciation is represented

linguistically in phoneme sequences, which contribute in defining the exact meaning of the whole speech context. Substitution of one phoneme unit for another unit results in making a distinction of the speech's meaning. In other words, replacing one phoneme with other phoneme affects one word's pronunciation and meaning within whole speech context. This distinction shows the uniqueness in every phoneme by its own unique fundamental frequency and formant frequencies, thus enabling ASR to distinguish and identify every phoneme at recognition stage. In addition, phoneme increases the efficiency of continuous ASR system by enabling ASR to recognize hundreds and thousands of words through matching phoneme sequences with the respective word's pronunciation.

However, phoneme does not free from the problems related to speech signal. Majority of the researchers still have issues on phoneme recognition performance as phonemes are too fragile in spectrum form (Salam, Mohamad and Salleh, 2011). In English language, most phones that manifest the phonemes' acoustic signal are too short (Flanagan, Allen and Hasegawa-Johnson, 2008) to be distinguished precisely. Similar to ASR problem, acoustic signal for each phoneme may prone to signal corruption from external noises and transmission distortion (Holmes and Holmes, 2001; Furui, 2009; Flanagan, Allen and Hasegawa-Johnson, 2008). This corruption is related to environment variability that affects the signal's quality and recognition performance. Speech signal variation from a wide variety of individuals adds up the difficulties in achieving high performance ASR, especially speaker-independent mode ASR (SI-ASR). These environmental and speaker variability are elaborated more in the problem background section.

1.2 Research Background

According to John Holmes and Wendy Holmes (2001), environment and speaker variation are two main causes of acoustic variations that contribute to signal distortion problems. Both of these variations that distort the speech signal are best illustrated by Professor Dr Sadaoki Furui (2009) in Figure 1.1. Environment

variation is related to the background signal distortion captured in both analog and digital form. Speaker variation is distinguished from the speakers with context difference and socio-physiology difference.

