Preservation of Tone in Right-Dominant Tone Sandhi: A Fragment of Disyllabic Tone Sandhi in Máodiàn Wú Chinese

Ruiqing Shen[1] *& Phil Rose*[2]
[1]Hong Kong University of Science & Technology
[2]ANU Emeritus Faculty
shenruiqings@hotmail.com, philjohn.rose@gmail.com

Abstract

Impressionistic and acoustic data are presented for the nine citation tones, and a small part of the disyllabic tone sandhi, of a speaker of the previously undescribed Chinese dialect of Maodian 毛店 from the Wuzhou 婺州 subgroup of Wu 吳. The data are used to refine the typology of the apparent right-dominant tone sandhi characteristic of the southern Wu and Min area. It is shown that not all word-final tones are the same as citation tones; and that therefore preservation of word-final tones cannot be criterial for right-dominance.
Index Terms: Tone Sandhi, Right-dominance, Tonal acoustics, Wu dialects, Wuzhou, Maodian.

1. Introduction

Language likes to exploit the polarity of metrical strength. One striking example is the typological difference, independent of segmental phonotactics, between right- and left-dominant tone sandhi systems found in the highly complex morphotonemics of the so-called *sandhi-zone* of China's eastern coastal provinces [1, 2]. In right-dominant varieties, it is the tones on the morphemes on the rightmost syllables of a word which determine the sandhi shape. The tone on the word-final syllable is said to be 'preserved', 'unchanged' or 'in agreement with' the citation tone, and tonal contrasts on the preceding syllables tend to be neutralised, although the neutralisation groupings are often bewilderingly complicated. Right-dominant varieties are said to be found in the southern Wu and Min dialects [3, p.287], but the exact distribution is not known. The variety described in this paper is located the southern Wu subgroup of Wuzhou, so it can be expected to be right-dominant.

But is right-dominance monolithic, or are there degrees to this typological parameter? We present data from a variety in the right-dominant area which appear to show the latter. We will focus in this paper on just one aspect of the relationship described as criterial for right-dominant sandhi: the extent to which the citation tone values are preserved on word-final syllables. For the purposes of this paper we define *preservation* thus. Preservation occurs *iif* word-final and citation tonal acoustics are intrinsically related, i.e. if the word-final tonal acoustics can be understood to be the same as the citation tone, once allowance is made for the effect of occurrence in word-final position and the expected perseverative assimilatory effect from the tone on the preceding syllable.

Since the disyllabic data are complicated we only have space to present two cases: one (relatively) simple, which is prototypically right-dominant; and one which is clearly not. We describe the citation tones first, then tones on disyllabic words with the simple case preceding the more complex.

2. Data

The data are from a high quality recording which was part of a wider survey of tones and tone sandhi in ca. 30 southern Wu varieties conducted by Professor W. Ballard in 1988 [4], and his generosity in making the data available is gratefully acknowledged here. The recording consists of three replicates each of ca. 40 monosyllabic and 190 disyllabic utterances elicited from a then 27 year old male speaker who was born and grew-up in Maodian until 17. A comparison of Wu dialect descriptions done in 1928 and 1992 [5,6] shows that they changed considerably in this ca. thirty year period, and more recent socio-phonetic findings [7] suggest that tonal change is accelerating, at least in metropolitan areas. In a sense, therefore, this description may also be partly considered a salvage operation. The digitised recordings – both citation tones and disyllabic tone sandhi – can be listened to on the second author's web-page (http://philjohnrose.net), where their individual and mean acoustics are also plotted.

3. Citation tones

In Chinese tonology, a citation tone is the tone given to a morpheme when its Chinese character, which may represent either a free or bound morpheme, is read out. The speaker has nine citation tones which may be described auditorily as follows (segmentals are transcribed phonemically). The **upper-mid level tone**, a reflex of Middle Chinese (MC) tone Ia**,** has a level pitch contour in the upper third of the pitch range, e.g. fi *fly* 飛, kɑ *liver* 肝, nuŋ *east* 東. The **lower-mid level tone**, from MC Ib, has level pitch in the lower third of the pitch range, e.g. bi *skin* 皮, dzua *tea* 茶, nia *year* 年. The **mid rising tone** (< MC IIa) has prolonged pitch in the mid pitch range with a final rise, e.g. siəu *arm* 手, nia *point* 點, hua *fire* 火. The **lower-mid rising tone** (< MC IVa) has the same delayed pitch rise a little below that of the mid rise tone e.g. siɛ *snow* 雪, ɓɛi *north* 北. The **low rising** tone (<MC IIb, IVb) has pitch which rises from low in the speaker's pitch range to mid with a prolonged initial component, e.g. bi *blanket* 被, zua *sit* 坐 , [ɔ] *to study* 學, dɑu *poison* 蝳. The **high falling** tone (<MC IIIa) has pitch which falls through the speaker's modal pitch range, e.g. si *four* 四, tʰɯa *to jump* 跳. The **depressed high falling tone** (<MC IIIb) has similar pitch to the high falling tone, but with a low onset which results in a convex pitch contour in the bottom two thirds of the pitch range. Examples are di *ground* 地, vɑ *rice* 飯, miɛ *face* 面. The **short stopped mid tone** (<MC IVa) has a short pitch in the lower-mid pitch range truncated by a glottal stop, e.g. kuaʔ *bone* 骨, tɕʰyaʔ *to come out* 出. The **short stopped low rise**

tone (<MC IVb) has a short rising pitch in the lower third of the pitch range truncated by a glottal stop, e.g. zaʔ *ten* 十.

This rather large number of observed tones relates primarily to the historical development of morphemes with tonal cognates of Middle Chinese so-called *entering tones* IVa and IVb. Originally, in Proto Wu say, these two tones had short duration and ended in a glottal stop. In many modern Wu dialects their reflexes retain these features and are still considered separate tones; but in other varieties the tones have lost their glottal stop and undergone further development. In some Wuzhou varieties the short tones have lengthened and merged with other tones; in others they have lengthened but remained separate by virtue of different pitch shapes [8, p.23]. Interestingly, the Maodian speaker provided a further variation on this theme, in that he clearly showed a merger of etymological tone IVb with tone IIb (the low rising tone), whilst keeping a lengthened version of etymological tone IVa separate (as the lower-mid rise tone).

This situation was further complicated, however, by a phenomenon, again said to be typical for Wuzhou, whereby some morphemes with etymological IVa and IVb tones have alternative phonological shapes [8, p.23]. One shape is conservative, preserving the short pitch ending in a glottal stop; the other is the innovative lengthened tone. This alternation was also shown by the Maodian speaker. For most IVa and IVb cognates he had innovative long reflexes. For a few IVa and IVb cognates, however, he retained a conservative short stopped tonal shape. Although this phenomenon is traditionally termed 文白異讀 *different colloquial and literary character readings*, there was nothing in the linguistic structure of any of the morphemes involved that would serve as an obvious conditioning factor. Thus, for example, he read the characters for the morphemes *bone*, *come out* and *ten* with short stopped tones, but, in the same formal elicitation session, those for *snow*, *put out* and *month* were given long tones. Indeed, Ballard's notes show some free variation, in that *bone* was also said with a long tone. Although the conditioning of such short forms remains elusive, therefore, it is clear that one has to deal with nine different tonal shapes.

Citation tone acoustics were quantified with the same method used in a previous study of a right-dominant Wu variety [9]. A wideband spectrogram was generated in *Praat*, together with its wave-form and superimposed F0. The token's tonally relevant F0 was then identified, extracted and modeled in *R* by an 8th order polynomial. This enabled F0 values to be sampled from the polynomial F0 curve with a sufficiently high sampling frequency (at 10% points of the curve as well as 5% and 95%) to capture the details of its time-course.

The mean tonal acoustics of the nine Maodian citation tones (F0 as function of duration) are shown in figure 1. The tonal F0 shapes have been plotted in two panels, as their complex configuration would have made it difficult to identify their shapes otherwise. In order to demonstrate that some IVb morphemes have indeed merged with reflexes of IIb, the low rising tone is plotted separately with a green dotted line for its IVb and a black dotted line for its IIb constituents. Their extreme similarity indicates provenance from the same synchronic tone.

