coustics'08
Paris
June 29-July 4, 2008
www.acoustics08-paris.org

euronoise

Vietnamese monophthong vowel production by native speakers and American adult learners

Matthew Winn[a], Allison Blodgett[b], Jessica Bauman[b], Anita Bowles[b], Lykara Charters[b], Anton Rytting[b] and Jessica Shamoo[b]

[a]University of Maryland College Park, Department of Hearing & Speech Sciences, 0100 Lefrak Hall, College Park, MD 20742, USA
[b]University of Maryland College Park, Center for Advanced Study of Language, 7005 52nd Ave, College Park, MD 20742, USA
mwinn@hesp.umd.edu

This study provides new data regarding the vowel space and duration contrasts of Vietnamese monophthongs. These data address conflicting descriptions found in the literature regarding orthographic *ư*, *ơ*, *â*, and *ă*. A modified vowel mapping analysis indicates that native speakers of Vietnamese produce unrounded *ư* and *ơ* as central, not back. The data also show quality differences for one pair of vowels in short-long opposition (*ơ-â*), but not the other (*ă-a*). Duration measurements of vowels and nasals in coda position are consistent with prior claims of temporal compensation in Vietnamese. In contrast, measurements of non-native vowel production reveal inaccurate and inconsistent vowel quality for the central vowels, and a complete absence of duration distinctions between long and short vowels.

1 Introduction

The purpose of this investigation is to clarify the nature of the Vietnamese vowel inventory. Particular issues of interest include the rounding and/or advancement distinction between orthographic *ư-u* and *ơ-ô* pairs, as well as the vowel quality and duration of short vowels *â* and *ă* in relation to their long counterparts, *ơ* and *a*.

A second purpose of this investigation is to evaluate the production of Vietnamese vowels by adult learners whose native language is English. Special attention is paid to orthographic *ư*, *ơ*, *â*, and *ă*, in terms of articulation by adult learners as compared to the native speakers.

2 Problematic descriptions of Vietnamese vowels

The Vietnamese vowel system contains several monophthongs that have been described consistently and that have a transparent orthography: *i*[i], *u*[u], *ô*[o], *o*[ɔ], *ê*[e], *e*[ɛ], *a*[a]. The other vowels, however, including orthographic *ư*, *ơ*, *â*, and *ă*, have been described using different, conflicting sets of features. For example, *ư* has been described as high back unrounded [ɯ] [1], high central unrounded [ɨ] [2, 3, 4], and high central [ɯ] [5]. Similarly, *ơ* has been described as back unrounded [ɤ] or [Λ] [1], as [ə] [2, 4], and as central [ɤ] [5]. These distinctions are obscured because rounding and tongue advancement are virtually identical in terms of acoustic quality [6, 7].

Some descriptions of Vietnamese phonology describe *ơ* and *â* as two vowels in long-short opposition, and similarly link *a* with *ă* [8, 5]. An alternative view asserts that all four vowels are distinct in quality [2, 9]. The current study explores the quality dimension of these pairs along with the nature of their duration oppositions.

3 Predicted problems for American adult learners

3.1 Non-front unrounded vowels

We predicted that *ư* and *ơ* would be difficult for native English speakers, because these two vowels have no clear equivalent in English. Literature on non-native speech production suggests that learners will approximate, but not accurately reach the acoustic target for these sounds, thus arriving at a "compromised" phone [10]. In English, lip rounding is phonologically redundant with backing, so non-native production of *ư* was expected to drift within the vowel space towards *u*, and *ơ* likewise towards *ô*.

3.2 Vowel duration contrasts

We also predicted that the contrastive short and long vowel durations would be an obstacle for adult learners because English does not exhibit phonemic quantity differences independent of consonant voicing environments. This means that adult learners should have difficulty producing long (*ơ* and *a*) and short (*â* and *ă*) vowel categories.

4 Methods

4.1 Participants

Native speaker participants included 3 Northern dialect speakers (1 female, 2 male) and 1 Southern dialect speaker (female). All were originally from Vietnam and had been living in an English-speaking country for 6 to 26 years. They ranged in age from 42 to 64, and all had experience teaching Vietnamese as a foreign language to adults.

