Disyllabic parameterisation of Vietnamese tonal F0 trajectories in likelihood ratio-based forensic voice comparison

Michael Carne[1] & Shunichi Ishihara[1]
[1]The Australian National University, Canberra, Australia
michael.carne@anu.edu.au, shunichi.ishihara@anu.edu.au

ABSTRACT

The paper presents a preliminary investigation of the performance of acoustic-phonetic based forensic voice comparison features derived from tonal fundamental frequency (F0) trajectories parameterised over disyllabic words, rather than individual syllables as is typically done. Its aim was to see whether the disyllabic parameterisation led to any improvements in the validity and reliability of the voice comparison using speech data from 10 native Vietnamese speakers. Polynomial functions were modelled to speakers' raw F0 trajectories and feature vectors constructed from the model coefficients. Likelihood ratios were calculated using the multivariate kernel density likelihood ratio (MVKD-LR) formula and performance assessed using the log-likelihood ratio cost function (Cllr) (measured in bits) and Bayesian credible interval (BCI). The best performing disyllabic system achieved a Cllr of 0.40, representing a 0.15 and 0.25 bit improvement in validity over the syllable based systems.

1 INTRODUCTION

Fundamental frequency (F0) is a commonly used feature in acoustic-phonetic approaches to forensic voice comparison (FVC). A majority of previous work examining F0 in FVC has focused on the potential of so called long-term F0 features, based on the long term mean or other distributional characteristics of speakers' F0 (e.g. [6,7]). This focus is largely a consequence of the languages explored, where the function of F0 is primarily non-linguistic. Likelihood ratio-based (LR-based) FVC studies examining the potential of tonal acoustics are relatively few. Those that exist are for Cantonese (e.g. [8, 16, 19]) and Thai [13]. These studies have shown features based on speakers' tonal F0 trajectories from monosyllables to perform relatively poorly, unless combined with other features.

Since languages differ in the way speaker-specific information is encoded, it is important to test features independently for different languages [14]. There are no existing LR-based FVC studies for Vietnamese based on acoustic-phonetic features; this is one motivation for the current study. The main aim though is to determine whether parameterisation of tonal F0 trajectories over disyllabic words offers better FVC performance than the typically monosyllabic parameterisation. This is motivated by a relatively recent study demonstrating improvements in Cantonese FVC applying this approach [16]. The likelihood ratio approach (LR approach) is used to test to what extent this applies to disyllabic words in Vietnamese as well. The performance of the disyllabic F0 parameterisation for two test words and their constituent syllables are evaluated separately in terms of their validity (accuracy) and reliability (precision).

1.1 The likelihood ratio approach

The LR approach is increasingly recognised as logically and legally correct framework for assessing voice evidence [5, 15]. LR-based FVC involves calculating $p(E|H_{ss})$ the probability of the voice evidence (E_{Sp}) assuming it originates from the same speaker (H_{ss}) vs. $p(E|H_{ds})$, which is the probability assuming the samples are from different speakers (H_{ds}). These are expressed as ratio of conditional probabilities: $p(E_{Sp} \mid H_{ss})/p(E_{Sp}|H_{ds})$. The numerator of the LR expresses the similarity between the suspect and offender samples, divided by their typicality [14].

The result is a measure of the relative strength of evidence in favour of one of the hypotheses. The LRs direction and magnitude are proportional to the strength of evidence: an LR greater than unity (i.e. 1) gives support in favour of H_{ss}, while an LR less than unity supports H_{ds}. On a logarithmic scale, log-likelihood ratio values < 0 favour the H_{ds} and those > 0 support H_{ss}.

1.2 Vietnamese tonal system

In this study speech data was elicited from Vietnamese speakers from the Da Nang region, Viet Nam. The tonal system of speakers from this region have been categorised as speakers of the Southern Vietnamese dialect [18], however, differences in the realisation of tones in terms of contrastive laryngeal features are noted [18, p.75]. The register and contours of Vietnamese tones reported for Da Nang speakers are summarised in Table 1 (adapted from [18]).

tone name	Register/ Contour	Tone numbers	Phonation type
ngang	mid-level	33	modal
sắc	high-rising	35	modal
huyền	low-falling	21	modal
hỏi ~ ngã	mid-concave	214	+glottal ~ +creak
nặng	low	212	+creak

Table 1: Vietnamese tones and corresponding phonation types. Tone numbers indicate the relative pitch of the tones between targets (scale is 1-5).

