# A Likelihood-Ratio Based Forensic Voice Comparison in Standard Thai

# Supawan Pingjai

## **August 2019**

A thesis submitted for the degree of Doctor of Philosophy

The Australian National University

© Copyright by Supawan Pingjai 2019

All Rights Reserved

# **Declaration**

I hereby declare that this thesis is entirely my own work, except where acknowledgement to the contrary is made in the text. I also confirm that I have fully cited and referenced all material and results that are not original.

SIGNED:	
Supawan Pingjai	
DATE	

## Acknowledgements

I want to express my deep sense of gratitude to Dr Shunichi Ishihara for his valuable guidance and encouragement during the production of this thesis. He devoted a lot of his time to addressing the many difficulties I encountered during my research.

It would not have been possible to complete this thesis without the financial support of the University of Phayao, Thailand. I would similarly like to thank my parents for their financial support during my stay in Australia, and for the love they have shown throughout.

I am also very thankful to Dr Philip Rose, who inspired me to conduct this Forensic Voice Comparison research. I express my deep and sincere gratitude to Dr Rose for his suggestions at various stages of my thesis. His guidance immensely contributed to the evolution of my ideas.

This thesis received editorial input from Maxine McArthur and Bert Peeters. I would like to express my sincere thanks and deep appreciation for their professional support.

I also acknowledge, with thanks, the moral support of the ANU Thai Association in Canberra.

Finally, yet importantly, I wish to thank all the informants who contributed many hours of their time to participate in three recording sessions. Without their patience, the speech database for this PhD thesis would not have eventuated.

### **Abstract**

This research uses a likelihood ratio (LR) framework to assess the discriminatory power of a range of acoustic parameters extracted from speech samples produced by male speakers of Standard Thai. The thesis aims to answer two main questions: 1) to what extent the tested linguistic-phonetic segments of Standard Thai perform in forensic voice comparison (FVC); and 2) how such linguistic-phonetic segments are profitably combined through logistic regression using the FoCal Toolkit (Brümmer, 2007). The segments focused on in this study are the four consonants /s, tch, n, m/ and the two diphthongs [5i, ai].

First of all, using the alveolar fricative /s/, two different sets of features were compared in terms of their performance in FVC. The first comprised the spectrum-based distributional features of four spectral moments, namely mean, variance, skew and kurtosis; the second consisted of the coefficients of the Discrete Cosine Transform (DCTs) applied to a spectrum. As DCTs were found to perform better, they were subsequently used to model the consonant spectrum of the remaining consonants. The consonant spectrum was extracted at the center point of the /s, tch, n, m/ consonants with a Hamming window of 31.25 msec.

For the diphthongs [ɔi] - [nɔi L] and [ai] - [mai HL], the cubic polynomials fitted to the F2 and F1-F3 formants were tested separately. The quadratic polynomials fitted to the tonal F0 contours of [ɔi] - [nɔi L] and [ai] - [mai HL] were tested as well. Long-term F0 distribution (LTF0) was also trialed.

The results show the promising discriminatory power of the Standard Thai acoustic features and segments tested in this thesis. The main findings are as follows.

- 1. The fricative /s/ performed better with the DCTs ( $C_{llr} = 0.70$ ) than with the spectral moments ( $C_{llr} = 0.92$ ).
- 2. The nasals /n, m/ ( $C_{llr} = 0.47$ ) performed better than the affricate /tc<sup>h</sup>/ ( $C_{llr} = 0.54$ ) and the fricative /s/ ( $C_{llr} = 0.70$ ) when their DCT coefficients were parameterized.
- 3. F1-F3 trajectories ( $C_{llr} = 0.42$  and  $C_{llr} = 0.49$ ) outperformed F2 trajectory ( $C_{llr} = 0.69$  and  $C_{llr} = 0.67$ ) for both diphthongs [5i] and [ai].

- 4. F1-F3 trajectories of the diphthong [5i] ( $C_{llr} = 0.42$ ) outperformed those of [ai] ( $C_{llr} = 0.49$ ).
- 5. Tonal F0 ( $C_{llr} = 0.52$ ) outperformed LTF0 ( $C_{llr} = 0.74$ ).
- 6. Overall, better results were obtained when DCTs of /n/ [na: HL] and /n/ [noi L] were fused. ( $C_{llr} = 0.40$  with the largest consistent-with-fact  $SSLog_{10}LR = 2.53$ ).

In light of the findings, we can conclude that Standard Thai is generally amenable to FVC, especially when linguistic-phonetic segments are being combined; it is recommended that the latter procedure be followed when dealing with forensically realistic casework.

# **Table of contents**

Declaration	i
Acknowledgements	ii
Abstract	iii
Table of Contents	v
List of Tables	xi
List of Figures	xiii
Chapter 1 Introduction	1
1.1 Introduction	1
1.2 A brief history of the study of speaker recognition	1
1.3 Types of speaker recognition: Speaker identification/verification/forensic voice comparison	2
1.4 What is Forensic Voice Comparison?	5
1.5 Motivations	9
1.6 The research approach	10
1.6.1 Traditional vs automatic parameters	10
1.6.2 Statistical modelling techniques	11
1.7 Linguistic-phonetic segments and acoustic parameters	12
1.8 Research questions	14
1.9 Thesis outline	14
1.10 Summary	15
Chapter 2 Literature review	16
2.1 Introduction	16
2.2 Ideal features of forensic scientific evidence	16
2.2.1 Lack of control over variation	17
2.2.1.1 Between-speaker variation	17
2.2.1.2 Within-speaker variation	18
2.2.2 Reduction in dimensionality	20
2.3 Speech variation and the role of probability theory	21

2.4 What is the Likelihood Ratio?	. 22
2.4.1 Why forensic experts should limit themselves to the calculation of Likelihood Ratio (LR)	. 29
2.5 The Thai legal system	. 29
2.6 Shift to a new paradigm	. 31
2.7 Standard Thai and other main dialects: Phonetics and phonology	. 34
2.7.1 Standard Thai consonant phonemes	. 35
2.7.2 Standard Thai clusters	. 35
2.7.3 Standard Thai vowels	. 36
2.7.4 Standard Thai tones	. 37
2.8 Northern Thai dialect	. 38
2.8.1 Northern Thai consonant phonemes	. 38
2.8.2 Northern Thai clusters	. 39
2.8.3 Northern Thai vowels	. 39
2.8.3.1 Monophthongs	. 39
2.8.3.2 Diphthongs	. 39
2.8.4 Northern Thai tones	. 39
2.9 Southern Thai dialect	. 40
2.9.1 Southern Thai consonant phonemes	. 40
2.9.2 Southern Thai clusters	. 41
2.9.3 Southern Thai vowels	. 42
2.9.4 Southern Thai tones	. 42
2.10 Northeastern Thai dialect	. 43
2.10.1 Northeastern Thai consonant phonemes	. 43
2.10.2 Northern Thai clusters	. 43
2.10.3 Northeastern Thai vowels	. 43
2.10.3.1 Monophthongs	. 43
2.10.3.2 Diphthongs	. 44
2.10.4 Northeastern Thai tones	. 44
2.11 Speech signal representation	. 44
2.11.1 General feature extraction process: Short-time analysis	. 44
2.12 Source of individualizing information: Spectral, phonotactic, prosodic and idiolectal levels	. 46
2.13 DCT-smoothed spectrum vs cepstrally smoothed spectrum	. 47
2.14 Mel- and Bark-scaled DCT (cepstral) coefficients	. 52
2.14.1 Hertz-scaled DCT (cepstral) coefficients fitted to a raw spectrum	. 52

	2.14.2 Bark-scaled DCT (cepstral) coefficients fitted to a raw spectrum	. 54
	2.15 Spectrum of the fricatives	. 57
	2.15.1 Articulatory description of the fricatives	. 58
	2.15.2 Previous acoustic studies of English fricatives	. 59
	2.15.3 Previous FVC research on the English fricative /s/	. 62
	2.15.3.1 Acoustic measurement	. 63
	2.15.3.2 Experimental results of /s/	. 63
	2.16 Spectrum of the affricates	. 63
	2.16.1 Articulatory description of affricates	. 64
	2.16.2 Previous acoustic studies of the affricates	. 64
	2.16.3 Previous FVC studies of affricates	. 65
	2.17 Spectrum of the nasals /m/, /n/, and /ŋ/	. 65
	2.17.1 Articulatory description of nasals	. 66
	2.17.2 Previous acoustic studies of the nasals	. 66
	2.17.3 Previous FVC studies of nasals	. 67
	2.17.3.1 FVC results of English /n/	. 68
	2.17.3.2 FVC results of English /m/	. 68
	2.17.3.3 FVC results of English /ŋ/	. 68
	2.18 Fundamental frequency (F0)	. 69
	2.18.1 Background knowledge	. 69
	2.18.2 Tonal F0	. 71
	2.18.3 Previous FVC research on tonal F0	. 71
	2.18.4 Long-term F0 distribution	. 72
	2.18.5 Previous FVC studies on LTF0	. 72
	2.19 Formant trajectory	. 74
	2.19.1 Background knowledge	. 74
	2.19.2 Previous FVC research	. 74
	2.20 Summary	. 79
C	Chapter 3 Methodology	. 80
	3.1 Introduction	. 80
	3.2 Likelihood ratio (LR) as the logical framework	. 80
	3.3 Statistical tools: the MVLR formula	. 81
	3.4 Speech corpus	. 84
	3.4.1 Informants	. 85
	3.4.2 Elicitation	. 85

	3.5 Tippett plots	88
	3.6 Logistic regression calibration	90
	3.7 Why logistic regression is better than the Gaussian models	91
	3.8 Metric for assessing the validity (accuracy) of MVLR	92
	3.9 Logistic regression fusion	94
	3.10 Summary	96
C	hapter 4 Pilot FVC studies using Standard Thai diphthongs	97
	4.1 Introduction	97
	4.2 Pilot study on the Standard Thai diphthongs [i:aw], [u:a] and [u:a]	97
	4.2.1 Parameters and informants	97
	4.2.2 Results	98
	4.3 Pilot study on the Standard Thai diphthongs [o:i] and [ə:i]	. 101
	4.3.1 Parameters and informants	. 101
	4.3.2 Results	. 102
	4.4 Pilot study on Standard Thai diphthongs [ai] and [u:a]	. 104
	4.4.1 Parameters and informants	. 104
	4.4.2 Results	. 104
	4.5 Summary	. 106
C	hapter 5 Results of the spectral moments of /s/ and cepstral coefficients (CCs) of /s, teh,	
	n, m/	. 108
	5.1 Introduction	. 108
	5.2 Segmentation criteria	. 108
	5.2.1 Segmentation of /s/	. 109
	5.2.2 Segmentation of /tc <sup>h</sup> /	. 111
	5.2.3 Segmentation of /n/ - [noi L]	. 112
	5.2.4 Segmentation of /n/ - [na: HL]	. 114
	5.2.5 Segmentation of /m/ - [mai HL] 'no'	. 115
	5.3 Spectral mean, variance, skew, kurtosis (spectral moments)	. 116
	5.4 Results of the spectral moments of /s/	. 118
	5.4.1 Distribution of the spectral moments using histograms	. 119
	5.4.2 ANOVA results	. 120
	5.4.3 MVLR results	
	5.4.4 DCT results	. 132
	5.4.4.1 Fricative /s/ extracted from the word [sa:m LH] 'three'	. 132
	5.4.4.2 Affricate /teh/ extracted from the word [tehai HL] 'yes'	. 135

	5.4.4.3 Nasal /n/ extracted from the particle [noi L]	. 137
	5.4.4.4 Nasal /n/ extracted from the word [na: HL thi: HL] 'duty'	. 139
	5.4.4.5 Nasal /m/ extracted from the word [mai HL] 'no'	. 142
	5.5 Overall comparisons and discussions	. 144
	5.6 Summary	. 145
(	Chapter 6 Results of the formant trajectories of the diphthongs [ɔi] - [nɔi L] and [ai] - [mai HL]	. 147
	6.1 Introduction	. 147
	6.2 Why were the F2 trajectories of the diphthongs [ɔi] - [nɔi L] and [ai] - [mai HL] experimented on?	. 147
	6.3 Informants	. 148
	6.4 Segmentation	. 148
	6.4.1 Formant trajectories of [ɔi] - [nɔi L]	. 148
	6.4.2 Formant trajectoires of [ai] - [mai HL]	. 150
	6.4.3 Formant tracking errors and manual correction	. 151
	6.4.4 Discarding of poor recording speech samples	. 154
	6.4.5 Formant trajectories of the diphthong [ɔi]	. 155
	6.4.6 Polynomial curve fitting (cubic polynomials)	. 156
	6.4.7 Formant trajectories of the diphthong [ai] - [mai HL]	. 158
	6.4.8 Polynomial curve fitting of the diphthong [ai] - [mai HL]	. 159
	6.5 Experimental results: F2 trajectory of diphthongs [oi] - [noi L] and [ai] - [mai]	. 160
	6.6 Experimental results: F1-F3 trajectories of diphthongs [ɔi] - [nɔi L] and [ai] - [mai HL]	. 163
	6.7 Discussion	. 165
	6.8 Summary	. 166
(	Chapter 7 Results of the fundamental frequency (F0): Long-term F0 (LTF0) and tonal F0	. 167
	7.1 Introduction	. 167
	7.2 Long-term fundamental frequency (LTF0)	. 167
	7.2.1 Data extraction	. 168
	7.2.2 Standard Thai LTF0 distribution plots	. 169
	7.3 Experimental results when using LTF0	. 170
	7.3.1 LTF0: all six features	. 171
	7.3.2 LTF0: the four spectral moments	. 172
	7.3.3 LTF0: model F0 and modal density	. 174
	7.4 Experimental results when using the 10% percentile technique	. 175

7.5 Tonal F0 of [ɔi] - [nɔi L] and [ai] - [mai HL]	17
7.5.1 F0 tracking of [ɔi] - [nɔi L]	78
7.5.2 F0 tracking of [ai] - [mai HL]	79
7.5.3 F0 contours of [ɔi] - [nɔi L]	31
7.5.4 F0 contours of [ai] - [mai L]	31
7.5.5 Polynomial curve fitting of tonal F0 for [ɔi] - [nɔi L]	32
7.5.6 Polynomial curve fitting of tonal F0 for [ai] - [mai HL]	33
7.6 Informants	33
7.7 Experimental results when using tonal F0	34
7.7.1 The Tippett plot of [ɔi] - [nɔi L]	34
7.7.2 The Tippett plot of [ai] - [mai HL]	35
7.8 Summary	37
Chapter 8 Conclusions and recommendations for future research	38
8.1 Introduction	38
8.2 Answers to the research questions	38
8.3 Future research	<b>)</b> 3
8.3.1 Speech corpus	<b>)</b> 3
8.3.2 Parameters 19	<b>)</b> 4
8.3.2.1 Voice onset time (VOT) of a stop $/k^h/$	<b>)</b> 4
8.3.2.2 Trill /r/ and liquid /l/	<b>)</b> 4
8.3.3 Statistical tools and data extraction techniques	<b>)</b> 5
8.4 Implications for the forensic academic community	<b>)</b> 5
8.5 Summary	<del>)</del> 6
References 19	€7
Appendix A Recording manuals	17
Appendix B F1-F3 values of [5i] plotted against a normalized time scale (100 msec) 23	35
Appendix C F1-F3 trajectories of [5i] plotted together with cubic polynomials	10
Appendix D F1-F3 values of [ai] plotted against a normalized time scale (100 msec) 24	15
Appendix E F1-F3 trajectories of [ai] plotted together with cubic polynomials	50
Appendix F Tonal F0 values of [5i] plotted against a normalized time scale (100 msec)	55
Appendix G Tonal F0 values of [5i] plotted together with a quadratic polynomial curve fitting	50
Appendix H Tonal F0 values of [ai] plotted against a normalized time scale (100 msec)	55
Appendix I Tonal F0 values of [ai] plotted together with a quadratic polynomial curve	
fitting	70

# **List of Tables**

Table 1: Acoustic parameters and linguistic-phonetic segments	13
Table 2: Sources of within-speaker variation	18
Table 3: Verbal equivalents of LRs	26
Table 4: Standard Thai consonant phonemes	35
Table 5: Standard Thai consonant clusters	36
Table 6: Northern Thai consonant phonemes	38
Table 7: Northern Thai tones	39
Table 8: Consonant phonemes of the Southern Thai dialect (Phang-Nga)	41
Table 9: Southern Thai clusters	41
Table 10: Southern Thai tones	42
Table 11: Northeastern Thai consonant phonemes	43
Table 12: Northeastern Thai Tones	44
Table 13: Acoustic parameters, datasets and filter conditions for the fricative /s/	62
Table 14: Frequencies and their corresponding spectral data extracted at the midpoint of the token /s/ spoken by Speaker 8, Session 1	
Table 15: ANOVA results for speaker and/or session (N = 56 in 4000 and 8000 Hz) on each spectral moment calculated from /s/	
Table 16: Bonferroni's pairwise comparisons for a Skew (m3) of /s/	. 121
Table 17: Log <sub>10</sub> LR, C <sub>llr</sub> , and EER values of the fricative spectra /s/ according to the combine parameters (leftmost column) measured at the temporal midpoint of the fricative /s/ from 56 speakers, in 500-4000 Hz as well as 500-8000 Hz conditions	L
Table 18: Calibrated Log <sub>10</sub> LR, C <sub>llr</sub> , and EER of the fricative /s/ - [sa:m LH] when its 15 DC and 20 DCTs were parameterized in both Hertz and Bark scales, in the 500-8000 Hz filter	
Table 19: Calibrated Log <sub>10</sub> LR, C <sub>llr</sub> , and EER of the affricate /te <sup>h</sup> / - [te <sup>h</sup> ai HL] when its 15 DO and 20 DCTs were parameterized in both Hertz and Bark scales, respectively	CTs
Table 20: Calibrated Log <sub>10</sub> LR, C <sub>llr</sub> , and EER of the nasal /n/ - [noi L] when its 15 DCTs and 20 DCTs were parameterized in both Hertz and Bark scales, respectively	
Table 21: Calibrated Log <sub>10</sub> LR, C <sub>llr</sub> , and EER of the nasal /n/ - [na: HL] when its 15 DCTs are 20 DCTs were parameterized in both Hertz and Bark scales, respectively	
Table 22: Calibrated Log <sub>10</sub> LR, C <sub>llr</sub> , and EER of the nasal /m/ - [mai HL] when its 15 DCTs at 20 DCTs were parameterized in both Hertz and Bark scales, respectively	
Table 23: Ranking order of the linguistic-phonetic segments experimented on in terms of the $C_{llr}$ and EER values (from low to high) with their corresponding acoustical parameters	

Table 24: Calibrated Log <sub>10</sub> LR, C <sub>llr</sub> , and EER values when cubic polynomial coefficients from the diphthongs [ɔi] - [nɔi L] and [ai] - [mai HL] were parameterized, respectively 160
Table 25: Calibrated $Log_{10}LR$ , $C_{llr}$ , and EER values when cubic polynomial coefficients of the diphthongs [ $\mathfrak{i}$ i] - [ $\mathfrak{n}\mathfrak{i}$ i L] and [ $\mathfrak{a}$ i] - [ $\mathfrak{m}\mathfrak{i}$ i HL] were parameterized, respectively
Table 26: Log <sub>10</sub> LR, C <sub>llr</sub> , and EER values when mean, SD, skew, kurtosis, modal F0, and modal density were combined according to different patterns
Table 27: Largest consistent-with-fact/contrary-to-fact SSLRs and DSLRs of Standard Thai when all six LTF0 features were parameterized
Table 28: Largest consistent-with-fact/contrary-to-fact SSLRs and DSLRs of Standard Thai when the four spectral moments (mean, SD, skew, kurtosis) were parameterized
Table 29: Largest consistent-with-fact/contrary-to-fact SSLRs and DSLRs of Standard Thai when modal F0 and model density were parameterized
Table 30: Largest consistent-with-fact/contrary-to-fact SSLRs and DSLRs of Standard Thai LTF0 when their distribution was captured by the 10% percentiles and parameterized in a Hertz scale
Table 31: Calibrated Log <sub>10</sub> LR, C <sub>llr</sub> , and EER values when a quadratic polynomial was fitted to the F0 contours of [ɔi] - [nɔi L] and [ai] - [mai HL], respectively

# **List of Figures**

Figure 1: Standard Thai monophthongs3	36
Figure 2: Standard Thai diphthongs3	37
Figure 3: F0 contours of the five Standard Thai tones for the same segmental sequence [pa:] 3	37
Figure 4: The South of Thailand4	10
Figure 5: Feature extraction process	15
Figure 6: A raw spectrum (black) extracted from the temporal midpoint of an oral vowel /i/ and its corresponding Hertz-scaled DCT curve fitting (red) using 512 data points	18
Figure 7: Cepstral analysis5	50
Figure 8: A DCT-smoothed signal (cepstrally smoothed spectrum) superimposed on the original spectrum (in black) by summing up the first 30 half-cycle cosine waves on a Hz scale (in red) of /s, teh, m, n/	53
Figure 9: A DCT-smoothed signal (cepstrally smoothed spectrum) superimposed on the original spectrum (in black) by summing up the first 30 half-cycle cosine waves in a Bark scale (in red) of /s, tch, m, n/	54
Figure 10: A DCT-smoothed signal (cepstrally smoothed spectrum) of /s, teh, m, n/ from Speaker 1's 1st session, plotted by summing up the first 30 half-cycle cosine waves in a Hertz scale uttered on 1st and 2nd repeats (in black), and Speaker 1's 2nd session, similarly consisting of 1st and 2nd repeats (in red).	55
Figure 11: A DCT-smoothed signal (cepstrally smoothed spectrum) of /s, teh, m, n/ plotted by summing up the first 30 half-cycle cosine waves in a Hertz scale from Speaker 2's 1st session (1st – 5th repeats, in black) and Speaker 3's 1st session (1st – 5th repeats, in red) 5	57
Figure 12: MVLR formula (Aitken & Lucy, 2004)	32
Figure 13: Floor plan of the recording rooms	36
Figure 14: Tippett plot of 20 Hertz-scaled DCTs of /m/ - [mai HL]	39
Figure 15: Tippett plot of the 15 Hertz-scaled DCTs of /teh/ [tehai HL]9	)2
Figure 16: Tippett plot of [i:aw] - [li:aw H] when [F2, F3, F4] were fitted with cubic polynomials (reproduced from Pingjai et al., 2013).	99
Figure 17: Tippett plot of [u:a] - [phu:a? HL] when [F2, F3, F4] were fitted with cubic polynomials (reproduced from Pingjai et al., 2013).	00
Figure 18: Tippett plot of [u:a] - [su:an L] when [F2, F3, F4] were fitted with cubic polynomials (reproduced from Pingjai et al., 2013).	00
Figure 19: Tippett plot of [o:i M] - [do:i M] when its F2 trajectory was fitted by cubic polynomials	)2
Figure 20: Tippett plot of [ə:i] - [khə:i M] when its F2 trajectory fitted by cubic polynomials	13

Figure 21: Tippett plot of [ai] - [te"ai HL] when its F0 contour was fitted by quadratic polynomials	105
Figure 22: Tippett plot of [u:a] - [ru:am HL] when its F0 contour was fitted by linear regression.	. 105
Figure 23: Label tier (top), waveform (middle), and spectrogram (bottom) of the phrase "every three weeks", with overlaid formants. The section highlighted in grey in the label tier shows the target segment /s/	
Figure 24: Label tier (top), waveform (middle), and spectrogram (bottom) of the phrase "every three weeks", with overlaid formants. The section highlighted in grey in the label tier shows the target segment /s/	
Figure 25: Label tier (top), waveform (middle), and spectrogram (bottom) of part of the sentence frame "This is because we do not have any responsibility", with overlaid formatracking values. The section highlighted in grey in the label tier shows the target segment /teh/	t
Figure 26: Label tier (top), waveform (middle), and spectrogram (bottom) of part of the sentence frame "This is because we do not have any responsibility", with overlaid formatracking values. The section highlighted in grey in the label tier shows the target segment /n/	t
Figure 27: Label tier (top), waveform (middle), and spectrogram (bottom) of a target segment /n/ - [noi L]	. 114
Figure 28: Label tier (top), waveform (middle), and spectrogram (bottom) of the target segment /n/ - [na: HL thi: HL]	. 115
Figure 29: Label tier (top), waveform (middle), and spectrogram (bottom) of the target segment /m/ - [mai HL].	. 116
Figure 30: Histograms of the spectral mean, variance, skew, and kurtosis of /s/ uttered by 56 speakers	. 119
Figure 31: Tippett plots for SS and DS comparisons when mean, variance, skew and kurtosis were parameterized in the 500-4000 Hz (left) and 500-8000 Hz (right) band-pass filters.	. 124
Figure 32: Tippett plots for SS and DS comparisons when two or three of the four parameter (as indicated on top of each of the plots) were combined in a 500-4000 Hz band-pass filter	
Figure 33: Tippett plots for SS and DS comparisons when two or three of the four parameter (as indicated on top of each of the plots) were combined in a 500-8000 Hz band-pass filter	
Figure 34: The means (circles) and ranges of spectral mean	. 129
Figure 35: The means (circles) and ranges of spectral variance.	130
Figure 36: The means (circles) and ranges of spectral skew	130
Figure 37: The means (circles) and ranges of spectral kurtosis	131

Figure 38: Tippett plot of the best performing parameter, 15 Bark-scaled DCTs of /s/ - [sa:m LH]	133
Figure 40: Tippett plot of the best performing parameter, 20 Bark-scaled DCTs of $/tc^h/-$ [ $tc^h$ ai HL].	136
Figure 41: Tippett plot of the worst performing parameter, 15 Hertz-scaled DCTs of $/te^h/-[te^h \ ai \ HL]$ .	137
Figure 42: Tippett plot of the best performing parameter, 20 Hertz-scaled DCTs of /n/ - [noi L].	138
Figure 43: Tippett plot of the worst performing parameter, 20 Bark-scaled DCTs of /n/ - [noi L].	139
Figure 44: Tippett plot of the best performing parameter, 15 Bark-scaled DCTs of /n/ - [na: HL].	141
Figure 45: Tippett plot of the worst performing parameter, 20 Bark-scaled DCTs of /n/ - [na: HL].	141
Figure 46: Tippett plot of the best performing parameter, 20 Hertz-scaled DCTs of /m/ - [mai HL].	143
Figure 47: Tippett plot of the worst performing parameter, 15 Bark-scaled DCTs of /m/ - [mai HL].	144
Figure 49: Label tier (top), waveform (middle), and spectrogram (bottom) of [ai] - [mai HL] with the corresponding formant frequencies tracking	150
Figure 50: Label tier (top), waveform (middle), and spectrogram (bottom) of [ɔi] - [nɔi L] with the corresponding formant tracking containing some errors (in the last one-third)	151
Figure 51: Label tier (top), waveform (middle), and spectrogram (bottom) of [ɔi] - [nɔi L] with the corresponding formant tracking after manual correction	152
Figure 52: Label tier (top), waveform (middle), and spectrogram (bottom) of [ai] - [mai HL] with the corresponding formant tracking containing some errors.	
Figure 53: Label tier (top), waveform (middle), and spectrogram (bottom) of [ai] - [mai HL] with the corresponding formant tracking after manual correction	153
Figure 54: Label tier (top), waveform (middle), and spectrogram (bottom) of [ɔi] - [nɔi L] containing many formant tracking errors	154
Figure 55: Speaker 1's F1-F3 trajectories of the diphthong [ɔi] - [nɔi L] plotted against normalized duration (100 msec).	155
Figure 56: Speaker 1 and Speaker 9's F1-F3 trajectories of the diphthong [ɔi] - [nɔi L] plotted against normalized duration (100 msec).	156
Figure 57: F2 trajectory values of [ $\circ$ i] - [ $\circ$ i] (black dots), plotted together with its cubic polynomial curve fitting (dotted red line) of $(0.000531)x^3 + 0.155159x^2 + (-15.291889)x + 1539.427481$ .	157
Figure 60: F2 trajectory values (black dots) of the diphthong [ai] - [mai HL] plotted together with its cubic polynomial fitting (dotted red line) of (-0.000854)x <sup>3</sup> + (-0.023227)x <sup>2</sup> + (16.052403)x + 1589.501831	

Figure 61: Tippett plot of [5i]'s second formant (F2) trajectory when cubic polynomials were parameterized.	. 161
Figure 62: Tippett plot of [ai]'s second formant (F2) trajectory when cubic polynomials were parameterized.	. 162
Figure 63: Tippett plot of [ɔi]'s F1-F3 trajectories when cubic polynomials were parameterized.	. 164
Figure 64: Tippett plot of [ai]'s F1-F3 trajectory when cubic polynomials were parameterized.	. 165
Figure 65: Label tier, speech waveforms, and F0 tracking. Each u in the label tier represents an utterance of speech samples used to extract LTF0	. 168
Figure 66: LTF0 distribution plots extracted from Speakers 1-4. Blue and red curves represent the first and second recording sessions, respectively.	. 169
Figure 67: Tippett plot of LTF0 when all six LTF0 features (mean, SD, skew, kurtosis, modal F0 and modal density) were parameterized.	. 171
Figure 70: Tippett plot of LTF0 when its distribution was captured by the 10% percentiles and parameterized in a Hertz scale.	. 176
Figure 71: Label tier (top), waveform and overlaid F0 tracking (middle), and spectrogram (bottom) for the target segment [ɔi] - [nɔi L]	. 178
Figure 72: Label tier (top), waveform and overlaid F0 tracking (middle), and spectrogram (bottom) for the target segment [ai]	. 179
Figure 73: Label tier (top), waveform and overlaid formant tracking (middle), and F0 tracking (bottom) for the target segment [ɔi] - [nɔi L]. There are some F0 tracking errors the middle of [ɔi]	
Figure 74: Label tier (top), waveform and overlaid formant tracking (middle), and F0 tracking after manual correction (bottom) for the target segment [ɔi] - [nɔi L]	. 180
Figure 75: F0 contours of the diphthong [ɔi] - [nɔi L] plotted against normalized duration (100 msec) for Speakers 1 and 9	. 181
Figure 76: F0 contours of the diphthong [ai] - [mai HL] plotted against normalized duration (100 msec) for Speakers 1 and 9	
Figure 77: F0 values (black dots) of the diphthong [5i] - [n5i L] plotted together with its quadratic polynomial curve fitting (dotted red line) in normalized duration (100 msec)	. 182
Figure 80: Tippett plot of [ai] - [mai HL] when its tonal F0 contours were parameterized by a quadratic polynomial	
Figure 81: Tippett plot for the fused and calibrated LRs for /n/ - [na: HL] and /n/ - [noi L]	

## Chapter 1

#### Introduction

#### 1.1 Introduction

It is clear from our everyday experience that humans are able to quite effectively recognize or identify familiar speakers such as family members, friends, and celebrities by their voices (Atal, 1972, p. 1687; Nolan, 2001, p. 276). Such scenarios could include, for example, recognizing someone who is speaking on the phone from just a relatively short utterance such as *hello*, or recognizing someone who is speaking but whom we cannot directly see (Bricker et al., 1971). This "normal everyday ability" or "naïve speaker recognition" (Nolan, 2001, p. 276), prompts us to investigate how acoustic cues can reliably encode information about a speaker.

#### 1.2 A brief history of the study of speaker recognition

As early as 1962, Kersta (1962) developed a spectrographic methodology in order to compare speech samples. This was colloquially called "voiceprint". Kersta claimed that such spectrographic patterns obtained from speech samples are "uniquely different enough to make any given speech sample identifiable with the same accuracy that fingerprint identification enjoys" Kersta (1962, p. 1355). His assumption is principally based on the notion that the vocal tract size and the manner of articulation are uniquely different among individuals. As such, Kersta (1962, p. 1355) strongly believed that the simple visual matching of spectrographic patterns could be used to identify speakers. This theory is easy to dispute in the present day, however, through the use of state-of-the-art speech analysis tools such as Praat (Boersma & Weenink, 2003) and the EMU speech database system (Cassidy, 1999). These tools can empirically show that even a sound that linguistically seems identical to another produced by the same speaker, and uttered just a few seconds apart, is different acoustically (e.g. having a different frequency (Hz) and relative amplitude (dB)). This means that the resulting differences can be audible and measurable (Ladefoged & Johnson, 2014, pp. 207-208). Vanderslice (1969) also affirmed that some speech spectrograms that looked similar were actually produced by different speakers, and some that looked different were produced by the same speakers. In Chapter

2, I will elaborate more on how speech varies both within and between speakers, making speech difficult to discriminate forensically.

The "voice print method" by Kersta (1962, p. 1355), just described, has been challenged by many researchers such as Bolt et al. (1970), who questioned its reliability and validity as well as its applicability in investigations and in judicial applications. In particular, the questions researchers like Bolt et al. (1970) raise are about the probabilities of false identification and of correct identification, as well as the probabilities of correct identification under various conditions, ranging from controlled situations to forensically realistic ones.

In the early 1970s, many experiments concerning acoustic parameters in automatic speaker recognition were undertaken. These experiments can be placed into two categories – speaker *verification* and speaker *identification* (Bricker et al., 1971, p. 1427). Specifically, speaker verification is the process of rejecting or accepting the identity claimed by a speaker, whereas speaker identification is the process of assigning an unknown utterance to a known speaker/group (or of leaving it unassigned, as the case may be) (ibid.). Among the automatic speaker identification experiments done between 1960 and 1970 are those by Glenn and Kleiner (1968), Pruzansky (1963), Pruzansky and Mathews (1964), and Wolf (1970), where the first of these studies employed parameters extracted from the nasal spectrum, and the second and third used parameters based on a spectrum obtained from whole words; the last study, by Wolf, instead employed phonological features as parameters for testing. All of these speaker identification studies showed a similar rate of success of around 90% correct recognition (or more). However, since these studies were exploratory experiments in the early days of automatic speaker recognition research, they were controlled using a small population of 10-30 voices recorded in a laboratory environment and obtained from a reading style; as such these favorable factors might have contributed to such a high success rate (ibid.).

# 1.3 Types of speaker recognition: Speaker identification/verification/forensic voice comparison

When considering the use of standardized terminology within the field, which may be confusing to newcomers, it is useful to look at definitions of the term *speaker recognition* 

itself and its classifications (Gonzalez-Rodriguez, Ortega-Garcia, & Sanchez-Bote, 2002; Meuwly, 2004; Nakasone & Beck, 2001; Nolan, 1983; Rose, 2002, 2006). We will firstly begin here with a definition of speaker recognition by Atal (1976, p. 460), who states that:

The term speaker recognition refers to any decision-making process that uses some features of the speech signal to determine if a particular person is the speaker of a given utterance which will include tasks such as identification, verification, discrimination and authentication of speakers.

From the above, *speaker recognition* can be defined as the task of using acoustic features to discriminate between the speech samples of different speakers. Indeed, Rose and Clermont (2001, p. 33) point out that *identification* is the end result of a process of discrimination. That is, if speech samples are discriminated as coming from the same speaker, the suspect can be identified as the offender; if not then no identification is possible (ibid.). Therefore, discriminating a speech sample does not necessarily entail identification. In contrast, speaker verification, which is the most common task in speaker recognition, is a process of accepting or rejecting the identity claimed by an unknown whose utterances are to be compared against those in the stored reference samples whose identities are claimed (Nolan, 1983, p. 8; Rose, 2002, p. 85). Usually the result of the verification process can be one of the following: correct acceptance; correct rejection; false acceptance; false rejection (Rose, 2002, p. 85). False acceptance is considered the most serious result in speaker verification, as the security system verifies the speaker is an imposter and not who he claims to be. False rejection happens when a security system incorrectly denies a bona fide speaker access into his or her installation. Regarding the term *authentication*, which is not directly relevant to the current work, this involves tasks such as determining if speech samples have been digitally edited or not (Rose, 2002, p. 2).

However, as outlined in a position statement resulting from the collaborative effort of a number of researchers and forensic practitioners working in the United Kingdom (French, Nolan, Foulkes, Harrison, & McDougall, 2010), the term *recognition* and its two main categories, i.e. *identification* and *verification*, are not appropriate for the current research; instead the term *comparison*, which is regarded as neutral amongst these terms, will be adopted (Rose & Morrison, 2009, p. 7). This is because the terms *recognition*, *identification* and *verification* imply that we can give a definitive answer to the question of whether a suspect is guilty or not. In other words, these words imply that *categorical* 

decisions can be made in answer to the question: are the unknown and known samples from the same speaker? For a forensic expert to categorically decide whether a suspect is guilty or not is, in fact, logically impossible and legally inappropriate (Rose, 2002, p. 89). It is intuitively easy and straightforward to understand why it is legally inappropriate. This is because only the trier-of-fact would be in the position where the ultimate decision (i.e. the suspect being guilty or not) is to be made. Thus, in this scenario, forensic experts trying to give a decision on guilt would be seen as usurping the duties of the fact finders (e.g. judges or juries). Although in Chapter 2, I will explain in detail why it is logically impossible to make such categorical decisions, one has to consider all relevant information to the case (e.g. all different types of evidence presented). However, it is impossible for experts to consider all the information relevant to the case at hand; thus, it is impossible for forensic experts to derive a categorical decision (or an ultimate decision). Having said that, it is far easier to derive a categorical rejection where, for example, it is sometimes clear to trained linguists that speech samples sound too different from one another (e.g. different sexes, languages, accents) (Rose, 2002, pp. 64-65). It is considered that it will usually be obvious to the trained linguists what language a speaker is speaking, whether a speaker is a male or a female, and broadly what accent a speaker is using (Morrison, Enzinger, & Zhang, 2016; Rose, 2002). But it is also reported that this is not a trivial task (Hughes & Rhodes, 2018). Moreover, such a categorical rejection may be arrived at with a closed set of comparisons (ibid.).

Following on from the above discussion concerning the terms selected for use in the current work, Rose and Morrison (2009, p. 7) point out that *speech samples*, not *speakers*, are the things that are being compared. Thus, the term 'forensic *voice* comparison' (FVC) should be adopted instead of 'forensic *speaker* comparison'. Rose (2002, p. 278) defines *voice* as "vocalisations (i.e. sound produced by a vocal tract) when thought of as made by a specific individual and recognisable as such". In this definition, voice is produced by a speaker and needs to be perceived as such by its speaker. Since the voice is the only item to be compared, not any other aspect, e.g. DNA profiling or finger print analysis of the speaker, I will use the term 'forensic *voice* comparison' (FVC) in the current thesis. Its exact meaning is discussed in §1.4.

#### 1.4 What is Forensic Voice Comparison?

FVC, as the name suggests, concerns the task of comparing speech samples, usually of the offender and the suspect(s), under two competing hypotheses, one of which is the prosecution hypothesis, according to which the speech samples are from the same speaker, while the other one is the defense hypothesis, according to which the speech samples are from different speakers (Robertson & Vignaux, 1995; Champod & Meuwly, 2000; Aitken & Taroni, 2004; Rose & Morrison, 2009). The results of such comparisons are dependent on the ratio of two conditional probabilities: firstly, how likely the evidence occurs under the prosecution hypothesis; and secondly, how likely the same evidence occurs under the defense hypothesis; these probibilities make up the so-called "strength of voice" evidence, which can be mathematically expressed as follows:

$$LR = \frac{p(E|Hp)}{p(E|Hd)}$$

#### **Equation 1**

In equation 1, p stands for probability, Hp stands for prosecution hypothesis (the two speech samples come from the same speaker), Hd stands for defense hypothesis (the samples come from different speakers), and E for speech evidence.

The above ratio between the two probabilities is called *a likelihood ratio* (LR); it shows the strength of the evidence. In equation 1, the difference between the offender and suspect samples is considered as *E*. This means that an LR tells us how much more likely (*p*) a difference between the offender and suspect samples is on the assumption that they have come from the same speaker (*Hp*) than on the assumption that they have come from different speakers (*Hd*) (Robertson & Vignaux, 1995; Champod & Meuwly, 2000; Aitken & Taroni, 2004; Rose & Morrison, 2009, p. 6). For example, an LR of 20 means that it is 20 times more likely that we will observe similarities/differences between speech samples that are from the same speaker rather than from different speakers. Another way of understanding the LR is that the numerator calculates the similarities/differences between the speech samples of the suspect and those of the offender; the denominator calculates the typicality of the speech samples being compared, i.e. how likely it is that by chance we will find speech samples that are similar to those of the suspects and the offenders in a relevant population sample (Rose, 2002, p. 58). We should be aware that the strength

of voice evidence or LR depends not only on the similarities/differences between speech samples, but also on how typical they are (ibid.). One of the arguments of the current thesis concerns the proper role of forensic experts when dealing with LR. I will argue throughout that forensic experts should limit themselves only to its evaluation.

In this section, I will briefly illustrate how we can calculate LRs using voice-related matter. Suppose that a court wants to know whether given speech samples are from the same or from different speakers. Imagine further that the numerator of the LR formula is 90% and the denominator is 10%; the ratio between them, which is an LR, is 9. The LR value of 9 means that the voice evidence is 9 times more likely to be from the same speaker (Hp) than from different speakers (Hd). The LR will be elaborated on more in Chapter 2, particularly in the context of the Bayesian Theorem.

Given the previous definitions, FVC is actually a subtype of speaker verification because both of them deal with two hypotheses: whether two speech samples are likely to be from the same or from different speakers. However, there are three main differences between the two terms. Thus, it is worth contrasting the term *FVC* on the one hand and the term *speaker verification* on the other, with reference to the three crucial differences (1. reference data; 2. categorical decision and threshold; and 3. control over samples) that I will outline below.

#### 1) Reference data

In speaker verification, the referenced speech samples are from a known population, such as a company's clientele and employees, and their speech properties (whether they are from the same or different speakers) are known (Rose, 2002, p. 88). As such, the threshold needed to discriminate speech samples produced by one and the same speaker from speech samples produced by different speakers can be set directly using the (known) reference data (as we know which speech samples are from which speakers). This means that the threshold is not only maximized by the between-speaker distances (which are directly estimated from the known reference data), but also that it can be updated by changes in the known reference data (Furui, 1981, p. 258).

In contrast, the referenced speech samples in FVC are not known beforehand and the corresponding acoustic properties can only be *approximated* (Broeders, 1995). Thus, the constitution of the reference data for FVC depends on circumstances (Rose, 2002, p. 89).

As Gold and Hughes (2014, p. 295) point out: "without knowing who the offender is, it is not possible to be certain what the population is of which he is a member". In agreement with Gold and Hughes's statement, Rose points out that it is still plausible for competent linguists to approximately delimit a population that is relevant to the offender using the information available (e.g. age, sex, accent) (Rose, 2002, pp. 64-65, 84). Selecting appropriate reference data is a very important aspect of the LR framework, which needs to be addressed here. First, as Drygajlo, Meuwly, and Alexander (2003) point out, the suspected speaker reference data should be as much as possible equivalent to that of the questioned speaker. This includes, for example, the equivalence in speaking style, quantity of the recorded speech and the technical characteristics of speech recordings. More discussion on refining the reference data specified in the defense hypothesis can be found, among others, in Morrison, Enzinger, and Zhang (2016), Hicks et al. (2015) and Hughes and Rhodes (2018). Second, in order to assess the extent of within-speaker variations, the suspected speaker reference data should be collected on two separate occasions in order to model the within-speaker variations, which always vary over time (ibid., p. 691). Third, FVC should be conducted on the open-set speakers, as opposed to the closed set, in order to avoid a misleading FVC calculation (Champod & Meuwly, 2000, p. 196). According to Champod and Meuwly (2000, p. 196), "it seems particularly unfair to disclose only the identity of the best candidate without providing the evidence obtained for the others". Last but not least, the task of FVC should also be conducted on a large potential population database as the reliable statistical models rely heavily on large population databases (Drygajlo et al., 2003, p. 691). However, there are many theoretical and practical issues that should be addressed. I will discuss these separately in Chapter 3, but will now return to my discussion of the differences between speaker verification and FVC.

#### 2) Categorical decision and threshold

In speaker verification, the ultimate goal is to give a categorical answer (yes/no) to the question: are the speech samples that are being verified the same as those in previously stored reference data? In such a verification system, as discussed above, the decision thresholds need to be calculated from a known and pre-set population so that (tested) speech samples with differences not exceeding a certain threshold are considered as being from the same speaker, but also that samples with differences greater than the threshold

are considered to have come from different speakers (Rose, 2002, p. 90). In the forensic field, this categorical answer, which is the *probability of a hypothesis* (H) given the speech evidence (E) or P(H/E), is seen as a *logical flaw* as it transposes the conditional probabilities of the LR (Equation 1 above) (Robertson & Vignaux, 1995, pp. 19-20). That is, the probability of evidence under the competing hypotheses P(E/H) is transposed to the probability of hypotheses given the evidence P(H/E). The experts should not be giving such categorical answers about guilt or innocence.

With FVC, if an expert tries to give such a categorical answer, which is the *probability of* a hypothesis given the speech evidence or P(H/E), he or she needs to take into account all information relevant to the case (which is usually not known by experts). This, however, would be seen as a legal usurpation (Morrison, 2009a). However, we need to be aware that such a categorical decision is possible in real casework if FVC experts conduct a closed-set comparison and when there are speech samples that sound very dissimilar (e.g. originating from different sexes) (Kinoshita, Ishihara, & Rose, 2009, p. 103). I will now discuss the last difference between speaker verification and FVC.

#### 3) Control over samples

In speaker verification, there is a very good level of control over reference data, i.e. speakers are asked to read prescribed texts, which are stored and retrievable as templates from an automatic speech recognition system (Rose, 2002, pp. 90-91). Such templates contain high speaker-specific parameters (this is done deliberately) (Broeders, 1995, p. 156). Additionally, those doing the tests have a high degree of control (high degree of comparability) over the samples being tested. That is, speakers who wish to be verified are cooperative with the test and repeat the phrases of any desired reference templates (ibid.). For FVC, in contrast, the questioned samples might be incriminating speech recorded during a robbery, whereas the speech samples obtained from suspects might be from a police interview and intimidating. Asking suspects to utter the same incriminating lines of text might be construed as a means of obtaining comparability. However, this is not the case as no one ever says the same thing twice in a way that has exactly the same acoustic/physical properties (Rose, 2002, p. 10).

Having presented the three main differences (reference data, categorical decision making and threshold, and control over samples) between speaker verification and FVC, I will now discuss the motivations for conducting this thesis.

#### 1.5 Motivations

In Thailand, like in other countries, many crimes are committed using mobile phones. Criminal cases in which digital forensic evidence obtained from the mobile phone carrier led to an arrest of suspects are now increasingly reported (Ngamkham & Nanuam, 2015). Unfortunately, to the best of my knowledge, no legal cases in which voice evidence was admitted in court have been publicly reported in the media. This does not mean that the Thai authorities are not interested in voice evidence in criminal proceedings. Indeed, I myself was asked via e-mail, in February 2015, by a defendant's lawyers, to proffer voice evidence to a court in Chiang Mai, Thailand. The defendant, who had been accused of bank fraud, perjury, and extortion, hoped that expert opinions concerning voice evidence might be of value for the adjudication of their case. Such an FVC demand suggests that the number of FVC experts needed in Thailand is growing. Thus, my first motivation for conducting this research is to help fulfill this requirement.

My second motivation for conducting this thesis and choosing Standard Thai is my background as a Thai native speaker. Being a linguist who has the ability to comprehend the forensic samples under investigation yields many advantages. As Rose (2002, p. 333) explains, a native speaker automatically knows which sounds realize which phonemes, as well as what constitutes the typicality of a language peculiar to a community. This means that as a trained phonetician, who is a speaker and listener of the language under examination, I am able to interpret the speech complexity that is normally to be found in forensic samples. Another reason for choosing Standard Thai is that it is standardized and used nation-wide in Thailand, i.e. it is taught in educational institutions, used in the media, and described by grammar books and dictionaries (Tingsabadh & Abramson, 1993). As such, it can be said that Standard Thai is spoken by the majority of people in Thailand (approximately 20 million, as compared to the 15 million, 6 million and 4.5 million natives of Thailand who speak the Northeastern, Northern and Southern Thai dialects, respectively) (Lewis, Simons, & Fennig, 2013).

Given that there has been conspicuously little FVC research done on Standard Thai, except that of Thaitechawat and Foulkes (2011), the third motivation for undertaking this thesis is to further empirically test the applicability of the LR framework to that language.

Based on these three motivations, I conduct the current work in order to search for the specific parameters that might potentially contain high speaker-specificity, and that are of forensic use in Standard Thai. Thus, it is appropriate for me to now discuss the approach employed in the current thesis.

#### 1.6 The research approach

In this section, I aim to make clear the research approach employed in the current thesis, by illustrating the differences in parameters, on the one hand, and the statistical modelling techniques used, on the other.

#### 1.6.1 Traditional vs automatic parameters

To begin with, the current thesis employs the traditional approach, which comprises both auditory and acoustic analyses. This means that, before proceeding to an acoustic analysis, forensic experts will first listen to incriminating speech samples of unknown origin, taken for instance from recordings embedded in CCTV footage, and speech samples of suspects, i.e. of known origin, taken for instance from conversations during a police interview, to judge the quality and the comparability of the speech recordings (in an attempt to extract comparable words, phrases, etc). The auditory-acoustic approach was revealed as the most popular both in an INTERPOL survey of the use of speaker identification by law enforcement agencies, by Morrison, Sahito, et al. (2016), and in a survey on forensic speaker comparison, by Gold and French (2011). The two surveys were different in terms of the use of differing types of respondents: all the respondents of the former survey were from law enforcement agencies, but half of the latter were from government labs, while the rest were academics and private practitioners. The other difference between the two surveys is that the former revealed the usefulness of FVC both for the investigative as well as for forensic applications, while the latter reported the usefulness of FVC for forensic applications only.

As Rose (2002) and Enzinger (2009) point out, the choice of parameters used in FVC is in part language-specific. Thus, conducting FVC research on Standard Thai is also

justifiable due to the need to add to our existing knowledge of the extent to which Standard Thai, which is a tonal language, lends itself to FVC as compared to other languages, with different parameters found to contain highly individualizing information. For the purposes of this thesis, various segments, consonantal (/s/, /te<sup>h</sup>/, /n/, /m/) as well as vocalic (diphthongs [5i], [ai]), were selected, from which different acoustic information (the spectrum, fundamental frequency (F0) and formant trajectories) was extracted depending on the segments. These parameters were chosen based on promising results reported in the relevant literature, as I will show in detail in Chapter 2. However, to give an overview, the traditional parameters, which are the F0 and formant trajectories, will be tested in this study because they have been proven to work well in many languages, including Standard Thai, cf. Thaitechwat and Foulkes (2011). Furthermore, they are directly correlated to articulatory and auditory phonetic features (Rose & Clermont, 2001). Apart from the traditional parameters, spectrum, which is one of the most popular parameters in ASR, will also be tested (Franco-Pedroso, Gonzalez-Rodriguez, Gonzalez-Dominguez, & Ramos, 2012). Spectrum is one way of looking at speech waveform in terms of the amounts of energy that are present at particular frequencies, a so-called "frequency-domain representation" (Rose, 2002, p. 199). Spectrum is relatively easy to extract as opposed to other traditional linguistic parameters such as formants. As such, the spectrum of /s/, in particular, has been chosen for the current experiment as it has recently been reported to have promising FVC results in English by Kavanagh (2012). Spectrum extracted from nasals will also be investigated, as it is said that nasal spectrum contains considerable amounts of individualizing information (Enzinger & Zhang, 2011; Glenn & Kleiner, 1968; Su, Li, & Fu, 1974; Wolf, 1972; Yim & Rose, 2012).

Following on from the above, it can be concluded that the current research is *traditional*, in the sense that traditional acoustic parameters from the same comparable phonemes are tested (Rose, 2011, p. 5900). However, the spectrum-based features, which are common in automatic speech recognition, are also trialed in the current work in order to see whether either automatic or traditional parameters perform better in testing.

#### 1.6.2 Statistical modelling techniques

The statistical modelling technique used in the current thesis is the multivariate likelihood ratio (MVLR), which is more often used to calculate LR in semi-automatic FVC than the GMM-UBM (Gaussian Mixture Model - Universal Background Model) (Rose, 2002, p.

89). Having said that, the MVLR may also be used with automatic parameters, e.g. Melfrequency cepstral coefficients (MFCCs), and the GMM-UBM with traditional parameters (cf. Zhang, Morrison, & Thiruvaran, 2011) – but this possibility is not being pursued here. Thus, the current thesis adopts *semi-automatic* principles: it requires human experts to listen to and select speech segments that are of use for FVC analysis. Furthermore, the experts themselves interpret the FVC results and present their testimony to a court of law (ibid.).

#### 1.7 Linguistic-phonetic segments and acoustic parameters

Many scholars (e.g. Boves, 1998; Andrews, Kohler, Campbell, & Godfrey, 2001; Doddington, 2001) have argued that significant improvements to *automatic speaker recognition* hinge on the discovery of acoustic parameters that contain highly individualizing information. This applies to *forensic voice comparison* as well, where forensic experts are expected to calculate, with as high a level of validity as possible, the probability of observing speech evidence under competing hypotheses. Focussing on Standard Thai, this thesis aims to identify which of its linguistic segments (consonants, vowels and tones/tonal F0) perform best in FVC and which of their acoustic parameters (both traditional and automatic) potentially contain the most highly individualizing information.

It is of course beyond the scope of any single study to examine all of a language's phones/allophones and potential parameters available for FVC. The selection of various acoustic parameters (F0, format trajectories, spectrum) in this thesis is motivated, as previously discussed, by the fact that Standard Thai is a tonal language and that promising results for tonal F0 and formant trajectories of Standard Thai were reported in Thaitechawat and Foulkes (2011) (100% correct classification rate using discriminant analysis (DA)). Spectrum, too, is selected for testing in this thesis, in addition to traditional linguistic-phonetic parameters, as promising results have long been reported in automatic speaker recognition (ASR) literature that details the use of this method (cf. Franco-Pedroso et al., 2012).

Table 1 (overleaf) shows the specific acoustic parameters explored in this thesis; it is categorized according to linguistic-phonetic segments. Table 1 shows that, in this thesis,

I explore various acoustic parameters, extracted from different linguistic-phonetic segments. Since the spectral moments (mean, variance, skew and kurtosis) of the English consonants /s, m, n, n/ have been found by Kavanagh (2012) to contain promising

Acoustic parameters	Linguistic-phonetic segments
1. Spectrum 1.1 Spectral moments (mean, variance, skew, kurtosis) 1.2 Cepstral coefficients (CCs)	1. /s/, /te <sup>h</sup> /, /n/, /m/ 2. /s/, /te <sup>h</sup> /, /n/, /m/
2. Fundamental Frequency (F0) 2.1 Tonal F0 2.2 LTF0 (Long-term F0)	2.1 Diphthongs [ɔi], [ai] 2.2 Fax task
3. Formant trajectory 3.1 Formant trajectories (F1-F3)	3.1 Diphthongs [ɔi], [ai]

**Table 1:** Acoustic parameters and linguistic-phonetic segments

individualizing information, it will be prudent to test FVC performance in Standard Thai with reference to the fricative /s/, the affricate /tch/ and the nasals /n/ and /m/ (I will discuss how to select these linguistic-phonetic segments in greater detail in Chapter 2). Apart from spectral moments, the ceptral coefficients (CCs) approximating the spectral shape of /s/, /te<sup>h</sup>/, /n/, and /m/ are also trialed. Additionally, based on the promising results of the discriminant analysis (DA) of Standard Thai tones and formants by Thaitechawat and Foulkes (2011), the traditional parameters, constituting the tonal fundamental frequency, together with the formant trajectories (F1-F3) of the diphthongs [5i], [ai], will be tested as well. These fundamental frequency (F0) and formant trajectories will be extracted using the polynomial coefficients. I will proceed based on the research findings achieved, among others, by McDougall and Nolan (2007) and Zhang, Morrison, Ochoa, and Enzinger (2012), which suggest that these formant trajectories contain more individualizing information than the static formant values measured at the temporal midpoint of a vowel. In order to find additional F0 features, which might be of FVC use, the distribution of long-term fundamental frequency (LTF0) extracted from an information exchange task (two informants having a conversation based on obfuscated information given in a fax message) will also be trialed in order to see how it performs in Standard Thai.

#### 1.8 Research questions

Following on from my discussion of the acoustic parameters and linguistic-phonetic segments listed in Table 1, the first aim of my thesis is to examine to what extent such acoustic parameters and linguistic-phonetic segments work in Standard Thai FVC. Thus, the following are the specific research questions I pursued in this study.

- 1.1 How well does the spectrum extracted from the consonants /s/, /tch/, /n/, /m/, and modelled using two different techniques, firstly by means of the so-called spectral moments (mean, variance, skew, kurtosis) and secondly by means of the coefficients of a discrete cosine transform (DCTs), perform in Standard Thai FVC; and which parameterization techniques perform better?
- 1.2 How well do the diphthongs [5i] and [ai]'s tonal F0 contours and the first three formant trajectories modelled by polynomials perform in Standard Thai FVC and which diphthong performs better?
- 1.3 How well do the six long-term F0 (LTF0) parameters that relate to the shape of the F0 distribution (1. mean; 2. standard deviation (SD); 3. skew; 4. kurtosis; 5. modal F0; and 6. modal density) perform in Standard Thai FVC and which LTF0 parameter performs better?

The second aim of this thesis is, through an interpretation of fusion results (whereby two or more parameters are fused or combined), to further add to the research findings pursued in the current work. The specific question is:

2. How can the linguistic-phonetic segments tested in this thesis be profitably combined?

#### 1.9 Thesis outline

In Chapter 2, I present an overview of the existing literature relating to Bayes' theorem. I will also review the effectiveness of forensic voice comparison (FVC) and the so-called paradigm shift to provide background knowledge about a Likelihood Ratio (LR) framework. Then, I will discuss Standard Thai sound systems. Many previous studies on FVC, in particular the literature on which I based my decision to select the specific parameters for Standard Thai, will be summarized.

In chapter 3, the statistical tools that I employed to calculate LRs, i.e. the multivariate likelihood ratio (MVLR), and to assess the performance of the FVC system, i.e. log-likelihood ratio cost ( $C_{llr}$ ), will be extensively discussed. Then, I will describe the speech corpus design, which involves the selection of informants and the elicitation method. Finally, I review the basic concepts of logistic regression calibration and fusion.

In chapter 4, three pilot studies on 1) the Standard Thai (phonetic) diphthongs [i:aw], [u:a] and [u:a], 2) the Standard Thai (phonetic) diphthongs [o:i] and [o:i], and 3) the Standard Thai (phonetic) diphthongs [ai] and [u:a], all of which are conducted as parts of the current work, are presented to justify the use of the selection of traditional parameters (F0 and formant trajectories).

In chapter 5, I illustrate the annotation of the target segments /s, te<sup>h</sup>, n, m/ and diphthongs [5i], [ai]. I review the basic concepts of mean, variance, skew, and kurtosis (also known as *spectral moments*). I present the results of the FVC experiments with /s, te<sup>h</sup>, n, m/. This is followed by a discussion about the linguistic-phonetic in relation to the forensic perspectives.

In chapter 6, I present and discuss the results of the FVC experiments with the diphthongs [5i] and [ai], with a focus on formant trajectories.

In chapter 7, I first present and discuss the results of the FVC experiments with the diphthongs [5i] and [ai], with a focus on the tonal F0. This will be followed by the presentation and discussion of the LTF0.

In chapter 8, I present the answers to the research questions and the overall findings of the thesis. Finally, I outline opportunities for future research, which can build upon the current work.

## 1.10 Summary

In this chapter, I have introduced the background of the current study, including the scope and the details of what will be covered in the thesis. The primary purpose of this chapter was to make sure that the term 'forensic voice comparison' or FVC is clearly explained from the outset.

## Chapter 2

#### Literature review

#### 2.1 Introduction

In this chapter, I will first explain the less-than-ideal factors that make voices difficult to discriminate forensically. I aim here to explain how forensic experts should deal with speech evidence, and how such speech evidence should be integrated into the case at hand. Then, the justification of the Bayesian theorem, specifically the LR framework, for the current research will be extensively discussed. After that, I introduce Standard Thai sound systems as well as those of other Thai dialects. Then, I will review previous FVC studies employing the spectral moments (mean, variance, skew, and kurtosis), coefficients of the Discrete Cosine Transform (DCTs), tonal F0, LTF0, and the formant trajectories, whilst introducing and explaining the basic acoustic knowledge that is required for understanding the results of this thesis.

#### 2.2 Ideal features of forensic scientific evidence

In an ideal forensic scientific system, individuals can be easily discriminated because of features that are: 1) unique (individuals can be distinguished based on these features); 2) unambiguous; 3) more or less probable with the features than without ("able to place individuals at a crime scene"); 4) unchanging; and 5) relatively easy and economical to operate (Robertson & Vignaux, 1995, p. 6). However, it is difficult to find such features in evidence in real-world circumstances. This is because pieces of evidence supposedly constituting ideal features may satisfy some but not others (ibid.). As for the fourth ideal feature just described, the suspect's speech may sound hesitant during a police interview but the offender may sound aggressive or provocative in CCTV footage showing an incriminating act. Such different communicative settings, involving two speech samples with different sounding emotions and different background noise levels (loud vs quiet), and recorded at a different time of the day, may show different F0 values due to differing acoustic properties. Anger in a voice and loud background noise will increase F0 levels to a different degree (Braun, 1995; Klasmeyer & Sendlmeier, 2013; Laukkanen, Vilkman, Alku, & Oksanen, 1996; Rose, 2002; Williams & Stevens, 1972). Likewise, F0 levels rise

from morning to afternoon (Garrett & Healey, 1987; Rantala, Vilkman, & Bloigu, 2002). Therefore, voice quality will change depending on various internal and external factors. In §2.2.1, I summarize the two main variations that voices inherently carry, i.e. withinand between-speaker variations, which make speech evidence difficult to discriminate forensically. Besides variations, I will explain the issue of dimensionality (§2.2.2).

#### 2.2.1 Lack of control over variation

In forensically realistic conditions, the first factor that makes speech samples difficult to discriminate forensically is their level of variation. As Rose (2002, p. 19) points out, variation is typical in voices, and this variation always occurs both within and between speakers. One example of within-speaker variation is when two speech samples obtained from the same speaker, uttered a few seconds apart, are found to be acoustically different. This is also true for speech samples uttered by different speakers, i.e. in the case of between-speaker variation (ibid.).

#### 2.2.1.1 Between-speaker variation

Voice differences produced by different speakers are termed *between-speaker variation* (Rose, 2002, p. 10). In this section, I discuss the factors that make different speakers speak differently with reference to a range of different models of sources of between-speaker variation found in the literature. As shown below, between-speaker variation is characterized differently by various scholars:

Differences in voices stem from two broad bases: organic and learned differences (Wolf, 1972, p. 2045)

Organic and learned differences are the sources of intertalker variability (Tosi, 1979, p. 55)

We can tentatively categorise speaker-diagnostic variables in terms of two basic distinctions: organic versus acquired or learned, and individual versus group (Garvin & Ladefoged, 1963, p. 194)

Acoustic parameters of speech reflecting speaker identity must be derived either from the unique physiological characteristics of the speaker's vocal apparatus or from idiosyncrasies in his manner of speaking

(Glenn & Kleiner, 1968, p. 368)

Given the models quoted above, it is easy to understand that different people have different voices, due to: 1) anatomical or organic differences, i.e. differently sized vocal

cords and different shape of the vocal tract; and 2) learned/acquired differences in their manner of speaking – how a speaker habitually speaks, i.e. a speaker's *idiosyncratic* usage. With respect to organic differences, the length and mass of vocal cords in women is generally shorter and lighter than in males (Stevens, 2000, p. 5), resulting in the higher pitch of female voices (180-300 Hz) compared to male voices (90-140 Hz) (Rose, 2003, p. 4102).

#### 2.2.1.2 Within-speaker variation

It could be argued that if someone were to constantly speak in the same way, it would naturally be easier to identify that individual. However, as Rose (2002, p. 10) points out, no one ever speaks in exactly the same way twice. That is, there are factors that cause even the same individuals to speak differently; this is called *within-speaker variation* (ibid.). It follows, then, that FVC is feasible on condition of *small* within-speaker variation and *large* between-speaker variation. Otherwise, FVC evidence would be near to useless, because within- and between-speaker variations are equally likely.

I will now discuss some of the factors that cause within-speaker variation. The factors listed in Table 2 are typical causes of within-speaker variation.

#### Within-speaker variation

- Emotions
- Linguistic message
- Social settings
- Health
- Elapsed time between recordings
- Technical factors

Table 2: Sources of within-speaker variation

(Based on Wolf 1972, p. 2045; Tosi, 1979, p. 55; Garvin & Ladefoged, 1963, p. 194; Glenn & Kleiner, 1968, p. 368; Nolan, McDougall, De Jong, & Hudson, 2006; Butterfint, 2004; Loakes & McDougall, 2004)

I will begin with emotional factors. When a speaker feels sorrow, F0 levels will decrease considerably by up to 30 Hz; when a speaker is angry, his/her F0 level will increase (Johnstone & Scherer, 2000; Paeschke, Kienast, & Sendlmeier, 1999; Williams & Stevens, 1972). Second, the F0 level or a perceived pitch by a listener will also change as

a function of the linguistic message. That is, there is a falling pitch in the word studying in the English statement: "She is studying", but there is a rising pitch in the last word of the English yes/no question: "Is she studying?" (Rose, 2002, p. 19). Moreover, F0 tends to be higher for reading tasks than for spontaneous speaking (Braun, 1995, p. 17). This is important to remember when conducting FVC analysis. That is, the FVC experts should select speech samples that are comparable in terms of syntactic, semantic, idiomatic and stylistic elements. This should be done to ensure comparability, because different speaking styles, among other factors, may result in different acoustic properties in a voice. Third, same-speaker voices may change as a function of the interlocutors, i.e. speakers tend to accommodate their speech in social interaction with their interlocutors (Rose, 2002, p. 20). Such examples can be found in everyday conversation. For example, when we talk to a child, our voice may, in some situations, tend to be high-pitched to make our speech sound like that of the child. However, when we give a presentation in front of a classroom, our voice may become low-pitched to convey a sense of discretion. This kind of pitch convergence should be considered when conducting FVC analysis, to ensure comparability remains valid (ibid.). Fourth, a speaker's state of health may also cause within-speaker variation. It has been found by Klingholz, Penning, and Liebhardt (1988) and Schiel and Heinrich (2009) that moderate levels of intoxication increase the F0 standard deviation by up to 100%. Additionally, the influence of legal drugs for use in the treatment of cancer as well as steroids and androgens used in the treatment of femaleto-male transsexuals are found to lower voice pitch (Braun, 1995, p. 15). Stress is also found to affect F0 levels quite consistently with regard to within-speaker variation (but it affects between-speaker variation levels differently) (Hecker, Stevens, von Bismarck, & Williams, 1968). Fifth, when the elapsed time of one of the two recordings is longer than the other, the within-speaker speech samples will sound different to those lasting a shorter period of time (Rose, 2002, p. 20). Given the facts just described, it would be prudent to conduct FVC research with non-contemporaneous speech (speech recorded in two sessions separated by at least a week or up to two months for the current work) when testing the accuracy and reliability of the FVC system for Standard Thai.

Lastly, there are technical factors that may cause same-speaker speech samples to sound different (Künzel, 2001; Rose & Simmons, 1996). They include the disguising of one's voice, either through a lowering/raising of F0 level, or a change in register – language variation according to use as defined by Halliday, McIntosh, and Strevens (1964), e.g.

modal vs falsetto (Braun, 1995). Moreover, differences in measurement/analysis process such as the different quantizations or differing tape recorder speeds are also found to cause F0 level differences (Braun, 1995, p. 10).

# 2.2.2 Reduction in dimensionality

The second effect of real-world conditions on FVC is a reduction in dimensionality or reduction in dimension number (Rose, 2002, p. 25). By "dimension number", or dimensionality, I mean the number of (acoustic) dimensions per speech sample (ibid.). The number of dimensions considered for the purposes of FVC is inherently variable and "not all dimensions are equally powerful" (Rose, 2002, p. 16). However, if a decision were made to assess a series of speech samples in terms of three dimensions, e.g. the F0 value of one of its vowels, for example /e/ (Dimension 1), the F-pattern of one of its diphthongs, for example /ia/ (Dimension 2), and one of the spectral moments of a nasal, for example /n/ (Dimension 3), then one would expect, in ideal circumstances, all samples to be three-dimensional. In practice, though, the suspect's speech samples may lack the vowel /e/, which makes them quantifiable only on Dimensions 2 and 3, whereas the offender's speech samples may lack diphthongs, which makes them quantifiable only on Dimensions 1 and 3. According to Rose (2002, pp. 21-22), a reduction in dimensionality means that some powerful features or parameters that are of forensic use might not be available due to many factors. Apart from a reduction in dimensionality, the distortion of dimensions is another real-world condition that makes voices difficult to discriminate forensically (ibid.). An example of this is telephone signals that are degraded by a telephone line or distorted by 1) technical factors such as a low-quality tape recording and 2) the effect of echoic rooms and background noise (ibid.). To put it another way, forensic samples are never available in optimal conditions due to various inherited and/or unavoidable reasons (ibid.). Rather, an expert should be aware what factors affect different dimensions and which dimensions are more resistant to distortion (ibid.). The last thing to consider, though it is of no less importance than the factors listed above, is the time available for forensic experts to evaluate as many potential features as possible (ibid.). Some powerful features, even though they contain much individualizing information, might need to be excluded from the analysis because of time constraints.

# 2.3 Speech variation and the role of probability theory

Given the within- and between-speaker variations just described, we can see that forensic experts need to find ways of properly dealing with these kinds of variations/uncertainties. To put this in context, when speech evidence is tendered to a court, interested parties (police/prosecution, trier-of-fact) want to know whether the speech recordings of known and unknown origin were produced by the same speaker or not. In the Thai legal context (which will be explained in more detail in §2.5), the courtrooms may request a person who has specialized knowledge beyond that of the courts, i.e. a so-called *expert witness*, to give his/her testimony related to the matters at hand (Wannasaeng, 2008). The role of an expert witness is to provide a testimony to help the courts draw certain inferences to reach their conclusion, as opposed to a testimony that will lead to an accusation being made against an offender or to the elimination of a suspect (ibid.). Of course, a hypothesis such as whether given speech recordings were produced by the same speaker or not can be either true or false; no one can be sure about its truth (Robertson & Vignaux, 1995, p. 13). One of the best ways to deal with such uncertainties is through the use of probability models (Aitken & Stoney, 1991), which are a "rational measure of the degree of belief in the truth of an assertion based on the information" (Robertson & Vignaux, 1995, p. 14). In the current thesis, the probability model that will be used to deal with speech variations is called the Bayesian theorem or Bayes' theorem (I will discuss this in detail in §2.4). Before going further, I will explain how I interpret the word *probability* in the current thesis.

Probabilities are often interpreted as either 1) frequencies of repeated or long-running events (for example: What is the probability that we will get heads after tossing a fair coin infinitely many times? As probability scores are between 0 and 1 inclusive, with 1 meaning that an event is certain to happen while 0 means that an event is excluded from happening, the answer is 0.5); or 2) descriptions of beliefs (describing our degree of belief about uncertain situations); or 3) betting preferences (what kind of bets we are willing to make) (Bertsekas & Tsitsiklis, 2002; Murphy, 2012, p. 28).

As Bertsekas and Tsitsiklis (2002) point out, whether a prediction will be any good or not depends on the probability model that we choose to employ. My interpretation of the probability theory employed in this thesis, called the *Bayesian interpretation*, is that it can *quantify our degree of belief about uncertain situations* (Robertson & Vignaux, 1995,

p. 17). In the Bayesian interpretation, the above example of coin tossing will be interpreted like this: "it is equally likely that the coin will land on heads or tails in the next toss" (Bertsekas & Tsitsiklis, 2002; Murphy, 2012, p. 27). Bayes' probability model provides a systematic way of thinking about and describing our beliefs concerning uncertainty, which is one of its big advantages (Murphy, 2012, p. 27), particularly where evidence is not obtainable in a numerical format. In order to evaluate the degree of our belief in a hypothesis such as: "Speech samples from the suspect were in fact made by the offender", which is true or false (we cannot be sure either way), we need to know about the similarity and typicality of the offender's and the suspect's speech samples. The ratio of similarity (between the offender's and the suspect's speech samples) and typicality (how likely it is that speech samples like those of the suspect or offender are to be found by chance in a relevant population sample) is called a *likelihood ratio* or *LR*; it allows the strength of voice evidence to be evaluated. §2.4 introduces the framework of the Likelihood Ratio used in this study.

# 2.4 What is the Likelihood Ratio?

As previously mentioned, in a typical FVC scenario, a recording of incriminating speech samples, e.g. a CCTV recording of a telephone bomb threat from an unknown speaker (offender), is compared with recordings from one or more known speakers; the latter are being assessed as potential suspects whose speech samples are obtained, for instance, during a police interview. The task of forensic experts is to calculate the *strength of this evidence* or *LR* (Robertson & Vignaux, 1995).

In order to do so, two probabilities are taken into account. First, the probability of the similarity/difference between the suspect's and the offender's speech samples, given the same-speaker hypothesis (prosecution hypothesis). Second, the probability of observing the same evidence (the similarity/difference between the speech samples) under the different-speaker hypothesis (defense hypothesis) (Rose & Morrison, 2009, p. 6). The ratio of these two probabilities is called a *likelihood ratio* or *LR*, and is part of Bayes' theorem (Robertson & Vignaux, 1995, p. 17), which is shown below:

#### Bayes' theorem

$$\frac{p(Hp|E)}{p(Hd|E)} = \frac{p(Hp)}{P(Hd)} \times \frac{p(E|Hp)}{p(E|Hd)}$$
Posterior odds Prior odds Likelihood ratio

In this theorem p stands for probability, Hp for prosecution hypothesis, Hd for defense hypothesis, and E for (speech) evidence.

