Dialect experience in Vietnamese tone perception[a)]

James Kirby[b)]
Department of Linguistics, Phonology Laboratory, University of Chicago, 1010 East 59th Street, Chicago Illinois 60637

(Received 1 October 2009; revised 16 December 2009; accepted 26 January 2010)

This study investigated the perceptual dimensions of tone in Vietnamese and the effect of dialect experience on listener's prelinguistic perception of tone. While Northern Vietnamese tones are cued by a combination of pitch and voice quality, Southern Vietnamese tones are purely pitch based. 30 listeners from two Vietnamese dialects (10 Northern, 20 Southern) participated in a speeded AX discrimination task using northern stimuli. The resulting reaction times were used to compute an INDSCAL multidimensional scaling solution and were submitted to hierarchical clustering analysis. While the analysis revealed a similar three-dimensional perceptual space structure for both listener groups, corresponding roughly to f_0 offset, voice quality, and contour type, the relative salience of these dimensions varied by dialect: Southern listeners were more likely to confuse tones produced with nonmodal voice quality, whereas Northern listeners found tones with similar pitch excursions to be more confusable. The results of hierarchical clustering of the stimuli further support an analysis where low-level perceptual similarity is influenced by primary dialect experience.
 [DOI: 10.1121/1.3327793]

PACS number(s): 43.71.Hw, 43.71.Bp [MAH] Pages: 3749–3757

I. INTRODUCTION

Adult listeners, in marked contrast to very young infants, often have difficulty identifying or discriminating phonetic contrasts which are not used distinctively in their native language (Lisker and Abramson, 1970; Werker and Tees, 1984; Strange, 1995). Even years of extensive training or exposure may not be sufficient to acquire native or near-native sensitivity to non-native contrasts. For example, native Japanese speakers, for whom the American English /r/-/l/ distinction is notoriously difficult to perceive and produce, may continue to have difficulty discriminating these sounds even after years of exposure (MacKain *et al.*, 1980). This difficulty extends to suprasegmental contrasts as well. A large body of research indicates that speakers of a lexical tone language such as Thai or Mandarin Chinese are better at discriminating tones from an unfamiliar tone language than are speakers of languages such as English, which do not exploit lexical tone (Gandour and Harshman, 1978; Gandour, 1983; Lee and Nusbaum, 1993). For instance, Wayland and Guion (2004) presented data from experiments showing that both native Mandarin and Taiwanese speakers were better at discriminating a Thai tonal contrast than were native English speakers. They argued that this is due to the Chinese listeners' native experience with using f_0 as a cue to lexical contrasts, which allowed them to extend this ability to a new language by mapping non-native tones onto native tone categories.

Like languages, dialects may also differ in the number and kind of phonetic contrasts they distinguish. In a study comparing standard and Swiss French, Miller and Grosjean (1997) argued that the ability to use an acoustic dimension to discriminate between existing (native dialect) phonological categories depends on native experience with contrastive use of that dimension. In Swiss French, unlike standard French, duration plays a decisive role in maintaining vowel distinctions: duration covaries with spectral information in the /o/-/ɔ/ contrast (as also occurs in standard French) and in the /ɑ/-/a/ contrast (largely lost in standard French), but is used to maintain a distinction between forms that have become homophones in standard French, such as /iː/-/i/ (*vie*, *vit*) and /aː/-/a/ (*voie*, *voix*). Despite the fact that temporal and spectral information covary for at least one of the contrasts for standard French speakers, those speakers did not appear to make use of temporal information in vowel identification (see also Gottfried and Beddor, 1988). This led the authors to conclude that there are constraints on the use of potentially relevant phonetic information in perception that derive, at least in part, from L1 experience. In other words, when presented with an acoustic cue with which they had at least some previous experience but which was not a primary bearer of contrast, listeners did not appear to attend to that cue.

