MALAY STATISTICAL PARAMETRIC SPEECH SYNTHESIS WITH INTELLIGIBILITY IMPROVEMENT USING ARTIFICIAL INTELLIGENCE

LAU CHEE YONG

A thesis submitted in fulfilment of the
requirements for the award of the degree of
Doctor of Philosophy (Biomedical Engineering)

Faculty of Biosciences and Medical Engineering
Universiti Teknologi Malaysia

JANUARY 2015

Dedicated to my mum and dad,

my brother and sisters,

and my beloved friends.

ACKNOWLEDGEMENT

First of all, I would like to thank my supervisor in this PhD study which is Dr Tan Tian Swee. He has given me a lot of chances and guidance in completing and fulfilling this study. He is the one who always provide me advices whenever I have problem in my research.

I also wish to deliver my heartiest gratitute to my supervisor and mentor, Profesor Simon King and Dr Oliver Watts during my attachment in project Simple4all at Centre of Speech Technology Research (CSTR) in Informatic Forum, University of Edinburgh, United Kingdom. Professor Simon King has guided me throughout the attachment in Edinburgh and Dr Oliver Watts has helped me in solving a lot of technical problem especially in Active Learning and and given me a lot of attentive care in my research.

Besides, I would like to thank my fabulous laboratory mate and they are Lum Kin Yun, Gan Hong Seng, Nizam Mazenan and Leong Kah Meng. They have given me a plenty of mental support and made me had a great time in my laboratory time.

Last but not least, I would like to thank my parents and sisters as they have provided me precious moral support to make me travel until this far. They are always the one that I want to make them proud. Thank you.

ABSTRACT

Speech synthesis is important nowadays and could be a great aid in various applications. So it is important to build a simple, reliable, light-weight, ease of use speech synthesizer. However, conventional speech synthesizers require tedious human efforts to prepare high quality recorded database, and the intelligibility of synthetic speech may decrease due to the appearance of polyphone (character with more than 1 pronunciation) because the speech synthesizer may not contain the definition of the polyphones. Moreover, the ready speech synthesizers in market are mostly built in Unit Selection method, which is large in database size and relying on Malay linguist knowledge. In this study, statistical parametric speech synthesis method has been adopted using lab speech and free speech data harvested online. The intelligibility improvement has been achieved using Active Learning and Feedforward Neural Network with Back-Propagation. The amount of training data used remained the same throughout this study. The result was evaluated using perception test. The listening test showed that the intelligibility of synthetic speech has been improved about 20%-30% using the artificial intelligence technique. Volunteers were invited to take part in Active Learning experiment. The result showed no controversy between the result done by volunteers and the correct answer. In conclusion, a light-weight Malay speech synthesizer has been created without relying on Malay linguist knowledge. Using free source as training data can ease the human effort in preparing training database and using artificial intelligence technique can improve the intelligibility of synthetic speech under the same amount of training data used.

ABSTRAK

Sintesis ucapan adalah penting pada hari ini dan boleh menjadi bantuan yang besar untuk pemulihan masalah menghasilkan ucapan. Jadi adalah penting untuk membina pensintesis yang mudah, boleh dipercayai dan mudah alih. Walau bagaimanapun, pensintesis ucapan konvensional memerlukan banyak usaha manusia untuk menyediakan data rakaman, dan kejelasan ucapan sintetik mungkin berkurangan akibat kemunculan *polyphone* (watak dengan lebih daripada 1 sebutan) dalam perkataan yang berbeza kerana pensintesis ucapan tersebut mungkin tidak mengandungi definisi maklumat *polyphone*. Selain itu, pensintesis ucapan yang terdapat dalam pasaran kebanyakannya dibina dengan kaedah Pemilihan Unit, menyebabkan saiz pangkalan data yang besar dan bergantung kepada pengetahuan ahli bahasa Melayu. Dalam kajian ini, statistik parametrik kaedah sintesis ucapan telah digunakan menggunakan sumber bebas yang boleh didapatkan daripada internet secara percuma. Peningkatan kejelasan telah dicapai dengan menggunakan beberapa teknik *Artificial Intelligence (AI)* seperti *Active Learning (AL)* dan *Feedforward Neural Network (FNN)* dengan *Back-Propagation (BP)*. Jumlah data latihan yang digunakan adalah tetap sama sepanjang kajian ini. Keputusan ini telah dibandingkan dengan data terlatih yang direkodkan. Ujian menunjukkan bahawa kejelasan ucapan sintetik telah bertambah kira-kira 20% - 30% menggunakan teknik AI tersebut. Sukarelawan-sukarelawan telah dijemput untuk mengambil bahagian dalam eksperimen pembelajaran aktif. Hasilnya menunjukkan tiada sebarang kontroversi antara penutur asli dan berbilang sukarelawan. Kesimpulannya, ucapan pensintesis Melayu yang ringan telah dicipta tanpa bergantung kepada pengetahuan ahli bahasa Melayu. Dengan menggunakan sumber bebas sebagai data latihan boleh mengurangkan usaha manusia dalam penyediaan data latihan dan menggunakan teknik AI boleh meningkatkan kejelasan ucapan sintetik di bawah jumlah data latihan yang sama.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AL	-	Active Learning
BIC	-	Bayesian Information Criterion
BP	-	Back-Propagation
CVC	-	Consonant Vowel Consonant
CV	-	Consonant Vowel
DRT	-	Diagnostic Rhyme Test
EM	-	Expectation-Maximization
FNN	-	Feedforward Neural Network
GMM	-	Gaussian Mixture model
HMM	-	Hidden Markov Model
HTK	-	Hidden Markov model Toolkit
LLR	-	Log Likelihood Ratio
MFCC	-	Mel-frequency Cepstral Coefficient
MOS	-	Mean Opinion Score
MRT	-	Modified Rhyme Test
PDF	-	Probability Distribution Function
QBB	-	Query-by-Bagging
QBC	-	Query-by-Committee
SM	-	Standard Malay
STRAIGHT	-	Speech Transformation and Representation using
	-	Adaptive Interpolation of weiGHTed spectrum
SUS	-	Semantically Unpredictable Sentences
VAD	-	Voice Activity Detection
WER	-	Word Error Rate