The F0 shapes of the individual tones are clear and generally correspond fairly well to their pitch descriptions. The two short stopped tones (brown) can be seen to have a duration of about half that of the unstopped tones, and also to have very similar onsets to their corresponding long tones (green). The lower-mid level tone (blue) appears to have a slightly depressed onset extending for the first 10 csec. or so. One clear area of disagreement between the F0 and tonal pitch is in the high falling and depressed high falling tones (red). Their offsets, between ca. 105 Hz and 110 Hz, lie considerably below the two low rising tones that sound to lie near the bottom of the speaker's range. Including these falling tone offsets as tonally relevant will have the effect of distorting the way the F0 represents tonal pitch, and they are best considered as idiosyncratic offset perturbations (some speakers end their falling tones with a glottal stop or creak; others, like this Maodian speaker, have a gradual offset to modal phonation). In the following sections, we describe tone sandhi in words ending with morphemes which carry the lower-mid level and mid rising citation tone.

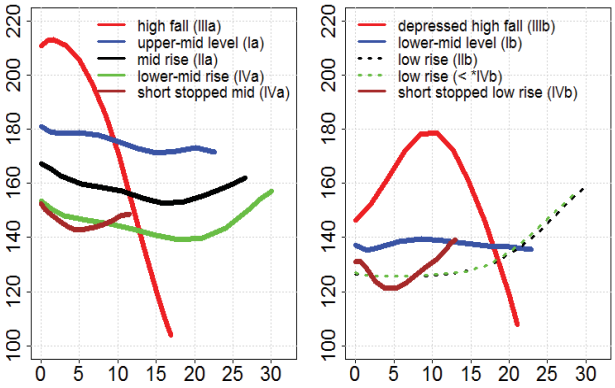

Figure 1: *Mean F0 (Hz) of the speaker's nine isolation tones plotted against mean duration (csec.).*

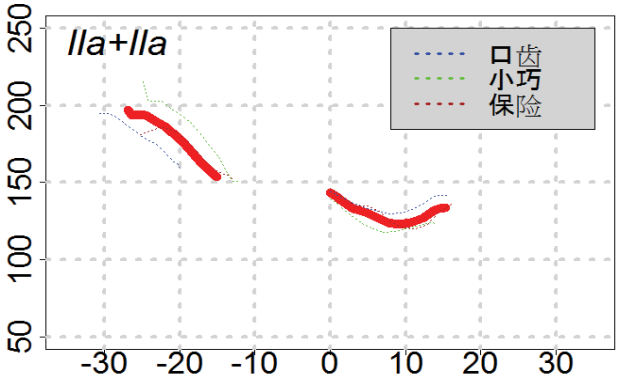

Figure 2. *Tonal acoustics of three Maodian words with mid rising morphotonemes on both syllables. Thick solid line = mean F0, dotted line = individual tokens' F0. X axis = duration (csec.) y axis = F0 (Hz).*

4. Disyllabic tone sandhi: procedure

The same procedure for extracting the tonal acoustics of the disyllabic words was used as in [9], which sampled F0 as a function of the word's segmental structure – first Rhyme, intervocalic consonant and second Rhyme – using 8th order polynomial modeling in *R*. Three different words were measured for each etymological tonal combination, and their mean values calculated. Figure 2 shows the individual values and their mean for three words with the mid rising morphotoneme on both syllables. These have a high falling pitch on the first syllable followed by a low dipping pitch on

the word-final syllable. (One example was /ɓɯa ɕie/ [52.212] *insure*, which is a synonym compound consisting of the bound morpheme {保 *protect* ɓɯa 323} and the free morpheme {險 *danger* ɕie 323}. F0 is plotted as a function of absolute duration aligned at onset of second-syllable Rhyme (csec.0). The fairly tight clustering of the individual words' F0 values is typical.

5. Disyllabic tone sandhi: a minimally complex example

To demonstrate the mechanics of the least complex tone sandhi in Maodian disyllabic words, we examine combinations with underlying mid rising tone on the word-final syllable, and all tones on the preceding syllable. Examples are given in table 1. The procrustean Chao five-point scale transcribing tonal pitch is intended as convenient abbreviation only.

Table 1. *Examples of Maodian speaker's tone sandhi in words with underlying mid rising tone on word-final syllable. Pitch representations are color-coded with figure 3.*

word-final mid-dipping isolation tone [323] preceded by …			
… upper-mid level morphotoneme [44] on S1		… lower-mid level morphotoneme [22] on S1	
kɔ kʰɔ 33.323	高考 *college entrance exam*	ɔŋ ɕy 23.323	洪水 *flood*
… mid rising morphotoneme [323] on S1		… short mid stopped morphotoneme [3̲] on S1	
ɓɯa ɕie 43.212	保險 *insure*	tɕʰyə̆ kʰəu 4̲.323	出口 *export*
… high falling morphotoneme [51] on S1		… depressed high falling morphotoneme [241] on S1	
ɗe ɓi 33.323	對比 *compare*	ẓ nia 32.212	字典 *dictionary*
… lower-mid rising morphotoneme [212] on S1		… low rising morphotoneme [13] on S1	
ɓa kuo 43.212	百果 *all kinds of fruits*	ba kuo 32.212	白果 *ginko*

Table 1 indicates, firstly, five pitch shapes for the first syllable tone: upper-mid level [33], low rising [23], high and mid falling [43], [32], and short stopped high [4̲]. These shapes reflect several complex mergers. The mid level [33] tone is the realization of a merger between the underlying high falling tone and the underlying upper-mid level tone. The high falling [43] tone is the realization of a merger between mid rising and lower-mid rising tones. The mid falling [32] tone represents a merger between underlying low rising and depressed high falling tones. The low rising [23] tone corresponds to underlying low level tone, and the short high [4̲] tone corresponds to short stopped mid. Note that none of the resulting first syllable tonal allomorphs corresponds to its citation shape. A high falling citation tone is realized as mid level, for example, and a high rising citation tone is realized as high falling. The same sort of first tone complexity was also demonstrated for the Wu dialect of Wencheng [9], and appears typical.

In contrast to the first syllable tones, the word-final tone is straightforward. Table 1 shows it has two surface forms, both with delayed rising pitch like the corresponding citation tone. One is in the mid pitch range, represented as [323], and one slightly lower "[212"]). The conditioning is very largely clear: the lower version occurs after preceding falling pitched tones, the higher elsewhere. Exactly the same intrinsic perseverative allotony conditioned by a [+/- fall] on the preceding syllable occurs in Wencheng [9], showing that even in right-dominant systems the weak tone can influence the strong. The lower allotone also occurs after the short high tone, which has a falling F0, but is too short to have a pitch contour. The conditioning is not clear in this case.

Figure 3 shows the mean tonal acoustics corresponding to the shapes in table 1 (colour-coding is used for the first-syllable tones). The mid rising citation tone acoustics are also shown. Five different mean F0 shapes – two falling, one level, one rising and one short – can be seen for the first syllable tone corresponding to the five pitch shapes just described. The two dipping F0 shapes corresponding to the two intrinsic word-final allotones can be seen lying a little lower than the citation tone. Figure 3 shows the word-final tone can be considered as a mid rising citation target intrinsically perturbed by co-articulation with the preceding syllable tones: a clear case of word-final preservation of tone.