Non-native speaker participants included 3 Northern dialect learners (all male) and 3 Southern dialect learners (1 female, 2 male). They ranged in age from 30 to 50. All had been studying Vietnamese intensively (i.e., at least 5 hours a day), but for varying lengths of time. Their weeks of training ranged from 14 to 32. All 10 participants resided in the Washington, DC, area at the time of recording.

4.2 Stimuli

Targets comprised 102 real words and used 8 monophthongs: *i, ư, u, ơ, ô, a, â, ă*. Six vowels appeared with all possible tones for each of three syllable types: open (e.g., *ba, bà, bạ, bá, bả, bã*), stop-final (e.g., *bạt, bát*), and nasal-final (e.g., *bang, bàng, bạnh, báng, bảng, vãn*). Consistent with Vietnamese phonology, two vowels (*â* and *ă*) appeared in stop-final and nasal-final syllables only.

To the extent possible, we matched targets for initial and final segments within syllable type and within vowel. We attempted to maintain consistent consonant place and manner, but, when necessary, sacrificed one or both in the interest of ensuring that all target stimuli were real words.

Speakers were recorded in a sound-dampened room using Sound Forge 7.0 (22 kHz, 16 bit, mono), a Yamaha 01V96 digital mixing console with no effects settings, and a Neumann TLM 103 microphone.

4.3 Procedure

Participants produced 3-word sentences in response to individual target words that appeared on a computer screen in red, blue, black, or purple. For example, if the target word *bang* appeared in blue, the speaker said *Từ bang xanh* ("the word *bang* is blue"). Participants had access to the written color names as they completed 8 practice trials and then two lists of words. Each list contained all 102 targets, which were pseudo-randomized such that the vowel, tone, and color of the word always changed from one trial to the next. Four additional targets occurred on each list. Three were non-adjacent repetitions of existing targets but in a narrow contrastive context (i.e., in the same color as the immediately preceding word). This added one token each of *i, ư, u, ơ, ô,* and *a*. The fourth addition (*ma*) occurred in list-final position and was never included in analyses. Targets that were paired with *xanh* and *tím* (purple) on List 1 were paired with *đen* (black) and *đỏ* (red), respectively, on List 2, and vice versa. Participants thus produced two repetitions of each target word, but novel utterances each time. In this self-paced task, participants could repeat any utterance before advancing to the next word. When speakers did repeat, we analyzed only the final repetition. Each speaker ultimately produced 12 tokens of each vowel (*i, ư, u, ơ, ô, a, â,* and *ă*) in nasal-final syllables (which were not used in formant analyses); an additional 4 tokens of each vowel in stop-final syllables; and 13 tokens of each of the long vowels (*i, ư, u, ơ, ô,* and *a*) in open syllables.

5 Analysis

Using Praat [11], we marked relevant onsets and offsets based on audio and visual inspection of each waveform and spectrogram. A vowel quality region was bounded by the onset and offset of well-defined formant structure. A vowel duration region had less stringent criteria; even if spectral structure was not conducive to formant analysis, vowel duration endpoints fell at the end of vowel production. For stop-final syllables, the vowel offset coincided with the acoustic signature of the vocal tract closure (quick transition and transient glottalization). For nasal-final syllables, vowel offset was judged at the cessation of robust high-frequency resonance accompanied by a shift in formant structure. As previously mentioned, formant analyses excluded nasal-final targets because of complications introduced with potential nasalization, particularly from non-native speakers. Formant analyses of the short vowels *â* and *ă* came from stop-final syllables only. The short duration of stop-final syllables induced hypoarticulation for peripheral vowels, essentially creating bimodal distributions for single vowel categories which were not germane to the current investigation.

Formant values at the midpoint of the vowel quality region were automatically extracted by Praat scripts. When voice quality of the tone interrupted the midpoint of the vowel, the formant extraction was done manually from a steady-state portion of the vowel. In these and all other cases where manual formant analysis was performed, judgments were made in accordance with Hillenbrand [12].