Just as language-specific phonetic structure may encode both segmental and suprasegmental distinctions, so too may dialect-specific phonetics. While words in Tokyo Japanese manifest a pitch-accent pattern salient to listeners (Minematsu and Hirose, 1995; Cutler and Otake, 1999), other varieties of Japanese are "accentless" (Otake and Cutler, 1999). The cues to tonal contrasts in many Southeast Asian languages such as Khmu' (Suwilai, 2001; Svantesson and House, 2006; Abramson *et al.*, 2007) and Vietnamese (Vũ, 1981, 1982) have also been shown to vary by dialect. For example, the tonal inventory of Northern Vietnamese (NVN) contains six tones, distinguished by both f_0 and voice quality cues, whereas the Southern Vietnamese (SVN) dialect contains just five tones, distinguished only (or at least chiefly) by

[a)]Portions of this work were previously presented in May 2009 at the 157th Meeting of the Acoustical Society of America 2nd ASA Special Workshop on Speech, Portland, OR, and the 19th Annual Meeting of the Southeast Asian Linguistics Society, Ho Chi Minh City, Vietnam.
[b)]Electronic mail: jkirby@uchicago.edu

f_0 (Vũ, 1981, 1982; Nguyễn and Edmonson, 1998). Based on the results of an identification experiment, Brunelle (2009) argued that Northern listeners are more sensitive to voice quality cues than are Southern listeners, who have less native experience with these cues. Similarly, words which are distinguished by f_0 differences in the Northern and Western dialects of Khmu' are distinguished by voicing contrasts in the Eastern dialect. Svantesson and House (2006) showed that Eastern Khmu' speakers neither produce nor perceive f_0 differences at the lexical level.

These differences do not seem to greatly impair mutually intelligibility between dialects, however. Indeed, it has recently been argued that differences in production between dialects are not always indicative of differences in perception (Sumner and Samuel, 2009). This raises the question of the extent to which primary dialect experience plays a role in low-level phonetic perception. That is, the production differences between Northern and Southern Vietnamese may be negligible from a perceptual standpoint: Southern listeners may simply be ignoring the voice quality cues present in Northern speech, focusing instead on the pitch-based cues shared by both dialects. If this is the case, we might expect Southern listeners to process Northern tones solely on the basis of their f_0 profiles, and patterns of confusion would be based purely on f_0 information. Indeed, while the study of Brunelle (2009) shows that Southern listeners frequently ignore voice quality information when making identification judgments, it also suggests that they are at least somewhat sensitive to voice quality cues. The question then arises of whether Southern listeners' prelinguistic processing of these cues is the same as that of Northern listeners.

The identification task employed in Brunelle, 2009 is an example of an "off-line" paradigm, where participants are allowed to reflect on the stimulus before responding. As a result, listeners had adequate time to access lexical and/or metalinguistic information, which may obscure potential differences in prelinguistic processing. In order to obtain responses driven by physical (acoustic, prelinguistic) properties of the stimulus, an "on-line" task, in which participants respond immediately and/or subconsciously, may be more appropriate. The present study used a speeded AX discrimination paradigm designed to encourage "phonetic listening," thus enabling an assessment of sensitivity to lower-level acoustic-phonetic information (Werker and Tees, 1984) and the extent to which differences in this sensitivity (if any) may be attributed to dialect background.

II. BACKGROUND

A. Acoustic properties of Vietnamese tones

The Vietnamese language consists of multiple mutually intelligible dialects which nevertheless exhibit considerable lexical, segmental, and tonal variation. The Northern dialect, typified by educated Hanoi speech, is the official standard, and due to media exposure, schooling, etc., most Vietnamese speakers are at least somewhat familiar with this form of the language. NVN distinguishes six tones in open syllables, combining pitch contours with a distinction between modal and breathy and/or laryngealized voice qualities (Andreev and Gordina, 1957; Thompson, 1965; Đoàn, 1977; Vũ, 1981, 1982; Hoàng, 1986, 1989; Nguyễn and Edmonson, 1998; Brunelle, 2003; Phạm, 2003; Michaud, 2004; Brunelle, 2009). Like many Southeast Asian languages, Vietnamese restricts the inventory of tones which may appear in closed syllables. These will not be considered here; see Michaud, 2004 for details.

TABLE I. Tones in Northern and Southern Vietnamese. Chao tone numbers are modified from Vũ (1982).