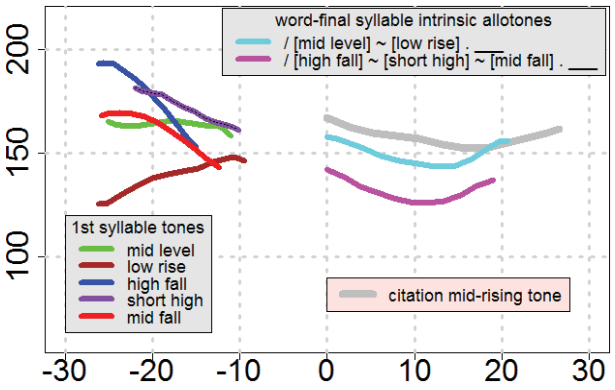

Figure 3. *Mean tonal acoustics of Maodian disyllabic words with underlying mid rising tone on word-final syllable. X-axis = duration (csec.) aligned at onset of second syllable rhyme, y-axis = F0 (Hz).*

6. Disyllabic tone sandhi: a more complex example

Table 2 gives examples of words with underlying lower-mid level [22] tone on the word-final syllable, and the corresponding acoustics are shown in figure 5. The mean acoustics of the low level citation tone have also been plotted. It is evident that the situation is more complex for both first syllable and word-final tones. Unlike before the mid rising tone just discussed there are no first syllable mergers. There are eight different tonal shapes on the first syllable. Their F0 shapes of are all clear in figure 4. Again none match their citation tones. Risking confusion, we list them here (they are colour-coded with table 2 to help matching). As in the previous section, the lower-mid level citation tone corresponds to a low rising tone (brown), the mid rising tone corresponds

to a high fall (blue), the high falling citation tone corresponds to mid level (green), and the depressed high fall corresponds to a mid fall (yellow). Unlike the previous section, the upper-mid level citation tone corresponds to a high rising tone (red), the low rise tone corresponds to a low concave tone (magenta), the mid short stopped citation tone corresponds to a short high rise (purple) and the short stopped rising tone corresponds to a low level (orange). None of these look like morphotonemic alternations easily generalizable with conventional tone features.

Table 2. *Examples of Maodian speaker's tone sandhi in words with lower-mid level morphotoneme* [22] *on word-final syllable. Pitch representations are color-coded with figure 4.*

word-final lower-mid level morphotoneme [22] preceded by …	
… upper-mid level morphotoneme [44] on S1	… lower-mid level morphotoneme [22] on S1
tʰia dʑɔ **45.41** 天橋 *flyover*	jɔ muɑ **24.51** 羊毛 *wool*
… mid rising morphotoneme [323] on S1	… low rising morphotoneme [13] on S1
ɓuɑ dʑ **53.21** 保持 *preserve*	du bi **243.31** 肚皮 *stomach*
… high falling morphotoneme [51] on S1	… depressed high falling morphotoneme [241] on S1
tʰa bɐŋ **33.334** 太平 *peace*	di dʑiəu **32.22** 地球 *globe*
… lower-mid rising morphotoneme [212] on S1	… short stopped low rise morphotoneme [12] on S1
tsŏ dʑiəu **34.51** 足球 *soccer*	zĕ dəu **2.223** 石頭 *stone*

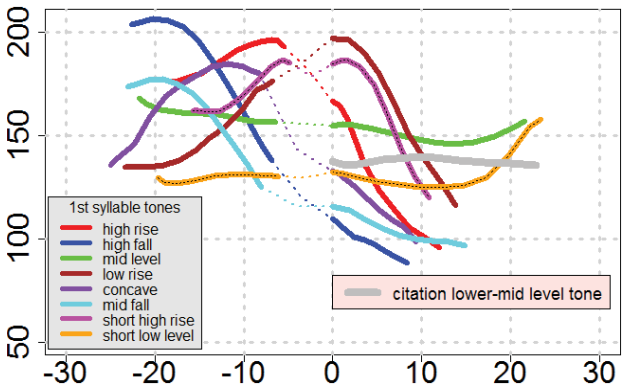

Figure 4. *Mean tonal acoustics of Maodian disyllabic words with underlying lower-mid level tone on word-final syllable. X-axis = duration (csec.) aligned at onset of second syllable rhyme, y-axis = F0 (Hz). Dotted lines indicate F0 on intervocalic consonant.*

The word-final tonal pitch shapes fall into three classes. There are, firstly, four pitches falling from different heights to low: [51], [41], [31] and [21]. The onsets of the five F0 shapes corresponding to these four falling pitches are clearly all determined by the trajectory of the preceding tone, and therefore can be considered intrinsic variants of a falling pitched tone (or even realisations of no tone, with their falling pitch conditioned by a low boundary tone). Secondly, there are two pitches rising to mid: [224] and [334]; these too are intrinsically determined by the height of the preceding tone. Finally, there is one low level [22] pitch, its percept corresponding to the clear leveling out of the F0. There are thus three extrinsic allotones of the word-final lower-mid level tone: falling, rising and level, only the last of which, just, can be considered a case of preservation. Moreover, a case can be made for combinations with the entirely predictable falling word-final allotone to be instantiations of *strong-weak* metrical structure rather than the *weak-strong* structure implied by right dominance. Two of the shapes – 53.21 and 243.31 – are reminiscent of spread high falling and depressed high falling first-syllable tones, except there is no explanation for where such first-syllable tones might have come from: (depressed) high falling and convex tones exist in Maodian to be sure, but table 2 shows they are related to first-syllable [33] and [32] shapes! It seems that, for these data at least, right dominance is not monolithic, and word-final tone preservation cannot be considered criterial for it.

7. Summary

A small part of the tone sandhi behavior has been described for a speaker of Wuzhou Wu, and the data interrogated for typical right-dominant behavior using impressionistic description and quantified acoustics. Complex first syllable behavior typical of right dominant systems was observed. The mid rising tone was shown to be preserved word-finally, but not the lower-mid level tone, showing that preservation of word-final tone is not invariant in putatively right-dominant systems. Clearly, right-dominance is worthy of further study.

8. Acknowledgements

Many thanks to our two anonymous reviewers!

9. References

[1] Ballard, W., "Wu, Min and a Little Hakka – Tone Sandhi: Right and Left", Cahiers de Linguistique Asie Orientale 13:3-34, 1984.

[2] Zhang J., "A directional asymmetry in Chinese tone sandhi systems", J. East Asian Linguistics, 16:259-302, 2007.

[3] Pan W., "An Introduction to the Wu dialects", in Wang S. [Ed.] Languages and Dialects of China, Journal of Chinese Linguistics Monograph 3:237-293, 1991.

[4] Ballard, W., "Oujiang Wu Tone Sandhi: Visi-Pitch Results", Chinese Language and Linguistics 1: Chinese Dialects. Symposium Series of the Institute of History and Philology Academia Sinica 2, Taipei, 41-66, 1992.

[5] Chao Y. 趙元任, 現代吳語的研究 *Studies in the Modern Wu Dialects*, Tsing Hua College Research Institute Mono. 4, 1928.

[6] Qian N. 钱乃荣, 当代吴语研究 [*Studies in the Contemporary Wu Dialects*], Shanghai Educational Press, 1992.

[7] Zhang J., A Sociophonetic Study on Tonal Variation of the Wúxī and Shànghǎi Dialects, LOT Netherlands Graduate School of Linguistics, 2014.

[8] Fu G. 傅国通, Fang S. 方松熹, Cai Y. 蔡勇飞, Bao S. 鲍士杰, Fu Z. 傅佐之, 浙江吴语分区 [Wu dialect subgroups of Zhejiang], Zhejiang Linguistics Society, 1985.

[9] Rose, P., "Complexities of Tonal Realisation in a Right-Dominant Chinese Wu dialect – Disyllabic Tone Sandhi in a Speaker from Wencheng", Journal of the South East Asian Linguistics Society 9:48-80, 2016.

Towards a Better Understanding of Regional Variation in Standard Australian English: Analysis and Comparison of Tasmanian English Monophthongs

Rael Stanley
University of Melbourne
raels@student.unimelb.edu.au

Abstract

Using phonetic analysis, this investigation looks at the acoustic properties of Tasmanian English vowels, as produced by speakers of that variety of English from the Austalk corpus. It compares the formant values of monophthongal vowel targets to published formant values Melbourne and Sydney vowels. The aim of the study is to give a first outline of the vowel space of Tasmanian English, to determine whether there is any regional variation between Tasmanian and mainland vowel realisation, and to compare what differences there are in vowel realisations for older and younger speakers of Tasmanian English.

Index Terms: Regional variation, Tasmanian English, vowels

1. Introduction

This study is looking at the accent produced by speakers of Tasmanian English, in comparison to varieties of Australian English spoken in other Australian states. As vowels are the phonemes most responsible for accent variation [1], it will be focussing on how production of them is similar to or different from vowels produced elsewhere.

1.1 Research Questions

The bulk of work done on regional accent variation for Australian English has focussed on the larger population centres in the country, such as Sydney, Melbourne, Adelaide, and Perth. However, there is a paucity of data for smaller areas, particularly the capital of Tasmania: Hobart.