It is this quantified LR that creates a logical link between the measurement of speech similarity/difference (E) and a competing hypothesis (H). That is, we calculate how much more likely such speech evidence, which is based on the similarity/difference between the offender's and the suspect's samples, is likely to be from the same speakers p(E/Hp), rather than from different speakers p(E/Hd) (Rose & Morrison, 2009, p. 6).

I will now explain in more detail (from right to left of the equality sign) what Equation 2 means. The *prior odds*, which is the ratio of the probabilities of two competing hypotheses (the probability that the same-speaker hypothesis is true is divided by the probability that the different-speaker hypothesis is true), are considered in forensic science before taking the evidence into account (Robertson & Vignaux, 1995, p. 17). Calculating prior odds is the purview of the trier-of-fact because these prior odds depend on the initial assumptions plus the changes of belief in the probability of the hypotheses, based on the evidence already presented (Morrison, 2009a, p. 300). As such, the priors are usually not known by forensic experts (ibid.), and thus, when following Bayes' theorem, they cannot calculate the posterior odds (posterior odds = prior odds x LR) (Li & Rose, 2012). In stark contrast, forensic experts can calculate the *likelihood ratio*, which is the probative value or the strength of evidence in favor of the hypothesis (Robertson & Vignaux, 1995, p. 21). As Morrison (2009a, p. 300) stated, the task of forensic experts is to limit themselves the calculation of the likelihood ratio, which is the ratio of speech similarities/differences being compared, while taking into account their typicality with reference to the speech of a relevant population, "from which the true perpetrator of the crime could conceivably have come", Morrison, Ochoa, and Thiruvaran (2012, p. 64). Rose (2002), Nolan et al. (2006) and Loakes (2008) further explain that the suspect's (or suspects') speech samples should be similar-sounding or "at least not too differentsounding", in terms of sex, language, and accent, to those of the offender (Rose, 2002, p. 97). When the *prior odds* and the *likelihood ratio* are combined using Equation 2, it gives us the *posterior odds* – the posterior probabilities of a guilty versus not guilty hypothesis (ibid.). The posterior odds are what the court would like to know.

Prior odds can have a substantial effect on the strength of evidence because a different prior odd will yield a different posterior odd (Robertson & Vignaux, 1995, p. 18). In this section, I will try to illustrate in more detail why this is the case. Imagine, for example, that two people (including the suspect) were recorded by CCTV cameras carrying a backpack in the vicinity of a crime scene right before a bomb blast. In this case, the probability of the hypothesis that the suspect is the offender will be 1/2 or 0.5. This probability can be converted into a prior odd "by dividing the probability by one minus the probability" (Rose, 2002, p. 63), or, in mathematical terms, 0.5 / (1 - 0.5) = 0.5 / 0.5= 1, which means that the hypothesis that the suspect is the offender is as likely to be true as the defense hypothesis (that the suspect is not the offender). When this prior odd is combined with a LR of, say, 10, the posterior odd is (1 x 10 =) 10. At this point it is possible to say that it is 10 times more likely that the suspect is the offender. In contrast, when for example five males (including the suspect) have been recorded by CCTV cameras instead of two, the probability of the hypothesis that the suspect is the offender is reduced to 1/5 or 0.2, which is a prior odd of (0.2 / 1 - 0.2 =) 0.25. When this is combined with an LR of 10, it gives a posterior odd of (1/4 x 10 =) 2.5, which in turn gives less support (two point five times less, to be precise) to the hypothesis that the suspect was the offender than in the previous example (ibid.). Given the example just described, we see that the prior odds have an important effect on the posterior odds (i.e. they give more or less support to the hypothesis that the suspect is an offender), even though the LR is the same (ibid.).

Typically, in an FVC scenario, a police officer (a layperson with respect to FVC) will listen to the recordings of speech samples from the offender and decide whether they sound sufficiently similar to those of a particular suspect (Morrison, Ochoa, & Thiruvaran, 2012, p. 64). If they do sound sufficiently similar, the police officer will submit the two recordings (of speech samples of the suspect and the offender) for further FVC analysis (ibid.). A same-speaker hypothesis is then generated (ibid.). In contrast, if the police officer thinks that the speech samples of an offender and a suspect do not sound

similar enough, he or she will not submit the two recordings to FVC experts and a samespeaker hypothesis will not be generated (ibid.).

Apart from prior odds, which have a substantial effect on the calculation of scientific evidence, the selection of an appropriate alternative or defense hypothesis (denominator of LR) also has an enormous effect on LRs or the strength of evidence (Robertson & Vignaux, 1995, pp. 33-50). This is because the alternative or defense hypothesis (which can take many forms), and which directly relates to the selection of background samples (the denominator of LR), changes the prior odds (Morrison, Ochoa, & Thiruvaran, 2012, p. 64). Usually, a defense hypothesis is generated as: "the speech samples are from some other speaker" (Morrison, Ochoa, & Thiruvaran, 2012, p. 64). Preferably, the defense hypothesis should be more specific than simply stating: some other speaker, in order to define a relevant population group (ibid.). For example, if the defense hypotheses were that 1) the speech samples are from other male speakers with high-pitched voices on a particular island and 2) it is a suspect's brother, the relevant population or background speech samples would need to be changed accordingly. That is, under the first defense hypothesis, the recordings of male speakers with high-pitched voices on a particular island will be selected; under the second defense hypothesis, the recordings of a suspect's brother will be considered.

Intuitively, the posterior odds will significantly increase for the second defense hypothesis (it is a suspect's brother), as the prior odds are bigger (even odds) than the former (100/1 against it being male speakers with high-pitched voices on a particular island). That is, with the defense hypothesis that the suspect's brother made the call, for example, the posterior odds would be derived by timing a prior odd of 1 with an LR value of say 50, giving posterior odds of 50 in favor of a hypothesis that the suspect's brother's voice is indeed the one heard in the incriminating speech samples. However, under the defense hypothesis that other male speakers with high-pitched voices, on a particular island, made a telephone call, let us scale up the LR value accordingly to 100. Then the posterior odds become  $(1/100 \times 100 =) 1$ , suggesting that it is equally likely that every male with a high-pitched voice on a particular island is responsible for the incriminating speech samples. Given these examples, we can appreciate that the selection of a defense hypothesis can alter the strength of voice evidence (Morrison, Ochoa, & Thiruvaran, 2012, p. 64).

To interpret derived numerical LRs, it is common practice to use  $\log_{10}LR$ , where unity is set at 0 (Rose, 2006). The greater the deviation from unity either way, the greater the strength of evidence supporting either the prosecution or the defense hypotheses (Robertson & Vignaux, 1995, p. 17). In contrast, if a  $\log_{10}LR$  reading is close to or equal to unity, the examined evidence provides only limited support for either hypothesis and is thus regarded as useless or unhelpful (ibid.). Since LRs or the strength of evidence are assessed using numeric values, verbal equivalents of such LR values (both linear LRs and  $Log_{10}LRs$ ) have been proposed, among others, by Champod and Evett (in Rose, 2002, p. 62). The verbal equivalents in Table 3 make the experts' numerical analyses and interpretations more understandable for the court (ibid.). Linear LRs and their corresponding  $Log_{10}LRs$ , supporting both the prosecution and defense hypotheses, are presented in the first and second columns, respectively, whereas the verbal equivalents are shown in the third column.

Likelihood Ratio	Log <sub>10</sub> Equivalent	Possible interpretation	
> 10 000	> 4	Very strong	
1000 to 10 000	3 to 4	Strong	
100 to 1000	2 to 3	Moderately strong	support for
10 to 100	1 to 2	Moderate	the prosecution hypothesis
1 to 10	0 to 1	Limited	
1 to 0.1	0 to −1	Limited	
0.1 to 0.01	−1 to −2	Moderate	
0.01 to 0.001	0.001 –2 to –3 Moderately strong		support for the defense hypothesis
0.001 to 0.0001	-3 to -4	Strong	the defense hypothesis
< 0.0001	> -4	Very strong	

**Table 3:** Verbal equivalents of LRs (Adapted from Champod & Evett, 2000, p. 240)

Before I go further, I should point out that there is no consensus agreement on the use of the scaling of LRs (whether linear or log LRs). Although it is more intuitive than a linear LR of 1000, which, instead of 3, should be used to reflect "moderately strong evidence" (Rose, 2002, p. 62), there are advantages in using a common log LR over a linear LR. A linear LR = 1, suggests *useless* evidence and since a common log of 1 is 0, a threshold of  $Log_{10}LR = 0$ , rather than LR = 1, is preferable as the value of zero is

naturally linked to the notion of its "worthlessness" (Rose, 2002, p. 62). Moreover, a common log LR is preferred because it is relatively easier to use when combined with LRs obtained from other independent pieces of evidence (ibid.). For example, a log of  $10^4$  (LR of 10000) and a log of  $10^3$  (LR of 1000) can be added in the form of a common log  $10^4$  + log  $10^3$ , giving the overall log LR =  $10^7$  (this is equivalent to multiplying an LR of 10000 with a LR of 1000, which also gives the overall LR = 100000000 or log  $10^7$ ).

I will now discuss how to read the numeric LR values. Ideally, when two speech samples produced by the same speakers are compared, the same-speaker comparisons (or SS comparisons) should yield log LRs greater than 0. In addition, when two speech samples produced by different speakers are compared, those different-speaker comparisons (DS comparisons), should be lower than 0 (Rose, 2002, p. 62). As such, if speech evidence with  $log_{10}LR > +4$  is tendered to the court, it can be translated into "There is very strong evidence to support the prosecution hypothesis that the speech samples are more likely from the same speaker than from different speakers" (Rose, 2002, p. 62). In contrast, if  $log_{10}LR < -4$  is presented to the court, it is interpreted as "There is very strong evidence to support the defense hypothesis that the speech samples are more likely to come from different speakers than from the same speaker" (ibid.). It should be noted from Table 3 that the mathematical symbol used to indicate a negative value that is smaller than  $log_{10}LR = -4$  is > (greater than), not < (less than). This means that the pure magnitude or size of a number (after the negative sign) is taken into account. Since it is difficult to explain how to read Tippett plots without actually looking at one, I will provide more detail in §3.5.

Surveys (Morrison, Sahito, et al., 2016; Gold & French, 2011) on the use of FVC among its practitioners around the globe show the same result: the presentation of conclusions using *non-numeric LRs* is more popular than that of conclusions using *numeric LRs* (Morrison, Sahito, et al., 2016, pp. 96-97). Moreover, *subjective judgement* approaches (whereby phoneticians form a qualitative opinion based on the auditory analysis of the speech recordings and/or the spectrographic analysis) are more popular in FVC than quantitative statistical approaches (whereby phoneticians carry out a quantitative analysis and make use of a statistical model to calculate the strength of evidence) (Morrison, Sahito, et al., 2016, p. 94).

In this section, I discuss the negative criticism that has been levelled against the use of verbal interpretations of LRs. Firstly, there is the danger that interpretations of verbal equivalents will vary among different interested parties, such as the jury, expert witnesses, counsel, and the judiciary (Robertson & Vignaux, 1995, pp. 55-57). For example, the expression "moderately supports" might be interpreted differently among these interested parties. Secondly, words are inadequate when the evidence becomes stronger than: "very strong", but less than certain (ibid.). Third, descriptive phrases/adjectives such as "very strong" and "good" evidence cannot be combined (ibid.). A solution to such controversies concerning the presentation of LRs is offered in Robertson and Vignaux (1995, p. 57), who suggest that the LR numbers, together with their verbal equivalents, should be presented to the court.

It is also important to note that there is some negative criticism concerning the use of the Bayesian approach. Firstly, since the prior odds, by nature, are not known, calculating the prior odds under the Bayesian framework is said to be too subjective an approach (Rose, 2002, p. 74). To be more precise, since the prior odds are indeterminate, different investigators will come up with different prior odds and, as such, different posterior odds results (Lindley, 1990, p. 45). Another criticism of the estimation of prior odds under the Bayesian framework is that it contradicts the presumption of innocence (Gigerenza et al., 1989, p. 264). Under the Bayesian framework, presumption of innocence means a zero probability of guilt. Thus, whatever the number of LRs, when combined with zero prior odds, the posterior odds will also be zero (ibid.). The solution to the determination of prior odds is that, as previously discussed, the experts should limit themselves to the calculation of LRs (Robertson & Vignaux, 1995). However, in practice, it is possible that an expert will be invited by an opposing side to calculate different prior odds under different assumptions; as such it is prudent for experts to prepare themselves for this eventuality (Rose, 2002, p. 74). As explained by Champod and Meuwly (2000, p. 199), not knowing the priors is rather beneficial for the experts so as to prevent them from making false evaluations of scientific evidence, due to the expectation effect. Secondly, under legal systems where the defense is not allowed to disclose its line of defense in advance, the calculation of an LR is not possible (Robertson & Vignaux, 1995, pp. 210-211). The solution to this is that the expert be prepared for the calculation of LR under a different defense hypothesis (ibid.). Thirdly, the logical and mathematical complexity of the Bayesian approach makes it difficult to explain to the court (Evett, 1991, p. 14;

Gigerenza et al., 1989). As such, it is recommended that an expert explain the logic and methodology used under the Bayesian framework and explain how the results are derived (Morrison, 2009a). I will now discuss in detail the reasons why forensic experts should limit themselves to the calculation of the likelihood ratio.

# 2.4.1 Why forensic experts should limit themselves to the calculation of Likelihood Ratio (LR)

There are both logical and legal reasons why forensic experts should limit themselves to the calculation of LR. As Morrison (2009a, p. 300) explains, it is *not logically possible* for forensic experts to calculate the posterior odds and say, for example, that speech samples "are highly likely to be from the same speaker given the similarity of speech samples". Since the expert is not able to access all evidential information relevant to the case, they cannot estimate the prior odds and hence derive the posterior odds as these are the product of the LR and the prior odds (Rose, 2002, p. 56).

Such an attempt to calculate the posterior probability by forensic experts would be seen as a legal *violation* or *usurpation* of the duty of the judges or jury (Morrison, 2009a, p. 300). Thus, an expert logically cannot and legally must not estimate the posterior odds. To avoid these issues, the expert should instead calculate the likelihood ratio – *the* probability of the evidence given the competing hypotheses  $(p(E/H)/p(E/\overline{H}))$ , not the probability of the hypothesis given the evidence (p(H/E)) (Rose & Morrison, 2009, p. 4).

# 2.5 The Thai legal system

The Thai legal system is based on Civil law, which has evolved from Roman law. All regulations are recorded in writing. There are three levels of courts in the Thai legal system. These are 1) the Courts of First Instance, which include, for example, the family court and the juvenile court; 2) the Courts of Appeal; and 3) the Supreme Court (Tiamjan, 2006). What I am concerned with here is the tendering of scientific evidence, particularly voice evidence, to the criminal court of Thailand by experts. I will first define who should be an expert by referring to the *criminal procedure code*, Chapter V (Tiamjan, 2006, p. 144):

Any person having, by profession or otherwise, expert knowledge on any subject such as science, art, work of skill, commerce, medicine or foreign law, and whose opinion may be of value for the adjudication of a case may, in the course of an inquiry, preliminary examination or trial, be a witness in matters such as the examination of the body or mind of the injured person, alleged offender or accused, or of handwriting, or the carrying out of experiments or other works.

From the above quote, we can see that an expert witness may be any person who has specialized knowledge, beyond that of the court, in order to help the court understand matters related to the case at hand. Such expertise may be acquired through the expert's profession, training and experience from almost any discipline or endeavor. The quote above can be further interpreted to include FVC experts as their expertise is gained from qualifications, research and experience. The *criminal procedure code* (Tiamjan, 2006, p. 144), Chapter V further states the following:

The Court may order the expert to submit his opinion in writing, but he shall be required to appear and give testimony in corroboration of such a written opinion. A copy thereof shall be served on the parties not less than three days before the date fixed for giving evidence.

From the above, it can be inferred that an expert witness may be asked to present a testimony in writing, which may or may not be followed by an oral presentation (as requested by the court or other parties in the case). Although there is no specific legislation that regulates the use of voice evidence, there are three guidelines for an expert witness to follow. First, the expert witness should give opinions based solely on their expertise (Wannasaeng, 2008). For example, a fingerprint expert cannot give testimony on DNA testing. Second, an expert witness must not give testimony about an issue that will lead to an accusation being made against an offender or the elimination of a suspect (ibid.). This is reasonable in the sense that an expert witness is privy to the rest of the information relevant to the case at hand. Thus, an expert witness should give his or her opinion to help the court draw certain inferences allowing it to reach a conclusion but not an opinion on the ultimate issue (ibid.). This is where the LR framework fits in the context of Standard Thai, where the task of an expert witness is limited to calculation of the likelihood of the evidence (E) being valid based on competing hypotheses (H). Third, an expert witness should give his opinion based on facts, and the information used should be supported by theory or academic reasoning (ibid.). It follows that the likelihood ratio framework fits well in the current work as it satisfies how the scientific evidence should be evaluated under the Thai legal system (i.e. the experts are not permitted to give a categorical answer). The criteria for the admissibility of scientific evidence in the Thai legal system just described are congruent with those reported in other countries such as the U.S. and Canada (Campbell, 2014) as we shall see in the following discussion.

# 2.6 Shift to a new paradigm

Responding to the call for an objective procedure to evaluate and present forensic scientific evidence with empirically demonstrable scientific validity (e.g. National Research Council (2009) and President's Council of Advisors on Science and Technology (U.S.) (2016)), many disciplines within the forensic science community have increasingly sought to use quantitative methods. Some rulings regarding the admissibility of scientific evidence in the U.S. (Daubert v. Merrell Dow Pharmaceuticals Inc, 1993 (Abboud, 2017)) – in which the court ruled that for evidence to be accepted as scientific evidence, the error rate of a methodology used for forensic analyses must remain within acceptable levels, and the methodology itself must be empirically testable so that others can validate it – are the original driving force for this change in the evaluation and presentation of evidence in the forensic sciences.

The use of LRs for conveying expert opinions to the decision makers, such as the court or juries, has been supported and recommended by relevant communities (see e.g. Aitken et al., 2011; Evett, 1998). In Aitken et al. (2011), which is a position statement signed by 31 individuals and supported by the Board of the European Network of Forensic Science Institutes (ENFSI), LR is described as "the most appropriate foundation for assisting the court".

Similarly, Rose (2002) strongly recommends the use of the LR framework in FVC. To be more precise, the use of objective measurements, databases that reflect the true population of the speakers, and statistical models is preferred (over subjective opinions formed by phoneticians), in order to test the validity and reliability of FVC systems under forensically realistic conditions (ibid.). As Gonzalez-Rodriguez, Rose, Ramos, Toledano, and Ortega-Garcia (2007) report, much forensic testimony has been presented in the form of expert opinions where a hard match (individualization), with categorical opinion or the use of verbal scales concerning the probability of a hypothesis, given the evidence p(H|E), is reported. Such forensic analyses lack scientific rigor (are not transparent), and are

inherently unfalsifiable (not testable) (Gonzalez-Rodriguez et al., 2007, p. 2104). The driving force behind this change from expert opinions to a more *testable* and *replicable* form of testimony, at least in methodology, has been dubbed *a paradigm shift* (not a Kuhnian paradigm shift) (Evett, 1998, p. 2105). Such a paradigm shift towards more scientifically sounding techniques is provoked by the success of using DNA profiling within the forensic domain (ibid.). To emulate the approach used to objectively quantify forensic DNA, in compliance with the admissibility criteria, which is impelled by the U.S. supreme court, the forensic testimony should satisfy *all or most* of the following criteria, based on the decision of the judge to be admitted to the court of law (Black, Ayala, & Saffran-Brinks, 1993).

- i) Whether the theory or technique can be, and has been, tested.
- ii) Whether the technique has been published or subjected to peer review.
- iii) Whether actual or potential error rates have been considered.
- iv) Whether the technique is widely accepted within the relevant scientific community.

(Quoted from Robertson & Vignaux, 1995, p. 205)

The LR framework will now be discussed in relation to the above admissibility criteria to justify the use of LR in this thesis.

First, LR meets the first admissibility criterion as it has been accepted as a logical framework (Morrison, 2012, p. 17). Since the LR framework is independent of its approaches, which according to Morrison (2012, pp. 16-18), can be typified as subjective (the phoneticians form qualitative opinions using an auditory analysis and/or a spectrographic analysis) vs objective (the phoneticians use quantitative measurements and statistical models to evaluate the strength of evidence), the validity (accuracy) and reliability (precision) for each of these approaches must be tested (ibid.) (I shall discuss the accuracy and reliability of this in Chapter 3). Thus, we see that the LR framework is a logical approach to use as its validity and reliability are tested regardless of the way it will be implemented (either subjectively or objectively) by different practitioners.

The second admissibility criterion is *the publication of peer-reviewed papers about the* use of this technique. As previously touched upon, the use of the likelihood ratio has been accepted among members of the forensic speech science community, and it has been adopted by many within that group (Aitken & Taroni, 2004; Balding, 2005; Champod &

Meuwly, 2000; Evett, 1991, 1998; Friedman, 1996; Gonzalez-Rodriguez et al., 2006; Good, 1991; Aitken & Stoney, 1991; Robertson & Vignaux, 1995). There is no doubt that peer-reviewed publications about the likelihood ratio framework will continue to be produced within the forensic speech science community.

The third criterion is the known or potential error rates of the technique. As Robertson and Vignaux (1995, p. 208) point out: "An error does occur when a test produces a result which it ought not to produce, owing to some contamination of the sample, a mistake in technique, or an undetected variation in testing conditions". As we shall see in Chapter 3, the multivariate likelihood ratio (MVLR), used to calculate LRs or the strength of voice evidence, was originally developed for glass fragments, where the invariant nature of the evidence is one of the assumptions in the formula (Aitken & Taroni, 2004). This assumption does not hold in FVC, where speech samples vary even when they are produced by the same speaker. As such, voice evidence "cannot get any closer than minimally different" (Rose, 2002, p. 21). Theoretically, the DSlog<sub>10</sub>LR and SSlog<sub>10</sub>LR should cross at the threshold at  $log_{10}LR = 0$ . However, this is not always the case in FVC. In part, this is due to the different nature of the data that are experimented on using the same MVLR statistical tool: glass fragments (invariant) vs speech samples (variant, i.e. variable over time). Errors may therefore arise in FVC systems if LRs mistakenly support the counterfactual hypothesis, i.e. if same-speaker samples are wrongly discriminated as coming from different speakers and vice versa. Such errors, which I have termed scores in the current work, can be calibrated into true LRs by means of logistic regression calibration; we assess their accuracy using the log-likelihood-ratio-cost ( $C_{llr}$ ).

I will now discuss the last criterion that makes LR admissible for use – *general acceptance within the relevant scientific community*. Robertson and Vignaux (1995, p. 208) explain that "acceptance within the relevant scientific community will usually follow from the testing and replication of such experiments". To date, many empirical experiments have been carried out on LR-based FVC to test the realistic level of forensic application, in a number of languages including Cantonese (Li & Rose, 2012; Chen & Rose, 2012; Yim & Rose, 2012), Chinese (Zhang & Tan, 2008; Zhang, Morrison, & Thiruvaran, 2011), English (Morrison, 2009b; Rose, 2003; Rose, Warren, & Watson, 2006) and Japanese (Kinoshita et al., 2009; Rose, Osanai, & Kinoshita, 2003).

So far, we have seen that Bayes' theorem or the LR-based approach is suitable for evaluating and interpreting speech evidence. There is another important characteristic of LR that justifies its use in FVC or forensic evidential science in general. That is, when the LR values from any pieces of evidence are not correlated, they can be multiplied via Naïve Bayes' theorem to get the overall LR (Rose, 2002, p. 61). For example, an LR of 8 from blood type and another one of 2 from voice evidence results in an overall LR of  $16 (= 2 \times 8)$ . Not only different pieces of evidence, but also the LRs obtained based on, say, two different properties of the same evidence, can be combined if they are not correlated. For example, the formant and F0 values (which are obtained from the same vocal tract of a speaker) can be combined via the logistic regression fusion proposed by Brümmer and du Preez (2006). I will discuss this alongside statistical analysis in Chapter 3.

To sum up, we have seen many advantages of using the Bayesian framework in this thesis. Firstly, the Bayesian framework allows forensic experts to obtain and present numerical and meaningful values of the weight of evidence to the court in the form of LRs (Robertson & Vignaux, 1995). Secondly, there is a clear distinction between the role of forensic experts and that of fact finders, leaving the court to incorporate the priors into their decision-making process (ibid.). Given these advantages, the degree of belief in the conditional probabilities under two competing hypotheses is assessed in a scientific way using the LR framework (Gonzalez-Rodriguez et al., 2007, p. 2105).

# 2.7 Standard Thai and other main dialects: Phonetics and phonology

This section describes the phonetics and phonology of Standard Thai, a dominant language taught in schools, used in the media, and described by grammar books and dictionaries (Tingsabadh & Abramson, 1993). The phonetics and phonology of other major dialects, grouped according to regions, will also be discussed to explain the linguistic situation in Thailand. They are: 1) the Lanna Thai or Northern Thai dialect, spoken in the north of Thailand by 6,000,000 native speakers (Lewis et al., 2013); 2) the Isan or Northeastern Thai dialect, spoken in the northeastern part of Thailand by 15,000,000 native speakers; and 3) the Southern Thai dialect, spoken in the South of Thailand by 4,500,000 native speakers (ibid.). The presentation of the phonetics and phonology of other Thai dialects can help the reader appreciate which linguistic-phonetic

features, e.g. consonant and vowel phonemes as well as tones, are shared between the dialects and Standard Thai, and which are different. In a forensically realistic scenario, for example, a speech sample of Standard Thai retrieved from a relatively short recording has not necessarily been produced by a speaker of Standard Thai who has a "Bangkok accent". There might be some other linguistic-phonetic cues, e.g. the different vowel phoneme realization of the Standard Thai diphthong [oi], that suggest that the author of the questioned speech sample is more likely to be from a different speech community, e.g. from a person who has a Northern Thai accent.

# 2.7.1 Standard Thai consonant phonemes

Table 4 shows the Standard Thai consonant phonemes listed according to their place (horizontal axis) and manner (vertical axis) of articulation.

Consonants	Bilabial	Labio- dental	Alveolar	Post- alveolar	Palatal	Velar	Glottal
		uentai		arveorar			
Plosive	p, p <sup>h</sup> ,b		t, t <sup>h</sup> , d			k, k <sup>h</sup>	3
Nasal	m		n			ŋ	
Fricative		f	S				h
Affricate				tc, tc <sup>h</sup>			
Trill			r				
Approximant					j	W	
Lateral approx.			1				

**Table 4:** Standard Thai consonant phonemes (Adapted from Tingsabadh & Abramson, 1993, p. 24)

In Standard Thai, a stressed syllable may be represented as  ${}^{T}C(C)VC$  or  ${}^{T}C(C)VV(C)$ , where C is a consonant, CC a consonant cluster, V a short vowel, VV a long vowel, and  ${}^{T}$  a tone (adapted from Onsuwan, 2005, p. 5). Stressed syllables consist of a tone and up to two initial consonants followed by a short vowel and a coda, or a long vowel with optional coda (ibid.). As shown in Table 4, 21 consonant phonemes – 9 stops, 2 affricates, 3 fricatives, 3 nasals, 2 liquids (lateral, trill), and 2 glides – are possible in the onset. Only nine, /p, t, k, ?, m, n,  $\mathfrak{n}$ , j, w/, are allowed in the coda (ibid.). These final consonants, which are stops, are unreleased and phonetically transcribed as [p', t', k'].

#### 2.7.2 Standard Thai clusters

There are twelve consonant clusters in Standard Thai, which are listed in Table 5.

No.	1st phoneme	2 <sup>nd</sup> phoneme	Clusters
1	/p/	/ <b>r</b> /	/pr/
2	/t/	/r/	/tr/
3	/k/	/r/	/kr/
4	$/p^{h}/$	/r/	/p <sup>h</sup> r/
5	/t <sup>h</sup> /	/r/	/t <sup>h</sup> r/
6	$/k^h/$	/r/	$/k^h r/$
7	/p/	/1/	/pl/
8	/k/	/1/	/kl/
9	$/p^{ m h}/$	/1/	$/p^{h}l/$
10	$/\mathrm{k}^{\mathrm{h}}/$	/1/	/k <sup>h</sup> l/
11	/k/	/w/	/kw/
12	$/k^h/$	/w/	$/k^hw/$

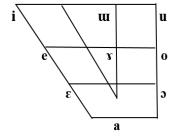
**Table 5:** Standard Thai consonant clusters (Adapted from Sriyaphai, 2013, p. 78)

The columns in Table 5 reveal that three different patterns can be observed. In the first pattern, only six of the nine stops, /p, t, k,  $p^h$ ,  $t^h$ ,  $k^h$ , can be followed by a trill /r. In the second, a bilabial /p/ or a velar stop /k/ (both aspirated and unaspirated) are followed by a liquid /l/. In the third pattern, only a velar stop /k/ (both aspirated and unaspirated) can be followed by a velar glide, giving in total 12 consonant clusters.

## 2.7.3 Standard Thai vowels

The vowel space of the Standard Thai monophthongs and diphthongs is shown in Figure 1 and Figure 2.

#### 2.7.3.1 Monophthongs



**Figure 1:** Standard Thai monophthongs (Adapted from Tingsabadh & Abramson, 1993, p. 25)

Figure 1 shows the vowel space of the nine Standard Thai monophthongs. There are three front vowels (/i, e,  $\epsilon$ /), three central vowels (/u,  $\gamma$ , a/) and three back vowels (/u,  $\rho$ ,  $\rho$ /). The monophthongs are contrastive in length, e.g. [ci:p L] "to flirt" vs [cip L] "to sip".

## 2.7.3.2 Diphthongs

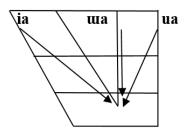


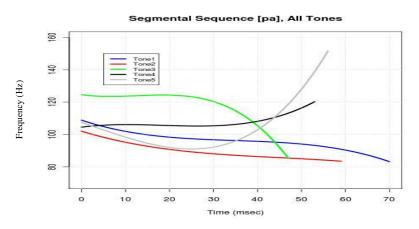
Figure 2: Standard Thai diphthongs

(Adapted from Tingsabadh & Abramson, 1993, p. 25)

Figure 2 shows the vowel space of the three Standard Thai diphthongs. Each diphthong consists of a high vowel (/i, uı, u/) followed by a low vowel (/a/ or /a:/). As such, /ia/ has two allodiphthongs [ia] and [i:a], /uua/ has [uua] and [u:a], and /uua/ has [uua] and [u:a]. The shorter versions ([ia], [uua], [uua]) occur in closed syllables while the longer ones ([i:a], [u:a], [u:a]) occur in open syllables. Short and long allodiphthongs do not differ in their diphthongal quality (Roengpitya, 2012, p. 53).

#### 2.7.4 Standard Thai tones

Standard Thai tones can be categorized into two groups, static and dynamic (Abramson, 1962; Naksakul, 1998). The first group consists of low tones [`], mid tones [no phonetic symbol] and high tones [´]; the second group consists of rising tones [`] and falling tones [^]. Instead of using diacritics to indicate tones, I will symbolize low tone as [L], mid tone as [M], high tone as [H], rising tone as [LH] and falling tone as [HL]. Figure 3 shows all five tones as uttered by a male speaker.



**Figure 3:** F0 contours of the five Standard Thai tones for the same segmental sequence [pa:]. The mid tone [M] is navy blue, the low tone [L] is red, the falling tone [HL] is green, the high tone [H] is black, and the rising tone [LH] is grey.

(Reproduced from Pingiai, 2011, p. 15)

Each of these five tones was plotted using the averaged F0 values from ten tokens. The x-axis represents the time that elapsed in milliseconds (msec) and the y-axis shows the frequency in Hertz (Hz). Duration was not normalized, so as to allow readers to observe the original shapes of each of the five tones.

Tone 1 [M] (navy blue) gradually drops in frequency from an onset of ca. 110 Hz to an offset of ca. 82 Hz. The entire duration of Tone 1 lasts ca. 70 msec. Tone 2 [L] (red) also shows a slightly falling contour along its entire time-course, which lasts about 60 msec. Tone 3 [HL] (green) is fairly stable at around 125 Hz during the first 30 msec before it sharply falls down towards its offset (at ca. 85 Hz). Tone 4 [H] (black) starts at ca. 105 Hz and is fairly stable around this frequency, before it gradually rises towards its offset (ca. 120 Hz). Tone 5 [LH] (grey) starts at ca. 110 Hz and shows a concave contour during the first 40 msec, before it sharply rises towards its offset (155 Hz).

## 2.8 Northern Thai dialect

Reasons for including the phonetics and phonology of other major dialects in Thailand, grouped according to regions of Thailand, were given above. In this section, I present the consonants, vowels and tones of the Lanna Thai or Northern Thai dialect, which is spoken in the north of Thailand. The description of Northern Thai phonetics and phonology (consonants, vowels, and tones, respectively) is based on earlier work by different Thai scholars. Presentations of the Pak Tai or Southern Thai and the Isan or Northeastern Thai dialects follow in §§2.9 and 2.10.

#### 2.8.1 Northern Thai consonant phonemes

Table 6 lists the Northern Thai consonant phonemes. /c\*/ is classified as plosive in Rungruengsri (1991) and Wimolkasem (2006), although its place and manner of articulation is the same as that of Standard Thai /tc/.

Consonants	Bilabial	Labiodental	Alveolar	Palatal	Velar	Glottal
Plosive	b, p, p <sup>h</sup>		d, t, t <sup>h</sup>	c*	k, k <sup>h</sup>	3
Fricative		f	S			h
Nasal	m		n	n	ŋ	
Lateral			1			
Approximant	W			j		

**Table 6:** Northern Thai consonant phonemes (Adapted from Wimolkasem, 2006, p. 9)

There are 20 consonant phonemes in the Northern Thai dialect (as opposed to 21 in Standard Thai). There are four nasals, /m, n,  $\eta$ ,  $\eta$ , as opposed to three, /m, n,  $\eta$ , in the standard language. There is no trill /r/ in the Northern Thai phoneme inventory, /r/ being realized as [h] (most commonly) or [l] (Wimolkasem, 2006, p. 30). Thus, the word for the verb *study* is [ri:an M] in Standard Thai, but [hi:an M], in Northern Thai. Eleven consonants (/b, d, c, p<sup>h</sup>, t<sup>h</sup>, k<sup>h</sup>, f, s, h, n, l/) only occur word-initially; the nine others (/p, t, k, m, n,  $\eta$ , w, y, ?/) can occur both word-initially and word-finally (ibid.).

#### 2.8.2 Northern Thai clusters

In the northern Thai dialect, /w/ can occur after the ten initial consonants /?, k,  $k^h$ , c, t,  $\eta$ , n, s, l, y/ to form clusters (Wimolkasem, 2006, pp. 10-12). By contrast, in Standard Thai, there are three consonant phonemes /r, l, w/ that can occur after /p, t, k,  $p^h$ ,  $t^h$ ,  $k^h$ /.

#### 2.8.3 Northern Thai vowels

I will now discuss the vowel phonemes of the Northern Thai dialect.

### 2.8.3.1 Monophthongs

The Northern Thai dialect has eighteen monophthongs: /i, i:, e, e:,  $\varepsilon$ ,  $\varepsilon$ ;,  $\omega$ ,  $\omega$ ;,  $\tau$ ;, a, a:, u, u: o, o:, o, o:/, each of which is distinctive in length. These monophthongs have the same vowel quality as those of Standard Thai (Wimolkasem, 2006, p. 41).

#### 2.8.3.2 Diphthongs

There are six diphthongs in the Northern Thai dialect (as opposed to three in Standard Thai). Unlike Standard Thai diphthongs, they are distinctive in length (Roengpitya, 2012).

#### 2.8.4 Northern Thai tones

There are also six tones in the Northern Thai dialect (as opposed to five in Standard Thai) (Rungruengsri, 1991). They are shown in Table 7.

1.	Mid
2.	Rising
3.	Low
4.	Falling
5.	High-falling
6.	High

**Table 7:** Northern Thai tones (After Rungruengsri, 1991, p. 251)

Following the categorization of tones (static vs dynamic) used by Abramson (1962) and Naksakul (1998), Northern Thai has three static and three dynamic tones. The main difference between the Northern Thai dialect and Standard Thai in terms of tones is that the latter has only one falling tone but the former has two: falling and high-falling. In comparison to Standard Thai, five of the Northern Thai tones are the same as those found in Standard Thai, the exception being the high-falling tone.

# 2.9 Southern Thai dialect

In this section I provide information about the Southern Thai phonology (consonants, vowels, and tones, respectively), based on earlier work done by different Thai scholars.

# 2.9.1 Southern Thai consonant phonemes

The Southern Thai dialect is spoken in the south of Thailand, from Chumporn to Narathiwat provinces (fourteen provinces in total) (Nookua, 2012, p. 28), as shown on the map in Figure 4. There are also dialectal varieties according to regions such as those in the Southernmost provinces of Pattani, Yala, Narathiwat and some parts of Songkhla (Nookua, 2012, p. 28).



**Figure 4:** The South of Thailand (Source: Google, n.d.)

The main distinction between the Southern Thai dialect and Standard Thai concerns its phonology (tones) and its lexicon rather than its syntax (Nookua, 2012, p. 28). Additionally, there is no recognized orthography in this dialect (ibid.). In this section, the phonology of Phang-Nga has been chosen to represent the Southern Thai dialect (as Phang-Nga is mostly cited in the relevant literature). Thus, Table 8 contains the consonant phonemes of the Southern Thai dialect (Phang-Nga).

Consonants	Bilabial	Labiodental	Alveolar	Palatal	Velar	Glottal
Plosive	b, p, p <sup>h</sup>		d, t, t <sup>h</sup>		k, k <sup>h</sup>	3
Affricate				c, c <sup>h</sup>		
Fricative			S			h
Nasal	m		n		ŋ	
Lateral			1			
Trill			r			
Approximant	W			j		

**Table 8:** Consonant phonemes of the Southern Thai dialect (Phang-Nga) (Adapted from Wilaisak, n.d.)

Unlike Standard Thai and the Northern Thai dialects, there is no fricative /f/ in Southern Thai phonology (Phang-Nga). Moreover, /ŋ/ does not occur word-initially in this dialect (Wilaisak, n.d.).

# 2.9.2 Southern Thai clusters

We now look at the clusters of the Southern Thai dialect. They are listed in Table 9.

Clusters	l	r	W
p	pl	pr	-
p <sup>h</sup>	p <sup>h</sup> l	p <sup>h</sup> r	-
b	bl	br	-
t	-	tr	-
k	kl	kr	kw
k <sup>h</sup>	k <sup>h</sup> l	k <sup>h</sup> r	k <sup>h</sup> w
m	ml	mr	-

 Table 9: Southern Thai clusters

(Adapted from Wilaisak, n.d.)

Table 9 shows that, interestingly, a nasal /m/ and a plosive /b/ can be followed by a lateral /l/ or a trill /r/ to form a cluster. This is not the case in the Northern Thai and Standard Thai dialects.

#### 2.9.3 Southern Thai vowels

## 2.9.3.1 Monophthongs

There are eighteen monophthongs in the Southern Thai dialect; they are the same as those found in Standard Thai and in the Northern Thai dialect (Wilaisak, n.d.).

## 2.9.3.2 Diphthongs of the Southern Thai Dialect

Similar to Standard Thai, there are three diphthongs /ia, wa, ua/ in the Southern Thai dialect (Wilaisak, n.d.). However, unlike those of the Northern Thai dialects, these three diphthongs are not distinctive in length in the Southern Thai dialect.

#### 2.9.4 Southern Thai tones

In the Southern Thai dialect (Phang-Nga), there are seven contrastive tones, the highest number found among the dialects discussed so far. Southern Thai tones are shown in Table 10.

1.	High rising-falling
2.	High
3.	Mid rising-falling
4.	Mid
5.	Low rising-falling
6.	Low rising
7.	Low falling

**Table 10:** Southern Thai tones (Adapted from Wilaisak, n.d.)

There are two static and five dynamic tones in the Southern Thai dialect. A unique aspect of this dialect are the rising-falling tones (high rising-falling, mid rising-falling, and low rising-falling), which do not exist in the other dialects. Low tones also have rising and falling tones. The high and mid tones have allotones – level or rising. These allotones are determined by the initial consonant classifications of high, middle, and low (Li, 1966). That is, a {high, mid}-rising allotone is determined by high consonants in checked or 'dead' syllables (i.e. those ending in /p, t, k/) with short vowels, and a {high, mid}-level variant is determined by high consonants in open syllables or checked syllables with long vowels (Gedney, 1972, p. 424).

## 2.10 Northeastern Thai dialect

This section provides a description of the Northeastern Thai phonology (consonants, vowels, and tones), based on earlier work done by different Thai scholars.

# 2.10.1 Northeastern Thai consonant phonemes

Northeastern Thai has twenty initial consonants (based on speakers of Ubon Ratchathani province) as shown in Table 11.

Consonants	Bilabial	Labiodental	Alveolar	Palatal	Velar	Glottal
Plosive	b, p, p <sup>h</sup>		d, t, t <sup>h</sup>	c*	k, k <sup>h</sup>	3
Fricative		f	S			h
Nasal	m		n	n	ŋ	
Lateral			1			
Approximant	W			j		

**Table 11:** Northeastern Thai consonant phonemes (Adapted from Wilaisak, n.d.).

#### 2.10.2 Northern Thai clusters

There are no clusters in this dialect (Wilaisak, n.d.).

## 2.10.3 Northeastern Thai vowels

#### 2.10.3.1 Monophthongs

There are eighteen monophthongs in the Northeastern Thai dialect. These eighteen monophthongs are the same as those found in other Thai dialects discussed so far.

#### **2.10.3.2 Diphthongs**

There are three diphthongs in the Northeastern Thai dialect, as in the other dialects, with the exception of the Northern Thai dialect. In some varieties of the Northeastern dialect, /uua/ does not exist (Luemsai, 2001; Wilaisak, n.d.). In other words, in some areas of the Northeast only two diphthongs are used: /ia/ and /ua/.

#### 2.10.4 Northeastern Thai tones

Northeastern Thai tones range from four to seven, depending on the region (Luemsai, 2001). As Smalley (1994, p. 89) points out, within the Northeastern area, some small differences in tonal systems may be found within certain provinces, districts, and even villages; as a result of this, one can identify the area where a person comes from based on these tonal qualities. The tonal system illustrated in Table 12 is that of Ubon Ratchathani province; it comprises six tones (Wilaisak, n.d.). A mid-rising tone is the distinctive tone found in this region (compared to Standard Thai).

1.	Low
2.	Mid
3.	Mid-rising
4.	High
5.	Falling
6.	Rising

**Table 12:** Northeastern Thai Tones (Adapted from Wilaisak, n.d.)

# 2.11 Speech signal representation

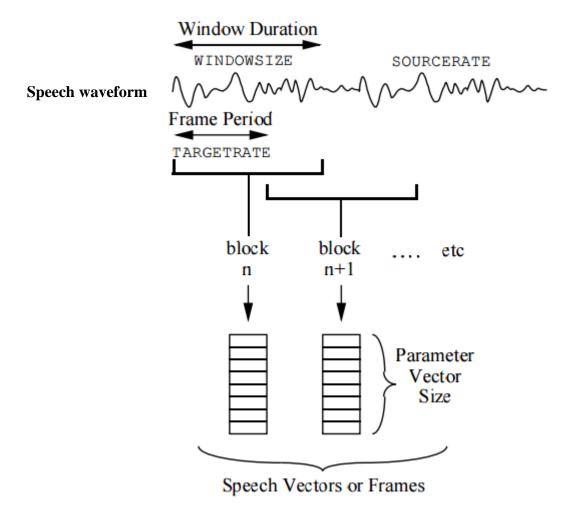
Having introduced this background knowledge on FVC, including the LR framework, the Thai legal system and Thai phonetics and phonology, I will now present a literature review detailing my choice of acoustic parameters tested in Standard Thai. As such, the following subsections first provide essential knowledge about speech feature extraction. This will be followed by a review of previous FVC research using the spectrum, F0, LTF0, and formants as parameters.

# 2.11.1 General feature extraction process: Short-time analysis

Although the FVC work in this thesis uses the semi-automatic method, where both a computer-based and human-supervised analysis are involved, the speech feature

extraction process used in Automatic Speech Recognition (ASR) is introduced here to show the basic principles adhered to when speech is parameterized in the current work. It should be pointed out here that the current work uses a *front-end analysis* as opposed to a *back-end analysis*, whereby "the acoustic signal is converted into a sequence of acoustic feature vectors" (Meseguer, 2009, p. 14), as we shall see below.

Figure 5 shows the speech feature extraction process used in automatic speech recognition (ASR).



**Figure 5:** Feature extraction process

(Source: Young et al., 2002, p. 59)

The aspect of most importance for the current work is how the speech signal is converted into a sequence of feature vectors and then used to compare speakers. To begin with, speech is decomposed into a sequence of short time frames, a so-called short-time analysis (Meseguer, 2009, p. 12). That is, each of these frames is converted into a speech

vector that contains speech information labeled as block n and block n+1 in . The segment of the waveform to be analyzed is referred to as window (Young et al., 2002, p. 58). This window duration is independent from the frame period, where the former is usually larger than the latter (ibid.). In this thesis, Hamming window is chosen as it offers better frequency resolution (Meseguer, 2009, p. 13). Once we get the resultant acoustic feature vectors, they are used to model a speaker (e.g. offender or suspect), and consequently to calculate LRs.

# 2.12 Source of individualizing information: Spectral, phonotactic, prosodic and idiolectal levels

This section reviews previous studies with particular consideration given to the acoustic parameters of 1) spectral moments and the cepstrum of /s, tch, n, m/, 2) tonal F0, 3) long-term F0 (LTF0), and 4) the formant trajectories, which are the focus of the present thesis. It is well known in phonetics and other disciplines, such as engineering, speech signal processing and physics, that sources of individualizing information can be found at the spectral, phonotactic, prosodic, and idiolectal levels, where spectrum is regarded as the lowest-level feature and idiolect as the highest (Reynolds et al., 2003, p. 260).

Spectrum, the lowest-level feature, is said to be directly related to the dynamic configuration of the vocal tract (Castro, 2007, p. 20) because it is one of the processes used to separate source and filter components according to the source-filter model. It is actually a mathematical abstract and, when combined, can approximate speech waveforms. This being the case, spectrum has been commonly found to be a parameter in automatic speaker recognition (ASR) in the last decade, and it has been demonstrated that it contains much speaker-specific information (ibid.). It should be noted here that spectrum is not a common parameter in traditional FVC as opposed to traditional features such as formant frequencies, whose F1 inversely correlates with vowel height and whose F2 correlates with the vowel backness/rounding (Rose, 2002; Nolan, 1983).

At the (second lowest) phonotactic level, speaker-specific information is extracted from phonemes, syllables and their realizations (Castro, 2007, p. 21). Information at this level is said to contain much language-dependent variability, as phonotactics is fundamentally concerned with the freedom and restrictions with which phonemes are combined (ibid.).

Much of the naturalness of a speaker's voice is said to be situated at the prosodic level. This level contains pitch, loudness, and rhythm, which make speech sound natural and human-like (Halliday, 2015).

Lastly, there is the idiolectal level, where individualizing information is extracted by the use of particular words, grammar and pronunciation unique to the speaker (Castro, 2007, p. 21). This highest level is assumed to yield optimal results in speaker recognition, as it contains information distinctive to an individual (ibid.). However, in this thesis, the sources of individualizing information exploited are from the acoustic parameters at the spectral and phonetic, rather than the prosodic and idiolectal, levels, and they are extracted from the segmental consonants /s/, /tch/, /m/, /n/ and the diphthongs [ɔi] and [ai]. The following sections present an overview of the body of literature surrounding the acoustic parameters of the spectral moments and cepstrum (§§2.15–2.17), tonal F0 and LTF0 (§2.18), and the formant trajectories (§2.19), respectively.

# 2.13 DCT-smoothed spectrum vs cepstrally smoothed spectrum

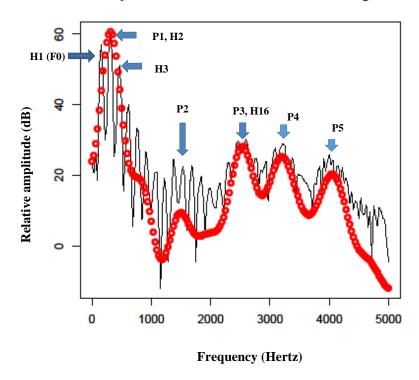
In this thesis, a discrete cosine transformation (DCT) is applied to a raw spectrum (I will explain this in §2.14) extracted from various consonants. In applying DCTs, a short-time Hamming window of 31.25 ms (sampling frequency of 16,000 Hz / a signal of length *N* = 512 points) was selected in the EMU speech database system (Harrington, 2010, p. 276). The reason for choosing a Hamming window is that it is said to offer a better frequency resolution (Meseguer, 2009, p. 13). With these DCT coefficient outputs, various experiments were conducted. To execute the DCT analysis, the **dct** () function is used as an argument inside **fapply** () in the EMU/R library (Cassidy, 1999). For example, a cepstrally smoothed spectrum with 30 coefficients (dct) for a spectral matrix /s/, extracted at the temporal midpoint in the frequency range between 500–8000 Hz (s\_dft\_500\_8000.5), is given by dct\_s = fapply (s\_dft\_500\_8000.5, dct, 29, fit = T), where fit is an argument for cepstrally smoothed spectra.

A DCT, which is very much like a Discrete Fourier Transformation (DFT), decomposes a speech signal into a set of sinusoids; when summated, it reconstructs the original speech signal (Harrington, 2010, pp. 304-305). A DCT is a set of sinusoids "at *half*-cycles, that is, at  $k = 0, 0.5, 1, 1.5 \dots \frac{1}{2}(N-1)$ , rather than, as for the DFT, at *integer* cycles (k = 0, 1,

 $2, \dots (N-1)$ " (ibid.). Moreover, the output of a DCT (as opposed to that of a DFT), is a set of sinusoids with no phase, namely a cosine wave. Hence the name, Discrete Cosine Transformation (ibid.). The amplitudes of these cosine waves are called *DCT coefficients* (Harrington, 2010, p. 310). With these DCT coefficients of the raw spectrum as fecture vectors, a series of experiments were carried out (see Chapter 5). The DCT amplitudes are usually labeled from 0 to N-1; "the more that are summed, the more the resulting signal approximates the original spectrum" (ibid.). This motivates us to use 1) the coefficient zero ( $k_0$ ) up to coefficient 14 ( $k_{14}$ ) (15 DCT coefficients in total), and 2)  $k_0$ - $k_{19}$  (20 DCT coefficients in total), as they are considered to be sufficient to correlate with the speaker's vocal tract (ibid.).

To introduce the basic concept of spectral analysis, let us look at the jagged profile of a spectrum extracted from the temporal midpoint of an oral vowel /i/ and its DCT fitting. Figure 6 further investigates a 512-point dB spectrum of this vowel, sampled at 16,000 Hz.

#### Spectrum of /i/ and its DCT curve fitting



**Figure 6:** A raw spectrum (black) extracted from the temporal midpoint of an oral vowel /i/ and its corresponding Hertz-scaled DCT curve fitting (red) using 512 data points.

The x-axis represents the frequencies in Hertz (Hz); the y-axis is the relative amplitude of the spectrum in Decibels (dB). As this segment is an oral vowel, a low-pass filter of 5000 Hz is sufficient to capture the first four formants that contribute to its phonetic distinction.

#### According to Harrington (2010, p. 309):

The 512-point window is easily wide enough so that harmonics appear: there is a gradual rise and fall due to the presence of formants and superimposed on this is a jaggedness produced by the fundamental frequency and its associated harmonics.

A close inspection of Figure 6, which is a two-dimensional plot of relative amplitude (yaxis) against frequency (x-axis), reveals a jagged profile of a vowel /i/ in black and its grosser structure, which consists of five major peaks in red. Each of the local spikes (black) is a harmonic with a given frequency and relative amplitude. These harmonics in even spacing frequencies are the sinusoidal components (Harrington, 2010, pp. 304-305). The first harmonic with the lowest frequency, labeled H1, is the fundamental frequency (F0) – the rate of the repetition of the vocal cords per second (Rose, 2002, p. 244). It is quite difficult to observe the frequency at which this H1 occurs, given the frequency axis. But this can be visually approximated as anywhere in the vicinity of 100–500 Hz. The second higher spike in frequency, which is double that of the F0 at 60 dB, is called H2. The next, higher in frequency, is the third harmonic, H3, and so on (ibid., p. 205). Mathematically, the harmonics occur at whole-number multiples of the F0 (ibid., p. 205). A grosser structure (the dotted red line in Figure 6) can be obtained when the jagged profile of the harmonics is *smoothed* (only lower DCT coefficients, which exclude fundamental frequency and harmonics information, are summated) (Harrington, 2010, p. 309). There are five major peaks, labeled P1-P5, which are the frequencies where the air in the supralaryngeal vocal tract is vibrating at its maximum amplitude. In acoustic phonetics, these peaks are termed formant center frequencies. As Rose (2002, p. 206) states, "the frequencies of the lowest three major peaks are the primary correlatives of vowel quality".

It should be made clear at the outset that, in speech technology, the output of a DCT analysis applied to a spectrum is considered to be a very close approximation to that of a *cepstral analysis*, notwithstanding their minor differences (Milner & Shao, 2006). Rose and Clermont (2001, p. 31) explain that cepstrum is a smoothed spectrum.

Thus, the following aims to explain minor differences between cepstrum and spectrum in terms of their extraction process. Cepstral analysis is the inverse of a Fourier

transformation of the log-spectrum (Meseguer, 2009; Bogert, Healy, & Tukey, 1963; Nair, Alzqhoul, & Guillemin, 2014); it is illustrated in the diagram in Figure 7.

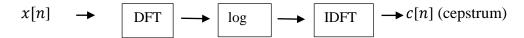


Figure 7: Cepstral analysis

(Adapted from Gutierrez-Osuna, 2017, p. 2)

In Figure 7, x[n] stands for the speech signal. First, a discrete Fourier transformation (DFT) is applied to the speech signal x[n] to convert acoustic features from the time domain into their corresponding representation in the frequency domain. Second, the logarithm function (log) is applied to the magnitude component of speech (and the phase is thrown away) to convert it into the "quefrency domain or the cepstral domain, which is similar to the time domain" (Meseguer, 2009; Gutierrez-Osuna, 2017; vlab.amrita.edu, 2011). Third, the inverse discrete Fourier transformation (IDFT) is performed to the product of the second stage (log) to get the cepstral coefficients (c[n]) (ibid.). These cepstral outputs can separate the source (glottal excitation) and the filter (vocal tract) as represented by the higher and lower cepstral coefficients, respectively (Kumar & Lahudkar, 2015).

A DCT analysis, on the other hand, proceeds differently. When a DCT (a mathematical operation) is applied to a spectrum, the speech signal is decomposed into a set of cosine waves at half-cycles (Harrington, 2010, p. 304). The resultant amplitude of such cosine waves is referred to as the DCT coefficients. The outputs of the cepstral and DCT analyses are actually the spectrum, which represents the vocal tract filter. In this regard, cepstral coefficients are considered a close approximation of the DCT coefficients, although they are different in terms of their extraction process.

Thus, *DCT coefficients* are essentially *cepstral coefficients* and a *DCT-smoothed spectrum* is a *cepstrally-smoothed spectrum* (Harrington, 2010, p. 306). This smoother version of the original signal, i.e. the *DCT-smoothed spectrum* or *cepstrally-smoothed spectrum*, results when the harmonics (I will explain this in the subsequent section) are at high frequencies due to vocal fold vibrations that are filtering out, reflecting only the shape of the vocal tract (ibid.).

The benefits of obtaining these cepstral coefficients, among others, are that 1) they can separate source and filter, as indicated above (i.e. high coefficients approximate the glottal excitation while low coefficients approximate the vocal tract), and 2) they are very compact in representing the spectral envelope (Gutierrez-Osuna, 2017, pp. 8-10). In what follows, I justify in more detail why cepstrum is very attractive in traditional FVC, based on the three main advantages explained in Rose (2013a).

The first advantage of using a cepstrum as a parameter in FVC is its great power. Rose (2013a, p. 192) empirically shows that, given the same data, cepstrum lends itself to a five times stronger LR magnitude than formant frequencies alone. Rose (2013a) further explains that this great power of cepstrum is probably due to more information captured from the whole of the spectral envelope. That is, there is a greater chance of picking up more speaker-specific information (ibid.). Rose (2013a, p. 84) points out that the cepstral-spectral envelope of a vowel not only reflects the vocal tract dimensions (in its F-patterns), but also the phonatory activity of the source (in its spectral slope) and the tract compliance (in its formant bandwidths).

The second advantage of cepstrum (over formants) is that it is much easier to extract. Additionally, in forensically realistic casework, any speech recordings obtained from a degraded transmission channel will typically distort or lose speaker-individualizing information. As such, it is more difficult to find acoustic variables that are continuous in nature (other than, among others, the cepstrum and duration variables). For example, it is empirically found in Hughes, Foulkes, and Wood (2016) that the duration of the hesitation marker "um" (together with the formant trajectories) can improve the validity of the Melfrequency cepstral coefficients (MFCC)-based ASR.

The third advantage of cepstrum is its use as a complementary feature to potentially add the strongest possible evidence when combined with other parameters under the LR framework (J. Holmes, W. Holmes, & Garner, 1997; Rose, 2013a). For example, an expert can combine the formant frequencies from an easy-to-extract vowel with the cepstral coefficients of a difficult nasal segment (Rose, 2013a, p. 85).

Last but not least, a cepstrum is preferred over formants because the cepstral coefficients can be used to quantify both voiced and voiceless speech segments, while it is harder to extract formants in voiceless sounds (Harrington, 2010, p. 316).

Regardless of the main advantages of cepstrum in FVC, there are also some caveats about its use. Firstly, cepstrum is sensitive to channel transmission (Rose, 2013a, p. 84). When the shape of the spectrum is perturbed in telephone transmissions, for example, all cepstral coefficients will change (ibid.). Secondly, cepstrum generally lacks "interpretability in terms of speech production" (ibid.), i.e. it is just a mathematical operation that, when combined, gives the best approximation of a reconstructed smoothing spectral shape (Clermont & Itahashi, 2000).

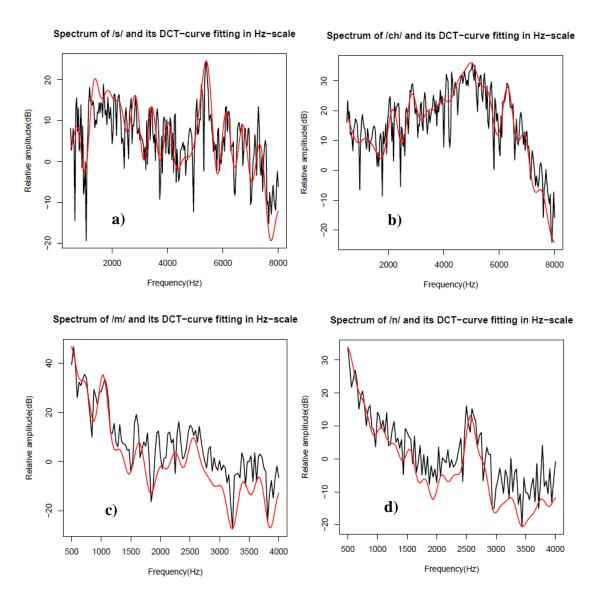
# 2.14 Mel- and Bark-scaled DCT (cepstral) coefficients

Apart from using the perceived pitch or frequency in Hertz as a scale to represent a spectrum, Mel and Bark scales can be used to warp the frequency in Hz into one that corresponds more closely to what humans hear (Harrington, 2010, p. 312). In this thesis, I parameterized a spectrum using the frequency in Hertz and Bark scales to identify their discriminatory power and determine the best-performing parameter.

The motivation to use auditory Bark scales was not only the fact that they correlate to the frequency processed in human ears, but also that *fewer* Bark-scaled DCT (cepstral) coefficients are needed to efficiently distinguish among different phonetic categories than when working with a Hz scale, as reported in automatic speech recognition (Meseguer, 2009, p. 19). Fewer coefficients, based on perception models, may imply less computational cost in terms of experimental time in a forensically realistic world. In what follows, I show how well the DCT coefficients fit the spectra in both Hertz and Bark scales, using data from the current thesis.

# 2.14.1 Hertz-scaled DCT (cepstral) coefficients fitted to a raw spectrum

Figure 8 (overleaf) shows a raw spectrum (in black) extracted at the temporal midpoint of each of the target segments /s,  $tc^h$ , m, n/, plotted together with its DCT curve fitting (in red) in a Hertz scale. The speech samples are from male speakers of Standard Thai and are excerpted from the current FVC corpus (for details, see Chapter 3). The x-axis is the frequency in Hz and the y-axis is the relative amplitude in decibels (dB). As shown in Figure 8, a DCT or cepstrally smoothed version of the spectrum excludes in principle the contribution from the source, i.e. the summation does not include the higher frequency cosine waves (only  $k_0$ - $k_{29}$  were included) that encode information about the F0 and



**Figure 8:** A DCT-smoothed signal (cepstrally smoothed spectrum) superimposed on the original spectrum (in black) by summing up the first 30 half-cycle cosine waves on a Hz scale (in red) of /s, te<sup>h</sup>, m, n/.

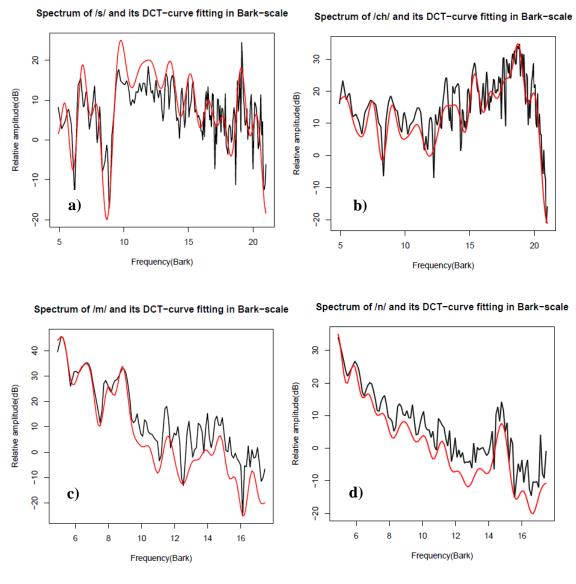
Note: /te<sup>h</sup>/ is labeled as /ch/.

harmonics (Harrington, 2010, p. 310). To simulate realistic forensic conditions of a possible telephone transmission effect, a speech signal is high-passed at the same cut-off frequency of 500 Hz. However, a signal is low-passed with a different cut-off frequency: 4000 Hz for the nasals /m, n/ (Figures 8c and 8d), and 8000 Hz for the fricative /s/ and affricate /tch/ (Figures 8a and 8b). This is because the acoustic energy of nasals can be observed and measured in a low frequency range due, among other factors, to a relatively narrow opening nasal cavity (Reetz & Jongman, 2011; Stevens, 2000, p. 489). In contrast, the acoustic energy of the fricatives and affricates can be observed in as high a frequency as 8000 Hz due to the frication noise made when air hits the teeth and the palate before it

is released (Bolt et al., 1973; Stevens, 2000, p. 379). Figure 8 shows that a DCT-smoothed spectrum in a Hertz scale provides a better fit against the raw spectrum of the fricative /s/ and the affricate /tch/ (Figures 8a and 8b) than against that of the nasals /m/ and /n/ (Figures 8c and 8d). Worse DCT approximations for nasal spectra, especially in a high frequency range of ca. 1500-4000 Hz, may have different causes, including the fact that the low acoustic energy resulting in the frequency information was not well captured.

# 2.14.2 Bark-scaled DCT (cepstral) coefficients fitted to a raw spectrum

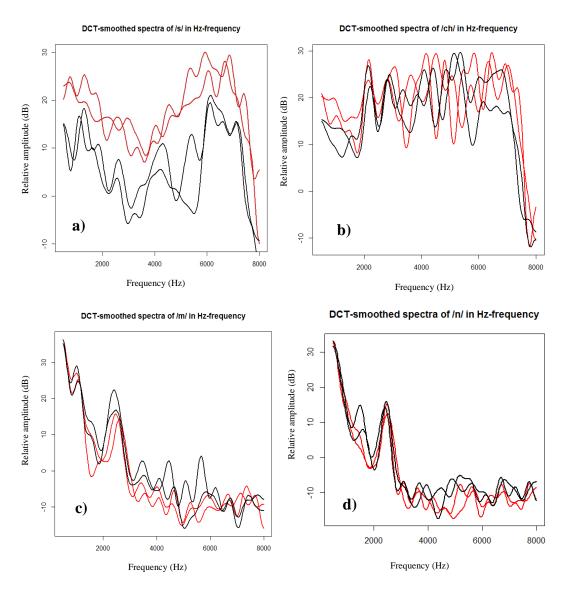
Figure 9 shows a raw spectrum (in black), extracted at the temporal midpoint of each of the target segments /s, te<sup>h</sup>, m, n/, plotted together with its DCT curve fitting (in red) in a



**Figure 9:** A DCT-smoothed signal (cepstrally smoothed spectrum) superimposed on the original spectrum (in black) by summing up the first 30 half-cycle cosine waves in a Bark scale (in red) of /s, tch, m, n/. Note: /tch/ is labeled as /ch/.

Bark scale. Figures 9a to 9d show that a raw spectrum of /m/ is best fitted by a cepstrally smoothed spectrum in a Bark scale by summing up the first 30 half-cycle cosine waves. In comparison with Figure 8c, the higher frequency region of Figure 9c is better approximated in a Bark scale than in a Hertz scale.

Figure 10 reproduces more samples of a raw spectrum extracted from Speaker 1, one of our male informants. It shows the speaker's 1<sup>st</sup> session (1<sup>st</sup> and 2<sup>nd</sup> repeats, in black) and 2<sup>nd</sup> session (1<sup>st</sup> and 2<sup>nd</sup> repeats, in red). Although data collection involved five repeats per session, only two repeats were plotted below to make visual inspection of within-speaker



**Figure 10:** A DCT-smoothed signal (cepstrally smoothed spectrum) of /s, te<sup>h</sup>, m, n/ from Speaker 1's 1<sup>st</sup> session, plotted by summing up the first 30 half-cycle cosine waves in a Hertz scale uttered on 1<sup>st</sup> and 2<sup>nd</sup> repeats (in black), and Speaker 1's 2<sup>nd</sup> session, similarly consisting of 1<sup>st</sup> and 2<sup>nd</sup> repeats (in red).

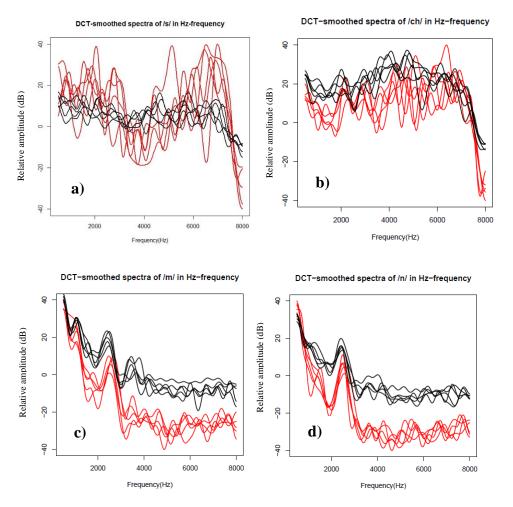
Note: /tch/ is labeled as /ch/.

variation easier. Each spectrum was extracted from the temporal midpoint of the target consonants; the x-axis is the frequency in Hertz, extracted from a frequency range of 500-8000 Hz, and the y-axis is the relative amplitude in dBs. The high-pass filter of 500 Hz is set to exclude potential distortion of the signal due to realistic transmission effects (Künzel, 2001).

Figure 10 illustrates within-speaker variation as shown by the spectral shape of each of the target segments of Speaker 1's 1<sup>st</sup> session (1<sup>st</sup> and 2<sup>nd</sup> repeats, in black) and 2<sup>nd</sup> session (1<sup>st</sup> and 2<sup>nd</sup> repeats, in red). A quick glance at the plots tells us that the spectra extracted from both sessions (black and red) uttered by Speaker 1 show a similar trend in their curvature for each of the target segments/s, tch, m, n/. Moreover, it can be seen in Figures 10c and 10d that, for the spectra of /m, n/ uttered during Speaker 1's 1<sup>st</sup> session (1<sup>st</sup> and 2<sup>nd</sup> repeats, in black) and 2<sup>nd</sup> session (1<sup>st</sup> and 2<sup>nd</sup> repeats, in red), there is greater overlap within each segment. In Figures 10a and 10b, there is less overlap in the spectra for /s/ and / tch/, showing much greater within-speaker variation.

The smaller within-speaker variations found in nasals are ideal and theoretically should provide a better result than the within-speaker variations found in /s/ and /te<sup>h</sup>/, if the degree of between-speaker difference is identical. Having said that, the curvature of a spectrum as found between speakers needs to be investigated, and I shall do so below. Figure 11 (overleaf) shows the curvature, in a Hertz scale, of the spectra uttered by two of our informants, Speaker 2 (1<sup>st</sup> session, 1<sup>st</sup> – 5<sup>th</sup> repeats, in black) and Speaker 3 (1<sup>st</sup> session, 1<sup>st</sup> – 5<sup>th</sup> repeats, in red). Not only the degree of between-speaker variation is shown, but also the within-speaker variation for each of the target segments /s, te<sup>h</sup>, m, n/. In order to see clearly how the curvature of each spectrum differs between speakers, all five repeats from session one uttered by Speaker 2 (in black) are plotted against those of Speaker 3 (in red) in a Hertz scale.

Looking at Figure 11, we can see different degrees of overlap for the curvatures found in Speakers 2 and 3 (plotted in black and red). In the case of the nasals /m, n/, the curvatures are clearly different from each other, especially in the higher frequencies (ca. 3000 Hz onwards). However, the spectra of the fricative /s/ and the affricate /tch/ show significant overlap from low to high frequencies, resulting in larger between-speaker variation than in the case of the nasals. Based on Figures 10 and 11, the nasals /m, n/ have smaller within-speaker and larger between-speaker variation than the fricative /s/ and the affricate



**Figure 11:** A DCT-smoothed signal (cepstrally smoothed spectrum) of /s, tch, m, n/ plotted by summing up the first 30 half-cycle cosine waves in a Hertz scale from Speaker 2's 1st session (1st – 5th repeats, in black) and Speaker 3's 1st session (1st – 5th repeats, in red).

Note: /tch/ is labeled as /ch/.

/te<sup>h</sup>/. As such, it is predicted that the nasals /m, n/ will perform better than the fricative /s/ and the affricate /te<sup>h</sup>/. This being the case, the relevant literature needs to be reviewed before we can confirm our final decision with respect to the selection of the target segments, as we shall see below.

# 2.15 Spectrum of the fricatives

To date, much acoustic research on spectrum has been focused on the mapping of distinct spectral patterns with phonetic characterizations, i.e. place and manner of articulation (Harrington, 2010), especially those of the English stop consonants (Kavanagh, 2012). There are considerably less FVC studies that searched for speaker specificity contained in other consonants such as the fricatives and affricates. It is established for English that

male and female speakers differ in their observed acoustic differences in the production of the fricative /s/ (Stuart-Smith, Timmins, & Wrench, 2003). The factors that are expected to contribute to such acoustic differences are: 1) the biological sex, i.e. a female's smaller vocal tract; 2) a female's shorter resonance cavity in front of the constriction; and 3) the more fronted articulation of the /s/ in females than in males (ibid.). These facts lead us to explore the Standard Thai fricative /s/ when searching for speaker-specific information that might be contained in its spectrum.

In addition to the use of /s/, the Standard Thai affricate /te<sup>h</sup>/ is also considered to be one of the FVC parameters in the current thesis. The reasons are as follows. First, based on advice from Rose (personal communication, 2015), commonly used words such as those with the same meaning as the English "yes" and "no" are worth exploring as they are most frequently used in everyday conversation. As such, the Standard Thai word [te<sup>h</sup>ai - HL] 'yes' is selected. Second, since dentition is involved in the articulation of the affricate /te<sup>h</sup>/ (Rose, 2011, p. 5900), which is assumed to vary largely between speakers, the airstream of /te<sup>h</sup>/ is expected to be individually distinctive. Although there have not been any results reported in the FVC literature concerning the use of /te<sup>h</sup>/, I believe it will be prudent to test the discriminatory performance of /te<sup>h</sup>/ in Standard Thai, due to the above reasons.  $\S 2.15.1$  introduces the articulatory description of the fricatives. This is followed by a literature review of various acoustic studies as well as previous FVC research, with particular attention given to the fricative /s/.  $\S 2.16$ , where we review previous FVC research on the spectrum of the affricate /te<sup>h</sup>/ and the nasals /m, n,  $\eta$ /, follows the same general outline.