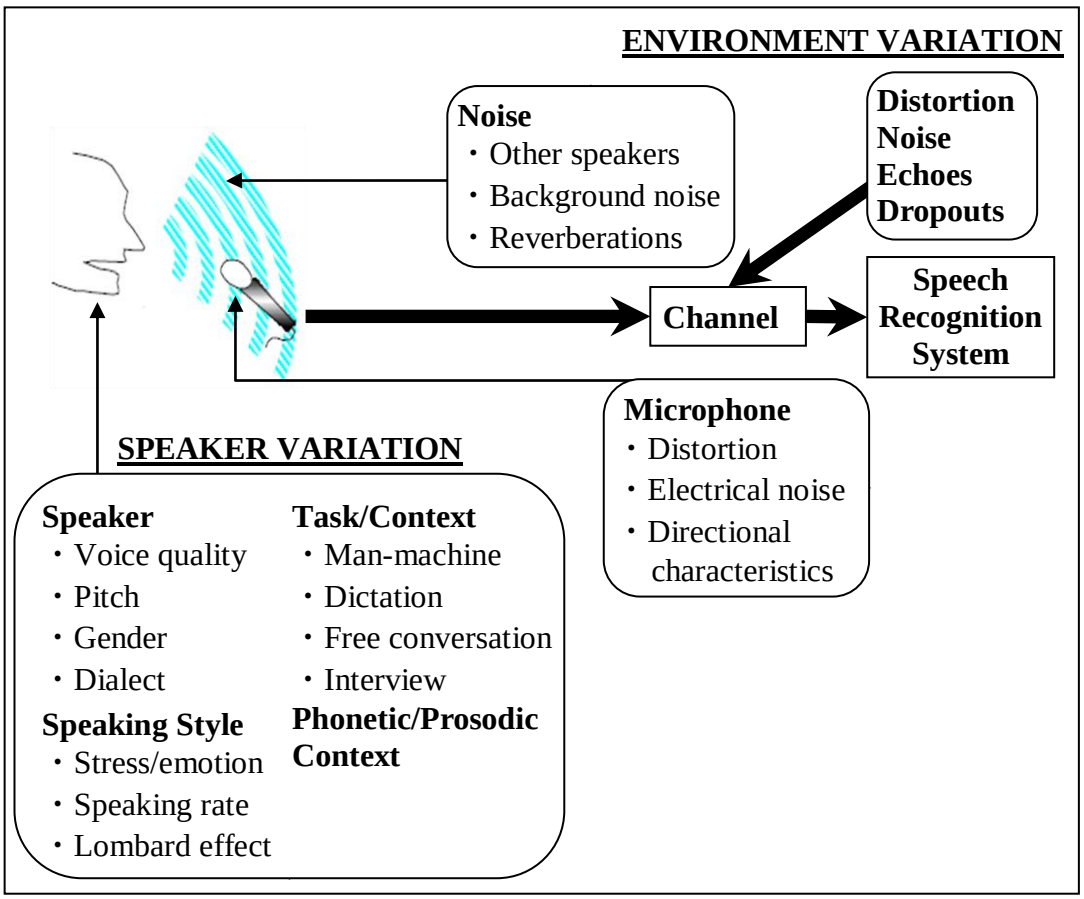

Figure 1.1 Main Causes of Acoustic Variations in Speech (adapted from Furui, 2009)

Environment variation is divided into noise and channel variation. As most ASR systems operate under noisy environment with different quality type of microphones, the signal that feeds into the recognizer will not be the closest match with the one that is uttered by the speaker. During the signal transmission from speaker's mouth to recognizer, the external noise from the surrounding may present and capture together with the speech signal to the recognizer. Beside the captured noise, the channel variation manifested through the medium of transmission between the microphone and recognizer, causing alteration of the transmitting signal's

characteristic and turning signal partially into dropouts or echoes. The medium used for transmitting signal can be either via wired and wireless transmissions or through electronic devices such as telephone channel and microphone transducer.

Speaker variation comes in two different types of variation, namely intra-speaker and inter-speaker variability (Holmes and Holmes, 2001; Dileep and Sekhar, 2012). Intra-speaker variability is related to variation within the same speaker caused by each speaker's emotion and stress (Schroeter, 2008). This variation affects the speaker to speak differently out of normal habitual way of speaking. Inter-speaker variability is related to the socio-linguistic and physiological differences between different human speakers (Umesh, 2011; Schroeter, 2008). Factors such as voice quality, educational level, regional dialect background, articulator habits and gender are the general example of inter-speaker variability.

Although environment and speaker variations are two different causes of acoustic variations, there is no denying that certain environment types lead to the speaker variation (Holmes and Holmes, 2001). Speaker's dialect and accent are parts of the inter-speaker variation that depends on the region and culture background the speaker lives in. Lombard effect, part of the intra-speaker variation, is the effect that causes the speaker to change speaking style and increase speech volume in the presence of acoustic noise from the surrounding area. The context based environment also influences the way speaker utters a speech, such as dictation, free conversation, man-machine interaction and interview session. Despite all the mentioned environment types, this environment variation can be controlled and fixed as ASR operates best when it is operated in the same environment condition. If the whole ASR training session is done on the near identical condition with its operating environment condition, then ASR needs to prioritize on reducing speaker variation in speech signal.

Speaker variation can be observed through SI-ASR systems. As SI-ASR systems are designed with the aim to minimize acoustic differences between speeches from various different speakers, the inter-speaker variability is taken into

account as the main source of speech differences (Umesh, 2011). For this reason, inter-speaker variability becomes the main research focus of the problem that is related to physiological and socio-linguistic differences. In general, each involving speakers from all walks of life possess different and unique voice of their own, making it distinguishable by every human speaker. Despite the differences, human speakers able to identify one single speech from different speakers, while SI-ASR machine faces difficulty to perform accurate recognition on that similar speech. This difficulty leads us to seek for the approach to compensate inter-speaker variability.

There are two general approaches known to compensate inter-speaker variation, namely adaptation and normalization. The difference between adaptation and normalization approaches lies on the SI-ASR component that these approaches are operated and implemented. Speaker adaptation approach makes the SI-ASR trained speech model reference to closely match with the targeted speaker by allowing a small amount of data from targeted speaker to transform existing model reference. On the other hand, speaker normalization approach normalized speech spectrum, minimizing the speaker variation effect from the incoming speech signal by focusing on front-end signal preprocessing that closely collaborates together with feature extraction part of the SI-ASR.