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

1.1 Background Study

Speech synthesis is a method of converting written text into spoken speech (Sproat *et al.*, 1995; Dutoit and Stylianou, 2003; Dutoit, 1997). This process is also known as Text-to-Speech (TTS) generation. It is a reversion of speech recognition (Rabiner, 1989) which recognizes speech and transcribes the speech into text. From time to time, the evolution of speech synthesis has made speech synthesizers robust and reliable in handling many applications such as telephony services, screen readers for the blind or visually impaired, navigation systems and many more (Lemmetty, 1999). For medical purposes, this technique could provide a substitute for mute people to communicate with other people. A famous example of a person with a speech disability is the theoretical physicist Stephen Hawking (Larsen, 2005). He is almost entirely paralyzed and uses synthetic speech to communicate with others. In order to build a high quality speech synthesizer, the development should take care of the following aspects:

1. **Naturalness** (Taylor, 2009). People are sensitive to speech, not only by the words spoken but how the person speaks. Mechanical or robotic synthetic voices are annoying and irritating after a long time listening to that type of voice. Therefore, one of the goals for a speech synthesizer is to generate natural sounding speech.
2. **Intelligibility** (Benoit *et al.*, 1996). The key significance of a speech synthesizer is to deliver messages. A good speech synthesizer can replace human efforts and take over many areas of speech. There is no point building a speech synthesizer if it produces speech that we cannot understand. Therefore, speech intelligibility is an important factor to be considered when making high quality speech synthesizers.

3. **Able to produce novel speech** (Taylor, 2009). Normally the quality of speech synthesizers depend on the condition of the training data. The way to design and produce a high quality training database is highly sophisticated. However, a good speech synthesizer should be able to speak any novel words beyond the training data. It is less practical if the speech synthesizer is only able to speak utterances within the training corpus. Moreover, the uttered novel words should also be natural and intelligible to listeners.

In short, a speech synthesis system should be efficient, be able to produce intelligible speech, and sound natural for novel words (Tabet and Boughazi, 2011).

With the improvement of computer technology nowadays, speech synthesis has evolved from knowledge-based into data-based (Black *et al.*, 2007). Speech synthesizer can be built from a sufficient amount of human speech data. One of the example of data-based speech synthesizers is Statistical Parametric Speech Synthesis. It is a data-based speech synthesis method and it has gained more and more attentions recently. It models the data of parametric representations of natural speech and generates similar sounding speech segments during synthesis. This is in contradiction to the Unit Selection method (Conkie, 1999) which keeps the speech data unmodified and generating synthetic voices using natural speech data. However, experiments have shown that the synthetic voice generated using the Statistical Parametric Speech Synthesis method is natural and intelligible. In the Blizzard Challenge 2005 (Bennett and Christina, 2005) and 2006 (Clark *et al.*, 2006), a common speech database was provided to participants to build synthetic voices. The results showed that the synthetic speech generated using the Statistical Parametric Speech Synthesis method was preferred due to its naturalness. The synthetic speech was intelligible and understandable to the listeners and it was proven using the Word Error Rate (WER) score (Zechner and Waibel, 2000). This result has shown that Statistical Parametric Speech Synthesis is capable to synthesize good quality speech.

Besides, Statistical Parametric Speech Synthesis also offers several advantages which increases its flexibility and extends speech technology:

1. Unit Selection chooses a finite unit from its database. It may face a problem of choosing inadequate examples. This can be viewed as a lack of database coverage. However, Statistical Parametric Speech Synthesis generates speech using statistical data. Therefore, it has better acoustic space coverage than the Unit Selection method and a wider range of units are available.