## 2.15.1 Articulatory description of the fricatives

During the production of a fricative, there is a narrow constriction at some point along the vocal tract, from the larynx to the lips, with an approximate area of 0.1 cm<sup>2</sup> (Abramson, 1962; Shadle, 2012). In the case of the fricative /s/, the frication noise is generated when the airstream hits the teeth. Ladefoged and Maddieson (1986, p. 57) point out that:

...in a fricative a variation of one millimeter in the position of the target for the crucial part of the vocal tract makes a great deal of difference. There has to be a very precisely shaped channel for the turbulent airstream to be produced. [In] a stop closure the strength of the closure does not have to be constant throughout the gesture. But in many fricatives...an exactly defined shape of the vocal

tract has to be held for a noticeable period of time. These demands result in a fricative such as s having a greater constancy of shape in varying phonetic contexts, in comparison with the corresponding stops t, d, and nasals n (Bladon & Nolan, 1977; Lindblad, 1980; Subtelny & Oya, 1972).

We can summarize the information in this quote by saying, first of all, that, during the production of a fricative, a constriction must be formed (by active and passive articulators) in a very precisely controlled manner, in such a way that sufficient air flow rates meet the aerodynamic conditions to change laminar into turbulent airflow (Catford, 1977, pp. 183-201; Clark & Yallop, 1990). A second passage from the above quote ("These demands result in a fricative such as s having a greater constancy of shape in varying phonetic contexts") also has strong relevance to FVC. It means that the fricative /s/ is relatively free from a neighboring phonetic context, which would result in a relatively constant acoustic realization. It is also well known from the literature (e.g. Stevens, 2000) that the front teeth contribute to the acoustic characteristics of fricatives as they deflect the airflow that produces additional turbulence in the (dental, alveolar and post-alveolar) fricatives. Additionally, since dentition is expected to vary considerably between individuals, the fricative /s/ is in principle a good candidate for Standard Thai FVC. Although a promising FVC performance of the fricative /s/ has been previously reported for English (for an extensive review, see §2.15.3), the fricative /s/ has not yet been tested in Standard Thai.

## 2.15.2 Previous acoustic studies of English fricatives

In the literature, many acoustic cues have been reported for the identification of the exact place of articulation, with a distinction within the fricative class. Such acoustic cues include: 1) the spectral properties of the frication; 2) the amplitude of the frication; 3) the frication duration; and 4) the spectral properties of the transition from the fricative into the following vowels (Hughes & Halle, 1956; Zhang & Tan, 2008). Fricatives can also be grouped according to their place of articulation, their voicing, and the velocity of airflow (sibilants and non-sibilants) (Ladefoged & Johnson, 2014). In general, the sibilant fricatives /s, z,  $\int$ ,  $\int$ , which are perceptually high-pitched, are reported to be different from the non-sibilant ones /f, v,  $\theta$ ,  $\partial$ /, which are perceptually low-pitched, on the basis of spectrum, amplitude and duration of the frication noise (Jongman, Wayland, & Wong, 2000, p. 1253). Within each class, /s, z/ may be distinguished from / $\int$ ,  $\int$ / with respect to their place of articulation by spectral properties of the frication, while /f, v/ may be

distinguished from  $\theta$ ,  $\delta$  on the basis of spectral properties of the fricative-vowel transition (ibid., p. 1254).

In Jongman et al. (2000), both static and dynamic parameters were tested to classify the place of articulation of the English fricatives /s, z,  $\int$ ,  $\int$ , v,  $\partial$ ,  $\partial$ . I will limit the discussion to the static parameters that are relevant to the current work – spectral moments. *Spectral moments* are the mean, variance, skew and kurtosis. *Mean* and *variance* are defined as the average energy concentration and its range, respectively (Jongman et al., 2000, p. 1253). *Skew* is a symmetrical indicator for the energy distribution (ibid., p. 1253). That is, a skew of zero means that there is a symmetrical distribution of energy around the mean, but a positive skew is obtained when the right tail of the distribution extends further than the left tail (ibid., p. 1253) whereas a negative skew is obtained when the left tail of the distribution extends further than the right tail (ibid., p. 1253). Finally, *kurtosis* is an indicator of energy peakedness (ibid., p. 1253). Positive kurtosis means a well-defined spectrum with clear spectral peaks, while negative kurtosis indicates a flat spectrum without clear peaks (ibid.).

Jongman et al. (2000) found that spectral peak, mean, variance and kurtosis were significantly higher in females than in males. In contrast, the spectral skew of females was significantly lower than that of males (ibid.). This indicates that the spectra obtained from females were more well-defined: they displayed clearer peaks with a concentration of energy in the higher frequencies when compared to those obtained from males (Jongman et al., 2000). In light of the findings presented above, spectral information (spectral mean, variance, skew and kurtosis) of the fricative /s/ was proved to significantly reflect gender differences between males and females. As such, it is prudent to further examine the individualizing information, which might be contained in a fricative /s/ produced by Standard Thai male speakers. The reasons to select this particular alveolar fricative /s/ to represent the Thai fricative class /f, s, h/ are as follows. Although a labiodental fricative /f/ might be a good candidate, as it is phonetically (partly) dependent on dentition, empirical research on the typical French fricatives /f, s, š/ reveals that the tongue's body shape, when producing the labiodental fricative /f/, is influenced to a greater extent by the following vowels, whereas less influence is found for a dental fricative /s/ and palatalo-alveolar fricative /š/ (Stevens, 2000, p. 379). I therefore focus on

/s/ in this thesis. Needless to say, the investigation of other fricatives is a topic for future research.

With regard to the articulation of the glottal stop /h/, the vocal cords are opened wide enough to let the airstream pass through the glottis (but not vibrate) to produce the unvoiced frication (Rose, 2002, p. 239). In terms of source-filter theory, this turbulent airstream in the glottis acts as a source of energy that gets the air in the vocal tract to vibrate at its resonant frequencies (ibid.). A fricative /h/ is produced with "a spread glottal configuration with no significant constriction above the laryngeal region" (Stevens, 2000, p. 423). In other words, there is no modification of the airstream through the passage in the vocal tract. This being the case, /h/ is expected to contain less individualizing information than the other two fricatives /s/ and /f/, and as such is excluded from the current experiment.

Let us look at another acoustic study, by Stuart-Smith et al. (2003), on the Glaswegian English /s/ to justify why speech from a single biological sex should be examined in the first place in the current thesis. This experiment examined whether 1) the observed differences in the acoustic characteristics of /s/, obtained from males and females, differ according to biological 'sex'; 2) the acoustic characteristics of /s/ were influenced by 'gender', in terms of its role as a socio-cultural construct, in addition to biological 'sex'; and 3) to what extent the factor of class contributes to these differences (ibid.). The speech spectra were obtained from 31 speakers who were then divided equally according to their social groups: 1) middle-class women; 2) middle-class girls; 3) working-class women; 4) working-class girls; 5) middle-class men; 6) middle-class boys; 7) working-class men; 8) working-class boys.

The results revealed that 1) to a certain degree, men and women exhibit differences in their pronunciation of /s/; 2) gender (social factor) is a more explanatory variable than sex (biological factor), as there is a clearly observable difference between working-class and middle-class women pronouncing /s/; and 3) working-class women (as opposed to middle-class) tend to pronounce "retracted" variants of /s/ in the same way as men. The findings show that the fricative /s/ carries not only gender but also socio-cultural information, and that it bears some important forensic relevance in, for example, approximating the linguistic community of individuals. Given the differences in pronunciation of the Glaswegian English /s/ by males and females, it is prudent to

examine the FVC performance based on a single sex first (males have been selected for this thesis) in order to clearly observe the individualizing information that might be contained in male speech. Needless to say, female speech will be explored in future work.

## 2.15.3 Previous FVC research on the English fricative /s/

This section briefly reviews more recent research on FVC by Kavanagh (2012), with particular reference to the fricative /s/ (Kavanagh's research used the spectral properties of various English consonants to explore speaker-specific information in the English fricative /s/, and the nasals /m, n, ŋ/). To the best of my knowledge, Kavanagh (2012) is one of the first to undertake FVC research testing spectral moments extracted from a range of English consonants. Given Kavanagh's (2012) promising results, spectral parameters (spectral moments and cepstral coefficients) are chosen as parameters in the current work, and I will discuss this below. However, I will first provide a general description of the databases and experimental conditions used in Kavanagh's (2012) study. In Kavanagh (2012), there were five acoustic parameters for the fricative /s/ as shown in Table 13.

Acoustic Parameters	Datasets	Filter conditions
1. Normalized duration	30-speaker set	1. 500-4000 Hz
2. Center of gravity (COG) or spectral mean		2. 500-8000 Hz
3. Standard deviation (SD)	18-speaker set	1. 500-4000 Hz
4. Skew		2. 500-8000 Hz
5. Kurtosis		3. 500-16000 Hz
		4. 500-22050 Hz

**Table 13:** Acoustic parameters, datasets and filter conditions for the fricative /s/ (Adapted from Kavanagh, 2012, p. 348)

The second column in Table 13 shows that Kavanagh's (2012) speech corpora were divided into two datasets: 1) a full 30-speaker set; and 2) an 18-speaker set. To simulate the telephone transmission effect, there were two filter conditions for the full 30-speaker set (500-4000 Hz and 500-8000 Hz) and four for the 18-speaker set (500-4000 Hz, 500-8000 Hz, 500-16000 Hz, 500-22050 Hz).

#### 2.15.3.1 Acoustic measurement

There were both static and dynamic measurements for the segment /s/ in Kavanagh (2012). For the static measurement, spectral properties were calculated using a single 40 msec Kaiser 2 window, centered on the midpoint of the segment. For the dynamic measurement, 20 ms windows were used at the onset, midpoint, and offset of each segment of /s/. No pre-emphasis was applied to the spectra of the fricative /s/.

#### 2.15.3.2 Experimental results of /s/

#### 2.15.3.2.1 Results of the static measurements

Only the LR results of Kavanagh (2012, pp. 373-378) are reported in this section, as LR experiments are the most relevant to the current thesis, as we shall see now. In general, in the 30-speaker set, the LR results from both the 4000 and the 8000 Hz filter conditions were comparatively similar. However, the LRs done at 4000 Hz were considered to provide better results on the basis of a lower false positive rate and a lower  $C_{llr}$  (ibid.). In the 18-speaker set, 4000 Hz was also considered to perform better than the other three filter conditions, with a higher proportion of  $log_{10}LRs \ge +4$ , relatively low false positives (20%), false negatives (6%), EER (17%) and the second-lowest  $C_{llr}$  (0.55) (ibid.).

## 2.15.3.2.2 Results of the dynamic analysis

The dynamic results did not show much improvement over those obtained in the static measurements (Kavanagh, 2012, pp. 346-386). Kavanagh (2012) pointed out that better results might have been achieved using a single 40 msec Kaiser 2 window, centered on the midpoint of the segment (static measurement), rather than the 20 msec windows placed at the onset, midpoint, and offset (dynamic measurement) of each segment of /s/.

Given the promising results described above, for example with regard to the EER (the error rate in discriminating same speakers from different speakers), which is ca. 17%, I decided to work on the spectral properties extracted at the midpoint (static measure) of the voiceless alveolar fricative /s/ to search for speaker-specific information.

# 2.16 Spectrum of the affricates

This section provides an articulatory description and summarizes previous acoustic and FVC studies done on affricates. As previously mentioned, the selection of the Standard

Thai voiceless aspirated alveolo-patatal affricate /teh/ is motivated by Rose's (2002) suggestion that commonly used words, such as those with the same meaning as the English "yes" and "no", should also be explored. As such, apart from the nasal /m/ extracted from the word [mai HL] 'no', the affricate /teh/, extracted from the word [tehai HL] 'yes', was also selected. The literature surrounding the affricates is reviewed in §§2.16.2 and 2.16.3.

### 2.16.1 Articulatory description of affricates

Articulation of affricates involves the initial rapid release of a complete occlusion, formed by the articulators at the anterior end of the constriction (Stevens, 2000, p. 412). After this release, there is a period of frication, i.e. a constriction, that is formed "immediately posterior to the point of release", "is maintained for a few tens of milliseconds and is then released" (ibid.). In other words, an affricate is a stop followed by a fricative (Kent, 2002) and the articulators responsible for making affricates can be divided into two parts: 1) the anterior part, which forms the closure; and 2) a longer posterior part, which causes the frication (Stevens, 2000, p. 412). During this frication, the shape of the constriction must be adjusted so that the flow of air hits the appropriate obstacle (ibid.), such as the teeth in the case of the Standard Thai affricate / tch /.

#### 2.16.2 Previous acoustic studies of the affricates

There are various acoustic studies that use spectral moment features to contrast between places of articulation of the affricates (cf. Mays and Beckman, 2008). Liu, Tseng, and Tsao (2000) found that the Chinese affricates /ʧ, ʧ<sup>h</sup>/ and fricative /ɛ/ can be contrasted by their acoustic correlatives. That is, the Chinese affricates /ʧ, ʧ<sup>h</sup>/ have a higher initial burst but a shorter frication period than the fricative /ɛ/ (ibid.). In two Catalan dialects, the articulatory differences between the affricates /ʧ, dʒ/ "are better specified at the frication than at closure" (Recasens & Espinosa, 2007, p. 143).

In Urdu, the acoustic cues of affricates were tested in four native middle-aged (18-22 year old) speakers (Sheikh, n.d.). It was found that 1) only minute differences between the duration of closure and of friction are observed in Urdu; 2) for a voiced affricate /dʒ/, some of the speakers extend the voicing to the frication portion while others do not; 3) the ratio of the closure and friction duration varies significantly across speakers, i.e. some elongate the frication while others have longer closures; 4) no significant difference

between the duration of the preceding and final vowel is observed; 5) no exact relation is found between the duration of the following vowel and the frication part; 6) the ratio of F4 and F3, when going into the closure, is found to be sufficient to distinguish between speakers (although it is not explained how to measure and calculate the ratio); and 7) no actual pattern of the intensities (relative amplitudes) during the frication can be observed (this is hence regarded as an insignificant acoustic cue to distinguish between speakers) (ibid.).

So far, a general trend that can be observed from the literature review summarized above is that the duration ratio between the stop and frication portions of the affricates, on the one hand, and the frication portion itself, on the other, are the potential cues for distinguishing between speakers. This is not surprising as the frication portion itself provides more information (than a stop portion does) about the articulatory configuration, as reflected in spectra, e.g. the burst spectra of English /tf/ are relatively flat (Stevens, 1993, 2000). Since measuring the duration ratio between the stop and the fricative portions of the affricates seems to be too laborious for the current work under a limited timeframe, only the spectral moments (mean, variance, skew, and kurtosis) of the Standard Thai affricate /tch/ are to be tested and compared with the other Standard Thai consonant segments of /s, m, n/.

## 2.16.3 Previous FVC studies of affricates

To date, not much FVC and ASR research on affricates has been reported. However, n English affricate /tf/ was tested in Franco-Pedroso et al. (2012), together with various other consonants, and the MFCC was extracted. This ASR experiment tested the 2006 Speaker Recognition Evaluation (SRE) datasets, during which 219 male speakers were tested against the training datasets of 367 male speakers. It was found that  $C_{llr}$  and EER values were relatively high at 0.98 and 43.53, respectively. As previously indicated, a Standard Thai affricate /tch/, extracted from the word [tchai HL] 'yes', which is commonly used in everyday conversation, is worth exploring in FVC, although previous ASR studies have also reported such high  $C_{llr}$  and EER values.

# 2.17 Spectrum of the nasals /m/, /n/, and $/\eta/$

This section firstly aims to introduce the articulatory movements involved in the production of the nasals /m/, /n/, and  $/\eta/$ . The phonetic experiments, including the acoustic

and perceptual tests that exploited the individualizing information contained in nasals to identity speakers, are presented next. After that, previous FVC experiments are reported for English nasals.

## 2.17.1 Articulatory description of nasals

In the course of articulating nasals, the velum is lowered, if it is not already open (Ladefoged & Johnson, 2014). There is a complete closure at some point along the vocal tract (at the labials for /m/, at the alveolar ridge for /n/, and at the velum for /ŋ/), where there is no increased pressure behind such oral constriction (Stevens, 2000, p. 287). From an acoustic perspective, the vocal cords are held together and are vibrating to generate a voiced sound source (Ladefoged & Johnson, 2014). A flow of air and the majority of voice energy pass through the nasal cavity (ibid.). Sometimes, this energy passes through the constriction in the vocal tract, which helps modify the distinctive sound qualities for each of the nasals (ibid.). Once the nasal stop is finished, an oral closure may be released with no audible noise as the air pressure passes through the nose (ibid.). The selection of the nasals /m, n/ as parameters in the current thesis is justified based on previous acoustic and FVC studies, as shown in the following sections.

### 2.17.2 Previous acoustic studies of the nasals

Amino, Sugawara, and Arai (2006) investigated speaker individuality through an analysis of the nine Japanese consonants /t, d, s, z, r, j, m, n, n/. This experiment involved both perceptual and acoustic tests. In the perceptual test, the fourth syllable of a carrier sentence 'aCaCaCa', which contained each of the nine consonants, was manually excerpted and used as a stimulus. Five subjects who were familiar with the speakers were required to identify them. The study revealed that the subjects could identify the speakers better when the stimuli were nasal rather than oral sounds. There was also a tendency for voiced sounds (as opposed to their voiceless counterparts) to provide better speaker identification (ibid.).

In the acoustic test, the cepstral distance between pairs of the nine selected consonants /t, d, s, z, r, j, m, n, p/ was used to investigate the consonants' contribution to speaker individuality. The ratios of between- to within-speaker distance were calculated using the F-ratio metric. It was found that the ratios were greatest for the nasals /m, n/ and smallest for the stops /t, d/ (ibid.). Both the perceptual and acoustic experiments carried out by

Amino, Sugawara, and Arai (2006) therefore demonstrate that nasals carry higher individualizing information than other consonants. This provides justification for working on nasals. Although previous FVC studies on nasals present limitations in terms of the number of subjects involved, they provide results similar to those reported above.

## 2.17.3 Previous FVC studies of nasals

Nasals are known to have 1) low F1 due to a long resonant cavity (including pharyngeal, oral plus nasal branches); 2) low amplitude due to the relatively narrow opening of the nasal cavity; 3) increased formant bandwidth as the energy is absorbed by the walls of nasal and oral cavities; and 4) anti-formants (Reetz & Jongman, 2011, pp. 194-195), rendering extraction of accurate acoustic information difficult, especially in poor recording conditions. Besides, nasals are known to be subject to channel transmission effects (Rose, 2013a). Despite all these characteristics, it is still prudent to exploit the small amount of information available on this subject, to see how the nasals perform in FVC. This section summarizes previous FVC research in order to justify the use of the nasals /m, n/ as parameters in the current thesis.

Nasals have been found to contain promising speaker-specific information since the 1970s (Glenn & Kleiner, 1968; Su et al., 1974; Wolf, 1972). In a speaker identification study by Wolf (1972), /m, n/ ranked second and third, respectively, after F0. Moreover, Su et al. (1974) found that the spectral transition between /m/ and a following vowel contains highly idiosyncratic characteristics, and can be used to identify speakers better than the spectral transition between /n/ and a following vowel.

More recent LR-based FVC experiments that investigate the effectiveness of nasals in discriminating speech samples include those carried out by Yim and Rose (2012), who compared the effectiveness of the Japanese mora nasal /N/ and the Cantonese syllabic nasal /m/. In their experiment, the spectral envelopes were fitted using the cepstral coefficients and used as parameters (see §2.13 for the detailed explanation of cepstral coefficients). The results showed that the Japanese mora nasal and the Cantonese syllabic nasal yielded promising results.

More recent FVC research on the nasals /m, n, n/ was conducted by Kavanagh (2012). The acoustics of nasals were extracted by a single 40 msec Kaiser 2 window at the midpoint. §§2.17.3.1 and 2.17.3.2 summarize the results of /n/ and /m/ in terms of the

magnitude of the derived LR and a cost-based metric, namely log-likelihood ratio cost  $(C_{llr})$  and values (of which a detailed explanation is given in Chapter 3). For /ŋ/, only the descriptive results, excluding LRs and  $C_{llr}$ , were reported, as recalled in §2.17.3.3. It is worth remembering that the parameters used for nasals are 1) normalized duration; 2) center of gravity (COG) or spectral mean; 3) standard deviation (SD); 4) skew; and 5) kurtosis.

## 2.17.3.1 FVC results of English /n/

In this section I report on the FVC results of /n/ achieved by Kavanagh (2012, pp. 202-251). It was found that /n/ performs very well; the lowest C<sub>llr</sub> is at 0.47 (ibid.). As evident from the lowest C<sub>llr</sub> obtained for /n/ in this section, and for /m/ in §2.17.3.2, the two-parameter combinations were more promising than the individual parameters; when all predictors were parameterized, the results were again less promising. As such, it is prudent that in the current work I test the spectral parameters (mean, variance, skew, kurtosis) in order to look for which *segments* and *combinations* perform better than others, rather than testing the individual parameters or fusing all predictors together. This decision is supported by Rose (2002, p. 18), who states that the discriminatory power of each acoustic parameter is not equal (one might be more or less powerful than the other). Since time is usually limited for FVC investigations, it is worthwhile to find out which acoustic parameters have optimal power.

### 2.17.3.2 FVC results of English /m/

In terms of  $C_{llr}$ , two-parameter combinations such as COG plus SD, or COG plus spectral Peak, as opposed to the combination of all available parameters, performed best for /m/. They produced fairly strong consistent-with-fact LR values (e.g. Log<sub>10</sub> LR  $\geq \pm 4$ , Kavanagh (2012, pp. 151-201).

### 2.17.3.3 FVC results of English /ŋ/

Since the available tokens of  $/\eta$ / were relatively limited, no estimate for LRs was undertaken in Kavanagh (2012, pp. 252-289), except for the descriptive studies. ANOVAs were used to assess potential speaker identity on acoustic measures of the velar nasal  $/\eta$ /. The results confirmed once again that two-parameter combinations, such as COG and SD, were relatively high in *F*-ratios, which in turn suggests further investigation could be done in searching for speaker specificity. Based on the promising LR results of

the English /m/ and /n/, reported by Kavanagh (2012) and other scholars discussed so far, I think it will be prudent to search for speaker specificity that might be contained in the nasals /m, n/ of Standard Thai.

## 2.18 Fundamental frequency (F0)

## 2.18.1 Background knowledge

F0 is the "acoustical correlate of rate of vocal cord vibration" (Rose, 2002, p. 244). The acoustical F0 values are determined by the length and mass of the vocal cords of a given speaker (ibid., 245). Specifically, F0 has an inverse relationship to the length and mass of the vocal cords (ibid., 246): higher F0 values will result when shorter and lighter vocal cords are vibrating and vice versa (ibid.). Typically, males have thicker and longer vocal cords than females (ibid.). As a result, the F0 values of males are usually lower than those of females. The presence or absence of F0 also functions to contrast the voicing in speech segments (Rose, 2003, p. 4102). That is, when the cords are vibrating, the resulting sound is voiced. However, when the cords are not vibrating, the sound is voiceless (ibid.). It should be noted that *F0* and *pitch* are different, as the latter is the *perceptual descriptor* of the former (Rose, 2003, p. 4098). We explain the difference between *acoustical F0* and *auditory pitch* below.

Producing the English utterance "This is a train to Bangkok" with an increasing rate of vocal cord vibration (F0) on the last word to produce an *auditory rising pitch* signals a question (Nolan, 2014). On the other hand, decreasing the rate of vocal cord vibration (F0) on the last word to produce an *auditory falling pitch* signals a statement (ibid.). This use of pitch is called *intonation* and indicates, among others, discourse function (ibid.). In contrast, when pitch is used to distinguish the meaning of the words in languages such as Standard Thai, a different pitch will convey an entirely different word. For example, [pa: L] means "forest" and [pa: HL] means "aunt". This use of pitch is called *tone* (Rose, 2003, p. 4102).

In traditional FVC, F0 is one of the popular parameters that are expected to yield large between- to within-speaker variation (Rose, 2002, pp. 244-46; Nolan, 1983, p. 124). This might be due to the fact that the length and mass of vocal cords are biologically determined (Hudson, De Jong, McDougall, Harrison, & Nolan 2007). There are many

prima facie attractive features of F0 for FVC as it is 1) relatively robust in transmission channels; 2) not adversely affected by poor recording quality; and 3) easy to measure and extract as compared to other features such as F-patterns (F1 and higher formants), which are easily distorted by transmission channels (Hudson et al., 2007; Nolan, 1983). In early automatic speaker recognition (ASR) literature, promising results of F0 from various studies such as Atal (1972) were reported. However, as has been shown in the literature, F0 can be affected by many factors. These involve emotional states, heath conditions, linguistic genres, background noise and voice disguise (Elliott, 2000; Maekawa, 1998; Watanabe, 1998), all of which make voices difficult to discriminate forensically.

I elaborate more on voice disguise here as all other factors that affect F0 values have previously been discussed in §2.3.2. It is necessary for forensic experts to understand the basic concept of voice disguise that can affect the F0 values under investigation. Künzel (2007, p. 290) explains that voice disguise can be any kind of falsetto, pertinent creaky voice, the act of whispering, faking a foreign accent, or the pinching of one's nose while speaking. All of these result in a slight increase of mean F0 in both males and females (Künzel, 2001, p. 172). Synthesizing of voices using electronic devices, on the other hand, has been reported very rarely. What has been reported is that using someone else's voice and editing such speech on a computer has caused a lot of trouble in speaker recognition (ibid.). Künzel (2007) studied the effects of different kinds of voice disguise on auditory and automatic (acoustical) speaker recognition (ASR). He sampled speech data from 50 males and 50 females who read a written text that was designed to contain semantic, idiomatic and stylistic elements typical of a kidnapper's telephone call. The results show a correlation between the F0 of a speaker's natural speech behavior and the way in which speaker disguises his/her voice. That is, a speaker with a higher-than-normal F0 tends to raise his/her F0; one with a lower-than-normal F0 is likely to end up with a creaky voice (ibid.). The latter is clearly found in males more than in females, who are generally reluctant to dramatically change their voices (ibid.). The findings also show that pinching one's nose results in a slight increase in the mean F0 in both sexes (ibid.).

Voice imitation can be regarded as another form of voice disguise. A study on using dialect imitation for the purposes of voice disguise was conducted by Markham (2007). He conducted an experiment on eight Swedish speakers who were recorded speaking their native dialects and imitating three other speakers' voices. From the auditory analysis, the

effectiveness of dialect immitation depends on how well the imitator obscures one's native dialect, as opposed to how well one can convincingly imitate a given dialect. Zetterholm (2007), too, conducted auditory and acoustical experiments on Swedish impersonators (two professional imitators and one amateur); they imitated between six and nine target voices of well-known male Swedish politicians and TV personalities. One of the findings obtained in the auditory analysis was that the impersonators were flexible enough in their imitations to achieve the different target speakers' pitch (ibid., p. 198). This was confirmed by the findings of the acoustical experiment, which showed that the averaged F0 values of "some of the voice imitations are quite close to the target voices" (ibid., p. 200). This means that the F0 values, among others, are the fundamental speech features that imitators use to imitate other people's voices. As such, voice disguise is a very crucial factor; interference caused by voice disguise must be ruled out before speech samples are subjected to further FVC analysis as this will affect the alternative hypotheses and hence the strength of voice evidence.

#### 2.18.2 Tonal F0

So far, there have been multiple FVC experiments exploring the temporary structure of F0, i.e. how an F0 changes over a short period of time (Rose, 2002, p. 248). This short-term F0 is different from the long-term one (LTF0), where the average F0 values over a long stretch of speech are statistically estimated to determine their distribution (Rose, 2002, p. 248). FVC experiments using temporal F0 values have been previously conducted in English (Hudson et al., 2007), Chinese (Zhang & Enzinger, 2013), Standard Thai (Pingjai, 2011) and Cantonese (Li & Rose, 2012; Wang & Rose, 2012); whereas the first two studies extracted F0 values from spontaneous speech, the remainder extracted F0 from read speech.

### 2.18.3 Previous FVC research on tonal F0

This section briefly reviews the use of temporal variations of F0 as parameters in traditional FVC. Rose (2013b) empirically proved, from real casework, that temporal variations of F0 were powerful and could be used as evidence in a \$150 million telephone bank fraud case. The F0 time course was sampled at 1) the midpoint of the vowel /u:/, in *too*, and 2) the first target, midpoint, and peak of the vowel /æ/, in *bad*. When LRs of these *too bad* F0 values were combined with LRs from other acoustical parameters, such as the F-patterns of /o/, /u:/, /æ/, and the spectrum of /s/, extremely good LRs of ca. 11

million (log<sub>10</sub>LR > 4) were obtained. However, such high LRs should be approached with some *skepticism*, especially since LR calculation, at the time, did not take the correlation between acoustical segments into account. As such, Rose (2013b) discarded the putatively correlated LRs (LRs from the F-patterns in *not too bad*), and derived smaller LRs of 300,000 instead. Regarding Standard Thai, promising results were previously reported in Pingjai (2011), where the tonal F0 extracted from the read-out speech was parameterized by polynomical curves. As such, it is prudent to further investigate new segments, which can canvas more variations of F0 in Standard Thai, namely the Standard Thai (phonetic) diphthongs [5i, ai].

## 2.18.4 Long-term F0 distribution

Apart from tonal F0, long-term F0 (LTF0) will also be tested in the current thesis. LTF0 are the statistical tools used to model the distribution of F0 values. They consist of the 1) mean; 2) standard deviation (SD); 3) skew; 4) kurtosis; 5) modal F0; and 6) modal density over a long stretch of speech (Rose, 2002, p. 248). The question that may arise is how much speech we need to ensure that it characterizes a speaker, not the linguistic content, while keeping in mind that these long-term characterizations are valid only for a particular occasion (Rose, 2002, pp. 248-262). Nolan (1983, pp. 13,123) suggested that at least 60 seconds is needed to analyze such long-term characterizations. In contrast, Rose (1991, p. 241) found from seven Chinese dialects that less than 60 seconds is sufficient. However, Nolan (1983) and Rose (1991) agree that the amount of speech needed for longterm distribution might vary from language to language. In the current work, an utterance of one minute, as opposed to a shorter one, was chosen in accordance with the suggestion made in Nolan (1983). This is based on the assumption that longer durations will produce better individualizing information. It should be noted that the first four measures of LTF0 (i.e. mean, standard deviation (SD), skew, and kurtosis) are essentially the four moments previously referred to in §2.15.2. The last two measures are the mode (the most often occurring value) of F0 and F0's kernel probability density (the area under such F0 values), respectively. We will elaborate more on this when we discuss the LTF0 results.

### 2.18.5 Previous FVC studies on LTF0

A study conducted by Kinoshita (2005) showed that the *mean* of LTF0 itself was not promising. Three years later, Kinoshita, Ishihara, and Rose (2008) proved that by combining all six LTF0 distribution properties (mean, SD, skew, kurtosis, modal F0, and

modal density) promising FVC results with a lower EER of 10.7% were obtained. Kinoshita and Ishihara (2010) further improved their experiment by using four different F0 extractions and three different models to approximate the F0 distribution from spontaneous speech samples produced by 201 Japanese males. Specifically, F0 values were extracted in a Hertz scale and a logarithmic scale; delta features or the dynamic information of the F0 sequences were also parameterized (ibid.). It should be pointed out that these delta or dynamic features are very popular in automatic speaker recognition (ASR) as better accuracy was achieved when these delta features (dynamic features) were added to the static cepstral features such as MFCCs (Mel-frequency cepstral coefficients) (cf. Furui, 1986; Kumar, Kim, & Stern, 2011). It will be appropriate to point out the difference between the dynamic and static features here. The dynamic features, such as delta features and percentiles, can capture the F0 distribution better than static metrics such as mean and mode (Kinoshita & Ishihara, 2010). This is because delta features can approximate the F0 distribution at every 10% and 15% interval (ibid.). This means that more data were obtained with these dynamic features as opposed to a single data point extracted by the static metrics.

As such, in Kinoshita and Ishihara (2010), the delta F0 features were parameterized and defined as the difference between the two adjacent HzF0 values ( $\Delta$ HzF0<sub>i</sub> = HzF0<sub>i</sub> – HzF0<sub>i+1</sub>), and when one of these was equal to 0, they were excluded from the data. In addition, the shape of the F0 distribution was captured not only by the *six LTF0 statistical tools* but also by the *percentile techniques* (10% and 15% percentiles). The percentiles were measures of the F0 values at every 10% and 15% interval of the probability density function (PDF). By using percentiles, Kinoshita and Ishihara (2010) expected to accurately approximate the non-unimodal distribution of F0. To sum up, there were *four* F0 measured scales (HzF0,  $\Delta$ HzF0, Log<sub>10</sub>F0,  $\Delta$ Log<sub>10</sub>F0) and *three* F0 distribution models (six LTF0 measures, 15% percentiles and 10% percentiles), which resulted in a total of 12 permuted tests. The results showed that the percentile-based technique with a non-linear scale ( $\Delta$ Log<sub>10</sub>F0) performed best as it yielded consistent results and the 10% percentile was considered the most effective in terms of its reliability (EER 2.49%) (ibid.).

Based on these promising results, I decided to use six LTF0 measures (mean, SD, skew, kurtosis, modal F0, and modal density) as well as percentile-based techniques to model

the F0 distribution to capture speaker specificity in Standard Thai spontaneous speech. It is also appealing to use long-term distribution in the current work as the concept of long-term distribution analysis over a long stretch of speech can be applied not only to F0 but also to any other acoustical-phonetic parameters (Rose, 1991).

## 2.19 Formant trajectory

## 2.19.1 Background knowledge

Formants are the acoustical outputs of the transfer function that "correlate with the size and shape of the vocal tract" (Rose, 2002, p. 211). The formant with the lowest frequency is called F1, followed by F2 and so on (ibid.). The frequency at which there is maximum amplitude of energy is called the formant center frequency, which occurs roughly in 1000 Hz intervals for adult male speakers (ibid.). It is well known that each formant frequency has distinct properties. That is, F1 inversely correlates with vowel height and F2 correlates with the vowel backness/rounding (Nolan, 1983; Rose, 2002). The formant trajectory of diphthongs and triphthongs is of interest in FVC research (Morrison & Kondaurova, 2009; Zhang, Morrison, & Thiruvaran, 2011). This is because diphthongs and triphthongs involve up to two and three vocalic targets, respectively, giving up to three formants for each of these vocalic targets (Rose et al., 2006). As such, it is reasonable to assume that the formant trajectory of diphthongs and triphthongs contains more individualizing information than that of monophthongs. §2.19.2 reviews a number of previous studies on FVC employing the formant trajectory or the dynamic behavior of vowel formants as parameters.

## 2.19.2 Previous FVC research

McDougall (2004) tested the dynamic feature of formants or formant trajectory of the Australian English (AE) diphthong /ai/ extracted from five native speakers. F1-F3 through /ai/ were examined in equidistant time, normalized at 10% intervals and statistically tested using discriminate analysis (DA). The findings indicated that correct classification rates were often achieved between 88 and 95%, with the best performing being the nuclear-stressed /ai/ (as opposed to a non-nuclear fast speech /ai/). Although subjected to a DA experiment (as there was no further LR calculation), the results confirmed that the formant dynamics of an AE diphthong /ai/ can distinguish individual

differences in pronouncing this diphthong, as evidently shown by a high classification rate of at least 88%. In addition, the discriminatory power was improved when the increased numbers of variables both from a given formant or additional formants were combined.

Kinoshita and Osanai (2006), too, conducted an FVC experiment using the formants extracted from the AE diphthong /aɪ/ as parameters. The first target (T1), the second target (T2) and the slope of F2 in the glide between T1 and T2 of the diphthong /aɪ/ from 10 speakers were tested against three different speech styles ("Word" style, "Spelling" style, and "Sentences" style). It was found that the angle of the F2 slope was not robust against the three different speaking styles. However, comparatively speaking, the F2 slope performed as well as the T1 and T2 targets. The combination of all three parameters (T1, the slope of glide, and T2) yielded an equal error rate (EER) as low as 13.71% (as opposed to an EER ranging between 35.26% and 37.80% with one parameter, and between 16.83% and 32.02% with two parameters).

Rose et al. (2006) examined the formant trajectories of AE /aɪ/ sampled from 25 male native speakers. In their experiment, the test data were independent of the reference data: the F-pattern of the trajectories of the diphthong /aɪ/ from the 25 male AE speakers was tested against that from the Bernard (1967) dataset, which included 170 adult males. The necessity for forensic evidence evaluation means that the test data should be taken from a relevant population. Rose et al. (2006) decided to extrinsically evaluate the test data for these two reasons. Firstly, they aimed to see how well the Bernard dataset's F-pattern for the diphthong /aɪ/ represented the population of male AE speakers. Secondly, they aimed to see what kind of results they would get when this data, which was recorded a long time ago (in 1967), was tested against more recent data. The results turned out to be good with calibrated EERs between 8% and 10%.

Similarly, Morrison and Kinoshita (2008) conducted an FVC experiment using the formant trajectories of the AE monophthong /o/ from 27 Australian males. The target /o/s were embedded in reading sentences of the type "Hoe, H-O-E spells hoe." The first three formants (F1-F3) of each of the first and last words in this sentence frame were analyzed. The formant trajectories were fitted using 1) the quadratic and cubic polynomial coefficients and 2) the quadratic and cubic DCTs. Moreover, the formant trajectories were scaled in Hz and log-Hz, which were measured in both absolute and equalized time scales.

It was found that using F1 through F3 yielded substantially lower  $C_{llr}$  than using only F2 and F3, suggesting that F1 adds much speaker-specific information to F2-F3. In addition, equalized duration yielded lower  $C_{llr}$  than absolute duration. The results also showed that there were small differences between using coefficients and DCTs, quadratic and cubic orders and log-Hertz and linear-Hertz scales. The best performing parameter was found to be the cubic polynomials fitted to Hertz scales for equalized durations of F1-F3/F2-F3 trajectories. This being the case, cubic polynomials fitted to the normalized duration of F1-F3 in a Hertz scale will be trialed in this thesis.

Morrison and Kondaurova (2009) further examined the formant trajectories extracted from the diphthongs /aɪ/, /eɪ/, /oʊ/, /aʊ/, and /ɔɪ/ of 27 AE males. Each of these vowels was fitted with the polynomials and DCTs in different orders. The diphthongs were measured in linear-Hertz and log-Hertz frequency scales in both original and equalized time durations. The results showed that 1) DCTs outperformed polynomials; 2) third-order as opposed to lower-order curves yielded better results; 3) curves fitted using a linear-Hertz frequency scale outperformed those in a log-Hertz scale; and 4) curves that were fitted in equalized duration outperformed those in the original time scale. The diphthongs were ranked in order from best to worst as /eɪ/, /aɪ/, /oʊ/, /ɔɪ/, /aʊ/. Moreover, fusing these vowels using two formants (F2, F3) and three formants (F1-F3) yielded similar results, suggesting that the performance would not be extremely compromised by excluding F1 in forensic casework. Fusion also resulted in lower Ctlr and a complete separation between SS and DS comparisons was achieved.

Rose and Winter (2010) conducted the first FVC experiment on female voices using both the Gaussian mixture model-universal background model (GMM-UBM) and Multivariate likelihood ratio (MVLR) (we will elaborate on this in the following chapter). They parameterized the first three formants (F1-F3), extracted from the five long monophthongs of 20 "general" AE speakers. The results showed that MVLR outperformed GMM-UBM, judging from the lower EER and C<sub>llr</sub> values. Specifically, EERs less than 1% and a C<sub>llr</sub> of 0.04 were achieved in the fused MVLR system. All SS comparisons were also correctly discriminated with MVLR. Rose and Winter (2010) pointed out that such superior results of MVLR might be attributed to an estimate of *overall* rather than *specific* between- and within-speaker variance and that there might be

"less correlation between the MVLRs than the LRs from the GMM" (Rose & Winter, 2010, p. 45).

However, the findings of Rose and Winter (2010) contradict the results obtained by Morrison (2011), who empirically showed that the GMM-UBM substantially outperformed MVLR in terms of its reliability and validity when several different acoustical-phonetic units were fused. The inconsistency of the findings might be due to the fact that Morrison (2011) used smaller parameters (four coefficients) but larger amounts of data (16-20 tokens per speaker of each recording) (Zhang et al., 2011, p. 2283).

Zhang et al. (2011) are among the first researchers who tested the performance of the formant trajectories of the Chinese triphthong /iau/ extracted from a relatively large corpus of 60 female speakers. The first 20 speakers were used to produce background data to model the distribution of the features in the population, the next 20 speakers were used for the development data to train the logistic regression weight (we will explain this in Chapter 3), and the last 20 speakers were used for the test data. Two statistical tools were tested, i.e. MVLR and GMM-UBM, where the former is common in traditional FVC and the latter is popular in automatic speaker recognition.

In the MVLR test carried out by Zhang et al. (2011, p. 300), the different order DCT coefficients fitted to F1 through F3 trajectories were initially trialed in the development set. Once this was done, zeroth through fourth DCT coefficients of F2 and F3 were finally chosen. LRs were then calculated for each speaker pair in the development set and these LRs were used as the weights for logistics regression (pooled procedures for calculation of the calibration weights were used, i.e. those of acoustical-phonetic and automatic systems). The LRs for the test set were then evaluated and calibrated using these weights from the development set.

LRs from the test data of MVLR and those of GMM-UBM were fused in the second test. The results showed that incorporating the acoustical-phonetic /iau/ to a fully automatic system substantially improved the performance over a single automatic system in terms of validity, i.e. the  $C_{llr}$  of the fused system was about one third that of a single automatic system.

Zhang, Morrison, Enzinger, and Ochoa (2012) further tested the validity and reliability of the formant trajectories of Standard Chinese /iau/ when the recording conditions were mismatched, i.e. high-quality versus degraded conditions, using both microphone and mobile-to-landline conditions. Additionally, they tested the reliability of a human-supervised formant tracking system using a *FORMANT MEASURER* (Morrison & Nearey, 2011) versus the five fully automatic formant trajectory measurements. The results showed that the human-supervised measurements always outperformed the fully automatic formant trajectory measurements in terms of the C<sub>llr</sub> and LR values (after each of these measurements was fused with the MFCC baseline system extracted over the entire speech-active portion of each recording) (Zhang et al., 2012, pp. 11-29). Based on these findings, any formant tracking errors found in the current thesis are to be manually corrected (as I shall explain in Chapter 5).

Zhang et al. (2012, p. 29) also pointed out that any recordings obtained via mobile transmission channels were particularly problematic for a fully automatic formant tracker, hence they yielded worse results for both automatic baseline MFCC and human-supervised systems. In addition, there was also a tendency for *human-supervised measurements* to contribute more to FVC improvement than the MFCC-on-/iau/ system, when two same-channel conditions involving a landline telephone were involved. Having said that, there is a caveat that one should not generalize these results to any other phonemes, languages, or the gender of speakers (ibid.).

Among the more recent LR-based FVC experiments using formant trajectories as parameters are those carried out in Cantonese (Chen & Rose, 2012; Jialin & Rose, 2012; Li & Rose, 2012). These empirical studies showed that the formant trajectories of the Cantonese /ɔy/ (i.e. /ɔ/, a low back rounded vowel, followed by /y/, a high front rounded vowel), /iau/, and /ei/ yielded Log<sub>10</sub>LRs  $\leq$  2 with a C<sub>llr</sub> of 0.55, 0.6, and 0.46, respectively. When Log<sub>10</sub>LRs from /ɔy/ formant trajectories were fused with those of /iau/, the FVC performance was improved, resulting in log<sub>10</sub>LRs  $\leq$  2 with a C<sub>llr</sub> of 0.44 (ibid.). When the formant trajectories (F2-F3) of the /i/ rime were tested, log<sub>10</sub>LRs  $\leq$  1 with the C<sub>llr</sub> of 0.65 were obtained (ibid.).

Given the above empirical findings, I decided to use the formant trajectories extracted from the Standard Thai (phonetic) diphthongs [5i, ai] under human-supervised

measurements. This means that the formants were manually corrected if there were any tracking errors. Of course, this task seems to be arbitrary in that it is based on the visual judgments of a researcher, which further depends on his/her expertise. This being the case, the criteria used for the formant tracking correction were initially set out to ensure its reliability and consistency, as we shall see later. In the current work, the cubic polynomials fitted to the equalized duration of the first through third formant trajectories (F1-F3) of the Standard Thai (phonetic) diphthongs [ɔi, ai] were tested. The selection of all these parameters, i.e. third order polynomials fitted to the normalized duration of the formant trajectory, were based on the literature summarized above.

## **2.20 Summary**

In this chapter, we have discussed the factors that make less-than-ideal speech evidence difficult to discriminate forensically. In this regard, Bayes theorem and the LR framework have been justified as the theoretical and conceptual framework for this thesis. Standard Thai sound systems and Thai legal systems were also introduced in this chapter. This was followed by a literature review of multiple FVC studies, with particular reference to the use of 1) spectrum of the fricatives, affricates and nasals; 2) tonal fundamental frequency (F0); 3) long-term fundamental frequency (LTF0); and 4) formant trajectories. Based on these promising results reported in the literature, the following are tested in the current FVC work: 1) the spectrum of Standard Thai fricative /s/, affricate /te<sup>h</sup>/ and nasals /m/ and /n/; 2) the tonal F0 of Standard Thai (phonetic) diphthongs [5i, ai]; 3) long-term fundamental frequency (LTF0) of the information exchange task (involving two informants having a conversation based on obfuscated information given in a fax message and a relatively long spontaneous speech); and 4) formant trajectories of Standard Thai (phonetic) diphthongs [5i, ai].

# **Chapter 3**

# Methodology

### 3.1 Introduction

In this chapter I aim to discuss the five main concepts that are important in the current thesis: 1) the MVLR formula; 2) the speech corpus; 3) calibration; 4) fusion; and 5) the assessment metric employed. First, this chapter aims to review the concept of MVLR (Aitken & Lucy, 2004), a statistical tool that was originally developed to assess the strength of glass fragments evidence. Apart from its use for glass fragment assessment (cf. van Es, Wiarda, Hordijk, Alberink, & Vergeer, 2017), MVLR has been applied in other areas of forensics, for example the assessment of handwriting (cf. Bozza, Taroni, Marquis, & Schmittbuhl, 2008), finger print (cf. Neumann et al., 2007), text (cf. Ishihara, 2017) and voice evidence (cf. Morrison, 2009b). The second aim of this chapter is to explain the *protocol for the collection of databases for FVC research in Standard Thai*. Third, the mathematical notation of how to calibrate the derived scores into true LRs will be described. Fourth, the fusion of the derived LRs from different forensic systems, i.e. different sets of linguistic-phonetic parameters will be illustrated. Fifth, the metrics, the *log-likelihood-ratio-cost* or C<sub>llr</sub>, which is used to assess the validity of such LR outputs, will be explained. A leave-one-out cross-validation will also be discussed in this regard.

## 3.2 Likelihood ratio (LR) as the logical framework

The LR framework has been proposed as a standard framework for evaluating forensic science evidence (Aitken & Taroni, 2004; Evett & Buckleton, 1996; Lindley, 1977). As I discussed in Chapter 2, there are many advantages to using the Bayesian theorem as my model. Firstly, the LR, which is part of the Bayesian theorem, allows forensic experts to calculate and present the numerical and meaningful values of weight of evidence to the court (Robertson & Vignaux, 1995). Secondly, there is a clear distinction between the role of forensic experts and that of fact finders, leaving the court to incorporate the priors into their decision-making process (ibid.). Given these advantages, the LR framework has been widely accepted as a logical and legal framework in the forensic science community for evaluating forensic scientific evidence (Gonzalez-Rodriguez et al., 2007, p. 2105).

### 3.3 Statistical tools: the MVLR formula

In the current experiment, Aitken & Lucy (2004)'s Multivariate Likelihood Ratio (MVLR) formula is used to calculate the weight/strength of evidence.

The MVLR formula was developed at Edinburgh University's Joseph Bell Center for Forensic Statistics and Legal Reasoning (Aitken & Lucy, 2004). The original MVLR assumed normal distribution of speech samples, however this assumption cannot be made as speech data is likely to be non-normally distributed (Rose, 2002, p. 321; Alderman, 2005, p. 22). As such, Aitken and Lucy's (2004) MVLR formula is updated and includes a kernel density function to deal with the actual distribution of speech samples that can deviate from normality. Kernel density function is a combination of normal distributions; when they are combined, they can better approximate the actual distribution of speech samples. It is important to note that only the reference data (denominator) can be modeled with kernel density in this version of MVLR as the distributions of the suspect and offender samples (numerator) may be too sparse to be modelled with anything other than normal assumptions. Speech evidence normally needs to be quantified by multiple sets of parameters (e.g. F0 and formants) and these parameters are usually correlated as they are produced by the same vocal tract. Thus, Aitken and Lucy's (2004) MVLR is suitable for use in the current thesis as a forensic expert can take into account the correlation between the parameters extracted from speech evidence (Rose, 2013a, p. 92). When multiple LRs are derived from different FVC systems, i.e. different sets of parameters (F0 and formants) from the same set of speakers, these can be fused by using the Focal toolkit proposed by Brümmer (2007), so the results will show the overall strength of speech evidence.

In the MVLR formula in Figure 12, the numerator approximates the distribution of the offender and suspect speech samples (p (E|Hp)) using *the normal or Gaussian distribution model* (Rose, 2013a, pp. 94-95) while the denominator approximates the distribution of the reference data (p (E|Hd)) using *a kernel density function*. For this reason, the formula is sometimes named a multivariate kernel density (MVKD) formula (ibid.).

numerator of MVLR = 
$$(2\pi)^{-p}|D_{l}|^{-1/2}|D_{2}|^{-1/2}|C^{-1/2}(mh^{p})^{-1}|D_{l}|^{-1} + D_{2}^{-1} + (h^{2}C)^{-1}|^{-1/2}$$

$$\times \exp\left\{-\frac{1}{2}(v_{1} - \overline{v}_{2})^{T}(D_{1} + D_{2})^{-1}(\overline{v}_{1} - \overline{v}_{2})\right\}$$

$$\times \sum_{i=1}^{m} \exp\left[-\frac{1}{2}(v^{*} - \overline{x}_{i})^{T}\left\{(D_{l}^{-1} + D_{2}^{-1})^{-1} + (h^{2}C)^{T}(v^{*} - \overline{x}_{i})\right\}$$

$$(2\pi)^{-p}|C^{-1}(mh^{p})^{-2}\prod_{l=1}^{2}\left[|D_{l}|^{-1/2}|D_{l}^{-1} + (h^{2}C)^{-1}|^{-1/2}\right] \times \sum_{l=1}^{m} \exp\left[-\frac{1}{2}(\overline{v}_{l} - \overline{x}_{i})^{T}(D_{l} + h^{2}C)^{T}(\overline{v}_{l} - \overline{x}_{i})\right]$$
where  $U, C$  = within-, between-speaker variance/covariance matrices;  $n_{1}, n_{2}$  = number of replicates per speaker  $m$  = number of speakers in reference population;  $p$  = number of assumed correlated variables per speaker  $D_{l} = D_{1}, D_{2}$  = offender, suspect var/cov matrices =  $n_{1}^{-1}U, n_{2}^{-1}U$ 

$$h$$
 = optimal smoothing parameter for kernel density =  $(4/(2p+1))^{1/(p+4)}m^{-1/(p+4)}m^{-1/(p+4)}$ 

$$\overline{v}_{l} = \overline{v}_{1}, \overline{v}_{2} = \text{offender, suspect means; } y^{*} = (D_{1}^{-1} + D_{2}^{-1})^{1}(D_{1}^{-1} \overline{v}_{1} + D_{2}^{-1} \overline{v}_{2})$$

$$\overline{x}_{l}$$
 = within-speaker means of reference population.

Figure 12: MVLR formula (Aitken & Lucy, 2004)

 $m \rightarrow$  number of speakers in the background data

 $n_i \rightarrow$  number of tokens from each speaker in the background data

 $n_l \rightarrow$  combined number of tokens from suspect and offender data

 $p \rightarrow$  number of speech features

 $x_{ij} \rightarrow \text{background data measurement}$ 

 $y_{ij} \rightarrow$  suspect and offender data

 $D_1 \rightarrow \text{variance/covariance matrix of offender data}$ 

 $D_2 \rightarrow \text{variance/covariance matrix of suspect data}$ 

 $U \rightarrow$  within-speaker covariance matrix

 $C \rightarrow$  between-speaker covariance matrix

 $h \rightarrow smoothing parameter$ 

 $\overline{y}_I \rightarrow$  mean of offender data

 $\overline{y}_2 \rightarrow$  mean of suspect data

Some further explanation of the above mathematical notations is called for. The withinspeaker variation of the observation  $i^{th}$ , in the background data  $x_i$ , assumes normal distribution  $(x_{ij}|\theta_i, U) \approx N(\theta_i, U)$ , where N stands for normal distribution,  $\theta_i$  is the mean and U is a variance/covariance matrix U, i = 1, 2, ..., m and j = 1, 2, ..., n (Murphy, 2012; Rose, 2013a, pp. 94-95). Between-speaker variation is modelled by the kernel density function with a mean  $\mu$  and variance/covariance C (Rose, 2013a, p. 95). The similarity/dissimilarity of speech samples is measured by distances. This is expressed inside the exponent in the numerator of MVLR as  $(\bar{y}_I - \bar{y}_2)^T (D_I + D_2)^{-1} (\bar{y}_I - \bar{y}_2)$ , which is called the *Mahalanobis* distance (ibid.). We also observe that the inverted variance/covariance matrices of the suspect and offender,  $(D_I + D_2)^{-1}$ , are included to decorrelate the individual variables and equalize their contribution to the LR (Khodai-Joopari, 2006, p. 145). We further observe from the above MVLR that it involves the *Mahalanobis* distance of the offender and suspect mean vectors and variance ratio of the suspect and offender (Rose, 2013a, p. 93). Other complexities of the formula concern the kernel density function (such as the h smoothing parameter) to model the background data as well as the scaling of LR (ibid.).

Since the MVLR formula was originally developed for glass fragments with three or four input parameters, the use of many acoustical parameters (as is the case for FVC, e.g. eight coefficient values of the cubic polynomial fitted to F1 and F2) can result in under- or over- estimation of LRs (Morrison, 2009a). When the input parameters are larger than about four (due to sparse data with many parameters), the smoothing process required for the kernel density is likely to become difficult (Nair et al., 2014, p. 91), which will cause further computation problems in the inverses of the matrices. As such, erroneous LRs may result (ibid.). Such computation weaknesses in the MVLR formula may be due to many causes. As pointed out by Nair et al. (2014, p. 91), 1) the matrices of the offender and suspect,  $D_1$  and  $D_2$ , are required extensively in the MVLR algorithm; 2) the inverses of these offender and suspect matrices,  $D_1^{-1}$  and  $D_2^{-1}$ , are also used at several stages; and 3) these matrices,  $D_1^{-1}$  and  $D_2^{-1}$ , are then converted again as in the following term:

$$-\frac{1}{2} \left( y * - \overline{x}_i \right)^{\mathrm{T}} \left\{ \left( D_1^{-1} + D_2^{-1} \right)^{-1} + \left( h^2 C \right) \right\}^{-1} \left( y * - \overline{x}_i \right)$$

Since the aim of the current thesis is to explore the discriminatory power of Standard Thai FVC rather than to test the robustness of the computational algorithm, Aitken and Lucy's (2004) MVLR is chosen to assess the LRs because 1) the multiple input parameters (as compared to one parameter at a time by the Univariate Kernel Density; cf. Lindley, 1977) can be calculated at once; 2) MVLR takes correlations between parameters into account; and 3) promising FVC results using MVLR are widely reported in the available literature, as reviewed in Chapter 2.

## 3.4 Speech corpus

Experiments conducted to test and justify the real-world application of FVC should correspond as closely as possible to real-world conditions. Rose (2002) outlines several conditions necessary for FVC testing. One of them is the use of natural conversation. A second one, prompted by the need to perform within-speaker comparisons, is the use of different, non-contemporaneous recording sessions. Furui, Itakura, and Saito (1972) discovered that, whereas a long-term spectrum extracted over a period of between two/ three days and three weeks was found to be stable, significant shifts appeared when extraction occurred over a longer period of time. Likewise, Rose and Clermont (2001) empirically found that speech samples extracted from a single recording session achieved a 10% higher correct discrimination rate than those extracted from non-contemporaneous sessions (with an interval of at least a year). As such, it is important for the current thesis to simulate forensically realistic conditions where speech samples are separated. The third condition outlined by Rose (2002) for adequate FVC testing is the use of stratified sampling, i.e. a sample drawn from the same population, specifically of the same sex or with the same accent. In addition, experiments should be conducted on data obtained through different transmission channels, such as mobile phones and telephone landlines, to test levels of transmission-channel mismatch (ibid.).

To date, there have been several speech corpus development projects in Thailand. One of them was undertaken at the National Electronics and Computer Technology Center (NECTEC) some years ago (Kasuriya, Sornlertlamvanich, Cotsomrong, Kanokphara, & Thatphithakkul, 2003; Sornlertlamvanich & Thongprasirt, 2001). The NECTEC researchers collected spontaneous speech samples from a relatively large population (248 speakers). However, this database was put together for the purposes of speech recognition research, which means that the speech recordings were obtained in a single session. As previously mentioned, non-contemporaneous speech samples are necessary in FVC studies because the offender and suspect samples are non-contemporaneous. There is therefore a need to build up a corpus of speech samples that can be used to satisfy the above three requirements. In order to achieve such a goal, the researchers went back to Thailand to collect speech samples during three separate sessions with 60 male speakers. There were three tasks for each speaker to complete: 1) an information exchange task; 2)

a map task; and 3) the reading of sentences and words. The *protocol for the collection of databases for FVC research in Standard Thai* is described below.

### 3.4.1 Informants

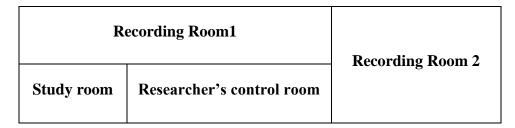
To model the characteristics of speech in my chosen population group, i.e. speakers of Standard Thai, I recruited 60 male informants, a number kept low on account of several constraints, such as time and budget limitations, but high enough to ensure FVC accuracy. As has been reported in the literature (e.g. Ishihara & Kinoshita, 2008; Rose, 2002; Hughes & Foulkes, 2014), FVC accuracy (validity) becomes relatively stable once around 30 speakers are involved. 60 informants are therefore regarded as a respectable number of speakers used for the purposes of the current thesis. I decided to collect speech samples from 60 males, aged between 22 and 60 years old, in a university setting. Informants were mostly students and staff at Thammasat University, Tha Prachan Campus, Bangkok, Thailand. This population was sampled because 1) they are native speakers of the Central Thai dialect, who represent the true population under investigation; and 2) it was convenient for me to recruit volunteers from this institution within a specific time frame. The reason male informants were chosen was that, statistically, males are more likely to commit crimes than females (Steffensmeier & Allan, 1996).

Recordings were conducted in a good-quality (not studio-quality) language laboratory at Thammasat University, Tha Prachan Campus, during three separate sessions. Standard recording conditions, which will be discussed in §3.4.2, were set out to ensure that speech data was of the same quality and produced using the same equipment. The protocol for speech data collection followed the *Protocol for the collection of databases of recordings* for forensic-voice-comparison research and practice by Morrison, Rose, and Zhang (2012). The number of the speakers from whom data were collected differed depending on the experiments.

### 3.4.2 Elicitation

Each informant was provided with recording manuals (see Appendix A) explaining all the tasks that they were required to complete. The handout was written in Thai. Two of the three tasks involved telephone conversations between two informants. Each of the two informants and I were in three acoustically separated rooms partitioned by a glass window (so that communication with the informants remained possible). That is, there

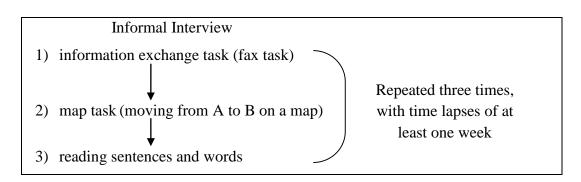
was one room for each informant and one room for myself to monitor the sound recordings. A floor plan of the recording room is shown in Figure 13.



**Figure 13:** Floor plan of the recording rooms

The microphones used were high-quality lapel microphones. They were connected via high-quality cable to a Roland® UA-25EX USB Audio Capture card, which was in turn connected to the researcher's laptop. *Audacity* software (Mazzoni & Dannenberg, 2000) was used to record speech samples. I was also equipped with headphones in order to monitor possible problems during recording sessions, such as background noises, poorly placed microphones that were likely to result in low speaker volume, or a misunderstanding of instructions among informants. Corrective action, such as the adjustment of microphone position and repeated explanation of instructions to informants, were undertaken as needed.

The computer used was a Lenovo laptop with a battery backup. The incoming signals were stored as WAV files at a sampling frequency of 44.1 kHz and at a 16-bit amplitude resolution. These speech signals were then downsampled to 16 kHz for the experiment. One speaker was recorded using input channel 1 and the other speaker was recorded using input channel 2. The different tasks each informant was required to perform are shown below. Similar to the protocol for speech data collection, the elicitation procedures (fax and map tasks) followed the *Protocol for the collection of databases of recordings for forensic-voice-comparison research and practice* by Morrison, Rose, and Zhang (2012).



Data collection first involved an interview between the informants and myself. In each of the recording sessions, two informants who knew each other (a deliberate decision, as it is easier to elicit a natural conversation from people known to one another) were asked to come to the recording room and provide information about themselves (nickname, education, place of birth, where they had been living, and the language they had used since they were young). In this way, the informants would feel more at ease and get more accustomed to my interview style.

Next, each informant was given a recording manual containing all the instructions for each task (see Appendix A). The first task was an *Information exchange task*. The rationale for selecting this task, as discussed previously, was to elicit as much natural speech as possible containing words and numbers commonly found in everyday conversation. Each informant was given a partially obfuscated fax message and was asked to exchange information about it with the other informant. It should be noted that fragments that were illegible for one informant were legible for the other and vice versa. The informants were asked to have a conversation on the internal telephone provided in the recording room and to write down the obfuscated information on the sheet provided.

In the second task, a map task, one of the two informants was asked to give his interlocutor directions to different places at Thammasat University's Rangsit campus. A map was provided to this end (see Appendix A). The interlocutor asked three questions, for example, how to get to Building No.59 if the starting point was Building No.9. The building *numbers* and building *names* were the targets.

In the last task, each informant was asked to read 36 sentences out loud. The sentences covered at least six tokens of several target consonants and vowels (/s, te<sup>h</sup>, m, n/ and [ɔi, ai]) and tones (low and high-falling). Some example sentences in which the target segments /s, te<sup>h</sup>, m, n/ and [ɔi, ai] are embedded are shown below, together with an English translation. The full reading list can be found in Appendix A.

1. /pʰrśʔ <u>mâi tcʰâi</u> nâ: tʰî: ʔá rai sàk <u>nòi</u>/

'This is because we do not have any responsibility.'

2. /læʔ mi: ka:n mun wi:an sà-ma:-teʰík| nai klum tæ làʔ klum| tʰúk săm ʔa:-tʰít |pʰŵ:aʔ hâ:i pʰû: ri:an mi: o:-kàt tʰî: teà tʰam ŋa:n| rû:am kàp pʰû: ri:an ʔwin ʔwin/ 'There is a member rotation for each group every three weeks in order for the students to participate with others.'

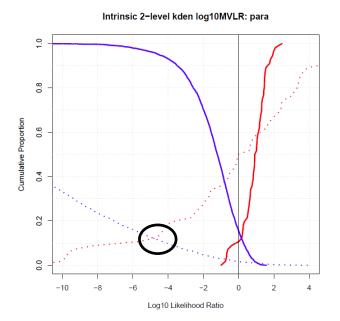
In total, 72 phonetic targets (6 target segments x 6 repeats x 2 tones) were excerpted from the recorded speech and used in the current work. There are several good reasons for recording speech in this way. First, different speaking-style mismatches are available, i.e. information exchange tasks vs reading tasks, to test forensically realistic conditions, as pointed out by Rose (2002). Second, sufficient tokens extracted from the reading tasks, which are embedded in the same phonological environment and recorded in the same recording conditions, are made available for use in experiments. This being the case, it is relatively easy to compare the results obtained from the different FVC systems, i.e. different target linguistic-phonetic segments extracted from the same recording conditions. (We can even compare the experimental results with those collected in clean conditions, and with those where speech samples will eventually become degraded, to test the effect of transmission mismatches in future research).

After reading the 36 sentences, participants were asked to read out the three most frequently used words (i.e. [thi: HL], a preposition meaning 'at'; [ka:n M], a prefix meaning '-ness'; and [nai M], a preposition meaning 'in') in a citation manner and six times in a row (see Appendix A). Instructions pertaining to pronunciation were kept to a minimum. The scripted words and sentences targeted in the third task were drawn from the LOTUS corpus developed at the NECTEC (National Electronics and Computer Technology Center) in Thailand by Kasuriya et al. (2003). LOTUS comprises the fewest sentences containing all possible phonemes in Thai. The vocabulary of LOTUS is taken from written text from Thai language magazines, encyclopedias, and journals.

# 3.5 Tippett plots

Having introduced the MVLR formula and speech corpus collected for use in the current thesis, I will now explain how to interpret the MVLR results using a Tippett plot (sometimes called a reliability or probability distribution plot). Tippett plots are now the

conventional way of presenting FVC results. Figure 14 shows the Tippett plot of 20 Hertz-scaled DCTs of /m/ - [mai HL].



**Figure 14:** Tippett plot of 20 Hertz-scaled DCTs of /m/ - [mai HL]. The (red) curves rising to the right represent the cumulative proportion of the SS (same speaker) comparisons, with log<sub>10</sub>LRs equal to or less than the value indicated on the x-axis, while the (blue) curves rising to the left represent the cumulative proportion of the DS (different speaker) comparisons, with log<sub>10</sub>LRs equal to or greater than the value indicated on the x-axis. Dotted lines and solid lines represent the uncalibrated and calibrated SS and DS Log<sub>10</sub>LRs, respectively.

An explanation of how to read the Tippett plot in Figure 14 is provided in the figure's caption (this practice will be adopted throughout the thesis) and is repeated here for the reader's convenience. The (red) curves rising to the right represent the cumulative proportion of the SS (same speaker) comparisons, with  $\log_{10}LRs$  equal to or less than the value indicated on the x-axis, while the (blue) curves rising to the left represent the cumulative proportion of the DS (different speaker) comparisons, with  $\log_{10}LRs$  equal to or greater than the value indicated on the x-axis. Dotted lines and solid lines represent the uncalibrated and calibrated SS and DS  $Log_{10}LRs$ , respectively.

In Figure 14, there are two types of errors, false negatives (SS comparisons, which were wrongly discriminated as coming from different speakers) and false positives (DS comparisons, which were wrongly discriminated as coming from the same speakers). There is a trade-off relationship between these two errors. That is, the number of false positives decreases while the number of false negatives increases when a  $Log_{10}LR = 0$ 

threshold moves up the LR scales (Enzinger, 2009, p. 47). Having said that, there is a point where the two errors are equal, which is the equal error rate (EER).

As evident in Figure 14, the two dotted curves cross at around  $log_{10}LR = -4.5$  (as marked by the black circle) and we get an EER of ca. 11%. This means that about 11% of both SS and DS comparisons were wrongly evaluated. Obviously, the scores (represented as the dotted curves) are not well-calibrated as 1) the crossing point of the EER is not at the theoretical threshold log<sub>10</sub>LR=0, but instead shifted to the left; 2) there are values of SS and DS comparisons that cross the threshold, suggesting wrongly discriminated speech samples (Rose, 2013a, pp. 97-98). The reasons why the LRs were not well-calibrated when MVLR was implemented for speech samples might be the inherent variability found in speech samples (as previously discussed in Chapter 2). As pointed out before, the MVLR formula was originally developed for glass fragments, where the within-speaker variance/covariance U is assumed to be constant (Aitken & Lucy, 2004). In other words, glass fragments are invariant in nature (they do not change over time). This is in contrast with the characteristics of voice evidence, where acoustical features are variant (e.g. voice changes according to health and emotions; cf. §2.2.1). Additionally, the MVLR formula is originally suitable for only three or four input parameters where a large background population is also assumed (ibid.). However, in the case of voice evidence, many input acoustical variables are evaluated, e.g. 15 cepstral coefficients extracted from only a small number of the population sample. Since the outputs of MVLR are not well-calibrated, we need a calibration method for speech evidence.

## 3.6 Logistic regression calibration

Following on from §3.5, the uncalibrated LRs, called *scores* and abbreviated as s, are to be converted into true LRs by the logistic-regression line in a logged odd space using Equation 3, where the weights ( $\alpha$  and  $\beta$ ) are usually obtained from scores based on an independent set of data (Morrison, 2013, p. 184). In other words, the s for the test set are calibrated by monotonically shifting (by amount  $\alpha$ ) and scaling (by amount  $\beta$ ) in a logged odd space using the weights from other speech samples, rather than those used for test comparisons, which are pooled together as background data.

#### $\log(LR) = \alpha + \beta s$

#### **Equation 3** (reproduced from Morrison, 2013, p. 184)

In this equation,  $\alpha$  is the Y intercept (where the line crosses the Y axis at x = 0),  $\beta$  is the slope of the line (how steep the line is) and s is the scores. The scores will be perfectly calibrated when the equation line is  $\log(LR) = 0 + 1$  x s (Morrison, 2013, p. 184), as the line perfectly crosses the Y axis at x = 0. Let us imagine a scenario where the scores are not perfectly calibrated and conversion into LRs is necessary. Suppose that the training data were shifted to the left, thus the straight line in a logged odd space is shifted one unit to the left (ibid.). To convert a score into a log LR, the equation line then becomes  $\log(LR) = 1 + 1$  x s. Imagine another scenario where the within-group variance of the data is increased (ibid.) by a factor of 4, and the slope of the line has quartered. In order to convert such a score into a log LR, the equation line becomes  $\log(LR) = 0 + 0.25$  x s. Once we calibrate these scores into LRs, we need another step to assess the accuracy of such LR outputs. Before I go further, I will break the flow of my discussion to explain why logistic regression in a logged odds space is preferred over the Gaussian models for calibrating the scores into LRs.

## 3.7 Why logistic regression is better than the Gaussian models

This section explains why logistic regression is preferred over Gaussian models. The reason is that logistics regression is *discriminative*: it models the *boundary* between same-speaker and different-speaker comparisons (Morrison, 2013, pp. 184-185). That is, logistic regression models can be shifted by amount  $\alpha$  and scaled by amount  $\beta$  as a function of a straight line, as shown in Equation 3 (ibid.). In contrast, Gaussian models are *generative*: they model the distribution for each of the same-speaker and different-speaker comparisons (ibid., pp. 182-183). Adding a very high score of 10 in the training data, which is known to be from the same speaker, will move the Gaussian model for same-speaker scores to the right (ibid., pp. 184-185). The variances for both SS and DS Gaussians will increase as well (ibid.). Logistic regression is the model least affected by such extremely high training data (s = 10), which is far from the boundary (ibid.). This being the case, logistic-regression calibration, particularly the FoCal toolkit (Brümmer & du Preez, 2006), is used in the current thesis to convert the scores into LRs.

To show the performance of the calibration process, let us look at the Tippett plot of the 15 Hertz-scaled DCTs of /tch/ - [tchai HL] in Figure 15.

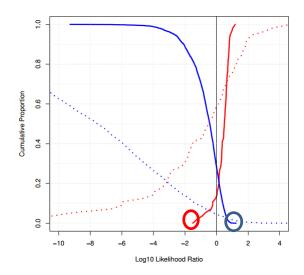


Figure 15: Tippett plot of the 15 Hertz-scaled DCTs of /te<sup>h</sup>/ [te<sup>h</sup>ai HL]. The (red) curves rising to the right represent the cumulative proportion of the SS (same speaker) comparisons, with the log<sub>10</sub>LRs equal to or less than the value indicated on the x-axis, while the (blue) curves rising to the left represent the cumulative proportion of the DS (different speaker) comparisons, with the log<sub>10</sub>LRs equal to or greater than the value indicated on the x-axis. Dotted lines and solid lines represent the uncalibrated and calibrated SS and DS Log<sub>10</sub>LRs, respectively.

In Figure 15, the most misleading Log<sub>10</sub>LRs for SS and DS comparisons (dotted lines; outside the range of the x-axis) were –13.99 and 5.04, respectively. After calibration, these were significantly reduced to –1.51 and 1.23 (marked by the red and navy blue circles). If a forensic expert presented the uncalibrated results to a court, such a misleading DSLR of 5.04 would inevitably lead the court to give undue weight to the voice evidence and this might lead to the conviction of a suspect as guilty, based on this voice evidence alone. Not only the magnitude of the misleading LRs was reduced, the magnitude of the correct SSLRs was also reduced from 4.67 (which previously suggests "very strong" support for the SS hypothesis) to 1.17 (suggesting "limited" support for the SS hypothesis). So far, we observe that the calibration process reduces the strength of evidence or the magnitude of both correct and incorrect LRs.

# 3.8 Metric for assessing the validity (accuracy) of MVLR

Accuracy is defined as the closeness of a given magnitude to its true value (Morrison, 2011, p. 92). As previously mentioned, the EER is the binary metric used to assess the

misprediction rate of the FVC system. The EER% metric indicates that the percentage of correct FVC predictions (LR > e) and incorrect FVC predictions (LR < e), where e is a threshold for SS comparisons that were wrongly discriminated as coming from different speakers and DS comparisons that were wrongly discriminated as coming from the same speakers (Rose, 2002, p. 13), are equal. However, since EER is based on a binary decision (correct vs. not correct), and it does not consider the actual magnitude of LRs (including both factual and counterfactual LRs) or calibration performance (Enzinger, 2009, p. 51), other metrics such as  $C_{llr}$ , which has been recently proposed in LR-based speaker recognition systems (Brümmer & du Preez, 2006), are selected to assess the system performance in the current thesis. To begin with,  $C_{llr}$  attaches cost to the LR scale (Morrison, 2011, p. 94) and the FVC system, which yields smaller  $C_{llr}$  and has better accuracy than the FVC system, which yields greater  $C_{llr}$  (ibid.).

