DISCRIMINATION OF SPEAKERS USING TONE AND FORMANT DYNAMICS IN THAI

Sukanya Thaitechawat & Paul Foulkes

University of York, UK
st584@york.ac.uk; paul.foulkes@york.ac.uk

ABSTRACT

Dynamic properties of speech have been identified as offering greater potential than static features to discriminate between speakers. They may therefore offer useful evidence in forensic speaker comparison analyses. This exploratory study assesses the speaker specificity of diphthong and tone trajectories in Thai. Data were analysed from five male standard Thai speakers. Discriminant analysis (DA) yielded excellent results, with up to 100% correct attribution. The addition of tone data improved DA results based on formants alone.

Keywords: forensic phonetics, Thai, tone, formants, discriminant analysis

1. INTRODUCTION

It is now accepted that voice alone cannot be used to establish a speaker's identity with absolute certainty. Despite the continuing use of the term 'voiceprint', there is no indelible and uniquely identifying feature of any voice that is the equivalent of a fingerprint.

However, individual voices clearly differ from one another. Vocal features are routinely identified in speaker comparison cases, and such evidence may be used, albeit with a degree of caution, to help establish whether a suspect could or could not have been the speaker in an evidential recording. One of the most important goals of forensic phonetics is therefore to identify those features of speech or voice that *best* distinguish speakers from one another. In principle such features display little within-speaker variation (i.e. they are relatively stable across time and phonological context), but large cross-speaker variation.

In automatic speaker recognition research, the discriminatory power is tested of abstract, holistic parameters extracted from the speech signal. In forensic phonetics, by contrast, attention has been given to the discriminatory power of specific vocal features, especially segments and fundamental frequency. Some of the most promising results have been reported in studies examining dynamic properties of speech, such as diphthong trajectories and spectral change across segment sequences, e.g. [2, 4]. The hypothesis underlying such research is that cross-speaker variation is most likely to be observed in the transition between segments rather than in the physical centre of segments (as traditional vowel-midpoint analysis might show, for example). This is because segment targets are specified by the phonology, and are thus shared by speakers, while individuals are free to travel their own articulatory paths between targets.

The present study develops this line of research by examining dynamic features in Thai diphthongs. In addition to formant analysis it also considers dynamic properties of tone. Previous work on speech dynamics has dealt mainly with segmental properties. However, like segments, tone varies systematically as a function of dialect and sociolinguistic background [1]. It is thus reasonable to hypothesize that tonal features may also vary according to the individual speaker. The study is furthermore the first published forensic phonetic work on Thai.

2. THAI PHONOLOGY

Like any language, Thai displays regional and social variation. We therefore limit our discussion to Standard Thai as spoken in Bangkok. This variety is considered prestigious, is taught as a model in schools, and is used in mainstream media.

Thai has 9 monophthongs, all with contrastive length, and three falling diphthongs, /ia, ua, ɯa/. It has five tones, which are generally numbered and labelled as shown in Table 1, illustrating a minimal set with reference to /k^{h}a/.

Table 1: minimal set of contrastive tones, /k^{h}a/.

tone	gloss
1 mid	*to be stuck*
2 low	*galangal*
3 falling	*to kill*
4 high	*to trade*
5 rising	*leg*

Despite being considered level tones, there is generally a degree of f0 movement observable in tones 2 and 4 [3]. Phonotactic constraints mean that not all tones occur with all syllable structures.

3. HYPOTHESES

Adopting the assumption that dynamic properties offer the best potential for speaker discrimination [4], our analysis focused on the three diphthongs. The dynamic properties of all five tones were analysed. We predicted that the contour tones 3 and 5 would yield better discrimination than the level tones since they involve a greater degree of acoustic change, and thus more freedom for individual variation. We further hypothesized that the addition of tonal information would improve discrimination results based on vowel formant properties alone.

4. METHOD

4.1. Speakers

Five male speakers of Standard Thai were recruited, aged 24-28.