2. The Statistical Parametric Speech Synthesis method stores the statistical data of the acoustic model whereas the Unit Selection method stores real speech segments. Therefore, the Statistical Parametric Speech Synthesis method can achieve a smaller footprint than the Unit Selection method. For example, the footprint of voices of Nitech HMM-Based Speech Synthesis System in Blizzard Challenge 2005 is less than 2MB (Zen *et al.,* 2007).

3. The Statistical Parametric Speech Synthesis method is more robust than the Unit Selection method. This is because the real speech database of the Unit Selection method may suffer from noise and fluctuation disturbances due to the recording surroundings and the recording of a real human's speech may not practically cover all the phonetic possibilities. However, research has shown that Statistical Parametric Speech Synthesis method can resolve these problems (Yamagishi *et al.*, 2008).

4. The representation of speech in the Statistical Parametric Speech Synthesis method is statistical data of the spectrum, duration and excitation. Therefore these parameters can be separately modified and monitored.

5. The voice characteristics, emotions and speaking styles of synthetic speech can be transformed into Statistical Parametric Speech Synthesis. This is the key flexibility of this method. The transformation can be done by utilizing adaptation (Masuko *et al.*, 1997), eigenvoice (Kuhn *et al.*, 2000), interpolation (Yoshimura *et al.*, 1997) and multiple regression (Miyanaga *et al.*, 2004).

6. Statistical Parametric Speech Synthesis uses statistical principles that are defined in mathematical frameworks. The tuning parameters are lesser than the Unit Selection method which requires manual tuning and settings for various control.

1.2 Problem Statement

In order to build a reliable speech synthesizer especially targeted to Malaysian, the following problems should be considered.

1. The available speech synthesizers in the market are mostly in English. There are not many Malay speech synthesizers ready for Malaysian. The available Malay speech synthesizers are larger in file size (>25MB) (Tan, 2009; Lim, 2013) which is not peactical to be used in light-scale embedded system (Kim *et al.*, 2006).

2. The process of preparing training data in building a speech synthesizer is

sophisticated and cumbersome. It involves gathering words from sources, constructing suitable scripts which includes all the phonemes in the Malay language, the recording of scripts and the recording of sessions should be conducted in a high quality recording studio. It is expensive to construct a real speech database over a long period of time.

3. Conventionally, to build a speech synthesizer requires the knowledge of language expert to precisely draw the boundary of every phoneme because phonemes are the basic synthesis unit for a speech synthesizer. However, consulting a language expert adds extra workload and it is expensive to do so.
4. The intelligibility of synthetic speech is the main concern in every speech synthesizer. Most of the speech synthesizers might face the problem of low intelligibility especially in synthesizing words which are not found in database.

The aim of this research is to solve the aforementioned problems and create a reliable Malay speech synthesizer. Several techniques have been applied to resolve the problems and it will be explained in Chapter 3 and 4.

1.3 Objectives

This study is aiming to solve the related problems in building a speech synthesizer. Therefore, the objectives are:

1. To build a Malay speech synthesizer with a low footprint (data size).
2. To alleviate the problem of preparing database in Statistical Parametric Speech Synthesis System by including free data harvested online.
3. To exclude the dependency of linguist in building speech synthesizer.
4. To improve the synthetic speech intelligibility using Active Learning (AL) and Feedforward Neural Network (FNN) with Back-Propagation (BP) while the same amount of training data was used.

The block diagram of this study is shown in Figure 1.1.

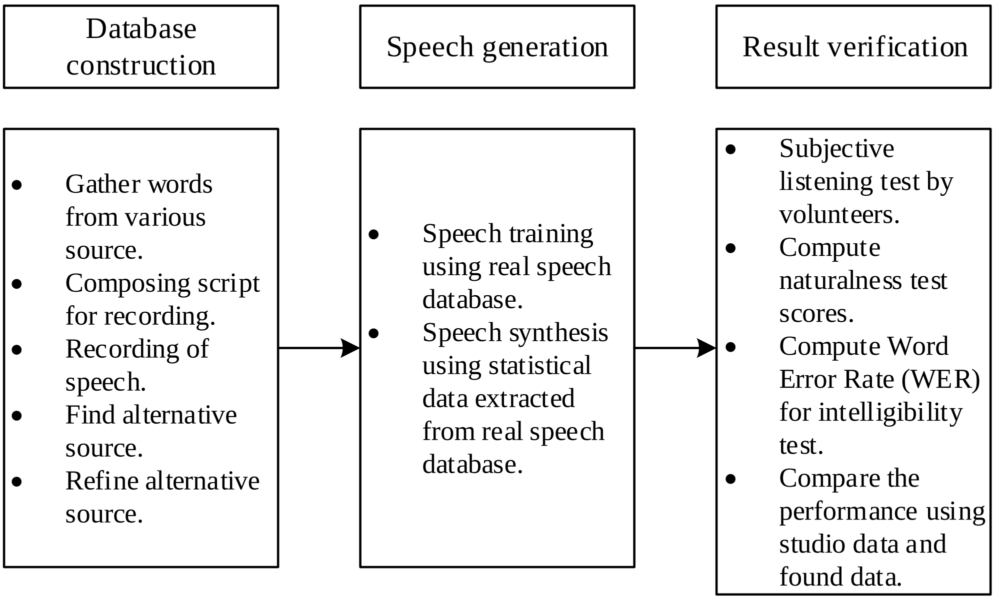

Figure 1.1: Mapping of procedures in this study

1.4 Scope of the Study

This study follows several scopes and they are:

1. The Malay speaking style used in this study is Standard Malay (SM) (Seman and Jusoff, 2008) which is the usual Malay speaking style spoken by Malaysians. No other accents like Kelantan Malay, Ulu Muar Malay and so on were used throughout this study. The reason Standard Malay is going to be used is to make the speech synthesizer suitable to be used in almost every area of speech, for example, speech rehabilitation, education, or any speech emitting devices like computer and smart phones. Standard Malay is also easily understandable by almost every Malaysian.
2. The invited speaker for the recording of the database is a Malay adult native speaker. This is to ensure the database contained the correct Malay pronunciations and that the voice is mature. Correct pronunciation can improve the synthetic speech intelligibility, therefore the synthetic speech would be easily understood.

3. The free training data harvested online is clear in pronunciation, low in background noise and no overlapped with any other voices or music.
4. The synthetic speech synthesized in this study would be in normal reading style. No any other voice tone would be incurred like happy, sad or angry emotions.

1.5 Thesis Organization

Chapter 1 briefly introduced the background of the study. It gave a basic overview on speech synthesis technologies and briefly talked about the state-of-the-art Statistical Parametric Speech Synthesis. It also presented the problem statements, objective and the scope of this study.

Chapter 2 provided a literature review of this study. It included a basic overview on the Malay language. The history of the speech synthesizer was introduced in this chapter in a timeline fashion. Comparisons between state of the art speech synthesizers were also discussed. A decision was made on which type of speech synthesizer was used in this research and the reasons. The technical review on statistical parametric speech synthesizer which was used in this thesis was presented from the basic model applied in this method until how it produces synthetic speech sounds. A brief discussion on how speech synthesizer can help people was presented within this chapter. The evaluation methods available were overviewed and only one evaluation approach was selected based on the suitability and effectiveness. How the result was statistically compared was also introduced in this chapter.

Chapter 3 is the Methodology used in this study. It involved how the training database was constructed, how the free source was obtained online, how the modifications were done to the found data, how the Artificial Intelligence techniques (Feedforward Neural Network with Back Propagation and Active Learning) was applied, how the front end processing was conducted, how the speech training and speech synthesis works, and how the listening test was carried out to test the quality of synthetic speech.

Chapter 4 is the Result and Discussion section. It showed the accuracy of classifiers trained with Feedforward Neural Network with Back Propagation and Active Learning. It also presented the listening test result of both Naturalness Test and Intelligibility Test in the experiments involving Found Data, Feedforward

Neural Network with Back Propagation and Active Learning. The total footprint or total file size of the speech synthesizers was displayed in detail. The significant difference test result was also calculated and compared and this chapter was concluded with discussions for all the experiments and benchmark with other Malay speech synthesizers.

Chapter 5 outlined the conclusion and explained the contributions of this study. The future work was also presented in the end of this chapter.

REFERENCES

Abe, N. and Mamitsuka, H. (1998). Query learning strategies using boosting and bagging. In *Proc. International Conference on Machine Learning*. 1–9.

Abraham, A. (2005). *Artificial Neural Networks*. Wiley Online Library.

Alias, F., Formiga, L. and Llora, X. (2011). Efficient and reliable perceptual weight tuning for unit-selection text-to-speech synthesis based on active interactive genetic algorithms: A proof-of-concept. *Speech Communication*. 53(5), 786–800.

Allen, J., Hunnicutt, M. S., Klatt, D. H., Armstrong, R. C. and Pisoni, D. B. (1987). *From text to speech: the MITalk system*. Cambridge University Press.

Alman, D. H. and Ningfang, L. (2002). Overtraining in back-propagation neural networks: A CRT color calibration example. *Color Research & Application*. 27(2), 122–125.

Ballard, K. J. (2001). Response generalization in Apraxia of speech treatments: taking another look. *Journal of Communication Disorders*. 34(1-2), 3–20.

Barra-Chicote, R., Yamagishi, J., King, S., Montero, J. M. and Macias-Guarasa, J. (2010). Analysis of statistical parametric and unit selection speech synthesis systems applied to emotional speech. *Speech Communication*. 52(5), 394–404.

Bebis, G. and Georgiopoulos, M. (1994). Feed-forward neural networks. *Potentials, IEEE*. 13(4), 27–31.

Bennett, C. L. (2005). Large scale evaluation of corpus-based synthesizers: results and lessons from the blizzard challenge 2005. In *Interspeech*. 105–108.

Benoit, C., Grice, M. and Hazan, V. (1996). The SUS test: A method for the assessment of text-to-speech synthesis intelligibility using Semantically Unpredictable Sentences. *Speech Communication*. 18(4), 381–392.