Before I explain how to assess in detail the accuracy of the LR outputs, the concept of Cllr proposed by DeGroot and Fienberg (1983) in the form of a weather forecast, needs to be reviewed. When a forecaster makes a weather prediction for a given location during a specified time interval of the day, i.e. that it will or will not rain tomorrow (ibid.), the accuracy of such competing hypotheses is assessed by the strictly proper scoring rules, which can be thought of as *cost functions*: assigning penalty to the confidence level given to a particular hypothesis or the posterior probability (Brümmer & du Preez, 2006). These strictly proper scoring rules are optimized based on 1) "the probabilistic value of the forecast" and 2) "the true hypothesis which actually occurred" (Ramos-Castro, Gonzalez-Rodriguez, & Ortega-Garcia, 2006, p. 3). That is, if a forecaster states with high probability that it will rain tomorrow (probabilistic value), but it turns out that it does not actually rain (true hypothesis), a high cost is given to such a prediction and vice versa (ibid.). Thus, it is this probabilistic value, and how far it deviates from the true hypothesis, that optimizes the strictly proper scoring rules. Since C<sub>llr</sub> is actually the *strictly proper* scoring rules in FVC (Brümmer & du Preez, 2006), it is chosen as a metric to assess the validity of the FVC system in the current work. One of the great benefits of  $C_{llr}$ , among others, is that it can separately assess the discrimination power of LRs through discrimination loss (C<sub>llr</sub><sup>min</sup>) and the calibration (i.e. how far the individual LRs deviate from the truly occurring LRs, given that we know each LR value is from the same or from different speakers) through calibration loss ( $C_{llr}^{cal}$ ). As such, a well-calibrating system,

 $C_{llr}^{cal}$ , is smaller than  $C_{llr}^{min}$  (Ishihara, 2017, p. 189). In the current thesis, the FoCal Toolkit (Brümmer, 2007) is used to computationally evaluate the  $C_{llr}$  values.

Let us now discuss the  $C_{llr}$  formula by looking at Equation 4.

$$C_{llr} = \frac{1}{2} \left( \frac{1}{Nss} \sum_{i=1}^{Nss} \log_2 \left( 1 + \frac{1}{LRssi} \right) + \frac{1}{Nds} \sum_{j=1}^{Nds} \log_2 \left( 1 + LRdsj \right) \right)$$

**Equation 4:** C<sub>llr</sub> equation by Brümmer and du Preez (2006)

In this equation, LRss is a likelihood ratio for same-speaker comparison, LRds is a likelihood ratio for different-speaker comparison, Nds = number of different-speaker comparisons, Nss = number of same-speaker comparisons.

Equation 4 simply takes the mean of all the SSLRs (as shown on the left side within the outer brackets) and the means of all DSLRs (as shown on the right side within the outer brackets). Thus,  $C_{llr}$  is the mean of these two means.  $C_{llr}$  was formulated to severely penalize misleading LRs according to the degree of deviation from unity (Rose, 2013a, p. 101). Suppose we get a DSLR of 1500 (which is known to be a counterfactual analysis). This needs to be penalized to a degree that will depend on the magnitude of the derived counterfactual LR (ibid.). Thus, with  $\log_2(1+1500) \cong 10.55$ , one would be 1500 times more likely to get a difference between the suspect and offender samples, assuming that they had come from the same speaker (although they are in fact from different people) (ibid.). This, then, means that a value of 10.55 gives a high contribution to the average of all the different-speaker LRs and the overall  $C_{llr}$  values (ibid.). It should be noted that  $C_{llr}$  does not reward correct LRs, even though such correct LRs are high (ibid.). In other words, LRs with higher magnitude in support of a contrary-to-fact hypothesis will be attached with higher cost and vice versa (Gonzalez-Rodriguez et al., 2007, p. 1).

## 3.9 Logistic regression fusion

In this section, I discuss the final concept needed to understand the experiments conducted in this thesis, i.e. logistic regression fusion. Logistic regression fusion is employed in the current work to combine the parallel sets of scores from different FVC systems. Different FVC systems can be 1) the automatic vs. traditional FVC systems and 2) the different acoustical-phonetic systems, e.g. the system that extracts F0 from different vowels, or the

system that extracts the formant and F0 values from the same vowel (Morrison, 2013, p. 174). In this study, LRs obtained from the best performing parameters will be fused.

The procedure of logistic regression fusion is the same as that of logistic regression calibration. To explain this further, we firstly train the model using the training scores, which are known to be the results (scores) of SS and DS comparisons using the background data. The fusion and calibration weights were obtained in a cross-validated manner from the resultant scores. This means that in order to fuse and calibrate the scores into true LRs, the other scores were pooled together as training data for estimating the calibration weight.

For Fusion, the only requirement is that "each system must produce a score for each training comparison and for each test comparison" (Morrison, 2013, p. 189). The test scores of the different linguistic-phonetic parameters can then be fused and calibrated using the compatible training scores from each FVC system, where the linear regression fusion takes correlation into account (ibid.). Ideally, the test and training data should be independent of each other in order to avoid over-estimation of FVC performance (Kinoshita & Ishihara, 2014, p. 202)

Notably, fusion does not ensure the best performance when the high-performance LRs are fused (Franco-Pedroso et al., 2012, p. 188). This is because an LR from, e.g., the spectra of a nasal /m/ might be highly correlated with another LR from the spectra of a fricative /s/ as they are produced by the same vocal tract for each speaker to be fused. This means, more or less the same end results might be achieved, as there may be no complementary parameters. In our case, cepstral coefficients from the best performing linguistic segments /m, n/ from 57 speakers (as opposed to 60 speakers, due to poor recording condition and mispronunciation of the speakers) will be fused using Equation 5.

$$\log(LR) = \alpha + \beta_1 s_1 + \beta_2 s_2 + \dots + \beta_n s_n$$

**Equation 5** (reproduced from Morrison, 2013, p. 189)

In this equation,  $s_1$ ,  $s_2$ , ...,  $s_n$  are the scores from the first to  $n^{th}$  FVC systems, and  $\beta_1$ ,  $\beta_2$ , ...,  $\beta_n$  are the logistic-regression-coefficient weights according to the training scores for scaling. The logistic-regression-coefficient weights for shifting are  $\alpha$ .

The output of the fusion is thus the linear weighting of scores from multiple systems, which is actually the calibrated LRs (ibid., p. 187).

# **3.10 Summary**

This chapter has firstly presented the MVLR formula, which is used to calculate the strength of voice evidence (LR) in the current thesis. It subsequently described the current protocol for speech database collection for Standard Thai FVC. The speech database is not only meant for the current experiments, but will also help instigate further FVC research in Thailand. Next, the calibration process used to convert the scores into LRs was explained, and this was followed by discussion of the C<sub>llr</sub> metric used to test the accuracy of the LR outputs. Lastly, I explained the fusion of the FVC systems obtained from various acoustical segments using logistic regression.

# Chapter 4

# **Pilot FVC studies using Standard Thai diphthongs**

## 4.1 Introduction

This chapter reviews the three pilot studies I extracted from natural but controlled speech in 2013, which tested the FVC performance of the formant trajectories and tonal F0 of Standard Thai diphthongs. The purpose of these preliminary studies, which were conducted separately from those reported in Chapters 5-7, was to explore the discriminatory power of the formant trajectories and tonal F0 of Standard Thai diphthongs. All speech samples in these three pilot studies were selected from the speech corpus collected for use in the current thesis (see Chapter 3). The first study tested the performance of the F-patterns (F1-F4) of the Standard Thai diphthongs [i:aw], [u:a] and [u:a] from 15 native speakers, randomly selected from the corpus. The second study tested the F2 trajectories of the diphthongs [o:i] and [ə:i] from 54 native speakers. In the third study, the tonal F0 values of the diphthongs [ai] and [u:a] were extracted from 54 speakers and then used in the experiment. Since these three preliminary studies were conducted separately from the rest of the thesis (Chapters 5-7), the original Tippett plots presented at the 21st International Congress on Acoustics in 2013 are used to display the results. They use different conventions from those found in the subsequent chapters.

# 4.2 Pilot study on the Standard Thai diphthongs [i:aw], [u:a] and [u:a]

The first pilot study (Pingjai, Ishihara, & Sidwell, 2013) was published in the *Proceedings* of Meetings on Acoustics ICA2013 (Vol. 19, No. 1, p. 060043) and was titled: A Likelihood Ratio-based forensic voice comparison using formant trajectories of Thai diphthongs. I, the first author, conducted this research under the supervision of the second, while the third gave some advice. Before presenting the FVC results using Tippett plots, I will comment on the extracted parameters and the number of informants tested.

## **4.2.1** Parameters and informants

The F1-F4 trajectories of the diphthongs [i:aw], [u:a] and [u:a] were extracted from the randomly selected 15 male native speakers of Standard Thai in the corpus (see §3.4) by the Praat sound program (Boersma & Weenink, 2003). The coefficients of the cubic

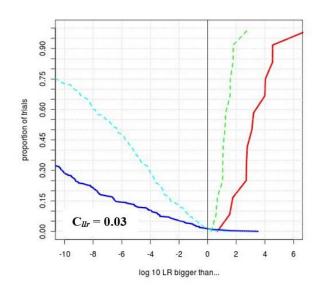
polynomials that were fitted to the F1-F4 trajectories of [i:aw], [u:a] and [u:a] were used as parameters. It should be noted that the samples of [i:a] were taken from occurrences in spontaneous speech of [li:aw H], "to turn left/right". Since approximant [w] in [li:aw] phonetically behaves like a vowel (no air turbulence in the air stream; cf. Ladefoged & Johnson, 2014), the acoustical values of [i:aw] were extracted instead of [i:a]. These three vocalic targets involved in [i:aw] were expected to exhibit greater between- to within-speaker variation as opposed to those of [u:a] and [u:a], which have two vocalic targets. The diphthongs [u:a] and [u:a], on the other hand, were taken from the read-out speech embedded in the words [phu:a? HL] 'in order to' and [su:an L] 'part/portion', respectively. Three different sets of formants were experimented on in the first pilot study: [F1, F2, F3], [F2, F3, F4], [F2, F3].

### **4.2.2 Results**

This section presents the experimental results in terms of Tippett plots. Discussion follows. Only the best performing parameters are presented.

Figures 16 to 18 show the Tippett plots of [i:aw], [u:a], and [u:a] when their cubic polynomials fitted to [F2, F3, F4] were used as parameters. The (dotted green and solid red) curves rising to the right represent the cumulative proportion of the SS comparisons, with the log<sub>10</sub>LRs equal to or less than the value indicated on the x-axis, while the (dotted light blue and solid navy blue) curves rising to the left represent the cumulative proportion of the DS comparisons with the log<sub>10</sub>LRs equal to or greater than the value indicated on the x-axis. Solid lines represent the uncalibrated SSLog<sub>10</sub>LRs and DSLog<sub>10</sub>LRs, while dotted lines represent the calibrated SSLog<sub>10</sub>LRs and DSLog<sub>10</sub>LRs (a leave-one-out cross-validation; see §3.6 on the difference between calibrated and uncalibrated LRs). Before proceeding to the discussion on how to read the results using conventional Tippett plots, the reader is recommended to go back to §2.4 (on how to translate Log<sub>10</sub>LRs into their verbal equivalents as proposed by Champod and Evett, 2000, p. 240) and §3.5 (on how to interpret the results using the Tippett plots).

Figure 16 (overleaf) shows that *all* SS comparisons were correctly discriminated. The best calibrated consistent-with-fact  $SSLog_{10}LRs$  obtained were  $SSLog_{10}LRs \le 3$  (dotted green line), suggesting "moderately strong" support for the same-speaker hypothesis. For DS comparisons, ca. 70% had calibrated consistent-with-fact  $DSlog_{10}LRs \le -4$  (dotted



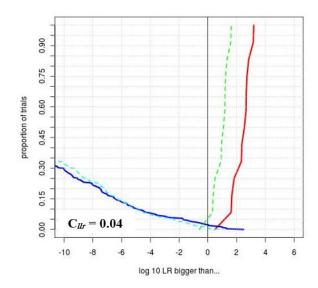
**Figure 16:** Tippett plot of [i:aw] - [li:aw H] when [F2, F3, F4] were fitted with cubic polynomials (reproduced from Pingjai et al., 2013).

The (green and red) curves rising to the right represent the cumulative proportion of the SS (same speaker) comparisons, with the log<sub>10</sub>LRs equal to or less than the value indicated on the x-axis, while the (navy blue and light blue) curves rising to the left represent the cumulative proportion of the DS (different speaker) comparisons, with the log<sub>10</sub>LRs equal to or greater than the value indicated on the x-axis. Dotted lines and solid lines represent the calibrated and uncalibrated SS and DS Log<sub>10</sub>LRs, respectively.

light blue line), suggesting "very strong" support for the defense hypothesis. The  $C_{llr}$  value for the segment [i:aw] - [li:aw H], when its [F2, F3, F4] were parameterized, was low at 0.03.

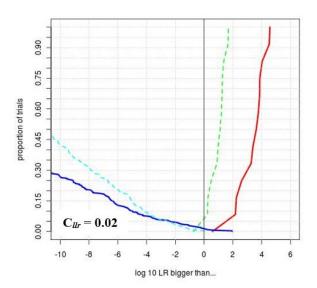
The results of [ui:a] - [ $\text{p}^{\text{h}}\text{ui:a?}$  HL], when its [F2, F3, F4] were used as parameters, are presented in Figure 17 (overleaf). Figure 17 reveals that ca. 5% of the calibrated SS comparisons (dotted green line) were incorrectly discriminated as coming from different speakers. The best calibrated consistent-with-fact SSlog<sub>10</sub>LRs obtained were SSlog<sub>10</sub>LRs  $\leq 2$ , suggesting "moderate" support for the same-speaker hypothesis. For DS comparisons, ca. 93% had calibrated consistent-with-fact DSlog<sub>10</sub>LRs  $\leq -4$ , suggesting "very strong" evidence in support of the defense hypothesis. The  $C_{llr} = 0.04$  for [ui:a] - [ $\text{p}^{\text{h}}\text{ui:a?}$  HL] is marginally higher than that of [i:aw], which is  $C_{llr} = 0.03$ .

The results of [u:a] - [su:an L] when its [F2, F3, F4] were used as parameters are presented in Figure 18 (overleaf). The  $C_{llr}$  value was lowest at 0.02 when its [F2, F3, F4] were used as parameters. About 5% of SS comparisons (dotted green line) were wrongly discriminated as coming from different speakers and the best calibrated consistent-with-fact  $SSLog_{10}LRs$  obtained were  $SSLog_{10}LRs \le 2$ , suggesting "moderate" support for the



**Figure 17**: Tippett plot of [w:a] - [phw:a? HL] when [F2, F3, F4] were fitted with cubic polynomials (reproduced from Pingjai et al., 2013).

The (green and red) curves rising to the right represent the cumulative proportion of the SS (same speaker) comparisons, with the log<sub>10</sub>LRs equal to or less than the value indicated on the x-axis, while the (navy blue and light blue) curves rising to the left represent the cumulative proportion of the DS (different speaker) comparisons, with the log<sub>10</sub>LRs equal to or greater than the value indicated on the x-axis. Dotted lines and solid lines represent the calibrated and uncalibrated SS and DS Log<sub>10</sub>LRs, respectively.



**Figure 18:** Tippett plot of [u:a] - [su:an L] when [F2, F3, F4] were fitted with cubic polynomials (reproduced from Pingiai et al., 2013).

The (green and red) curves rising to the right represent the cumulative proportion of the SS (same speaker) comparisons, with the log<sub>10</sub>LRs equal to or less than the value indicated on the x-axis, while the (navy blue and light blue) curves rising to the left represent the cumulative proportion of the DS (different speaker) comparisons, with the log<sub>10</sub>LRs equal to or greater than the value indicated on the x-axis. Dotted lines and solid lines represent the calibrated and uncalibrated SS and DS Log<sub>10</sub>LRs, respectively.

same-speaker hypothesis. For DS comparisons, ca. 90% of [u:a] - [su:an L] gave calibrated consistent-with-fact  $DSlog_{10}LRs \le -4$ , suggesting "very strong" evidence in support of the defense hypothesis.

All in all, with [F2, F3, F4], [i:aw], [u:a] and [u:a] performed comparatively well in terms of C<sub>llr</sub>. Specifically, the tested linguistic-phonetic parameters can be ranked in terms of  $C_{llr}$  values from low to high as [u:a] ( $C_{llr} = 0.02$ ), [i:aw] ( $C_{llr} = 0.03$ ), [u:a] ( $C_{llr} = 0.04$ ), respectively. Moreover, we observe that such  $C_{llr}$  values are marginally different. The underlying reason for [w:a] to perform worst in terms of C<sub>llr</sub> might be the articulatory movements it involves as compared to those of the other two vowels. As for [u:a], two vocalic targets (a high central unrounded vowel [u:] and a low central unrounded vowel [a]) are involved, but the tongue moves from high to low in central position without lip rounding. This is in contrast to the other two vowels, [u:a] and [i:aw], where the first vocalic target of [u:a] and [i:aw] (a high back rounded vowel [u:] and high front unrounded vowel [i:]) moves to a low central target [a], providing more space for the tongue to canvas in the vocal tract. Although [i:aw] ranked second in terms of  $C_{llr}$ , a magnitude of calibrated  $SSlog_{10}LRs \le 3$  was obtained, which is larger than for the other two segments, [u:a] and [u:a] (calibrated  $SSlog_{10}LRs \le 2$ ). Greater magnitude in terms of calibrated SSLog<sub>10</sub>LRs in the case of [i:aw] might be due to the fact that the vowellike "w" adds more individualizing information to the performance of [i:a] alone. That is, the triphthong [i:aw] might be pronounced with greater lip-rounding. Thus, different degrees of lip rounding (labialization) have presumably contributed to the higher individualizing information gained for [i:aw]. All in all, we can conclude that Standard Thai diphthongs were generally amenable to FVC.

# 4.3 Pilot study on the Standard Thai diphthongs [o:i] and [ə:i]

The second pilot study tested the FVC performance of the diphthongs [o:i] and [ə:i] embedded in the words [do:i M] 'by, with' and [khə:i M] 'used to', respectively. These speech samples were extracted from a reading task.

## **4.3.1** Parameters and informants

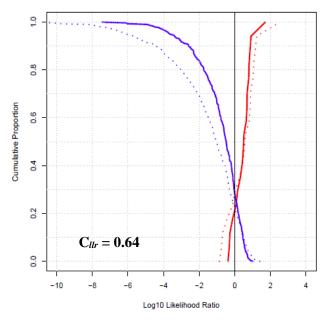
The (linear, quadratic, cubic) polynomials fitted to the F2 trajectories of the diphthongs [0:i] - [do:i M] and [9:i] - [kh9:i M] were obtained from 54 male native speakers. Since

low F1 values of the high vowels /i:, I, u/ are usually compromised by a telephone band-pass effect (Künzel, 2001, p. 89) and higher formants (F3) can also be affected by a mobile phone transmission channel (Rose et al., 2006, p. 331), only the F2 trajectories of the diphthongs [o:i] - [do:i M] and [ə:i] - [khə:i M] were chosen to simulate realistic conditions in FVC in this second pilot study. The duration of the F2 trajectories was also used as an additional parameter in order to see the improvement of FVC performance.

## **4.3.2 Results**

This section adopts the same convention as the previous one for the presentation of experimental results. The Tippett plots of the best performing parameters are presented first. Discussion follows.

Figure 19 shows that the F2-trajectories of [o:i] - [do:i M] performed best when its cubic polynomials (where duration was not included) were used as parameters.



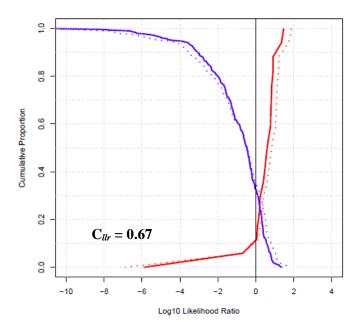
**Figure 19:** Tippett plot of [o:i M] - [do:i M] when its F2 trajectory was fitted by cubic polynomials.

The (red) curves rising to the right represent the cumulative proportion of the SS (same speaker) comparisons, with the  $\log_{10}LRs$  equal to or less than the value indicated on the x-axis, while the (blue) curves rising to the left represent the cumulative proportion of the DS (different speaker) comparisons, with the  $\log_{10}LRs$  equal to or greater than the value indicated on the x-axis. Dotted lines and solid lines represent the uncalibrated and calibrated SS and DS  $\log_{10}LRs$ , respectively.

Interestingly, adding the duration for [0:i] - [do:i M] did not improve the performance as the  $C_{llr}$  value was marginally higher at 0.66 (as opposed to 0.64 when duration was not

included). Setting the  $\log_{10}LR = 0$  as the threshold, ca. 75% of SS comparisons were correctly discriminated with calibrated consistent-with-fact  $SSlog_{10}LRs \le 2$ , suggesting "moderate" support for the SS hypothesis. For DS comparisons, only ca. 2% had calibrated consistent-with-fact  $DSlog_{10}LRs \le -4$ , suggesting "very strong" support for the defense hypothesis.

The best results of [a:i] -  $[k^ha:i M]$  when its F2 trajectory was fitted by cubic polynomials are presented in the Tippett plot in Figure 20.



**Figure 20:** Tippett plot of [ə:i] - [k<sup>h</sup>ə:i M] when its F2 trajectory fitted by cubic polynomials plus duration were parameterized.

The (red) curves rising to the right represent the cumulative proportion of the SS (same speaker) comparisons, with the log<sub>10</sub>LRs equal to or less than the value indicated on the x-axis, while the (blue) curves rising to the left represent the cumulative proportion of the DS (different speaker) comparisons, with the log<sub>10</sub>LRs equal to or greater than the value indicated on the x-axis. Dotted lines and solid lines represent the uncalibrated and calibrated SS and DS Log<sub>10</sub>LRs, respectively.

Figure 20 shows that the F2 trajectory of [ə:i] - [ $k^h$ ə:i M] performed best when cubic polynomials plus duration were parameterized, as a lower  $C_{llr} = 0.67$  was obtained, as opposed to  $C_{llr} = 0.78$  when only the F2 trajectory of [ə:i] - [ $k^h$ ə:i M] (duration was not included) was parameterized. Setting the  $log_{10}LR = 0$  as the threshold, ca. 90% of the SS comparisons were correctly discriminated as being from the same speakers. For DS comparisons, ca. 5% had calibrated consistent-with-fact  $DSlog_{10}LRs \le -4$ , suggesting "very strong" support for the defense hypothesis. As can be observed from Figures 19

and 20, the magnitude of the calibrated consistent-with-fact  $SSLog_{10}LRs$  is very similar between [o:i] - [do:i M] and [ə:i] - [k<sup>h</sup>ə:i M], i.e.  $SSlog_{10}LRs \le 2$ . On the other hand, the magnitude of the contrary-to-fact SSLRs is greater for [ə:i] - [k<sup>h</sup>ə:i M] than for [o:i] - [do:i M]. That is, the strongest contrary-to-fact  $SSlog_{10}LR = -5.9$  was obtained for [ə:i] - [k<sup>h</sup>ə:i M] while it was a contrary-to-fact  $SSlog_{10}LR = -0.3$  for [o:i] - [do:i M]. Such a strong contrary-to-fact  $SSlog_{10}LR$  contributed to a higher  $C_{llr} = 0.67$  for [ə:i] than for [o:i], where it is  $C_{llr} = 0.64$ .

## 4.4 Pilot study on Standard Thai diphthongs [ai] and [u:a]

A third pilot study tested the FVC performance of the F0 extracted from the diphthongs [ai] - [tehai HL] 'yes' and [u:a] - [ru:am HL] 'to share, to participate', using a reading task.

## **4.4.1 Parameters and informants**

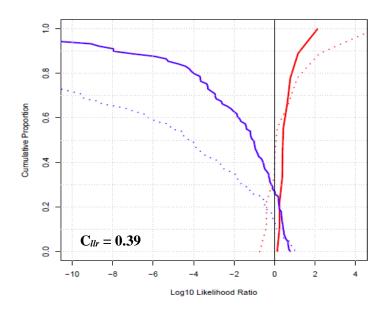
The (linear, quadratic, cubic) polynomials were fitted to the falling F0 contours of [ai] - [te<sup>h</sup>ai HL] and [u:a] - [ru:am HL]. There were 54 male speakers for [ai] and 30 speakers for [u:a].

#### 4.4.2 Results

This section presents experimental results in terms of Tippett plots of the best performing parameters.

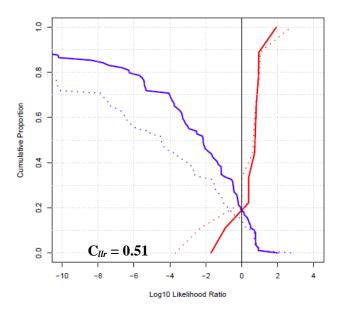
Figure 21 (overleaf) shows that all SS comparisons were correctly discriminated for [ai] - [tchai HL] when its F0 contour was fitted by the quadratic polynomials. This suggests that the second order polynomials sufficiently approximated the falling F0 contour of [ai] - [tchai HL], while the third order contour might be overfitted. A  $C_{llr} = 0.39$ , which is considered relatively low, was obtained. For DS comparisons, ca. 28% were wrongly discriminated as coming from the same speakers and ca. 20% had calibrated consistent-with-fact  $DSlog_{10}LRs \le -4$ , suggesting "very strong" support for the defense hypothesis.

The plot in Figure 22 (overleaf) indicates that the F0 contour of the diphthong [u:a] - [ru:am HL] performed best with linear polynomials (straight line). This is unexpected, as the literature suggests that a high-falling tone trajectory is likely to be best captured by



**Figure 21:** Tippett plot of [ai] - [tchai HL] when its F0 contour was fitted by quadratic polynomials.

The (red) curves rising to the right represent the cumulative proportion of the SS (same speaker) comparisons, with the log<sub>10</sub>LRs equal to or less than the value indicated on the x-axis, while the (blue) curves rising to the left represent the cumulative proportion of the DS (different speaker) comparisons, with the log<sub>10</sub>LRs equal to or greater than the value indicated on the x-axis. Dotted lines and solid lines represent the uncalibrated and calibrated SS and DS Log<sub>10</sub>LRs, respectively.



**Figure 22:** Tippett plot of [u:a] - [ru:am HL] when its F0 contour was fitted by linear regression.

The (red) curves rising to the right represent the cumulative proportion of the SS (same speaker) comparisons, with the log<sub>10</sub>LRs equal to or less than the value indicated on the x-axis, while the (blue) curves rising to the left represent the cumulative proportion of the DS (different speaker) comparisons, with the log<sub>10</sub>LRs equal to or greater than the value indicated on the x-axis. Dotted lines and solid lines represent the uncalibrated and calibrated SS and DS Log<sub>10</sub>LRs, respectively.

higher order (quadratic and cubic) polynomials, which can approximate the "U" and "S" shaped trajectory (Morrison, 2008, pp. 252-255). The results suggest that a high-falling (HL) tone trajectory of [u:a] - [ru:am HL] cannot be assumed for the diphthong [u:a] - [ru:am HL]. This finding also suggests that the vowel [a] in such a long diphthong [u:a] has undergone vowel reduction (Abramson, 1962, p. 76), resulting in [u:ə], which is best captured by a regression line rather than with higher order polynomials. With regards to SS comparisons, calibrated consistent-with-fact  $\log_{10} LRs \le 2$  were obtained for [u:a] - [ru:am HL], suggesting "moderate" support for the SS hypothesis. For DS comparisons, ca. 30% had calibrated consistent-with-fact  $DSlog_{10}LRs \le -4$ , suggesting "very strong" support for the defense hypothesis.

The results in Figures 21 and 22 show similar trends in terms of the strength of evidence (calibrated consistent-with-fact  $SSlog_{10}LRs \le 2$ ), except that all the same-speaker speech samples of [ai] - [tchai HL] were correctly discriminated when their F0 contours were fitted by quadratic polynomials. Moreover, the magnitude of the contrary-to-fact DSLR was larger for [u:a] than for [ai]. In terms of  $C_{llr}$ , [ai] - [tchai HL] performed better than [u:a] - [ru:am HL] as it yielded a lower  $C_{llr} = 0.39$  as compared to a  $C_{llr} = 0.51$ .

Based on the three pilot studies, we see that the formant trajectories of the Standard Thai (phonetic) diphthongs [i:aw], [u:a], [u:a], [o:i], [ə:i] and the tonal F0 of [ai] and [u:a] contain promising speaker-specific information. Further investigation into other Standard Thai diphthongs is thus warranted to see if similar results would be obtained, and which diphthong performs better than the other in terms of  $C_{llr}$  and  $log_{10}LR$  magnitude. I therefore decided to investigate the FVC performance of the formant trajectories (F1-F3) and tonal F0 of the Standard Thai diphthongs [ɔi] - [nɔi L] and [ai] - [mai HL], as we shall see in subsequent chapters.

# 4.5 Summary

So far, we have shown that the formant trajectories of the Standard Thai (phonetic) diphthongs [i:aw], [u:a], [o:i], [o:i] and tonal F0 contours of Standard Thai (phonetic) diphthongs [ai] and [u:a] are generally amenable to Standard Thai FVC. This is the best result in terms of  $C_{llr}$  value, when as low a level as 0.02 was obtained when F2-F4 of [u:a] were parameterized with cubic polynomials. Moreover, all SS comparisons

were correctly discriminated for [i:aw] with a low  $C_{llr} = 0.03$ . With regards to the tonal F0 contours fitted by polynomials, the lowest  $C_{llr} = 0.39$  was obtained when F0 contours of [ai] - [tchai HL] were parameterized with quadratic polynomials. All SS comparisons were also correctly discriminated for [ai] - [tchai HL]. Given these promising results, the first through third formant trajectories (F1-F3) (as opposed to only F2) and the tonal F0 contours of the Standard Thai diphthongs [5i] - [n5i L] and [ai] - [mai HL] will be tested further in the current work.

# Chapter 5

# Results of the spectral moments of /s/ and cepstral coefficients (CCs) of /s, te<sup>h</sup>, n, m/

## **5.1 Introduction**

This chapter reports on the strength of voice evidence (LR) of a spectrum extracted from the midpoint of the linguistic-phonetic target segment /s/ using a statistical analysis of mean, variance, skew, kurtosis (i.e. the *spectral moments*) tested in ANOVA and MVLRs, respectively. Apart from the spectral moments, cepstral coefficients (CCs) extracted from the segments /s, tch, n, m/ will also be tested in MVLRs. Thus, this chapter is divided into three main parts. The first part discusses the segmentation criteria used to locate the starting and end points of the target segments. The second part reviews the basic knowledge of spectrum and the statistical concepts of spectral moments (mean, variance, skew, kurtosis). Then, the results of spectral moments tested in ANOVA are first presented as a preliminary analysis before those of MVLR. In the third part I report the MVLR results of cepstral coefficients (CCs) based on the C<sub>llr</sub>, Log<sub>10</sub>LR and EER values, in both Hertz and Bark scales. Discussion follows.

## 5.2 Segmentation criteria

The criteria used to annotate the Standard Thai segments /s,  $te^h$ , n, m/ are presented in this section. I also give brief background information about the reasons why these particular segments were chosen for the current work. As mentioned in Chapter 2, the study conducted by Kavanagh (2012) showed that the fricative /s/ had promising discriminatory power in (British) English. Moreover, the English fricative /s/ had a lower  $C_{llr} = 0.88$  than the fricative /f/, whose  $C_{llr} = 0.97$  (as reported in the ASR research undertaken by Franco-Pedroso et al., 2012, which tested many English phonemes produced by male speakers of the NIST-SRE datasets using GMM-UBM). Additionally, FVC research by Rose (2013a), who based his study on cepstral spectra extracted from the Japanese alveolopalatal fricative [ $\epsilon$ ] (which is similar to the English palato-alveolar fricative /f/), anticipates promising results, especially when GMM-UBM and MVLR are fused ( $C_{llr} = 0.26$ ; EER = 74%). To elaborate, half of the DS comparisons were smaller than

DSlog<sub>10</sub>LR = -2 in the MVLR system. In contrast, half the DS comparisons were smaller than DSlog<sub>10</sub>LR = -1 in the GMM-UBM system. The largest counterfactual DSlog<sub>10</sub>LR = 2 was obtained for MVLRs but DSlog<sub>10</sub>LR = 5 was obtained for GMM-UBM, which is undesirable as, from a legal perspective, an expert would wish to avoid convicting an innocent person on the basis of such speech evidence alone (ibid., p. 5902). This contradicts the results of SS comparisons where GMM-UBM outperformed MVLR in its magnitude, i.e. about 25% of SS comparisons had  $log_{10}LR \ge 3$ . Rose (2011) concluded that LRs obtained from the Japanese alveolo-palatal fricative [ $\epsilon$ ] are likely to be of use in FVC, provided that they are combined with the LRs from other segments. Based on the above findings, it is prudent to search for speaker specificity that might be contained in the Standard Thai voiceless fricative /s/.

Given the above, I was aware how important it would be to make sure that all of the target segments experimented on in the current work, /s, tch, n, m/, are extracted from the same phonological environments to assure their comparability. Having said that, I am also aware that the choice of these Standard Thai segments will mean the results obtained might be different from those of previous studies in the FVC literature; this is because the individualizing information contained might be language-specific.

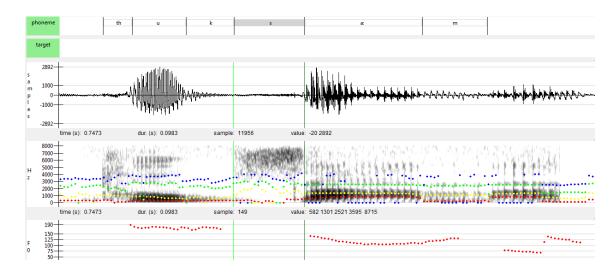
In §§5.2.1 to 5.2.5, I describe the criteria used to locate the target segments /s, tch, n, m/, respectively.

## **5.2.1 Segmentation of /s/**

The fricative /s/ was extracted from the word [sa:m LH] used in the sequence

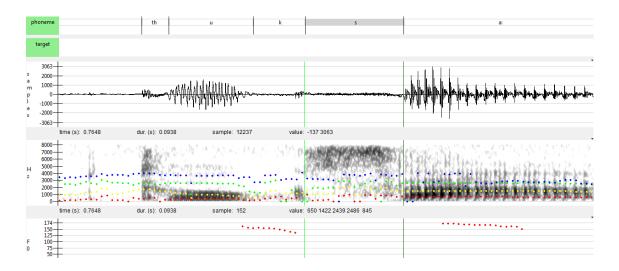
"every three weeks"

The word [sa:m LH] 'three' was chosen because it is a numeral, and as such it is frequently used in everyday conversation. Additionally, in this sequence /s/ occurs after a voiceless unaspirated stop /k'/, which makes it easier to segment since it has clear boundaries, as shown in the spectrogram in Figure 23 (overleaf). The segmentation of /s/ involved the simultaneous consultation of the audio-speech waveform and its



**Figure 23:** Label tier (top), waveform (middle), and spectrogram (bottom) of the phrase "every three weeks", with overlaid formants. The section highlighted in grey in the label tier shows the target segment /s/.

spectrograms. The beginning of /s/ is located at the point where the aperiodic noise first appeared in its spectrograms and the waveforms. The fricative offset was clearly defined at the endpoint of the frication noise and simultaneously at the beginning of a periodicity of the following vowel /a:/. However, there are instances where a voiceless stop /k/ was released after a vowel /u/ - [thuk], as shown in Figure 24.



**Figure 24:** Label tier (top), waveform (middle), and spectrogram (bottom) of the phrase "every three weeks", with overlaid formants. The section highlighted in grey in the label tier shows the target segment /s/.

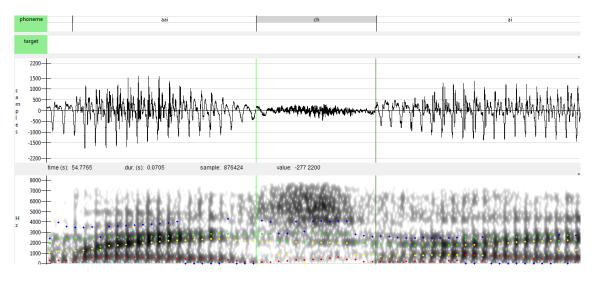
Figure 24 shows a label tier, a waveform, and a spectrogram of the phrase "every three weeks", with overlaid formants. It shows an instance where a voiceless stop /k/ is released after a vowel /u/ in [thuk]. The annotation criterion for the starting point of /s/ was

subsequently changed so that it would come *after* a released stop, as indicated in the first vertical green line in Figure 24. As is well known, the fricative energy utilized by a participant is expected to range from ca. 2000-8000 Hz or higher (Stevens, 2000). However, in forensically realistic conditions, /s/ energy reaching above 4000 Hz might not be recorded by the equipment due to the telephone band-pass filter (Byrne & Foulkes, 2007). It is therefore prudent to operate with two frequency bands: 500-4000 Hz and 500-8000 Hz, to test such realistic telephone-transmission conditions (Kavanagh, 2012); a lower band-pass of 500 Hz is found to be particularly advisable for the same reasons.

# 5.2.2 Segmentation of /tch/

The Standard Thai affricate /tch/ was extracted from the word [tchai HL] 'yes', used in the following sentence frame:

The above sentence can be translated in English as "This is because we do not have any responsibility". The highlighted /**m**, **te**<sup>h</sup>, **n**/ are the target segments and the underlined words represent stress. The acoustical properties of the target [te<sup>h</sup>] - [te<sup>h</sup>ai HL], excerpted from the above sentence frame, are shown in Figure 25, where the target segment [te<sup>h</sup>ai HL] is marked as [c<sup>h</sup>ai HL], as in the EMU speech database system (Cassidy, 1999).



**Figure 25:** Label tier (top), waveform (middle), and spectrogram (bottom) of part of the sentence frame "This is because we do not have any responsibility", with overlaid formant tracking values. The section highlighted in grey in the label tier shows the target segment /te<sup>h</sup>/.

An affricate can be defined as a sequence consisting of a stop plus a fricative (Turk, Nakai, & Sugahara, 2006). However, such an abrupt change in the amplitude of /teh/ cannot be clearly observed, since the duration of the stop part is too short. As such, the onset of [teh] was indicated by the start of an aperiodic waveform at a 2000-8000 Hz frequency and by aperiodic energy, as marked by the first vertical green line above. The offsets of [teh] were then marked at the cessation of the aperiodic waveform in conjunction with the F2 onset of the following vowel.

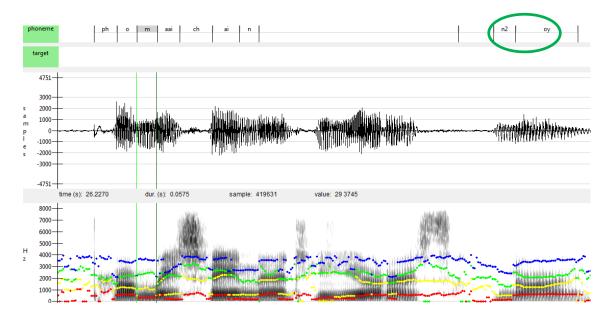
## 5.2.3 Segmentation of /n/ - [noi L]

Before going into the segmentation criteria of the target segment /n/ - [nɔi L], which is one of the particles most commonly used in spoken Standard Thai language, let us briefly discuss the usage of this sentence-final particle. Its primary function is as an "action-inducement utterance", hence its use in commands, invitations and suggestions (Cooke, 1989). According to Cooke (1989, p. 3), Standard Thai particles can occur in sentence-final position in sequences of up to six particles, whereas they are less likely to occur in a word-medial position. In addition, there are at least four particles that signal questions, three that signal commands, and half a dozen that are used for conversational or situational responses, more than half a dozen that signal speaker-addressee relationships, and some for other types of information (ibid., p. 2). These different particles can have their variants in terms of shades of meaning and form (ibid., p. 2). Having said that, the function of [nɔi L], selected in this thesis, is to soften commands and requests and it occurs sentence-finally. Some examples are given below.

kho [LH] 1. na:m [H] noi [L] **IPA** bring water please word-by-word gloss "Bring me some water, please" meaning 2. pai[M] klai [M] klai [M] noi [L] **IPA** far please word-by-word gloss go far "Go away from me, please" meaning

All the [noi L] samples used in this experiment were extracted from the following sentence frame, which was previously mentioned:

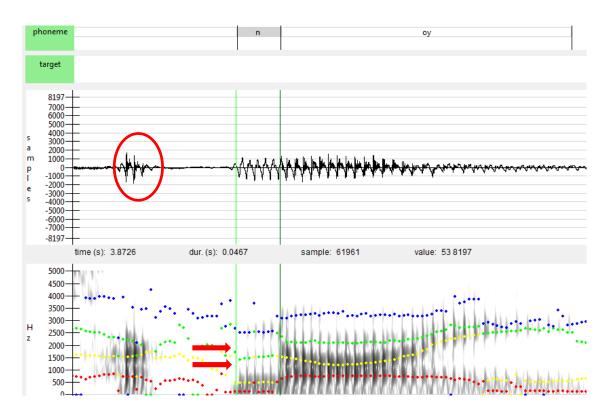
The acoustical properties of the target /n/ - [noi L], excerpted from the above sentence frame, are shown in Figure 26. The target segment [noi L] was marked as [n2oy].



**Figure 26:** Label tier (top), waveform (middle), and spectrogram (bottom) of part of the sentence frame "This is because we do not have any responsibility", with overlaid formant tracking values. The section highlighted in grey in the label tier shows the target segment /n/.

Since all [noi L] samples extracted for this experiment occur sentence-finally, as marked by the green oval, more duration is guaranteed, as it is well known that sentence-final words are normally stressed in Standard Thai (cf. Abramson,1962; Naksakul, 1998). This is confirmed by the duration of at least 117.40 msec and a maximum of 459.91 msec for the vowel [oi] - [noi L] in this experiment (the tonal F0 and formant trajectory of [oi] will also be experimented on, as we shall see later). The segmentation criteria of /n/ - [noi L] are explained below.

Nasals are acoustically described as having 1) lower amplitude relative to the adjacent vowels; 2) stronger energy in the low frequency range as opposed to high frequencies; and 3) a very low F1 in the frequency range between 250-300 Hz for male speakers (Mannell, 2009). In Figure 27 (overleaf), since the target /n/ - [noi L] is preceded by a short pause after a released stop [k] - [sak L] (as marked by the red oval), its starting point



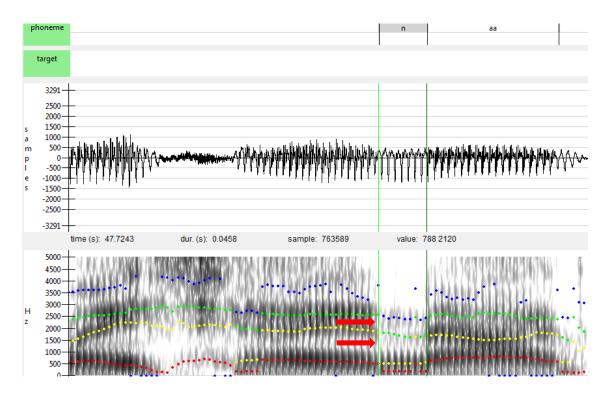
**Figure 27:** Label tier (top), waveform (middle), and spectrogram (bottom) of a target segment /n/ - [noi L].

Red dots represent the first formant frequencies (F1), yellow dots represent the second formant frequencies (F2), green dots and blue dots represent the third and fourth formant frequencies (F3 and F4), respectively. Note: [5] is labeled as [6].

was identified as the point where the vertical dark band spectrogram was observed after the short pause (as indicated by the first vertical green line). In this regard, we can also observe that anti-resonances occurred in the frequency ranges of ca. 1300 Hz and 1800 Hz, respectively (as marked by red arrows). The end point of /n/ was marked at the point of release of the oral closure, as indicated by the increased amplitude, which was marked by the second vertical green line.

## 5.2.4 Segmentation of /n/ - [na: HL]

In this section, I describe the segmentation of the second nasal /n/. I decided to experiment with this second /n/, extracted from the word [na: HL thi: HL], in order to see if /n/, when followed by the vowel [a:], yielded different results from /n/ followed by a diphthong [bi]. These two [n]s will shed light on whether different phonological contexts yield different FVC results. Figure 28 (overleaf) shows a label tier, a waveform, and a spectrogram of the alveolar nasal [n] extracted from the word [na: HL thi: HL].



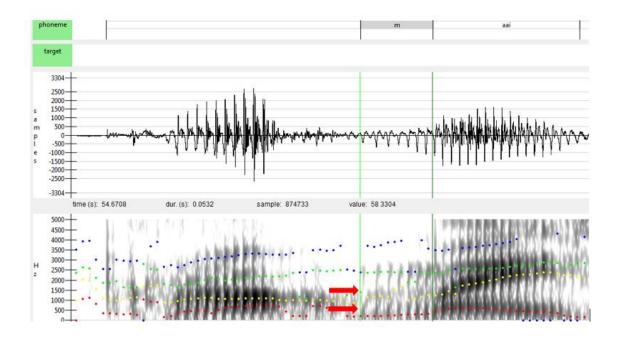
**Figure 28:** Label tier (top), waveform (middle), and spectrogram (bottom) of the target segment /n/ - [na: HL  $t^h$ i: HL].

Red dots represent the first formant frequencies (F1), yellow dots represent the second formant frequencies (F2), green dots and blue dots represent the third and fourth formant frequencies (F3 and F4), respectively. Note: [a:] is labeled as [aa].

The onset and offset of the segment [n] - [na: HL thi: HL], as indicated by the vertical green lines, are fairly straightforward due to the formant frequencies that are clearly attenuated by the nasal anti-resonant frequencies occurring between ca. 1400 Hz and 2300 Hz (as indicated by the red arrows).

## 5.2.5 Segmentation of /m/ - [mai HL] 'no'

Figure 29 (overleaf) shows the segmentation of the target segment [m] - [mai HL]. The vowel /ai/ is labeled as [aai]. We can see that for this experiment, where /m/ - [mai HL] was preceded by a glottal stop [?] - phro? [H] (cf. §5.2.2), it is slightly difficult to identify the starting point of /m/. However, the starting point of the bilabial nasal /m/ - [mai HL] was placed at the point where the anti-resonances were observed. In this case, the first anti-resonance occurred at ca. 500 Hz and the second anti-resonance was at ca. 1500 Hz, as indicated by the red arrows. The endpoint was located once the F2 onset of the following vowel [ai] was clearly observed in the vertical dark band spectrogram (indicated by the second vertical green line).



**Figure 29:** Label tier (top), waveform (middle), and spectrogram (bottom) of the target segment /m/ - [mai HL].

Red dots represent the first formant frequencies (F1), yellow dots represent the second formant frequencies (F2), green dots and blue dots represent the third and fourth formant frequencies (F3 and F4), respectively.

# 5.3 Spectral mean, variance, skew, kurtosis (spectral moments)

In this section, I introduce the basic concepts of spectral mean, variance, skew, and kurtosis. As previously mentioned (see Chapter 1), I will, for convenience, call these statistical measures the four *spectral moments* (specmoments<sup>m</sup>), where mean is m = 1, variance is m = 2, skew is m = 3, and kurtosis is m = 4. These abbreviations will be interchangeably used throughout this thesis. As explained by Forrest, Weismer, Milenkovic, and Dougall (1988), the purpose of using these spectral moments for analyzing speech spectra is to reduce the spectral information into a smaller number of parameters. In this study, the four spectral moments are computationally calculated using the following formula:

specmoments<sup>m</sup> = 
$$\frac{\sum f(x-k)^m}{\sum f}$$

Formula 1 (taken from Harrington, 2010, p. 298)

In this formula,  $\int$  is the extracted spectral data, x is the frequency at which the spectral data was extracted, k = 0 for m = 1;  $k = \text{specmoments}^1$  when m = 2, 3 and 4 (Harrington,

2010, p. 298). For illustrative purposes, I will use spectral data measured at the midpoint of the target segment /s/ from one of the study's informants, Speaker 8, Session 1, as shown in Table 14.

Frequency in Hz (sampled at 16 kHz)	Spectral data (dB values)
500	26
531.25	25
562.5	21
593.75	22
625	10
656.25	18
687.5	20
718.75	21
750	20
781.25	20
812.5	19
843.75	10
875	15
906.25	13
:	:
:	:
:	:
4000	2

**Table 14:** Frequencies and their corresponding spectral data extracted at the midpoint of the token /s/ spoken by Speaker 8, Session 1.

Table 14 shows the excerpted data of frequencies and their corresponding spectral values from Speaker 8, Session 1. The sampling points are at 31.25 msec intervals but rounded to the nearest integer from 500 Hz up to 4000 Hz. Thus, the mean (m1) can be put into a formula as follows.

Mean = 
$$\frac{26(500-0)^1 + 25(531-0)^1 + 21(562-0)^1 + \dots + 2(4000-0)^1}{25+25+21+\dots + 2} = 2302$$

From the above, the mean (m1) is 2302, which means that the spectral energy exerted by Speaker 8, Session 1, is concentrated at around 2302 Hz.

Let us now move on to the variance (m2), which shows the *ranges* of spectral mean values. To calculate variance, I change the value of m = 1 to m = 2 and k = mean (m1) (Harrington, 2010, p. 298) in the formula shown below.

Variance = 
$$\frac{26(500-2302)^2 + 25(531-2302)^2 + ... + 2(4000-2302)^2}{26+...+25+...+2} = 1162990$$

Thus, the variance (m2) is 11629890 Hz. It needs to be noted that we will get high values of variance if the spectra are more diffuse instead of being concentrated at a certain frequency (ibid.).

To calculate skew (m3), a value 3 is assigned to m and a mean (m1) is assigned to k as shown below.

$$Skew = \frac{26(500 - 2302)^3 + 25(531 - 2302)^3 + ... + 2(4000 - 2302)^3}{26 + 25 + ... + 2} = 509067010$$

The above derived value has to be divided by a *variance* (m2) raised to the power of 1.5 (Harrington, 2010, p. 298). As such, skew (m3) is  $\frac{509067010}{1162990(m2)^{^{1}}.5} = 0.4058925$ . This skew value is positive, i.e. the energy is not symmetrical around the mean but there are more values or tail on the left of the distribution (ibid.). In other words, the spectral energy is concentrated in the low frequency range. Note that *mean* (m1), which is 2302 (Hz), is correlated with *skew* (m3) (ibid.).

To calculate kurtosis (m4), the value of 4 is assigned to m and a mean (m1) is assigned to k (Harrington, 2010, p. 298). Furthermore, the derived value needs to be divided by the square of variance (m2) and subtracted by 3 (to normalize the distribution) (ibid.). As such, kurtosis can be put into a formula as follows.

Kurtosis = 
$$\frac{\frac{26(500-2302)^4+25(531-2302)^4+...+2(4000-2302)^4}{26+25+...+2}}{(1162990)^2-3} = -1.3091892$$

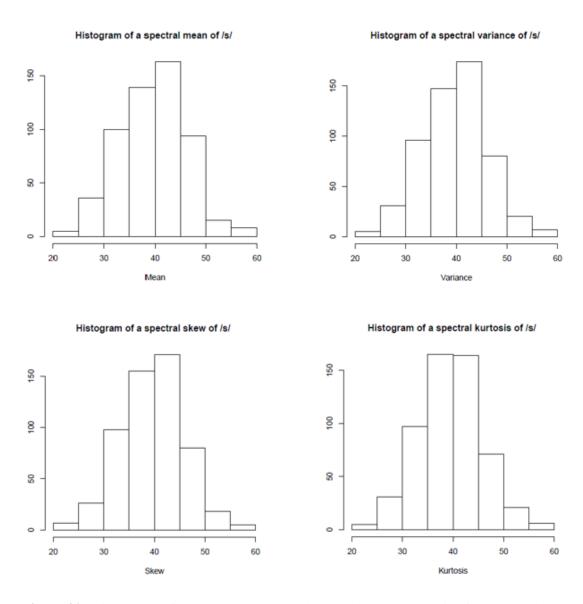
The negative value indicates that the spectrum is flat without clear peaks (ibid.).

# 5.4 Results of the spectral moments of /s/

In this section I report the results for the Standard Thai alveolar fricative /s/ when its spectral moments were parameterized. §5.4.1 shows the distribution of the spectral mean, variance, skew, and kurtosis of /s/, using histograms. The corresponding ANOVA results are presented in §5.4.2. Finally (§§5.4.3 and 5.4.4), I will present and discuss the MVLR and DCT results.

# 5.4.1 Distribution of the spectral moments using histograms

Figure 30 shows the histograms of the mean, variance, skew and kurtosis of the Standard Thai alveolar fricative /s/. They each show a similar shape with near-normal, slightly right-skewed distribution (more values can be observed on the left) (Bertsekas & Tsitsiklis, 2002). This means that the peaks of the spectral mean, variance, skew and kurtosis are slightly off the center of the distribution (ibid.), which suggests that the overall distribution of the /s/ spectrum is close to normal and worth further analysis with kernel density distribution, which can handle the non-normally distributed data.



**Figure 30:** Histograms of the spectral mean, variance, skew, and kurtosis of /s/ uttered by 56 speakers.

## **5.4.2 ANOVA results**

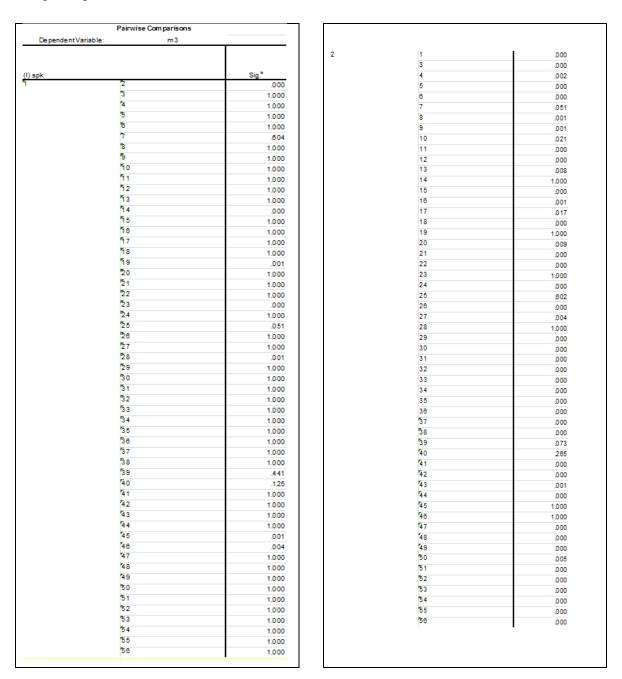
A mean (obtained from five repeats for each speaker/session, i.e. 56 speakers x 2 sessions) of the mean, variance, skew and kurtosis was statistically compared using ANOVA in GenStat (Payne, 2009). The reason to initially run these results in ANOVA was to observe their discriminatory power and judge if these spectral parameters are worth calculating LRs for in the MVLR. As previously mentioned, /s/ energy below 500 Hz and above 4000 Hz might not be available due to the telephone band-pass filter in forensically realistic conditions. It is therefore prudent to divide the experiment into two frequency bands, 500-4000 and 500-8000 Hz (Kavanagh, 2012). Table 15 shows the results; *p*-values at the level < 0.05 indicate a *significant* difference between speakers and/or sessions.

Variate	Factor(s)	500-4000 Hz	500-8000 Hz		
		<i>p</i> -values	<i>p</i> -values		
Mean	speaker	<.001	<.001		
	session	0.646	0.315		
	speaker x session	<.001	<.001		
Variance	speaker	<.001	<.001		
	session	0.013	0.431		
	speaker x session	<.001	<.001		
Skew	speaker	<.001	<.001		
	session	0.461	0.303		
	speaker x session	<.001	<.001		
Kurtosis	speaker	<.001	<.001		
	session	0.411	0.692		
	speaker x session	<.001	<.001		

**Table 15:** ANOVA results for speaker and/or session (N = 56 in 4000 and 8000 Hz) on each spectral moment calculated from /s/. Blue text indicates significant p values at the level < .05.

Table 15 shows that all spectral moments are significant for the *speaker* factor, i.e. there are significant *p* values at level .001 in both band-pass filters (500-4000 Hz, 500-8000 Hz). Likewise, when *speaker plus session* was included as a factor to compare the mean for each of the spectral moments, we get the same results. Yet, the spectral moments are not statistically significant for *session*. This implies that the within-speaker differences between the two sessions are not significant. For illustrative purposes, I therefore

employed the Bonferroni test to observe where the significance lies between speaker pairs using the spectral skew (m3) from the 500-4000 Hz filter. Results are shown in Table 16.



**Table 16:** Bonferroni's pairwise comparisons for a skew (m3) of /s/

The first column of the left and right sections in Table 16 shows the excerpted data where the skews of Speaker 1 and Speaker 2 were compared with those of the other 55 speakers. The right column indicates significant p values at level < .05. We can infer from the above results that Speaker 1 significantly differs from the other seven speakers (i.e. speakers 2,

14, 19, 23, 28, 45, 46) when a skew was parameterized within the 500-4000 Hz frequency range. Similarly, the skew (m3) of Speaker 2 is found to be statistically different from that of another 49 speakers (i.e. all 55 speakers except speakers 14, 19, 23, 28, 45, 46).

Based on the findings above, we can say that Speaker 1 is perhaps very typical (in terms of skew value), but Speaker 2 is not. From the auditory impression, the speech samples of the alveolar fricative /s/ uttered by 54 informants, all of whom are young male university students (aged 22 years old), sound similar enough to make it difficult to distinguish them by auditory analysis alone. Having said that, speech samples of the other two informants, who are in their 50s, sound dissimilar to the rest of the population. That is, the alveolar fricative /s/ seems to be carefully articulated, resulting in a clear and longer frication noise in the case of these two informants. This might be due to idiosyncratic factors: the older generation typically tends to speak in a clear manner with low speech tempo. Overall, the differences found between speakers (cf. Table 15) is statistically significant when each of the mean, variance, skew, and kurtosis was parameterized in ANOVAs. As such, it is worth including all these spectral moments of /s/ in an MVLR calculation.

### **5.4.3 MVLR results**

Table 17 (overleaf) shows the MVLR results according to calibrated  $Log_{10}LR$ ,  $C_{llr}$ , and EER values. I will first discuss the overall results with reference to this table. Detailed discussion about the proportion and magnitude of calibrated  $Log_{10}LRs$ ,  $C_{llr}$  and EER values will follow, with reference to the Tippett plots shown below.

The results in Table 17 are presented according to the possible combination (from two to four) of the parameters mean, variance, skew, and kurtosis in the 500-4000 Hz and 500-8000 Hz filters. The parameters were combined as shown in the leftmost column of Table 17 to test if different numbers of parameters yield different results and to ascertain which parameters perform better. Since calibration is an important aspect of the performance of an LR-based FVC system, comparing such uncalibrated and calibrated LRs is crucial to subsequently point out the derived magnitude of  $C_{llr}$  values that are useful for subsequent ranking of the FVC performance for each parameter. Generally, in Table 17, many of the highest calibrated consistent-with-fact  $SSlog_{10}LRs$  only provide "limited support" for the prosecution hypothesis that the speech samples were more likely to be from the same

	4000 Hz				8000 Hz			
Parameters	Calibrated LOG <sub>10</sub> LR		Cllr	EER	Calibrated LOG <sub>10</sub> LR		Cllr	EER
	SS	DS			SS	DS		
Mean, Variance, Skew, Kurtosis	≤ 0.41	≥ -3.23	0.83	39%	≤ 0.35	≥ -4.04	0.85	42%
Mean, Variance, Kurtosis	≤ 0.32	≥ -3.20	0.86	45%	≤ 0.27	≥ -3.84	0.86	42%
Mean, Variance, Skew	≤ 0.45	$\geq -2.90$	0.84	40%	≤ 0.37	≥ -5.54	0.85	44%
Variance,Skew,Kurtosis	≤ 0.27	≥ -3.13	0.87	44%	≤ 0.33	$\geq$ -3.77	0.86	42%
Mean,Skew,Kurtosis	≤ 0.32	$\geq -3.30$	0.87	45%	≤ 0.45	≥ -4.84	0.85	39%
Mean,Variance	≤ 0.33	≥ -2.92	0.87	43%	≤ 0.28	≥ -5.09	0.87	44%
Mean,Skew	≤ 0.53	$\geq -3.87$	0.87	43%	≤ 0.30	$\geq -2.80$	0.92	50%
Mean,Kurtosis	≤ 0.23	≥ -2.83	0.89	51%	≤ 0.38	≥ -4.61	0.87	43%
Variance,Skew	≤ 0.30	≥ -2.57	0.88	45%	≤ 0.28	$\geq -4.13$	0.89	49%
Variance, Kurtosis	≤ 0.28	≥ -1.95	0.88	50%	≤ 0.35	≥ -4.08	0.85	45%
Skew,Kurtosis	≤ 0.23	≥ -2.86	0.91	51%	≤ 0.36	≥ -4.88	0.86	44%

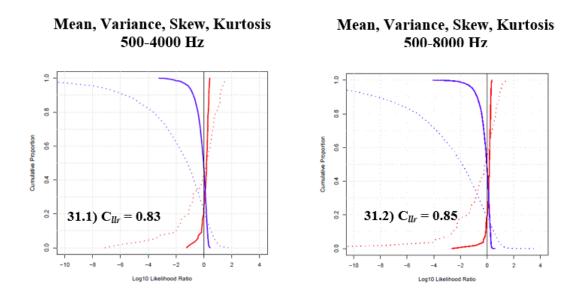
**Table 17:** Log<sub>10</sub>LR, C<sub>llr</sub>, and EER values of the fricative spectra /s/ according to the combined parameters (leftmost column) measured at the temporal midpoint of the fricative /s/ from 56 speakers, in 500-4000 Hz as well as 500-8000 Hz conditions. The values highlighted in blue and red show the best and worst Log<sub>10</sub>LR, C<sub>llr</sub>, and EER values, respectively.

speaker than from different speakers. For DS comparisons, the best calibrated consistent-with-fact  $DSlog_{10}LR = -5.54$  suggest "very strong" support for the defense hypothesis that the samples were more likely to be from different speakers than from the same speaker.

With respect to the 500-4000 Hz filter, the lowest  $C_{llr} = 0.83$  was obtained when all four spectral features (mean, variance, skew, and kurtosis) were parameterized. The highest  $C_{llr} = 0.91$  was obtained when skew plus kurtosis were used. Table 17 shows a general trend in the obtainment of the  $C_{llr}$  values: those with two parameters yielded higher  $C_{llr}$  (as compared to those with three or four parameters). From this, we can say that the results deteriorated as the number of parameters decreased. With respect to the 500-8000 Hz filter, the smallest  $C_{llr} = 0.85$  was obtained when 1) all four spectral moments; 2) mean, variance, skew; 3) mean, skew, kurtosis; and 4) variance and kurtosis were trialed. The highest  $C_{llr} = 0.92$  was obtained when mean plus skew were parameterized. The findings in the 500-8000 Hz filter agree with those in the 500-4000 Hz filter range: *all four spectral parameters* should be parameterized if the spectral features are to be of use in FVC, as

any combination of two parameters yields worse  $C_{llr}$  results (except for Variance and Kurtosis using a 500-8000 Hz filter) than combination involving three or four parameters.

I now discuss in more detail the proportion and magnitude of the LRs. The Tippett plots with two or three parameters are presented in Figures 32 and 33. Those in Figure 31 reflect parameterization of all four parameters in the 500-4000 and 500-8000 Hz band-pass filters.



**Figure 31:** Tippett plots for SS and DS comparisons when mean, variance, skew and kurtosis were parameterized in the 500-4000 Hz (left) and 500-8000 Hz (right) band-pass filters. The (red) curves rising to the right represent the cumulative proportion of the SS (same speaker) comparisons, with the log<sub>10</sub>LRs equal to or less than the value indicated on the x-axis, while the (blue) curves rising to the left represent the cumulative proportion of the DS (different speaker) comparisons, with the log<sub>10</sub>LRs equal to or greater than the value indicated on the x-axis. Dotted lines and solid lines represent the uncalibrated and calibrated SS and DS Log<sub>10</sub>LRs, respectively.

In the case of the Tippett plot of the 500-4000 Hz filter in Figure 31.1, the  $C_{llr} = 0.83$  was obtained when mean, variance, skew and kurtosis were all parameterized. The worst uncalibrated contrary-to-fact  $SSlog_{10}LR$  (for SS comparisons) and  $DSlog_{10}LR$  (for DS comparisons) were  $SSlog_{10}LR = -7.09$  and  $DSlog_{10}LR = 1.63$ . After calibration, these were significantly reduced to  $SSlog_{10}LR = -1.25$  and  $DSlog_{10}LR = 0.45$ . In the 500-8000 Hz filter (Figure 31.2), on the other hand, the highest uncalibrated contrary-to-fact  $SSlog_{10}LR = -16.11$  and the  $C_{llr} = 0.85$  were obtained. The uncalibrated  $SSLog_{10}LRs$  and  $DSLog_{10}LRs$  (dotted red and blue lines) were significantly reduced in magnitude after calibration (solid lines). Specifically, ca. 78% of DS comparisons (dotted blue line) had

Log<sub>10</sub>LRs  $\leq$  -4, but after calibration one of these had DSlog<sub>10</sub>LRs less than -4 (DSlog<sub>10</sub>LR = -4.04). Likewise, ca. 60% of uncalibrated SSLRs (dotted red line) had log<sub>10</sub>LR  $\leq$  2. This was reduced to only log<sub>10</sub>LR  $\leq$  0.35 after calibration. Based on these findings, we can conclude that when all four spectral parameters are used, the results obtained from the 500-4000 Hz filter are better in terms of a lower  $C_{llr}$  and smaller misleading SSLRs. However, the correct LRs are also weakened.

Figure 32 reproduces the Tippett plots that show the results obtained when two or three spectral parameters were combined in the 500-4000 Hz filters. It shows that the overall results for the 500-4000 Hz filter were not promising as the  $C_{llr}$  values were relatively high, between 0.83 and 0.91. However, if we compare these  $C_{llr}$  values with those of the English /s/, which had a  $C_{llr}$  = 0.88 in Franco-Pedroso et al. (2012), the Standard Thai fricative /s/ still performed in roughly the same manner as the English /s/. Having said

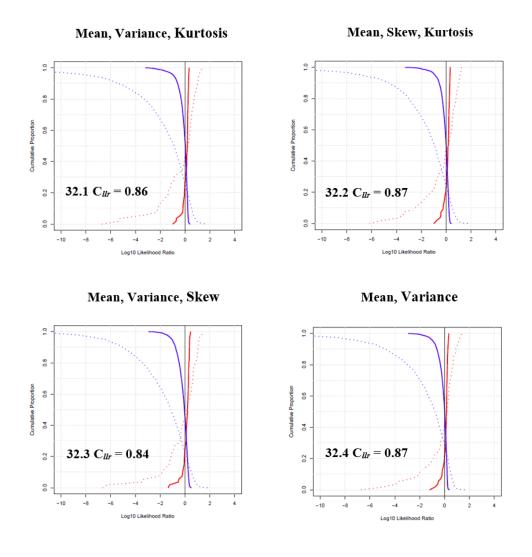
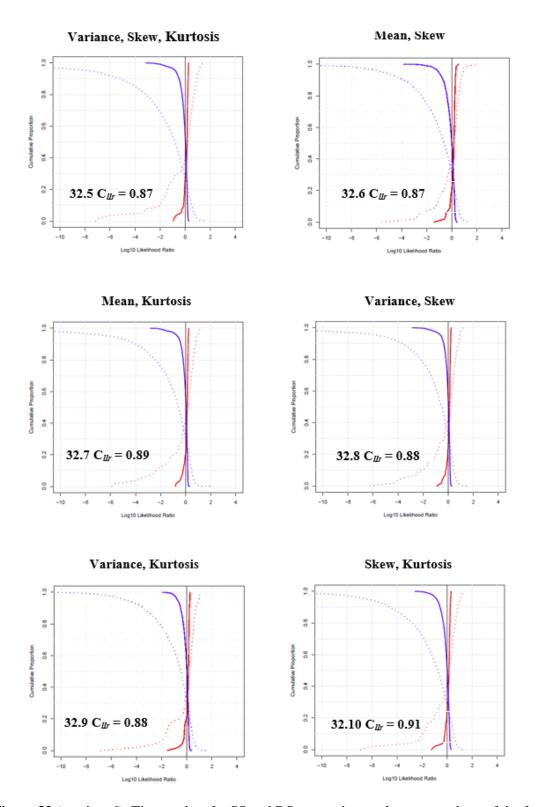


Figure 32 (continued overleaf)



**Figure 32** (continued): Tippett plots for SS and DS comparisons when two or three of the four parameters (as indicated on top of each of the plots) were combined in a 500-4000 Hz band-pass filter.

The (red) curves rising to the right represent the cumulative proportion of the SS (same speaker) comparisons, with the  $\log_{10}LRs$  equal to or less than the value indicated on the x-axis, while the (blue) curves rising to the left represent the cumulative proportion of the DS (different speaker) comparisons, with the  $\log_{10}LRs$  equal to or greater than the value indicated on the x-axis. Dotted lines and solid lines represent the uncalibrated and calibrated SS and DS  $\log_{10}LRs$ , respectively.

that, we cannot directly compare results obtained here with those obtained previously. This is because the different  $C_{llr}$  obtained might be attributed to many different factors, e.g. the contrasting number of speakers trialed or different recording room conditions or different equipment (e.g. type of microphones).

Figure 33 reproduces the Tippett plots that show the results obtained when two or three spectral parameters were combined in the 500-8000 Hz filters.

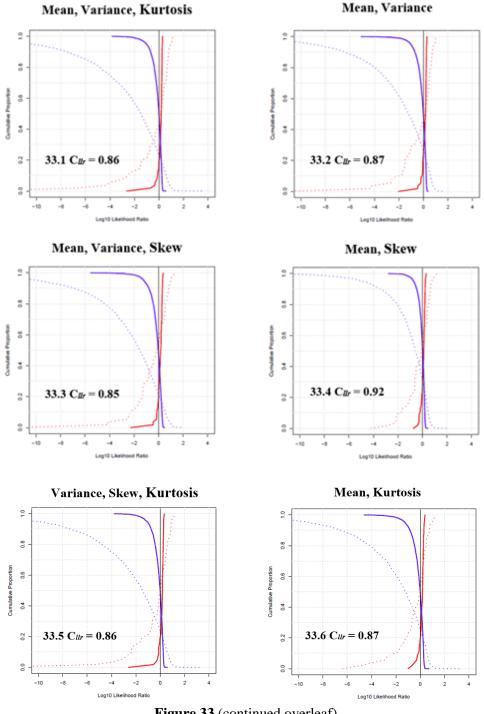
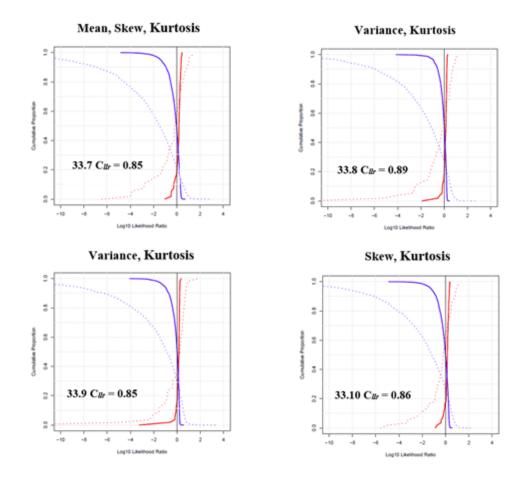


Figure 33 (continued overleaf)



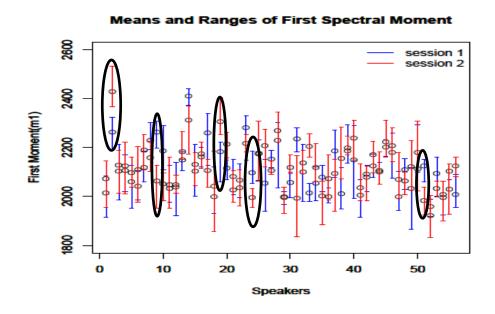
**Figure 33** (continued): Tippett plots for SS and DS comparisons when two or three of the four parameters (as indicated on top of each of the plots) were combined in a 500-8000 Hz band-pass filter.

The (red) curves rising to the right represent the cumulative proportion of the SS (same speaker) comparisons, with the log<sub>10</sub>LRs equal to or less than the value indicated on the x-axis, while the (blue) curves rising to the left represent the cumulative proportion of the DS (different speaker) comparisons, with the log<sub>10</sub>LRs equal to or greater than the value indicated on the x-axis. Dotted lines and solid lines represent the uncalibrated and calibrated SS and DS Log<sub>10</sub>LRs, respectively.

