

# **Automatic Speaker Classification Based on Voice Characteristics**

A thesis submitted for the degree  
of Master of Information Sciences (Research) of  
the University of Canberra

Phuoc Thanh Nguyen

December 2010

# Summary of Thesis

Gender, age, accent and emotion are some of speaker characteristics being investigated in voice-based speaker classification systems. Classifying speaker characteristics is an important task in the fields of Dialog, Speech Synthesis, Forensics, Language Learning, Assessment, and Speaker Recognition.

It is well known that reducing classification error rate has been a challenge in those research fields. This research thesis investigates new methods for speech feature extraction and classification to meet this challenge. Extracted speech features range from traditional features in speech recognition such as mel-frequency cepstral coefficients (MFCCs) to recently developed prosodic and voice quality features in speaker classification such as pitch, shimmer and jitter. Feature selection was then performed to find a more suitable feature set for building speaker models. For classification methods, feature weighting vector quantisation, Gaussian mixture models (GMMs), Support Vector Machine (SVM) and Fuzzy Support Vector Machine (FSVM) are investigated. Those new feature extraction and classification methods are then applied to gender, age, accent and emotion classification. Four well-known data sets including Australian National Database of Spoken Language (ANDOSL), aGender, EBO-DB, and FAU AIBO are used to evaluate those methods.

The contributions of this thesis to classification of speaker characteristics include:

1. The use of different speech features. Up to 1582 features and transliteration have been investigated.
2. Application of new feature selection method. Correlation based feature subset selection with SFSS was employed to eliminate redundant features because of large databases.
3. The use of fuzzy SVM (FSVM) as a new speaker classification method. FSVM assigns

a fuzzy membership value as a weight to each training data point to allow the decision boundary to move to overlapping regions to reduce empirical errors.

4. A detailed comparison of speaker classification performance for GMMs, SVM and FSVM.
5. A depth investigation on the relevance of feature type for classification of age and gender. Extensive experiments are performed to determine which features in the speech signal are suited to representation of age and gender in human speech.
6. Classification of age, gender, accent, and emotion characteristics is performed on four well-known data sets including ANDOSL, aGender, EBO-DB and FAU AIBO.

## Certificate of Authorship of Thesis

Except where clearly acknowledged in footnotes, quotations and the bibliography, I certify that I am the sole author of the thesis submitted today entitled —

### **Automatic Speaker Classification Based on Voice Characteristics**

I further certify that to the best of my knowledge the thesis contains no material previously published or written by another person except where due reference is made in the text of the thesis.

The material in the thesis has not been the basis of an award of any other degree or diploma except where due reference is made in the text of the thesis.

The thesis complies with University requirements for a thesis as set out in *Gold Book Part 7: Examination of Higher Degree by Research Theses Policy, Schedule Two (S2)*.

Refer to <http://www.canberra.edu.au/research-students/goldbook>

Signature of Candidate

Signature of chair of the supervisory panel

Date

# Acknowledgements

First and foremost, I would like to thank my supervisor, A/Prof. Dat Tran, for his enormous support during my study at the University of Canberra. I am also thankful for his valuable guidance both in research and life, his encouragement and attention to important research milestones and events, his very quick response to my questions, and his patience to help me enhance the thesis.

I would also like to thank my co-supervisor, A/Prof. Xu Huang, for his encouragement, advice, support and suggestions on research plans. I am also thankful for his patience to revised my thesis and careful feedbacks.

I would also like to thank the Faculty of Information Sciences and Engineering for supporting conference travels and maintaining the excellent computing facilities which were crucial for carrying out my research. Thanks to staff members as well as research students for discussions and seminars. A grateful thanks to Prof. John Campbell for his Research Proposal and Research Methodologies courses. A warm thanks to Mr. Hanh Huynh for his interesting discussions about life and encouragement. A special thanks to Trung Le for his valuable discussions.