Black, A. and Campbell, N. (1995). Optimising selection of units from speech databases for concatenative synthesis. In *European Conference on Speech*

Communication and Technology (EUROSPEECH). 581–584.

Black, A. W., Zen, H. and Tokuda, K. (2007). Statistical parametric speech synthesis. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2007*, vol. 4. IV–1229.

Borman, S. (2004). *The expectation maximization algorithm: A short tutorial.* Technical report.

Bray, J. H. and Maxwell, S. E. (1985). *Multivariate Analysis of Variance*. 54. Sage.

Breiman, L. (1996). Bagging predictor. *Machine Learning*. 24, 123–140.

Breuer, S. and Abresch, J. (2004). Phoxsy: multi-phone segments for unit selection speech synthesis. In *Interspeech*.

Cannito, M. P., Chorna, L. B., Kahane, J. C. and Dworkin, J. P. (2014). Influence of Consonant Voicing Characteristics on Sentence Production in Abductor Versus Adductor Spasmodic Dysphonia. *Journal of Voice*. 28(3), 394–e13.

Chapelle, O., Scholkopf, B. and Zien, A. (2006). *Semi-supervised learning*. vol. 2. MIT press Cambridge.

Clark, R., Richmond, K., Strom, V. and King, S. (2006). Multisyn Voices for the Blizzard Challenge 2006. In *Proc. Blizzard Challenge Workshop (Interspeech Satellite)*. Pittsburgh, USA.

Clark, R. A. J., Richmond, K. and King, S. (2007). Multisyn: Open-domain unit selection for the Festival speech synthesis system. *Speech Communication*. 49(4), 317–330.

Cohn, D. A., Ghahramani, Z. and Jordan, M. I. (1996). Active learning with statistical models. *Journal of Artificial Intelligence Research*. 4, 129–145.

Collins, J. T. (1998). *Malay, world language: a short history*. Dewan Bahasa dan Pustaka.

Conkie, A. (1999). Robust unit selection system for speech synthesis. In *137th Meeting of the Acoustical Society of America*. 978.

Dayhoff, J. E. and DeLeo, J. M. (2001). Artificial Neural Networks. *Cancer*. 91(S8), 1615–1635.

Deller, J. R., Proakis, J. G. and Hansen, J. H. L. (1993). *Discrete-Time Processing of Speech Signals*. Macmillan, New York.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical*

Society. Series B (Methodological). 39(1), 1–38.

Demuth, H., Beale, M. and Hagan, M. (2008). *Neural Network Toolbox User's Guide*. Math Works.

Donovan, R. E. (1996). *Trainable Speech Synthesis*. Ph.D. Thesis. Cambridge University Engineering Department.

Dutoit, T. (1997). *An Introduction to Text-to-Speech Synthesis*. vol. 3. Springer.

Dutoit, T. and Stylianou, Y. (2003). *Handbook of Computational Linguistics*. Oxford University Press.

Dwyer, K. and Kondrak, G. (2009). Reducing the Annotation Effort for Letter-to-phoneme Conversion. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1*. ACL '09. Stroudsburg, PA, USA: Association for Computational Linguistics, 127–135.

Efron, B. (1969). Student's t-test under symmetry conditions. *Journal of the American Statistical Association*. 64(328), 1278–1302.

Ekpenyong, M., Urua, E.-A., Watts, O., King, S. and Yamagishi, J. (2014). Statistical parametric speech synthesis for Ibibio. *Speech Communication*. 56, 243–251.

Enderby, P. (2013). *Chapter 22 - Disorders of communication: Dysarthria*, Elsevier, vol. 110. 273–281.

Forney, J., G.D. (1973). The viterbi algorithm. *Proceedings of the IEEE*. 61(3), 268–278.

Freund, Y. H., Seung, S., Shamir, E. and Tishby, N. (1997). Selective sampling using the query by committee algorithm. *Machine Learning*. 28, 133–168.

Furui, S. (1974). An analysis of long-term variation of feature parameters of speech and its application to talker recognition. *Electronics and Communications in Japan*. 57(12), 34–42.

Gardiner, C. W. (1985). *Handbook of stochastic methods*. vol. 3. Springer, Berlin.

Gimenez de los Galanes, F. M., Savoji, M. H. and Pardo, J. M. (1994). New algorithm for spectral smoothing and envelope modification for LP-PSOLA synthesis. IEEE International Conference on Acoustic, Speech, and Signal Processing, 573–576.

Gold, B., Morgan, N. and Ellis, D. (2000). *Speech and Audio Signal Processing: Processing and Perception of Speech and Music, 2nd Edition*. US: John Wiley and Sons, Inc.

Hansen, J. H. and Clements, M. A. (1991). Constrained iterative speech enhancement with application to speech recognition. *IEEE Transactions on Signal processing*. 39(4), 795–805.

Hecht-Nielsen, R. (1989). Theory of the backpropagation neural network. In *International Joint Conference on Neural Networks (IJCNN), 1989*. 593–605.