Among these approaches, Vocal Tract Length Normalization (VTLN) is the known method in speaker normalization approach that is widely used for compensating inter-speaker variability in SI-ASR (Holmes and Holmes, 2001; Furui, 2009; Umesh, 2011; Giuliani, Gerosa and Brugnara, 2006; Lee and Rose, 1996; Lee and Rose, 1998; Young *et al.*, 2006; Young, 2008). VTLN focuses on scaling or warping the speech spectrum linearly by shifting the positions of the spectral peaks corresponding to the formant frequencies, known as frequency warping approach. It is based on the facts that variations in human's vocal-tract length are manifested in speech spectrum, causing the positions of spectral peaks to be shifted in frequency in an approximately linear fashion (Young, 2008). This spectral shifting is done within the frequency boundary according to the estimated warp factor.

1.3 Problem Statement

The usual VTLN approach in estimating optimal warp factor and frequency boundary is based on every single specific speaker. However, this approach leads to repetition process in finding optimal value for warping per speakers which increase computational cost. This existing evidence is based on the physiological differences argument by which this speaker-specific warp factor is often estimated within a small limited range of factor value using maximum likelihood (ML) approach (Umesh, 2011; Lee and Rose, 1996; Lee and Rose, 1998). Even though this estimation yielded significant recognition performance as reported by other researchers, the task of finding every single warp factor per speaker brings disadvantage especially for SI-ASR systems with wide variety of individual speakers. This causes to have multiple different values of warp factors for normalizing each speaker. Therefore, it is important that this research is extended to not only to reduce the number of multiple warp factor values but also to increase the estimation range of warp factor value for this VTLN method.

1.4 Research Question

Based on the problem discovered in estimating optimal frequency warp factor and frequency boundary, the main question arises for this research: "How to find the optimal warp factor without having to consider the differences between each speaker?" This research question addresses two considerations; seeking optimal frequency warp factor, and applying that warp factor onto multiple speakers. Through these considerations, multi-speaker frequency warping approach is proposed as better alternative normalization approach toward obtaining one optimal warp factor for every single speaker. Beside usage reduction in computational resource, this proposed research not only aspires to remove the needs for one optimal warp factor per speaker but also able to normalize all speakers and yield similar or better recognition results than speaker-specific VTLN.

1.5 Objectives

In order to better focus within the context of the problem statement and achieve expected results, it is important to clearly define the objectives of this research. This research seeks:

i. To use Hidden Markov Model (HMM) as a standard recognition platform to prepare and obtain baseline recognition result for experimentation on inter-speaker variability problem.
ii. To introduce new enhancement approach in obtaining optimal frequency warping value for VTLN within HMM to compensate inter-speaker variability.
iii. To compare the new warping approach with baseline HMM based on phoneme and word recognition performance.

1.6 Scopes

For the purpose in fulfilling the above objectives, the scopes for this research need to be clearly identified the research exploration area within designated boundary. The definition of research scopes is important as a guideline toward overall research progress, aiding the research progress to stay focused. These scopes include the following aspect:

i. This research only focus on inter-speaker variability, thus any environmental related variability is not included in this experiment to avoid bias result, as justified from problem background in Section 1.1, page 5.
ii. This research emphasized on tackling the physiological difference of the speakers and not gender difference. The non-inclusive gender

difference scope in this research is justified through literature review in Section 2.3.1, page 18.

iii. All experiments are conducted in continuous phoneme level speech recognition setting as this research uses large vocabulary sized standard TIMIT continuous speech corpus as the main speech data source (Section 4.3.1).

iv. The research work progress is to be executed using HMM Toolkit (HTK) as the main speech recognizer (Section 4.4).

1.7 Significant of Research

i. This research will be a significant work of discovery towards an alternative implementation approach of VTLN.

ii. This research problem and idea are highlighted based on the literature reviews done regarding to variation in human speech production as well as the way conventional VTLN is implemented in SI-ASR systems.

iii. With the defined research objectives and significant results obtained from the experiments, it is hoped that new knowledge, understanding and verification related to the existing VTLN is attained.

iv. At the same time, this new significant perspective on multi-speaker frequency warping is gained and new opportunity is open up to explore this type of normalization further beyond the phoneme level in SI-ASR systems.

v. This completed research thesis will serve as a milestone for further expansion on multi-speaker frequency warping as well as the knowledge sharing of this finding to every researcher involved in speaker normalization process.