Tone	Northern		Southern	
ngang (A1)	ba	44	ba	44
huyền (A2)	ba	21	ba	21
sắc (B1)	ba	24	ba	35
nặng (B2)	ba̱	22g	ba	212
hỏi (C1)	ba̱	3g2	ba	214
ngã (C2)	ba̱	3g5	ba	214

SVN is a term often used to refer to the colloquial speech of Ho Chi Minh City. Less research has been devoted to the study of SVN tones (but see Thompson, 1959; Trần, 1967; Gsell, 1980; Vũ, 1982; Hoàng, 1986), in part due to the fact that Ho Chi Minh City has seen a large influx of immigrants from other dialect regions since the 1940s, making it difficult to find "pure" SVN speakers. However, the acoustic data available confirm that SVN distinguishes just five tones. SVN speakers may mimic NVN voice quality distinctions in certain situations, but colloquial SVN speech does not make contrastive use of distinctive voice quality (Vũ, 1982). While the level *ngang* (A1), falling *huyền* (A2), and rising *sắc* (B1) tones are nearly identical to their NVN counterparts, the realization of the remaining tones is rather different, as seen in Table I and Fig. 1. The contour tones *hỏi* (C1) and *ngã* (C2) have merged in SVN into a single tone with a falling-rising pitch contour and little to no trace of laryngealization (these tones are arbitrarily indicated as C1 in Fig. 1). The SVN *nặng* (B2) tone has a similar type of pitch excursion, although its final rise is less extreme; in contrast to the NVN low *nặng* tone, which is short and marked by strong final laryngealization, the SVN *nặng* lacks

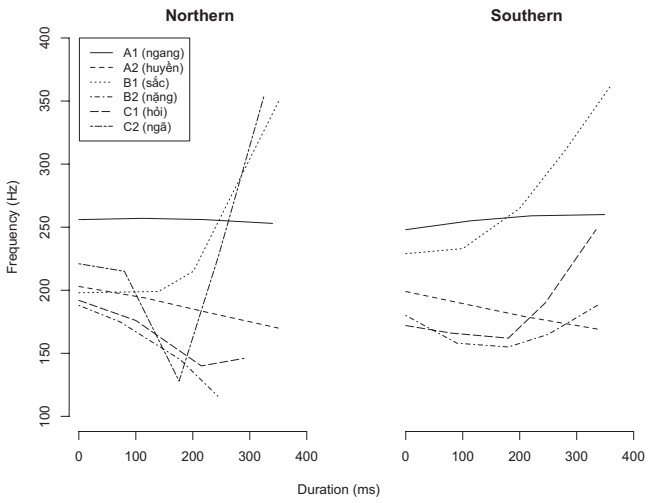

FIG. 1. Average f_0 contours for Northern (Hanoi) and Southern (Ho Chi Minh City) Vietnamese tones (modified from Hoàng, 1986, 1989).

any laryngealization, and perhaps as a result has a longer average duration compared to its NVN counterpart.

B. Discrimination and identification

In Brunelle's (2009) study, NVN and SVN listeners participated in an identification task using resynthesized NVN speech. The results indicated that while NVN and SVN listeners both use a compact set of perceptual cues when identifying tones, these sets are not identical. In particular, NVN listeners relied more on voice quality cues than SVN listeners, which is not surprising if SVN tones are distinguished solely by f_0 differences. While identification tasks can help identify the points on a stimulus continuum where listeners have a perceptual category boundary, they reveal less about the relative saliency of acoustic dimensions in the perceptual space. In addition, identification tasks are off-line, allowing subjects to make considerable use of (or allow interference by) their L1 linguistic knowledge. In order to learn more about the relative salience of the perceptual dimensions in each dialect, this study employed a speeded-response discrimination (AX) task, where subjects made rapid decisions if two stimuli were the same or different. By using a low-memory load task such as speeded AX, the hope was to encourage the use of language-independent, on-line phonetic processing, as opposed to a phonemic response based on L1 (or here, D1) categorization processes (Werker and Logan, 1985).