Separated from the mainland of Australia by the waters of Bass Strait, Tasmania is a mountainous island that was, for much of its history post-British-colonisation, isolated from the rest of Australia's population by more than simply the great distances found between populations centres on the mainland but also the broad and treacherous sea.

As physically isolating barriers are a common factor in regional variation of languages and dialects [2], it is somewhat surprising that there has been so little research into differences between Tasmanian English and the other varieties spoken in Australia, and this study seeks to redress this, with the following research questions:

1. What are the acoustic properties of Australian English short and long monophthongs as spoken by Tasmanians?
2. Are there differences in these acoustic properties that are present according to age categories?
3. Is there regional variation present between the vowels analysed in this study, and those for Melbourne and Sydney, in studies by Cox and Billington [3] & [4] respectively?

2. Method

2.1 Data Collection and Participants

21 male and 18 female speakers of Hobart English produced vowels in citation form (/hVd/), and were extracted from the list of Tasmanian speakers, who had completed all their schooling in Tasmania, from the AusTalk corpus [5]. A total of 1,096 vowel tokens were downloaded for analysis via Alveo online [6].

The speech data for each group, male and female, was separated into younger and older speakers, with younger speakers being aged between 20 and 39 years, and the older speakers aged 60 years and over, excluding those speakers aged 40 to 59 years in order to more clearly see what effects age have on the Tasmanian English accent. Table 1, below, shows the distribution of speakers across age categories and gender.

Table 1. *Showing number of speakers of each gender, according to age category*

Gender	Number of Younger Speakers	Number of Older Speakers
Female	11	7
Male	13	8

2.2 Data Labelling

The vowels chosen for analysis were the short and long monophthongs of Australian English, because these are the vowels shared by the analyses Performed by Cox and Billington [3] & [4] that I am comparing my data with.

Segments for each repetition of each cited form were automatically labelled through use of the application WebMAUS Basic [7].

2.3 Analysis

The open-source data manipulation program RStudio [8] was used to automatically identify the formant values for F1 and F2 at the vowel midpoints and map them to ellipse plots for comparison between the different groups by both age and gender of the speakers. Any outliers in the data sets were identified visually from these plots, and had their formants manually checked and adjusted, where necessary, as suggested by Harrington [9] using Praat [10].

Linear Mixed Model (LMM) tests were run, using the packages lme4 [11] and multcomp [12] to provide statistical significance data on the differences in these values. The LMM tests were run on both F1 and F2 values, using age category and gender as fixed effects, with speaker as a random effect for both.

Mean formant data collected for each vowel for male and female groups of the younger speakers was compared to mean formant data for the same monophthongal vowels collected in [3] and [4].

3. Results

Running LMM tests for age category as a fixed effect showed more sporadic statistical significance across the mean F1 and F2 data for males and females. The only vowels that showed statistically significant differences in both F1 and F2 by age category were /æ/ for both male and female speakers, and /e:/ for female speakers, only. Those vowels which displayed statistical significance in mean F1 value across age categories were female speakers' productions of /e, ɜ:, o:/ and male speakers' productions of /e:, e, ɜ:/, while the vowels showing differences of statistical significance in mean F2 values were female speakers' productions of /ɐ:, ɔ, ʊ, ʉ:/ and male speakers' productions of /ɐ:, ɐ, ʉ:/.

3.1 The Short Vowels of Tasmanian English

A pair of ellipse plots of the short vowels produced by the younger and older groups of Tasmanian females in citation forms is presented in Figures 1 and 2, below. These plots have been chosen, as they most clearly display the points of interest in these data sets.

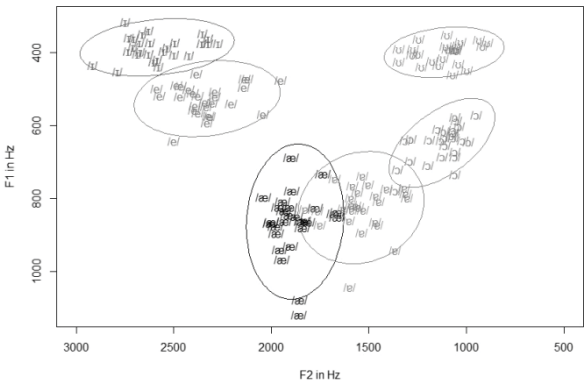

Figure 1: *Ellipse plots of the F1/F2 values of short vowels produced by younger female speakers of Tasmanian English*

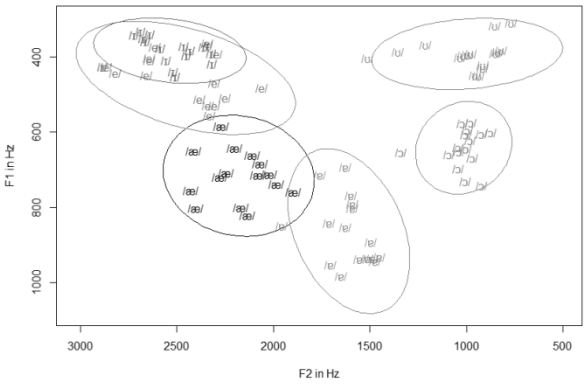

Figure 2: *Ellipse plots of the F1/F2 values of short vowels produced by older female speakers of Tasmanian English*

As can be seen in the above pairs of plots, there is a notable retraction and lowering of the /æ/ ("had" vowel) by the group of younger speakers, as compared to the older group. The mean values for F1 and F2 in the younger group are 863Hz and 1878Hz, respectively, while the same mean values for the older group are 716Hz and 2179Hz. The higher F1 (i.e. a more open vowel) values for younger speakers – a difference of 147Hz ($p \leq 0.001$) – between the two groups of speakers patterns with findings in [13], showing that short front vowels appear to be undergoing a reversal of the sound change that for a time saw the vowels raising. Further, the retraction is clear – a difference of 301Hz ($p \leq 0.001$) in the mean F2 values of the groups – also patterning with the findings [13].

Another point to note is the comparatively large overlaps shown by the ellipses for the "head" and "hid" vowels, /e/ and /ɪ/ respectively, in older speakers. The ellipse for production of /ɪ/ is almost entirely covered by the ellipse for production of /e/. However, for individual speakers there is a clear difference between vowel height in production of /e/ and /ɪ/ - /ɪ/ vowels for a speaker have a mean F1 value that's 73Hz lower than the corresponding speaker's /e/ vowel ($p \leq 0.001$).

These same patterns are seen when comparing the data for older and younger groups of male speakers

Overall, it can be seen that both genders of younger speakers appear to be reversing previously found raising and fronting of /æ/, patterning with the data found in [13]. While older speakers of both genders have /e/ and /ɪ/ productions that are very closely clustered (the females more so than the males), the individual speaker productions of these vowels are distinct from one another. And, while younger female speakers have considerably more distinct ellipse plots than their older counterparts, this is only true with regards the high front vowels for male speakers (who also a very broad range of realisations for /æ/, not seen in older speakers).

3.2 Frequency Data for the Short and Long Monophthongs of Tasmanian English for Younger Speakers

In this section, a description of the vowels of Tasmanian English is put forward. This incorporates the data shown in the previous sections, as well as the below figure (3) showing the mapped vowel spaces for both male and female younger speakers.

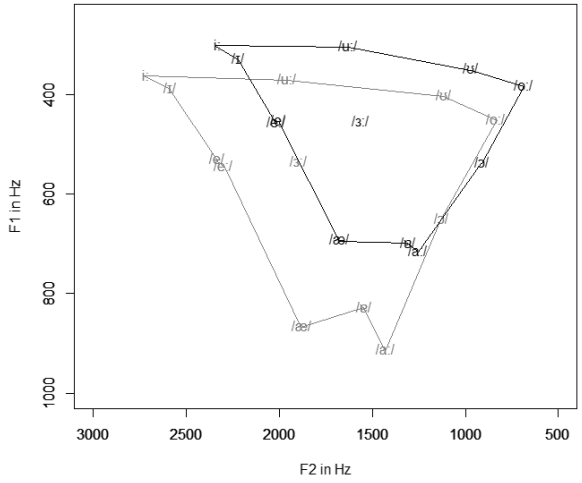

Figure 3: *A plot of the mean F1 and F2 midpoint values for the long and short vowels produced by younger female (black) and male (grey) speakers from Hobart.*

The most readily obvious gender difference seen in this figure is the difference in relative position for the low vowels: /ɐ:/ and /ɐ/ are where we see the first big differences between the two data sets. For male speakers, both these vowels are sitting very close together, with the short vowel only slightly higher and more forward than the long vowel. However, for the female speakers, the short vowel /ɐ/ is considerably higher than the long vowel /ɐ:/, being even higher than /æ/. In this way, the short vowel /ɐ/, for female speakers patterns similarly to how it does for Melbourne data in [4], while the long vowel /ɐ:/ is lower than /æ/, putting it in a relative position more similar to the Sydney data in [3].