Judging by Figure 33, worse results were obtained compared to the other filter. That is, the largest misleading uncalibrated  $SSlog_{10}LR = -16.11$  was obtained when all four spectral parameters were parameterized. The largest misleading  $DSlog_{10}LR = 3.58$  was also yielded when mean, variance and kurtosis were used (vs the misleading  $DSlog_{10}LR = 2.02$  in the 500-4000 Hz with mean + kurtosis parameters). Another observation with respect to this filter (500-8000 Hz) is that there is a trade-off between  $C_{llr}$  values and the contrary-to-fact  $SSLog_{10}LRs$  obtained. That is, when the former was low, the latter got high, e.g. Variance + Kurtosis had the lowest  $C_{llr} = 0.85$  but highest misleading  $SSlog_{10}LR = -15.64$ . Comparing the proportion of  $DSlog_{10}LRs$  with those in the 500-4000 Hz filter, a similar result was found: 80% of uncalibrated  $DSlog_{10}LRs$  were less

than –4. However, this proportion shrunk to 1% after calibration. Likewise, the magnitude of all calibrated SSLog<sub>10</sub>LRs obtained was less than 0.5, suggesting that this evidence, either in support of a SS hypothesis or DS hypothesis, would not be useful. Such a conservative manner of LRs might result from the extreme misleading scores, e.g. SSlog<sub>10</sub>LR = –15.64 (Variance + Kurtosis) obtained for /s/. That is, the logistic-regression-calibration weights calculated from such extreme misleading scores might cause an extensive scaling for logistic-regression-calibration (Ishihara, 2017, p. 191). To better deal with these outliers or extrapolation errors, the normalized Bayes error-rate (NBE) (Vergeer, van Es, de Jongh, Alberink, & Stoel, 2016) should be trialed in future work to limit the sensible minimum and maximum scores in the LR systems.

So far, we can see that the overall results of the spectral mean, variance, skew, and kurtosis of /s/ were not very promising (highest  $C_{llr}$  at 0.92). This being the case, I continued my search for the underlying factors that might contribute to such a modest performance of the Standard Thai /s/. To do so, I plotted the means and ranges for each of the four spectral parameters obtained from a 500-4000 Hz filter in order to see the ratio of within- to between-speaker variation obtained. Figures 34 to 37 show the means and ranges of spectral mean, variance, skew, and kurtosis, respectively, plotted for each of the 56 speakers. The vertical blue lines represent the spectral ranges of session 1, whereas the red lines represents the spectral ranges of session 2. A circle plotted in the middle of each vertical line represents the mean spectral value for that session for a given speaker.



**Figure 34:** The means (circles) and ranges of spectral mean.

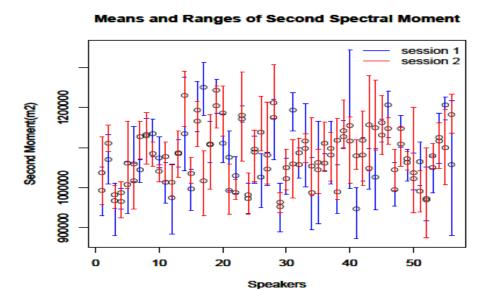


Figure 35: The means (circles) and ranges of spectral variance.

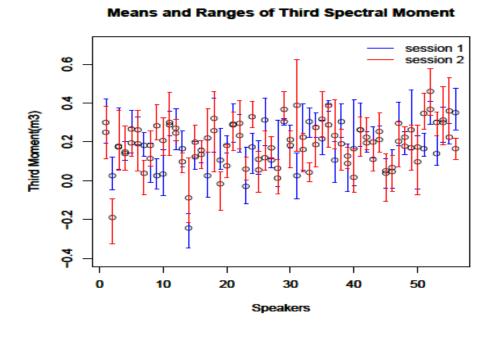


Figure 36: The means (circles) and ranges of spectral skew.

#### 

Figure 37: The means (circles) and ranges of spectral kurtosis.

Judging visually by the ranges above, we observe that the spectral values obtained from between-speaker variation overlap (in the y-axis) within a certain range for each of the parameters. An example is the mean values in Figure 34, which are clustered in a frequency range between 2000 Hz and 2300 Hz. Likewise, the within-speaker variation shows large ranges. An example is the complete separation between sessions (complete separation of blue and red lines) in Speakers 2, 9, 19, 24, and 51 for a spectral mean in Figure 34 (although ANOVA results showed that such between-session variation was not statistically significant). Thus, there is a large range found within speakers, which make the between-speaker differences less significant. Based on these findings, I found out that the spectral moments of /s/ performed only in a conservative manner. This being the case, a decision was made to further test FVC performance of the fricative /s/ using DCT parameterization. If the DCTs worked better than the spectral moments for /s/, I would then apply the DCT parameterization technique to other consonants. Moreover, since the marginal difference between the best  $C_{llr} = 0.83$  vs  $C_{llr} = 0.85$  (together with the same EER = 39%) were obtained for the 500-4000 Hz and 500-8000 Hz filter bands for spectral moments of /s/, I decided to test the FVC performance of the DCTs extracted from a single 500-8000 frequency band, as we shall see below.

#### **5.4.4 DCT results**

The number of speakers tested for each of the target segments /s, te<sup>h</sup>, n, m/ is slightly different, as some tokens were mispronounced by some of the informants. Low amplitudes, due to poor recording, were also found. Mispronounced target segments and poorly recorded low amplitude segments were discarded (I shall illustrate this further in Chapters 6-7).

#### 5.4.4.1 Fricative /s/ extracted from the word [sa:m LH] 'three'

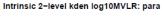
As before, the results of the calibrated Log10LR, Cllr and EER values for the fricative /s/, when its 15 and 20 DCTs were parameterized in both Hertz and Bark scales, in the 500-8000 Hz filter, are tabulated first. This is followed by a presentation of the best and worst results using the Tippett plots in Figures 38 and 39. 56 speakers were tested.

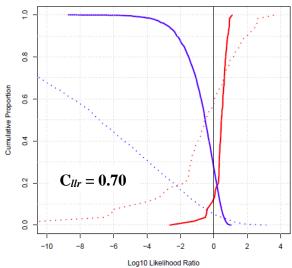
		Her	tz		Bar			ark		
Parameters	Calib	rated	$\mathbf{C}_{llr}$	EER	Calibrated		$\mathbf{C}_{\mathit{llr}}$	EER		
	LOG <sub>10</sub> LR				LOG <sub>10</sub> LR					
	SS	DS			SS	DS				
/s/ 15 coeffs	≤ 0.96	≥ -6.9	0.75	19	≤ 1.10	≥ -8.70	0.70	19		
/s/ 20 coeffs	≤ 0.93	≥ -6.83	0.77	23	≤ 1.07	≥ -7.62	0.72	20		

**Table 18:** Calibrated Log<sub>10</sub>LR,  $C_{llr}$ , and EER of the fricative /s/ - [sa:m LH] when its 15 DCTs and 20 DCTs were parameterized in both Hertz and Bark scales, in the 500-8000 Hz filter. The best  $C_{llr}$  and EER values are highlighted in blue; the worst are shown in red.

Table 18 shows that when a spectrum of /s/ was measured at the midpoint in a Bark scale and its 15 DCTs were tested, we get the lowest  $C_{llr} = 0.70$  and the lowest EER = 19%. However, when 20 Hertz-scaled DCTs were parameterized, the  $C_{llr}$  got worse at 0.77, with the highest EER of 23. When we compare these results with those from the spectral moments ( $C_{llr} = 0.83-0.92$ ), DCTs performed better than spectral moments (although the magnitude of LRs was comparatively the same, i.e.  $SSLog_{10}LRs \le 1$ ).

I will now look in more detail at the proportion and magnitude of Log<sub>10</sub>LRs using the Tippett plots of the 15 Bark-scaled DCTs of /s/ - [sa:m LH] (Figure 38).





**Figure 38:** Tippett plot of the best performing parameter, 15 Bark-scaled DCTs of /s/ - [sa:m LH].

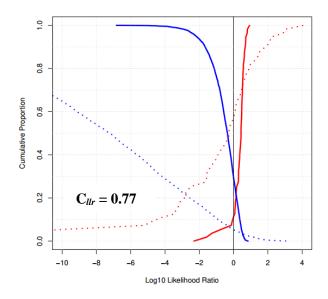
The (red) curves rising to the right represent the cumulative proportion of the SS (same speaker) comparisons, with the log<sub>10</sub>LRs equal to or less than the value indicated on the x-axis, while the (blue) curves rising to the left represent the cumulative proportion of the DS (different speaker) comparisons, with the log<sub>10</sub>LRs equal to or greater than the value indicated on the x-axis. Dotted lines and solid lines represent the uncalibrated and calibrated SS and DS Log<sub>10</sub>LRs, respectively.

With 15 Bark-scaled DCTs, the highest misleading SSLog<sub>10</sub>LR obtained was SSLog<sub>10</sub>LR = -16.80. After calibration, this was reduced to SSLog<sub>10</sub>LR = -2.63. The largest consistent-with-fact SSLR, which was formerly SSLog<sub>10</sub>LR = 3.71, was reduced to SSLog<sub>10</sub>LR = 1.10 after calibration. With respect to DS comparisons, the highly misleading DSLog<sub>10</sub>LR = 3.20 was reduced to DSLog<sub>10</sub>LR = 1.03. Similarly, 70% of uncalibrated DSLRs were less than -4 but this were reduced to only ca. 1% after calibration. Based on these results, there is a trade-off between the magnitude and proportion of the contrary-to-fact LRs and the correct ones. That is, the misleading DSLRs and SSLRs were significantly reduced both in magnitude and proportion, but interestingly the magnitude and proportion of the correct DSLRs and SSLRs were reduced in the same way.

As discussed in §5.4.2, the speech samples of the alveolar fricative /s/ uttered by 54 of the informants, all of whom are young males (aged 22 years old), sound dissimilar to those of the other two informants, who are in their 50s. The auditory impression is that the more senior informants pronounce the alveolar fricative /s/ in a clear manner, resulting

in a clear and longer fricative duration. Such small between-speaker differences observed in the pronunciation of the fricative /s/ for the two informants who are in their 50s might contribute to the resultant calibrated  $SSLog_{10}LR = 1.10$ , which can only provide "limited" support for the same-speaker hypothesis and only 1% of calibrated  $DSLog_{10}LR \le -4$  give "very strong" support for the DS hypothesis.

I will now describe the worst results for the fricative /s/, which were obtained when the 20 Hz-scaled DCTs were parameterized (Figure 39).



**Figure 39:** Tippett plot of the worst performing parameter, 20 Hertz-scaled DCTs of /s/ - [sa:m LH].

The (red) curves rising to the right represent the cumulative proportion of the SS (same speaker) comparisons, with the log<sub>10</sub>LRs equal to or less than the value indicated on the x-axis, while the (blue) curves rising to the left represent the cumulative proportion of the DS (different speaker) comparisons, with the log<sub>10</sub>LRs equal to or greater than the value indicated on the x-axis. Dotted lines and solid lines represent the uncalibrated and calibrated SS and DS Log<sub>10</sub>LRs, respectively.

The largest misleading  $SSlog_{10}LR = -19.97$  was reduced to only  $SSlog_{10}LR = -2.31$  after calibration. We also observe that the strongest uncalibrated correct  $SSlog_{10}LR = 4.08$  (suggesting "very strong" evidence in support of the same-speaker hypothesis) was reduced to only  $SSlog_{10}LR = 0.93$ , which suggests "useless" speech evidence in support of both SS and DS hypotheses. For DS comparisons, the same results were obtained as with the previous parameter (15 Bark-scaled DCTs): the magnitude and proportion was  $DSlog_{10}LR \le -4$ , mostly reduced to 1% after calibration.

From the above, we can see that using the DCT coefficients fitted to a spectrum (extracted from the midpoint) of the consonants /s/ yields better results ( $C_{llr}$  values between 0.70 and 0.77) than those of the spectral mean, variance, skew and kurtosis ( $C_{llr}$  values between 0.83 and 0.92). This being the case, I decided to search for the individualizing information that might be found in the segments /te<sup>h</sup>, n, m/; I did this by using their DCT coefficients extracted from the midpoint. The corresponding results are reported below.

#### 5.4.4.2 Affricate /tch/ extracted from the word [tchai HL] 'yes'

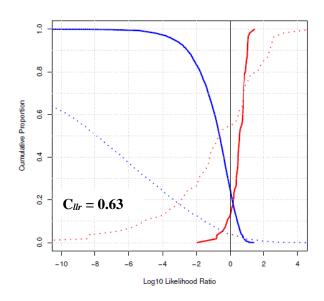
In this section I show the results of the Standard Thai affricate /teh/ - [tehai HL] when its 15 and 20 DCTs were parameterized in both Hertz and Bark scales in the 500-8000 Hz filter. Calibrated Log10LR, Cllr, and EER values are presented in Table 19 followed by the Tippett plots of the best and worst performing parameters. 57 speakers were tested.

		Her	rtz			Baı	ark		
Parameters	Cali	brated	$\mathbf{C}_{llr}$	EER	Calibrated		$\mathbf{C}_{llr}$	EER	
	LO	$G_{10}LR$			LOG <sub>10</sub> LR				
	SS	DS			SS	DS			
/teh/ 15 coeffs	≤ 1.17	≥ -9.28	0.71	22	≤ 1.39	≥ −10.11	0.68	20	
/teh/ 20 coeffs	≤ 1.36	≥ -10.04	0.66	19	≤ 1.46	≥ −11.62	0.63	19	

**Table 19:** Calibrated Log<sub>10</sub>LR,  $C_{llr}$ , and EER of the affricate /te<sup>h</sup>/ - [te<sup>h</sup>ai HL] when its 15 DCTs and 20 DCTs were parameterized in both Hertz and Bark scales, respectively. The best  $C_{llr}$  and EER values are highlighted in blue; the worst are shown in red.

Table 19 shows the results of the Standard Thai affricate  $/te^h/$  extracted from the word  $[te^hai HL]$  'yes'. For the best results, the lowest  $C_{llr} = 0.63$  and EER = 19% were obtained when 20 Bark-scaled DCTs were tested. The worst results were obtained with the highest  $C_{llr} = 0.71$  and highest EER = 22% when 15 DCTs were parameterized in a Hertz scale. So far, we have observed that none of the experimental settings yield the greatest consistent-with-fact SSLRs, which exceed  $log_{10}LR = 2$ .

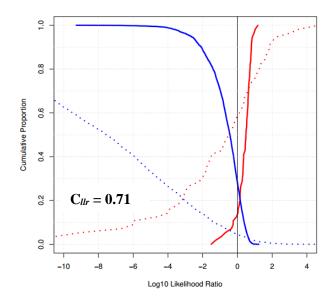
Figures 40 and 41 show the best and worst performing Tippett plots for  $/tc^h/$  -  $[tc^hai HL]$  based on  $C_{llr}$  and EER values.



**Figure 40:** Tippett plot of the best performing parameter, 20 Bark-scaled DCTs of /teʰ/ - [teʰ ai HL]. The (red) curves rising to the right represent the cumulative proportion of the SS (same speaker) comparisons with the log<sub>10</sub>LRs equal to or less than the value indicated on the x-axis, while the (blue) curves rising to the left represent the cumulative proportion of the DS (different speaker) comparisons, with the log<sub>10</sub>LRs equal to or greater than the value indicated on the x-axis. Dotted lines and solid lines represent the uncalibrated and calibrated SS and DS Log<sub>10</sub>LRs, respectively.

Figure 40 shows the Tippett plot when 20 Bark-scaled DCTs were parameterized. The highest uncalibrated contrary-to-fact SSLog<sub>10</sub>LR and DSLog<sub>10</sub>LR obtained were -13.94 and =4.58, respectively. After calibration, these were reduced to SSLog<sub>10</sub>LR = -1.97 and DSLog<sub>10</sub>LR = 1.41. Not only was the magnitude of Log<sub>10</sub>LR reduced but also its proportion; ca. 73% of DSLog<sub>10</sub>LRs  $\leq -4$  was reduced to 2% after calibration. The strongest correct SSLR and DSLR, which were formerly SSLog<sub>10</sub>LR = 4.92 and DSLog<sub>10</sub>LR = -42.05, were significantly reduced to SSLog<sub>10</sub>LR = 1.46 and DSLog<sub>10</sub>LR = -11.62, respectively. The EER obtained for 20 Bark-scaled DCTs of  $/tg^h/$  was 19%.

With the 15 Hertz-scaled DCT parameter in Figure 41 (overleaf), the worst contrary-to-fact Log<sub>10</sub>LRs for SS and DS comparisons were SSLog<sub>10</sub>LR = -13.99 and DSLog<sub>10</sub>LR = 5.04, respectively. These were significantly reduced to SSLog<sub>10</sub>LR = -1.51 and DSLog<sub>10</sub>LR = 1.23 after calibration. The magnitude of correct SSLRs was reduced from SSLog<sub>10</sub>LR = 4.67 (which suggests "very strong" support for the SS hypothesis) to SSLog<sub>10</sub>LR = 1.17 (which suggests "limited" support for the SS hypothesis). Similarly, ca. 77% of correct DSLRs  $\leq -4$  was reduced to 1% after calibration. Comparing the best and worst results in terms of  $C_{llr}$  and EER values in this 500-8000 filter band, we see that the best parameter, which was the 20 Bark-scaled DCTs, gave a lower  $C_{llr} = 0.63$  and



**Figure 41:** Tippett plot of the worst performing parameter, 15 Hertz-scaled DCTs of /tch/ - [tch ai HL].

The (red) curves rising to the right represent the cumulative proportion of the SS (same speaker) comparisons, with the log<sub>10</sub>LRs equal to or less than the value indicated on the x-axis, while the (blue) curves rising to the left represent the cumulative proportion of the DS (different speaker) comparisons, with the log<sub>10</sub>LRs equal to or greater than the value indicated on the x-axis. Dotted lines and solid lines represent the uncalibrated and calibrated SS and DS Log<sub>10</sub>LRs, respectively.

EER = 19%, although many of the SS comparisons had the same  $SSlog_{10}LR \le 2$  and ca. 1% of the DS comparisons had  $DSlog_{10}LR \le -4$ .

#### 5.4.4.3 Nasal /n/ extracted from the particle [noi L]

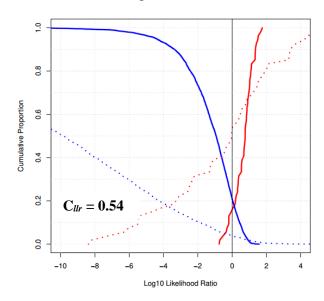
In this section I show the results of using DCTs extracted from the midpoint of a nasal consonant /n/ - [noi L]. [noi L] is a Thai particle put at the end of a request to lessen an imperative statement. 55 speakers were tested.

	Hertz			Bark				
Parameters	Calibrated LOG <sub>10</sub> LR		$\mathbf{C}_{llr}$	EER	Calibrated LOG <sub>10</sub> LR		$\mathbf{C}_{llr}$	EER
	SS	DS			SS	DS		
/n/ 15 coeffs	≤ 1.79	≥ −12.99	0.54	19	≤ 1.78	≥ −11.98	0.56	20
/n/ 20 coeffs	≤ 1.76	≥ -12.45	0.54	18	≤ 1.75	≥ -10.51	0.58	20

**Table 20:** Calibrated Log<sub>10</sub>LR,  $C_{llr}$ , and EER of the nasal /n/ - [noi L] when its 15 DCTs and 20 DCTs were parameterized in both Hertz and Bark scales, respectively. The best  $C_{llr}$  and EER values are highlighted in blue; the worst are shown in red.

Table 20 shows that, in general terms, the best  $SSlog_{10}LRs \le 2$  were obtained, which indicates "moderate" support for the prosecution hypothesis that the speech samples are more likely to be from the same speaker than from different speakers. In contrast, for DS comparisons, the best  $DSlog_{10}LRs \le -4$  suggest "very strong" support for the defense hypothesis that the speech samples are more likely to be from different speakers than from the same speaker. We also see in Table 20 that, when 20 Hertz-scaled DCTs were parameterized, the lowest  $C_{llr}$  derived is 0.54 with an EER of 18. Roughly the same results were obtained,  $C_{llr} = 0.54$  and EER = 19%, with 15 Hertz-scaled DCTs. However,  $C_{llr}$  and EER are higher on the Bark scale: a  $C_{llr}$  of 0.56 and 0.58 and an EER = 20% when 15 and 20 Bark-scaled DCTs were parameterized, respectively.

The Tippett plots of the best and worst performing parameters of /n/ - [noi L], based on its  $C_{llr}$  and EER values, are shown in Figures 42 and 43.

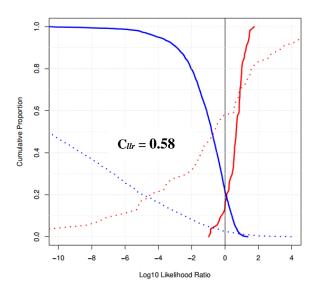


**Figure 42:** Tippett plot of the best performing parameter, 20 Hertz-scaled DCTs of /n/ - [noi L]. The (red) curves rising to the right represent the cumulative proportion of the SS (same speaker) comparisons, with the  $\log_{10}$ LRs equal to or less than the value indicated on the x-axis, while the (blue) curves rising to the left represent the cumulative proportion of the DS (different speaker) comparisons, with the  $\log_{10}$ LRs equal to or greater than the value indicated on the x-axis. Dotted lines and solid lines represent the uncalibrated and calibrated SS and DS  $\log_{10}$ LRs, respectively.

Apart from having a lower  $C_{llr} = 0.54$  than those of the fricatives /s/ and /tch/, as previously shown, the Tippett plot of the best performing parameter (20 Hertz-scaled DCTs) for the nasal /n/ - [noi L] (Figure 42) also shows the largest magnitude of the calibrated consistent-with-fact  $SSLog_{10}LR = 1.76$  (consistent-with-fact uncalibrated  $Log_{10}LR = 1.76$ )

5.74). After calibration, about 8% of DS comparisons had  $Log_{10}LR \le -4$ , which suggests "very strong" support for the different-speaker hypothesis. The EER for the 20 Hertz-scaled DCTs of /n/ - [noi L] was 18%.

Using 20 Bark-scaled DCTs extracted from /n/ - [noi L] (Figure 43), a marginally higher  $C_{llr} = 0.58$  was obtained than in the case of 20 Hz-scaled DCTs ( $C_{llr} = 0.54$ ). After calibration, ca. 4% of DSLRs had  $DSlog_{10}LRs \leq -4$ . The largest calibrated consistent-with-fact SSLR obtained was  $SSlog_{10}LR = 1.75$  ( $SSLog_{10}LR = 6.70$  before calibration). Although this was the worst performing parameter, it yielded only a marginally higher EER = 20% than the best performing parameter for /n/ - [noi L], which was 18%.



**Figure 43:** Tippett plot of the worst performing parameter, 20 Bark-scaled DCTs of /n/ - [noi L].

The (red) curves rising to the right represent the cumulative proportion of the SS (same speaker) comparisons, with the  $\log_{10}LRs$  equal to or less than the value indicated on the x-axis, while the (blue) curves rising to the left represent the cumulative proportion of the DS (different speaker) comparisons, with the  $\log_{10}LRs$  equal to or greater than the value indicated on the x-axis. Dotted lines and solid lines represent the uncalibrated and calibrated SS and DS  $\log_{10}LRs$ , respectively.

#### 5.4.4.4 Nasal /n/ extracted from the word [na: HL thi: HL] 'duty'

In this section, I look at another instance of /n/ - [na: HL], embedded in the word [na: HL thi: HL] 'duty'. I then compare the FVC performance of the two /n/s. 55 speakers were tested.

		Her	·tz			Ba	ırk			
Parameters	Cali	brated	$\mathbf{C}_{llr}$	EER	Calibrated		Calibrated		$\mathbf{C}_{llr}$	EER
	LO	G <sub>10</sub> LR			LOG <sub>10</sub> LR					
	SS	DS			SS	DS				
/n/ 15 coeffs	≤ 2.40	≥ −14.57	0.49	18	≤ 1.99	≥ −15.15	0.47	15		
/n/ 20 coeffs	≤ 1.73	≥ -12.71	0.50	13	≤1.87	≥ −10.97	0.51	18		

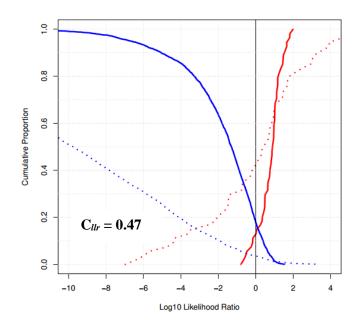
**Table 21:** Calibrated Log<sub>10</sub>LR,  $C_{llr}$ , and EER of the nasal /n/ - [na: HL] when its 15 DCTs and 20 DCTs were parameterized in both Hertz and Bark scales, respectively. The best  $C_{llr}$  and EER values are highlighted in blue; the worst are shown in red.

Table 21 shows that the best results were obtained when 15 Bark-scaled DCTs were parameterized: the lowest  $C_{llr} = 0.47$  and an EER = 15%. These are the best results obtained so far when compared to those of the previous target segments /s/ - [sa:m LH], /tch/ - [tchai HL/, and /n/ - [noi L]. The worst results were obtained when 20 Bark-scaled DCTs were used as parameters: the highest  $C_{llr} = 0.51$  and an EER = 18%. However, using 15 and 20 DCTs in a Hertz scale produced marginally higher  $C_{llr}$  values of 0.49 and 0.50, respectively. The best calibrated consistent-with-fact SSlog<sub>10</sub>LR = 2.40 was obtained when 15 DCTs in a Hertz scale were parameterized. The worst  $C_{llr} = 0.51$  and an EER = 18 obtained for this experimental setting are better than those of the best performing parameters of the previous target segments, /s/ - [sa:m LH] ( $C_{llr} = 0.70$  and EER = 19), /tch/ - [tchai HL] ( $C_{llr} = 0.63$  and EER = 19), and /n/ - [noi L] ( $C_{llr} = 0.54$  and EER = 18).

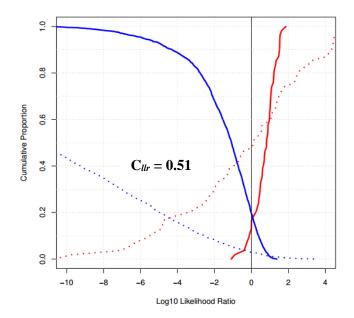
The Tippett plots of the best and worst performing parameters of /n/ - [na: HL], based on the  $C_{llr}$  and EER values, are presented in Figures 44 and 45 (overleaf).

Figure 44 shows that, apart from the lowest  $C_{llr} = 0.47$ , the best calibrated consistent-with-fact  $SSlog_{10}LR = 1.99$  was obtained when 15 Bark-scaled DCTs of /n/ - [na: HL] were parameterized (compared to those of the previous target segments, /s/ - [sa:m LH], /te<sup>h</sup>/ - [te<sup>h</sup>ai HL/, and /n/ - [noi L]); ca. 14% of DS comparisons had calibrated consistent-with-fact  $DSLog_{10}LRs$  less than -4. These results of  $C_{llr}$ , the magnitude and proportion of calibrated  $SSLog_{10}LRs$  and  $DSLog_{10}LRs$ , are the best obtained so far.

The worst  $C_{llr} = 0.51$  was obtained for /n/ - [na: HL] when its 20 Bark-scaled DCTs were parameterized (Figure 45). The largest calibrated consistent-with-fact  $SSlog_{10}LR = 1.87$ 



**Figure 44:** Tippett plot of the best performing parameter, 15 Bark-scaled DCTs of /n/ - [na: HL]. The (red) curves rising to the right represent the cumulative proportion of the SS (same speaker) comparisons, with the log<sub>10</sub>LRs equal to or less than the value indicated on the x-axis, while the (blue) curves rising to the left represent the cumulative proportion of the DS (different speaker) comparisons, with the log<sub>10</sub>LRs equal to or greater than the value indicated on the x-axis. Dotted lines and solid lines represent the uncalibrated and calibrated SS and DS Log<sub>10</sub>LRs, respectively.



**Figure 45:** Tippett plot of the worst performing parameter, 20 Bark-scaled DCTs of /n/ - [na: HL]. The (red) curves rising to the right represent the cumulative proportion of the SS (same speaker) comparisons, with the log<sub>10</sub>LRs equal to or less than the value indicated on the x-axis, while the (blue) curves rising to the left represent the cumulative proportion of the DS (different speaker) comparisons, with the log<sub>10</sub>LRs equal to or greater than the value indicated on the x-axis. Dotted lines and solid lines represent the uncalibrated and calibrated SS and DS Log<sub>10</sub>LRs, respectively.

was obtained with this parameter and ca. 11% of DS comparisons had calibrated consistent-with-fact  $DSlog_{10}LRs \le -4$  (which is marginally smaller than for the 15 Bark-scaled DCTs parameter).

Given the best Tippett plots of /n/ - [nɔi L] and /n/ - [na: HL] in Figures 42 and 44, we see that /n/ - [na: HL] outperforms /n/ - [nɔi L], showing a lower  $C_{llr}$  = 0.47, as opposed to the 0.54 that was obtained for /n/ - [nɔi L]. Additionally, the largest calibrated consistent-with-fact  $SSLog_{10}LR = 1.99$  obtained for /n/ - [na: HL] was marginally stronger than for /n/ - [nɔi L] (calibrated consistent-with-fact  $SSLog_{10}LR = 1.76$ ). The underlying reason why /n/ - [na: HL] outperforms /n/ - [nɔi L] in terms of  $C_{llr}$  might be that a contour tone of /n/ - [na: HL] potentially provides more space for the speakers to exhibit their individualizing information than the level tone of /n/ - [nɔi L]. In addition, more jaw opening for [a:] - [na: HL] as for /ɔi/ - [nɔi L] might contribute to a better performance of /n/ - [na: HL] than of /n/ - [nɔi L].

#### 5.4.4.5 Nasal /m/ extracted from the word [mai HL] 'no'

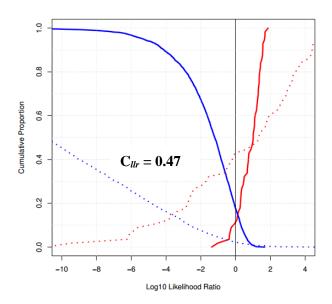
Table 22 shows that relatively good results were obtained when both the 20 Hertz- and 20 Bark-scaled DCTs of the nasal /m/ - [mai HL] were parameterized, as the Cllr values obtained were 0.47 and 0.49, respectively. 57 speakers were tested.

		Her	rtz			Bai	·k		
Parameters	Cali	brated	$\mathbf{C}_{llr}$	EER	Calibrated		$\mathbf{C}_{llr}$	EER	
	LO	G <sub>10</sub> LR			LOG <sub>10</sub> LR				
	SS	DS			SS	DS			
/m/ 15 coeffs	≤ 1.73	≥ -13.84	0.53	13	≤ 1.72	≥ -13.95	0.54	14	
/m/ 20 coeffs	≤ 1.87	≥ −15.14	0.47	14	≤ 1.96	≥ −15.61	0.49	13	

**Table 22:** Calibrated Log<sub>10</sub>LR, C<sub>llr</sub>, and EER of the nasal /m/ - [mai HL] when its 15 DCTs and 20 DCTs were parameterized in both Hertz and Bark scales, respectively. The best C<sub>llr</sub> and EER values are highlighted in blue; the worst are shown in red.

Using fewer DCTs had a negative effect, as was evident from the higher  $C_{llr}$  values of 0.53 and 0.54 when 15 DCTs (as opposed to 20 DCTs), in both Hertz and Bark scales, were parameterized. Figures 46 and 47 (overleaf) show the Tippett plots of the best and worst performing parameters based on the  $C_{llr}$  and EER values of /m/ extracted from the word [mai HL] 'no'.

As shown in Figure 46, the greatest (calibrated) strength of evidence obtained for SS comparisons was  $SSLog_{10}LR = 1.87$ ; for DS comparisons it was  $DSLog_{10}LR = -15.14$ , of which ca. 10% had  $DSLog_{10}LRs \le -4$ . The lowest levels obtained for  $C_{llr} = 0.47$  and EER = 14% were acceptably low and were obtained with this set of parameters.

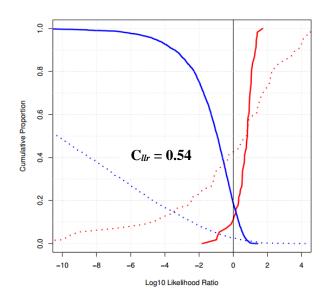


**Figure 46:** Tippett plot of the best performing parameter, 20 Hertz-scaled DCTs of /m/ - [mai HL].

The (red) curves rising to the right represent the cumulative proportion of the SS (same speaker) comparisons, with the log<sub>10</sub>LRs equal to or less than the value indicated on the x-axis, while the (blue) curves rising to the left represent the cumulative proportion of the DS (different speaker) comparisons, with the log<sub>10</sub>LRs equal to or greater than the value indicated on the x-axis. Dotted lines and solid lines represent the uncalibrated and calibrated SS and DS Log<sub>10</sub>LRs, respectively.

Figure 47 (overleaf) shows that a marginally higher  $C_{llr} = 0.54$  was obtained with 15 DCTs extracted in a Bark scale for /m/ - [mai HL]. The largest SSLR obtained was  $SSlog_{10}LR = 1.72$  (providing "limited" support for a SS hypothesis) and only ca. 8% of DS comparisons had  $DSLog_{10}LRs \le -4$ .

So far we see that, using the DCTs extracted from the midpoint of /s,  $tc^h$ , n, m/, the  $C_{llr}$  values obtained range from 0.47 for /n, m/ to 0.77 for /s/. With reference to Table 23 (§5.5), we see a general trend emerging from the findings. Table 23 ranks the linguistic-phonetic segments experimented on in terms of their  $C_{llr}$  and EER values (from low to high) with their corresponding acoustical parameters.



**Figure 47:** Tippett plot of the worst performing parameter, 15 Bark-scaled DCTs of /m/ - [mai HL].

The best  $C_{llr}$  and EER values are highlighted in blue; the worst are shown in red. The (red) curves rising to the right represent the cumulative proportion of the SS (same speaker) comparisons, with the  $\log_{10}LRs$  equal to or less than the value indicated on the x-axis, while the (blue) curves rising to the left represent the cumulative proportion of the DS (different speaker) comparisons, with the  $\log_{10}LRs$  equal to or greater than the value indicated on the x-axis. Dotted lines and solid lines represent the uncalibrated and calibrated SS and DS  $\log_{10}LRs$ , respectively.

#### 5.5 Overall comparisons and discussions

The best results of using the DCT coefficients fitted to a spectrum are summarized in Table 23. As expected, nasals performed the best, with the lowest  $C_{tlr} = 0.47$  for /n/ - [na: HL] in a Hertz scale. This finding agrees with what is reported in Amino et al. (2006), who investigated the level of speaker individuality found in the nine Japanese consonants /t, d, s, z, r, j, m, n, jn/. They found that nasals performed the best (cf. §2.21). However, I cannot directly compare the results obtained in the current thesis with these from previous studies, due to different experimental settings, e.g. different numbers of speakers (5 speakers for Japanese vs 57 speakers for Standard Thai) and the statistical techniques used (*F*-ratio metric for Japanese vs MVLR for Standard Thai). The Standard Thai affricate /tch/ ( $C_{tlr} = 0.63$ , EER = 19%) performed better than the English affricate /tf/ ( $C_{tlr} = 0.98$  and EER = 44%) as reported in Franco-Pedroso et al. (2012). If we compare the results of the Standard Thai fricative /s/ ( $C_{tlr} = 0.70$ , EER = 19%) with those of previous FVC experiments, the Standard Thai fricative /s/ performed much worse than the English /s/ (lowest  $C_{tlr} = 0.55$  and EER = 17%) (cf. Kavanagh, 2012).

I now turn to a detailed discussion of the best performing parameter for each of the target segments. The overall results shown in Table 23 confirm that the nasals /n, m/ perform better (lowest  $C_{llr} = 0.47$ , largest consistent-with-fact  $SSLog_{10}LR = 2.40$ , ca. 14% of DS comparisons  $\leq -4$ ).

Linguistic- phonetic segments	$\mathbf{C}_{llr}$	EER	Acoustical parameters
/n/ - [na: HL]	0.47	15	15 Hertz-scaled DCTs
/m/ - [mai HL]	0.47	14	20 Hertz-scaled DCTs
/n/ - [nɔi L]	0.54	18	20 Hertz-scaled DCTs
/te <sup>h</sup> / - [te <sup>h</sup> ai HL]	0.63	19	20 Hertz-scaled DCTs
/s/ - [sa:m LH]	0.70	19	15 Hertz-scaled DCTs

**Table 23:** Ranking order of the linguistic-phonetic segments experimented on in terms of their  $C_{llr}$  and EER values (from low to high) with their corresponding acoustical parameters.

The best  $C_{llr}$  values for each of the parameters trialed can be ranked in order from low to high as shown in Table 23. The nasals /n, m/ performed the best ( $C_{llr} = 0.47$ ), the affricate /tch/ performed marginally worse than the nasals ( $C_{llr} = 0.54$ ) and the fricative /s/ performed the worst ( $C_{llr} = 0.70$ ). In addition, the findings show that when the DCTs are parameterized in a Hertz scale they outperform those in a Bark scale. This implies that warping the spectral information in a perception scale might not be beneficial for extracting individualizing information in Standard Thai. Additionally, if a nasal /n/ is to be of use in Standard Thai FVC, /n/ should be extracted from a word such as [na: HL], instead of a particle such as [nɔi L].

#### 5.6 Summary

In this chapter I first presented the segmentation procedure of the Standard Thai /s,  $te^h$ , n, m/. Then the spectral moments (mean, variance, skew, and kurtosis) of /s/ were statistically analyzed using ANOVA, then MVLR. Although the ANOVA outputs gave significant results for the spectral moments of /s/ (p < 0.05), the derived FVC values showed only "limited support" at best for SS comparisons, with the highest  $C_{llr} = 0.92$ . Additionally, only 1% of the calibrated consistent-with-fact DSLog<sub>10</sub>LR was below –4. This means that the magnitude of such derived LRs was fairly weak. As such, the DCTs

for /s/ were further experimented on to see if the DCT parameters perform better than the spectral moments. The results showed that the DCTs performed better than the spectral moments on the basis of  $C_{llr}$  and EER values with the greatest calibrated consistent-with-fact  $SSLog_{10}LR \leq 2.40$ . Thus, I decided to use the DCT based parameters for the other segments of Standard Thai /tch, n, m/.

### Chapter 6

# Results of the formant trajectories of the diphthongs [ɔi] - [nɔi L] and [ai] - [mai HL]

#### **6.1 Introduction**

This chapter first presents the FVC results when the F2 trajectory and F1-F3 trajectories of the diphthongs [ $\mathfrak{s}i$ ] - [ $\mathfrak{n}\mathfrak{s}i$  L] and [ $\mathfrak{a}i$ ] - [ $\mathfrak{m}\mathfrak{a}i$  HL] were parameterized by cubic polynomials. After explaining the underlying reasons why these diphthongs were chosen, I show how to annotate such phonetic targets and explain how to correct the corresponding formant tracking errors. Regarding the results of the F2 trajectory,  $C_{llr} = 0.67$  and  $C_{llr} = 0.69$  were obtained for [ $\mathfrak{s}i$ ] - [ $\mathfrak{n}\mathfrak{s}i$  L] and [ $\mathfrak{a}i$ ] - [ $\mathfrak{m}\mathfrak{s}i$  HL], respectively. However, the FVC performance significantly improved when F1 and F3 trajectories were added in addition to F2, resulting in the lowest  $C_{llr} = 0.42$  for [ $\mathfrak{s}i$ ] - [ $\mathfrak{n}\mathfrak{s}i$  L] and  $C_{llr} = 0.49$  for [ $\mathfrak{s}i$ ] - [ $\mathfrak{m}\mathfrak{s}i$  HL].

### 6.2 Why were the F2 trajectories of the diphthongs [ɔi] - [nɔi L] and [ai] - [mai HL] experimented on?

It is well known that the measurement of F1 is quite often subject to different transmission channels (Künzel, 2001). It is therefore generally assumed that F1 should be excluded from FVC analyses. This is because the intrinsically low F1 values of the high vowels /i:, I, II/I are shifted upwards by the telephone effect, which might result in a faulty F2 measurement (Künzel, 2001, p. 89). Byrne and Foulkes (2007) also examined the effect of mobile-telephone transmission on first through third formant values (F1, F2, F3), which were measured at the temporary stable midpoint of the vowels. They found that the mean F1 values were 29% higher in telephone conditions than they were in speech recorded using a direct high-quality microphone. Chen, Shen, Campbell, and Schwartz (2009) also empirically tested the effect of mobile and landline conditions on formants from fully automatic formant measurements and F2 values were found to be lower in mobile conditions. However, Byrne and Foulkes (2007) suggested that F2 and F3 can still be used in FVC.

In the first experiment on the diphthongs [5i] - [n5i L] and [6i] - [mai HL], only the F2 trajectories were tested (followed by the F1-F3 trajectories of the diphthongs [5i] - [n5i] L] and [ai] - [mai HL]). This is because F2 is most robust. As previously mentioned, in forensically realistic conditions, F1 values (of high and possibly mid vowels) are usually compromised by a telephone's band-pass, as is also the case for F3, especially when the telephone transmission is very bad (Rose et al., 2006, p. 331). Because of this compromise, only the F2 trajectory was chosen for the first experiment to simulate forensically realistic conditions. There are two reasons for choosing these particular two diphthongs. First, the diphthongs [5i] and [ai], traversed a large part of acoustical vowel space, so they were expected to exhibit greater between- to within-speaker variation. Specifically, /ɔ/ is a low-mid back vowel and /i/ is a high front vowel, thus these two vocalic targets provide more space for speakers to exhibit their variation, due to much movement in the vocal tract. Therefore, more individualizing information can be gained from this wide articulatory movement, which might further contain more speaker-specific information useful for FVC. Similarly, the diphthong [ai] is also interesting as it involves two widely separated articulatory targets, i.e. a low central vowel [a] and a high front vowel [i]. Second, since [5i] occurs in a sentence-final position, more duration is guaranteed because it tends to be stressed (as previously mentioned in §5.2.3, the duration of [n] - [noi L] was at least 117.40 ms and a maximum was 459.91 ms for this experiment).

#### **6.3 Informants**

All speech samples of [5i] - [n5i L] and [ai] - [mai HL] were extracted from 30 speakers.

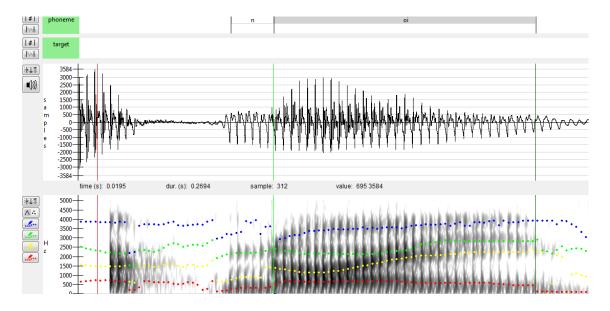
#### **6.4 Segmentation**

In what follows, I visually explain how I segmented the target diphthongs [5i] - [n5i L] (§6.4.1) and [ai] - [mai HL] (§6.4.2), using the displays of the EMU speech database system (Cassidy, 1999).

#### 6.4.1 Formant trajectories of [ɔi] - [nɔi L]

As mentioned previously, all the [ɔi] - [nɔi L] samples used in this experiment were extracted from the following sentence frame:

This sentence can be translated into English as "This is because we do not have any responsibility". The highlighted /ɔi/ is the target segment and the underlined words represent stress. Figure 48 reproduces the displays of the EMU speech database system (Cassidy, 1999) for the diphthong [ɔi] - [nɔi L]; it shows a label tier, a waveform, and a spectrogram with the corresponding formant tracking values.



**Figure 48:** Label tier (top), waveform (middle), and spectrogram (bottom) of [ɔi] - [nɔi L], with the corresponding formant frequencies tracking. Red, yellow, green, and blue dots represent F1, F2, F3, and F4, respectively.

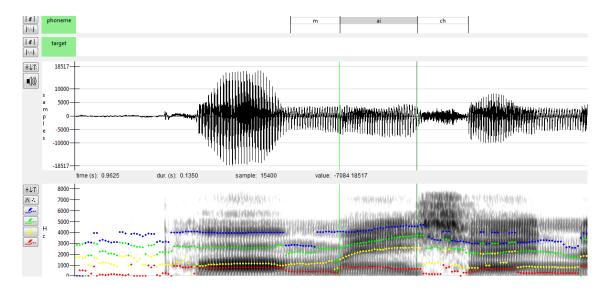
Note: [ɔi] is labeled as [oi].

A diphthong is defined as "two vocalic targets in a single syllable" (Rose, 2002, p. 241). As can be seen in Figure 48, the articulatory movement of [ɔi] nicely reflects the changing F-patterns (F1-F4) as the tongue moves smoothly from one target to another. The first target of /ɔ/ is realized as a low-mid back rounded vowel and the second target /i/ as a high front unrounded vowel. The beginning of the first target [ɔ] was located at the F2 vowel onset, as indicated by a clear dark band spectrogram (the first vertical green line). For [ɔ], we observe that F1 is high and is close to F2, indicating a back vowel (Rose, 2002, p. 241). Additionally, F1 is fairly constant and F2 is continuously increasing to converge with F3 throughout [ɔ]. Regarding the second diphthongal target /i/, F1 is a bit lower and F2 is higher than in the case of /ɔ/, indicating a high front vowel (ibid.). The offset of /i/ was thus located at the point where F2 reaches its maximum frequency as

indicated by the second vertical green line. The corresponding acoustical properties of the target segment [ai] - [mai HL] are presented in Figure 49.

#### 6.4.2 Formant trajectoires of [ai] - [mai HL]

Figure 49 reproduces the displays of the EMU speech database system (Cassidy, 1999) for the diphthong [ai] - [mai HL]; it shows a label tier, a waveform, and a spectrogram with the corresponding formant tracking values.



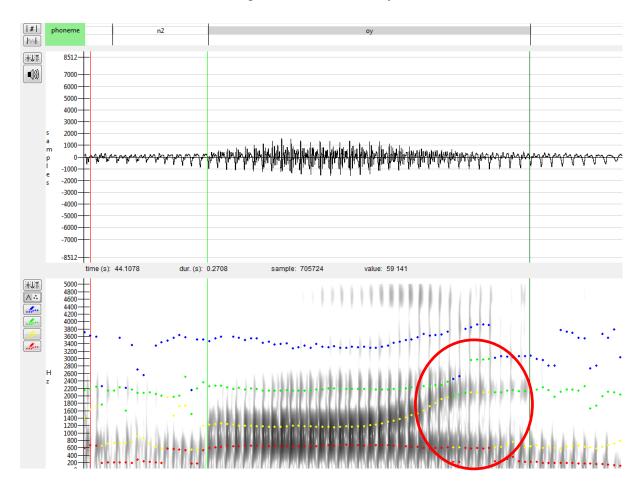
**Figure 49:** Label tier (top), waveform (middle), and spectrogram (bottom) of [ai] - [mai HL] with the corresponding formant frequencies tracking. Red, yellow, green, and blue dots represent F1, F2, F3, and F4, respectively.

The diphthong [ai] involves a low central vowel /a/ and a high front vowel /i/. The first vocalic target /a/ is phonetically realized as a short half-open vowel [v], as it is influenced by the height of the second vocalic target /i/ (Rose, 2002, pp. 241-242). We also observe that the F2 and F3 of this /a/ are not static but increasing as the tongue is moving for the next high front vowel /i/ (ibid.). As such, the beginning point of a diphthong [ai] was marked at F2 onset, where there was a sudden change in F-patterns between /m/ and /a/ as indicated by the first vertical green line in Figure 49. In other words, the starting point was located at the earliest point right after the low amplitude of the preceding /m/. The offset was marked when the F2 of /i/ has reached its maximum frequency in conjunction with the beginning of the frication noise of the following affricate /tch/, as indicated by the second vertical green line. Notable in Figure 49 is the individualizing variation in the segment following the [ai] vowel. As clearly seen in the third panel of Figure 49, an

alveolar affricate /tch/ is realized as the fricative alveolar /s/. This is interesting as it is auditorily perceived as an affricate by the researcher.

#### 6.4.3 Formant tracking errors and manual correction

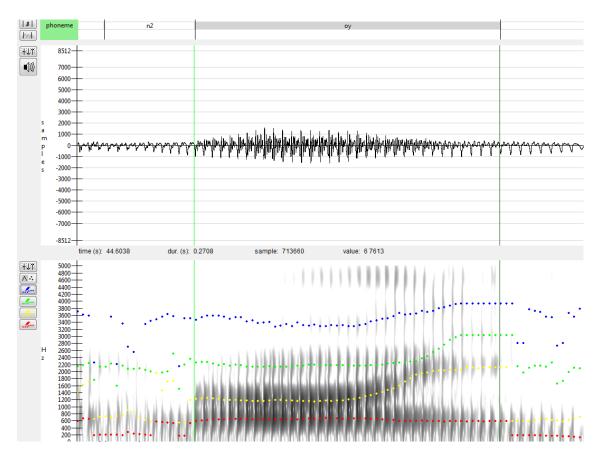
This section presents the formant tracking errors and corresponding manual corrections of the diphthong [ɔi] - [nɔi L]. Figure 50 shows the formant tracking errors while Figure 51 shows how such formant tracking errors were manually corrected.



**Figure 50:** Label tier (top), waveform (middle), and spectrogram (bottom) of [ɔi] - [nɔi L] with the corresponding formant tracking containing some errors (in the last one-third).

Note: [nɔi] is labeled as [n2oy].

Figure 50 shows that manual correction of [5i]'s formant tracking was needed as some errors were observed during the last one-third of this particular token (as indicated by the red circle). That is, some blue, green, yellow, and red dots, which represent F4 to F1, respectively, had shifted downwards from their correct formant tracking position. Figure 51 shows how such errors were corrected.

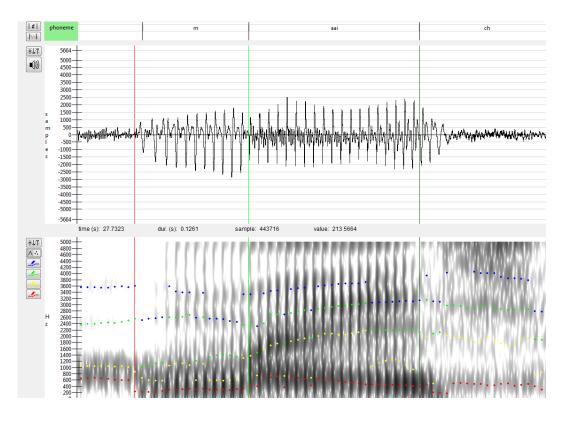


**Figure 51:** Label tier (top), waveform (middle), and spectrogram (bottom) of [ɔi] - [nɔi L] with the corresponding formant tracking after manual correction.

Note: [nɔi] is labeled as [n2oy].

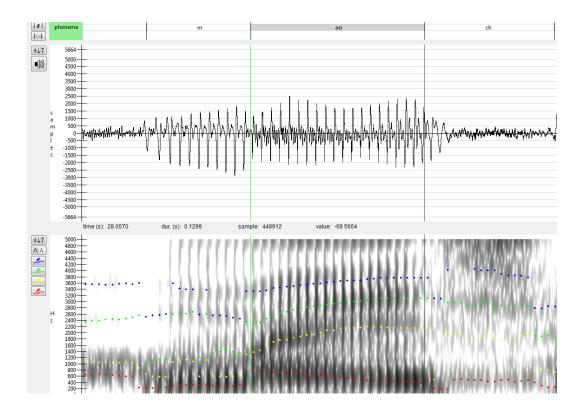
We observe from Figure 51 that, during the last one-third of this token, red, yellow, green and blue dots that represent F1 to F4, respectively, were manually shifted upwards. That is, red dots were shifted to the place where the yellow dots used to be. The same technique was applied to correct F2-F3. In contrast, it is not easy to correct F4 values, in this case because the highest (horizontal) dark band spectrogram is not easily observed. As such, F4 errors were manually shifted upwards using the preceding F4 values as the clues (F4 was not tested in the current experiment).

We now move on to the diphthong [ai], showing how to correct its formant trajectories. Figure 52 (overleaf) shows a label tier, a waveform, and a spectrogram of [ai] - [mai HL] with the corresponding formant tracking containing some errors. There were F-pattern tracking errors throughout the [ai] token. Again, some F4-F2 values (blue, green, and yellow dots) had shifted downwards especially towards the end of their trajectories. In contrast, fewer errors of F1 (red dots) were observed at the beginning portion. As such, the formant trajectory values were manually corrected as shown in Figure 53.



**Figure 52:** Label tier (top), waveform (middle), and spectrogram (bottom) of [ai] - [mai HL] with the corresponding formant tracking containing some errors.

Note: [ai] is labeled as [aai].



**Figure 53:** Label tier (top), waveform (middle), and spectrogram (bottom) of [ai] - [mai HL] with the corresponding formant tracking after manual correction.

Note: [ai] is labeled as [aai].

Figure 53 shows the formant trajectories of [ai] after some errors were manually corrected. As previously discussed for [bi], higher formant values were used as clues in correcting those of lower formants. That is, some erroneous red dots at the beginning of F1 were manually placed at positions where F2 values were formerly tracked by the EMU speech database system (Cassidy, 1999). This same technique was also applied to F2 and F3. Regarding F4, the highest (horizontal) dark band spectrogram was used to track F4 values (although F4 was not tested in the current experiment).

#### **6.4.4 Discarding of poor recording speech samples**

Some tokens of [ɔi] - [nɔi L] and [ai] - [mai HL] were excluded from the experiment because of low amplitude caused by poor recording quality. This might be the result of a poorly placed clipping microphone below the informants' chin and/or the mal-functioning Roland® UA-25EX USB Audio Capture card itself. The formant tracking errors triggered by low amplitude are shown in Figure 54, where the [ɔi] diphthong of one of our informants, Speaker 9, is used as an example. As discussed in §6.2, forensic recording conditions are notoriously bad and F2 is able to be extracted from very bad quality recordings, thus only the F2 trajectories will be tested in addition to the F1-F3 trajectories.

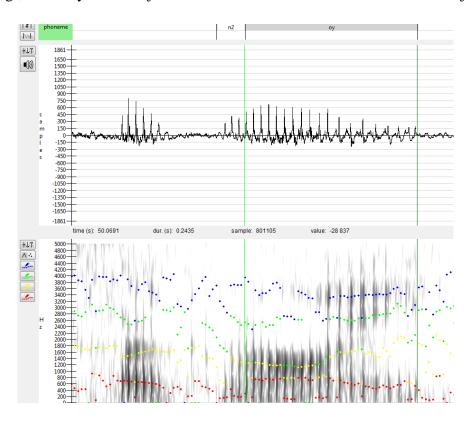


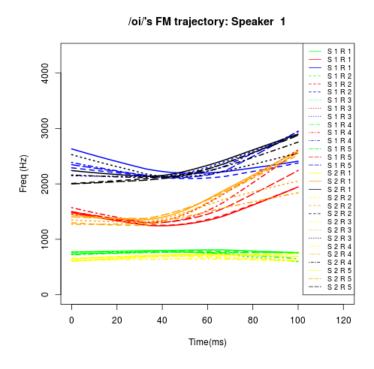
Figure 54: Label tier (top), waveform (middle), and spectrogram (bottom) of [ɔi] - [nɔi L] containing many formant tracking errors.

Note: [nɔi] is labeled as [n2oy].

Figure 54 clearly shows multiple errors in formant tracking, making it very difficult to identify the onset and offset of this particular token of [5i]. For this reason, tokens of this kind with extremely bad F-pattern tracking were discarded in the current experiment.

#### 6.4.5 Formant trajectories of the diphthong [5i]

This section shows the formant trajectories (F1-F3) of the diphthongs [ɔi] - [nɔi L] plotted against the normalized duration in 100 msec. Figure 55 shows the F1-F3 trajectories of [ɔi] - [nɔi L] for Speaker 1. Plots of F1-F3 trajectories for all 30 speakers can be found in Appendix B.

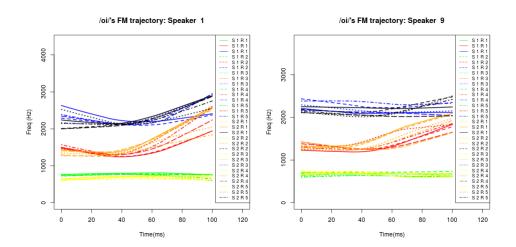


**Figure 55:** Speaker 1's F1-F3 trajectories of the diphthong [ɔi] - [nɔi L] plotted against normalized duration (100 msec).

S stands for session (1-2) and R stands for repeat (1-5). F1 is presented in green for session1 and yellow for session 2, F2 is presented in red for session 1 and orange for session 2, and F3 is presented in blue for session 1 and black for session 2. Note: /ɔi/ is labeled as /oi/.

Since it is crucial to make sure at the outset that the input data for the MVLR formula, i.e. the formant trajectories, are properly extracted, F1-F3 values are plotted accordingly. Thus, Figure 55 exhibits the ten trajectories (2 sessions x 5 repeats) for each of F1, F2 and F3, extracted from the diphthong [ɔi] - [nɔi L] uttered by Speaker 1. F1 trajectories (green and yellow) show relatively good consistency within the frequency range of ca. 500-800 Hz. F2-F3 trajectories, on the other hand, exhibit greater within-speaker variation, i.e. the concave shape of F2 (red and orange) varies a lot after 40 msec while

that of F3 (black and blue) exhibits much within-variation at the onsets and offsets. In Figure 56, between-speaker variation is illustrated for [ɔi] - [noi L] by means of the F1-F3 trajectories uttered by Speakers 1 and 9.



**Figure 56:** Speaker 1 and Speaker 9's F1-F3 trajectories of the diphthong [ɔi] - [nɔi L] plotted against normalized duration (100 msec).

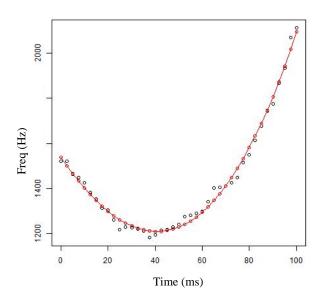
S stands for session (1-2) and R stands for repeat (1-5). F1 is presented in green for session 1 and yellow for session 2, F2 is presented in red for session 1 and orange for session 2, and F3 is presented in blue for session 1 and black for session 2. Note: /ɔi/ is labeled as /oi/.

Figure 56 shows the F1-F3 trajectories of the diphthong [5i] - [n5i L] uttered by Speaker 1 (left) and Speaker 9 (right). The F1 trajectories (green and yellow) of Speakers 1 and 9 show a relatively similar straight line within the frequency range of ca. 500-800 Hz. The F2 trajectories (red and orange) also show a similar concave contour although Speaker 1 shows the F2 offset at a higher frequency of ca. 2500 Hz. In the case of the F3 trajectories (black and blue), on the other hand, much between-speaker variation is observed. That is, F3 trajectories show a convex shape for Speaker 1 but are relatively straight for Speaker 9. In §6.4.6, the cubic polynomial curve fitting is presented for [5i] - [n5i L].

#### **6.4.6** Polynomial curve fitting (cubic polynomials)

In the current thesis, I approximate formant trajectories using a cubic polynomial function of the type  $ax^3 + bx^2 + cx + d$ , where a, b, c, and d are the coefficients and x is the time. The cubic polynomials can approximate the formant trajectory with an 'S' shaped trajectory, while lower order polynomials such as quadratic can only approximate the 'U' shaped trajectory and the linear polynomials can approximate only a straight line (Morrison, 2008, pp. 252-255). The cubic polynomials are selected in the current thesis as they adequately represent the formant trajectories of the current experimental data of

[oi] - [noi L], whose complex shape is either 'U' or 'S' (see §6.4.5 for the F1-F3 trajectory plots). As such, cubic polynomials fitted to the formant trajectories in normalized duration (100 msec) were used as parameters. The use of cubic polynomials fitted to normalized duration of the formant trajectories was justified by the results obtained in previous FVC studies (see §2.23.2), in which cubic polynomials fitted to equalized duration performed better than quadratic polynomials fitted to absolute duration (Morrison & Kinoshita, 2008; Morrison & Kondaurova, 2009). Moreover, the pilot study of Standard Thai diphthongs [i:aw], [u:a] and [u:a] presented in §4.1 also confirmed that, when cubic polynomials were used, the best results of  $C_{llr} = 0.02$ -0.04 were obtained. Figure 57 is an example of a cubic polynomial curve fitting, in which the F2 trajectory values of [oi] - [noi L] are plotted together with its polynomial fitting of  $(0.000531)x^3 + 0.155159x^2 + (-15.291889)x + 1539.427481$ . All F1-F3 trajectories of [oi] - [noi L] plotted together with cubic polynomials can be found in Appendix C.



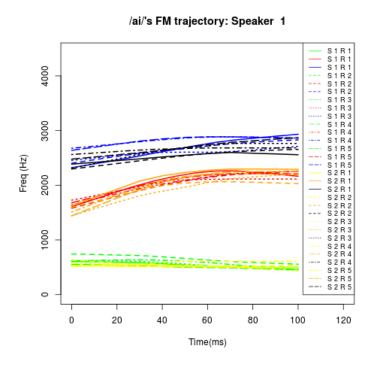
**Figure 57:** F2 trajectory values of [ $\mathfrak{i}$ i] - [ $\mathfrak{n}\mathfrak{i}$  L] (black dots), plotted together with its cubic polynomial curve fitting (dotted red line) of  $(0.000531)x^3 + 0.155159x^2 + (-15.291889)x + 1539.427481$ .

Figure 57 shows the F2 trajectory (black dots) of [ɔi] - [nɔi L], which is relatively well-fitted by the cubic polynomials (dotted red line). Trying to fit such an F2 trajectory with fourth order polynomials might come at the risk of overfitting what is essentially imprecise raw data. After the F2 trajectory values were extracted by the EMU speech database system, the formant trajectories of this particular token [ɔi] - [nɔi L] could be

parameterized as [0.000531, 0.155159, -15.291889, 1539.427481] and used as input for MVLR calculation. The formant trajectories of [ai] - [mai HL], together with its curve fitting, are presented in §6.4.7.

#### 6.4.7 Formant trajectories of the diphthong [ai] - [mai HL]

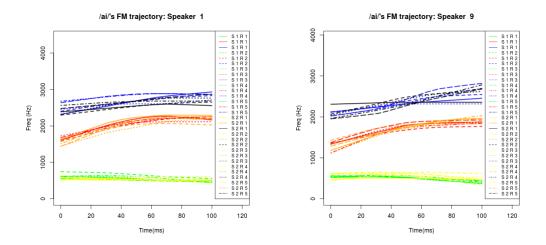
This section shows the formant trajectories (F1-F3) of the diphthong [ai] - [mai HL] plotted against normalized duration (100 msec). Plots of the F1-F3 trajectory of [ai] - [mai HL] for all 30 speakers can be found in Appendix D.



**Figure 58:** Speaker 1's F1-F3 trajectories of the diphthong [ai] - [mai HL] plotted against the normalized duration (100 msec).

S stands for session (1-2) and R stands for repeat (1-5). F1 is presented in green for session 1 and yellow for session 2, F2 is presented in red for session 1 and orange for session 2, and F3 is presented in blue for session 1 and black for session 2.

Figure 58 exhibits the ten trajectories (2 sessions x 5 repeats) for each of F1, F2 and F3 extracted from the diphthong [ai] - [mai HL] uttered by Speaker 1. The F1-F3 formant trajectories have considerably consistent contours. However, F1 (green vs yellow) and F2 (red vs orange) trajectories exhibit less between-session variation than the F3 (black and blue) trajectory. To illustrate between-speaker variation, Speaker 1 and Speaker 9's F1-F3 trajectories of the diphthong [ai] - [mai HL] are shown in Figure 59 (overleaf).



**Figure 59:** Speaker 1 and Speaker 9's F1-F3 trajectories of the diphthong [ai] - [mai HL] plotted against normalized duration (100 msec).

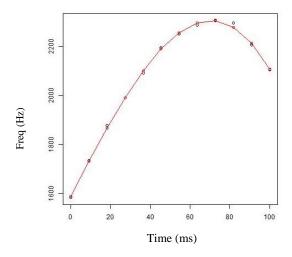
S stands for session (1-2) and R stands for repeat (1-5). F1 is presented in green for session 1 and yellow for session 2, F2 is presented in red for session 1 and orange for session 2, and F3 is presented in blue for session 1 and black for session 2.

Figure 59 shows Speaker 1 and Speaker 9's F1-F3 trajectories of the diphthong [ai] - [mai HL] plotted against normalized duration (100 msec). The F1 trajectories of Speaker 9 show marginally less between-session than those of Speaker 1 within ca. 500-800 Hz. F2 trajectories occupy a higher frequency range (ca. 1400-2200 Hz) for Speaker 1 than for Speaker 9 (ca. 1100-1800 Hz). Moreover, the F3 trajectories of Speaker 9, except a token from session 2 repeat 1 (black solid line), exhibit a slight 'S' shape in a lower frequency (ca. 1900-2800 Hz), while those of Speaker 1 show a slightly convex shape in a higher frequency range (ca. 2300-3000 Hz).

#### 6.4.8 Polynomial curve fitting of the diphthong [ai] - [mai HL]

This section shows how well cubic polynomials fitted the F2 trajectory of [ai] - [mai HL]. Duration was normalized to 100 msec. All F1-F3 trajectories of [ai] - [mai HL] plotted together with cubic polynomials can be found in Appendix E.

Figure 60 (overleaf) exemplifies a cubic polynomial curve fitting in which the formant trajectory of [ai] - [mai HL] is plotted together with its polynomial fitting of  $(-0.000854)x^3 + (-0.023227)x^2 + (16.052403)x + 1589.501831$ . We can see in Figure 60 that the cubic polynomials (dotted red line) fit fairly well to the F2 trajectory (black dots). §§6.5 and 6.6 present the results of 1) the F2 trajectory of [5i] - [noi L] and [ai] - [mai HL] and 2) the F1-F3 trajectories of [5i] - [noi L] and [ai] - [mai HL] in terms of Log10LR,



**Figure 60:** F2 trajectory values (black dots) of the diphthong [ai] - [mai HL] plotted together with its cubic polynomial fitting (dotted red line) of  $(-0.000854)x^3 + (-0.023227)x^2 + (16.052403)x + 1589.501831$ .

Cllr and EER values, respectively. Tippett plots will then be presented followed by discussion and comparison with the results of the pilot studies presented in Chapter 4.

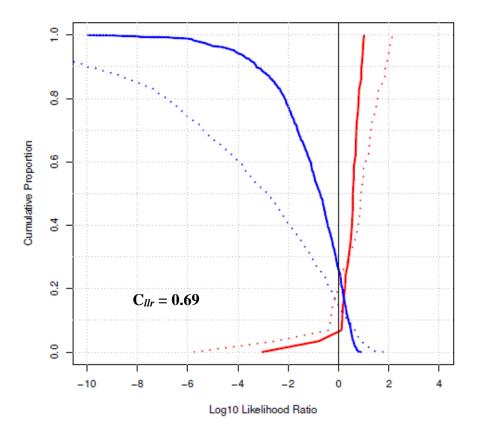
## 6.5 Experimental results: F2 trajectory of diphthongs [ɔi] - [nɔi L] and [ai] - [mai]

This section first tabulates the results of  $Log_{10}LR$ ,  $C_{llr}$  and EER when cubic polynomials fitted to F2 trajectories of the diphthongs [5i] - [n5i L] and [ai] - [mai HL] were parameterized. Then, a Tippett plot will be presented and this will be followed by discussion.

	Cubic Polynomials(F2)							
Diphthongs	Cal	ibrated	$\mathbf{C}_{llr}$	EER				
	LO	$G_{10}LR$						
	SS	DS						
[ɔi] - [nɔi L]	≤ 1.01	≥ -9.96	0.69	19				
[ai] - [mai HL]	≤ 1.38	≥ -19.91	0.67	17				

**Table 24:** Calibrated Log<sub>10</sub>LR,  $C_{llr}$ , and EER values when cubic polynomial coefficients from the diphthongs [ $\mathfrak{o}i$ ] - [ $\mathfrak{n}\mathfrak{o}i$  L] and [ $\mathfrak{a}i$ ] - [ $\mathfrak{m}\mathfrak{o}i$  HL] were parameterized, respectively.

Table 24 shows that the magnitude of the strongest consistent-with-fact  $SSLog_{10}LR$  for both [5i] - [n5i L] and [ai] - [mai HL] only moderately supports the SS hypothesis. As for [5i], the magnitude of its SSLRs is fairly weak; all of them are smaller than the strongest consistent-with-fact  $SSLog_{10}LR = 1.01$ . The diphthong [ai] - [ai HL] shows a similar trend in that the largest  $SSLog_{10}LR = 1.38$  was obtained, suggesting only "moderate" support for the SS hypothesis. For DSLRs, better results,  $log_{10}LR \le -4$ , which strongly support the DS hypothesis, were obtained for both [5i] - [n5i L] and [ai] - [mai HL]. We will look closely at the magnitude of LR values using Tippett plots in Figures 61 and 62.



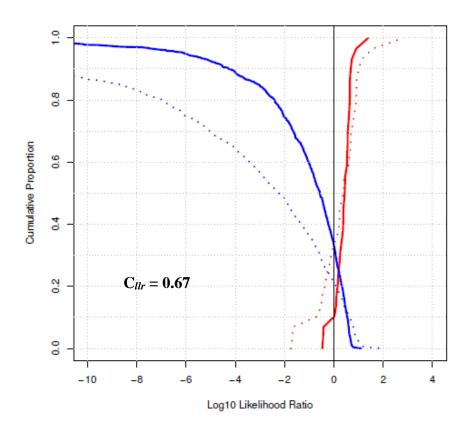
**Figure 61:** Tippett plot of [5i]'s second formant (F2) trajectory when cubic polynomials were parameterized.

The (red) curves rising to the right represent the cumulative proportion of the SS (same speaker) comparisons, with the log<sub>10</sub>LRs equal to or less than the value indicated on the x-axis, while those rising to the left (blue curves) represent the cumulative proportion of the DS (different speaker) comparisons, with the log<sub>10</sub>LRs equal to or greater than the value indicated on the x-axis. Dotted lines and solid lines represent the uncalibrated and calibrated SS and DS Log<sub>10</sub>LRs, respectively.

Figure 61 shows that, when [ $\circ$ i]'s F2 trajectory was parameterized with cubic polynomials, a  $C_{llr}$  value = 0.69 and an EER = 19% were obtained. After calibration, the magnitude of the misleading  $SSLog_{10}LR = -5.74$  and  $DSLog_{10}LR = 1.78$  was reduced to

SSLog<sub>10</sub>LR = -3.03 and DSLog<sub>10</sub>LR = 0.89. The largest calibrated SSLog<sub>10</sub>LR = 1.01 was obtained, suggesting only "moderate" support for the SS hypothesis. For DS comparisons, ca. 6% had DSLog<sub>10</sub>LRs  $\leq -4$ , suggesting "very strong" support for the defense hypothesis.

A Tippett plot of [ai]'s second formant (F2) trajectory when its cubic polynomials were parameterized is shown in Figure 62.



**Figure 62:** Tippett plot of [ai]'s second formant (F2) trajectory when cubic polynomials were parameterized.

The (red) curves rising to the right represent the cumulative proportion of the SS (same speaker) comparisons, with the log<sub>10</sub>LRs equal to or less than the value indicated on the x-axis, while the (blue) curves rising to the left represent the cumulative proportion of the DS (different speaker) comparisons, with the log<sub>10</sub>LRs equal to or greater than the value indicated on the x-axis. Dotted lines and solid lines represent the uncalibrated and calibrated SS and DS Log<sub>10</sub>LRs, respectively.

We see in Figure 62 that, when the F2 trajectory of [ai] - [mai HL] was parameterized by cubic polynomials, marginally lower  $C_{llr} = 0.67$  and EER = 17% (as opposed to  $C_{llr} = 0.69$  and EER = 19% for [ai] - [nai L]) were obtained. All SS comparisons for [ai]'s F2 trajectory gave (consistent-with-fact) SSLRs smaller than SSLog<sub>10</sub>LR = 1.38, suggesting "moderate support" for the SS hypothesis. For DS comparisons, 10% gave calibrated

consistent-with-fact  $DSlog_{10}LRs \le -4$ , showing "very strong" support for the DS hypothesis when the F2 trajectory of [ai] - [mai HL] was parameterized by cubic polynomials.

### 6.6 Experimental results: F1-F3 trajectories of diphthongs [ɔi] - [nɔi L] and [ai] - [mai HL]

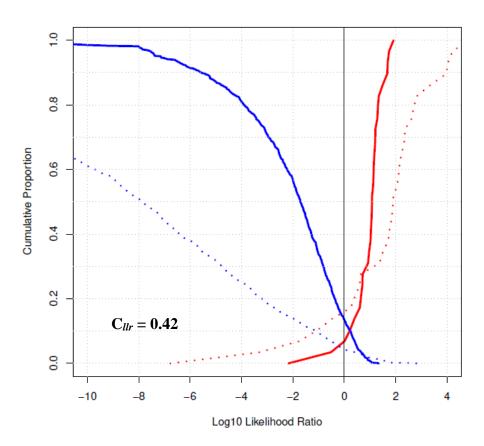
This section, where  $Log_{10}LR$ ,  $C_{llr}$  and EER values of cubic polynomials fitted to F1-F3 trajectories of the diphthongs [ $\sigma$ i] - [ $\sigma$ i] and [ $\sigma$ i] - [ $\sigma$ i] are parameterized, adopts the same convention as the previous one.