2 METHODS

2.1 Data, speakers and elicitation

Two non-contemporaneous voice recordings were elicited from 10 female speakers all aged between 18 and 20. The recordings were made using a Zoom© H4 hand recorder sampled at 44.1 kHz. The elicitation was designed to emulate the missing information task described in [12], which is intended to elicit controlled but relatively natural, running speech. In the present study, the participants' task was to alternate asking questions to complete a motorcycle price list. For example, participant A, might ask '*Honda XR50R địa chỉ người bán là gì?*' ('*What's the address of the Honda XR50R seller?'),* and participant B respond '*4 Phạm Ngũ Lão, Đà Nẵng'* ('*4 Pham Ngu Lao street Da Nang')*. For this preliminary study only two disyllabic words were selected from the resulting speech corpus: /ɗie^{212} ci^{214}/ (*address*) and /ŋɯɤj^{21}.ɓan^{35}/ (*seller*), both constituents of the same noun phrase: *địa chỉ người bán, ('[the] seller's address')*. In total 10 tokens for each speaker (five per session) for each word were identified aurally and extracted for further processing.

2.2 Feature extraction and parameterisation

Raw F0 values were extracted in *Praat* [1] using the autocorrelation method [2]. This was done for a total of 10 tokens (five per session) for each disyllable and constituent syllables for each speaker. F0 then was sampled at 10% intervals and polynomial functions fitted to the trajectories. The feature vectors for the FVC were subsequently derived from the polynomial model parameters. Selection of the order of polynomial and duration base (equalised vs. unequalised) in each case (i.e. for each test word and constituent syllable) was determined empirically by generating LRs and choosing the feature set exhibiting the highest validity. The specific methods applied are detailed in the remaining subsections.

2.3 MVKD-LR Estimation and calibration

LRs were calculated using the Multivariate Kernel Density LR (MVKD-LR) procedure [1]. Testing involved partitioning speakers into test pairs for different-speaker (DS) comparisons and same-speaker (SS) comparisons then calculating a LR for each test pair. Given 10 speakers, a total 90 DS and 10 SS comparisons were possible. The small number of speakers meant that a separate training, test and background database was not used. Instead leave-one-out cross-validation (LOOCV) was applied. For the DS comparisons, this involved removing the speaker pair being tested from the background sample; for the SS comparisons only the speaker being tested was excluded.

Logistic-regression calibration [4], commonly applied to FVC LRs, was used to calibrate the LRs and transform the output scores to LRs. Calibration is necessary because the raw output LRs of the MVKD-LR formula are typically poorly calibrated [11]. Further, the output LRs are in fact scores quantifying the similarity of the differences between samples with respect to their typicality, rather than true LRs [10]. Again, given the small dataset a cross validation procedure was used to derive the calibration weights.

2.4 Evaluation of performance

The log-likelihood ratio cost function (Cllr) [4] was used to evaluate system validity. Cllr measures information loss in *bits*. It provides a gradient measure of system accuracy reflecting the overall strength of evidence yielded by the FVC system, rather than simply the proportion of SS and DS comparisons correctly classified; the latter typically qualified by the Equal Error Rate [9]. Cllr penalises systems more for high contrary-to-fact LRs (i.e. systems with poor validity). Optimum validity is achieved when Cllr = 0 and decreases as Cllr approaches and exceeds 1 [9].

The Bayesian credible interval (BCI), the Bayesian equivalent of confidence intervals, assesses reliability. The method used to calculate the credible interval for these experiments follows [9]. The BCI requires at least two LRs for each unique comparison. Since there are only two recordings it was not possible to calculate a BCI for the SS comparisons. Therefore, the BCI could only be estimated for the DS comparisons.

3 RESULTS

The results for the best performing feature combinations for each disyllable and their constituent syllables are given in Table 2. The Cllr value represents

the lowest achieved from iterative FVC testing of the coefficient parameters from different polynomial orders (1-7), as well as duration base combinations.

Syllable/disyllable	Cllr	BCI	features
/ŋɯɤj^{21}/	0.80	±1.68	$a_{1,2}$*
/ɓan^{35}/	0.55	±10.15	$a_{1,2}$*
/ŋɯɤj^{21}.ɓan^{35}/	0.40	±4.61	$a_{0,}a_{1,2,3,4,5,6}$*
/ɗie^{212}/	0.80	±1.37	$a_{0,,}a_{1}$*
/ci^{214}/	0.95	±0.87	$a_{1,2}$**
/ɗie^{212} ci^{214}/	0.80	±2.88	$a_{1,2,3,4,}$**

Table 2: Results for best performing F0 trajectory features. a_0 = intercept, a_n = polynomial coefficient term; * = unequalised duration, ** = equalised duration.