3.3 Regional Variation

Below, in Figures 4 and 5, can be seen the comparative vowel spaces for female speakers from Hobart, Melbourne, and Sydney. Overall, the data for female speakers from Hobart have a general vowel space that appears retracted and raised, compared to the speakers from Melbourne and Sydney.

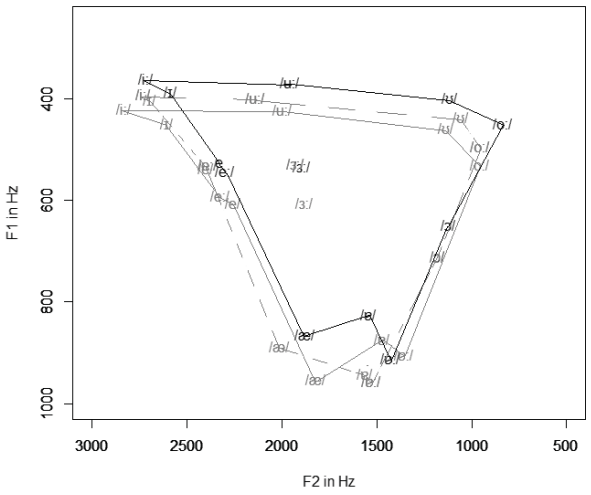

Figure 4: *Vowel spaces for Hobart female speakers (in black), Melbourne female speakers (in grey), and Sydney female speakers (the broken line)*

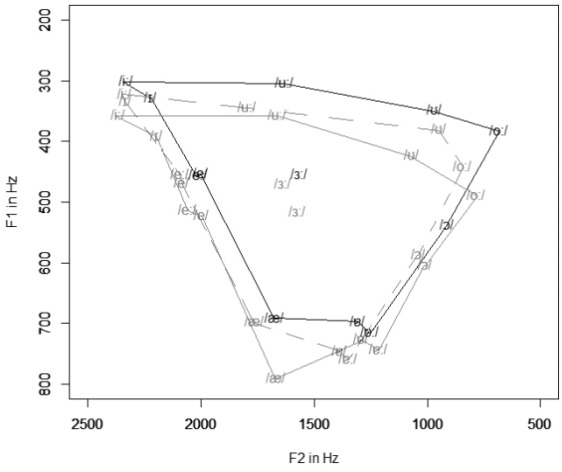

Figure 5: *Vowel spaces for Hobart male speakers (in black), Melbourne female speakers (in grey), and Sydney female speakers (the broken line)*

3.3.1 "Had" /æ/

The open front vowel /æ/ shows quite noticeable variation between the three states. The difference between Hobart and Melbourne is greater than that between Hobart and Sydney. Compared to speakers from Sydney, the speakers from Hobart show retraction and very slight raising in this vowel, with an F1 frequency difference of 23Hz and an F2 frequency difference of 135Hz. Compared to Melbourne data, /æ/ produced by speakers from Hobart is higher in the vowel space, the difference in F1 frequency between the two being 88Hz.

3.3.2"Hard" /ɐ:/ and "Hud" /ɐ/

These two vowels show a very similar pattern in their F1 and F2 frequencies between Hobart and Melbourne data sets. Both these vowels are almost plotted on top of one another for the speakers from Sydney. However, there is a clear difference in both the Hobart and Melbourne productions of these vowels, with the short vowel /ɐ/ being higher and further forward in the vowel space than the long vowel /ɐ:/. Both Hobart and Melbourne-based speakers have very similar targets for /ɐ:/ but Hobart speakers appear to produce an /ɐ/ that is higher and slightly further forward in the vowel space than speakers from Melbourne. The difference in F1 values for this vowel is 48Hz, and the difference in F2 values is 72Hz.

3.3.3 "Who'd" /ʉ/

Realisations of this vowel share a similar relationship between speakers from Hobart and speakers from Melbourne, as the vowel in "Hood". The difference isn't particularly great but the Hobart data set shows a lower mean F1 value for speakers from Hobart than is seen for speakers from Melbourne. The Sydney data, however sits roughly at the midpoint between Hobart and Melbourne, in F1 frequency value but is further forward in the vowel space, with a higher F2 value than either of the others. This pattern is repeated for male speakers.

4. Discussion

This study analysed variability in Standard Australian English spoken in Tasmania. In bringing Tasmania into the analysis, there were four research questions posed. This section examines the findings surrounding each of those questions.

4.1 A Description of the Vowel Space of Hobart English

The results of the acoustic analysis of the vowels of English as spoken in Hobart, showed vowel spaces shaped rather typically for Australian English, although there was some variability for some sounds. For example he "Had" vowel /æ/ is less open than that produced in Sydney or Melbourne (in particular), however it still shows patterning similar to that found in [13], in reversing the raising and fronting movement previously observed. This is a trend that is clearer in the F1/F2 frequency means for female speakers than it is for the male speakers, a point that is to be expected, considering that females have a tendency to lead males with regards to linguistic changes [14]. There is also a distinctly higher production of /ɐ/ as compared to /æ/ for female speakers than there is for male speakers.

4.2 Difference in Age Category

As discussed throughout the results, age differences occur for many of the vowels, indicating that the accent changes observed in [13] are, indeed, also occurring in a similar

manner in Tasmania as in other regions. Most vowels show differences in the mean formant values for one or both of F1 and F2 of their targets between younger and older groups, for both males and females. Overall, where the values show statistically significant differences, the younger speakers have higher mean formant frequency values than the older speakers, resulting in a lower and more front set of vowel realisations.

4.3 Regional Variations

There is some visible variation in the overall vowel plots between Tasmania, Victoria, and New South Wales, for both males and females, with what appears to be a general trend of a compressed vowel space for Tasmanian speakers, compared to their mainland counterparts. While this is not very great in extent, there are some larger differences to be seen in some of the low vowel targets, specifically the realisations of /æ/ show large variance between Tasmanian and Victorian female speakers of both genders, and there is a similarly large gap between production of /ɐ/between female Tasmanians and females from New South wales. There is also a clear difference in the data for the production of /o:/ between younger speakers from Tasmania and Victoria, with that vowel being produced noticeably higher and further back in the vowel space of Tasmanians, and speakers from New South Wales occupying a space in between the two. Conversely, the Tasmanian production of /ʉ:/ patterns similarly to Victorian production. That is, it's still quite a fronted vowel but not nearly to the same extent as that seen in New South Wales.

5. Conclusion

This study has added to the current knowledge on variation in Standard Australian English, by adding acoustic data for the vowels of Tasmanian English to the discussion, displaying vowel spaces for both male and female speakers. The study has shown evidence for some regional variation between reports on Standard Australian English as spoken in Sydney and Melbourne in some vowels, as well as variation between older speakers of this variety of Standard Australian English.

Being that this study was a one of the monophthongs, it would make sense that further study could be made into how the diphthongs of Tasmanian English pattern. Also, since vowel targets were analysed at a fixed point, a dynamic analysis of the vowel formants would likely be enlightening on this topic, in future.

6. References

[1] F. Cox, Australian English Pronunciation and Transcription, Melbourne: Cambridge University Press, 2012.

[2] R. Wardhaugh, An Introduction to Sociolinguistics, 5th ed., Carlton: Blackwell Publishing, 2006.

[3] F. Cox, "The Acoustic Characteristics of /hVd/ Vowels in the Speech of Some Australian Teenagers," *Journal of Australian Linguistics,* pp. 147-179, 2006.