Heiga, Z., Tomoki, T., Nakamura, M. and Tokuda, K. (2007). Details of the Nitech HMM-based speech synthesis system for the Blizzard Challenge 2005. *IEICE Transactions on Information and Systems*. 90(1), 325–333.

Hill, D. and Preucil, M. (1973). Control of an analog speech synthesieer by a time-shared digital computer. In *Proc. 7th AICA Congress, Prague*, vol. 1.

House, A. S., Williams, C., Hecker, M. H. and Kryter, K. D. (2005). Psychoacoustic speech tests: A modified rhyme test. *Journal of the Acoustical Society of America*. 35(11), 1899–1899.

Hugh-Munier, C. M., Scherer, K. R., Lehmann, W. and Scherer, U. (1997). Coping strategies, personality, and voice quality in patients with vocal fold nodules and polyps. *Journal of Voice*. 11(4), 452–461.

Hunt, A. J. and Black, A. W. (1996). Unit selection in a concatenative speech synthesis system using a large speech database. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 1996*, vol. 1. 373–376.

Iida, A. (2002). *A study on Corpus-based Speech Synthesis with Emotion*. Ph.D. Thesis. University of Keio, Japan.

Imai, S., Sumita, K. and Furuichi, C. (1983). Mel log spectrum approximation (MLSA) filter for speech synthesis. *Electronic and Communication in Japan*, 10–18.

Iversen, G. R. and Norpoth, H. (1987). *Analysis of variance*. 1. Sage.

Juang, B. H. and Rabiner, L. R. (1991). Hidden Markov Model for Speech Recognition. *Technometrics*. 33, 251–272.

Karaiskos, V., King, S., Clark, R. A. J. and Mayo, C. (2008). The Blizzard Challenge 2008. In *Proc. Blizzard Challenge Workshop*. September. Brisbane, Australia.

Kawahara, H. (2006). STRAIGHT, exploitation of the other aspect of VOCODER: Perceptually isomorphic decomposition of speech sounds. *Acoustical Science and Technology*. 27(6), 349.

Kawahara, H., Morise, M., Takahashi, T., Nisimura, R., Banno, H. and Irino, T. (2008). A unified approach for F0 extraction and aperiodicity estimation based on a temporally stable power spectral representation. In *ISCA Tutorial and Research Workshop (ITRW) on Speech Analysis and Processing for Knowledge Discovery. Aalborg*. 4–6.

Kempler, D. and Lancker, D. V. (2002). Effect of Speech Task on Intelligibility in Dysarthria: A Case Study of Parkinson's Disease. *Brain and Language*. 80(3), 449–464.

Klatt, D. H. (1987). Review of Text-to-Speech Conversion for English. *Journal of the Acoustical Society of America*. 82, 737–793.

Kuhn, R., Junqua, J.-C., Nguyen, P. and Niedzielski, N. (2000). Rapid speaker adaptation in eigenvoice space. *IEEE Transactions on Speech and Audio Processing*. 8(6), 695–707.

Kumar, G. H., Ravishankar, M., Nagabushan, P. and Anami, B. S. (2006). Hidden Markov model-based approach for generation of Pitman shorthand language symbols for consonants and vowels from spoken English. *Sadhana*. 31(3), 277–290.

Laganaro, M., Croisier, M., Bagou, O. and Assal, F. (2012). Progressive apraxia of speech as a window into the study of speech planning processes. *Cortex*. 48(8), 963–971.

Larsen, K. (2005). *Stephen Hawking: a biography*. Greenwood Publishing Group.

Lemmetty, S. (1999). *Review of speech synthesis technology*. Ph.D. Thesis. Helsinki University of Technology.

Lim, Y. C. (2013). *Optimization of Unit Selection Algorithm for Malay Speech Synthesis System*. Ph.D. Thesis.

Lim, Y. C., Tan, T. S., Shaikh Salleh, S. H. and Ling, D. K. (2012). Application of Genetic Algorithm in unit selection for Malay speech synthesis system. *Expert Systems with Applications*. 39(5), 5376–5383.

Lorenzo-Trueba, J., Watts, O., Barra-Chicote, R., Yamagishi, J., King, S. and Montero, J. M. (2012). Simple4all proposals for the albayzin evaluations in speech synthesis. In *Proc. Iberspeech 2012*.

Mamiya, Y., Yamagishi, J., Watts, O., Clark, R. A., King, S. and Stan, A. (2013). Lightly Supervised GMM VAD to Use Audiobook for Speech Synthesizer. In *Proc. International Conference on Acoustics, Speech, and Signal Processing*

(ICASSP).

Mantel, N. (1963). Chi-square tests with one degree of freedom; extensions of the Mantel-Haenszel procedure. *Journal of the American Statistical Association*. 58(303), 690–700.

Marreiros, A. C., Daunizeau, J., Kiebel, S. J. and Friston, K. J. (2008). Population dynamics: Variance and the sigmoid activation function. *NeuroImage*. 42(1), 147 – 157.