1.8 Thesis Structure

This thesis structure is arranged and divided into 6 different chapters, starting with this Chapter 1 as the Introduction of my research problem issue, as well as the research objectives, scopes and the significant of this research based on the result from the experiments. In Chapter 2, every literature review is conducted on the problem issue related to the methods and approaches used to compensate inter-speaker variability. Chapter 3 highlights the planned research methodology for this research. This methodology includes the preparation for experimental data and setup as well as the way research experiment is conducted to obtain expected results. All results and discussions on the research experiments' outcome and the validation of the research hypothesis are to be elaborated in Chapter 4 and 5. Finally, the conclusion for overall research is made in Chapter 6.

REFERENCES

Becker, T. (2007). The Influence of Intra-Speaker Variability in Automatic Speaker Identification. *Proc of IAFPA 2007*. Plymouth, UK.

Benesty, J., Sondhi, M.M., and Huang Y.T. (Eds.) (2008). *Springer Handbook of Speech Processing*. (1st ed.) Berlin, Heidelberg: Springer-Verlag.

Breen A., Bowers E., Welsh W. (1996). An Investigation into the Generation of Mouth Shapes for a Talking Head. *Proceedings of ICSLP 96 (4)*.

Cole, R.A., Mariani, J., Uszkoreit, H., Zaenen, A., Zue, V., Varile, G., and Zampolli, A. (Eds.) (1997). *Survey of the State of the Art in Human Language Technology*. New York, USA.: Cambridge University Press and Giardini.

Davis, S.B., and Mermelstein, P. (1980). Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences. *IEEE Trans. on Acoustics, Speech and Signal Processing*. 28(4), 357-366.

Deller, J., Proakis, J. and Hansen, J. (2000) *Discrete-Time Processing of Speech Signals*. (2nd ed.) New York, USA.: Wiley-IEEE Press.

Dileep, A.D. and Sekhar, C.C. (2012). Speaker Identification Using Intermediate Matching Kernel-Based Support Vector Machines. *Forensic Speaker Recognition: Law Enforcement and Counter-Terrorism*, eds. Neustein, A. and Patil, H.A., (1st ed.) Springer. 389-424.

Donovan R. (1996). Trainable Speech Synthesis. PhD. Thesis. Cambridge University Engineering Department, England.

Englund, C. (2004). *Speech recognition in the JAS 39 Gripen aircraft- adaptation to speech at different G-loads*, Master thesis, Department of Speech, Music and Hearing, Royal Institute of Technology, Stockholm.

Faria, A. and Gelbart, D. (2005). Efficient Pitch-based Estimation of VTLN Warp Factors. *Proc. INTERSPEECH*. 213-216.

Fernández, S., Graves, A. and Schmidhuber, J. (2008). Phoneme recognition in TIMIT with BLSTM-CTC. *Technical Report No. IDSIA-04-08, IDSIA, Manno-Lugano, Switzerland*. April.

Flanagan, J.L., Allen, J.B. and Hasegawa-Johnson, M.A. (2008). *Speech Analysis Synthesis and Perception*. (3rd ed.) Verlag, Berlin: Springer.

Furui, S. (2008). *Speaker recognition. Scholarpedia, 3(4):3715*. Retrieved on January 1, 2013, from http://www.scholarpedia.org/article/Speaker_recognition.

Furui, S. (2009). Generalization Problem in ASR Acoustic Model Training And Adaptation. *Proc. IEEE ASRU 2009. Merano, Italy*. December. 1-10.

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S. and Dahlgren, N. L. (1993). TIMIT Acoustic-phonetic Continuous Speech Corpus. *Linguistic Data Consortium, Philadelphia*.

Giuliani, D., Gerosa, M. and Brugnara, F. (2006). Improved Automatic Speech Recognition Through Speaker Normalization. *Computer Speech & Language*, 20 (1). 107-123.