The goals of the present study were to (a) corroborate the number and type of perceptual dimensions to tone in Northern and Southern Vietnamese and (b) determine the extent to which individual differences in tone perception are influenced by dialect background. If phonetic perception is not influenced by dialect experience, NVN and SVN listeners should respond similarly to NVN speech (i.e., show a similar pattern of errors in judgment and similar reaction times). It is predicted that when tones are more confusable (i.e., closer in perceptual space), participants will (i) make more mistakes when asked to tell whether they are the same or different and (ii) take longer to make the judgment: the shorter the perceptual distance, the longer the reaction time (RT) (Shepard, 1978; Nosofsky, 1992; Huang, 2007).

III. EXPERIMENT

A. Method

1. Stimuli

Stimuli were constructed from 16 kHz, 16 bit digital recordings of a 34-year-old female NVN speaker producing the carrier syllable /ba/ with each of the six NVN tones six times in isolation. Of the resulting 36 productions, three productions of each syllable-tone pair judged to be canonical by the author were selected and normalized (intensity: 75 dB; duration: 400 ms). In most cases, duration normalization entailed syllable shortening, although it meant lengthening in the case of the syllable bearing the 22g tone. Examples of the stimuli used are shown in Fig. 2.

2. Participants

A total of 30 native Vietnamese speakers participated in the experiment (10 NVN, 20 SVN). Northern participants were university students from the Hanoi Medical University, Hanoi University of Technology, or Hanoi National University of Education who had been born and raised in Hanoi; Southern participants were all students at the Vietnam National University of Social Sciences and Humanities who had been born and raised in Ho Chi Minh City. Mean age was 18.6 (s.d. 0.95) for Northern participants and 19.7 (s.d. 1.01) for Southern participants. All of the Northern participants and 15 of the Southern participants were female. None of the participants reported speech or hearing disorders; all had varying degrees of second language fluency in English, French, German, Spanish, Russian, Korean, Mandarin Chinese, and/or Japanese.

3. Procedure

Participants were first presented visually with instructions explaining the experiment, which asked them to indicate whether they had heard two identical syllables or two different syllables by pressing the "g" or "k" keys on a computer keyboard (chosen to correspond to the first letter of the Vietnamese terms for "same" and "different," respectively). A short supervised training session consisting of eight trials was conducted to familiarize participants with the experimental procedure. Participants then heard 42 stimulus pairs in five blocks for a total of 210 stimuli. In each trial, the pair was either identical (e.g., 44–44) or different (e.g., 44–22g). Each token was randomly selected from one of the three productions at each trial; for "same" trials, the same (acoustically identical) production token was used for both elements. Stimuli were played at a 300 ms interstimulus interval (ISI), with a 1000 ms response deadline and a 2000 ms intertrial interval (ITI), from a MacBook Pro laptop via Sony MDR-V600 headphones at a comfortable listening level. Presentation took place in various university classrooms. Judgment accuracy and RTs were recorded using PsyScope X B53 (MacWhinney *et al.*, 1997).

B. Results

All participants showed an understanding of the task as indicated by their lack of errors during the supervised familiarization session. Responses were analyzed for accuracy and reaction time and submitted to both individual differences scaling analysis (INDSCAL) and hierarchical cluster analysis.

1. Accuracy and reaction times

Out of a total of 6300 possible responses, 6038 responses were collected and used in the accuracy analysis. Of these, 1922 were "same" responses and 4116 were "different" responses, of which 1738 were "same" trials and 4300 were "different" trials. For the reaction time analysis, only correct different responses were analyzed; this amounted to 4007 trials, meaning that subjects incorrectly judged 261 "same" trials to be "different."

Table II shows the group error rates (in percent) and median reaction times (in ms) of the correct "different" re-