In the results, we focus on the relative performance of syllable vs. disyllabic parameterisation for each test word. For /ŋɯɤj^{21}.ɓan^{35}/, the lower Cllr value for the disyllabic parameterisation of the F0 trajectory in Table 2 indicates this system on the whole is delivering better strength of evidence than its constituent syllable based systems. A Cllr of 0.40 is achieved, representing a 0.15 and 0.25 bit improvement in validity over using the F0 trajectories from constituent syllables individually.

Notable too is a considerable narrowing of the BCI relative to the F0 trajectory over /ɓan^{35}/ (±4.61 vs. ±10.15), indicating an increase in system reliability in the disyllabic condition. Though wider than for /ŋɯɤj^{21}/ (BCI = ± 1.68), this is offset by the superior validity of the disyllabic system (Cllr = 0.40 vs. 0.80). The comparative performance of the syllable vs. disyllabic systems can also be appreciated visually in the tippet plots in Figures 1-4. For all figures, the cumulative proportion of trials is plotted on the y-axis against the LR ($\log_{10}$) on the x-axis.

Figure 1: Tippet plot /ŋɯɤj^{21}/

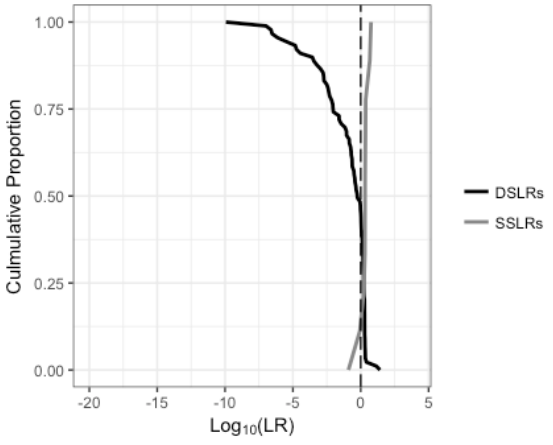


Figure 2: Tippet plot /ɓan^{35}/

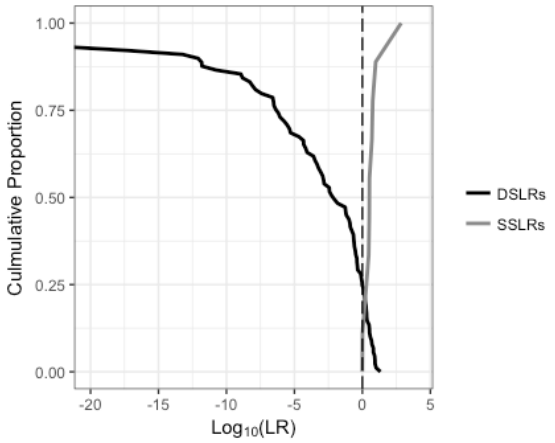


Figure 3: Tippet plot /ŋɯɤj^{21}.ɓan^{35}/

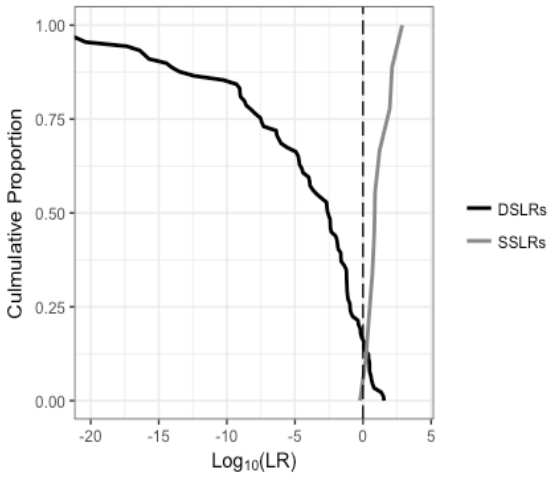
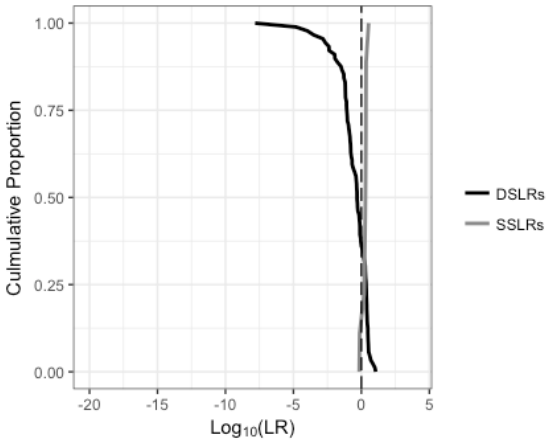