Haeb-Umbach, R. (1999). Investigations on Inter-Speaker Variability in the Feature Space. *Proceedings of the ICASSP'99*. 1, 397-400.

Hanson, H. M., and Chuang, E. S. (1999). Glottal Characteristics of Male Speakers: Acoustic Correlates and Comparison with Female Data. *The Journal of the Acoustical Society of America*. 106(2), 1064-1077.

Ho Ching-Hsiang. Speaker Modeling for Voice Conversion. Ph.D. Thesis. Brunel University; 2001.

Holmes, J. and Holmes, W. (2001). *Speech Synthesis and Recognition*. (2nd ed.) New Fetter Lane, London.: Taylor & Francis.

Huang, X., Acero, A., and Hon, H.-W. (2001). *Spoken Language Processing: a Guide to Theory, Algorithm, and System Development*. (1st ed.) Upper Saddle River, N.J.: Prentice-Hall.

Huang, Y., Benesty, J. and Chen, J. (2008). Dereveberation. *Springer Handbook of Speech Processing*, eds. Benesty, J., et al., (1st ed.) Berlin, Heidelberg: Springer-Verlag. 929-944.

Johnson, D. (2005). Modeling the Speech Signal. *Connexions*. April. Available at: http://cnx.org/content/m0049/2.29/.

Kahn, J., Audibert, N., Rossato, S. and Bonastre, J-F. (2010). Intra-speaker variability effects on Speaker Verification performance. *Proc of Odyssey 2010*. Brno, Czech Republic. 109-116.

Kinnunen, T. (2003). *Spectral Features for Automatic Text-Independent Speaker Recognition*. Doctor Philosophy, University of Joensuu, Finland.

Lee, K. F. and Hon, H. W. (1989). Speaker-Independent Phone Recognition Using Hidden Markov Models. *IEEE Trans. on Acoustics, Speech and Signal Processing*. 37(11), 1641-1648.

Lee, L. and Rose, R.C. (1996). Speaker Normalization Using Efficient Frequency Warping Procedures. *Proc. IEEE ICASSP 96.* 1, 353-356.

Lee, L. and Rose, R.C. (1998). A Frequency Warping Approach to Speaker Normalization. *IEEE transactions on speech and audio processing.* 6(1), 49-60.

Leggetter, C.J. and Woodland, P.C. (1995). Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models. *Computer Speech and Language*. 9(2), 171-185.

Liu, M., Zhou, X., Hasegawa-Johnson, M., Huang, T. S. and Zhang, Z. Y. (2007). Frequency Domain Correspondence for Speaker Normalization. *Proc. Interspeech*. 274-277.

Lopes, C., and Perdigão, F. (2011). Phone Recognition on the TIMIT Database. *Speech Technologies/Book*, 1, 285-302.

Mermelstein, P. (1976). Distance Measures for Speech Recognition, Psychological and Instrumental. *Pattern Recognition and Artificial Intelligence*, eds. Chen, C. H., New York, Academic. 374–388.

Mihelic, F. and Zibert, J. (Eds.) (2008). *Speech Recognition, Technologies and Applications.* (1st ed.) Vienna, Austria.: I-Tech.

O'Shaughnessy, D. (2008). Formant Estimation and Tracking. *Springer Handbook of Speech Processing*, eds. Benesty, J., et al., (1st ed.) Berlin, Heidelberg: Springer-Verlag. 213-228.

Panchapagesan, S. and Alwan, A. (2009). Frequency Warping for VTLN and Speaker Adaptation by Linear Transformation of Standard MFCC. *Computer Speech and Language*. 23, 42-64.

Peterson, G.E. and Barney, H.L. (1952). Control Methods Used In a Study of the Vowels. *Journal of the Acoustical Society of America*. 24, 175–184.

Pitz, M. and Ney, H. (2003). Vocal Tract Normalization as Linear Transformation of MFCC. *Eurospeech 2003*. 1445-1448.

Pye, D. and Woodland, P.C. (1997). Experiments in Speaker Normalization and Adaptation for Large Vocabulary Speech Recognition. *IEEE International Conference on Acoustics Speech and Signal Processing, ICASSP 97*. 1047-1050.

Rabiner, L. (1989). A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings of the IEEE*. 77(2), 257-286.