To approximate the input as it would be received by the auditory system, values were converted to Bark [13]. Additionally, each speaker's average Bark values were calculated for each vowel along the F1 and F2 planes. Each speaker initially provided 13 tokens each of *i, ư, u, ơ, ô,* and *a* and 4 tokens each of *â* and *ă* for formant analyses. We excluded any token greater than 2 standard deviations from the central target. One native speaker token (out of 344) and 13 non-native speaker tokens (out of 516) were excluded, with no more than 4 tokens attributed to a single speaker.

Each speaker's vowel space was normalized using a method inspired by Watt and Fabricius [14] and a revised form of Gerstman's [15] procedure described by Disner [16]. Maximum and minimum values for both F1 and F2 (from all vowel tokens) were identified for individual speakers and used as endpoints within which each vowel token had a position in that speaker's vowel space. Thus, for example, the high-front vowel [i] is likely to exhibit an F1 value close to the minimum ("0") end of the F1 scale, and an F2 value close to the maximum ("1") end of the F2 scale. Formant values are converted using the calculation in Eq.(1), where *Fx* is the formant value in Hz.

$$\text{F1 (norm)} = \frac{(Fx - MinF1)}{(Max\ F1 - Min\ F1)} \qquad (1)$$

Eq.(1) addresses the question, "within this speaker's vowel space, where does this particular vowel fall?" We used these values to describe each speaker's vowel space and also to compare articulation across speakers. Unlike the procedures described in Disner [16], the values were not rescaled along a synthetic Hertz continuum.

6 Results

6.1 Native speaker vowel quality

Figure 1 shows that *ư* and *ơ* are located in the middle of the F2 plane, suggesting that they are central (unrounded) vowels. Orthographic *ư* appears to be high central unrounded [ɨ], and *ơ* would be best represented by mid-central [ə]. The quality of *ă* is equal to that of *a*, suggesting that quality is not the primary cue for this distinction. The chart shows a clear height difference between *ơ* and *â*, the latter being the more open of the pair, which would be signified by [ɐ]. Thus, while duration appears to play a part in distinguishing this pair, it is accompanied by a quality difference as well.

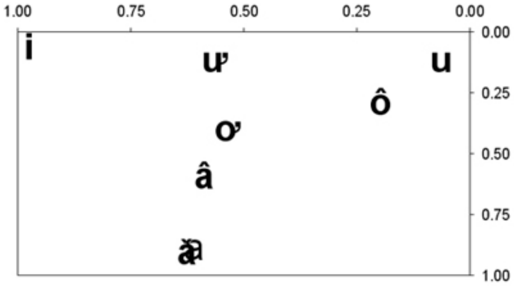

Fig.1 Vowel chart produced by data from all four native speakers recorded in this experiment, normalized according to the method described above.

6.2 Non-native speaker vowel quality

Figures 2 through 7 illustrate the six non-native speakers' average vowel articulations (solid letters) overlaid on the normalized native speaker vowel space (outlined letters, from Figure 1). Non-native speakers rarely achieve native-like separation of *ư* from *u* (the possible exceptions being Learners 02 and 07, but see Section 6.3). They also do not produce *u* as far back as the native speakers, but demonstrate clear overlap in the production of *i*. Separation of *ơ* and *ô* seems to be easier for this group of speakers, but vowel height is variable. In general, most errors were misarticulations of tongue advancement or rounding, as indicated by deviant F2 values.