[4] R. Billington, "Location, Location, Location! Regional Characteristics and National Patterns of Change in the Vowels of Melbourne Adolescents," *Australian Journal of Linguistics,* pp. 275-303, 2011.

[5] D. Burnham, D. Estival, S. Fazio, J. Viethen, J. Cox, R. Dale, S. Cassidy, J. Epps, R. Togneri, M. Wagner, Y. Kinoshita, R. Göcke, J. Arciuli, M. Onslow, T. Lewis, A. Butcher and J. Hajek, "Building an audio-visual corpus of Australian English: large corpus collection with an economical portable and replicable Black Box," in *Proceedings of the 12th Annual Conference o the International Speech Communication Association (Interspech, 2011)*, 2011.

[6] S. Cassidy, D. Estival, T. Jones, D. Burnham and J. Berghold, "The Alveo Virtual Laboratory: A Web Based Repository API," in *9th Language resources and Evaluation Conference (LREC 2014)*, Reykjavik, 2014.

[7] T. Kisler, F. Schiel and H. Sloetjes, "Signal processing via web services: the use case WebMAUS," in *Proceedings Digital Humanities*, Hamburg, Germany, 2012.

[8] RStudio Team, "RStudio: Integrated Development for R.," Boston, 2015.

[9] J. Harrington, The Phonetic Analysis of Spech Corpora, Munich: Blackwell Publishing, 2010.

[10] P. Boersma and D. Weenink, "Praat: doing phonetics by computer [Computer program]. Version 6.0.17," 2016. [Online]. Available: http://www.praat.org. [Accessed 21 April 2016].

[11] D. Bates, M. Mächler, B. Bolker and S. Walker, "Fitting Linear Mixed-Efects Models using lme4," *Journal of Statistical Software,* vol. 67, no. 1, pp. 1-48, 2015.

[12] T. Hothorn, F. Bretz and P. Westfall, "Simultaneous Inference in General Parametric Models," *Biometrical Journal,* vol. 50, no. 3, pp. 346-363, 2008.

[13] F. Cox and S. Palethorpe, "Reversal of short front vowel raising in Australian English," in *Proceedings of Interspeech 2008, 22nd-26th September 2008*, Brisbane, 2008.

[14] W. Labov, Principles of Linguistic Change, Oxford: Blackwell, 1992.

Background Specificity in Forensic Voice Comparison and Its Relation to the Bayesian Prior Probability

Michael Wagner[1], Yuko Kinoshita[2]

[1]Faculty of ESTeM, University of Canberra
[1]Research School of Computer Science, The Australian National University
[1]Quality and Usability Lab, Technical University of Berlin
[2]College of Arts and Social Science, The Australian National University
michael.wagner@canberra.edu.au; yuko.kinoshita@anu.edu.au

Abstract

This study investigates the effect of background data specificity on likelihood ratio and prior odds, and consequently on the posterior odds outcome. It is motivated by discussions on the correct choice of speaker recognition background, particularly in forensic voice comparison. We performed strictly controlled experiments with the ANDOSL database where background specificity is the sole independent variable. Results show that target and non-target scores are better separated with less specific background, but that in turn priors must be adjusted down. Because the risk of class recognition instead of individual recognition increases with lower background specificity, we suggest that the prior probability in the Bayes formula is factorised into one part that remains in the domain of the trier of fact – as is conventional – and another part that is related to the specificity of the assumed or agreed background.
Index Terms: Forensic voice comparison, Bayesian method, forensic prior probability, background specificity.

1. Introduction

The Bayesian approach used in forensic voice comparison (FVC) is similar in principle to that of non-forensic speaker authentication tasks. However, the interpretation of Bayesian likelihood ratios (LRs) is quite different in the FVC context.

In non-forensic speaker authentication, the Universal Background Model (UBM) has long been the standard method [1, 2, 3, 4]; contemporary methods, such as joint factor analysis, iVectors etc. are also implicitly based on the chosen UBM. Systems such as those prominent in recent NIST speaker recognition evaluations [5], generally use very large UBMs that represent speaker characteristics across dialects, accents and even languages spoken in multicultural societies, except that by common consensus, they are usually restricted to speakers of the same sex as that of the unknown speaker. To compensate for external factors, such as environmental noise and channel characteristics, speaker recognition scores are normalised with a cohort of speakers with similar attributes.

In FVC, the situation is somewhat more complicated. Within the context of the widely accepted Bayesian paradigm [6, 7], the forensic speech scientist estimates the likelihood ratio (LR) between the two likelihoods: 1) for the crime-scene recording to be consistent with the speaker model of the suspect (numerator of the LR) and 2) for the same recording to be consistent with the multi-speaker model of a background population (denominator of the LR). Although it is rarely performed explicitly in reaching the final decision, the trier of fact is required to combine the LR obtained from FVC with the prior odds $P(H_{so})/P(H_{do})$: the prior probability of the same-origin hypothesis (H_{so}–offender and suspect are the same person) versus the prior probability of the different-origin hypothesis (H_{do}–offender and suspect are different persons).

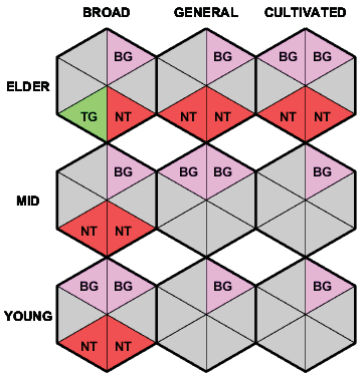

Figure 1. *Partitioning of the non-accented male speakers of ANDOSL into 9 subgroups, each of 6 speakers, and speaker partitioning into target (TG), non-target (NT) and background (BG) speakers.*

The determination of the prior odds is usually considered the domain of the court, as the forensic scientist does not have access to information on the case other than the voice recordings. Combining those prior odds with the forensic scientist's LR $p(X|H_{so})/p(X|H_{do})$ yields the posterior odds $P(H_{so}|X)/P(H_{do}|X)$ for the same-origin hypothesis versus the different-origin hypothesis according to the Bayes Rule [7]

$$\frac{P(H_{so}|X)}{P(H_{do}|X)} = \frac{P(H_{so})}{P(H_{do})}\frac{p(X|H_{so})}{p(X|H_{do})}. \quad (1)$$

In FVC, H_{do} plays a pivotal role in selecting the background population, and this has been the subject of much debate among forensic scientists. Some have argued that the background population should be tailored to the characteristics of the offender, the suspect or both [8]. It has also been argued that the background population should be based on those characteristics of the offender's voice that both prosecution and defence agree upon [6] or that it should represent a set of speakers sufficiently similar to the offender's voice that an investigating police officer would bother submitting voice samples for examination by a forensic scientist at all [9].

In forensic casework, the different-origin hypothesis defines the subpopulation to which the offender apparently belongs and to which any suspects should also belong. Those attributes of the subpopulation that can affect speech acoustics (e.g. language variety, gender, age range) will set the selection criteria for the background population data for the case. As should be clear from Eq. (1), the selection of the background affects the forensic scientist's LR through $p(X|H_{do})$ as well as the trier of fact's assignment of prior odds through $P(H_{do})$.

A commonly used imaginary forensic examination illustrates the two effects: Assuming a criminal investigation on an island of 100 inhabitants, if nothing other than being a member of the island population is known about the offender, the prior odds $P(H_{so})/P(H_{do})$ would be 1/99, and the forensic scientist should build a background model from a representative sample of the entire island population. If however, in addition, the offender were known to be a member of the female half of the population, the prior odds would increase to 1/49, and the forensic scientist would build the background model only from the female subpopulation. Any additional knowledge about the offender would further raise the prior odds and, at the same time, likely diminish the acoustic variance of the background.

While the characteristics of the background population such as gender or dialect should be consistent with an agreed H_{do}, in practice the forensic scientist's choices are often constrained by data availability, time and resources. Sometimes there is not even a clear reference to an agreed H_{do}. This is problematic for the validity of the Bayesian estimation of the posterior odds unless the scientist informs the trier of fact of this relationship between H_{do} and choice of background data on one hand, and LR, prior and posterior odds on the other.