Masuko, T., Tokuda, K., Kobayashi, T. and Imai, S. (1996). Speech synthesis using HMMs with dynamic features. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 1996*, vol. 1. 389–392.

Masuko, T., Tokuda, K., Kobayashi, T. and Imai, S. (1997). Voice characteristics conversion for HMM-based speech synthesis system. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 1997*, vol. 3. 1611–1614.

Meister, S. (1978). *The diagnostic rhyme test (DRT): An Air Force implementation*. Technical report. DTIC Document.

Meron, Y. and Hirose, K. (1999). Efficient weight training for selection based synthesis. In *European Conference on Speech Communication and Technology (EUROSPEECH)*.

Miyanaga, K., Masuko, T. and Kobayashi, T. (2004). A style control technique for HMM-based speech synthesis. In *Proc. International Conference on Spoken Language Processing (ICSLP)*, vol. 4.

Moulines, E. and Charpentier, F. (1990). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*. 9, 453–467.

Peach, R. K. and Tonkovich, J. D. (2004). Phonemic characteristics of apraxia of speech resulting from subcortical hemorrhage. *Journal of Communication Disorders*. 37(1), 77–90.

Pearson, V. A. H. (1995). Speech and language therapy: Is it effective? *Public Health*. 109, 143–153.

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., Jackson, L. B., Schafer, R. W. and Coker, C. H. (1971). A Hardware Realization of a Digital Formant Speech Synthesizer. In *IEEE Transaction on Communication Technology*, vol. 19. 1016–1020.

Rabiner, L. R. and Juang, B.-H. (1993). *Fundamentals of speech recognition*. vol. 14. PTR Prentice Hall Englewood Cliffs.

Riedi, M. P. (1998). *Controlling Segmental Duration in Speech Synthesis System*. Ph.D. Thesis. Swiss Federal Institude of Technology Zurich.

Rosner, B., Glynn, R. J. and Lee, M.-L. T. (2006). The Wilcoxon signed rank test for paired comparisons of clustered data. *Biometrics*. 62(1), 185–192.

Rutten, P., Coorman, G., Fackrell, J. and Van Coile, B. (2000). Issues in Corpus Based Speech Synthesis. IEEE Seminar on State of the art in speech synthesis, 1–7.

Saito, S. and Kazuo (1985). *Fundamentals of Speech Signal Processing*. Academic Press (New York).

Sang-Jin, K., Jong-Jin, K. and Minsoo, H. (2006). HMM-based Korean speech synthesis system for hand-held devices. *Consumer Electronics, IEEE Transactions on*. 52(4), 1384–1390.

Santen, J. P. H. v., Olive, J. P., Sproat, R. and Hirschberg, J. (1997). *Progress in Speech Synthesis*. Springer-Verlag New York, Inc. USA.

Schroeder, M. R. (1993). A brief history of synthetic speech. *Speech Communication*. 13, 231–237.

Seman, N. and Jusoff, K. (2008). Acoustic pronunciation variations modeling for standard Malay speech recognition. *Computer and Information Science*. 1(4), P112.

Sinclair, M. and King, S. (2013). Where are the challenges in speaker diarization? In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2013*. May. Vancouver, British Columbia, USA.

Spencer, K. A. and Rogers, M. A. (2005). Speech motor programming in hypokinetic and ataxic dysarthria. *Brain and Language*. 94(3), 347–366.

Sproat, R. W. and Olive, J. P. (1995). *Text-to-Speech Synthesis*. vol. 74. Wiley Online Library.

Stan, A., Bell, P. and King, S. (2012). A grapheme-based method for automatic alignment of speech and text data. In *SLT*. 286–290.

Stan, A., Watts, O., Mamiya, Y., Giurgiu, M., Clark, R., Yamagishi, J. and King, S. (2013). TUNDRA: A Multilingual Corpus of Found Data for TTS Research Created with Light Supervision. In *Proc. Interspeech*. August. Lyon, France.

Stan, A., Yamagishi, J., King, S. and Aylett, M. (2011). The Romanian speech synthesis (RSS) corpus: Building a high quality HMM-based speech synthesis system using a high sampling rate. *Speech Communication*. 53(3), 442–450.

Swee, T. T. and Salleh, S. H. S. (2008). Corpus-based Malay text-to-speech synthesis system. In *14th Asia-Pacific Conference on Communications (APCC), 2008*. 1–5.

Tabachnick, B. G., Fidell, L. S. *et al.* (2001). *Using multivariate statistics*. Allyn and Bacon Boston.

Tabet, Y. and Boughazi, M. (2011). Speech synthesis techniques. A survey. In *7th International Workshop on Systems, Signal Processing and their Applications (WOSSPA), 2011*. 67–70.

Tadmor, U. (2007). Grammatical borrowing in Indonesian. *Empirical Approaches to Language Typology*. 38, 301.

Tan, T. S. (2009). *Corpus-based Malay Text to Speech System*. Ph.D. Thesis.

Taylor, P. (2009). *Text-to-Speech Synthesis*. Cambridge University Press.