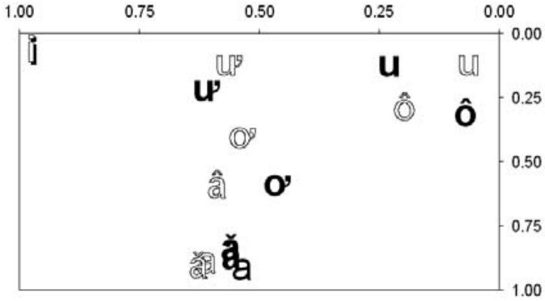

Fig.2 Normalized vowel chart produced by Learner 02

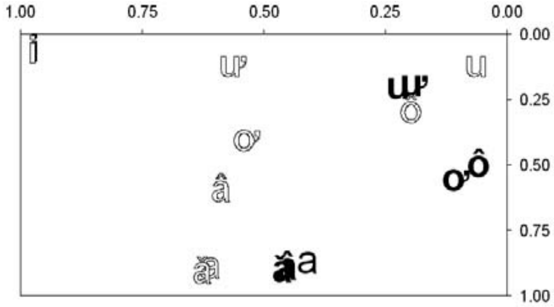

Fig.3 Normalized vowel chart produced by Learner 03

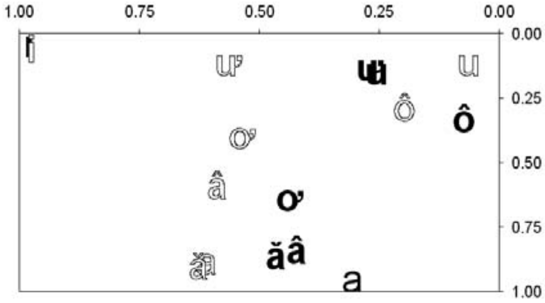

Fig.4 Normalized vowel chart produced by Learner 04

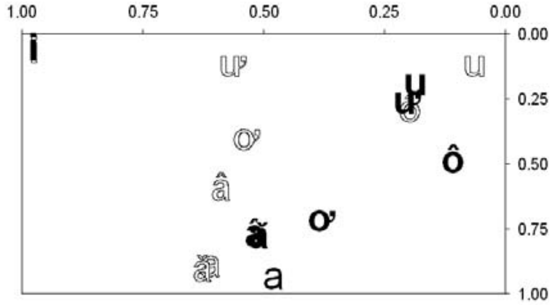

Fig.5 Normalized vowel chart produced by Learner 05

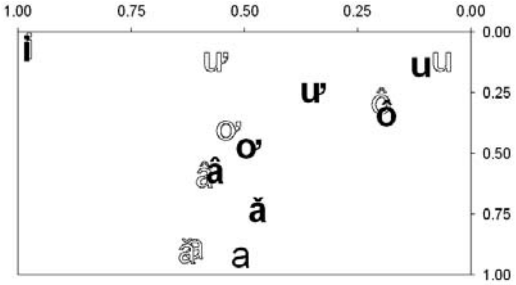

Fig.6 Normalized vowel chart produced by Learner 06

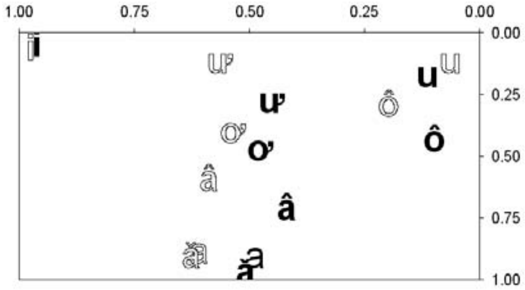

Fig.7 Normalized vowel chart produced by Learner 07

6.3 Vowel quality analysis

A variation of the d' measure was used as a metric of vowel category separation. The distance between means along one continuum (F2 for the *ư*/*u* and *ơ*/*ô* comparisons and F1 for the *â*/*ă* comparison) is measured in root mean square standard deviations calculated from each distribution. Higher d' values indicate a greater separation of categories.

As shown in Table 1, no learner showed native-like separation of the *ư* and *u* vowels. As shown in Tables 2 and 3, only Learner 07 approximated native-like category separation for the *ơ* and *ô* and the *â* and *ă* vowels.