4.2. Materials

Word lists were constructed to elicit minimal sets containing combinations of each diphthong and tone. For each diphthong*tone combination sets were devised with initial /b- p- l- s-/, and both open and closed syllables, /CVV/ and /CVVN/. The word lists therefore consisted of 120 items (3 vowels x 5 tones x 2 syllable types x 4 initials).

In most cases the target items were isolated words. To complete the minimal sets it was necessary to use some bisyllabic words, and to create a number of nonce words that were, however, phonologically possible in Thai.

4.3. Recordings

Recordings were conducted in a recording suite in two or three sessions over a two week period, to minimize speaker fatigue. Target words were shown in Thai script on a PowerPoint presentation. They appeared individually and in random order to avoid list effects. The recording administrator (the first author) judged the acceptability of each production and asked speakers to read the words again if necessary. Speakers read the lists five times each. There were therefore 3,000 tokens recorded (120 words x 5 speakers x 5 repetitions).

Recordings were made with a Neumann U87i cardioid microphone, situated ca. 15 cm from the speaker, direct into Adobe Audition v. 1.0 mounted on a standard PC running Windows XP. Mono recordings were made at a sample rate of 44.1 kHz, 16 bit depth.

4.4. Acoustic analysis

Each of the 3,000 tokens was edited into a separate .wav file in Praat v. 5.0.22. Segmentation of the vowel was then recorded in a text grid. Segmentation was performed with reference to both spectral and amplitude properties [6]. After /p-/ and /s-/ the start of the vowel was identified as the onset of periodic energy; after /b-/ it was identified coincident with the release burst of the stop; after /l-/ the start of the vowel was marked at the point in the spectrogram at which stronger amplitude formant structure became apparent, occasionally also with reference to a weak release transient. For /CVV/ tokens the end of the vowel was identified at the final vertical transient visible in the spectrogram. For /CVVN/ tokens the end of the vowel was marked where amplitude and formant structure weakened for the nasal.

Praat scripts were then applied to record values of f0, F1, F2 and F3 across the segmented vowels. The scripts performed time normalization by dividing the vowel duration into ten equal sections. Formant and f0 values were recorded at nine intervals: at 10%, 20%... 90% through the vowel.

For formant measurement, settings were generally applied to identify 5 formants in a 5.5 kHz range. Manual correction was made of erroneous results by adjusting the Praat settings or by hand measurement.

Praat's f0 measurement parameters were set to the range 70–300 Hz. Values were corrected where necessary by adjusting f0 settings or via manual calculation of the duration between glottal pulses (especially where low or falling tones descended into creak). 32 tokens were discarded (1.1% of the data) because of insurmountable difficulty in obtaining satisfactory measurements.

4.5. Discriminant analysis

Discriminant analysis (DA) was used to assess the power of the features analysed to discriminate between speakers. DA builds a predictive model for each speaker based on known data, then attributes unknown samples to these speaker models [5]. A statistic is reported summarising the

proportion of data correctly attributed. Thus a DA result of 100% indicates that all data have been correctly attributed. Chance performance in this data set of five speakers is 20%. DA was performed using SPSS version 17.

The 'leave-one-out method' was used, such that each token in a speaker's data set would be treated as an unknown sample, and the remaining data would be used to build the speaker model. The best performance is then reported by the DA. Following standard practice outliers were removed before DA proceeded [5]. A total of 83 outliers were removed, leaving 2,885 tokens for the DA.

Separate DA runs were conducted on each of 30 word types (3 vowels x 5 tones x 2 syllable types). A limitation of DA is that the maximum number of predictors in a speaker model must be less than the number of tokens produced by that speaker [5]. In our analysis we had 36 predictors (9 x 3 formant values + 9 f0 values), but a maximum of 20 tokens per word per speaker. Thus no more than 19 predictors could be used, and in some cases fewer because of discarded tokens. The lowest number of predictors used was 14. To judge which of the 36 possible predictors to include in any speaker*word type combination, ANOVAs were run on each set of data to identify F-ratios for the predictors. The predictors with the highest F-ratios were chosen for inclusion in the DA.