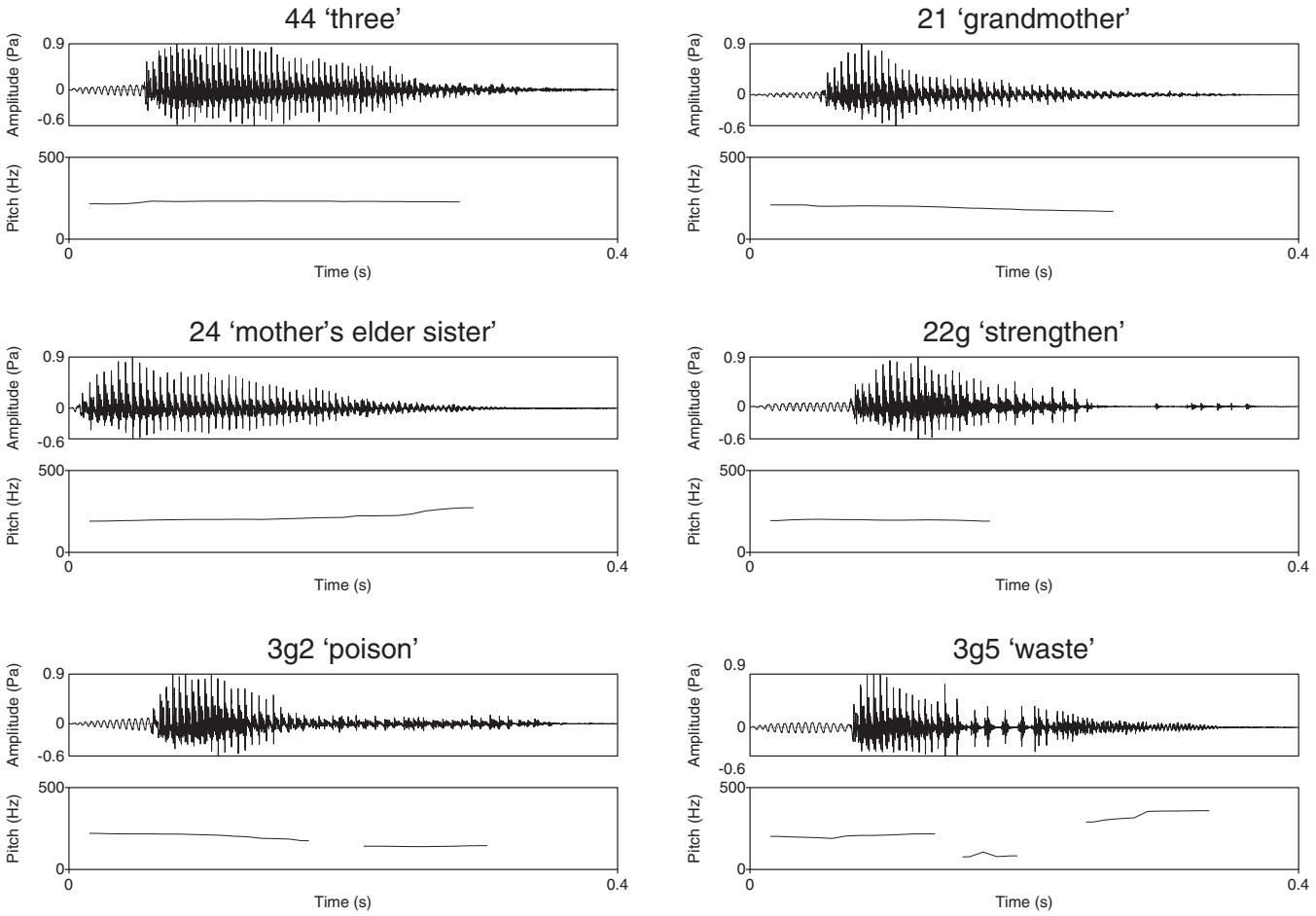

FIG. 2. Waveforms and f_0 tracks of stimuli used in discrimination experiment. Carrier syllable is /ba/ in all cases.

sponses for the nonidentical tone pairs (median values were used due to the right skewing of RT measurements). RT is measured from the onset of the second member of the stimulus pair. Generally speaking, error rates were fairly low, although SVN listeners tended to have lower error rates for all tone pairs except for the pairs 22g/3g2 and 3g2/3g5. The error rate among SVN listeners for the pair 22g/3g2 was particularly high.

TABLE II. Error rates (proportion of 1) and reaction times (in ms), AX discrimination task. Reaction times represent the average computed from the median of correct responses for each participant by tone group.