	F2 Norm Bark (StDev)	F2 Norm Bark (StDev)	d'
	Central Vowels	*Back Vowels*	
NS	ư: 0.548 (0.039)	u: 0.065 (0.040)	12.21
02	ư: 0.616 (0.101)	u: 0.235 (0.104)	3.71
03	ư: 0.196 (0.113)	u: 0.230 (0.063)	-0.39
04	ư: 0.273 (0.068)	u: 0.260 (0.119)	0.13
05	ư: 0.207 (0.144)	u: 0.191 (0.064)	0.18
06	ư: 0.352 (0.205)	u: 0.117 (0.058)	1.78
07	ư: 0.455 (0.161)	u: 0.118 (0.060)	3.05

Table 1
d' values showing F2 category separation of *ư* and *u* for native speakers (NS) and adult learners (02 to 07)

	F2 Norm Bark (StDev)	F2 Norm Bark (StDev)	d’
	Central Vowels	*Back Vowels*	
NS	ơ : 0.530 (0.029)	ô: 0.184 (0.027)	12.46
02	ơ : 0.470 (0.045)	ô: 0.067 (0.039)	9.62
03	ơ : 0.113 (0.058)	ô: 0.058 (0.039)	1.13
04	ơ : 0.444 (0.067)	ô: 0.070 (0.044)	6.7
05	ơ : 0.389 (0.114)	ô: 0.105 (0.039)	3.7
06	ơ : 0.499 (0.068)	ô: 0.184 (0.073)	4.48
07	ơ : 0.486 (0.030)	ô: 0.096 (0.024)	14.58

Table 2
d’ values showing F2 category separation of *ơ* and *ô* for native speakers (NS) and adult learners (02 to 07)

	F1 Norm Bark (StDev)	F1 Norm Bark (StDev)	d’
	Mid-open Vowels	*Open Vowels*	
NS	â: 0.617 (0.030)	ă: 0.908 (0.037)	8.72
02	â: 0.616 (0.101)	ă: 0.848 (0.109)	-0.28
03	â: 0.657 (0.048)	ă: 0.668 (0.052)	0.23
04	â: 0.848 (0.121)	ă: 0.877 (0.086)	0.28
05	â: 0.771 (0.043)	ă: 0.789 (0.098)	0.26
06	â: 0.580 (0.121)	ă: 0.747 (0.087)	1.60
07	â: 0.717 (0.032)	ă: 0.980 (0.018)	10.48

Table 3
d’ values showing F1 category separation of *â* and *ă* for native speakers (NS) and adult learners (02 to 07)

7 Vowel duration analysis

7.1 Native speakers

Vowel duration analysis was restricted to *ơ*/*â* and *a*/*ă*, the two pairs previously described as being in long/short opposition. For long vowels, we only analyzed nasal-final syllables, or stop-final syllables carrying the sắc or nạng tones, since short vowels are restricted to these conditions.

Each short/long vowel comparison was significant within the native speaker data for both stop-final and nasal-final syllables. Paired samples t-tests showed that, as a group, native speakers produced *ơ* and *a* vowels that were significantly longer than *â* and *ă* in stop-final syllables [$t(31) = -13.97$, $p<0.001$] and in nasal-final syllables [$t(95) = -20.40$, $p<0.001$] (matched for phonetic environment). The average duration ratio of long:short was 1.72:1 for stop-final syllables and 1.80:1 for nasal-final syllables.

In addition, paired-samples t-tests revealed that the native speaker group produced word-final nasal consonant segments that were significantly longer when preceded by short vowels [$t(95)=7.39$, p<0.001] than by long vowels. This temporal compensation helps to neutralize coda length across vowel types, perhaps to preserve syllable timing. These data support Pham’s [5] observation that nasals following short vowels recover some duration compromised by the shortness of the *â* and *ă* vowels.

	Avg long vowel dur (ms)	Avg short vowel dur (ms)	Ratio long:short
NS 01	186	94	1.98
NS 08	161	108	1.49
NS 09	151	78	1.94
NS 10	181	123	1.47
02	82	77	1.06
03	155	147	1.06
04	99	102	0.98
05	120	123	0.98
06	108	105	1.03
07	117	118	0.99

Table 4
Duration measurements for vowels in stop-final syllables for native speakers (NS) and adult learners (02 to 07)

	Avg long vowel dur (ms)	Avg short vowel dur (ms)	Ratio long:short
NS 01	221	113	1.95
NS 08	220	137	1.61
NS 09	183	95	1.93
NS 10	200	115	1.74
02	135	147	0.92
03	195	202	0.97
04	152	164	0.93
05	196	191	1.02
06	155	143	1.08
07	199	194	1.03

Table 5
Duration measurements for vowels in nasal-final syllables for native speakers (NS) and adult learners (02 to 07)

7.2 Non-native speakers

None of the non-native speakers exhibited any significant vowel duration differences for any syllable type; the average ratio was 1:1 and the largest long:short ratio shown by any non-native speaker was 1.08:1. Temporal compensation was not analyzed within the non-native speaker group because they exhibited no vowel duration differences which might trigger temporal compensation.