In the remainder of this paper, we thus examine how the specificity of the different-origin hypothesis and the corresponding selection of the background population affect the outcome of forensic voice comparison: firstly through $p(X|H_{do})$ and the resulting LR and secondly through $P(H_{do})$ and its effect on the prior odds estimated by the trier of fact.

2. Experiment

2.1. Data

The Australian National Database for Spoken Language (ANDOSL) [10] comprises Australian English speech data from a range of speakers, varying in age, sex, and their variety of Australian English. For this study, we utilise only the read-sentence data by the ANDOSL male native speakers. Within that population, we have a 3×3×6 partitioning into 3 age groups, elder, mid, young, the 3 sociolect groups of Australian English, broad, general, cultivated, on the basis of the tag provided in ANDOSL [11], and 6 speakers in each group, as is illustrated in Fig. 1.

Each of the 9 subpopulations is shown as a hexagon, and the 6 speakers of each subpopulation are shown as colour-coded triangles. The single target speaker is shown as the green triangle, tagged *TG*. The non-target speakers are shown as red triangles, tagged *NT*, 5 in the first row for the *elder* subpopulation and 5 in the first column for the *broad* subpopulation. The 12 background speakers, tagged *BG*, are shown as magenta triangles. This design ensures that there is no overlap between target, non-target and background speakers, hence avoiding a potential statistical bias.

Each of the 54 male speakers read 200 sentences that were designed to cover the entire acoustic-phonetic space of Australian English. Of those, 180 were used solely for training background models. Using 180 sentences for UBM training, we assume that the background models cover the acoustic-phonetic space of the background population sufficiently. Of the remaining 20 sentences, 10 were used solely for maximum-a-posteriori (MAP) adaptation of the target-speaker models, and 10 were used solely for the target and non-target testing. We consider using 10 sentences for GMM adaptation forensically realistic, given the typical constraints of FVC casework, where suspects are often uncommunicative during police interviews and provide precious little material for the adaptation of the target-speaker GMM.

Recordings of the 3×3×6×200=10,800 sentences are stored as wav files, sampled at 20,000 samples/s and 16 bits/sample. 12 mel-frequency cepstral coefficients (MFCC) and log energy were determined for 20ms windows shifted in 10ms steps. Derivative coefficients were discarded as the amount of data was insufficient for training higher-dimensional models. Using a simple absolute energy threshold, low-energy frames were eliminated and about 61% of the frames retained, yielding on average 313 feature vectors per sentence for the analysis.

2.2. Experimental design

The read-speech data in ANDOSL were produced under highly controlled conditions, and each speaker was recorded in a single session. ANDOSL is thus generally regarded as an inadequate database for FVC experiments. However, our experimental design turns this limitation into an advantage. The single-session nature of ANDOSL and the read-speech material enabled perfect control over the independent variables of our design. Being a single session recording eliminates extraneous variation such as intersession and channel variation as well as intra-speaker variation in health or emotion. Using speech material read from prepared texts, we exclude the variability in quantity and phonetic contents that is inevitable with spontaneous speech data. Therefore, we can reasonably interpret any effects on the output as being caused by the chosen background specificity–the independent variable of our design.

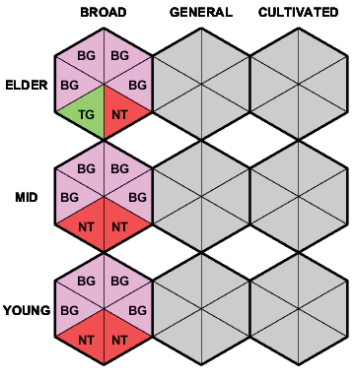

Figure 2. *Speaker verification in the broad subpopulation with specific background.*

We conducted altogether 4 experiments. In the first, the target speaker, the 5 non-target speakers and the 12 background speakers are all from the subpopulation of the *broad* sociolect speakers as shown in Fig. 2. In the second experiment, the target speaker, the 5 non-target speakers and the 12 background speakers are all from the *elder* subpopulation as shown in Fig. 3. Each experiment proceeded to build a UBM

from the background speakers, building a GMM for the target speaker, and determining LRs for target and non-target tests.

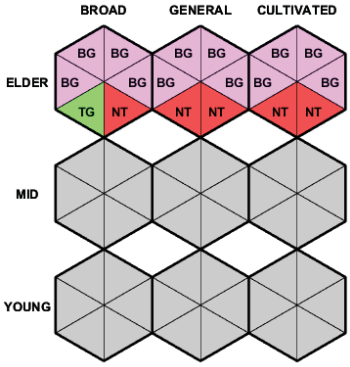

Figure 3. *Speaker verification in the elder subpopulation with specific background.*

We then repeated these 2 experiments altering the specificity of the UBM by building it from the background speakers drawn from the entire population of male speakers as shown in Fig. 1, while the other factors, i.e. target and non-target speakers and the sentence material, were kept identical. For the comparison of system performance, LLRs are usually normalised or calibrated for environmental or channel variation between recordings. However, the current study does not require such a step due to the strict control in experimental design.

Fig. 4 depicts our experimental design schematically. We constructed a UBM from the 180 designated sentences spoken by the 12 chosen background speakers. For the target speaker, we MAP-adapted this UBM to a target-speaker-specific GMM using the 10 designated adaptation sentences of the target speaker. Finally, 10 target trials were conducted with the designated test sentences of the target speaker, and 50 non-target trials were conducted with the same test sentences spoken by 5 non-target speakers. For each trial, we produced the sentence-mean log likelihoods and log likelihood-ratios with respect to the target GMM and the UBM.

For the *broad*-sociolect subpopulation, we conducted the following pair of experiments: firstly, we built a UBM from the 12 designated background speakers of the subpopulation as illustrated in Fig. 2. For the designated target speaker, we adapted the UBM to build the target-speaker GMM and conducted the target trials. Then we conducted the non-target trials with the designated 5 non-target speakers of the subpopulation. Both target and non-target trials consist of the designated 10 testing sentences for each target and non-target speaker. That experiment was repeated with the same target and non-target speakers, but with the 12 background speakers drawn from the full population as shown in Fig. 1 and the target-speaker GMMs adapted from that wider-background UBM.

An equivalent set of experiments was then conducted with the other subpopulation under investigation, *elder* speakers. Here, the background speakers were drawn from this subpopulation and from the full population as shown in Fig. 3 and again in Fig. 1. For both pairs of experiments, the independent variable is the specificity of the background: either specific to the subpopulation matching the test speaker characteristics or less specific by being a superset of that subpopulation–in our case the entire population of the male speakers in ANDOSL.

For each trial, we observe the log likelihoods (LL) $log\ p(X|H_{so})$ and $log\ p(X|H_{do})$ for each sentence X, each being the mean of the frame log likelihoods for the sentence. We also observe the resulting log likelihood-ratios $LLR = log\ p(X|H_{so}) - log\ p(X|H_{do})$. The statistics of the above LLs and LLRs are the dependent variables of the design, while background specificity is the independent variable.

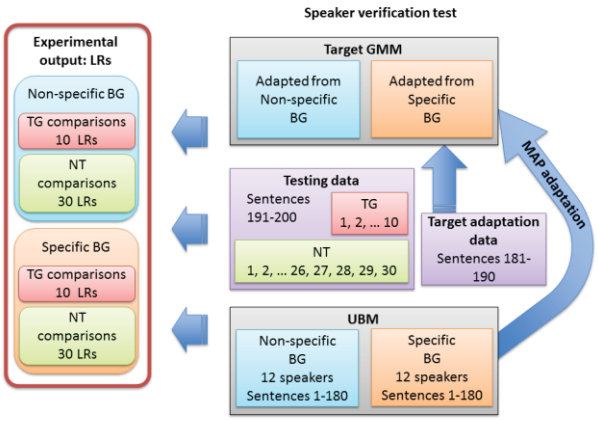

Figure 4. *Experimental speaker recognition system showing the parallel evaluations of non-specific background (blue boxes) and specific background (orange boxes)*

3. Results and discussion

Table 1 presents the mean LLRs for the target and non-target tests for the 2 sets of experiments and their mean differences ΔLLR as well as their log-likelihood-ratio costs C_{llr} [12]. The last 2 rows combine the 2 specific-background and the 2 non-specific-background experiments.