Taylor, P., Black, A. W. and Caley, R. (2002). *The Festival Speech Synthesis System: System Documentation Edition 1.4*. Technical report. Human Communication Research Centre.

Teoh, B. S. (1994). *The Sound System of Malay Revisited*. Dewan Bahasa dan Pustaka, Ministry of Education, Malaysia.

Thierry, D. (1993). *High Quality Text-to-Speech Synthesis of the French Language*. Ph.D. Thesis. Faculte Polytechnique de Mons.

Toda, T., Kawai, H. and Tsuzaki, M. (2004). Optimizing sub-cost functions for segment selection based on perceptual evaluations in concatenative speech synthesis. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2004*, vol. 1. I–657.

Tokuda, K., Heiga, Z. and Black, A. W. (2002a). An HMM-based speech synthesis system applied to English. In *IEEE Workshop on Speech Synthesis*. 227–230.

Tokuda, K., Masuko, T., Miyazaki, N. and Kobayashi, T. (2002b). Multispace probability distribution HMM. *IEICE Trans. Inform. Systems*, 455–464.

Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T. and Kitamura, T. (2000). Speech parameter generation algorithms for HMM-based speech synthesis. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2000*, vol. 3. 1315–1318.

Triola, M. F. (2010). Bayes' Theorem. *Elementary Statistics*. 11.

Vergin, R., O'shaughnessy, D. and Farhat, A. (1999). Generalized mel frequency cepstral coefficients for large-vocabulary speaker-independent continuous-speech recognition. *IEEE Transactions on Speech and Audio Processing*. 7(5), 525–532.

Viswanathan, M. and Viswanathan, M. (2005). Measuring speech quality for text-to-speech systems: development and assessment of a modified mean opinion score (MOS) scale. *Computer Speech & Language*. 19(1), 55–83.

Watts, O. (2012). *Unsupervised Learning for Text-to-Speech Synthesis*. Ph.D. Thesis. University of Edinburgh.

Watts, O., Stan, A., Clark, R., Mamiya, Y., Giurgiu, M., Yamagishi, J. and King, S. (2013). Unsupervised and lightly-supervised learning for rapid construction of TTS systems in multiple languages from 'found' data: evaluation and analysis. In *8th ISCA Workshop on Speech Synthesis*. August. Barcelona, Spain, 121–126.

Wildt, A. R. and Ahtola, O. (1978). *Analysis of covariance*. Sage Publications Beverly Hills, CA.

Wouters, J. (2001). *Analysis and Synthesis of Degree of Articulation*. Ph.D. Thesis. Katholieke Universiteit Leuven (KUL), Belgium.

Yamagishi, J. (2006). *An introduction to hmm-based speech synthesis*. Technical report. Technical report, Tokyo Institute of Technology.

Yamagishi, J., Ling, Z. and King, S. (2008). Robustness of HMM-based Speech Synthesis. In *Proc. Interspeech 2008*. September. Brisbane, Australia, 581–584.

Yap, M. J., Liow, S. J. R., Jalil, S. B. and Faizal, S. S. B. (2010). The Malay Lexicon Project: A database of lexical statistics for 9,592 words. *Behavior research methods*. 42(4), 992–1003.

Yelken, K., Gultekin, E., Guven, M., Eyibilen, A. and Aladag, I. (2010). Impairment of Voice Quality in Paradoxical Vocal Fold Motion Dysfunction. *Journal of Voice*. 24(6), 724–727.

Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T. and Kitamura, T. (1997). Speaker interpolation in HMM-based speech synthesis system. In *European Conference on Speech Communication and Technology (EUROSPEECH)*.

Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T. and Kitamura, T. (1998). Duration Modeling for HMM-based Speech Synthesis. *In: Proc. International*

Conference on Spoken Language Processing (ICSLP).

Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T. and Kitamura, T. (1999). Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis. In *Proc. European Conference on Speech Communication and Technology (EUROSPEECH)*. 2347–2350.

Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X. A., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V. and Woodland, P. (2009). *The HTK Book Version 3.4*.

Yu, X., Efe, M. O. and Kaynak, O. (2002). A general backpropagation algorithm for feedforward neural networks learning. *IEEE Transactions on Neural Networks*. 13(1), 251–254.

Zechner, K. and Waibel, A. (2000). Minimizing word error rate in textual summaries of spoken language. In *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*. Association for Computational Linguistics, 186–193.

Zen, H., Nose, T., Yamagishi, J., Sako, S., Masuko, T., Black, A. and Tokuda, K. (2007). The HMM-based speech synthesis system (HTS) version 2.0. In *Proceedings of the 6th International Speech Communication Association Workshop on Speech Synthesis*. 294–299.

Zen, H., Tokuda, K. and Black, A. W. (2009). Statistical parametric speech synthesis. *Speech Communication*. 51(11), 1039–1064.

Zhao, X.-M. (2013). Bayesian Information Criterion (BIC). *Encyclopedia of Systems Biology*, 73–73.