	Cubic polynomials (F1-F3)						
Diphthongs	Calil	orated	$\mathbf{C}_{llr}$	EER			
	LOG <sub>10</sub> LR						
	SS	DS					
[ɔi] - [nɔi L]	≤ 1.91	≥ −15.96	0.42	10			
[ai] - [mai HL]	≤ 1.84	≥ -25.06	0.49	18			

**Table 25:** Calibrated Log<sub>10</sub>LR, C<sub>llr</sub>, and EER values when cubic polynomial coefficients of the diphthongs [5i] - [noi L] and [ai] - [mai HL] were parameterized, respectively.

Table 25 shows a substantially lower  $C_{llr} = 0.42$  and  $C_{llr} = 0.49$  for the diphthongs [5i] - [n5i L] and [ai] - [mai HL] when the cubic polynomials of their F1-F3 trajectories (as opposed to only the F2 trajectory) were parameterized. Not only the  $C_{llr}$  values were lower but also the EER values. EER = 10% and EER = 18% were obtained for the diphthongs [5i] - [n5i L] and [6i] - [

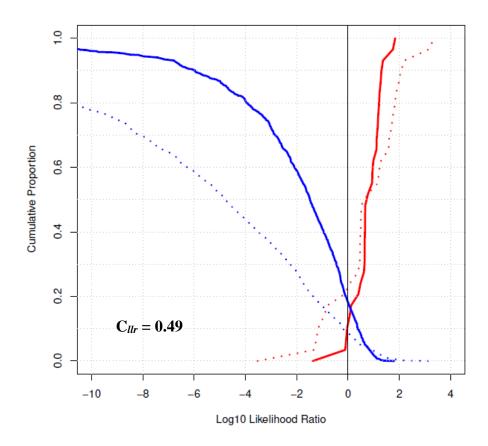
We will look closely at the magnitude of LR values using Tippett plots in Figures 63 and 64.



**Figure 63:** Tippett plot of [51]'s F1-F3 trajectories when cubic polynomials were parameterized. The (red) curves rising to the right represent the cumulative proportion of the SS (same speaker) comparisons, with the log<sub>10</sub>LRs equal to or less than the value indicated on the x-axis, while the (blue) curves rising to the left represent the cumulative proportion of the DS (different speaker) comparisons, with the log<sub>10</sub>LRs equal to or greater than the value indicated on the x-axis. Dotted lines and solid lines represent the uncalibrated and calibrated SS and DS Log<sub>10</sub>LRs, respectively.

Figure 63 reveals that, in comparison with the results of [5i] - [n5i L] using only the F2 trajectory as parameter, we get a better consistent-with-fact  $SSlog_{10}LR = 1.91$  when cubic polynomials fitted to the F1-F3 trajectories of [5i] - [n5i L] are parameterized. For DS comparisons, ca. 20% had calibrated consistent-with-fact  $DSlog_{10}LRs \le -4$  when F1-F3 trajectories of [5i] - [n5i L] were parameterized as compared to only ca. 6% using just the F2 trajectory.

Similarly, Figure 64 (overleaf) reveals that, in comparison to the results obtained for [5i] - [n5i L], when only the F2 trajectory was parameterized by cubic polynomials ( $C_{llr} = 0.67$  and EER = 17%), we obtained a substantially lower  $C_{llr} = 0.49$  but a marginally higher EER = 18% for [ai] - [mai HL], where all F1-F3 trajectories were parameterized by cubic polynomials. After calibration, the largest consistent-with-fact  $SSlog_{10}LR = 3.36$  was reduced in magnitude to  $SSlog_{10}LR = 1.84$ , suggesting only "moderate" support for



**Figure 64:** Tippett plot of [ai]'s F1-F3 trajectory when cubic polynomials were parameterized. The (red) curves rising to the right represent the cumulative proportion of the SS (same speaker) comparisons, with the log<sub>10</sub>LRs equal to or less than the value indicated on the x-axis, while the (blue) curves rising to the left represent the cumulative proportion of the DS (different speaker) comparisons, with the log<sub>10</sub>LRs equal to or greater than the value indicated on the x-axis. Dotted lines and solid lines represent the uncalibrated and calibrated SS and DS Log<sub>10</sub>LRs, respectively.

the SS hypothesis. For DS comparisons, ca. 20% had calibrated consistent-with-fact  $DSlog_{10}LRs \le -4$ , suggesting "very strong" support for the defense hypothesis.

#### 6.7 Discussion

We have observed, on the basis of Log<sub>10</sub>LR, C<sub>llr</sub> and EER values, that, for the same target diphthongs [5i] - [n5i L] and [ai] - [mai HL], FVC performance of the F1-F3 trajectories, which were parameterized by cubic polynomials, is superior to that of the F2 trajectory used on its own. In the case of [5i] - [n5i L], the best SSlog<sub>10</sub>LR = 1.91, DSlog<sub>10</sub>LRs  $\leq$  -4 (20% of DS comparisons), C<sub>llr</sub> = 0.42 and EER = 10% were obtained when F1-F3 trajectories were parameterized by cubic polynomials. A similar trend was observed for [ai] - [mai HL], with the best SSlog<sub>10</sub>LR = 1.84, C<sub>llr</sub> = 0.49 and EER = 18% when F1-F3

trajectories were parameterized as opposed to SSLog<sub>10</sub>LR = 1.38,  $C_{llr}$  = 0.67 and EER = 17% for [ai] - [mai HL] when only the F2 trajectory was parameterized. The first possible reason why the formant trajectories of [ɔi] - [nɔi L] outperformed those of [ai] - [mai HL] might be that the diphthong [ɔi] - [nɔi L] provided more acoustical vowel space for a speaker to exhibit greater individualizing information. That is, [ɔi] - [nɔi L] involves a mid-low *back* vowel [ɔ] and a high front vowel [i], whereas the diphthong [ai] involves a low *central* vowel [a] and a high front vowel [i]. The second possible reason for a better performance of [ɔi] - [nɔi L] might be that a longer duration of at least 117.40 msec and a maximum of 459.91 msec were parameterized for [ɔi] - [nɔi L].

The reason why an F1-F3 trajectory outperforms an F2 trajectory may be in part related to the additional individuating information gained from the articulatory movement reflected in F1, which inversely correlates with vowel height (Nolan, 1983; Rose, 2002). In addition, as Ladefoged and Johnson (2014, p. 207) stated, higher frequency formants "are not uniquely determined for each speaker, but they certainly are indicative of a person's voice quality". Thus, F3 potentially added much individualizing information due to a speaker's voice quality reflected in F3.

# **6.8 Summary**

In this chapter I have presented the FVC results when 1) the F2 trajectory and 2) the F1-F3 trajectories of the diphthongs [ $\mathfrak{s}i$ ] - [ $\mathfrak{n}\mathfrak{s}i$  L] and [ $\mathfrak{a}i$ ] - [ $\mathfrak{m}\mathfrak{s}i$  HL], fitted by cubic polynomials, were parameterized. The findings show that the results of the F1-F3 trajectories outperform those of the F2 trajectory in terms of Log<sub>10</sub>LR, C<sub>llr</sub> and EER values. The best C<sub>llr</sub> and EER values of [ $\mathfrak{s}i$ ] - [ $\mathfrak{n}\mathfrak{s}i$  L] were marginally lower than those of [ $\mathfrak{s}i$ ] - [ $\mathfrak{m}\mathfrak{s}i$ ] - [ $\mathfrak{m}\mathfrak{s}i$ ] + [ $\mathfrak{s}\mathfrak{s}\mathfrak{s}i$ ] trajectories were parameterized. The magnitude of the consistent-with-fact SSLRs is fairly weak, in that the strongest consistent-with-fact SSLog<sub>10</sub>LR = 1.91 for [ $\mathfrak{s}\mathfrak{s}i$ ] - [ $\mathfrak{m}\mathfrak{s}i$ ] -

# **Chapter 7**

# Results of the fundamental frequency (F0): Long-term F0 (LTF0) and tonal F0

## 7.1 Introduction

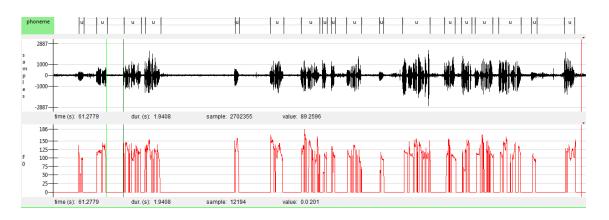
This chapter presents the FVC results using long-term fundamental frequency (LTF0) and tonal F0 as parameters. I first present the methodology I have used to show how the LTF0 (six LTF0 distribution parameters and the percentile-based technique) and tonal F0 were extracted and parameterized. The results are then presented and discussed on the basis of log<sub>10</sub>LRs, C<sub>llr</sub> values and Tippett plots. I conclude this chapter with comments regarding the parameterization of LTF0 in real casework.

# 7.2 Long-term fundamental frequency (LTF0)

Before we go further, the reader is encouraged to go back to §2.18.4 and §2.18.5; in these sections I outlined the rationale behind the use of LTF0 and I reviewed the relevant literature. To remind the reader, the current experiment aims to exploit the F0 features not only from the tonal F0 but also from the distribution of long-term fundamental frequency (LTF0). As such, the six LTF0 distribution parameters that relate to the shape of the F0 distribution were tested. They include: 1) mean; 2) standard deviation (SD); 3) skew; 4) kurtosis; 5) modal F0; and 6) modal density. As previously mentioned (§2.22.4), the first four measures of LTF0 (mean, SD, skew, and kurtosis) are essentially the four moments. The last two measures are the mode (the most often occurring value) of F0 and F0's kernel probability density (area under such F0 values), respectively. That is, the mean is defined as the average energy concentration and SD is how spread-out the energy is (Jongman et al., 2000, p. 1253). Skew is a symmetrical indicator for the energy distribution (ibid.). Kurtosis is an indicator of energy peakedness (ibid.). Modal density refers to the area where the modes are clustered around the mean. When modal density is clustered around the mean with larger standard deviation (SD), the distribution will be larger than a modal density with the same mean but lower standard deviation. If the modal density has a platykurtic distribution (plateau-like distribution), the modes are clustered around the edges (as opposed to the mean) of the distribution (Alderman, 2005, pp. 2833). In the current experiment, the modal F0 and its density values were estimated using the *KernSmooth* library (Wand & Jones, 1994) in R (Ihaka & Gentleman, 1996). The appropriate kernel density bandwidth was selected using the *dpik function* in *KernSmooth* library (Sheather & Jones, 1991; Wand & Jones, 1994).

#### 7.2.1 Data extraction

LTF0 values were extracted from spontaneous speech, collected by means of a fax task, and 53 speakers of Standard Thai were included in this experiment. As such, 53 SS comparisons and 1,378 DS comparisons were possible. As mentioned in §2.22.4, a one-minute long speech was chosen for testing LTF0 in Standard Thai. Before each of the 60-second speech samples was selected for FVC experimentation, it was visually judged to contain roughly the same number of utterances (*u*) separated by pauses as shown in Figure 65. Using EMU in counting a one-minute long speech does not help much when dealing with a lot of silence during the conversation. Visual judgement is therefore the best strategy, given that all utterances are further measured by EMU to ensure they are of approximately one-minute duration. Pauses were defined to be at grammatical junctures or semantically determined (Eisler, 1968, p. 13). Figure 65 shows a screenshot of the EMU used to cut up a one-minute utterance into the smaller chunks (*u*) by pauses.

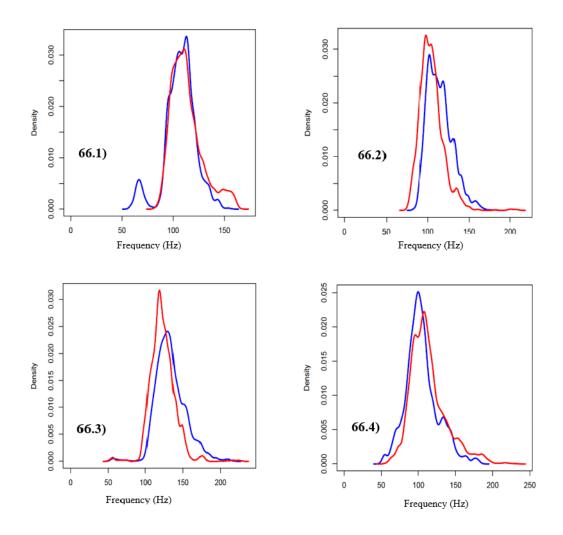


**Figure 65:** Label tier, speech waveforms, and F0 tracking. Each u in the label tier represents an utterance of speech samples used to extract LTF0.

In Figure 65, all cut-up utterances are labeled as u. F0 values of these utterances were then measured every 10 msec with a window length of 0.0075 by EMU and modeled using the LTF0 parameters.

## 7.2.2 Standard Thai LTF0 distribution plots

In order to observe more closely the distribution of Standard Thai LTF0, the charts in Figure 66 show the sample LTF0 distributions from four different informants, elicited during two different sessions. The first session is plotted in blue and the second in red.



**Figure 66:** LTF0 distribution plots extracted from Speakers 1-4. Blue and red curves represent the first and second recording sessions, respectively.

All the LTF0 distributions plotted in Figure 66 are more or less unimodal. For the first session (blue) of Figure 66.1, an additional peak is observed on the left. This bimodal distribution is "found to be very common due to creaky phonation" (Kinoshita & Ishihara, 2010, p. 50). As for the second session (red) of Figure 66.1, there is a clear mode that exists around the F0 of 100 Hz. Despite the similarity in the general shape of the LTF0 distribution between sessions, such additional peak (blue), as shown in Figure 66.1, will shift the values of the six LTF0 parameters significantly and hence produce different values (Kinoshita & Ishihara, 2010, p. 50). Moreover, small changes within each

distribution curve can also be observed, especially those in Figure 66.2 and Figure 66.4. Based on these dynamic variations, we can conclude that the percentile-based technique should also be used to capture such dynamic variations of the LTF0 distribution. Since the 10% percentile measure was found to be the most effective technique for capturing the LTF0 distribution in Kinoshita and Ishihara (2010, p. 53), the 10% percentile technique will also be tested in the current thesis. In sum, not only the six LTF0 parameters of LTF0 will be tested, but also the 10% percentile technique.

# 7.3 Experimental results when using LTF0

This section and the following present the results of using LTF0 and the 10% percentile technique on the basis of Log<sub>10</sub>LR,  $C_{llr}$  and EER, respectively. The Tippett plots will then be presented accordingly. The six LTF0 measures were split into three patterns to see to what extent the *modal F0* + *modal density* parameters improve the experimental results obtained on the basis of the four spectral moments, i.e. to determine which combinations work well and which perform more poorly. Table 26 shows the MVLR results when using as parameters: 1) all six LTF0-based features (the spectral moments, modal F0, and modal density); 2) the spectral moments by themselves; and 3) mode plus modal density.

All LTF0-based features			The four spectral moments			Modal F0 and density					
Calibrated		$\mathbf{C}_{llr}$	EER	Calibrated		$\mathbf{C}_{llr}$	EER	Calibrated		$\mathbf{C}_{llr}$	EER
$LOG_{10}LR$				LOG <sub>10</sub> LR				LOG <sub>10</sub> LR			
SS	DS			SS	DS			SS	DS		
≤ 1.06	≥ −11.01	0.74	28%	≤ 1.18	≥ −12.02	0.75	27%	≤ 1.01	≥ -15.17	0.74	25%

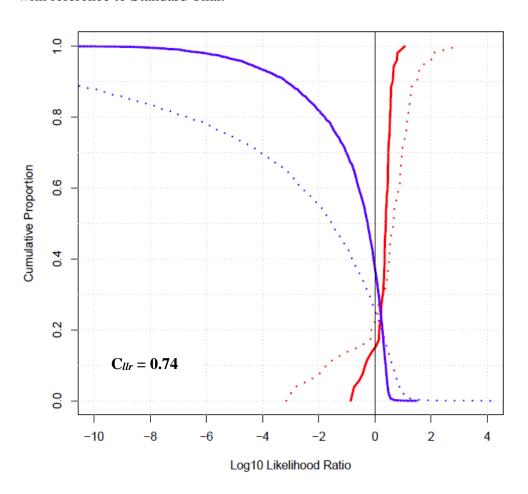
**Table 26:** Log<sub>10</sub>LR, C<sub>llr</sub>, and EER values when mean, SD, skew, kurtosis, modal F0, and modal density were combined according to different patterns.

Table 26 shows that the three sets of parameters performed at the same level since their calibrated Log<sub>10</sub>LRs,  $C_{llr}$  and EER values were comparatively similar. All three sets produced largest consistent-with-fact SSlog<sub>10</sub>LRs  $\leq 2$  and DSlog<sub>10</sub>LRs  $\geq -16$ . The highest  $C_{llr} = 0.75$  was obtained for the four moments, but it was only marginally higher than that of the other patterns:  $C_{llr} = 0.74$  for the six LTF0-based features and for the modal F0 and modal density, respectively. EER values of 25%, 27% and 28% were obtained for the modal F0 and modal density, the four moments and the six LTF0-based features. These comparatively similar results suggest that the six LTF0-based features

were actually so closely correlated that separating them did not produce significantly different results. The Tippett plots and summary tables in §§7.3.1 to 7.3.3 reflect results achieved following the parameterization of 1) all six LTF0-based features; 2) the four spectral moments (mean, standard deviation, skew, kurtosis); and 3) mode and modal density. Each plot and table is followed by discussion.

#### 7.3.1 LTF0: all six features

The Tippett plot in Figure 67 shows system performance subject to the parameterization of all six LTF0 features (mean, SD, skew, kurtosis, modal F0 and modal density). Table 27 summarizes the strongest consistent-with-fact and contrary-to-fact SSLRs and DSLRs with reference to Standard Thai.



**Figure 67:** Tippett plot of LTF0 when all six LTF0 features (mean, SD, skew, kurtosis, modal F0 and modal density) were parameterized.

The (red) curves rising to the right represent the cumulative proportion of the SS (same speaker) comparisons, with the log<sub>10</sub>LRs equal to or less than the value indicated on the x-axis, while the (blue) curves rising to the left represent the cumulative proportion of the DS (different speaker) comparisons, with the log<sub>10</sub>LRs equal to or greater than the value indicated on the x-axis. Dotted lines and solid lines represent the uncalibrated and calibrated SS and DS Log<sub>10</sub>LRs, respectively.

Standard Thai	Consistent-with-fact/ Contrary-to-fact SSLR	Consistent-with-fact/ Contrary-to-fact DSLR		
Uncalibrated	2.94/-3.15	-37.7/4.36		
Calibrated	1.06/-0.86	-11.01/1.49		

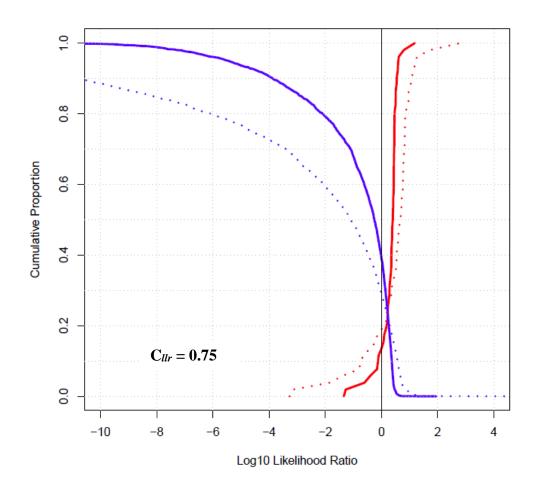
**Table 27:** Largest consistent-with-fact/contrary-to-fact SSLRs and DSLRs of Standard Thai when all six LTF0 features were parameterized.

As shown in Figure 67 and Table 27, the largest uncalibrated consistent-with-fact and contrary-to-fact SSLog<sub>10</sub>LR values (solid and dotted red lines, respectively) were SSlog<sub>10</sub>LR = 2.94 and SSlog<sub>10</sub>LR = -3.15. After calibration, the former was reduced to SSlog<sub>10</sub>LR = 1.06, suggesting "limited" support for the SS hypothesis. The latter was reduced substantially to SSlog<sub>10</sub>LR = -0.86. For DS comparisons, the magnitude of the largest uncalibrated contrary-to-fact DSlog<sub>10</sub>LR = 4.36 was substantially reduced to DSlog<sub>10</sub>LR = 1.49 after calibration. Moreover, 7% of the calibrated DSLRs (solid blue line) were greater than -4, which suggests "very strong" support for the defense hypothesis. The magnitude of the consistent-with-fact DSLRs is much greater than that of the SSLRs.

#### 7.3.2 LTF0: the four spectral moments

The results achieved when using the four spectral moments (mean, standard deviation, skew, kurtosis) as parameters are shown in Figure 68 and Table 28 (overleaf).

Figure 68 and Table 28 reveal that using only mean, standard deviation, skew, and kurtosis as parameters yielded a slightly worse  $C_{llr}$  value of 0.75 as compared to 0.74 in the previous experiment (where all six LTF0 features were used). After calibration, the magnitude of  $Log_{10}LRs$  was reduced in both SS and DS comparisons. For example, the largest uncalibrated consistent-with-fact  $SSLog_{10}LR = 2.79$  shrunk to  $SSLog_{10}LR = 1.18$ , suggesting "limited" support for the SS hypothesis; the largest uncalibrated consistent-with-fact  $DSLog_{10}LR = -32.85$  shrunk to  $SSLog_{10}LR = -12.02$ , suggesting "very strong" support for the defense hypothesis. The largest contrary-to-fact  $SSlog_{10}LR = -3.26$  and  $DSlog_{10}LR = 4.81$  were substantially reduced to  $SSlog_{10}LR = -1.33$  and  $DSlog_{10}LR = 1.94$ .



**Figure 68:** Tippett plot of LTF0 when the four spectral moments (mean, standard deviation, skew, kurtosis) were parameterized.

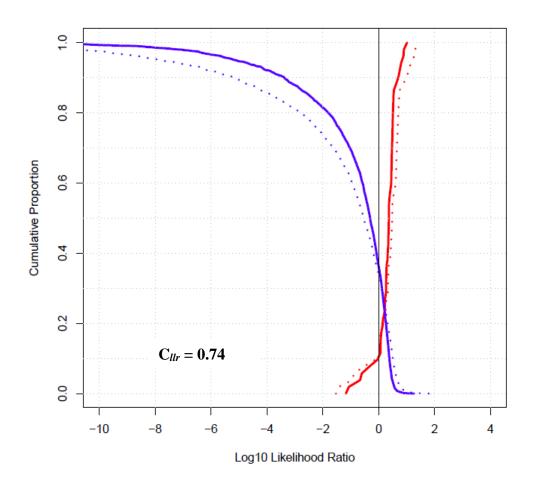
The (red) curves rising to the right represent the cumulative proportion of the SS (same speaker) comparisons, with the  $\log_{10}LRs$  equal to or less than the value indicated on the x-axis, while the (blue) curves rising to the left represent the cumulative proportion of the DS (different speaker) comparisons, with the  $\log_{10}LRs$  equal to or greater than the value indicated on the x-axis. Dotted lines and solid lines represent the uncalibrated and calibrated SS and DS  $\log_{10}LRs$ , respectively.

Standard Thai	Consistent-with-fact/ Contrary-to-fact SSLR	Consistent-with-fact/ Contrary-to-fact DSLR		
Uncalibrated	2.79 / -3.26	-32.85 / 4.81		
Calibrated	1.18 / -1.33	-12.02 / 1.94		

**Table 28:** Largest consistent-with-fact/contrary-to-fact SSLRs and DSLRs of Standard Thai when the four spectral moments (mean, SD, skew, kurtosis) were parameterized.

## 7.3.3 LTF0: model F0 and modal density

The Tippett plot in Figure 69 shows system performance subject to the parameterization of modal F0 and model density. Table 29 summarizes the strongest consistent-with-fact and contrary-to-fact SSLRs and DSLRs with reference to Standard Thai.



**Figure 69:** Tippett plot of LTF0 when modal F0 and model density were parameterized. The (red) curves rising to the right represent the cumulative proportion of the SS (same speaker) comparisons, with the log<sub>10</sub>LRs equal to or less than the value indicated on the x-axis, while the (blue) curves rising to the left represent the cumulative proportion of the DS (different speaker) comparisons, with the log<sub>10</sub>LRs equal to or greater than the value indicated on the x-axis. Dotted lines and solid lines represent the uncalibrated and calibrated SS and DS Log<sub>10</sub>LRs, respectively.

Standard Thai	Consistent-with-fact/	Consistent-with-fact/		
	Contrary-to-fact SSLR	Contrary-to-fact DSLR		
Uncalibrated	1.46 / -1.53	-22.94 /1.83		
Calibrated	1.01 / -1.17	-15.17 / 1.26		

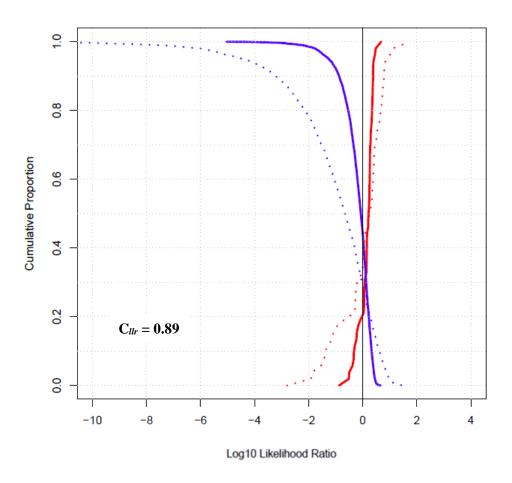
**Table 29:** Largest consistent-with-fact/contrary-to-fact SSLRs and DSLRs of Standard Thai when modal F0 and model density were parameterized.

Figure 69 and Table 29 show that, after calibration, the magnitude of log<sub>10</sub>LR values was not reduced as much as in the case of the two previous experiments. For instance, the uncalibrated contrary-to-fact  $SSlog_{10}LR = -1.53$  was reduced to  $SSlog_{10}LR = -1.17$  and the uncalibrated contrary-to-fact  $DSlog_{10}LR = 1.83$  to  $DSlog_{10}LR = 1.26$ . These relatively well-calibrated LRs reflected in a lower  $C_{llr}^{cal}$  (calibration loss) of 0.078, as compared to  $C_{llr}^{cal} = 0.095$  for the six LTF0 features and  $C_{llr}^{cal} = 0.104$  for the four moments (mean, SD, skew and kurtosis) (see §3.8 on calibration loss). Notably, when modal F0, as opposed to mean F0, was included in the experiment, better results in terms of calibration loss were obtained. This is confirmed by the findings in Hudson et al. (2007). These showed that mode (as opposed to mean) should be a truer indicator for capturing the characteristics of LTF0 distribution as the mean of means has been "pulled up" by higher F0 values (ibid.). Thus, in the current experiment, mode rather than mean is considered as better able to capture LTF0 distribution, which might further result in low calibration loss when modal F0 and modal density are parameterized. However, the C<sub>llr</sub> in this experiment was still high, at 0.74, which might be due in part to a weak magnitude of the consistent-with-fact LRs, resulting in higher penalties. Given the results that I have just described, Standard Thai LTF0 distribution is generally amenable to FVC provided that such LTF0 features are combined with other linguistic-acoustical segments. §7.4 presents the results of using the 10% percentile technique in modeling the dynamic variations of the LTF0 distribution.

## 7.4 Experimental results when using the 10% percentile technique

As shown in §7.2.2, dynamic variations were observed in Standard Thai LTF0 distribution plots, which further prompts us to use the 10% percentile technique for capturing such dynamic characteristics of LTF0 distribution. In the current thesis, the distribution of F0 values extracted from a long stretch of speech were modeled by the binned kernel density function using a bkde command with an appropriate bandwidth set by the dpik function of the R's Kern Smooth library (Sheather & Jones, 1991; Wand & Jones, 1994). A bkde function was run in R (Ihaka & Gentleman, 1996) to calculate the density value and the Hz for each of the 10 percentiles. Figure 70 (overleaf) shows the Tippett plot of LTF0 when its distribution was captured by the 10% percentiles and parameterized in a Hertz scale. The corresponding largest consistent-with-fact/contrary-to-fact SSLRs and DSLRs are shown in Table 30.

Surprisingly, using the percentile-based technique in a Hertz scale did not improve the results, over those of the six LTF0-based features, as was expected from the results shown in the literature. That is, the 10% percentiles technique yielded the highest  $C_{llr} = 0.89$  and highest EER = 30% when compared to those of the previous LTF0 experiments. This



**Figure 70:** Tippett plot of LTF0 when its distribution was captured by the 10% percentiles and parameterized in a Hertz scale.

The (red) curves rising to the right represent the cumulative proportion of the SS (same speaker) comparisons, with the log<sub>10</sub>LRs equal to or less than the value indicated on the x-axis, while the (blue) curves rising to the left represent the cumulative proportion of the DS (different speaker) comparisons, with the log<sub>10</sub>LRs equal to or greater than the value indicated on the x-axis. Dotted lines and solid lines represent the uncalibrated and calibrated SS and DS Log<sub>10</sub>LRs, respectively.

Standard Thai	Consistent-with-fact/ Contrary-to-fact SSLR	Consistent-with-fact/ Contrary-to-fact DSLR		
Uncalibrated	-2.80 / 1.75	-16.39 / 1.65		
Calibrated	0.67 / -0.87	-5.03 / 0.65		

**Table 30:** Largest consistent-with-fact/contrary-to-fact SSLRs and DSLRs of Standard Thai LTF0 when their distribution was captured by the 10% percentiles and parameterized in a Hertz scale.

contradicts what we found in the literature, where percentile-based techniques were proved to better capture the distribution and to provide greater detail (cf. Kinoshita & Ishihara, 2010). Additionally, for the experimental setting discussed in this section, the largest consistent-with-fact  $SSlog_{10}LR$  obtained was  $SSlog_{10}LR = 0.67$ , suggesting only "limited support" for the SS hypothesis. Moreover, only ca. 7% had  $DSlog_{10}LRs \leq -4$ , suggesting "very strong" support for the defense hypothesis. Thus, the results of LTF0 in Standard Thai were worse than those reported for Japanese, where the percentile-based techniques outperformed the base-line 6 LTF0 measures (ibid.).

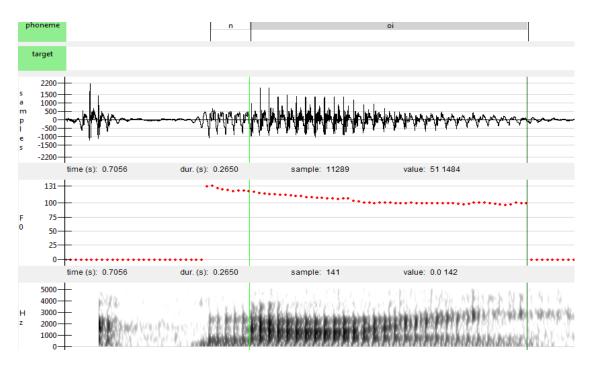
In §7.5, we look at the results achieved by exploiting tonal F0 for [ɔi] - [nɔi L] and [ai] - [mai HL].

# 7.5 Tonal F0 of [ɔi] - [nɔi L] and [ai] - [mai HL]

As mentioned in §4.4, where the linear, quadratic and cubic polynomials fitted to the falling F0 contours of [ai] - [tehai HL] were experimented with using MVLR, we achieved the best results when quadratic polynomials were parameterized. That is, all SS comparisons were correctly discriminated with the lowest  $C_{llr} = 0.39$ . For DS comparisons, ca. 28% were wrongly discriminated as coming from the same speakers and ca. 20% had  $DSlog_{10}LRs \le -4$ , suggesting "very strong" support for the defense hypothesis. It was also concluded from these findings that the quadratic polynomial sufficiently approximated the falling F0 contour of [ai] - [tehai HL], while the cubic polynomial might be overfitted. This being the case, it is prudent to conduct a further experiment on tonal F0 extracted from the diphthongs [5i] - [n5i L] and [ai] - [mai HL]. Before we go further, the reader is encouraged to refer to §6.2, where the rationale for choosing these particular segments is given. Briefly, [ji] - [noi L] has been chosen as this particular diphthong potentially provides more acoustical vowel space for informants to exhibit difference in articulation due to the two different vocalic targets that are involved, i.e. /ɔ/, which is a low-mid back vowel, and /i/, a high front vowel. Moreover, more duration is guaranteed with [5i] - [n5i L], where the final words tend to be stressed in this sentence-final position (Abramson, 1962; Naksakul, 1998). This assertion is supported by the fact that the duration of [n] - [noi L] was at least 117.40 msec and the maximum was 459.91 msec for this experiment. As for [ai] - [mai HL], although it has a high-falling contour similar to that of [ai] - [tehai HL], as previously shown in §4.3, it is prudent to test if the same falling contour, preceded by a different consonant from this commonly used word [ai] - [mai HL] 'no', performs better or worse than its counterpart [ai] - [tehai HL] 'yes'. The reader is also encouraged to go back to §6.4.1 and §6.4.2 for information on how the starting and end points of these target segments were located. In Emu labeler, a minimum F0 was pre-set to 50 Hz before all F0 values were extracted by an emu.query command in EMU-R (Cassidy, 1999), as shown in the following sections.

#### **7.5.1 F0 tracking of [5i] - [n5i L]**

Figure 71 shows a label tier, a waveform, an overlaid F0 tracking and a spectrogram of [ɔi] - [nɔi L].



**Figure 71:** Label tier (top), waveform and overlaid F0 tracking (middle), and spectrogram (bottom) for the target segment [ɔi] - [nɔi L]. The highlighted section in the label tier shows the target segment.

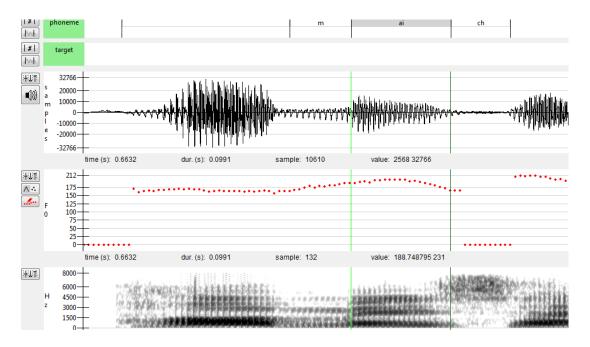
Note: [ɔi] is labeled as [oi].

Although low tones can be grouped with static tones (Abramson, 1962; Naksakul, 1998), it is evident from Figure 71 that a low tone shows a slightly falling contour along its entire time-course (from its onset at ca. 131 Hz to its offset at ca. 100 Hz). Since the figure is maximized for the sake of a visual inspection of F0 tracking, the x-axis, which indicates the time in msec, could not be included.

The following EMU labeler shows a display of the high-falling tones of [ai] - [mai HL].

## 7.5.2 F0 tracking of [ai] - [mai HL]

Figure 72 shows a label tier, a waveform, an overlaid F0 tracking and a spectrogram of [ai] - [mai HL]. It can be seen that the EMU speech database system is able to pick up the high-falling tone (red dots under the highlighted segment [ai]) relatively well.

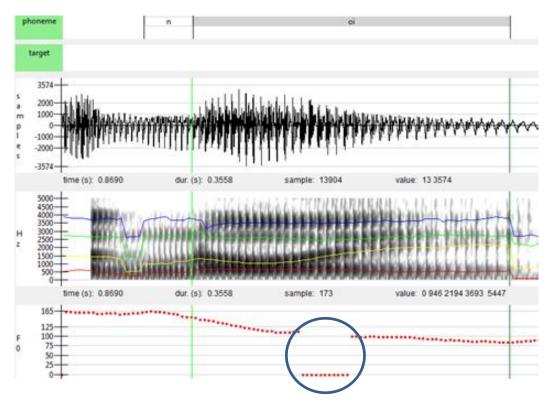


**Figure 72:** Label tier (top), waveform and overlaid F0 tracking (middle), and spectrogram (bottom) for the target segment [ai]. The highlighted section in the label tier shows the target segment.

However, as shown in Figure 73 (overleaf) for [ɔi] - [nɔi L], there are some instances where the tonal F0 values were not well tracked. A low tone of the target segment [ɔi] - [nɔi L] was not well tracked by the EMU speech database (as marked by the navy blue circle): some red dots in the middle of [ɔi] - [nɔi L]'s entire timecourse dropped down to 0 Hz. Therefore, a manual correction was performed, as shown in Figure 74.

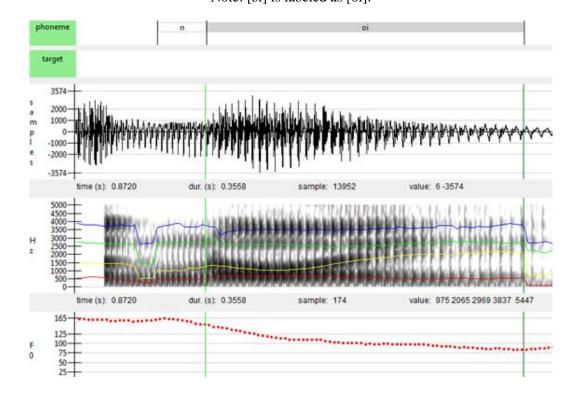
The last panel of Figure 74 shows F0 tracking after a manual correction was performed for [ɔi] - [nɔi L]. F0 tracking errors were manually corrected to be in approximately the same frequency range as that of the points preceding and following the F0 drops.

In §§7.5.3 and 7.5.4, the F0 values are plotted in a normalized duration (100 msec) so that the within-speaker and between-speaker variation of F0 contours for [5i] - [n5i L] and [ai] - [mai HL] can be more clearly observed.



**Figure 73:** Label tier (top), waveform and overlaid formant tracking (middle), and F0 tracking (bottom) for the target segment [ɔi] - [nɔi L]. The highlighted section in the label tier shows the target segment. There are some F0 tracking errors in the middle of [ɔi].

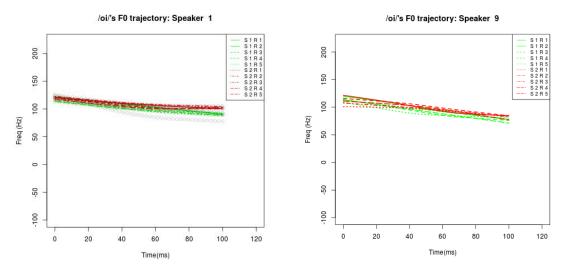
Note: [ɔi] is labeled as [oi].



**Figure 74:** Label tier (top), waveform and overlaid formant tracking (middle), and F0 tracking after manual correction (bottom) for the target segment [ɔi] - [nɔi L]. The highlighted section in the label tier shows the target segment.

Note: [ɔi] is labeled as [oi].

## 7.5.3 F0 contours of [5i] - [n5i L]



**Figure 75:** F0 contours of the diphthong [5i] - [n5i L] plotted against normalized duration (100 msec) for Speakers 1 and 9. F0 contours (2 sessions x 5 repeats) are represented in green for session 1 and red for session 2.

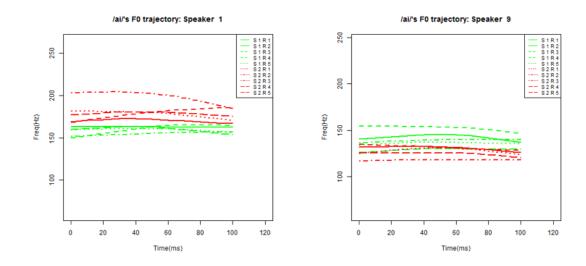
Note: [ɔi] is labeled as [oi].

Figure 75 reveals that the F0 contours of [ɔi] - [nɔi L] for both Speakers 1 and 9 show less within-speaker variation (as the green vs yellow lines overlap). However, Speaker 9 shows a marginal session-to-session difference during the first 20 msec. As for between-speaker variation, Speaker 1 exhibits his F0 contours within a marginally higher frequency range of ca. 90 Hz to ca. 125 Hz than Speaker 9 (between ca. 80 Hz and ca. 125 Hz). F0 contours of the diphthong [ɔi] - [nɔi L] plotted against normalized duration (100 msec) for all 30 speakers can be found in Appendix F.

The sample plots of the extracted F0 values of the diphthong [ai] - [mai HL] are shown in §7.5.4.

## **7.5.4 F0 contours of [ai] - [mai L]**

In Figure 76 (overleaf), much between-session variation is observed for Speaker 1, as all repeats of session 1 (green) are clearly separated from those of session 2 (red). As for between-speaker variation, Speaker 1 exhibits his contours within a higher frequency (ca. 150-200 Hz) than Speaker 9 (ca. 120-150 Hz). F0 contours of the diphthong [ai] - [mai HL] plotted against normalized duration (100 msec) for all 30 speakers can be found in Appendix H.

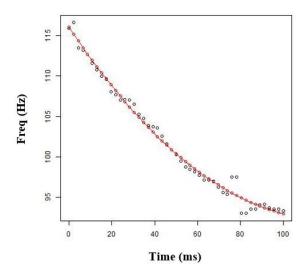


**Figure 76:** F0 contours of the diphthong [ai] - [mai HL] plotted against normalized duration (100 msec) for Speakers 1 and 9. F0 contours (2 sessions x 5 repeats) are represented in green for session 1 and red for session 2.

The sample F0 contours of the diphthongs [ɔi] - [nɔi L] and [ai] - [mai HL], plotted together with a quadratic polynomial curve fitting, are shown in §§7.5.5 and 7.5.6.

#### 7.5.5 Polynomial curve fitting of tonal F0 for [ɔi] - [nɔi L]

Figure 77 shows the F0 contour typical of the diphthong [ $\mathfrak{si}$ ] - [ $\mathfrak{nsi}$  L], which (in linguistic terms) has low tone. Phonetically, this [ $\mathfrak{si}$ ] - [ $\mathfrak{nsi}$  L] is not realized as such. Rather, a slightly falling contour is evident. Figure 77 is thus an example of a quadratic polynomial curve fitting in which the F0 contour of [ $\mathfrak{si}$ ] - [ $\mathfrak{nsi}$  L] is plotted together with its polynomial fitting of  $00.004838x^2 + (-1.476583)x + 223.981113$ . As discussed in §4.3,



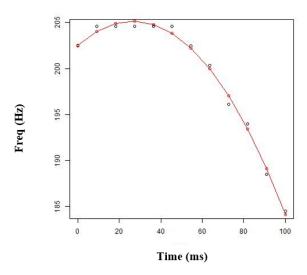
**Figure 77:** F0 values (black dots) of the diphthong [5i] - [n5i L] plotted together with its quadratic polynomial curve fitting (dotted red line) in normalized duration (100 msec).

the best results were obtained when a quadratic polynomial was used to approximate the falling F0 contour of [ai] - [tchai HL], while a cubic polynomial might be overfitted. This being the case, a quadratic polynomial was judged to sufficiently approximate a low tone of [bi] - [nbi L]. F0 values of the diphthong [bi] - [nbi L] for all 30 speakers, plotted together with its quadratic polynomial curve fitting, can be found in Appendix G.

A typical contour for a high-falling tone, as in [ai] - [mai HL], is shown in §7.5.6.

## 7.5.6 Polynomial curve fitting of tonal F0 for [ai] - [mai HL]

Figure 78 shows a high-falling contour typical of the diphthong [ai] - [mai HL].



**Figure 78:** F0 values (black dots) of the diphthong [ai] - [mai HL] plotted together with its quadratic polynomial curve fitting (dotted red line) in normalized duration (100 msec).

This is an example of a quadratic polynomial curve fitting in which the high-falling trajectory of [ai] - [mai HL] is plotted together with its polynomial fitting of 0.000233x<sup>2</sup> + (-0.035758)x + 163.815018. F0 values of the diphthong [ai] - [mai L] for all 30 speakers, plotted together with its quadratic polynomial curve fitting, can be found in Appendix I. §7.7 discusses the results of using this quadratic polynomial fitted to the F0 contours of the diphthongs [ɔi] - [nɔi L] and [ai] - [mai HL] on the basis of Log<sub>10</sub>LR, C<sub>llr</sub> and EER, respectively. A Tippett plot will then be presented, and discussion will follow.

#### 7.6 Informants

All F0 contours of [ɔi] - [nɔi L] and [ai] - [mai HL] were extracted from 30 speakers. As such, 30 SS comparisons and 435 DS comparisons were possible.

## 7.7 Experimental results when using tonal F0

Table 31 shows the results of calibrated  $Log_{10}LR$ ,  $C_{llr}$  and EER values when a quadratic polynomial was fitted to the F0 contours of [5i] - [n5i L] and [ai] - [mai HL].

	Quadratic				
Tonal F0	Cal	ibrated	$\mathbf{C}_{llr}$	EER	
	LO	G <sub>10</sub> LR			
	SS	DS			
[ɔi] - [nɔi L]	≤ 1.88	≥ -53.98	0.52	18	
[ai] - [mai HL]	≤ 0.45	≥ -10.64	0.93	33	

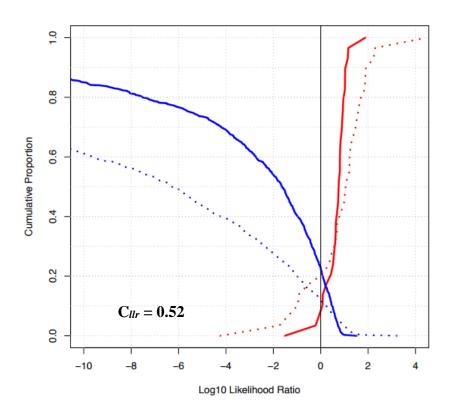
**Table 31:** Calibrated Log<sub>10</sub>LR,  $C_{llr}$ , and EER values when a quadratic polynomial was fitted to the F0 contours of [5i] - [n5i L] and [ai] - [mai HL], respectively.

Table 31 illustrates that much better  $C_{llr} = 0.52$  and EER = 18 were obtained when the tonal F0 values of the diphthong [ɔi] - [nɔi L] were parameterized (in the case of [ai] - [mai HL], the values were  $C_{llr} = 0.93$  and EER = 33). For SS comparisons, all calibrated consistent-with-fate SSLRs only "moderately" support the same-speaker hypothesis, with  $Log_{10}LRs \le 2$  for both [ɔi] - [nɔi L] and [ai] - [mai HL]. With respect to the DSLRs, there was "very strong" support for the DS hypothesis:  $log_{10}LR \le -4$  was obtained for the F0 contours in both diphthongs.

In §§7.7.1 and 7.7.2, I take a closer look at the magnitude of LR values using Tippett plots. The Tippett plots shown reflect the use of a quadratic polynomial to model the F0 contour of the diphthongs [5i] - [n5i L] and [ai] - [mai HL].

#### 7.7.1 The Tippett plot of [ɔi] - [nɔi L]

Figure 79 shows the Tippett plot of [5i] - [nɔi L] obtained when its tonal F0 contours were parameterized by a quadratic polynomial. It can be seen that the majority of SS comparisons (ca. 98%) had calibrated consistent-with-fact  $SSlog_{10}LRs \le 1$  and the largest  $SSlog_{10}LR$  obtained was  $SSlog_{10}LR = 1.88$ , suggesting only "limited" support for the SS hypothesis. After calibration, the contrary-to-fact  $SSlog_{10}LR = -4.23$  was reduced to  $SSlog_{10}LR = -1.51$ . Approximately 30% of DS comparisons gave calibrated consistent-with-fact  $DSlog_{10}LRs \le -4$ , suggesting "very strong" support for the defense hypothesis, whereas ca. 20% of DS comparisons were wrongly discriminated. After calibration, the contrary-to-fact  $DSlog_{10}LR = 2.73$  was reduced to  $DSlog_{10}LR = 1.51$ .

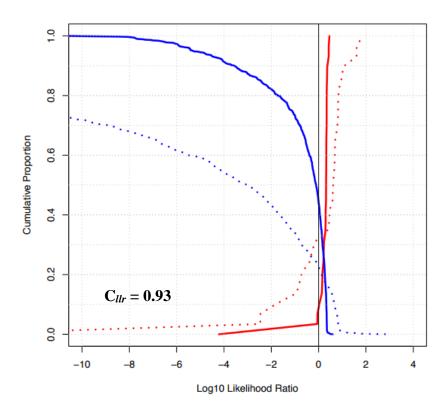


**Figure 79:** Tippett plot of [5i] - [n5i L] when its tonal F0 contours were parameterized by a quadratic polynomial.

The (red) curves rising to the right represent the cumulative proportion of the SS (same speaker) comparisons, with the log<sub>10</sub>LRs equal to or less than the value indicated on the x-axis, while the (blue) curves rising to the left represent the cumulative proportion of the DS (different speaker) comparisons, with the log<sub>10</sub>LRs equal to or greater than the value indicated on the x-axis. Dotted lines and solid lines represent the uncalibrated and calibrated SS and DS Log<sub>10</sub>LRs, respectively.

#### 7.7.2 The Tippett plot of [ai] - [mai HL]

Figure 80 shows the Tippett plot of [ai] - [mai HL] when its F0 contours, fitted by a quadratic polynomial, were parameterized. As can be seen in Figure 80, the F0 contours extracted from the diphthong [ai] - [mai HL] did not perform well because the  $C_{llr} = 0.93$  and EER = 33% that were obtained were considered relatively high. These high values may have resulted from the largest contrary-to-fact  $SSlog_{10}LR = -15.30$  and  $DSlog_{10}LR = 2.98$ . Both were substantially reduced, after calibration, to  $SSlog_{10}LR = -4.21$  and  $DSlog_{10}LR = 0.59$ . For DS comparisons, only 10% had calibrated consistent-with-fact  $Log_{10}LRs \le -4$ , suggesting "very strong" support for the defense hypothesis. Having said that, if the  $log_{10}LR = 0$  is set to the threshold, 50% of DS comparisons were wrongly discriminated as coming from the same speaker. Based on the above findings, when the tonal F0 values of [5i] - [n5i L] and [ai] - [mai HL] were used as parameters, the strength



**Figure 80:** Tippett plot of [ai] - [mai HL] when its tonal F0 contours were parameterized by a quadratic polynomial.

The (red) curves rising to the right represent the cumulative proportion of the SS (same speaker) comparisons, with the log<sub>10</sub>LRs equal to or less than the value indicated on the x-axis, while the (blue) curves rising to the left represent the cumulative proportion of the DS (different speaker) comparisons, with the log<sub>10</sub>LRs equal to or greater than the value indicated on the x-axis. Dotted lines and solid lines represent the uncalibrated and calibrated SS and DS Log<sub>10</sub>LRs, respectively.

of voice evidence was not that great: the largest consistent-with-fact SSLog<sub>10</sub>LR = 1.88 and SSLog<sub>10</sub>LR = 0.45 were obtained for [5i] - [noi L] and [ai] - [mai HL], respectively. Moreover, the  $C_{llr}$  = 0.93 was very high for [ai] as compared to  $C_{llr}$  = 0.52 for [5i] - [noi L]. In comparison to the results of the F0 contour obtained for [ai]- [ $te^h$ ai HL] and reported in §4.3 (where all SS comparisons of [ai] - [ $te^h$ ai HL] were correctly discriminated with the lowest  $C_{llr}$  = 0.39), the tonal F0 results of [ai] - [mai HL] revealed much worse calibrated Log<sub>10</sub>LR,  $C_{llr}$ , and EER values. The possible underlying reasons accounting for this bad FVC performance of [ai] - [mai HL] are twofold. First, a smaller number of speakers (30 speakers) was tested for [ai] - [mai HL] as opposed to 54 speakers for [ai] - [ $te^h$ ai HL]. Second, [ai] - [ $te^h$ ai HL] did not show much dynamic F0 contour (cf. Appendix H) typical for a high-falling tone, which might be the result of [ai] - [ $te^h$ ai HL] undergoing a vowel reduction (Abramson, 1962, p. 76).

# 7.8 Summary

In this chapter, I first illustrated how the LTF0 distribution of Standard Thai was parameterized using various parameters. They are 1) the six LTF0-based parameters (mean, SD, skew, kurtosis, modal F0, modal density) with three experimental settings (all six parameters, the four spectral moments, and modal F0 + modal density) and 2) the 10% percentile parameters measured in a Hertz scale. The results show that none of the six LTF0-based parameters performed well because of a  $C_{llr} = 0.74$ , with the largest calibrated consistent-with-fact SSLog<sub>10</sub>LR = 1.06 suggesting only "limited" support for the SS hypothesis. The findings of LTF0 of Standard Thai were then discussed in relation to those of Japanese. When the LTF0 distribution was captured by the 10% percentile technique in a Hertz scale, even higher  $C_{llr} = 0.89$  and smaller calibrated consistent-withfact  $SSLog_{10}LR = 0.49$  were obtained. This is surprising since the available literature had led us to expect that the use of the 10% percentile technique would improve the results. Subsequently, I presented the results using tonal F0 extracted from the diphthongs [5i] -[noi L] and [ai] - [mai HL]. Based on all these findings, tonal F0 would be more interesting than LTF0 in Standard Thai as the deriving  $C_{llr}$  obtained for [5i] - [n5i L] was lowest at 0.52 with the largest calibrated consistent-with-fact  $SSLog_{10}LR = 1.88$ . However, both LTF0 and tonal F0 in Standard Thai are still generally amenable to FVC. A better performance of LTF0 and tonal F0 might be achieved if they are combined with other linguistic-phonetic segments to give better overall strength of voice evidence.

# **Chapter 8**

## Conclusions and recommendations for future research

#### 8.1 Introduction

In this chapter I first answer the two research questions addressed in this thesis. Then, all experimental findings are summarized. Finally, ideas for future research based on the findings of the current work will be proposed.

## 8.2 Answers to the research questions

As mentioned in §1.8, this research addressed two specific questions.

The first question was to examine to what extent the acoustical parameters extracted from the linguistic-phonetic segments of /s/, /te<sup>h</sup>/, /n/, /m/ perform in Standard Thai FVC. It was proposed to model the spectrum extracted from the consonants /s/, /te<sup>h</sup>/, /n/, /m/ in two different ways: by means of the so-called spectral moments, on the one hand, and by means of the coefficients of the discrete cosine transform (DCTs), on the other. The use of two different parameterization techniques would allow us to compare what parameterization technique performs better.

The findings reveal that the derived FVC values for the spectral moments /s/ show only "limited" support at best for SS comparisons with the highest  $C_{llr} = 0.92$ . Additionally, only ca. 1% of calibrated DSLR was less than -4, suggesting "very strong" support for the defense hypothesis. This means that the magnitude of such derived LRs was relatively weak. The DCTs for /s/ were therefore further experimented on to see if DCT parameters perform better than the spectral moments. The results show that the DCTs outperformed spectral moments on the basis of  $C_{llr}$  and EER values with  $SSLog_{10}LR \le 2$ . Summarizing the experiments, the lowest  $C_{llr} = 0.47$  was obtained for /n/ - [na: HL] and /m/ - [mai HL] while the affricate /te<sup>h</sup>/ - [te<sup>h</sup>ai HL] performed marginally worse than the nasals ( $C_{llr} = 0.54$ ); the fricative /s/ - [sa:m LH] performed the worst ( $C_{llr} = 0.70$ ). The above findings confirm that the nasals /m, n/ contain more speaker-specificity than the fricative /s/ and affricate /te<sup>h</sup>/, as reflected in a greater magnitude of calibrated consistent-with-fact SSLR

(Log<sub>10</sub>LR = 1.99 for /n/ - [na: HL] and Log<sub>10</sub>LR = 1.87 for /m/ - [mai HL]) and lower  $C_{llr}$  values ( $C_{llr} = 0.47$  for /m/ - [mai HL] and  $C_{llr} = 0.47$  for /n/ - [na: HL]).

The first pilot study on the formant trajectories of the Standard Thai diphthongs [i:aw], [u:a] and [u:a] showed that they can be ranked in terms of  $C_{llr}$  values from low to high as [u:a] ( $C_{llr}$ = 0.02), [i:aw] ( $C_{llr}$ = 0.03), [u:a] ( $C_{llr}$ = 0.04), respectively, when their cubic polynomials fitted to [F2, F3, F4] were parameterized. Interestingly, all SS comparisons of [i:aw] were correctly discriminated and the best calibrated consistent-with-fact SSLog<sub>10</sub>LRs obtained for [i:aw] were SSLog<sub>10</sub>LRs  $\leq$  3, suggesting "moderately strong" support for the same-speaker hypothesis. As for [u:a] and [u:a], the best calibrated consistent-with-fact SSLog<sub>10</sub>LRs obtained for both [u:a] and [u:a] were SSLog<sub>10</sub>LRs  $\leq$  2, suggesting "moderate" support for the same-speaker hypothesis. However, at least ca. 90% of [u:a] and [u:a] had DSlog<sub>10</sub>LRs  $\leq$  -4 (as compared to ca. 70% for [i:aw]), suggesting "very strong" evidence in support of the defense hypothesis.

The second pilot study on the F2 trajectories of [o:i] - [do:i M] and [ə:i] - [khə:i M] found that [o:i] - [do:i M] performed best with a  $C_{llr}$  of 0.64 when its cubic polynomials (duration was not included) were parameterized. In contrast, [ə:i] - [khə:i M] performed best when its cubic polynomials plus duration were parameterized as lower  $C_{llr} = 0.67$  was obtained (as opposed to  $C_{llr} = 0.78$  when duration was not included). The best calibrated consistent-with-fact  $SSLog_{10}LRs$  obtained were  $SSLog_{10}LRs \le 2$  for both [o:i] and [ə:i], suggesting "moderate" support for the SS hypothesis. For DS comparisons, only ca. 2% of [o:i] and ca. 5% of [ə:i] had  $DSlog_{10}LRs \le -4$ , suggesting "very strong" support for the defense hypothesis.

The third pilot study on the tonal F0 contours of [ai] - [tehai HL] and [u:a] - [ru:am HL] fitted by (linear, quadratic and cubic) polynomials, the best  $C_{llr} = 0.39$  was obtained when F0 contours of [ai] - [tehai HL] were parameterized with quadratic polynomials. Moreover, all SS comparisons were correctly discriminated for [ai] - [tehai HL]. Regarding [u:a] - [ru:am HL], the best  $C_{llr} = 0.51$ , which is higher than that of [ai] - [tehai HL], was obtained when linear polynomials were parameterized. However, for DS comparisons, a higher proportion (30%) of [u:a] - [ru:am HL] had calibrated consistent-with-fact DSlog<sub>10</sub>LRs  $\leq -4$ , suggesting "very strong" support for the defense hypothesis, as compared to 20% in the case of [ai] - [tehai HL].

Findings related to the formant trajectories of the diphthongs [5i] - [noi L] and [ai] - [mai HL], when the F2 trajectory and F1-F3 trajectories, fitted by cubic polynomials, were parameterized in turn, showed that the results of F1-F3 trajectories outperformed those of the F2 trajectory on the basis of calibrated Log<sub>10</sub>LR,  $C_{llr}$  and EER values. That is, the best  $C_{llr}$  and EER values of [5i] - [noi L] were marginally lower than those of [ai] - [mai HL] ( $C_{llr}$ = 0.42 and EER = 10% vs  $C_{llr}$ = 0.49 and EER = 18%) when F1-F3 trajectories were parameterized. This is not surprising as F1 and F3 were expected to add more individualizing information (compared to exclusive use of the F2 trajectory) in the FVC experiment. However, the magnitude of the consistent-with-fact SSLRs was relatively weak. That is, the strongest consistent-with-fact SSLog<sub>10</sub>LR for [5i] - [noi L] and for [ai] - [mai HL] obtained were SSLog<sub>10</sub>LR = 1.91 and SSLog<sub>10</sub>LR = 1.84, respectively, both of which provided only "moderate" support for the SS hypothesis. In contrast, consistent-with-fact DSLRs obtained were  $\log_{10}$ LR  $\leq$  -4 for both [5i] - [noi L] (ca. 6% of DS comparisons) and [ai] - [mai HL] (ca. 10% DS comparisons), suggesting "very strong" support for the defense hypothesis.

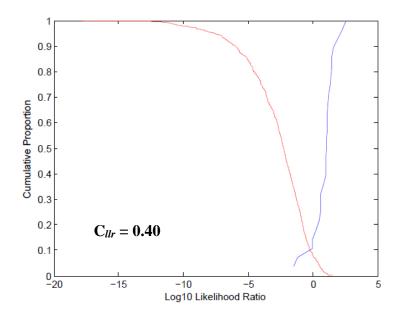
With regards to LTF0 and tonal F0, the results showed that none of the six LTF0 parameters performed well because the smallest  $C_{llr}$  values = 0.74 with the largest calibrated consistent-with-fact SSLog<sub>10</sub>LR = 1.06 (suggesting only "limited" support for the SS hypothesis) were obtained. When the LTF0 distribution was captured by the 10% percentile technique in a Hertz scale, an even higher  $C_{llr} = 0.89$  and smaller calibrated consistent-with-fact  $SSLog_{10}LR = 0.49$  were obtained. This is surprising, because the use of the 10% percentile technique did not improve these results, as the literature had led us to expect. The underlying reason for this is unclear. In contrast, the use of tonal F0 extracted from the diphthongs [5i] - [n5i L] and [6i] - [mai HL] yielded better results (as compared to LTF0) because the derived  $C_{llr}$  obtained for [5i] - [n5i L] was lowest at 0.52 with the largest calibrated consistent-with-fact  $SSLog_{10}LR = 1.88$ , suggesting "moderate" support for the SS hypothesis. For DS comparisons, ca. 30% of [5i] - [n5i L] had log<sub>10</sub>LR  $\leq$  -4, suggesting "very strong" support for the defense hypothesis, as opposed to ca. 10% of [ai] - [mai HL]. Although we might have expected a falling contour of [ai] - [mai HL] to outperform a level tone of [ɔi] - [nɔi L], a better performance of [ɔi] - [nɔi L] was obtained, which might be due to the fact that more duration (between 117.40 msec and a 459.91 msec) was involved. As such, more individualizing information might be picked up from such a longer duration. Based on the results gathered, we can conclude that Standard Thai acoustical parameters are generally amenable to Forensic Voice Comparison.

The performance of all parameters is summarized below on the basis of  $C_{llr}$  and LR values.

- 1. The fricative /s/ performed better when its DCTs were parameterized as opposed to the spectral moments.
- 2. The nasals /n, m/ performed better than the fricative /s/ and the affricate /tch/ when their DCTs were parameterized.
- 3. All F1-F3 trajectories outperformed the F2 trajectory for both diphthongs [oi] and [ai].
- 4. The F1-F3 trajectories of the diphthong [5i] outperformed those of [ai].
- 5. Tonal F0 performed better than LTF0.

I will now answer the second research question addressed in this thesis.

The second question was: through an interpretation of the fusion results, and adding to research findings pursued in the current work, how can such tested linguistic-phonetic segments be profitably combined? As previously mentioned, the number of speakers tested was different from one experiment to the next and since the numbers and order of speakers being fused and calibrated need to be the same, the segments of only a few parameters, namely the DCTs of /n/ - [na: HL] and /n/ - [noi L], were fused. Before we go any further, the reader is encouraged to go back to §3.9, which details the basic concept of fusion. The reader will recall that LRs were obtained in a leave-one-out manner, i.e. the test set of speech samples were excluded from the reference set. As such, the weights were calculated from the LR-like scores obtained in all the available data, except for a test pair of speakers used to monotonically shift and scale the scores in the test set to derive a true LR; this was achieved using the logistic regression technique by FoCal Toolkit (Brümmer, 2007). Figure 81 shows the Tippett plot when the best performing parameter (15 Hertz-scaled DCTs) of /n/ - [na: HL] and the best performing parameter (20 Hertz-scaled DCTs) of /n/ - [noi L] were fused.



**Figure 81:** Tippett plot for the fused and calibrated LRs for /n/ - [na: HL] and /n/ - [noi L]. The (blue) curve rising to the right represents the cumulative proportion of the SS (same speaker) comparisons, with the log<sub>10</sub>LRs equal to or less than the value indicated on the x-axis, while the (red) curve rising to the left represents the cumulative proportion of the DS (different speaker) comparisons, with the log<sub>10</sub>LRs equal to or greater than the value indicated on the x-axis.

Figure 81 shows that the results were significantly improved in terms of  $C_{llr}$ , EER and LR values. That is, lower  $C_{llr} = 0.40$  and lower EER = 10% values were obtained when the best performing parameter (15 Hertz-scaled DCTs) of /n/ - [na: HL] and the best performing parameter (20 Hertz-scaled DCTs) of /n/ - [noi L] were fused (as compared to  $C_{llr} = 0.47$  and EER = 15% of the best-performing parameter of /n/ - [na: HL] alone). Moreover, the largest calibrated consistent-with-fact SSLog<sub>10</sub>LR obtained was SSLog<sub>10</sub>LR = 2.53, suggesting "moderately strong" support for the SS hypothesis. Regarding DS comparisons, ca. 25% yielded  $Log_{10}LR \le -4$ , suggesting "very strong" support for the DS hypothesis. The greatest consistent-with-fact DSlog<sub>10</sub>LR obtained was DSlog<sub>10</sub>LR = -17.72. In general, it was found that fusion of the two target segments of /n/ - [na: HL] and /n/ - [noi L] significantly improved the results in terms of  $C_{llr}$ , EER and LR values. This suggests that the LRs from these two /n/s have a "sufficiently small correlation coefficient" (Franco-Pedroso et al., 2012), otherwise they would not contribute to better  $C_{llr}$ , EER and LR values. Moreover, the DCTs of /n/ - [na: HL] and /n/ - [noi L] may indeed carry complementary individualizing information.

## 8.3 Future research

There are many further steps that might be taken with regard to the current work. We could possibly fuse all acoustical parameters obtained from the same speakers in the current thesis to see the overall results. Other possible steps are listed below.

#### 8.3.1 Speech corpus

Since the current work only dealt with speech samples from male informants, those of females should be included in future in order to see how such speech samples perform in Standard Thai FVC. Likewise, the different dialects of the Northern, Northeastern and Southern Thai people should be trialed. As has been discussed in §§2.11-2.13, apart from Standard Thai, which is used as an official language to communicate nationwide, other major dialects are spoken in different areas of Thailand. Such major dialects are traditionally and geographically grouped as 1) the Lanna Thai or Northern Thai dialect; 2) the Isaan or Northeastern Thai dialect; and 3) the Pak Tai or Southern Thai dialect. As pointed out by Rose (2002, p. 333), little FVC research has been conducted using speech samples from speakers who are bilingual (in the broadest sense of speaking either two languages or two dialects). More insight into cross-linguistic FVC is needed for us to see to what extent they preserve the linguistic differences across such differing dialectal speech samples, which Rose (2002, p. 333) further suggests will depend on how well a speaker can command different dialects or languages. Since the Northern, Northeastern, Southern, and Standard Thai dialects "differ more in phonology and vocabulary than they do in grammar" (Tienmee, 1992, p. 229), it would be prudent to test to what extent these differences have been preserved cross-linguistically and which phonological features are distinctive to which dialects; such information might be of use forensically, at least for approximating a suspect's linguistic community. In addition, there are still many FVC research opportunities awaiting to be conducted in Standard Thai itself, for example, with speech samples obtained from mismatched conversational situations, speaking styles and emotional states. As speech segments do not occur in isolation but always in continuous/connected speech, coarticulation is another interesting area in Standard Thai FVC research. This is because neighboring segments may exert a degree of articulatory adaptation upon one another that might be unique to some individuals.

#### 8.3.2 Parameters

Apart from the linguistic-phonetic parameters explored in this thesis, more parameters should be tested. Possible interesting acoustical features to explore are shown below.

#### 8.3.2.1 Voice onset time (VOT) of a stop /kh/

Based on an acoustical analysis of voice onset time (VOT) of Standard Thai stops (threeway contrast), Gandour (1985) found that the distributions of VOT (minimum, maximum, range, mean and standard deviation) exhibited non-overlapping distribution patterns between speakers (speech samples were recorded from five speakers in a single session). Such large between-speaker variation (although from only a small number of speakers) has shown a potential for further FVC analysis. Based on the findings of Gandour (1985), VOT of a stop /kh/ embedded in the words [khrap H] and [kha? L] is appealing for future research, to test how well distributions of VOT can potentially discriminate betweenspeaker speech samples in Standard Thai FVC. There are several reasons why the words [khrap H] and [kha? L] are particularly interesting. First, they are Standard Thai particles, which are usually put at the end of a sentence to show a polite way of speaking (Slayden, 2009); because of their sentence-final position, more duration is guaranteed (Abramson, 1962; Naksakul, 1998). Second, since these particles, [khrap H] and [kha? L], are also used simply to say "yes" in response to a question (ibid.), they are commonly found in everyday conversation. Third, the particles [khrap H] and [kha? L] reflect the sex of the speaker. Males use [khrap H] while females use [kha? L] to show their biological sex, which will make it easier for an FVC expert to approximate the speaker's biological sex. Having said that, females are allowed to use [khrap H] and males are allowed to use [kha? L] (ibid.). Thus, both usage of [khrap H] and [kha? L] in Standard Thai (perhaps in conjunction with particles used in other dialects), would be worth investigating using FVC, as they are commonly found in everyday conversation.

#### 8.3.2.2 Trill /r/ and liquid /l/

It would also be interesting to see how the trill /r/ and liquid /l/ perform in Standard Thai as they might show idiosyncrasies typically found only in Thai speakers. From the sociolinguistic research conducted by Treyakul (1986), it was found 1) that the trills /r/ and /-r/ were mostly realized as [r] in the news announcement and passage reading styles (formal situations), whereas the lateral /l/ and /-l/ were mostly realized as [l] regardless

of stylistic variation; and 2) that the trill r was mostly realized as [1] and the trill -r deleted (>  $[\emptyset]$ ) in an informal situation (an interview).

A more recent linguistic study, which examined the relationship between the social variation of /r/ in Thai and /r/ in English in the speech of Bangkok Thai, is by Chunsuvimol (1992). First, it was found that /r/ in each of the two languages (Thai and English) has four main variants: 1) a tap [r]; 2) an approximant [x]; 3) a lateral [l]; and 4) r-lessness [Ø]. Whereas the first three variants occurred in both prevocalic positions and clusters, the fourth variant occurred in clusters only. Second, female speakers (as opposed to males) and speakers whose job was at a higher level, as well as those who had a longer background of speaking English, were found to use more prestigious variants ([r] in Thai and [1] in English). Such stylistic variation in the pronunciation of /r/ and /l/ might be of use for FVC purposes. Suppose that in the prevocalic position, there are 30 words in which a trill /r/ could have been articulated as the lateral /l/, but only 12 words where it in fact was. In another sample, there may be 40 lateral /l/ words, all except 10 said with a trill /r/ (hypercorrection that is defined as "an instance where an individual believes a linguistic rule has applied in a case where it has <u>not</u> actually applied" (Beebe, 1974, p. 355). Forensically, such difference between the incidence of (12/30 =) 40% and (30/40 =)=) 75% needs to be evaluated to answer if the 35% difference is more likely to be from the same or from different speakers.

#### 8.3.3 Statistical tools and data extraction techniques

A better approximation of within- and between-speaker variation may be possible with different statistical techniques such as GMM-UBM and Hidden Markov Models. Apart from using speech data from a laboratory recording, as in this thesis, speech data obtained during transmission channel mismatch (e.g. mobile-to-landline vs landline-to-landline) should be undertaken to test the effect of forensically realistic conditions.

# 8.4 Implications for the forensic academic community

Based on the results reported, we can say that Standard Thai is generally amenable to FVC. The results suggest that 1) the traditional parameters, which are tonal F0 and formant trajectories, as opposed to automatic parameters such as the spectrum, should be used as parameters for Standard Thai FVC; 2) nasal segments seem to be the best

candidates, as opposed to affricates and fricatives; 3) longer duration of the target segments can ensure that more individualizing information is captured; and 4) distance from the microphone when recording speech samples can affect EER values (Campbell, 2014).

Although it is too early at this stage to generalize the results obtained in the current thesis to the Thai legal context, it is appropriate to say that the results obtained can only be used as intelligence in police investigations, not as evidence in courts of law. FVC analyses must be conducted with caution and responsibility. Since no consensus has been reached for best practice in examining and reporting voice evidence in Thai courts, there should be collaboration among analysts and examiners to improve the usability and performance of FVC practices in Thailand. Participation in forensic-style evaluations should be extended to analysts. Last but not least, the reliability and validity of the approaches employed among forensic practitioners in Thailand need to be addressed before presenting voice evidence to the courts of law.

# 8.5 Summary

In this chapter, I have answered the two research questions. The fusion results of the two segments /n/ - [na: HL] and /n/ - [noi L] have been reported. Finally, suggestions for further FVC research in Standard Thai have been made.