Figure 4: Tippet plot /ɗie^{212} ci^{214}/



The magnitude of the $\log_{10}$LR is proportional to the strength of evidence in support of the same-speaker (H_{ss}) and different speaker hypotheses. LR values less than 0 favor the DS hypothesis and those greater than 0 favor the SS hypothesis. In Figure 3, the solid black line (indicating the DSLRs) saturates towards zero far

slower than those in Figures 1 and 2, indicating the comparisons evinced comparatively greater strength evidence than the syllable based comparisons. The same is true to a slightly lesser degree for the SSLRs.

The results for the second test word /ɗie^{214}.ci^{214}/ lend somewhat less support for the superiority of disyllabic parameterisation in terms of FVC performance. Here, while the disyllabic system showed a 0.15 bit improvement over the syllable based F0 trajectory for /ci^{214}/, there is no improvement over /ɗie^{212}/ - both the latter and the disyllabic system evince a Cllr of 0.80. However, in the disyllable F_0 trajectory system (Figure 4) an improvement in the strength of evidence for the DSLRs was apparent. Indeed, the maximum DSLR obtained from /ci^{214}/ and /ɗie^{212}/ were $\log_{10}$LRs of ~-3 and ~-4 respectively, whereas in the disyllable system it reaches a maximum $\log_{10}$LR of nearly -8.

Nonetheless, the results are nowhere near as good as for first test word (Cllr = 0.40 vs. 0.8). This is perhaps not surprising given the F0 trajectories for both constituent syllables yielded poor validity. Indeed, /ci^{214}/ offered the worst validity with Cllr close to 1 (Cllr = 0.95). While comparatively better validity is found for /ɗie^{212}/, the Cllr of 0.80 is indicative of poor discrimination between same and different speaker pairs. It is also worth noting there are a performance differences associated with F0 trajectories for the individual constituent syllables: the low-falling tone (/ŋɯɤj^{21}/) and low tone (/ɗie^{212}/) both had a Cllr of 0.8, the high-rising tone (/ɓan^{35}/) 0.55 and mid-concave (/ci^{214}/) 0.95.

4 DISCUSSION & CONCLUSIONS

These preliminary results provide evidence that basing feature vectors for the MVKD-LR on disyllabic tonal F0 trajectories does yield better performance than a syllable based parameterisation. While this was clearly the case for the disyllabic word /ŋɯɤj^{21}.ɓan^{35}/, it was not so clear for /ɗie^{212}.ci^{214}/. For the latter, no improvement (in terms of validity) was seen in the disyllabic condition over the /ɗie^{212}/ F0 trajectory features – though it did improve substantially on the tonal F0 features obtained from /ci^{214}/ and an improvement in the magnitude of DSLRs was apparent. These results are similar to those observed for Cantonese, where improvements using tonal F0 trajectories from disyllabic words over monosyllables have been demonstrated [16].

The monosyllabic based F0 trajectories in the Vietnamese experiments present here showed quite poor FVC performance in terms of validity. Cllr values were typically close to 1, indicating relatively poor discrimination. Again, this is a similar situation for Cantonese FVC using tonal F0 based on monosyllables [8]. Notable too in the present study are the reasonably large performance differences between tonal F0 trajectories for different constituent syllables.

This suggests that different tones may be better for discrimination than others. This has been demonstrated in FVC studies using tonal F0 in Thai ([13, 17]) and likely applies to Vietnamese tones as well. For the present data the tones do not occur in the same segmental environments so it is not possible to say which particular tone may be optimal for discrimination, since the role of intrinsic consonantal and vocalic influences on F0 cannot be separated. This is one limitation of the study.

Another is the small sample size used (only 10 speakers). The size of the reference population is important to ensuring accurate LRs, since it determines how well the probability densities of within- and between-speaker variation are modelled. The limited sample size could therefore be over or underestimating system validity. This is something that can only be addressed with a larger study involving more speakers.

This study tested the FVC performance of acoustic-phonetic based features for Vietnamese for the first time. It demonstrated some preliminary evidence that monosyllabic F0 trajectories yield relatively poor performance, though this can be improved by combining the trajectories of two tones i.e. through a disyllabic parameterisation. LR-based FVC is in practice based on more than a single acoustic feature vector. Therefore, fusing the output of the F0 trajectories with additional features from the test words, such as the F-pattern, would likely also improve performance. It is apparent though that better performance would be obtainable if the disyllabic F0 trajectory were used in this fusion. Future work is required to determine the discriminatory power of individual tonal F0 trajectories to inform the selection of disyllabic words for FVC in Vietnamese.