	Northern		Southern	
Tone pair	Error	RT	Error	RT
44/21	0.07	762	0.05	837
44/24	0.07	721	0.05	789
44/22g	0.07	877	0.03	832
44/3g2	0.07	813	0.03	814
44/3g5	0.11	769	0.02	796
21/24	0.09	817	0.02	831
21/22g	0.07	777	0.04	927
21/3g2	0.10	825	0.02	890
21/3g5	0.09	767	0.02	830
24/22g	0.03	788	0.03	818
24/3g2	0.07	853	0.02	846
24/3g5	0.07	797	0.04	916
22g/3g2	0.06	874	0.39	1049
22g/3g5	0.07	826	0.02	900
3g2/3g5	0.07	809	0.13	969

A binary accuracy value was calculated for each response (correct=1, incorrect=0) and used as the dependent variable in a logistic regression with within-subjects predictor STIMULUS PAIR and between-subjects predictor DIALECT. This revealed a main effect of STIMULUS PAIR $[\chi^2(14)=170.01, p<0.001]$ and an interaction between DIALECT and STIMULUS PAIR $[\chi^2(14)=95.56, p<0.001]$. Within-group comparisons on ACCURACY by STIMULUS PAIR did not reach significance for either dialect group. This may be due to the homogeneity of the error rates across stimulus pairs as well as to the low overall error rate.

A repeated measures analysis of variance was performed on the REACTION TIME of participants' correct "different" responses, with STIMULUS PAIR as the within-subject variable (15 levels) and DIALECT as the between-subjects variable (2 levels). There were significant effects of both STIMULUS PAIR $[F(1,14)=8.06, p<0.001]$ and DIALECT $[F(1,1)=75.38, p<0.001]$. The interaction did not reach significance $[F(1,14)=1.58, p=0.076]$.

Two types of *post hoc* pairwise comparisons were performed: between-group ("do reaction times of Northern listeners differ from that of Southern listeners for a given stimulus pair?") and within-group ("do listeners of a given dialect group take longer to respond to some stimulus pairs

TABLE III. Between-group comparisons of log-transformed reaction time results. Significance codes: *=<0.05; **=<0.01; ***=<0.003 (Bonferroni corrected α).

Tone pair	t	df	p	Tone pair	t	df	p
44/21	−2.50	219.72	*	21/3g2	−2.20	182.91	*
44/24	−3.36	208.17	***	21/3g5	−3.06	200.90	***
44/22g	−0.37	206.72		24/22g	−1.60	195.55	
44/3g2	−2.25	201.41	*	24/3g2	−2.49	222.98	*
44/3g5	−2.19	201.87	*	24/3g5	−2.27	196.46	*
21/24	−2.50	35	*	22g/3g2	−4.16	197.75	***
21/22g	−3.53	203.76	***	22g/3g5	−2.61	224.63	**
				3g2/3g5	−4.77	230.72	***

than others?"). Because raw reaction times naturally skew rather heavily, log-transformed values were used to ensure normality.

Between-group comparisons of the tone pairs indicated that the confusability of a given stimulus pair differed by dialect group. In particular, reaction times of NVN and SVN listeners differed significantly for the tone pairs 44/24, 21/22g, 21/3g5, 22g/3g2, and 3g2/3g5, as shown in Table III. The difference between these last two pairs was highlighted by the within-group comparisons, which confirmed the different treatment of laryngealized tone pairs by NVN and SVN listeners. SVN listeners found the 22g/3g2 pair to be significantly more confusable than all other stimulus pairs, with the possible exception of the pair 3g2/3g5. The rate of confusion among NVN listeners for the 22g/3g2 pair was also significantly different, or approached significance, for many tone pairs. SVN listeners, however, also found the pair 3g2/3g5 to be significantly confusable with several other pairs, whereas NVN listeners did not seem to find this pair significantly more confusable than any other. Table IV shows the results of two sets of pairwise comparisons. As no other within-group pairwise comparisons approached significance for either group, the results have been omitted for brevity.

2. Multidimensional scaling

To explore the factors affecting listeners' tone perception, a multidimensional scaling analysis was performed on

TABLE IV. Partial within-group comparisons for log-transformed reaction time results. Significance codes: .=<0.05; *=<0.01; **=<0.001; ***=<0.0005 (Bonferroni corrected α).