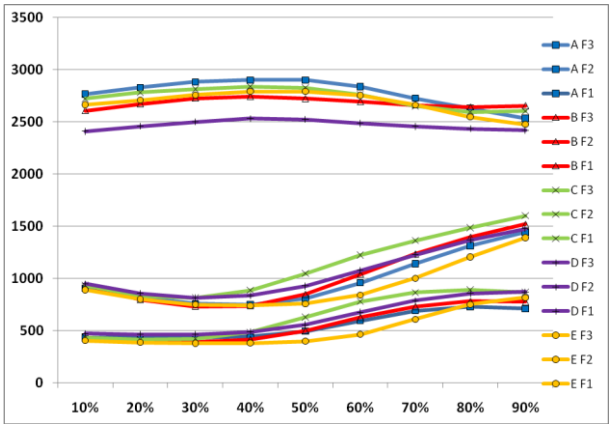
Three DA runs were then conducted for each word type using the best predictors: (i) solely from formants, (ii) solely from f0, and (iii) using the best combined set of f0 and formant data.

5. RESULTS

5.1. Illustrative example: formants

As expected, formant patterns varied across the speakers. Variation was found both in frequencies and dynamic pattern, i.e. the speed and trajectory of formant movement. Figure 1 shows mean formant values for /ua/ (all tones combined). It can be seen that all five speakers show similar F1 values at the start of the vowel, but differ considerably between the 50-70% points (i.e. the transition into the second element of the diphthong). Speakers C and E depart markedly from the F1 values of the other speakers. In F2 divergence between speakers can be observed from the 40% point, with C and E separated by around 300 Hz throughout the remainder of the vowel. The rise in F2 begins much earlier for C than E. F3 varies in both frequency and degree of movement. D's F3 is much lower than that of other speakers, and also flat (cf. A and E).

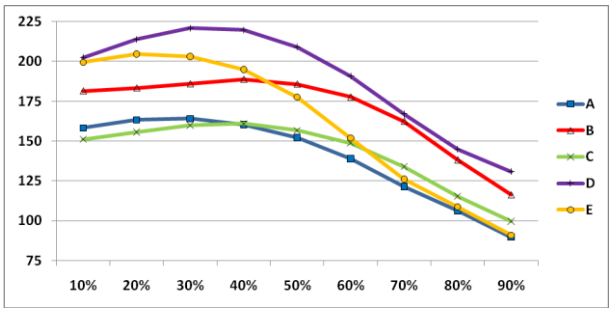
Figure 1: mean values of F1, F2 and F3 (Hz), /ua/.



5.2. Illustrative example: tone

Tonal patterns also showed variation across individuals. Figure 2 illustrates differences in the contours of falling tones. All speakers in fact show a small rise at the start of the vowel, with the subsequent descent varying markedly in both steepness and timing. For instance, speaker C's tone falls by 61 Hz on average, compared with 104 Hz for speaker E. The descent for B occurs around the 50% point, while A and E show a considerably earlier drop.

Figure 2: mean values of f0 (Hz), falling tone.



5.3. Discriminant analysis

DA results are summarized in Table 2. Results are shown for runs using formant data only, f0 data only, and the best combination of predictors.

Overall the DA results are very high, with several cases of 100% correct attribution. For formant data alone results are generally above 90%. For f0 data alone scores are always lower than those for the corresponding formant data, ranging from 72-88% correct. DA runs with combined formant and f0 data generally show an improvement on formant data alone. Rising tone

yields the best results using any predictor set (average correct 98% with formant + f0 data).