ACKNOWLEDGEMENTS

This study is part of the first author's honours thesis at the Australian National University. We especially thank The University of Foreign Languages (Da Nang University) for facilitating data collection in Viet Nam. We also thank the reviewers for their useful comments.

REFERENCES

[1] Aitken, C.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(4), 109-122. doi: 10.1046/j.00H35-9254.2003.05271.x

[2] Boersma, P. & Weenink, D. (2018). Praat: doing phonetics by computer [Computer software]. Version 6.0.37

[3] Boersma, P. (1993). Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. *Proceedings of the Institute of Phonetic Sciences* 17, 97–110. University of Amsterdam.

[4] Brümmer, N., & Du Preez, J. (2006). Application-independent evaluation of speaker detection. *Computer Speech & Language*, 20(2), 230-275.

[5] Drygajlo, A. Jessen, M., Gfroerer, S., Wagner, I., Vermeulen, J. & Niemi, T. (2015). Methodological Guidelines for Best Practice in Forensic Semiautomatic and Automatic Speaker Recognition. European Network of Forensic Science Institutes, Wiesbaden, Germany.

[6] Kinoshita, Y., Ishihara, S. & Rose, P. (2008). Beyond the Long-term Mean: Exploring the Potential of F0 Distribution Parameters in Forensic Speaker Recognition. In Brummer, N. (ed.). *Proceedings of the 2008 Odyssey Speaker and Language Recognition Conference*, 8.

[7] Kinoshita, Y. & Ishihara, S. (2010). F0 can tell us more: speaker classification using the long term distribution. *Proceedings of Thirteenth Australasian International Conference on Speech Science and Technology 2010*, 50-53.

[8] Li, J.J. & Rose, P. (2012). Likelihood Ratio-based Forensic Voice Comparison with F-pattern and Tonal F0 from the Cantonese /eu/ Diphthong. In Cox, F., Demuth, K. Lin, S, Miles, K., Palethrope, S., Shaw, J. & I. Yuen (Eds.). *Proceedings of the 14th Australasian International Conference on Speech Science and Technology,* 201-204.

[9] Morrison, G.S. (2011). Measuring the validity and reliability of forensic likelihood-ratio systems. *Science & Justice*, 51(3), 91–98. doi: 10.1016/j.scijus.2011.03.002

[10] 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, doi: 10.1080/00450618.2012.733025

[11] Morrison, G.S. & Kinoshita, Y. (2008). Automatic-Type Calibration of Traditionally Derived Likelihood Ratios: Forensic Analysis of Australian English /o/ Formant Trajectories. *Proceedings of the 9th Annual Conference of International Speech Communication Association*, 1501-1504.

[12] 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.

[13] Pingjai, S. (2011). Forensic Voice Comparison in Thai: A Likelihood Ratio-Based Approach Using Tonal Acoustics. Unpublished MA thesis. The Australian National University, Canberra.

[14] Rose, P. (2002). Forensic Speaker Identification. London, UK: Taylor & Francis.

[15] Rose, P. (2005). Forensic speaker recognition the beginning of the twenty first century. *Australian Journal of Forensic Sciences*, 37, 49-72.

[16] Rose, P. & Wang, X. (2016). Cantonese forensic voice comparison with higher-level features: likelihood ratio-based validation using F-pattern and tonal F0 trajectories over a disyllabic hexaphone. *Proceedings of the Odyssey 2016 Speaker and Language Recognition Workshop*, 326-333.

[17] Thaitechawat, S., & Foulkes, P. (2011). Discrimination of speakers using tone and formant dynamics in Thai. In *Proceedings of the 17th International Congress of Phonetic Sciences,*1978-1981.

[18] Vũ Thanh Phương (1981). The acoustic and perceptual nature of tone in Vietnamese. Unpublished Ph.D. thesis. Australian National University, Canberra.

[19] Wang, C. & Rose, P. (2012). Likelihood Ratio-based Forensic Voice Comparison with Cantonese /i/ F-pattern and Tonal F0. In F. Cox, K. Demuth, S. Lin, K. Miles, S. Palethrope, J. Shaw & I. Yuen (eds.), *Proceedings of the 14th Australasian International Conference on Speech Science and Technology*, 209-212.