More importantly, I would like to thank the HCMC University of Pedagogy, Viet Nam for providing me the scholarship which enabled me to undertake this research at the University of Canberra. I would like to express my gratitude to all my lecturers and colleagues at the Faculty of Mathematics and Informatics, HCMC University of Pedagogy. I wish to express my warm and sincere thanks to Dr. Nguyen Thai Son and Msc. Ly Anh Tuan, Faculty of Mathematics and Informatics, HCMC University of Pedagogy for their important guidance, support and encouragement during my first steps in the Faculty.

I devote my deepest gratitude to my parents for their unlimited love and support. They have encouraged me throughout the years of my study. The most special thanks belong to my wife Huyen, for her understanding about my leaving during all these years of my absence, her selfless love and support all along and encouragement.

# Contents

<b>Summary of Thesis</b>	<b>ii</b>
<b>Acknowledgements</b>	<b>v</b>
<b>Abbreviation</b>	<b>xiii</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Speaker Characteristics and Their Applications . . . . .	1
1.2 Gender, Age, Accent and Emotion Classification . . . . .	2
1.3 Research Problems . . . . .	4
1.4 Contributions of the Thesis . . . . .	6
1.5 Organisation of the Thesis . . . . .	7
<b>2 Literature Review</b>	<b>9</b>
2.1 The Speaker Classification System . . . . .	9
2.2 Sound Generation and Speech Signal . . . . .	11
2.3 Feature Extraction . . . . .	14
2.3.1 Spectral features . . . . .	14
Linear Prediction Analysis . . . . .	14
Formants . . . . .	15
Line Spectrum Pair . . . . .	15
Mel-Frequency Cepstral Coefficients . . . . .	15
2.3.2 Prosodic Features . . . . .	16
Pitch . . . . .	16

---

Energy . . . . .	17
Duration . . . . .	17
Zero Crossing Measure . . . . .	18
Probability of Voicing . . . . .	18
2.3.3 Voice Quality Features . . . . .	18
Jitter and Shimmer . . . . .	19
Harmonics-to-Noise Ratio . . . . .	20
2.3.4 Delta and Acceleration Coefficients . . . . .	20
2.3.5 Static Features . . . . .	20
2.3.6 Discussion . . . . .	20
2.4 Feature Selection . . . . .	22
2.5 Classification Methods . . . . .	24
2.5.1 Gaussian Mixture Models . . . . .	24
2.5.2 Support Vector Machine . . . . .	26
Binary Case . . . . .	26
Multi-class Support Vector Machine . . . . .	28
2.5.3 Discussion . . . . .	28
<b>3 Proposed Methods</b>	<b>30</b>
3.1 Fuzzy Support Vector Machine . . . . .	30
3.1.1 Calculating Fuzzy Memberships . . . . .	31
3.1.2 Fuzzy Clustering Membership . . . . .	31
3.1.3 The Role of Fuzzy Memberships . . . . .	32
3.2 Speaker Classification using Frame-level Features . . . . .	32
3.3 Speaker Classification using Static Features . . . . .	34
3.4 Feature Type Relevance in Age and Gender Classification . . . . .	35
<b>4 Experimental Results</b>	<b>37</b>
4.1 Data Sets . . . . .	37
4.1.1 ANDOSL . . . . .	38
4.1.2 aGender . . . . .	38

---

4.1.3	EMO-DB . . . . .	39
4.1.4	AIBO . . . . .	40
4.2	Accent Classification . . . . .	41
4.2.1	Parameter Settings for GMMs . . . . .	41
4.2.2	Parameter Settings for SVM . . . . .	43
4.2.3	Accent Classification Results Versus Age . . . . .	43
4.2.4	Accent Classification Versus Age and Gender . . . . .	45
4.3	Age, Gender and Emotion Classification Using Static Features . . . . .	46
4.4	Feature Type Relevance for Age and Gender Classification . . . . .	50
<b>5</b>	<b>Conclusions and Future Research</b>	<b>59</b>
5.1	Conclusions . . . . .	59
5.2	Future Research . . . . .	61
	<b>Appendices</b>	<b>62</b>
	<b>Publications</b>	<b>69</b>
	<b>References</b>	<b>70</b>