Table 2: DA results (% correct attribution).

tone	syllable	F1-3	f0	combined
1 mid	ia	89.7	77.3	89.7
	ua	94.8	84.4	100.0
	ɯa	93.8	79.4	92.8
	iaN	88.8	78.6	94.9
	uaN	97.9	81.4	99.0
	ɯaN	91.7	87.5	96.9
	average	92.8	81.4	95.6
2 low	ia	92.9	79.6	93.9
	ua	90.5	82.1	100.0
	ɯa	93.7	86.3	97.9
	iaN	87.8	73.3	92.2
	uaN	95.7	75.5	95.7
	ɯaN	93.6	87.2	98.9
	average	92.4	80.7	96.4
3 falling	ia	94.1	78.4	95.1
	ua	95.7	88.3	94.7
	ɯa	94.6	86.0	97.8
	iaN	94.3	78.2	94.3
	uaN	93.7	80.0	98.9
	ɯaN	97.1	81.4	97.1
	average	94.9	82.1	96.3
4 high	ia	95.8	81.1	96.8
	ua	90.8	72.4	95.5
	ɯa	96.9	82.5	100.0
	iaN	95.0	80.0	85.0
	uaN	100.0	79.2	96.9
	ɯaN	96.9	84.5	96.9
	average	95.9	80.0	95.2
5 rising	ia	88.0	87.1	100.0
	ua	96.0	83.8	99.0
	ɯa	96.9	80.4	100.0
	iaN	91.9	74.7	92.9
	uaN	96.9	85.6	97.9
	ɯaN	94.8	88.7	100.0
	average	94.1	83.4	98.3

6. DISCUSSION

The initial hypotheses are supported. The data show an excellent performance in discriminating between speakers using dynamic measurements.

Tonal variation was observed across speakers in both frequency and timing relative to segmental material. It was predicted that contour tones would discriminate better than level tones. This is borne out with respect to rising tone. Falling tone also yielded good DA scores, with the second best performance when f0 data alone were used in the DA run. It was further predicted that the addition of f0 data would improve the overall DA performance, which was indeed the case in respect of most word types. It is not surprising that f0 alone did not yield results as good as those from formants, since the f0 analysis was based on only 9 predictors compared with as many as 19 for the formant and combined analyses.

The results compare very well with those reported in other studies of segmental features. DA rates of 88-96% are reported for /aɪ/ for five speakers of Australian English [4], and a high of 88% for /jœː/ in Swedish [2].

7. CONCLUSION

This study furthers our understanding of individual phonetic variation. Individuals vary in the acoustic patterns they use in both vocalic and tonal features. The extent of cross-speaker variation and the within-speaker consistency of patterns means that measured data served to discriminate between speakers with a very high degree of success.

These preliminary data are promising, but are drawn from a small speaker sample and controlled laboratory speech. Further testing is required on larger data sets and spontaneous speech to assess the potential of the features in speaker comparison.

Finally, the study is the first to our knowledge to test claims made in forensic phonetics with reference to Thai. Work in this field has overwhelmingly concentrated on English and other widely-spoken European languages. We hope that others will continue to assess the robustness of the received wisdom with reference to typologically and sociolinguistically different languages.

8. REFERENCES

[1] Cruttenden, A. 1997. *Intonation (2nd ed.)* Cambridge: CUP.

[2] Eriksson, E., Sullivan, K. 2008. An investigation of the effectiveness of a Swedish glide + vowel segment for speaker discrimination. *Int. J. of Speech, Language and the Law* 15, 51-66.

[3] Gandour, J., Potisuk, S., Ponglorpisit, S., Dechongkit, S. 1991. Inter- and intra-speaker variability in fundamental frequency of Thai tones. *Speech Communication* 10, 355-372.

[4] McDougall, K. 2004. Speaker-specific formant dynamics: An experiment on Australian English /aɪ/. *Int. J. of Speech, Language and the Law* 11, 103-130.

[5] Tabachnick, B, Fidell, L. 2007. *Using Multivariate Statistics* (5th ed.). Boston: Allyn and Bacon.

[6] Turk. A., Nakai, S. Sugahara, M. 2006 Acoustic segment durations in prosodic research: A practical guide. In Sudhoff, S. et al (eds.), *Methods in Empirical Prosody Research*. Berlin: De Gruyter, 1-28.