Table 1. *Target and non-target LLRs and their differences.*

Target/UBM	TG LLR	NT LLR	ΔLLR	Cllr
Broad/Broad	2.664	-0.174	2.837	0.496
Broad/Non-specific	2.921	-0.174	3.096	0.491
Elder/Elder	2.682	-0.409	3.091	0.420
Elder/Non-specific	2.921	-0.311	3.232	0.438
Mean Specific	2.673	-0.291	2.964	0.458
Mean Non-specific	2.921	-0.243	3.164	0.464

The results show that the target and non-target scores are separated better for the non-specific background than for the specific background. Fig. 5 shows in addition the distribution of the numerator LLs (LLG) and the denominator LLs (LLU) in Eq. (1). The curves are based on the means and variances of the LLs and a normality assumption for the distributions.

The 2 largely overlapping narrow (green) Gaussians at the right of Fig. 5a show the distributions of the numerator LLs of the *broad* target trials against the specific UBM (dashed line) and against the non-specific UBM (full line). The specificity of the UBM seems to affect neither the mean nor the variance of those LLs. The other 2 narrow (black) Gaussians near the centre of Fig. 5a show the distributions of the denominator LLs of the *broad* target trials against the specific and non-specific UBMs. Those distributions show that the non-specific UBM produces distinctly smaller denominator LLs than the specific UBM. Since the numerator LLs are distributed almost identically, it follows that the non-specific UBM produces higher LLRs than the specific UBM.

Fig. 5b shows the corresponding 4 curves for numerator and denominator LLs against specific and non-specific UBMs for the *elder* subpopulation with the same trends as found for Fig. 5a. Each of the 2 figures also shows 4 wide Gaussians for

the non-target trials with the respective specific numerator (green) and denominator (black) LLs closely overlapping and the specific UBM (dashed) yielding a slightly larger mean LL than the non-specific UBM (full) as could be expected.

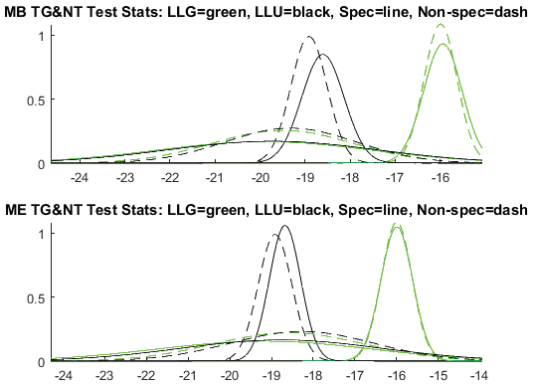

Figure 5. *LL statistics of target trials (narrow) and non-target trials (wide) against target GMM (green) and UBM (black): (a) Broad subset; (b) Elder subset (Note the almost complete overlap between the 4 pairs of wide curves in the 2 figures)*

In both cases, the background specificity affected non-target comparisons less. Our results show a larger distance between target and non-target scores for the non-specific UBM, which seems to be due to the larger denominator LLs of the non-target scores for the non-specific UBM. Table 1 shows no significant differences of C_{llr} between the specific and the non-specific background in either experiment.

From this rather small and therefore limited study, it appears that the specificity of the background is not of major consequence in FVC or even that a less specific background may be preferable for reasons of the slightly larger distance between target and non-target scores found here. However, this interpretation must be weighed against 2 other factors: Firstly, a less specific UBM has the tendency to turn the speaker recognition problem into a recognition of the subpopulation. In other words, there is a danger of recognising, for example, the language variety of a speaker instead of recognising the individual. And secondly, a less specific UBM implies a proportionately smaller prior probability for the defence hypothesis and correspondingly a less conclusive outcome in terms of the posterior odds of the analysis.

For example: an offender is only known to be a member of a population of 8 million Australian adult males. Non-acoustic evidence such as height, eye and hair colour place him in 10% of that population. Using Row 2 of Table 1, the forensic scientist reports a likelihood ratio of $e^{3.096} = 22.109$ against that background population with corresponding posterior odds of $1/7{,}999{,}999 \times 10\% \times 22.109 = 0.276 \times 10^{-6}$.

However, if on the acoustic evidence the forensic scientist determines that the offender is a member of the *broad* accent group of 2 million males and, according to Row 1 of Table 1, reports a likelihood ratio of $e^{2.837} = 17.064$ against that more specific background population, the corresponding posterior odds are $1/1{,}999{,}999 \times 10\% \times 17.064 = 0.853 \times 10^{-6}$, a value about 3 times larger than with non-specific background.

This example illustrates the case for factoring the prior odds into one part that represents the non-acoustic evidence (10% in our example) and another part that represents the size of the background population used by the forensic scientist.

4. Conclusions

A small-scale preliminary study was conducted to investigate how the specificity of the background population affects FVC, using age and sociolect specific subpopulations in the ANDOSL database. LLRs as well as the constituting numerator and denominator LLs were determined dependent on background specificity. Our small-scale experiments show that the denominator LLs for the target speaker were smaller for less specific UBMs and hence those LLRs were larger and the target-non-target separation was larger for less specific UBMs. However, a less specific background bears the risk of inadvertently performing class recognition instead of individual recognition. Further experiments with larger datasets and varying degree of specificity should be conducted.

Also, perhaps more importantly, we must be mindful that the choice of background population directly affects the determination of the prior odds and thus the interpretation of the forensic voice comparison by the trier of fact. In the example presented in this study, assuming equal distribution of the subpopulations, the choice of the full male population of ANDOSL for the UBM would imply prior odds 3 times smaller than the choice of the specific subpopulations.

It is therefore most important for the forensic scientist to report as precisely as possible the characteristics of the background database and its implications for the determination of the LRs and posterior odds of the forensic voice comparison.

5. References

[1] A.E. Rosenberg, J. DeLong, C.H. Lee, B.H. Juang, F.K. Soong, "The use of cohort normalized scores for speaker verification", *Proc. Int. Conf. on Spoken Language Processing*, 599-602, 1992.

[2] J.B. Millar, F. Chen, M. Wagner, X. Zhu, "The efficacy of cohort normalisation in a speaker verification task under different types of speech signal variance", *Proc. Austr. Int. Conf. on Speech Science and Technology*, 850-855, 1994.

[3] S. Furui, "Recent advances in speaker recognition", *Proc. 1st Int. Conf. on audio- and video-based biometric person authentication*, 237-252, 1997.

[4] D.A. Reynolds, T.F. Quatieri, R.B. Dunn, "Speaker verification using adapted Gaussian mixture models", *Digital Signal Processing*, 10, 19-41, 2000.

[5] G.R. Doddington, M.A. Przybocki, A.F. Martin, D.A. Reynolds, "NIST speaker recognition evaluation - overview, methodology, systems, results, perspective", *Speech Communication*, 31, 225-254, 2000.

[6] P. Rose, *Forensic speaker identification*, London: Taylor and Francis, 2002.

[7] G.S. Morrison, "Forensic voice comparison" in I. Freckelton and H. Selby [eds], *Expert Evidence*, Ch. 99, Sydney: Thomson Reuters, 2010.

[8] N. Brümmer, E. de Villiers, "What is the 'relevant population' in Bayesian forensic inference?", downloaded on 30 March 2016 from http://arxiv.org/pdf/1403.6008v1.pdf, 2014.

[9] G.S. Morrison, F. Ochoa, T. Thiruvaran, "Database selection for forensic voice comparison", *Proc. Odyssey 2012*, 62-77, 2012.

[10] J. Vonwiller, I. Rogers, C. Cleirigh, and W. Lewis, "Speaker and material selection for the Australian national database of spoken language", *Journal of Quantitative Linguistics*, 2, 177-211, 1995.

[11] J. Harrington, F. Cox, and Z. Evans, "An acoustic phonetic study of broad, general, and cultivated Australian English vowels," *Australian Journal of Linguistics*, 17:2, pp. 155-184, 1997.

[12] N. Brümmer, J. du Preez, 2006. "Application-independent evaluation of speaker detection," *Computer Speech & Language*, 20, 230–275.

Author Index