8 Discussion

Using a modified vowel normalization procedure, we were able to provide new data on the representation of orthographic *ư* and *ơ* in Vietnamese. Our data suggest that these vowels would be best represented as the central vowels [ɨ] and [ɘ], respectively. Additionally, our data contradict claims that *ơ* and *â* differ only in duration and, instead, support the existence of quality differences as well.

As predicted, adult learners struggled to produce the opposition between Vietnamese central and back vowels. They showed insufficient advancement separation of these vowels as compared to native speakers. We also predicted that adult learners would struggle to produce a duration distinction between Vietnamese long and short vowels. In fact, they failed to produce this distinction at all. This finding suggests that vowel duration contrasts are especially difficult for native speakers of English to learn.

Acknowledgments

We gratefully thank our speakers for participating and Bill Idsardi for technical assistance.

References

[1] M. Lindau, "Vowel features", *Language* 54, 541-563 (1978)

[2] L.C. Thompson, *A Vietnamese grammar*. Seattle: University of Washington Press (1965)

[3] M. Brunelle, *Coarticulation effects in Northern Vietnamese tones*. Unpublished manuscript, Cornell University (2003)

[4] M. Ferlus, "On the formation of the vowel system in Vietnamese", *Cahiers de Linguistique Asie Orientale,* 26, 37-51 (1997)

[5] A. Pham, *Vietnamese tone: A new analysis*. Outstanding Dissertations in Linguistics. New York: Routledge (2003)

[6] B.E.F. Lindblom, J.E.F. Sundberg, "Acoustical consequences of lip, tongue, jaw, and larynx movement", *J. Acoust. So. Am.* 50, 1166-1179 (1971)

[7] L. Lisker, M. Rossi, "Auditory and visual cueing of the [+/- rounded] feature of vowels", *Language and Speech* 35, 391-417 (1992)

[8] H. Nguyen, M. Macken, "Factors affecting the production of Vietnamese tones", *Studies in Second Language Acquisition* 30, 49-77 (2008)

[9] L.C. Thompson, "Saigon phonemics", *Language* 35, 454-476 (1959)

[10] J.E. Flege, J. Hillenbrand, "Limits on pronunciation accuracy in adult foreign language speech production", *J. Acoust. Soc.Am.* 76, 708-721 (1984)

[11] P. Boersma, D. Weenink, Praat: Doing phonetics by computer (Version 5.0.14) [Computer program]. Retrieved February 20, 2007, from http://www.praat.org/ (2007)

[12] J. Hillenbrand, L.A. Getty, M.J. Clark, K. Wheeler, "Acoustic characteristics of American English vowels", *J. Acoust. Soc. Am.*, 97, 3099-3111 (1995)

[13] E. Zwicker, E. Terhardt, "Analytical expressions for critical-band rate and critical bandwidth as a function of frequency", *J. Acoust. Soc. Am.* 68, 1523-1525 (1980)

[14] D. Watt, A. Fabricius, "Evaluation of a technique for improving the mapping of multiple speakers' vowel spaces in the F1 ~ F2 plane", In D. Nelson, *Leeds Working Papers in Linguistics and Phonetics,* 9:159-73 (2002)

[15] L.J. Gerstman, "Classification of self-normalized vowels", *IEEE Transactions on Audio and Electroacoustics AU-16*, 78-80 (1968)

[16] S.F. Disner, "Evaluation of vowel normalization procedures", *J. Acoust. Soc. Am.* 67, 253-261 (1980)