		Northern			Southern		
Pair 1	Pair 2	t	df	p	t	df	p
22g/3g2	44/21	3.67	160.85	***	5.41	253.24	***
	44/24	4.21	162.89	***	5.38	240.80	***
	44/22g	0.85	160.89		4.68	260.90	***
	44/3g2	3.31	159.97	**	5.28	261.19	***
	44/3g5	3.57	157.99	**	5.82	242.95	***
	21/24	3.23	153.94	*	5.29	234.84	***
	21/22g	2.59	158.26	*	3.97	224.21	***
	21/3g2	1.41	156.05		3.64	245.36	***
	21/3g5	3.93	160.57	***	5.57	230.18	***
	24/22g	2.57	164.22	*	5.70	218.78	***
	24/3g2	2.93	160.39	*	4.82	250.04	***
	22g/3g5	1.52	159.07		3.73	227.43	***
	3g2/3g5	2.71	159.99	*	2.00	270.29	.
3g2/3g5	44/21	0.95	158.89		3.42	346.72	***
	44/24	1.51	160.92		3.44	338.85	***
	44/22g	−1.73	158.82		2.68	354.66	*
	44/3g2	0.79	157.88		3.30	354.33	**
	44/3g5	0.85	156.00		3.79	342.59	***
	21/24	0.30	152.14		3.21	335.44	**
	21/22g	−0.25	156.38		1.80	323.95	
	21/3g2	−1.16	153.96		1.53	343.18	
	21/3g5	1.17	158.62		3.49	330.78	***
	24/22g	−0.18	162.29		3.58	319.08	***
	24/3g2	0.15	158.46		2.78	348.60	*
	22g/3g2	−2.71	159.99	*	−2.00	270.29	.
	22g/3g5	−1.39	157.29		1.55	328.07	

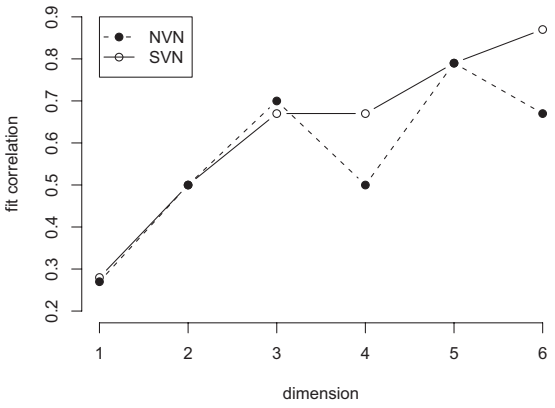

FIG. 3. Fit correlation (R^2) by MDS dimensionality solution.

the correct "different" responses. Following Shepard (1978) and Huang (2007), RTs were converted into perceptual distances using the reciprocal function

$$\text{distance} = \frac{1}{\log(\text{RT})} \tag{1}$$

with the results entered into a symmetrical distance matrix for each listener (averaged across both tone pairs in a group). Figure 3 shows the goodness of fit for multidimensional scaling (MDS) solutions of varying dimensionality. For both groups, there is a relatively clear elbow in the goodness-of-fit curve at three dimensions, and therefore a three-dimensional (3D) solution was initially calculated using three-way individual differences scaling (INDSCAL; Carroll and Chang, 1970) implemented in PRAAT (Boersma and Weenink, 2008). INDSCAL performs metric multidimensional scaling using separate matrices for each subject, assuming a common number of underlying perceptual dimensions across subjects. Note that INDSCAL is "dimensionally unique," meaning that the solution axes are fixed and cannot be arbitrarily transformed.

The resulting solutions are shown in Figs. 4 and 5. The interpretation of dimensions is similar for both dialect groups. The vertical axis corresponds to voice quality, with laryngealized tones forming one group and nonlaryngealized another. One horizontal axis may be taken to represented f_0 offset, with a group of "low" tones on one side and "high" on the other. The other horizontal dimension might be interpreted as "contour type;" although this interpretation is less clear-cut, the tones could be regarded as falling into three contour groups from left to right, "level," "complex," and "simple."

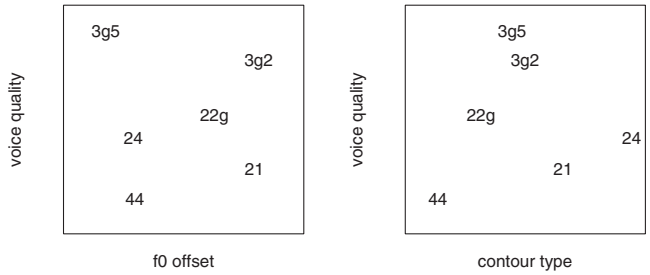

FIG. 4. 3D perceptual tone space for NVN listeners.