# References

- Abboud, A. (2017). Daubert v. Merrell Dow Pharmaceuticals, Inc. (1993). Retrieved from the Embryo Project Encyclopedia website: https://embryo.asu.edu/pages/daubert-v-merrell-dow-pharmaceuticals-inc-1993
- Abramson, A. S. (1962). The vowels and tones of Standard Thai: Acoustical measurements and experiments (Unpublished doctoral thesis), Indiana University, USA.
- Aitken, C. G. G., & Lucy, D. (2004). Evaluation of trace evidence in the form of multivariate data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 53(1), 109-122.
- Aitken, C. G. G., & Stoney, D. A. (1991). *The use of statistics in forensic science*. Chichester, UK: Ellis Horwood.
- Aitken, C. G. G., & Taroni, F. (2004). Statistics and the evaluation of evidence for forensic scientists. Chichester, UK: Wiley.
- Aitken, C., Berger, C. E., Buckleton, J. S., Champod, C., Curran, J., Dawid, A. P., ... Jackson, G. (2011). Expressing evaluative opinions: A position statement. *Science and Justice*, *51*(1), 1-2.
- Alderman, T. (2005). Forensic speaker identification: A likelihood ratio-based approach using vowel formants. Munich: LINCOM.
- Amino, K., Sugawara, T., & Arai, T. (2006). Idiosyncrasy of nasal sounds in human speaker identification and their acoustic properties. *Acoustical Science and Technology*, 27(4), 233-235.
- Andrews, W. D., Kohler, M. A., Campbell, J. P., & Godfrey, J. J. (2001). Phonetic, idiolectal and acoustic speaker recognition. *Odyssey 2001* (pp. 55-63). Retrieved from the ISCA website: https://www.isca-speech.org/archive\_open/archive\_papers/odyssey/pres/odys\_055\_p.pdf
- Atal, B. S. (1972). Automatic speaker recognition based on pitch contours. *Journal of the Acoustical Society of America*, 52(6B), 1687-1697.
- Atal, B. S. (1976). Automatic recognition of speakers from their voices. *Proceedings of the IEEE*, 64(4), 460-475.
- Balding, D. J. (2005). Weight-of-evidence for forensic DNA profiles. Chichester, UK: Wiley.

- Beebe, L. M. (1974). *Socially conditioned variation in Bangkok Thai* (Unpublished doctoral thesis), University of Michigan, USA.
- Bernard, J. (1967). Some measurements of some sounds of Australian English (Unpublished doctoral thesis), University of Sydney, Australia.
- Bertsekas, D. P., & Tsitsiklis, J. N. (2002). *Introduction to probability*. Belmont, MA: Athena Scientific.
- Black, B., Ayala, F. J., & Saffran-Brinks, C. (1993). Science and the law in the wake of Daubert: A new search for scientific knowledge. *Texas Law Review*, 72, 715-802.
- Bladon, R. A. W., & Nolan, F. (1977). A video-fluorographic investigation of tip and blade alveolars in English. *Journal of Phonetics*, *5*, 185-193.
- Boersma, P., & Weenink, D. (2003). Praat: A system for doing phonetics by computer [computer software]. Amsterdam: Institute of Phonetic Sciences, University of Amsterdam.
- Bogert, B. P., Healy, M. J. R., & Tukey, J. W. (1963). The quefrency analysis of time series for echoes: Cepstrum, pseudo-autocovariance, cross-cepstrum and saphe cracking. In M. Rosenblatt (Ed.), *Proceedings of the Symposium on Time Series Analysis* (pp. 209-243). New York: Wiley.
- Bolt, R. H., Cooper, F. S., David Jr, E. E., Denes, P. B., Pickett, J. M., & Stevens, K. N. (1970). Speaker identification by speech spectrograms: A scientists' view of its reliability for legal purposes. *Journal of the Acoustical Society of America*, 47(2B), 597-612.
- Bolt, R. H., Cooper, F. S., David Jr, E. E., Denes, P. B., Pickett, J. M., & Stevens, K. N. (1973). Speaker identification by speech spectrograms: Some further observations. *Journal of the Acoustical Society of America*, *54*(2), 531-534.
- Boves, L. W. J. (1998). Commercial applications of speaker verification: Overview and critical success factors. In *Proceedings of RLA2C*, *Avignon*, 150-159.
- Bozza, S., Taroni, F., Marquis, R., & Schmittbuhl, M. (2008). Probabilistic evaluation of handwriting evidence: Likelihood ratio for authorship. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *57*(3), 329-341.
- Braun, A. (1995). Fundamental frequency: How speaker-specific is it? *Beiträge zur Phonetik und Linguistik*, 64, 9-23.

- Bricker, P. D., Gnanadesikan, R., Mathews, M. V., Pruzansky, S., Tukey, P. A., Wachter, K. W., & Warner, J. L. (1971). Statistical techniques for talker identification. *Bell System Technical Journal*, 50(4), 1427-1454.
- Broeders, A. P. A. (1995). The role of automatic speaker recognition techniques in forensic investigations. *Proceedings of the XIIth International Congress of Phonetic Sciences, Stockholm: Vol. 3* (pp. 154-161).
- Brümmer, N. (2007). FoCal toolkit. Retrieved from http://www.dsp.sun.ac.za/nbrummer/focal
- Brümmer, N., & du Preez, J. (2006). Application-independent evaluation of speaker detection. *Computer Speech & Language*, 20(2), 230-275.
- Butterfint, Z. (2004). *Individuality in phonetic variation:* An investigation into the role of intra-speaker variation in speaker discrimination (Unpublished doctoral thesis), University of Manchester, UK.
- Byrne, C., & Foulkes, P. (2007). The 'mobile phone effect' on vowel formants. *International Journal of Speech, Language and the Law, 11*(1), 83-102.
- Campbell, J. P. (2014). Speaker recognition for forensic applications. *Odyssey 2014*.

  Retrieved from the University of Eastern Finland website: http://cs.uef.fi/odyssey2014/downloads/Odyssey\_Keynote\_Campbell\_20140616
  \_final\_v3.pdf
- Cassidy, S. (1999). The Emu Speech Database System [computer software]. Retrieved from http://emu.sourceforge.net/manual/chap.ssff.html
- Castro, D. R. (2007). Forensic evaluation of the evidence using automatic speaker recognition systems (Unpublished doctoral thesis), Universidad autónoma de Madrid, Spain.
- Catford, J. C. (1977). Fundamental problems in phonetics. Bloomington, IN: Indiana University Press.
- Champod, C., & Evett, I. W. (2000). Commentary on A. P. A. Broeders (1999) 'Some observations on the use of probability scales in forensic identification', Forensic Linguistics 6(2): 228–41. *International Journal of Speech, Language and the Law*, 7(2), 239-243.
- Champod, C., & Meuwly, D. (2000). The inference of identity in forensic speaker recognition. *Speech Communication*, 31(2), 193-203.

- Chen, A., & Rose, P. (2012). Likelihood ratio-based forensic voice comparison with the Cantonese triphthong /iau/. *Proceedings of the 14th Australasian International Conference on Speech Science and Technology, Sydney* (pp. 197-200).
- Chen, N. F., Shen, W., Campbell, J. P., & Schwartz, R. (2009). Large-scale analysis of formant frequency estimation variability in conversational telephone speech.

  Proceedings of INTERSPEECH-2009 (pp. 2203-2206).
- Chunsuvimol, B. (1992). Relationship between the social variation of (r) in Thai and (r) in English in the speech of Bangkok Thai speakers (Unpublished doctoral thesis), Chulalongkorn University, Thailand.
- Clark, J., & Yallop, C. (1990). An introduction to phonetics and phonology. Oxford: Blackwell.
- Clermont, F., & Itahashi, S. (2000). Static and dynamic vowels in a "cepstro-phonetic" sub-space. *Acoustical Science and Technology*, 21(4), 221-223.
- Cooke, J. R. (1989). Thai sentence particles: Forms, meaning and formal-semantic variations. *Southeast Asian Linguistics*, *12*, 1-90.
- DeGroot, M. H., & Fienberg, S. E. (1983). The comparison and evaluation of forecasters. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 32(1-2), 12-22.
- Doddington, G. R. (2001). Speaker recognition based on idiolectal differences between speakers. Paper presented at the Seventh European Conference on Speech Communication and Technology.
- Drygajlo, A., Meuwly, D., & Alexander, A. (2003). Statistical methods and Bayesian interpretation of evidence in forensic automatic speaker recognition. Paper presented at the Eighth European Conference on Speech Communication and Technology.
- Eisler, F. G. (1968). *Psycholinguistics: Experiments in spontaneous speech*. New York: Academic Press.
- Elliott, J. (2000). Comparing the acoustic properties of normal and shouted speech: A study in forensic phonetics. *Proceedings of the 8th Australasian International Conference on Speech Science and Technology* (pp. 154-159).
- Enzinger, E. (2009). Formant trajectories in forensic speaker recognition (Unpublished M.A. thesis), University of Vienna, Austria.

- Enzinger, E., & Zhang, C. (2011). *Nasal spectra for forensic voice comparison*. Paper presented at the Special Session on Forensic Acoustics, 162nd ASA Meeting, San Diego.
- Evett, I. W. (1991). Interpretation: A personal odyssey. In C. G. Aitken, & D. A. Stoney (Eds.), *The use of statistics in forensic science* (pp. 9-22). Chichester, UK: Ellis Horwood.
- Evett, I. W. (1998). Towards a uniform framework for reporting opinions in forensic science casework. *Science & Justice*, 38(3), 198-202.
- Evett, I. W., & Buckleton, J. S. (1996). Statistical analysis of STR data. In A. Carracedo, B. Brinkmann, & W. Bär (Eds.), 16th Congress of the International Society for Forensic Haemogenetics (Internationale Gesellschaft für forensische Hämogenetik e.V.), Santiago de Compostela, 12–16 September 1995 (pp. 79-86). Berlin: Springer.
- Forrest, K., Weismer, G., Milenkovic, P., & Dougall, R. N. (1988). Statistical analysis of word- initial voiceless obstruents: Preliminary data. *Journal of the Acoustical Society of America*, 84(1), 115-123.
- Franco-Pedroso, J., Gonzalez-Rodriguez, J., Gonzalez-Dominguez, J., & Ramos, D. (2012). Fine-grained automatic speaker recognition using cepstral trajectories in phone units. In C. Donohue, S. Ishihara, & W. Steed (Eds.), *Quantitative approaches to problems in linguistics: Studies in honour of Phil Rose* (pp. 185-195). Munich: LINCOM.
- French, P., Nolan, F., Foulkes, P., Harrison, P., & McDougall, K. (2010). The UK position statement on forensic speaker comparison: A rejoinder to Rose and Morrison. *International Journal of Speech, Language and the Law, 17*(1), 143-152.
- Friedman, R. D. (1996). Assessing evidence. Review of *Statistics and the evaluation of evidence for forensic scientists*, by C. G. G. Aitken, *Interpreting evidence: Evaluating forensic science in the courtroom*, by B. Robertson and G. A. Vignaux, and *Evidential foundations of probabilistic reasoning*, by D. A. Schum. *Michigan Law Review*, 94(6), 1810-1838.
- Fromkin, V., Rodman, R., & Hyams, N. (2010). *An introduction to language*. Boston: Cengage Learning.
- Furui, S. (1981). Cepstral analysis technique for automatic speaker verification. *Proceedings of the IEEE*, 29(2), 254-272.

- Furui, S. (1986). Speaker-independent isolated word recognition based on emphasized spectral dynamics. *Proceedings of the IEEE*, 11, 1991-1994.
- Furui, S., Itakura, F., & Saito, S. (1972). Talker recognition by the longtime averaged speech spectrum. *IECE Transactions*, *55*(10), 549-556.
- Gandour, J. (1985). A voiced onset time analysis of word-initial stops in Thai. *Linguistics* of the Tibeto-Burman Area, 8(2), 68-80.
- Garrett, K. L., & Healey, E. C. (1987). An acoustic analysis of fluctuations in the voices of normal adult speakers across three times of day. *Journal of the Acoustical Society of America*, 82(1), 58-62.
- Garvin, P. L., & Ladefoged, P. (1963). Speaker identification and message identification in speech recognition. *Phonetica*, *9*(4), 193-199.
- Gedney, W. J. (1972). A checklist for determining tones in Tai dialects. In M. Estellie Smith (Ed.), *Studies in linguistics in honor of George L. Trager*, 423-437.
- Gigerenza, G., Swijtink, Z., Porter, T., Daston, L., Beatty, J., & Krüger, L. (1989). *The empire of chance: How probability changed science and everyday life*. Cambridge, UK: Cambridge University Press.
- Gilbert, H. R., & Weismer, G. G. (1974). The effects of smoking on the speaking fundamental frequency of adult women. *Journal of Psycholinguistic Research*, 3(3), 225-231.
- Glenn, J. W., & Kleiner, N. (1968). Speaker identification based on nasal phonation. *Journal of the Acoustical Society of America*, 43(2), 368-372.
- Gold, E., & French, P. (2011). International practices in forensic speaker comparison. International Journal of Speech, Language and the Law, 18(2), 293-307.
- Gold, E., & Hughes, V. (2014). Issues and opportunities: The application of the numerical likelihood ratio framework to forensic speaker comparison. *Science & Justice*, 54(4), 292-299.
- Gonzalez-Rodriguez, J., Drygajlo, A., Ramos-Castro, D., Garcia-Gomar, M., & Ortega-Garcia, J. (2006). Robust estimation, interpretation and assessment of likelihood ratios in forensic speaker recognition. *Computer Speech & Language*, 20(2), 331-355.
- Gonzalez-Rodriguez, J., Ortega-Garcia, J., & Sanchez-Bote, J.-L. (2002). Forensic identification reporting using automatic biometric systems. In D. Zhang (Ed.), *Biometric solutions* (pp. 169-185). Boston: Springer.

- Gonzalez-Rodriguez, J., Rose, P., Ramos, D., Toledano, D. T., & Ortega-Garcia, J. (2007). Emulating DNA: Rigorous quantification of evidential weight in transparent and testable forensic speaker recognition. *Proceedings of the IEEE*, 15(7), 2104-2115.
- Good, I. J. (1991). Weight of evidence and the Bayesian likelihood ratio. Chichester, UK: Ellis Horwood.
- Gutierrez-Osuna, R. (2017). Introduction to speech processing: Cepstral analysis.

  Retrieved from the Texas A&M University website: http://research.cs.tamu.edu/prism/lectures/sp/l9.pdf
- Halliday, M. A. K. (2015). *Intonation and grammar in British English*. Berlin: Walter de Gruyter.
- Halliday, M. A. K., McIntosh, A., & Strevens, P. (1964). *The linguistic sciences and language teaching*. London: Longman.
- Harrington, J. (2010). Phonetic analysis of speech corpora. Chichester, UK: John Wiley.
- Hecker, M. H. L., Stevens, K. N., von Bismarck, G., & Williams, C. E. (1968).
  Manifestations of task- induced stress in the acoustic speech signal. *Journal of the Acoustical Society of America*, 44(4), 993-1001.
- Hicks, T., Biedermann, A., de Koeijer, J. A., Taroni, F., Champod, C., & Evett, I. W. (2015). The importance of distinguishing information from evidence/observations when formulating propositions. *Science & Justice*, 55(6), 520-525.
- Holmes, J. N., Holmes, W. J., & Garner, P. N. (1997). Using formant frequencies in speech recognition. Proceedings of Eurospeech 1997, Rhodes, Greece (pp. 2083-2086).
- Hudson, T., De Jong, G., McDougall, K., Harrison, P., & Nolan, F. (2007). F0 statistics for 100 young male speakers of Standard Southern British English. In *Proceedings of the 16th International Congress of Phonetic Sciences, Saarbrücken*. Retrieved from the ICPhS 2007 website: http://www.icphs2007.de/conference/Papers/1570/1570.pdf
- Hughes, G. W., & Halle, M. (1956). Spectral properties of fricative consonants. *Journal of the Acoustical Society of America*, 28(2), 303-310.
- Hughes, V., & Foulkes, P. (2014). Variability in analyst decisions during the computation of numerical likelihood ratios. *International Journal of Speech, Language and the Law*, 21(2), 279-315.

- Hughes, V., Foulkes, P., & Wood, S. (2016). Formant dynamics and durations of um improve the performance of automatic speaker recognition systems. *Proceedings of the 16th Australasian Conference on Speech Science and Technology (ASSTA)*.
   University of Western Sydney, Australia.
- Hughes, V., & Rhodes, R. (2018). Questions, propositions and assessing different levels of evidence: Forensic voice comparison in practice. *Science & Justice*, 58(4), 250-257.
- Ihaka, R., & Gentleman, R. (1996). R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, 5(3), 299-314.
- Ishihara, S. (2017). Strength of linguistic text evidence: A fused forensic text comparison system. *Forensic Science International*, 278, 184-197.
- Ishihara, S., & Kinoshita, Y. (2008). How many do we need? Exploration of the population size effect on the performance of forensic speaker classification. *Proceedings of INTERSPEECH-2008* (pp. 1941-1944).
- Jialin, P., & Rose, P. (2012). Likelihood ratio-based forensic voice comparison with the Cantonese diphthong /ei/ F-pattern. Proceedings of the 14th Australasian International Conference on Speech Science and Technology, Sydney (pp. 205-208)
- Johnstone, T., & Scherer, K. R. (2000). Vocal communication of emotion. In M. Lewis & J. Haviland (Eds.), *The handbook of emotion* (pp. 220-235). New York: Guilford.
- Jongman, A., Wayland, R., & Wong, S. (2000). Acoustic characteristics of English fricatives. *Journal of the Acoustical Society of America*, 108(3), 1252-1263.
- Kasuriya, S., Sornlertlamvanich, V., Cotsomrong, P., Kanokphara, S., & Thatphithakkul,
   N. (2003). Thai speech corpus for Thai speech recognition. *The Oriental COCOSDA*, *Singapore* (pp. 54-61).
- Kavanagh, C. (2012). New consonantal acoustic parameters for forensic speaker comparison (Unpublished doctoral thesis), University of York, UK.
- Kent, R. D. (2002). Acoustic analysis of speech. San Diego: Singular.
- Kersta, L. G. (1962). Voiceprint identification. *Journal of the Acoustical Society of America*, 34(5), 725-725.

- Khodai-Joopari, M. (2006). Forensic speaker analysis and identification by computer: A Bayesian approach anchored in the cepstral domain (Unpublished doctoral thesis), University of New South Wales, Australia.
- Kinoshita, Y. (2005). Does Lindley's LR estimation formula work for speech data? Investigation using long-term f0. *International Journal of Speech, Language and the Law*, 12(2), 235-254.
- Kinoshita, Y., & Ishihara, S. (2010). F0 can tell us more: Speaker verification using the long term distribution. *Proceedings of the 13th Australasian International Conference on Speech Science and Technology* (pp. 50-53).
- Kinoshita, Y., & Ishihara, S. (2014). Background population: How does it affect LR-based forensic voice comparison? *International Journal of Speech, Language and the Law*, 21(2), 191-224.
- Kinoshita, Y., Ishihara, S., & Rose, P. (2008). Beyond the long-term mean: Exploring the potential of F0 distribution parameters in traditional forensic speaker recognition. *Odyssey* 2008 (pp. 329-334).
- Kinoshita, Y., Ishihara, S., & Rose, P. (2009). Exploring the discriminatory potential of F0 distribution parameters in traditional forensic speaker recognition. International Journal of Speech, Language and the Law, 16(1), 91-111.
- Kinoshita, Y., & Osanai, T. (2006). Within speaker variation in diphthongal dynamics: what can we compare? *Proceedings of the 11th Australasian International Conference on Speech Science and Technology*, Auckland (pp. 112-117).
- Klasmeyer, G., & Sendlmeier, W. F. (2013). The classification of different phonation types in emotional and neutral speech. *International Journal of Speech, Language and the Law*, 4(1), 104-124.
- Klingholz, F., Penning, R., & Liebhardt, E. (1988). Recognition of low-level alcohol intoxication from speech signal. *Journal of the Acoustical Society of America*, 84(3), 929-935.
- Kumar, K., Kim, C., & Stern, R. (2011). Delta-spectral cepstral coefficients for robust speech recognition. *Proceedings of the 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 4784-4787).
- Kumar, P., & Lahudkar, S. L. (2015). Automatic speaker recognition using LPCC and MFCC. *International Journal on Recent and Innovation Trends in Computing and Communication*, *3*(4), 2106-2109.

- Künzel, H. J. (2001). Beware of the 'telephone effect': The influence of telephone transmission on the measurement of formant frequencies. *Forensic Linguistics*, 8(1), 80-99.
- Künzel, H. J. (2007). Effects of voice disguise on speaking fundamental frequency. *International Journal of Speech, Language and the Law, 7*(2), 150-179.
- Ladefoged, P., & Johnson, K. (2014). A course in phonetics. Boston: Cengage Learning.
- Ladefoged, P., & Maddieson, I. (1986). *Some of the sounds of the world's languages*. Los Angeles: Phonetics Laboratory, Department of Linguistics, UCLA.
- Laukkanen, A.-M., Vilkman, E., Alku, P., & Oksanen, H. (1996). Physical variations related to stress and emotional state: A preliminary study. *Journal of Phonetics*, 24(3), 313-335.
- Leemann, A., Mixdorff, H., O'Reilly, M., Kolly, M., & Dellwo, V. (2014). Speaker-individuality in Fujisaki model f0 features: Implications for forensic voice comparison. *International Journal of Speech, Language and the Law*, 21(2).
- Lewis, M. P., Simons, G. F., & Fennig, C. D. (2013). *Ethnologue: Languages of the World*. Texas: SIL International.
- Li, F. K. (1966). The relationship between tones and initials in Tai. In N. H. Zide (Ed.), Studies in Comparative Austro-Asiatic Linguistics (pp. 82-88). The Hague: Mouton.
- Li, J., & Rose, P. (2012). Likelihood ratio-based forensic voice comparison with F-pattern and Tonal F0 from the Cantonese /ɔy/ diphthong. Proceedings of the 14th Australasian International Conference on Speech Science and Technology (pp. 201-204).
- Lindblad, P. (1980). Svenskans sje-och tje-ljud i ett allmänfonetiskt perspektiv (Vol. 16). Lund, Sweden: LiberLäromedel/Gleerup.
- Lindley, D. V. (1977). A problem in forensic science. *Biometrika*, 64(2), 207-213.
- Lindley, D. V. (1990). The 1988 Wald memorial lectures: The present position in Bayesian statistics. *Statistical Science*, *5*, 44-65.
- Liu, H.-M., Tseng, C.-H., & Tsao, F.-M. (2000). Perceptual and acoustic analysis of speech intelligibility in Mandarin-speaking young adults with cerebral palsy. *Clinical linguistics & Phonetics*, *14*(6), 447-464.

- Loakes, D. (2008). A forensic phonetic investigation into the speech patterns of identical and non-identical twins. *International Journal of Speech, Language and the Law,* 15(1), 97-100.
- Loakes, D., & McDougall, K. (2004). Frication of /k/ and /p/ in Australian English: Interand intra-speaker variation. *Proceedings of the 10th Australasian International Conference on Speech Science and Technology* (pp. 171-176).
- Luemsai, S. (2001). *Northeastern Thai Dialect*. Bangkok: Department of Eastern Languages, Silpakorn University.
- Maekawa, K. (1998). Phonetic and phonological characteristics of paralinguistic information in spoken Japanese. *Proceedings of the Fifth International Conference on Spoken Language Processing, Sydney* (pp. 635-638).
- Mannell, R. (2009). Phonetics and phonology: Articulation of nasal stops. Retrieved from the Macquarie University website: https://clas.mq.edu.au/speech/phonetics/phonetics/consonants/nasal\_stops.html
- Markham, D. (2007). Listeners and disguised voices: The imitation and perception of dialectal accent. *International Journal of Speech, Language and the Law*, 6(2), 290-299.
- Mays, C., & Beckman, M. (2008). An acoustic study of affricates in the Songyuan dialect of Mandarin Chinese. Poster presented at the 22nd Annual CIC SROP Research Conference, Michigan State University, 24-26 July 2008.
- Mazzoni, D., & Dannenberg, R. (2000). Audacity [computer software]. Pittsburg: The Audacity Team.
- McDougall, K. (2004). Speaker-specific formant dynamics: An experiment on Australian English /aɪ/. *International Journal of Speech, Language and the Law, 11*(1), 103-130.
- McDougall, K., & Nolan, F. (2007). Discrimination of speakers using the formant dynamics of /u:/ in British English. *Proceedings of the 16th International Congress of Phonetic Sciences, Saarbrücken* (pp. 1825-1828).
- Meseguer, N. A. (2009). *Speech analysis for automatic speech recognition* (Unpublished M.A. thesis), Norwegian University of Science and Technology, Norway.
- Meuwly, D. (2004). Forensic speaker recognition: An evidence odyssey. *Odyssey* 2004.

- Milner, B., & Shao, X. (2006). Clean speech reconstruction from MFCC vectors and fundamental frequency using an integrated front-end. *Speech Communication*, 48(6), 697-715.
- Morrison, G. S. (2008). Forensic voice comparison using likelihood ratios based on polynomial curves fitted to the formant trajectories of Australian English /ai/. *International Journal of Speech, Language and the Law, 15*(2), 247-264.
- Morrison, G. S. (2009a). Forensic voice comparison and the paradigm shift. *Science & Justice*, 49(4), 298-308.
- Morrison, G. S. (2009b). Likelihood-ratio forensic voice comparison using parametric representations of the formant trajectories of diphthongs. *Journal of the Acoustical Society of America*, 125(4), 2387-2397.
- Morrison, G. S. (2011). A comparison of procedures for the calculation of forensic likelihood ratios from acoustic–phonetic data: Multivariate kernel density (MVKD) versus Gaussian mixture model–universal background model (GMM–UBM). *Speech Communication*, 53(2), 242-256.
- Morrison, G. S. (2012). The likelihood-ratio framework and forensic evidence in court: A response to R v T. *International Journal of Evidence and Proof, 16*(1), 1-29.
- Morrison, G. S. (2013). Tutorial on logistic-regression calibration and fusion: Converting a score to a likelihood ratio. *Australian Journal of Forensic Sciences*, 45(2), 173-197.
- Morrison, G. S., Enzinger, E., & Zhang, C. (2016). Refining the relevant population in forensic voice comparison: A response to Hicks et alii (2015) The importance of distinguishing information from evidence/observations when formulating propositions. *Science & Justice*, 56(6), 492-497.
- Morrison, G. S., & Kinoshita, Y. (2008). Automatic-type calibration of traditionally derived likelihood ratios: Forensic analysis of Australian English /o/ formant trajectories. *Proceedings of INTERSPEECH-2008* (pp. 1501-1504).
- Morrison, G. S., & Kondaurova, M. V. (2009). Analysis of categorical response data: Use logistic regression rather than endpoint-difference scores or discriminant analysis. *Journal of the Acoustical Society of America*, 126(5), 2159-2162.
- Morrison, G. S., & Nearey, T. M. (2011). FormantMeasurer: Software for efficient human-supervised measurement of formant trajectories. Retrieved from the first author's website: http://geoff-morrison.net/#FrmMes

- Morrison, G. S., Ochoa, F., & Thiruvaran, T. (2012). Database selection for forensic voice comparison. *Odyssey* 2012.
- Morrison, G. S., Rose, P., & Zhang, C. (2012). Protocol for the collection of databases of recordings for forensic-voice-comparison research and practice. *Australian Journal of Forensic Sciences*, 44(2), 155-167.
- Morrison, G. S., Sahito, F. H., Jardine, G., Djokic, D., Clavet, S., Berghs, S., & Dorny,
  C. G. (2016). INTERPOL survey of the use of speaker identification by law enforcement agencies. *Forensic Science International*, 263, 92-100.
- Murphy, K. P. (2012). *Machine learning: A probabilistic perspective*. Boston: The MIT Press.
- Nair, B. B. T., Alzqhoul, E. A. S., & Guillemin, B. J. (2014). Comparison between Melfrequency and complex cepstral coefficients for forensic voice comparison using a likelihood ratio framework. *Proceedings of the World Congress on Engineering and Computer Science, San Francisco: Vol. 1* (pp. 496-501).
- Nakasone, H., & Beck, S. D. (2001). Forensic automatic speaker recognition. *Odyssey* 2001.
- Naksakul, K. (1998). *Thai sound system*. Bangkok: Chulalongkorn University Press.
- National Research Council (2009). *Strengthening forensic science in the United States: A path forward.* Washington, DC: National Academies Press.
- Neumann, C., Champod, C., Puch- Solis, R., Egli, N., Anthonioz, A., & Bromage-Griffiths, A. (2007). Computation of likelihood ratios in fingerprint identification for configurations of any number of minutiae. *Journal of Forensic Sciences*, 52(1), 54-64.
- Ngamkham, W., & Nanuam, W. (2015, August 30). Phone data leads to bomb arrest.

  \*Bangkok Post.\* Retrieved from the newspaper's website: 
  http://www.bangkokpost.com/print/673708/
- Nolan, F. (1983). *The phonetic bases of speaker recognition*. Cambridge, UK: Cambridge University Press.
- Nolan, F. (2001). Speaker identification evidence: Its forms, limitations, and roles. Proceedings of the Conference "Law and Language: Prospect and Retrospect", Levi (Finnish Lapland) (pp. 1-19).
- Nolan, F. (2014). Intonation. Retrieved from the University of Cambridge website: http://www.ling.cam.ac.uk/francis/FN\_inton\_prepub.pdf

- Nolan, F., McDougall, K., De Jong, G., & Hudson, T. (2006). A forensic phonetic study of 'dynamic' sources of variability in speech: the DyViS project. *Proceedings of the 11th Australasian International Conference on Speech Science and Technology* (pp. 13-18).
- Nookua, S. (2012). The patterns of language use in the southernmost provinces of Thailand. กระแส วัฒนธรรม (*Cultural Trends*), 12(22), 26-35.
- Novak, A., Dlouha, O., Capkova, B., & Vohradnik, M. (1991). Voice fatigue after theater performance in actors. *Folia Phoniatrica et Logopaedica*, *43*(2), 74-78.
- Onsuwan, C. (2005). *Temporal relations between consonants and vowels in Thai syllables* (Unpublished doctoral thesis), University of Michigan, USA.
- Paeschke, A., Kienast, M., & Sendlmeier, W. F. (1999). F0-contours in emotional speech. In *Proceedinfs of the 14th International Congress of Phonetic Sciences: Vol. 2* (pp. 929-932).
- Payne, R. W. (2009). GenStat. Wiley Interdisciplinary Reviews: Computational Statistics, 1(2), 255-258.
- Pingjai, S. (2011). Forensic voice comparison in Thai: A likelihood ratio-based approach using tonal acoustics (Unpublished M.A. thesis), Australian National University, Australia.
- Pingjai, S., Ishihara, S., & Sidwell, P. J. (2013). A likelihood ratio-based forensic voice comparison using formant trajectories of Thai diphthongs. In *Proceedings of Meetings on Acoustics ICA2013* (Vol. 19, No. 1, p. 060043).
- President's Council of Advisors on Science and Technology (US) (2016). Report to the President, Forensic science in criminal courts: Ensuring scientific validity of feature-comparison methods. Executive Office of the President of the United States, President's Council of Advisors on Science and Technology.
- Pruzansky, S. (1963). Pattern-matching procedure for automatic talker recognition. Journal of the Acoustical Society of America, 35(3), 354-358.
- Pruzansky, S., & Mathews, M. V. (1964). Talker- recognition procedure based on analysis of variance. *Journal of the Acoustical Society of America*, 36(11), 2041-2047.
- Ramos-Castro, D., Gonzalez-Rodriguez, J., & Ortega-Garcia, J. (2006). Likelihood ratio calibration in a transparent and testable forensic speaker recognition framework. *Odyssey* 2006.

- Rantala, L., Vilkman, E., & Bloigu, R. (2002). Voice changes during work: Subjective complaints and objective measurements for female primary and secondary school teachers. *Journal of voice*, *16*(3), 344-355.
- Recasens, D., & Espinosa, A. (2007). An electropalatographic and acoustic study of affricates and fricatives in two Catalan dialects. *Journal of the International Phonetic Association*, 37(2), 143-172.
- Reetz, H., & Jongman, A. (2011). *Phonetics: Transcription, production, acoustics, and perception*. Chichester, UK: John Wiley.
- Reynolds, D., Andrews, W., Campbell, J., Navratil, J., Peskin, B., Adami, A., ... Godfrey, J. (2003). The SuperSID project: Exploiting high-level information for high-accuracy speaker recognition. *Proceedings of the 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, Hong Kong: Vol. 4* (pp. 784-787).
- Robertson, B., & Vignaux, G. (1995). *Interpreting evidence: Evaluating forensic science in the courtroom*. Chichester, UK: John Wiley.
- Roengpitya, R. (2012). Different durations of diphthongs in Thai: A new finding. Annual Meeting of the Berkeley Linguistics Society, 28(2), 43-54.
- Rose, P. (1991). How effective are long term mean and standard deviation as normalisation parameters for tonal fundamental frequency? *Speech Communication*, 10(3), 229-247.
- Rose, P. (2002). Forensic speaker identification. New York: Taylor & Francis.
- Rose, P. (2003). The technical comparison of forensic voice samples. In I. Freckelton, &H. Selby (Eds.), *Expert Evidence* (chapter 99). Sydney: Thomson Lawbook Company.
- Rose, P. (2006). Technical forensic speaker recognition: Evaluation, types and testing of evidence. *Computer Speech & Language*, 20(2), 159-191.
- Rose, P. (2011). Forensic voice comparison with secular shibboleths: A hybrid fused GMM-multivariate likelihood ratio-based approach using alveolo-palatal fricative cepstral spectra. *Proceedings of the 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5900-5903).
- Rose, P. (2013a). More is better: Likelihood ratio-based forensic voice comparison with vocalic segmental cepstra frontends. *International Journal of Speech, Language and the Law, 20*(1), 77-116.

- Rose, P. (2013b). Where the science ends and the law begins: Likelihood ratio-based forensic voice comparison in a \$150 million telephone fraud. *International Journal of Speech, Language and the Law*, 20(2), 277-324.
- Rose, P., & Clermont, F. (2001). A comparison of two acoustic methods for forensic speaker discrimination. *Acoustics Australia*, 29(1), 31-36.
- Rose, P., Kinoshita, Y., & Alderman, T. (2006). Realistic extrinsic forensic speaker discrimination with the diphthong/ai/. *Proceedings of the 11th Australasian International Conference on Speech Science and Technology* (pp. 329-334).
- Rose, P., & Morrison, G. S. (2009). A response to the UK position statement on forensic speaker comparison. *International Journal of Speech, Language and the Law,* 16(1), pp. 139-163.
- Rose, P., Osanai, T., & Kinoshita, Y. (2003). Strength of forensic speaker identification evidence: Multispeaker formant- and cepstrum-based segmental discrimination with a Bayesian likelihood ratio as threshold. *Forensic Linguistics*, 10(2), 179-202.
- Rose, P., & Simmons, A. (1996). F-pattern variability in disguise and over the telephone: Comparisons for forensic speaker identification. *Proceedings of the 6th Australasian International Conference on Speech Science and Technology* (pp. 121-126).
- Rose, P., Warren, P., & Watson, C. (2006). The intrinsic forensic discriminatory power of diphthongs. *Proceedings of the 11th Australasian International Conference on Speech Science and Technology* (pp. 64-69).
- Rose, P., & Winter, E. (2010). Traditional forensic voice comparison with female formants: Gaussian mixture model and multivariate likelihood ratio analyses. Proceedings of the 13th Australasian International Conference on Speech Science and Technology (pp. 42-45).
- Rungruengsri, U. (Ed.) (1991). *The Northern-Standard Thai Dictionary*. Chiang Mai, Thailand: Faculty of Humanities, Chiang Mai University.
- Saks, M. J., & Koehler, J. J. (2005). The coming paradigm shift in forensic identification science. *Science*, *309*(5736), 892-895.
- Schiel, F., & Heinrich, C. (2009). Laying the foundation for in-car alcohol detection by speech. *Proceedings of INTERSPEECH-2009* (pp. 983-986).

- Shadle, C. H. (2012). The aerodynamics of speech. In W. J. Hardcastle, J. Laver, & F. E. Gibbon (Eds.), *The handbook of phonetic sciences* (chapter 2). Chichester, UK: John Wiley.
- Sheather, S. J., & Jones, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society, Series B* (*Methodological*), 53(3), 683-690.
- Sheikh, M. J. (n.d.). Speaker dependent features in stops and affricates of native Urdu speakers. Lahore, Pakistan: Center for Research in Urdu Language Processing (pp. 102-110).
- Slayden, G. (2009). Central Thai phonology. Retrieved from http://www.thai-language. com/resources/slayden-thai-phonology.pdf
- Smalley, W. A. (1994). Linguistic diversity and national unity: Language ecology in *Thailand*. Chicago: University of Chicago Press.
- Sorensen, D., & Horii, Y. (1982). Cigarette smoking and voice fundamental frequency. *Journal of Communication Disorders*, 15(2), 135-144.
- Sornlertlamvanich, V., & Thongprasirt, R. (2001). Speech technology and corpus development in Thailand. *The Oriental COCOSDA Workshop* (pp. 44-47).
- Sriyaphai, W. (2013). *Thai linguistics (Phasasart PhasaThai)*. Bangkok: Chulalongkorn University Press.
- Steffensmeier, D., & Allan, E. (1996). Gender and crime: Toward a gendered theory of female offending. *Annual Review of Sociology*, 22(1), 459-487.
- Stevens, K. N. (1993). Modelling affricate consonants. *Speech Communication*, 13(1-2), 33-43.
- Stevens, K. N. (2000). Acoustic phonetics. Boston: The MIT Press.
- Stuart-Smith, J., Timmins, C., & Wrench, A. (2003). Sex and gender differences in Glaswegian /s/. *Proceedings of the 15th International Congress of Phonetic Sciences* (pp. 1851-1854).
- Su, L. S., Li, K. P., & Fu, K. (1974). Identification of speakers by use of nasal coarticulation. *Journal of the Acoustical Society of America*, *56*(6), 1876-1883.
- Subtelny, J. D., & Oya, N. (1972). Cineradiographic study of sibilants. *Folia Phoniatrica et Logopaedica*, 24(1), 30-50.

- Thaitechawat, S., & Foulkes, P. (2011). Discrimination of speakers using tone and formant dynamics in Thai. *Proceedings of the 17th International Congress of Phonetic Sciences, Hong Kong* (pp. 1978-1981).
- Tiamjan, B. (2006). *The criminal procedure code: Update 2005-2008*. Bangkok: Sout Paisal Press.
- Tienmee, W. (1992). Classification by tone shapes and by patterns of tonal splits and coalescenses. *Mon-Khmer Studies*, *21*, 229-236.
- Tingsabadh, M., & Abramson, A. S. (1993). Thai. *Journal of the International Phonetic Association*, 23(1), 24-28.
- Tosi, O. (1979). *Voice identification: Theory and legal applications*. Baltimore: University Park Press.
- Treyakul, S. (1986). Stylistic variations of "R" and "L" in Bangkok Thai: A study of the pronunciation of Bangkok F.M. radio newscasters (Unpublished M.A. thesis), Chulalongkorn University, Thailand.
- Turk, A., Nakai, S., & Sugahara, M. (2006). Acoustic segment durations in prosodic research: A practical guide. In S. Sudhoff, D. Lenertova, R. Meyer, S. Pappert, P. Augurzky, I. Mleinek, ... J. Schließer (Eds.), *Methods in empirical prosody research: Vol. 3*. New York: Walter de Gruyter.
- van Es, A., Wiarda, W., Hordijk, M., Alberink, I., & Vergeer, P. (2017). Implementation and assessment of a likelihood ratio approach for the evaluation of LA-ICP-MS evidence in forensic glass analysis. *Science & Justice*, *57*(3), 181-192.
- Vergeer, P., van Es, A., de Jongh, A., Alberink, I., & Stoel, R. (2016). Numerical likelihood ratios outputted by LR systems are often based on extrapolation: When to stop extrapolating? *Science & Justice*, *56*(6), 482-491.
- Vanderslice, R. (1969). The "voiceprint" myth. *Studies in language and language behavior: Progress report 8* (pp. 386-406). Ann Arbor: Center for Research on Language and Language Behavior, University of Michigan.
- vlab.amrita.edu. (2011). Cepstral analysis of speech. Retrieved from vlab.amrita.edu/?sub =3&brch=164&sim=615&cnt=1
- Wand, M. P., & Jones, M. C. (1994). *Kernel smoothing*. New York: Chapman and Hall/CRC.
- Wang, C. Y., & Rose, P. (2012). Likelihood ratio-based forensic voice comparison with Cantonese /i/ F-pattern and Tonal F0. *Proceedings of the 14th Australasian*

- International Conference on Speech Science and Technology (SST 2012) (pp. 209-212).
- Wannasaeng, P. (2008). Expert witnesses and environmental litigation: Legal seminar on using international law to resolve the environmental disputes. *Second anniversary of the Rabi Bhadanasak Research and Development Institute*, 68.
- Watanabe, T. (1998). Japanese pitch and mood. *Nihongakuho [Osaka University]*, 17, 97-110.
- Wilaisak, K. (n.d.). Thai dialects. Retrieved from http://cyberlab.lh1.ku.ac.th/elearn/faculty/human/hm19/lesson3.htm
- Williams, C. E., & Stevens, K. N. (1972). Emotions and speech: Some acoustical correlates. *Journal of the Acoustical Society of America*, 52(4B), 1238-1250.
- Wimolkasem, K. (2006). *Northern Thai dialect (Phasa Thai Thin Nuea)*. Bangkok: Department of Eastern Asian Languages, Silpakorn University.
- Wolf, J. J. (1970). Simulation of the measurement phase of an automatic speaker recognition system. *Journal of the Acoustical Society of America*, 47(1A), 83-83.
- Wolf, J. J. (1972). Efficient acoustic parameters for speaker recognition. *Journal of the Acoustical Society of America*, 51(6B), 2044-2056.
- Yim, A. C. S., & Rose, P. (2012). Are nasals better? Likelihood ratio-based forensic voice comparison with segmental cepstra from Cantonese and Japanese syllabic/mora nasals. Proceedings of the 14th Australasian International Conference on Speech Science and Technology (pp. 5-8).
- Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Povey, D. (2002). *The HTK book*. Cambridge: University of Cambridge.
- Zetterholm, E. (2007). Detection of speaker characteristics using voice imitation. In C. Müller (Ed.), *Speaker classification: Vol. 2* (pp. 192-205). Berlin: Springer.
- Zhang, C., & Enzinger, E. (2013). Fusion of multiple formant-trajectory-and fundamental-frequency-based forensic-voice-comparison systems: Chinese/ei/, /ai/, and/iau/. Proceedings of Meetings on Acoustics ICA2013 (Vol. 19, No. 1, p. 060044).
- Zhang, C., Morrison, G. S., Enzinger, E., & Ochoa, F. (2012). Laboratory report: Human-supervised and fully-automatic formant-trajectory measurement for forensic voice comparison Female voices. Sydney: FVC, EE&T, UNSW Laboratory Report.

- Zhang, C., Morrison, G. S., Ochoa, F., & Enzinger, E. (2012). Reliability of human-supervised formant-trajectory measurement for forensic voice comparison. *Journal of the Acoustical Society of America*, 133(1), EL54-EL60.
- Zhang, C., Morrison, G. S., & Thiruvaran, T. (2011). Forensic voice comparison using Chinese /iau/. *Proceedings of the 17th International Congress of Phonetic Sciences*, *Hong Kong* (pp. 2280-2283).
- Zhang, C., & Tan, T. (2008). Voice disguise and automatic speaker recognition. *Forensic Science International*, 175(2), 118-122.

# Appendix A Recording manuals

<u>คู่มือบันทึกเสียง</u>

**Recording manuals** 

กิจกรรมที่ 1: สนทนาแลกเปลี่ยนข้อมูล

**Activity 1: Information-exchange task** 

#### ท่านคือผู้พูด A

#### You are Speaker A

ด้านหลังกระดาษนี้เป็นเอกสารที่ส่งมาทางแฟกซ์ แต่เนื่องจากเครื่องรับ-ส่งแฟกซ์มีคุณภาพไม่ดี ให้ข้อมูลบางส่วน ในเอกสารนี้ไม่ชัดเจน คู่สนทนาของท่าน (ผู้พูด B) ก็ได้รับแฟกซ์ฉบับนี้เช่นกัน ซึ่งแฟกซ์ของผู้พูด B อาจมี คุณภาพชัดเจนกว่าของท่านหรือไม่ก็ได้ ต่อโทรศัพท์ไปยังผู้พูด B ตามหมายเลข 2357 จงสอบถามข้อมูลที่ ขาดหายไป และเขียนข้อมูลลงบนกระดาษที่ให้มาด้านหลัง

On the back of this document is a facsimile listing many fresh food products and their prices. Unfortunately, the fax is not in good quality, so some information is illegible for you but might be legible for your interlocutor (Speaker A). Ring Speaker B on ext. 2357 and ask for the information that is obfuscated to you. Then, write down that missing information on the same piece of paper provided.

\*หมายเหตุ โปรดใช้สรรพนามว่า A แทนตัวท่าน และ B แทนคู่สนทนา และวางโทรศัพท์ หลังจากทำกิจกรรมนี้เสร็จสิ้น

\*Please identify yourself as A and your interlocutor as B. Please hang up the phone after finishing this task.

## ราคาขายปลี่กลินค้า หมวดอาหารสดบางชนิดในกรุงเทพมหานคร

# ประจำวันที่ 27 พฤศจิกายน 2554

รายการ	หน่วย	6 กิ.ย. 54	27 พ.ย. 54
สุกรชำแหละ			
• เนื้อแดง (ตะโพก)	กก.	115.00-120.00	116.00-122.00
• เนื้อแดง (ไหล่)	กก.	115.00-120.00	119.75-124.00
		0.00.0.00	2 25 2 50
พริกขี้หนู (จินคา)	ขีด	8.00-9.00	7.00-8.00
เนื้อโค สันนอก/สะโพก	กก.	150.00-160.00	162.00-165.00
มะนาว(เบอร์ 3-4) ผลละ	ผล	1.75-2.00	1.00-2.00
ไก่สดทั้งตัว (ไม่รวมเครื่องใน)	กก.	70.00-72.00	69.00-70.00
กุ้งขาว (70-80 ตัว/กก.)	กก.	160.00-180.00	190.00-200.00
alognata danga a Madalas	ia! a	20.00 25.00	20.00.05.00
ปลานิล	กก.	70.00-80.00	70.00-75.00
ปลาทับทิม	กก.	80.00-90.00	75.00-80.00
ไข่เป็ด(กลาง)	ฟอง	4.90-5.00	4.90-5.00
181 1 A/1819 & V	911 O 0	2 (0 2 70	2 (2 2 2 2
ไข่ไก่(เบอร์ 3)	ฟอง	3.50-3.60	3.50-3.60
คะน้ำ	กก.	28.00-30.00	29.00-30.00
ผักบุ้งจิน	กก.	20.00-22.00	21.00-22.00
ผักกวางตุ้ง	กก.	12.00-15.00	12.00-15.00
กะหล่ำปลี	กก.	12.00-15.00	14.00-15.00
กะหล่ำดอก	กก.	16.00-17.00	17.00-19.00
		20 00 20 00	27 00 00 00
เห็คชิเมจิขาว	ขีด	29.00-30.00	27.00-29.00
สาลี่ทอง	กก.	48.00-50.00	49.00-52.00
് . മാര്ഥാവയാ	คค	SILTHESS IN	52.00-57.00

กิจกรรมที่ 2: บอกทิศทาง

**Activity 2: Giving directions** 

ท่านเป็นผู้คุ้นเคยกับการเดินทางภายในมหาวิทยาลัยธรรมศาสตร์ รังสิตเป็นอย่างดี แต่คู่สนทนาของท่านไม่คุ้นเคย กับการเดินทางภายในมหาวิทยาลัยธรรมศาสตร์ รังสิต ให้ท่านรอรับโทรศัพท์จากคู่สนทนา และตอบคำถาม เกี่ยวกับเส้นทางไปยังตึกต่างๆ (ตามแผนที่ที่แนบมาด้านหลัง)

You are familiar to travelling within Thammasat University (TU), Rangsit campus. But your interlocutor is new to TU. Wait for his call and give directions to different locations on Rangsit campus by referring to the map below.



#### กิจกรรมที่ 3: จงอ่านประโยคและคำที่กำหนดให้ (ให้ท่านอ่านตามความเร็วปกติเหมือนเวลาที่ท่านพูดโดยทั่วๆ ไป)

Activity 3: Read the following sentences and words (at a comfortable speaking rate). The vertical lines within sentences indicate the internal pauses. Please mark a short pause accordingly. (The rationale in doing this is to make sure that the extracted targets are from the same phonological environments).

#### 3.1 อ่านประโยคต่อไปนี้

- 1.โดยในระยะแรกแรก | จะมีลักษณะเปลี่ยนจาก | การใช้เครื่องมือแบบง่ายง่าย | มาใช้เครื่องจักรแทน |
- 1. do:i nai rá-yá ræk-ræk | tcà mi: lák-sà-nà plì:an tcà:k | ka:n tcʰá:i kʰrŵ:aŋ mw: bæp ŋâ:i ŋâ:i| ma: tcʰá:i kʰrŵ:aŋ tcàk tʰæ:n|
- 2. นิทรรศการนี้| จะแสดงให้เห็นวิวัฒนาการ| ของเครื่องมือเครื่องใช้ในการประกอบอาชีพ| ของคนในภูมิภาคตะวันตก| ตั้งแต่อดีตถึง ปัจจุบัน|
- 2. ni-tʰát-sà-ka:n ní:| tcà sà-dæŋ hâ:i hěn wí-wát-tʰá-na-ka:n| kʰɔ̃ŋ kʰrŵ:aŋ mw kʰrŵ:aŋ tcʰá:i nai ka:n prà-kòp ʔ:-tcʰî:p| kʰɔ̃ŋ kʰon nai pʰu:m-mí-pʰâ:k tà-wan-tòk| tâŋ tæ ʔà-dì:t tʰṁŋ pàt-tcù-ban|
- 3.สิบเอ็คจำนวนบัณฑิต| ที่มีความคิดเห็นว่า| วิชาภาษาอังกฤษที่เคยเรียน| ไม่เพียงพอในการประกอบอาชีพในปัจจุบัน| มีสัดส่วน ใกล้เคียงกับจำนวนบัณฑิต| ที่มีความคิดเห็นว่า| วิชาภาษาอังกฤษที่เรียนมาเพียงพอ|
- 3. sìp-ʔèt team nu:an ban dìt|  $t^h$ î: mi:  $k^h$ u:am  $k^h$ ít hěn wâ| wí-teha:  $p^h$ a-să ʔaŋ-krì:t  $t^h$ î:  $k^h$ əj ri:an| mâi  $p^h$ i:aŋ  $p^h$ ə nai ka:n prà-kòb ʔ:-tehî:p nai pàt-teù-ban| mi: sàt-suàn klâi- $k^h$ i:aŋ kàp team nu:an ban dìt|  $t^h$ î: mi:  $k^h$ u:am  $k^h$ ít hěn wâ | wí-teha:  $p^h$ a-să ʔaŋ-krì:t  $t^h$ î: ria:n ma  $p^h$ i:ang  $p^h$ ə|
- 4. เพราะไม่ใช่หน้าที่อะไรสักหน่อย
- 4. phró? mâi tchâi nâ: thî: ?á rai sàk nòi
- 5. และมีการหมุนเวียนสมาชิก| ในกลุ่มแต่ละกลุ่ม| ทุกทุกสามอาทิตย์| เพื่อให้ผู้เรียนมีโอกาสที่จะทำงาน| ร่วมกับผู้เรียนอื่นอื่น|
- 5. læʔ mi: ka:n mǔn wi:an sà-ma:-teʰík| nai klùm tæ làʔ klùm| tʰúk tʰúk sǎm ʔa:-tʰít  $|p^h\hat{u}:a$ ? hâ:i  $p^h\hat{u}:$  ri:an mi: o:-kàt tʰî: teà tʰam ŋa:n| rû:am kàp  $p^h\hat{u}:$  ri:an ʔẁn ʔẁn|

6.ความก้าวหน้าในทางเศรษฐกิจ| ที่ดำเนินไปทุกวันนี้| แท้จริงนั้น| เป็นลักษณะปลาใหญ่กินปลาเล็ก|

- 6. kʰwu:am kâ:w nâ: nai tʰa:ŋ sèt-tʰà-kìt| tʰî: dam nə:n pai tʰúk wan ní:| tʰǽ tɕɪŋ nán| pɛn lák-sà-nà| pla: yài kin pla: lɛ́k|
- 7.โดยในระยะแรกแรก| จะมีลักษณะเปลี่ยนจาก| การใช้เครื่องมือแบบง่ายง่าย| มาใช้เครื่องจักรแทน|
- 7. do:i nai rá-yá ræk-ræk | tcà mi: lák-sà-nà plì:an tcà:k | ka:n tchá:i khrŵ:aŋ mw: bæp ŋâ:i ŋâ:i| ma: tchá:i khrŵ:aŋ tcàk thæ:n|
- 8.นิทรรศการนี้ | จะแสดงให้เห็นวิวัฒนาการ | ของเครื่องมือเครื่องใช้ในการประกอบอาชีพ | ของคนในภูมิภาคตะวันตก | ตั้งแต่อดีตถึง ปัจจุบัน |
- 8. ni-t<sup>h</sup>át-sà-ka:n ní:| tcà sà-dæŋ hâ:i h**ěn wí-wát-t<sup>h</sup>á-na-ka:n**| k<sup>h</sup>ŏŋ k<sup>h</sup>rŵ:aŋ mw k<sup>h</sup>rŵ:aŋ tc<sup>h</sup>á:i nai ka:n prà-kòp ?:-tc<sup>h</sup>î:p| k<sup>h</sup>ŏŋ k<sup>h</sup>on nai p<sup>h</sup>u:m-mí-p<sup>h</sup>â:k tà-wan-tòk| tâŋ tæ ʔà-dì:t t<sup>h</sup>ឃŋ pàt-tcù-ban|
- 9.สิบเอ็คจำนวนบัณฑิต| ที่มีความคิดเห็นว่า| วิชาภาษาอังกฤษที่เคยเรียน| ไม่เพียงพอในการประกอบอาชีพในปัจจุบัน| มีสัคส่วน ใกล้เกียงกับจำนวนบัณฑิต| ที่มีความคิดเห็นว่า| วิชาภาษาอังกฤษที่เรียนมาเพียงพอ|
- 9. sìp-ʔèt team nu:an ban dìt|  $t^h$ î: mi:  $k^h$ u:am  $k^h$ ít hěn wâ| wí-teha:  $p^h$ a-să ʔaŋ-krì:t  $t^h$ î:  $k^h$ əj ri:an| mâi  $p^h$ i:aŋ  $p^h$ ə nai ka:n prà-kòb ʔ:-tehî:p nai pàt-teù-ban| mi: sàt-suàn klâi- $k^h$ i:aŋ kàp team nu:an ban dìt|  $t^h$ î: mi:  $k^h$ u:am  $k^h$ ít hěn wâ | wí-teha:  $p^h$ a-să ʔaŋ-krì:t  $t^h$ î: ria:n ma  $p^h$ i:ang  $p^h$ ə|
- 10. เพราะไม่ใช่หน้าที่อะไรสักหน่อย
- 10. phró? mâi tehâi nâ: thì: ?á rai sàk nòi
- 11. และมีการหมุนเวียนสมาชิก| ในกลุ่มแต่ละกลุ่ม| ทุกทุกสามอาทิตย์| เพื่อให้ผู้เรียนมีโอกาสที่จะทำงาน| ร่วมกับผู้เรียนอื่นอื่น|
- 11. læ? mi: ka:n mǔn wi:an sà-ma:-tcʰík| nai klùm tæ̀ là? klùm| tʰúk tʰúk sǎm ʔa:-tʰít |pʰŵ:aʔ hâ:i pʰû: ri:an mi: o:-kàt tʰî: tcà tʰam ŋa:n| rû:am kàp pʰû: ri:an ʔẁn ʔẁn|
- 12.ความก้าวหน้าในทางเศรษฐกิจ ที่ดำเนินไปทุกวันนี้ แท้จริงนั้น เป็นลักษณะปลาใหญ่กินปลาเล็ก
- 12. kʰwu:am kâ:w nâ: nai tʰa:ŋ sèt-tʰà-kìt| tʰî: dam nə:n pai tʰúk wan ní:| tʰǽ tcɪŋ nán| pɛn lák-sà-nà| pla: yài kin pla: lék|

- 13.โดยในระยะแรกแรก จะมีลักษณะเปลี่ยนจาก การใช้เครื่องมือแบบง่ายง่าย มาใช้เครื่องจักรแทน
- 13. do:i nai rá-yá ræk-ræk | teà mi: lák-sà-nà plì:an teà:k | ka:n teʰá:i kʰrŵ:aŋ mw: bæp ŋâ:i ŋâ:i| ma: teʰá:i kʰrŵ:aŋ teàk tʰæ:n|
- 14.นิทรรศการนี้| จะแสดงให้เห็นวิวัฒนาการ| ของเครื่องมือเครื่องใช้ในการประกอบอาชีพ| ของคนในภูมิภาคตะวันตก| ตั้งแต่อดีต ถึงปัจจุบัน|
- 14. ni-thát-sà-ka:n ní:| tcà sà-dæŋ hâ:i hẽn wí-wát-thá-na-ka:n| khōŋ khrŵ:aŋ mw khrŵ:aŋ tchá:i nai ka:n prà-kòp ?:-tchî:p| khōŋ khon nai phu:m-mí-phâ:k tà-wan-tòk| tâŋ tæ ʔà-dì:t thửŋ pàt-tcù-ban|
- 15.สิบเอ็ดจำนวนบัณฑิต| ที่มีความคิดเห็นว่า| วิชาภาษาอังกฤษที่เคยเรียน| ไม่เพียงพอในการประกอบอาชีพในปัจจุบัน| มีสัดส่วน ใกล้เคียงกับจำนวนบัณฑิต| ที่มีความคิดเห็นว่า| วิชาภาษาอังกฤษที่เรียนมาเพียงพอ|
- 15. sìp-?èt team nu:an ban dìt|  $t^h$ î: mi:  $k^h$ u:am  $k^h$ ít hěn wâ| wí-teha:  $p^h$ a-să ?aŋ-krì:t  $t^h$ î:  $k^h$ əj ri:an| mâi  $p^h$ i:aŋ  $p^h$ ə nai ka:n prà-kòb ?:-tehî:p nai pàt-teù-ban| mi: sàt-suàn klâi- $k^h$ i:aŋ kàp team nu:an ban dìt|  $t^h$ î: mi:  $k^h$ u:am  $k^h$ ít hěn wâ | wí-teha:  $p^h$ a-să ?aŋ-krì:t  $t^h$ î: ria:n ma  $p^h$ i:ang  $p^h$ ə|

16.เพราะไม่ใช่หน้าที่อะไรสักหน่อย

- 16. phró? mâi tehâi nâ: thì: ?á rai sàk nòi
- 17. และมีการหมุนเวียนสมาชิก| ในกลุ่มแต่ละกลุ่ม| ทุกทุกสามอาทิตย์| เพื่อให้ผู้เรียนมีโอกาสที่จะทำงาน| ร่วมกับผู้เรียนอื่นอื่น|
- 17. læ? mi: ka:n mǔn wi:an sà-ma:-tcʰík| nai klùm tæ là? klùm| tʰúk tʰúk sǎm ?a:-tʰít |pʰŵ:a? hâ:i pʰû: ri:an mi: o:-kàt tʰî: tcà tʰam ŋa:n| rû:am kàp pʰû: ri:an ?win| ?win|
- 18.ความก้าวหน้าในทางเศรษฐกิจ ที่คำเนินไปทุกวันนี้ แท้จริงนั้น เป็นลักษณะปลาใหญ่กินปลาเล็ก
- 18. k<sup>h</sup>wu:am kâ:w nâ: nai t<sup>h</sup>a:ŋ sèt-t<sup>h</sup>à-kìt| t<sup>h</sup>î: dam nə:n pai t<sup>h</sup>úk wan ní:| t<sup>h</sup>é tcɪŋ nán| pɛn lák-sà-nà| pla: yài kin pla: lék|
- 19. โดยในระยะแรกแรก | จะมีลักษณะเปลี่ยนจาก | การใช้เครื่องมือแบบง่ายง่าย | มาใช้เครื่องจักรแทน |
- 19. do:i nai rá-yá ræk-ræk | tcà mi: lák-sà-nà plì:an tcà:k | ka:n tcʰá:i kʰrŵ:aŋ mw: bæp ŋâ:i ŋâ:i| ma: tcʰá:i kʰrŵ:aŋ tcàk tʰæ:n|

- 20. นิทรรศการนี้| จะแสดงให้เห็นวิวัฒนาการ| ของเครื่องมือเครื่องใช้ในการประกอบอาชีพ| ของคนในภูมิภาคตะวันตก| ตั้งแต่ อดีตถึงปัจจุบัน|
- 20. ni-thát-sà-ka:n ní:| tcà sà-dæŋ hâ:i hẽn wí-wát-thá-na-ka:n| khōŋ khrŵ:aŋ mw khrŵ:aŋ tchá:i nai ka:n prà-kòp ?:-tchî:p| khōŋ khon nai phu:m-mí-phâ:k tà-wan-tòk| tâŋ tæ ʔà-dì:t thửŋ pàt-tcù-ban|
- 21.สิบเอ็ดจำนวนบัณฑิต| ที่มีความคิดเห็นว่า| วิชาภาษาอังกฤษที่เกยเรียน| ไม่เพียงพอในการประกอบอาชีพในปัจจุบัน| มีสัดส่วน ใกล้เกียงกับจำนวนบัณฑิต| ที่มีความคิดเห็นว่า| วิชาภาษาอังกฤษที่เรียนมาเพียงพอ|
- 21. sìp-?èt team nu:an ban dìt| t<sup>h</sup>î: mi: k<sup>h</sup>u:am k<sup>h</sup>ít hěn wâ| wí-te<sup>h</sup>a: p<sup>h</sup>a-să ʔaŋ-krì:t t<sup>h</sup>î: k<sup>h</sup>əj ri:an| mâi p<sup>h</sup>i:aŋ p<sup>h</sup>ə nai ka:n prà-kòb ʔ:-te<sup>h</sup>î:p nai pàt-teù-ban| mi: sàt-suàn klâi-k<sup>h</sup>i:aŋ kàp team nu:an ban dìt| t<sup>h</sup>î: mi: k<sup>h</sup>u:am k<sup>h</sup>ít hěn wâ | wí-te<sup>h</sup>a: p<sup>h</sup>a-să ʔaŋ-krì:t t<sup>h</sup>î: ria:n ma p<sup>h</sup>i:ang p<sup>h</sup>ə|
- 22. เพราะไม่ใช่หน้าที่อะไรสักหน่อย
- 22. phró? mâi tchâi nâ: thî: ?á rai sàk nòi
- 23. และมีการหมุนเวียนสมาชิก| ในกลุ่มแต่ละกลุ่ม| ทุกทุกสามอาทิตย์| เพื่อให้ผู้เรียนมีโอกาสที่จะทำงาน| ร่วมกับผู้เรียนอื่นอื่น|
- 23. læʔ mi: ka:n mǔn wi:an sà-ma:-teʰík| nai klùm tæ làʔ klùm| tʰúk tʰúk sǎm ʔa:-tʰít |pʰŵ:aʔ hâ:i pʰû: ri:an mi: o:-kàt tʰî: teà tʰam ŋa:n| rû:am kàp pʰû: ri:an ʔẁn ʔẁn|
- 24.ความก้าวหน้าในทางเศรษฐกิจ ที่คำเนินไปทุกวันนี้ แท้จริงนั้น เป็นลักษณะปลาใหญ่กินปลาเล็ก
- 24. k<sup>h</sup>wu:am kâ:w nâ: nai t<sup>h</sup>a:ŋ sèt-t<sup>h</sup>à-kìt| t<sup>h</sup>î: dam nə:n pai t<sup>h</sup>úk wan ní:| t<sup>h</sup>é teŋ nán| pɛn lák-sà-nà| pla: yài kin pla: lék|
- 25. โดยในระยะแรกแรก| จะมีลักษณะเปลี่ยนจาก| การใช้เครื่องมือแบบง่ายง่าย| มาใช้เครื่องจักรแทน|
- 25. do:i nai rá-yá ræk-ræk | tcà mi: lák-sà-nà plì:an tcà:k | ka:n tcʰá:i kʰrŵ:aŋ mw: bæp ŋâ:i ŋâ:i| ma: tcʰá:i kʰrŵ:aŋ tcàk tʰæ:n|
- 26. นิทรรศการนี้| จะแสดงให้เห็นวิวัฒนาการ| ของเครื่องมือเครื่องใช้ในการประกอบอาชีพ| ของคนในภูมิภาคตะวันตก| ตั้งแต่ อดีตถึงปัจจุบัน|
- 26. ni-tʰát-sà-ka:n ní:| tcà sà-dæŋ hâ:i hǐn wí-wát-tʰá-na-ka:n| kʰɔ̃ŋ kʰrŵ:aŋ mw kʰrŵ:aŋ tcʰá:i nai ka:n prà-kɔ̀p ʔ:-tcʰî:p| kʰɔ̃ŋ kʰon nai pʰu:m-mí-pʰâ:k tà-wan-tòk| tâŋ tæ̀ ʔà-dì:t tʰwॅŋ pàt-tcù-ban|

- 27.สิบเอ็คจำนวนบัณฑิต| ที่มีความคิดเห็นว่า| วิชาภาษาอังกฤษที่เคยเรียน| ไม่เพียงพอในการประกอบอาชีพในปัจจุบัน| มีสัดส่วน ใกล้เคียงกับจำนวนบัณฑิต| ที่มีความคิดเห็นว่า| วิชาภาษาอังกฤษที่เรียนมาเพียงพอ|
- 27. sìp-?èt team nu:an ban dìt|  $t^h$ î: mi:  $k^h$ u:am  $k^h$ ít hěn wâ| wí-teha:  $p^h$ a-să ?aŋ-krì:t  $t^h$ î:  $k^h$ əj ri:an| mâi  $p^h$ i:aŋ  $p^h$ ə nai ka:n prà-kòb ?:-tehî:p nai pàt-teù-ban| mi: sàt-suàn klâi- $k^h$ i:aŋ kàp team nu:an ban dìt|  $t^h$ î: mi:  $k^h$ u:am  $k^h$ ít hěn wâ | wí-teha:  $p^h$ a-să ?aŋ-krì:t  $t^h$ î: ria:n ma  $p^h$ i:ang  $p^h$ ə|
- 28. เพราะไม่ใช่หน้าที่อะไรสักหน่อย
- 28. phró? mâi tchâi nâ: thì: ?á rai sàk nòi
- 29. และมีการหมุนเวียนสมาชิก| ในกลุ่มแต่ละกลุ่ม| ทุกทุกสามอาทิตย์| เพื่อให้ผู้เรียนมีโอกาสที่จะทำงาน| ร่วมกับผู้เรียนอื่นอื่น|
- 29. læ? mi: ka:n mǔn wi:an sà-ma:-tcʰík| nai klùm tæ̀ là? klùm| tʰúk tʰúk sǎm ʔa:-tʰít |pʰŵ:a? hâ:i pʰû: ri:an mi: o:-kàt tʰî: tcà tʰam ŋa:n| rû:am kàp pʰû: ri:an ʔẁn ʔẁn|
- 30.ความก้าวหน้าในทางเศรษฐกิจ| ที่คำเนินไปทุกวันนี้| แท้จริงนั้น| เป็นลักษณะปลาใหญ่กินปลาเล็ก|
- 30. kʰwu:am kâ:w nâ: nai tʰa:ŋ sèt-tʰà-kìt| tʰî: dam nə:n pai tʰúk wan ní:| tʰǽ teɪŋ nán| pɛn lák-sà-nà| pla: yài kin pla: lɛ́k|
- 31. โดยในระยะแรกแรก | จะมีลักษณะเปลี่ยนจาก | การใช้เครื่องมือแบบง่ายง่าย | มาใช้เครื่องจักรแทน |
- 31. do:i nai rá-yá ræk-ræk | teà mi: lák-sà-nà plì:an teà:k | ka:n teʰá:i kʰrŵ:aŋ mw: bæp ŋâ:i ฏâ:i| ma: teʰá:i kʰrŵ:aŋ teàk tʰæ:n|
- 32.นิทรรศการนี้| จะแสดงให้เห็นวิวัฒนาการ| ของเครื่องมือเครื่องใช้ในการประกอบอาชีพ| ของคนในภูมิภาคตะวันตก| ตั้งแต่อดีต ถึงปัจจุบัน|
- 32. ni-tʰát-sà-ka:n ní:| tcà sà-dæŋ hâ:i hǐn wí-wát-tʰá-na-ka:n| kʰŏŋ kʰrŵ:aŋ mw kʰrŵ:aŋ tcʰá:i nai ka:n prà-kòp ʔ:-tcʰî:p| kʰŏŋ kʰon nai pʰu:m-mí-pʰâ:k tà-wan-tòk| tâŋ tæ ʔà-dì:t tʰṁŋ pàt-tcù-ban|
- 33.สิบเอ็ดจำนวนบัณฑิต| ที่มีความคิดเห็นว่า| วิชาภาษาอังกฤษที่เคยเรียน| ไม่เพียงพอในการประกอบอาชีพในปัจจุบัน| มีสัดส่วน ใกล้เคียงกับจำนวนบัณฑิต| ที่มีความคิดเห็นว่า| วิชาภาษาอังกฤษที่เรียนมาเพียงพอ|
- 33. sìp-?èt team nu:an ban dìt| t<sup>h</sup>î: mi: k<sup>h</sup>u:am k<sup>h</sup>ít hěn wâ| wí-te<sup>h</sup>a: p<sup>h</sup>a-să ?aŋ-krì:t t<sup>h</sup>î: k<sup>h</sup>əj ri:an| mâi p<sup>h</sup>i:aŋ p<sup>h</sup>ə nai ka:n prà-kòb ?:-te<sup>h</sup>î:p nai pàt-teù-ban| mi: sàt-suàn

klâi- $k^h$ i:aŋ kàp team nu:an ban dìt|  $t^h$ î: mi:  $k^h$ u:am  $k^h$ ít hěn wâ | wí-teha:  $p^h$ a-sǎ ʔaŋ-krì:t  $t^h$ î: ria:n ma  $p^h$ i:ang  $p^h$ ɔ|

34. เพราะไม่ใช่หน้าที่อะไรสักหน่อย

34. phró? mâi tehâi nâ: thì: ?á rai sàk nòi

35.และมีการหมุนเวียนสมาชิก| ในกลุ่มแต่ละกลุ่ม| ทุกทุกสามอาทิตย์| เพื่อให้ผู้เรียนมีโอกาสที่จะทำงาน| ร่วมกับผู้เรียนอื่นอื่น|

35. læ? mi: ka:n mǔn wi:an sà-ma:-teʰík| nai klùm tæ là? klùm| tʰúk tʰúk sǎm ?a:-tʰít |pʰŵ:a? hâ:i pʰû: ri:an mi: o:-kàt tʰî: teà tʰam ŋa:n| rû:am kàp pʰû: ri:an ?ẁn ?ẁn|

36.ความก้าวหน้าในทางเศรษฐกิจ| ที่ดำเนินไปทุกวันนี้| แท้จริงนั้น| เป็นลักษณะปลาใหญ่กินปลาเล็ก|

36. k<sup>h</sup>wu:am kâ:w nâ: nai t<sup>h</sup>a:ŋ sèt-t<sup>h</sup>à-kìt| t<sup>h</sup>î: dam nə:n pai t<sup>h</sup>úk wan ní:| t<sup>h</sup>é tcɪŋ nán| pɛn lák-sà-nà| pla: yài kin pla: lék|

#### 3.2 อ่านคำต่อไปนี้ (ห่างกันประมาณ 2 วินาที)

- 1. เรียน เรียน เรียน เรียน เรียน เรียน [ri:an mid] V. "to study"
- 2. ที่ ที่ ที่ ที่ ที่ ที่ [thi: HL] Prep. "at"
- 3. nis nis nis nis nis [ka:n mid] Prefix "-ness"
- 4. lu lu lu lu lu [nai mid] Prep "in"
- 5. คี คี คี คี คี คี **[di: mid]** ADJ "good"

#### กิจกรรมที่ 1: สนทนาแลกเปลี่ยนข้อมูล

#### **Activity 1: Information-exchange task**

#### ท่านคือผู้พูด B

#### You are Speaker B

ด้านหลังกระดาษนี้เป็นเอกสารที่ส่งมาทางแฟกซ์ แต่เนื่องจากเครื่องรับ-ส่งแฟกซ์มีคุณภาพไม่คื ทำให้ข้อมูลบางส่วนในเอกสารนี้ ไม่ชัดเจน คู่สนทนาของท่าน (ผู้พูด A) ก็ได้รับแฟกซ์ฉบับนี้เช่นกัน ซึ่งแฟกซ์ของผู้พูด A อาจมีคุณภาพชัดเจนกว่าของท่านหรือไม่ ก็ได้

ให้ท่านรอรับโทรศัพท์จากผู้พูด  $\mathbf A$  ให้สอบถามข้อมูลที่ขาดหายไป และเขียนข้อมูลลงบนกระดาษที่ให้มาด้านหลัง

On the back of this document is a facsimile listing many fresh food products and their prices. Unfortunately, the fax is not in good quality, so some information is illegible for you but might be legible for your interlocutor (Speaker A). Wait for a call from Speaker A and exchange the information that is obfuscated to you. Then, write down the missing information on the same piece of paper provided.

<sup>\*</sup>หมายเหตุ โปรดใช้สรรพนามว่า  ${f B}$  แทนตัวท่าน และ  ${f A}$  แทนคู่สนทนา และวางโทรศัพท์หลังจากทำกิจกรรมนี้เสร็จสิ้น

<sup>\*</sup>Please identify yourself as B and your interlocutor as A. Please hang up the phone after finishing this task.