# List of Figures

2.1	Structure of an automatic speaker classification system . . . . .	10
2.2	Structure of an automatic age classification system . . . . .	10
2.3	Frequency domain diagram of the source-filter explanation of the acoustics of a vowel (voiced) and a fricative (voiceless). The source spectrum (left), the vocal tract transfer function (middle), and the output spectrum (right), after Dellwo [16] . . . . .	11
2.4	Speech encoding process, after Young [70]. . . . .	13
2.5	Mel-Scale Filter Bank, after Young [70]. . . . .	16
2.6	Micro variations in vocal fold movements can be measured as shimmer (variation in amplitude) and jitter (variation in frequency), after Schotz [57]. . . . .	19
2.7	Speaker classification system . . . . .	25
2.8	Linear separating hyperplane for the non-separable data. The slack variable $\xi$ allows misclassified point. . . . .	27
3.1	Linear separating hyperplanes of SVM and FSVM for the non-separable data. The small membership $\lambda_i$ allows large error of misclassified point outside overlapping regions, hence the decision boundary tends to move to overlapping regions to reduce empirical errors in this region. . . . .	33
4.1	Accent classification for Broad, General and Cultivated groups. . . . .	42
4.2	Accent classification rates versus $C$ and $\gamma$ . . . . .	44
4.3	Accent classification versus age . . . . .	44

4.4	Accent classification versus age performed on male speakers . . . . .	45
4.5	Accent classification versus age performed on female speakers . . . . .	45

# List of Tables

2.1	Summary of the effects of several emotion states on selected acoustic features, after Ververidis [66]. Explanation of symbols: $>$ : increases, $<$ : decreases, $=$ : no change from neutral, $\nearrow$ : inclines, $\searrow$ : declines. Double symbols indicate a change of increased predicted strength. The subscripts refer to gender information: $M$ stands for males and $F$ stands for females. . . . .	22
4.1	Age and gender classes of the aGender corpus, where $f$ and $m$ abbreviate female and male, and $x$ represents children without gender discrimination. The last two columns represent the number of speakers/instances per set. . . . .	39
4.2	Distribution of emotions, data set EMO-DB . . . . .	40
4.3	Number of instances for the 5-class problem . . . . .	41
4.4	Standard deviation (%) of ACCENT classification from 10 experiments	43
4.5	Standard Deviation (%) of Accent classification Accuracy Versus Age Averaged on 10 experiments . . . . .	46
4.6	Paralinguistic feature set for Age and Gender classification, after Schuller [52]. . . . .	47
4.7	Emotion feature set for Emotion classification, after Schuller [51]. . .	47
4.8	Classification rates (%) of SVM and FSVM on the four data sets. . .	49
4.9	Classification rates (%) of SVM and FSVM on the four data sets with SFFS feature selection. . . . .	50
4.10	Classification rates of SVM and FSVM on the four data sets . . . . .	51

---

4.11	38 low-level descriptors with regression coefficients and 21 functionals.	53
4.12	Relevance of Low-Level-Descriptor types for all age and gender pairs using SVM (ANDOSL data set). . . . .	54
4.13	Relevance of Low-Level-Descriptor types for all age and gender pairs using SVM (aGender data set). . . . .	55
4.14	Relevance of Low-Level-Descriptor types for all age and gender pairs using FSVM (ANDOSL data set) . . . . .	56
4.15	Relevance of Low-Level-Descriptor types for all age and gender pairs using FSVM (aGender data set) . . . . .	57
4.16	Relevance of Low-Level-Descriptor types for all age and gender pairs using SVM. Averaging from Table 4.12 and Table 4.13 . . . . .	58
4.17	Relevance of Low-Level-Descriptor types for all age and gender pairs using FSVM. Averaging from Table 4.14 and Table 4.15 . . . . .	58

# Abbreviation

GMMs	Gaussian Mixture Models
SVM	Support Vector Machine
FSVM	Fuzzy Support Vector Machine
HTK	Hidden Markov Model Toolkit
HMM	Hidden Markov Model
FFS	Sequential Forward Floating Search
MFCCs	Mel-Frequency Cepstral Coefficients
LPC	Linear Prediction Coding