Rabiner, L. and Juang, B-H. (1993). *Fundamentals of Speech Recognition*. Prentice-Hall International.

Rabiner, L. and Juang, B. H. (2008). Historical Perspective of the Field of ASR/NLU. *Springer Handbook of Speech Processing*, eds. Benesty, J., et al., (1st ed.) Berlin, Heidelberg: Springer-Verlag. 521-538.

Rademacher, J., Wachter, M. and Mertins, A. (2006). Improved Warping-Invariant Features for Automatic Speech Recognition. *Proc. Int. Conf. Spoken Language Processing (Interspeech 2006 - ICSLP). Pittsburgh, PA, USA*. September, 1499-1502.

Reynolds, D.A., Campbell, W.M., Shen, W. and Singer, E. (2008). Automatic Language Recognition via Spectral and Token Based Approaches. *Springer Handbook of Speech Processing*, eds. Benesty, J., et al., (1st ed.) Berlin, Heidelberg: Springer-Verlag. 811-824.

Roger Jang, J.S. *Audio Signal Processing and Recognition*. Retrieved on January 2, 2013, from http://www.cs.nthu.edu.tw/~jang.

Roh, Y.W., Kim, J.H., Kim, D.J. and Hong, K.S. (2006). A Hybrid Warping Method Approach to Speaker Warping Adaptation. In Bloch, I., Petrosino, A. and Tettamanzi, A.G.B. (Eds.): *WILF 2005, LNAI 3849* (pp. 146-155). Berlin, Heidelberg: Springer-Verlag.

Rosenberg, A.E., Bimbot, F. and Parthasarathy, S. (2008). Overview of Speaker Recognition. *Springer Handbook of Speech Processing*, eds. Benesty, J., et al., (1st ed.) Berlin, Heidelberg: Springer-Verlag. 725-742.

Salam, M., Mohamad, D. and Salleh, S. (2011). Malay Isolated Speech Recognition Using Neural Network: A Work in Finding Number of Hidden Nodes and Learning Parameters. *The International Arab Journal of Information Technology*. 8(4), 364-371.

Schroeter, J. (2008). Principles of Speech Synthesis. *Springer Handbook of Speech Processing*, eds. Benesty, J., et al., (1st ed.) Berlin, Heidelberg: Springer-Verlag. 413-428.

Shen, W. and Reynolds, D. (2007). Improving Phonotactic Language Recognition with Acoustic Adaptation. *Proc. Interspeech*. August. 358-361.

Tsuge, S., Seida, K., Shishibori, M., Kita, K., Ren, F., Fukumi, M and Kuroiwa, S. (2007). Analysis of Variation on Intra-Speakers Speech Recognition Performances. *Proc of NLP-KE 2007*. Beijing, China. 387-392.

Umesh, S. (2011). Studies on Inter-speaker Variability in Speech and Its Application in Automatic Speech Recognition. *Indian Academy of Sciences, Sadhana*. 36(5), 853-883.

Wegmann, S., McAllaster, D., Orloff, J. and Peskin, B. (1996). Speaker Normalization on Conversational Telephone Speech. *Proc. IEEE ICASSP 96*. 1, 339-341.

Welling, L., Haeb-Umbach, R., Aubert, X. and Haberland, N. (1998). A Study on Speaker Normalization Using Vocal Tract Normalization and Speaker Adaptive Training. *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*. May. 797-800.

Yang, X., Millar, J.B. and Macleod, I. (1996). On The Sources of Inter- & Intra-Speaker In The Acoustic Dynamics of Speech. *Proc. ICSLP 96*. 3, 1792-1795.

Young, S. et al. (2006). The HTK Book. *Cambridge University Engineering Department*. (8th ed.)

Young, S. (2008). HMMs and Related Speech Recognition Technologies. *Springer Handbook of Speech Processing*, eds. Benesty, J., et al., (1st ed.) Berlin, Heidelberg: Springer-Verlag. 539-558.

Zhan, P. and Waibel, A. (1997). Vocal Tract Length Normalization for Large Vocabulary Continuous Speech Recognition. *CMU-CS-97-148, Carnegie Mellon University, Pittsburgh, PA.* May.