Fax 1B

### ราคาขายปลีกลินค้า หมวดอาหารสดบางชนิดในกรุงเทพมหานคร

# ประจำวันที่ 27 พฤศจิกายน 2554

รายการ	หน่วย	6 กิ.ย. 54	27 พ.ย. 54
สุกรชำแหละ			
• เนื้อแดง (ตะโพก)	กก.	115.00-120.00	116.00-122.00
• • • • • • • • • • • • • • • • • • • •	22.	115 00 120 00	110 75 104 00
มะนาว (เบอร์ 1-2) ผลละ	ผล	2.25-2.50	2.25-2.50
พริกขี้หนู (จินคา)	ขีด	8.00-9.00	7.00-8.00
เนื้อโค สันนอก/สะโพก	กก.	150.00-160.00	162.00-165.00
มะนาว(เบอร์ 3-4) ผลละ	ผล	1.75-2.00	1.00-2.00
ไก่สดทั้งตัว (ไม่รวมเครื่องใน)	กก.	70.00-72.00	69.00-70.00
20000 (20 00 %2/22)		100 00 100 00	100 00 000 00
ปลาทูนึ่ง (ขนาค 3 ตัว/เข่ง)	เข่ง	20.00-25.00	20.00-25.00
ปลานิล	กก.	70.00-80.00	70.00-75.00
2100000010001		22.22.22.22	25 00 00 00
ไข่เป็ด(กลาง)	ฟอง	4.90-5.00	4.90-5.00
ไข่ไก่(เบอร์ 2)	ฟอง	3.60-3.70	3.60-3.70
ไข่ไก่(เบอร์ 3)	ฟอง	3.50-3.60	3.50-3.60
คะน้ำ	กก.	28.00-30.00	29.00-30.00
ผักบุ้งจิน	กก.	20.00-22.00	21.00-22.00
	22.	12.00.15.00	12.00.15.00
กะหล่ำปลี	กก.	12.00-15.00	14.00-15.00
กะหล่ำคอก	กก.	16.00-17.00	17.00-19.00
เห็ดเออรินจิน	ขีด	29.00-30.00	27.00-29.00
ist and a state of	***	20 00 20 00	27 00 00 00
สาลี่ทอง	กก.	48.00-50.00	49.00-52.00
สาลี่แนชชื่	<b>ሰ</b> ስ.	50.00-55.00	52.00-57.00

#### กิจกรรมที่ 2: บอกทิศทาง

#### **Activity 2: Giving Directions**

ท่านไม่คุ้นเคยกับการเดินทางภายในมหาวิทยาลัยธรรมศาสตร์ รังสิต ขณะนี้**ท่านอยู่ที่ โรงพิมพ์มหาวิทยาลัยธรรมศาสตร์ (ตึกเลขที่** 

- 9) ท่านต้องการเดินทางไปยังสถานที่ต่างๆ ดังนี้
- 1. สถานีวิทยุกระจายเสียง มหาวิทยาลัยธรรมศาสตร์ (ตึกเลขที่ 50)
- 2. อาคารยิมเนเซียม 2 (ตึกเลขที่ 37)
- 3. สถานีบริการน้ำมัน ป.ต.ท. (ตึกเลขที่ 59)

เนื่องจากท่านไม่รู้ทางที่จะไปยังสถานที่ดังกล่าว ท่านจึงต่อโทรศัพท์หาเพื่อนตามหมายเลข 2358 เพื่อถามทิศทางไปยังสถานที่ ต่างๆ ข้างต้น

จงวาดแผนที่คร่าวๆ ตามคำบอกเล่าลงในกระดาษที่แนบมาด้านหลัง

Since you are not familiar with the directions in TU, Rangsit campus and now you are at Thammasat University's printing house (Building No.9), you would like to go to the following places.

- 1.TU Radio Station (Building No. 50)
- 2. Gymnasium 2 (Building No. 37)
- 3. PTT gas station (Building No. 59)

Please call your friend (Speaker A) by dialing ext. 2358 to ask for the directions. Please also draw a map of the above places in the space provided below.

#### กิจกรรมที่ 3: จงอ่านประโยคและคำที่กำหนดให้ (ให้ท่านอ่านตามความเร็วปกติเหมือนเวลาที่ท่านพคโดยทั่วๆ ไป)

Activity 3: Read the following sentences and words (at a comfortable speaking rate). The vertical lines within sentences indicate the internal pauses. Please give a short pause accordingly. (The rational in doing this is to make sure that the extracted targets are from the same phonological environments).

#### 3.1 อ่านประโยคต่อไปนี้

- 1.โดยในระยะแรกแรก | จะมีลักษณะเปลี่ยนจาก | การใช้เครื่องมือแบบง่ายง่าย | มาใช้เครื่องจักรแทน |
- 1. do:i nai rá-yá ræk-ræk | tcà mi: lák-sà-nà plì:an tcà:k | ka:n tcʰá:i kʰrŵ:aŋ mw: bæp ŋâ:i ฏâ:i| ma: tcʰá:i kʰrŵ:aŋ tcàk tʰæ:n|
- นิทรรศการนี้ | จะแสดงให้เห็นวิวัฒนาการ | ของเครื่องมือเครื่องใช้ในการประกอบอาชีพ | ของคนในภูมิภาคตะวันตก | ตั้งแต่อดีตถึง ปัจจุบัน |
- 2. ni-thát-sà-ka:n ní:| tcà sà-dæŋ hâ:i hěn wí-wát-thá-na-ka:n| khōŋ khrŵ:aŋ mw khrŵ:aŋ tchá:i nai ka:n prà-kòp ?:-tchî:p| khōŋ khon nai phu:m-mí-phâ:k tà-wan-tòk| tâŋ tæ ʔà-dì:t thwŋ pàt-tcù-ban|
- 3.สิบเอ็ดจำนวนบัณฑิต| ที่มีความคิดเห็นว่า| วิชาภาษาอังกฤษที่เคยเรียน| ไม่เพียงพอในการประกอบอาชีพในปัจจุบัน| มีสัดส่วน ใกล้เคียงกับจำนวนบัณฑิต| ที่มีความคิดเห็นว่า| วิชาภาษาอังกฤษที่เรียนมาเพียงพอ|
- 3. sìp-ʔèt team nu:an ban dìt|  $t^h$ î: mi:  $k^h$ u:am  $k^h$ ít hěn wâ| wí-teha:  $p^h$ a-să ʔaŋ-krì:t  $t^h$ î:  $k^h$ əj ri:an| mâi  $p^h$ i:aŋ  $p^h$ ə nai ka:n prà-kòb ʔ:-tehî:p nai pàt-teù-ban| mi: sàt-suàn klâi- $k^h$ i:aŋ kàp team nu:an ban dìt|  $t^h$ î: mi:  $k^h$ u:am  $k^h$ ít hěn wâ | wí-teha:  $p^h$ a-să ʔaŋ-krì:t  $t^h$ î: ria:n ma  $p^h$ i:ang  $p^h$ ə|
- 4. เพราะไม่ใช่หน้าที่อะไรสักหน่อย
- 4. phró? mâi tchâi nâ: thî: ?á rai sàk nòi
- 5. และมีการหมนเวียนสมาชิก| ในกล่มแต่ละกล่ม| ทกทกสามอาทิตย์| เพื่อให้ผู้เรียนมีโอกาสที่จะทำงาน| ร่วมกับผู้เรียนอื่นอื่น|
- 5. læʔ mi: ka:n mǔn wi:an sà-ma:-teʰík| nai klùm tæ làʔ klùm| tʰúk tʰúk sǎm ʔa:-tʰít  $|p^h\hat{u}:a\rangle$  hâ:i  $p^h\hat{u}:ri:an$  mi: o:-kàt tʰî: teà tʰam na:n| rû:am kàp  $p^h\hat{u}:ri:an$  ʔẁn ʔẁn|

6.ความก้าวหน้าในทางเศรษฐกิจ| ที่ดำเนินไปทุกวันนี้| แท้จริงนั้น| เป็นลักษณะปลาใหญ่กินปลาเล็ก|

- 6. kʰwu:am kâ:w nâ: nai tʰa:ŋ sèt-tʰà-kìt| tʰî: dam nə:n pai tʰúk wan ní:| tʰǽ tɕɪŋ nán| pɛn lák-sà-nà| pla: yài kin pla: lɛ́k|
- 7.โดยในระยะแรกแรก | จะมีลักษณะเปลี่ยนจาก | การใช้เครื่องมือแบบง่ายง่าย | มาใช้เครื่องจักรแทน |
- 7. do:i nai rá-yá ræk-ræk | tcà mi: lák-sà-nà plì:an tcà:k | ka:n tcʰá:i kʰrŵ:aŋ mw: bæp ŋâ:i ŋâ:i| ma: tcʰá:i kʰrŵ:aŋ tcàk tʰæ:n|
- 8.นิทรรศการนี้| จะแสดงให้เห็นวิวัฒนาการ| ของเครื่องมือเครื่องใช้ในการประกอบอาชีพ| ของคนในภูมิภาคตะวันตก| ตั้งแต่อดีตถึง ปัจจุบัน|
- 8. ni-t<sup>h</sup>át-sà-ka:n ní:| teà sà-dæŋ hâ:i hěn wí-wát-t<sup>h</sup>á-na-ka:n| k<sup>h</sup>ɔ̃ŋ k<sup>h</sup>rŵ:aŋ mw k<sup>h</sup>rŵ:aŋ te<sup>h</sup>á:i nai ka:n prà-kòp ?:-te<sup>h</sup>î:p| k<sup>h</sup>ɔ̃ŋ k<sup>h</sup>on nai p<sup>h</sup>u:m-mí-p<sup>h</sup>â:k tà-wan-tòk| tâŋ tæ̀ ʔà-dì:t t<sup>h</sup>ឃ்ŋ pàt-teù-ban|
- 9.สิบเอ็คจำนวนบัณฑิต| ที่มีความคิดเห็นว่า| วิชาภาษาอังกฤษที่เคยเรียน| ไม่เพียงพอในการประกอบอาชีพในปัจจุบัน| มีสัคส่วน ใกล้เกียงกับจำนวนบัณฑิต| ที่มีความคิดเห็นว่า| วิชาภาษาอังกฤษที่เรียนมาเพียงพอ|
- 9. sìp-ʔèt team nu:an ban dìt|  $t^h$ î: mi:  $k^h$ u:am  $k^h$ ít hěn wâ| wí-teha:  $p^h$ a-să ʔaŋ-krì:t  $t^h$ î:  $k^h$ əj ri:an| mâi  $p^h$ i:aŋ  $p^h$ ə nai ka:n prà-kòb ʔ:-tehî:p nai pàt-teù-ban| mi: sàt-suàn klâi- $k^h$ i:aŋ kàp team nu:an ban dìt|  $t^h$ î: mi:  $k^h$ u:am  $k^h$ ít hěn wâ | wí-teha:  $p^h$ a-să ʔaŋ-krì:t  $t^h$ î: ria:n ma  $p^h$ i:ang  $p^h$ ə|
- 10. เพราะไม่ใช่หน้าที่อะไรสักหน่อย
- 10. phró? mâi tchâi nâ: thì: ?á rai sàk nòi
- 11. และมีการหมุนเวียนสมาชิก| ในกลุ่มแต่ละกลุ่ม| ทุกทุกสามอาทิตย์| เพื่อให้ผู้เรียนมีโอกาสที่จะทำงาน| ร่วมกับผู้เรียนอื่นอื่น|
- 11. læ? mi: ka:n mǔn wi:an sà-ma:-tchík| nai klùm tæ là? klùm| thúk thúk săm ?a:-thít |phû:a? hâ:i phû: ri:an mi: o:-kàt thî: tcà tham ŋa:n| rû:am kàp phû: ri:an ?win ?win|
- 12.ความก้าวหน้าในทางเศรษฐกิจ ที่ดำเนินไปทุกวันนี้ แท้จริงนั้น เป็นลักษณะปลาใหญ่กินปลาเล็ก
- 12. kʰwu:am kâ:w nâ: nai tʰa:ŋ sèt-tʰà-kìt| tʰî: dam nə:n pai tʰúk wan ní:| tʰǽ tcɪŋ nán| pɛn lák-sà-nà| pla: yài kin pla: lék|

- 13.โดยในระยะแรกแรก| จะมีลักษณะเปลี่ยนจาก| การใช้เครื่องมือแบบง่ายง่าย| มาใช้เครื่องจักรแทน|
- 13. do:i nai rá-yá ræk-ræk | teà mi: lák-sà-nà plì:an teà:k | ka:n teʰá:i kʰrŵ:aŋ mw: bæp ŋâ:i ŋâ:i| ma: teʰá:i kʰrŵ:aŋ teàk tʰæ:n|
- 14.นิทรรศการนี้| จะแสดงให้เห็นวิวัฒนาการ| ของเครื่องมือเครื่องใช้ในการประกอบอาชีพ| ของคนในภูมิภาคตะวันตก| ตั้งแต่อดีต ถึงปัจจุบัน|
- 14. ni-thát-sà-ka:n ní:| tcà sà-dæŋ hâ:i hẽn wí-wát-thá-na-ka:n| khōŋ khrŵ:aŋ mw khrŵ:aŋ tchá:i nai ka:n prà-kòp ?:-tchî:p| khōŋ khon nai phu:m-mí-phâ:k tà-wan-tòk| tâŋ tæ ʔà-dì:t thửŋ pàt-tcù-ban|
- 15.สิบเอ็ดจำนวนบัณฑิต| ที่มีความคิดเห็นว่า| วิชาภาษาอังกฤษที่เคยเรียน| ไม่เพียงพอในการประกอบอาชีพในปัจจุบัน| มีสัดส่วน ใกล้เคียงกับจำนวนบัณฑิต| ที่มีความคิดเห็นว่า| วิชาภาษาอังกฤษที่เรียนมาเพียงพอ|
- 15. sìp-?èt team nu:an ban dìt|  $t^h$ î: mi:  $k^h$ u:am  $k^h$ ít hěn wâ| wí-teha:  $p^h$ a-să ?aŋ-krì:t  $t^h$ î:  $k^h$ əj ri:an| mâi  $p^h$ i:aŋ  $p^h$ ə nai ka:n prà-kòb ?:-tehî:p nai pàt-teù-ban| mi: sàt-suàn klâi- $k^h$ i:aŋ kàp team nu:an ban dìt|  $t^h$ î: mi:  $k^h$ u:am  $k^h$ ít hěn wâ | wí-teha:  $p^h$ a-să ?aŋ-krì:t  $t^h$ î: ria:n ma  $p^h$ i:ang  $p^h$ ə|

16.เพราะไม่ใช่หน้าที่อะไรสักหน่อย

- 16. phró? mâi tehâi nâ: thî: ?á rai sàk nòi
- 17. และมีการหมุนเวียนสมาชิก| ในกลุ่มแต่ละกลุ่ม| ทุกทุกสามอาทิตย์| เพื่อให้ผู้เรียนมีโอกาสที่จะทำงาน| ร่วมกับผู้เรียนอื่นอื่น|
- 17. læ? mi: ka:n mǔn wi:an sà-ma:-tcʰík| nai klùm tæ̀ là? klùm| tʰúk tʰúk sǎm ?a:-tʰít |pʰŵ:a? hâ:i pʰû: ri:an mi: o:-kàt tʰî: tcà tʰam ŋa:n| rû:am kàp pʰû: ri:an ?ẁn ?ẁn|
- 18.ความก้าวหน้าในทางเศรษฐกิจ| ที่ดำเนินไปทุกวันนี้| แท้จริงนั้น| เป็นลักษณะปลาใหญ่กินปลาเล็ก|
- 18. k<sup>h</sup>wu:am kâ:w nâ: nai t<sup>h</sup>a:ŋ sèt-t<sup>h</sup>à-kìt| t<sup>h</sup>î: dam nə:n pai t<sup>h</sup>úk wan ní:| t<sup>h</sup>é tcɪŋ nán| pɛn lák-sà-nà| pla: yài kin pla: lék|
- 19. โดยในระยะแรกแรก | จะมีลักษณะเปลี่ยนจาก | การใช้เครื่องมือแบบง่ายง่าย | มาใช้เครื่องจักรแทน |
- 19. do:i nai rá-yá ræk-ræk | tcà mi: lák-sà-nà plì:an tcà:k | ka:n tcʰá:i kʰrŵ:aŋ mw: bæp ηâ:i ηâ:i| ma: tcʰá:i kʰrŵ:aŋ tcàk tʰæ:n|

- 20. นิทรรศการนี้| จะแสดงให้เห็นวิวัฒนาการ| ของเครื่องมือเครื่องใช้ในการประกอบอาชีพ| ของคนในภูมิภาคตะวันตก| ตั้งแต่ อดีตถึงปัจจุบัน|
- 20. ni-thát-sà-ka:n ní:| tcà sà-dæŋ hâ:i hẽn wí-wát-thá-na-ka:n| khōŋ khrŵ:aŋ mw khrŵ:aŋ tchá:i nai ka:n prà-kòp ?:-tchî:p| khōŋ khon nai phu:m-mí-phâ:k tà-wan-tòk| tâŋ tæ ʔà-dì:t thửŋ pàt-tcù-ban|
- 21.สิบเอ็คจำนวนบัณฑิต| ที่มีความคิดเห็นว่า| วิชาภาษาอังกฤษที่เคยเรียน| ไม่เพียงพอในการประกอบอาชีพในปัจจุบัน| มีสัดส่วน ใกล้เคียงกับจำนวนบัณฑิต| ที่มีความคิดเห็นว่า| วิชาภาษาอังกฤษที่เรียนมาเพียงพอ|
- 21. sìp-?èt team nu:an ban dìt| t<sup>h</sup>î: mi: k<sup>h</sup>u:am k<sup>h</sup>ít hěn wâ| wí-te<sup>h</sup>a: p<sup>h</sup>a-să ʔaŋ-krì:t t<sup>h</sup>î: k<sup>h</sup>əj ri:an| mâi p<sup>h</sup>i:aŋ p<sup>h</sup>ə nai ka:n prà-kòb ʔ:-te<sup>h</sup>î:p nai pàt-teù-ban| mi: sàt-suàn klâi-k<sup>h</sup>i:aŋ kàp team nu:an ban dìt| t<sup>h</sup>î: mi: k<sup>h</sup>u:am k<sup>h</sup>ít hěn wâ | wí-te<sup>h</sup>a: p<sup>h</sup>a-să ʔaŋ-krì:t t<sup>h</sup>î: ria:n ma p<sup>h</sup>i:ang p<sup>h</sup>ə|
- 22. เพราะไม่ใช่หน้าที่อะไรสักหน่อย
- 22. phró? mâi tchâi nâ: thî: ?á rai sàk nòi
- 23. และมีการหมุนเวียนสมาชิก| ในกลุ่มแต่ละกลุ่ม| ทุกทุกสามอาทิตย์| เพื่อให้ผู้เรียนมีโอกาสที่จะทำงาน| ร่วมกับผู้เรียนอื่นอื่น|
- 23. læʔ mi: ka:n mǔn wi:an sà-ma:-teʰík| nai klùm tæ làʔ klùm| tʰúk tʰúk sǎm ʔa:-tʰít |pʰŵ:aʔ hâ:i pʰû: ri:an mi: o:-kàt tʰî: teà tʰam ŋa:n| rû:am kàp pʰû: ri:an ʔẁn ʔẁn|
- 24.ความก้าวหน้าในทางเศรษฐกิจ ที่ดำเนินไปทุกวันนี้ แท้จริงนั้น เป็นลักษณะปลาใหญ่กินปลาเล็ก
- 24. k<sup>h</sup>wu:am kâ:w nâ: nai t<sup>h</sup>a:ŋ sèt-t<sup>h</sup>à-kìt| t<sup>h</sup>î: dam nə:n pai t<sup>h</sup>úk wan ní:| t<sup>h</sup>é teŋ nán| pen lák-sà-nà| pla: yài kin pla: lék|
- 25.โดยในระยะแรกแรก| จะมีลักษณะเปลี่ยนจาก| การใช้เครื่องมือแบบง่ายง่าย| มาใช้เครื่องจักรแทน|
- 25. do:i nai rá-yá ræk-ræk | tcà mi: lák-sà-nà plì:an tcà:k | ka:n tcʰá:i kʰrŵ:aŋ mw: bæp ŋâ:i ŋâ:i| ma: tcʰá:i kʰrŵ:aŋ tcàk tʰæ:n|
- 26. นิทรรศการนี้| จะแสดงให้เห็นวิวัฒนาการ| ของเครื่องมือเครื่องใช้ในการประกอบอาชีพ| ของคนในภูมิภาคตะวันตก| ตั้งแต่ อดีตถึงปัจจุบัน|
- 26. ni-thát-sà-ka:n ní:| teà sà-dæŋ hâ:i hẽn wí-wát-thá-na-ka:n| khỏŋ khrŵ:aŋ mw khrŵ:aŋ tehá:i nai ka:n prà-kòp ?:-tehî:p| khỏŋ khon nai phu:m-mí-phâ:k tà-wan-tòk| tâŋ tæ ʔà-dì:t thửŋ pàt-teù-ban|

- 27.สิบเอ็คจำนวนบัณฑิต| ที่มีความคิดเห็นว่า| วิชาภาษาอังกฤษที่เคยเรียน| ไม่เพียงพอในการประกอบอาชีพในปัจจุบัน| มีสัดส่วน ใกล้เคียงกับจำนวนบัณฑิต| ที่มีความคิดเห็นว่า| วิชาภาษาอังกฤษที่เรียนมาเพียงพอ|
- 27. sìp-?èt team nu:an ban dìt|  $t^h$ î: mi:  $k^h$ u:am  $k^h$ ít hěn wâ| wí-teha:  $p^h$ a-să ?aŋ-krì:t  $t^h$ î:  $k^h$ əj ri:an| mâi  $p^h$ i:aŋ  $p^h$ ə nai ka:n prà-kòb ?:-tehî:p nai pàt-teù-ban| mi: sàt-suàn klâi- $k^h$ i:aŋ kàp team nu:an ban dìt|  $t^h$ î: mi:  $k^h$ u:am  $k^h$ ít hěn wâ | wí-teha:  $p^h$ a-să ?aŋ-krì:t  $t^h$ î: ria:n ma  $p^h$ i:ang  $p^h$ ə|
- 28. เพราะไม่ใช่หน้าที่อะไรสักหน่อย
- 28. phró? mâi tchâi nâ: thì: ?á rai sàk nòi
- 29. และมีการหมุนเวียนสมาชิก| ในกลุ่มแต่ละกลุ่ม| ทุกทุกสามอาทิตย์| เพื่อให้ผู้เรียนมีโอกาสที่จะทำงาน| ร่วมกับผู้เรียนอื่นอื่น|
- 29. læ? mi: ka:n mǔn wi:an sà-ma:-tcʰík| nai klùm tæ̀ là? klùm| tʰúk tʰúk sǎm ?a:-tʰít |pʰŵ:a? hâ:i pʰû: ri:an mi: o:-kàt tʰî: tcà tʰam ŋa:n| rû:am kàp pʰû: ri:an ?ẁn ?ẁn|
- 30.ความก้าวหน้าในทางเศรษฐกิจ| ที่คำเนินไปทุกวันนี้| แท้จริงนั้น| เป็นลักษณะปลาใหญ่กินปลาเล็ก|
- 30. khwu:am kâ:w nâ: nai tha:ŋ sèt-thà-kìt| thî: dam nə:n pai thúk wan ní:| thá teŋ nán| pen lák-sà-nà| pla: yài kin pla: lék|
- 31.โดยในระยะแรกแรก| จะมีลักษณะเปลี่ยนจาก| การใช้เครื่องมือแบบง่ายง่าย| มาใช้เครื่องจักรแทน|
- 31. do:i nai rá-yá ræk-ræk | teà mi: lák-sà-nà plì:an teà:k | ka:n teʰá:i kʰrŵ:aŋ mw: bæp ŋâ:i ฏâ:i| ma: teʰá:i kʰrŵ:aŋ teàk tʰæ:n|
- 32.นิทรรศการนี้| จะแสดงให้เห็นวิวัฒนาการ| ของเครื่องมือเครื่องใช้ในการประกอบอาชีพ| ของคนในภูมิภาคตะวันตก| ตั้งแต่อดีต ถึงปัจจุบัน|
- 32. ni-tʰát-sà-ka:n ní:| tcà sà-dæŋ hâ:i hěn wí-wát-tʰá-na-ka:n| kʰŏŋ kʰrŵ:aŋ mw kʰrŵ:aŋ tcʰá:i nai ka:n prà-kòp ʔ:-tcʰî:p| kʰŏŋ kʰon nai pʰu:m-mí-pʰâ:k tà-wan-tòk| tâŋ tæ ʔà-dì:t tʰṁŋ pàt-tcù-ban|
- 33.สิบเอ็ดจำนวนบัณฑิต| ที่มีความคิดเห็นว่า| วิชาภาษาอังกฤษที่เคยเรียน| ไม่เพียงพอในการประกอบอาชีพในปัจจุบัน| มีสัดส่วน ใกล้เคียงกับจำนวนบัณฑิต| ที่มีความคิดเห็นว่า| วิชาภาษาอังกฤษที่เรียนมาเพียงพอ|
- 33. sìp-?èt team nu:an ban dìt| t<sup>h</sup>î: mi: k<sup>h</sup>u:am k<sup>h</sup>ít hěn wâ| wí-te<sup>h</sup>a: p<sup>h</sup>a-să ?aŋ-krì:t t<sup>h</sup>î: k<sup>h</sup>əj ri:an| mâi p<sup>h</sup>i:aŋ p<sup>h</sup>ə nai ka:n prà-kòb ?:-te<sup>h</sup>î:p nai pàt-teù-ban| mi: sàt-suàn

klâi- $k^h$ i:aŋ kàp team nu:an ban dìt|  $t^h$ î: mi:  $k^h$ u:am  $k^h$ ít hěn wâ | wí-teha:  $p^h$ a-sǎ ʔaŋ-krì:t  $t^h$ î: ria:n ma  $p^h$ i:ang  $p^h$ ɔ|

34. เพราะไม่ใช่หน้าที่อะไรสักหน่อย

34. phró? mâi tehâi nâ: thì: ?á rai sàk nòi

35.และมีการหมุนเวียนสมาชิก| ในกลุ่มแต่ละกลุ่ม| ทุกทุกสามอาทิตย์| เพื่อให้ผู้เรียนมีโอกาสที่จะทำงาน| ร่วมกับผู้เรียนอื่นอื่น|

35. læ? mi: ka:n mǔn wi:an sà-ma:-teʰík| nai klùm tæ là? klùm| tʰúk tʰúk sǎm ?a:-tʰít |pʰŵ:a? hâ:i pʰû: ri:an mi: o:-kàt tʰî: teà tʰam ŋa:n| rû:am kàp pʰû: ri:an ?ẁn ?ẁn|

36.ความก้าวหน้าในทางเศรษฐกิจ| ที่ดำเนินไปทุกวันนี้| แท้จริงนั้น| เป็นลักษณะปลาใหญ่กินปลาเล็ก|

36. kʰwu:am kâ:w nâ: nai tʰa:ŋ sèt-tʰà-kìt| tʰî: dam nə:n pai tʰúk wan ní:| tʰǽ tɛɪŋ nán| pɛn lák-sà-nà| pla: yài kin pla: lɛ́k|

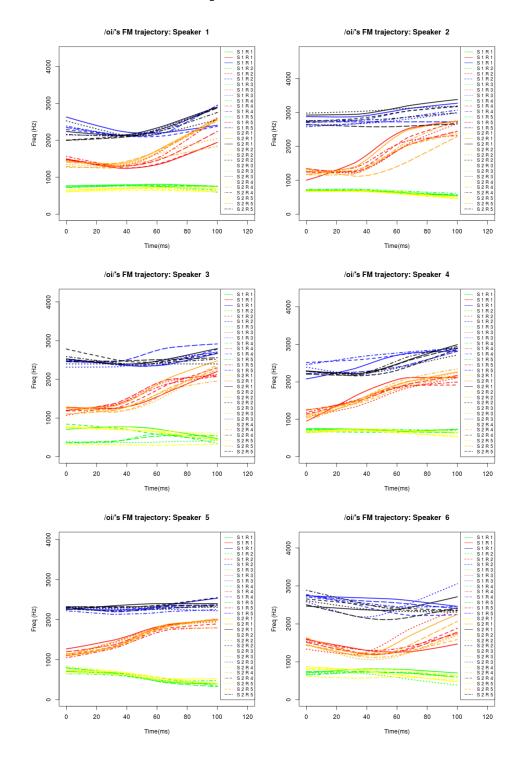
#### 3.2 อ่านคำต่อไปนี้ (ห่างกันประมาณ 2 วินาที)

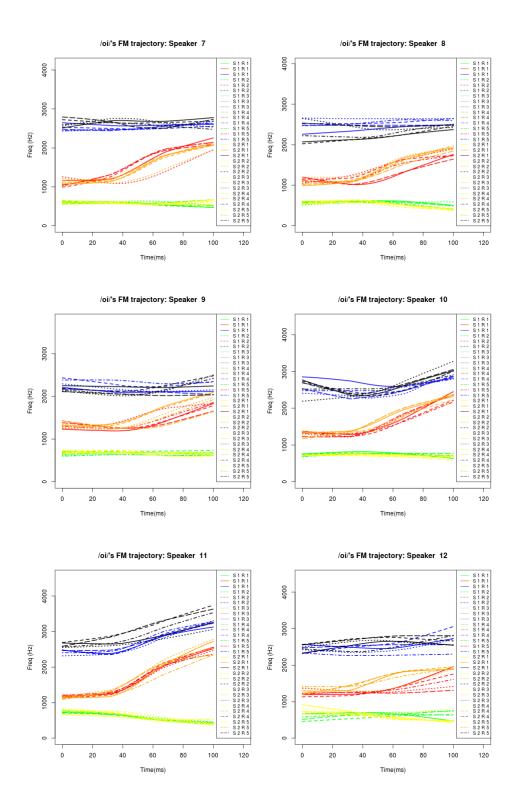
- 1. เรียน เรียน เรียน เรียน เรียน [ri:an mid] V. "to study"
- 2. ที่ ที่ ที่ ที่ ที่ ที่ [thi: HL] Prep. "at"
- 3. nis nis nis nis nis [ka:n mid] Prefix "-ness"
- 4. lu lu lu lu lu [nai mid] Prep "in"
- 5. คี คี คี คี คี คี **[di: mid]** ADJ "good"

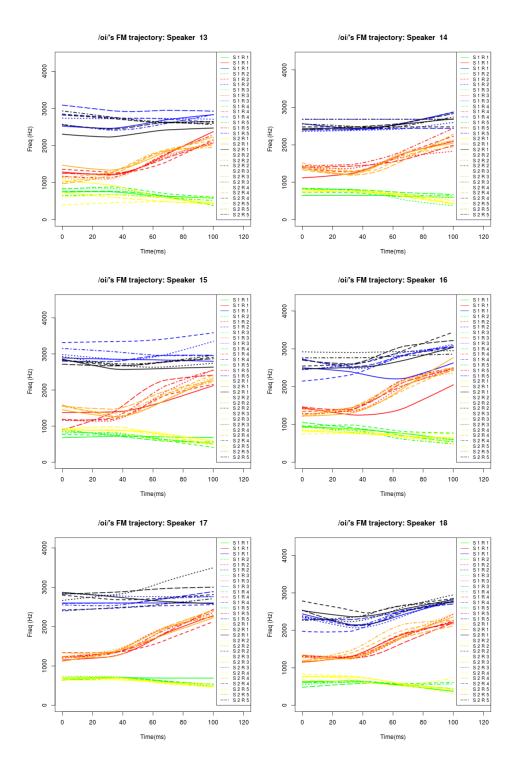
### Appendix B

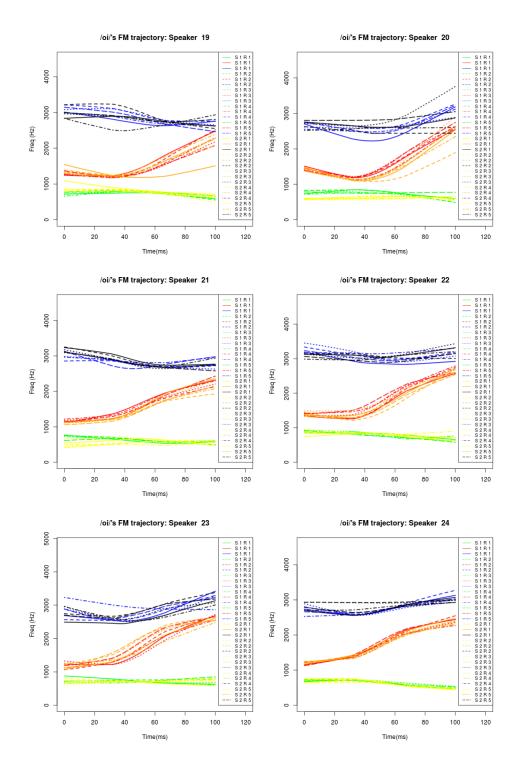
### F1-F3 values of [5i] plotted against a normalized time scale (100 msec)

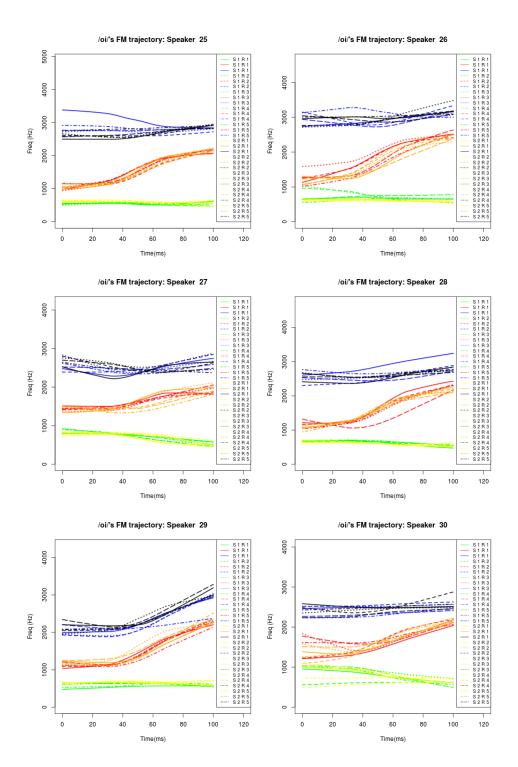
Note: [ɔi] is labeled as [oi] in the plots.







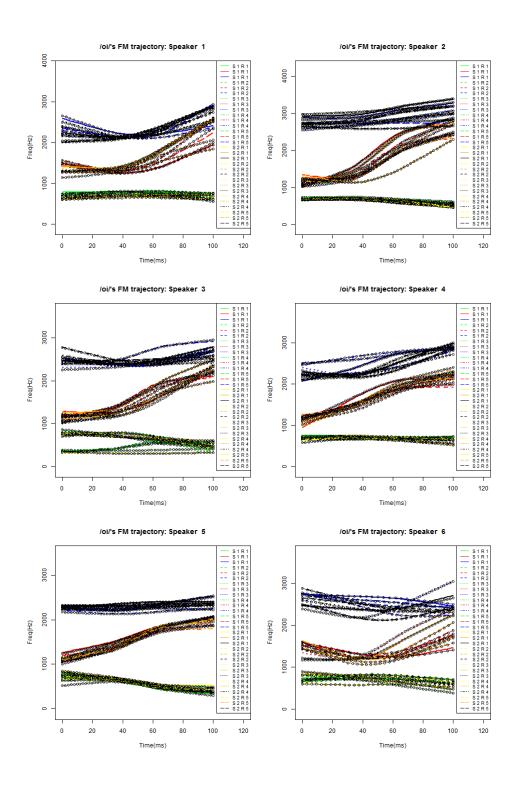


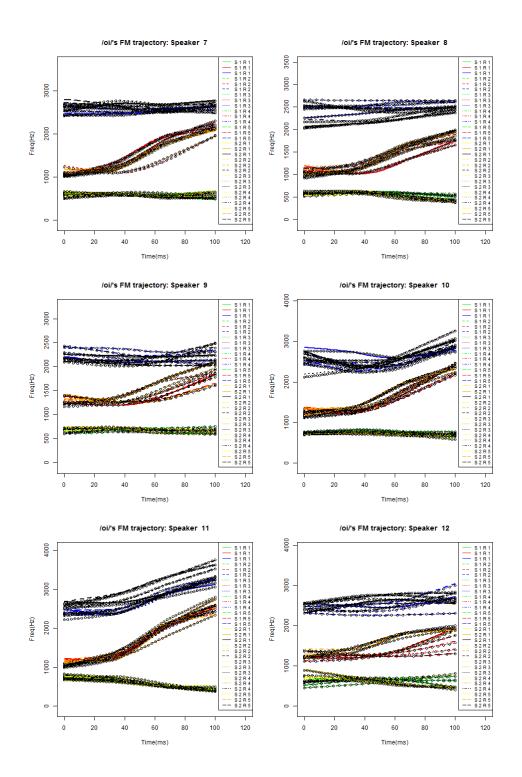


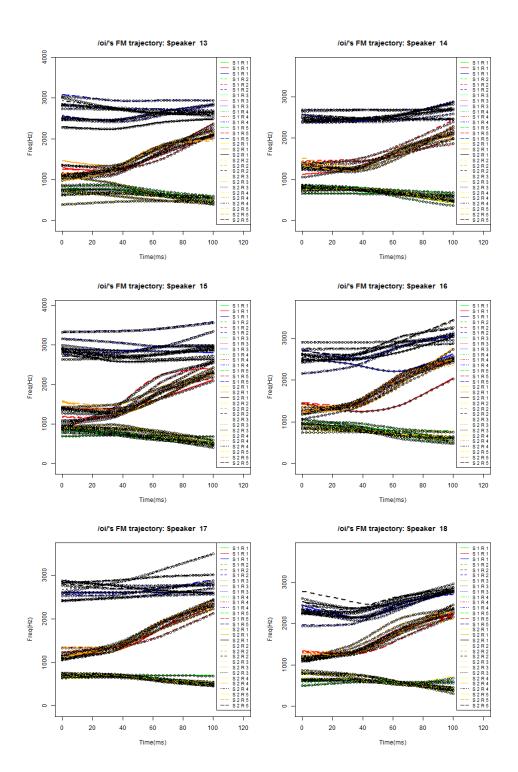
# **Appendix C**

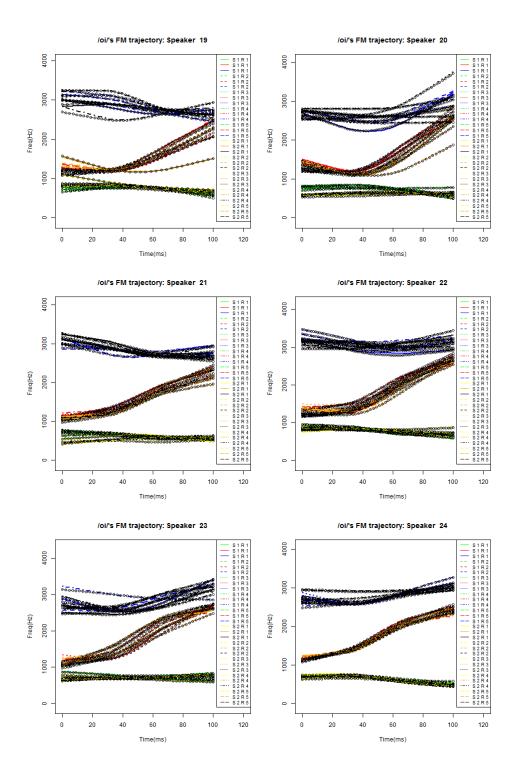
# F1-F3 trajectories of [5i] plotted together with cubic polynomials

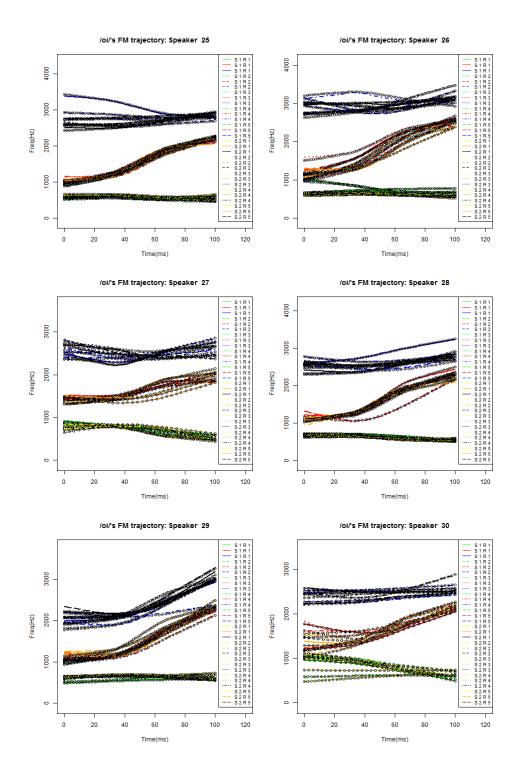
Note: [ɔi] is labeled as [oi] in the plots.





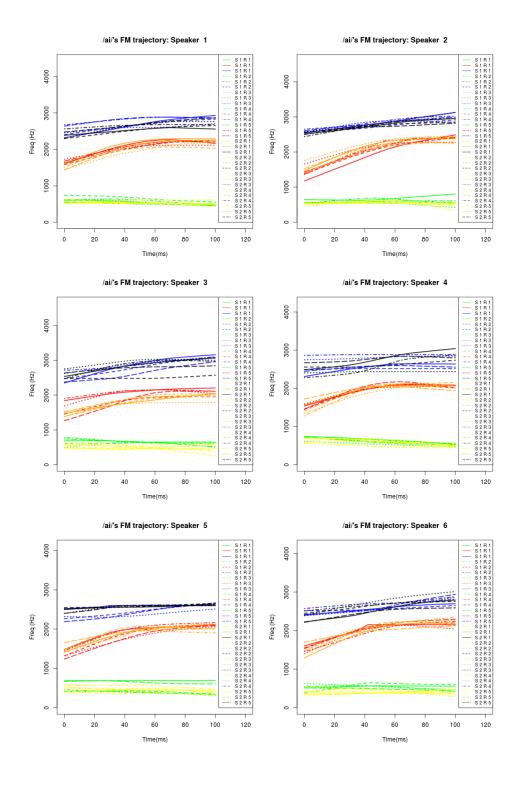


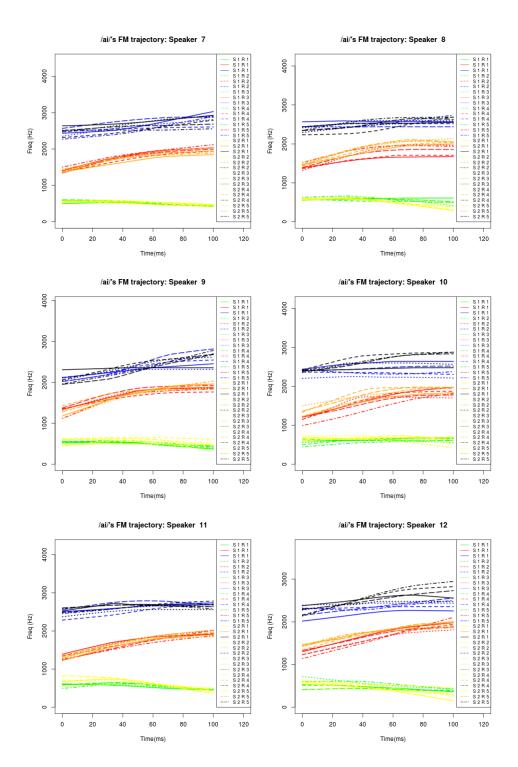


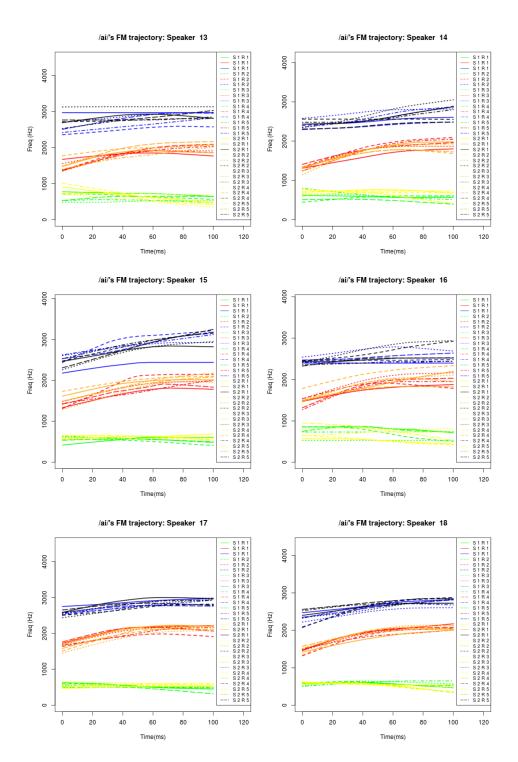


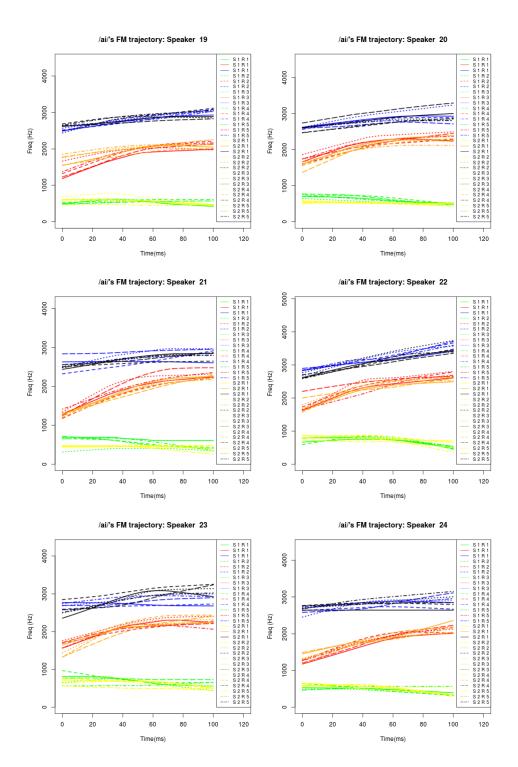
# Appendix D

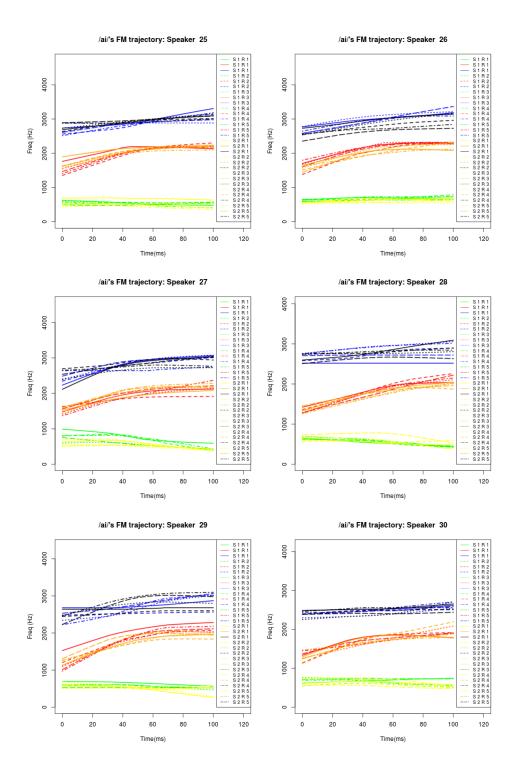
# F1-F3 values of [ai] plotted against a normalized time scale (100 msec)





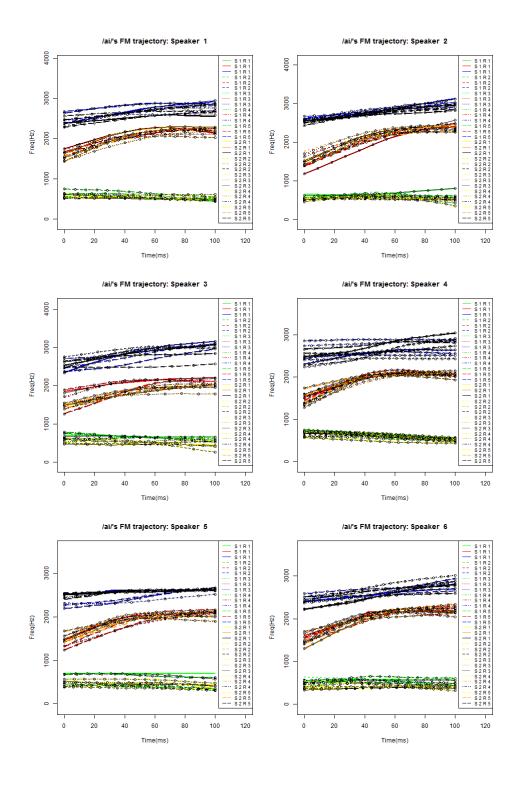


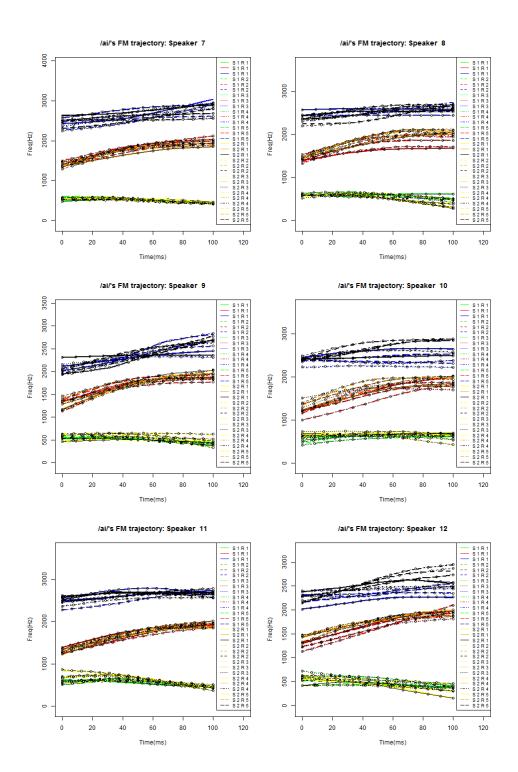


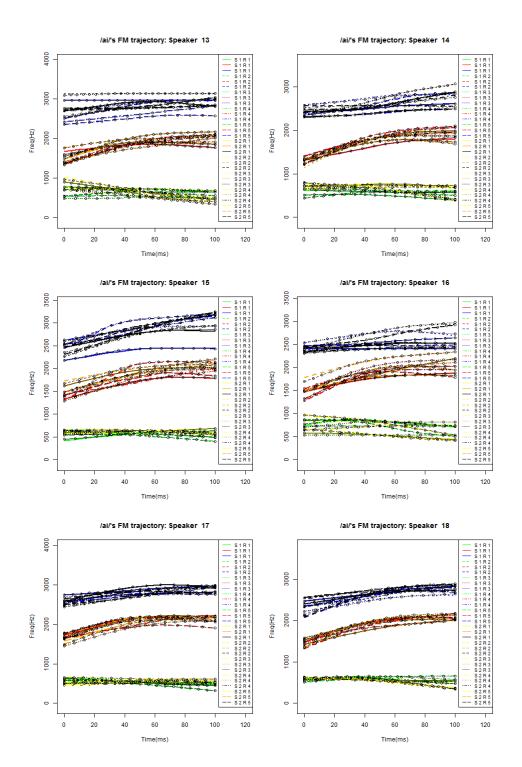


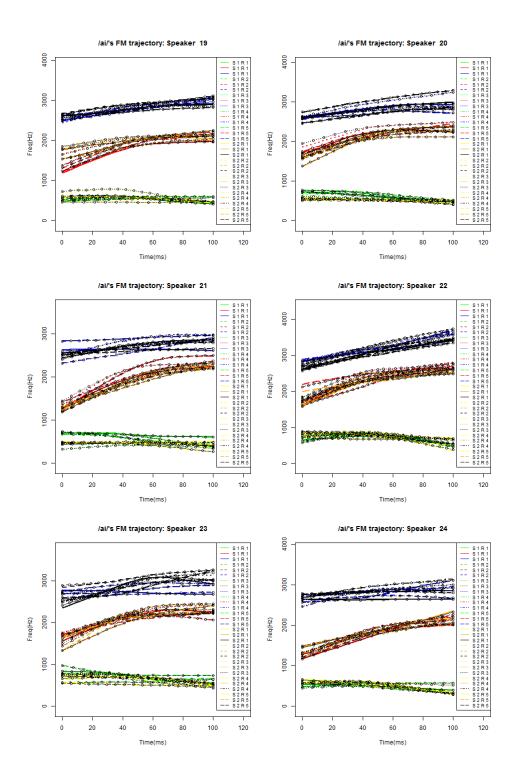
# Appendix E

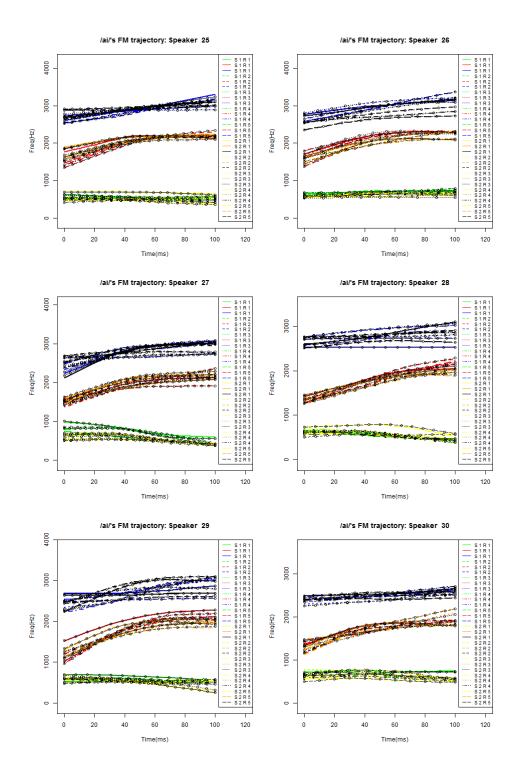
# F1-F3 trajectories of [ai] plotted together with cubic polynomials







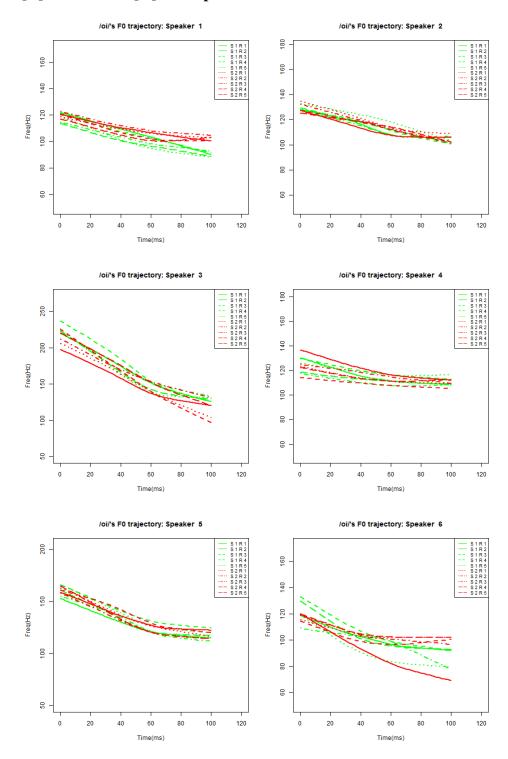


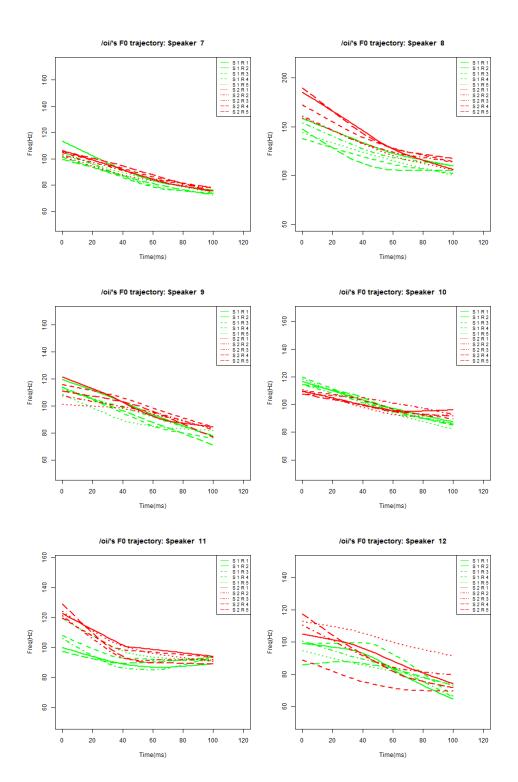


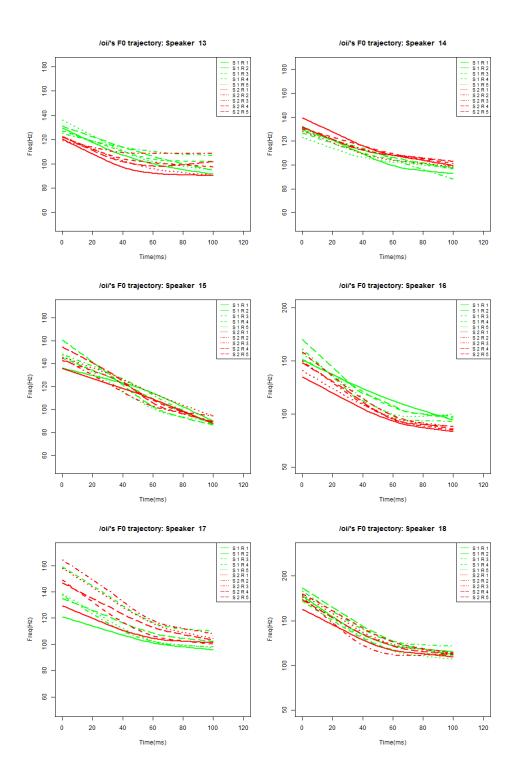
# Appendix F

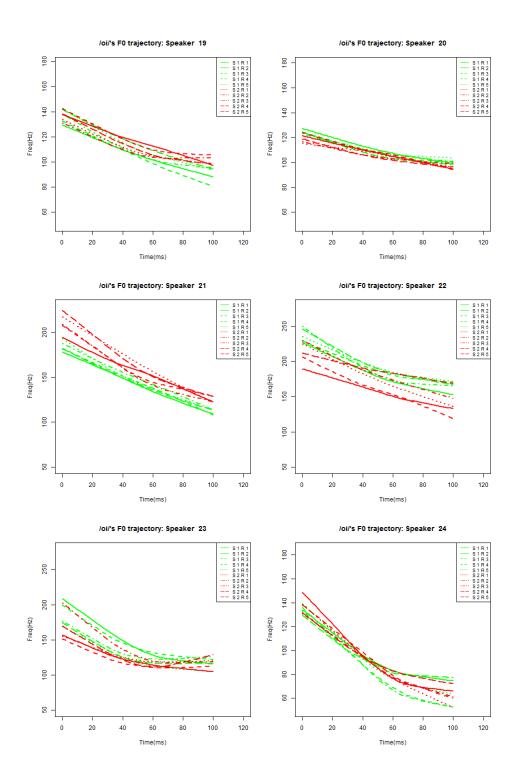
# Tonal F0 values of [5i] plotted against a normalized time scale (100 msec)

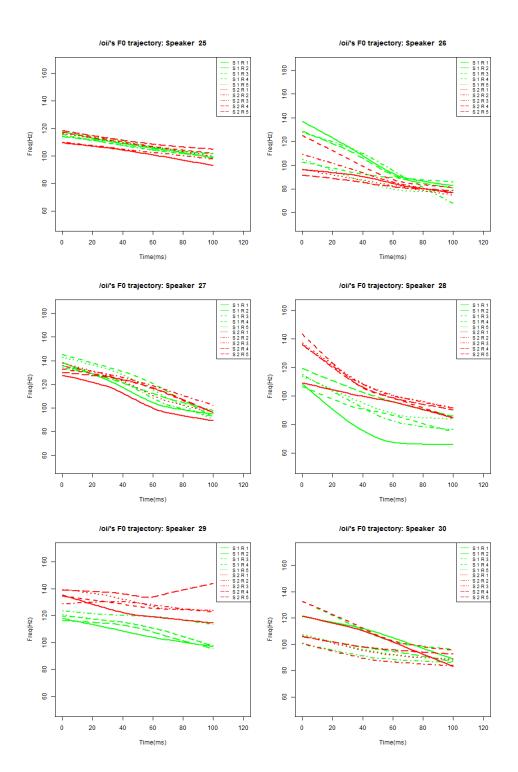
Note: [ɔi] is labeled as [oi] in the plots.







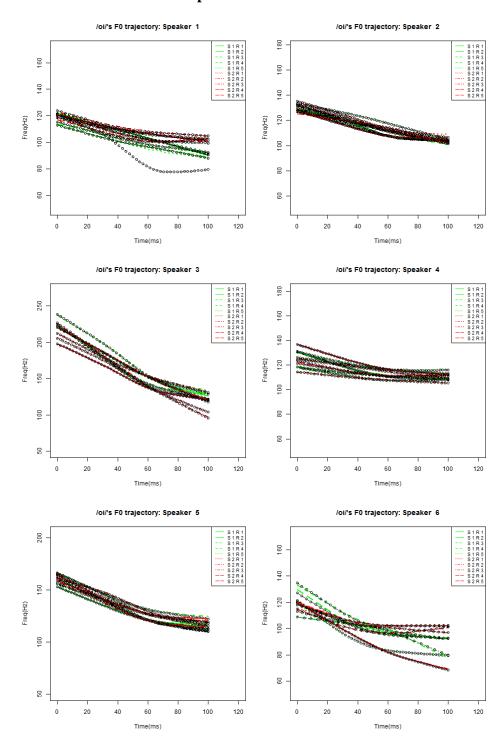


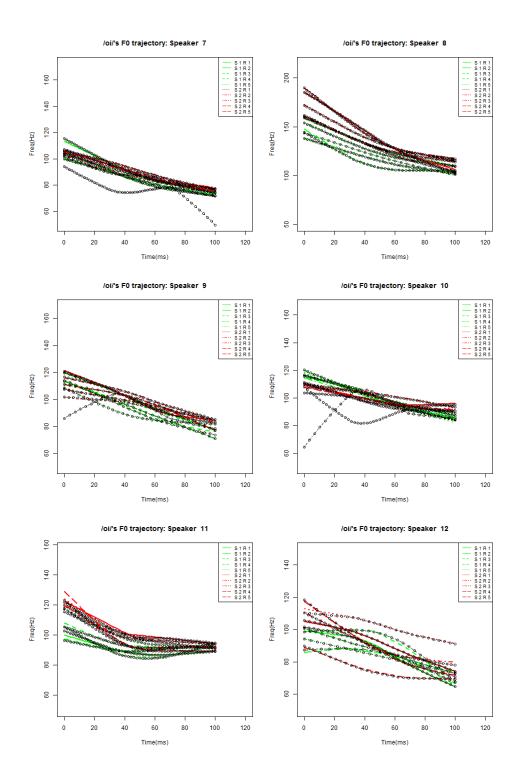


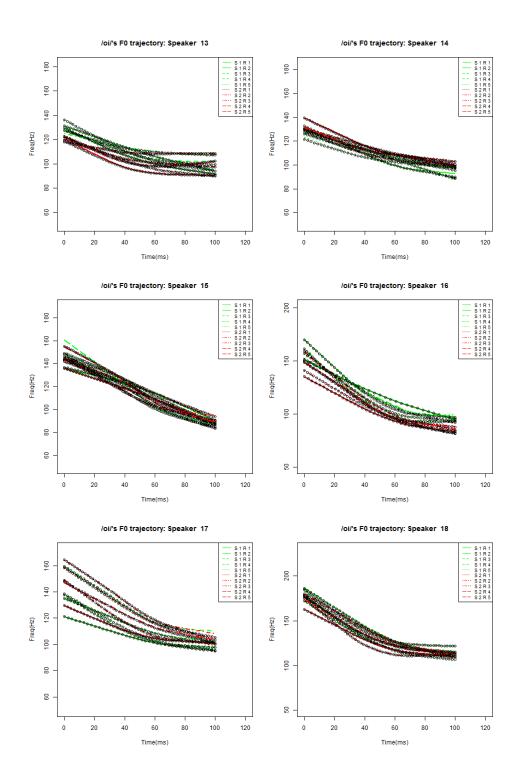
Appendix G

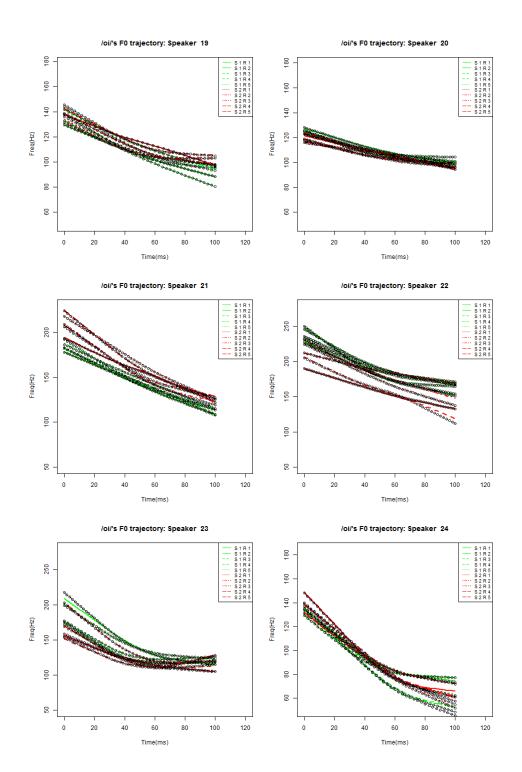
Tonal F0 values of [əi] plotted together with a quadratic polynomial curve fitting

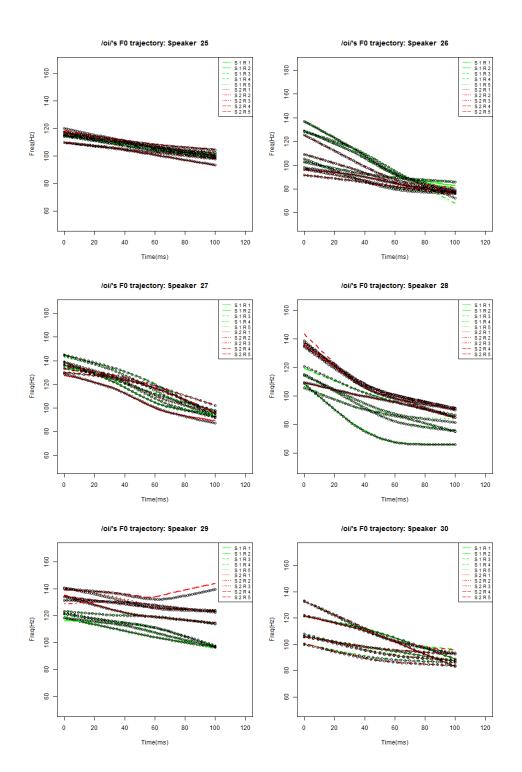
Note: /ɔi/ is labeled as /oi/ in the plots.



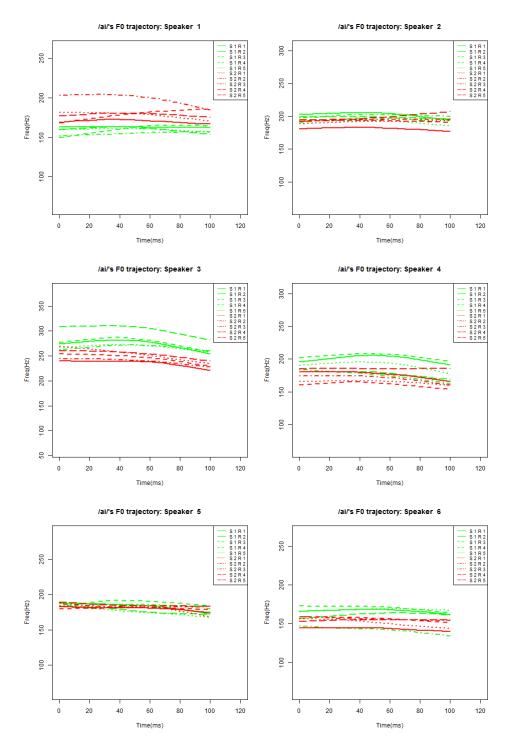


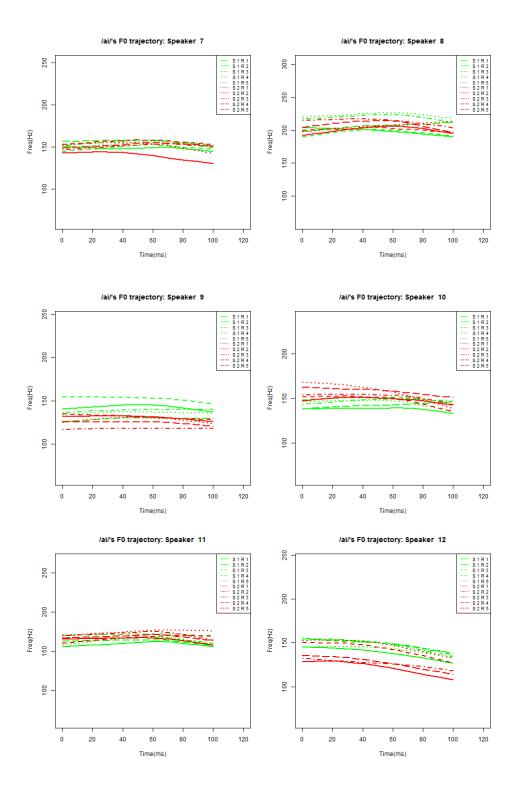


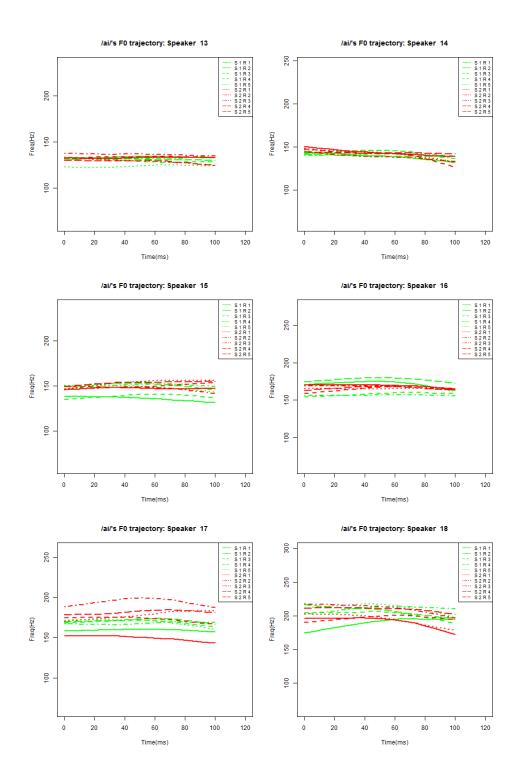


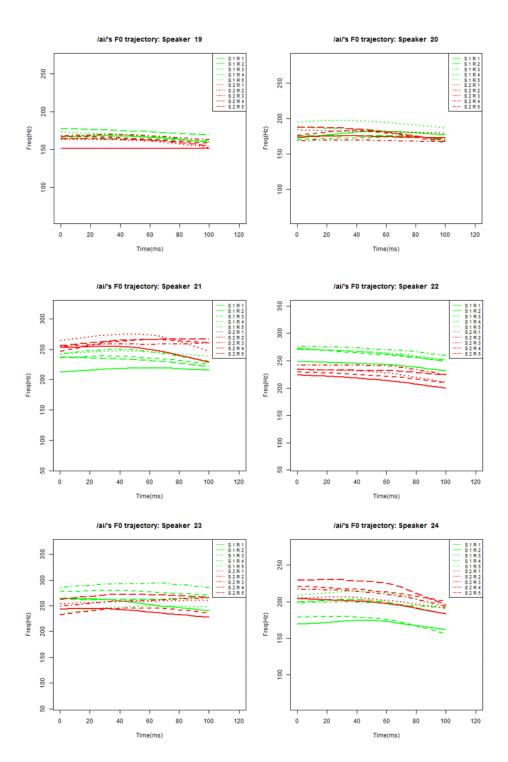


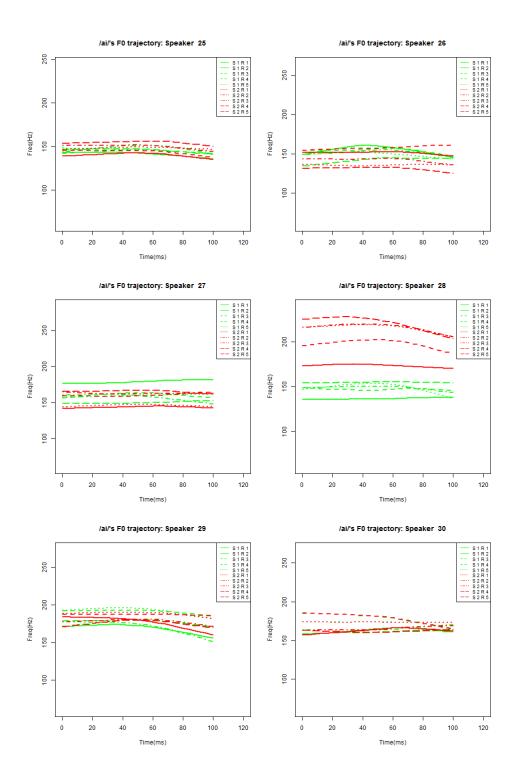
# Appendix H Tonal F0 values of [ai] plotted against a normalized time scale (100 msec)











# Appendix I Tonal F0 values of [ai] plotted together with a quadratic polynomial curve fitting