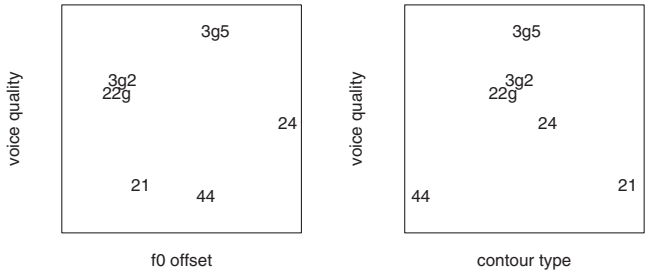

FIG. 5. 3D perceptual tone space for SVN listeners.

The individual subject weights are shown in Figs. 6 and 7. While no clear pattern may be discerned from the distribution of SVN subject weights, NVN listeners appear to form two groups with respect to their weighting of the f_0 offset dimension. Subsequent *post hoc* analysis of the two groups did not reveal any significant differences between them. That is, the MDS solutions computed for each subset of NVN listeners are highly similar to those computed for the entire group. It may simply be the case that, because fewer NVN subjects were tested, an artificial gap in the weights is observed, and that the addition of more subjects would reveal a more evenly distributed set of weights, as observed for the SVN listeners.

3. Hierarchical clustering

In order to assist with linking the geometric properties of the MDS solutions to substantive acoustic features, the reaction time data were also submitted to a hierarchical clustering analysis. The results of complete linkage clustering are shown here, although similar results were obtained using other types of agglomeration methods (average linkage clustering, Ward's minimum variance, etc.). Complete linkage clustering is often appropriate for cases when objects form natural groups, as they are expected to in this case.

The results of the hierarchical clustering are shown in Fig. 8. The perceptual dissimilarity between any two tones is represented by the sum of the lengths of the fewest number of vertical branches connecting the two nodes. The analysis reveals differences in both the structure and perceptual similarity of the tone space between listener groups. Northern listeners found the high-offset tones, as a group, distinct from all other tones, particularly low-offset tones. The group of tones sharing a laryngeal voice quality component did not form a unified node, suggesting that NVN listeners are well versed at perceiving this cue. Nevertheless, the laryngealized low tones emerged as extremely similar. This may be due to

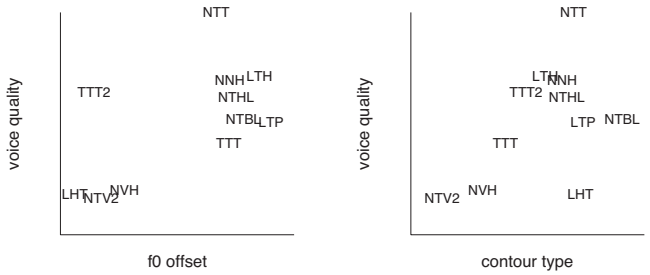

FIG. 6. Subject space of the 3D INDSCAL solution for NVN listeners.

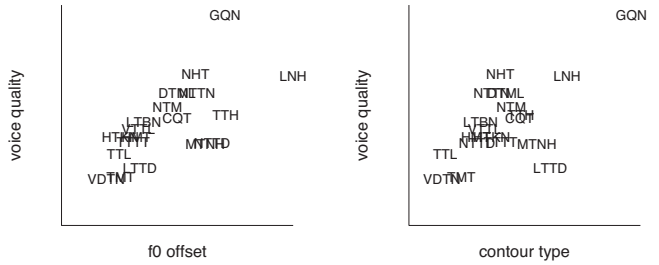

FIG. 7. Subject space of the 3D INDSCAL solution for SVN listeners.

the fact that, since durational cues were neutralized in this experiment, the normally short 22g tone was difficult to distinguish from 3g2 (see Brunelle, 2009).

For SVN listeners, two main nodes emerge, based on the voice quality of the tones. Within the nonmodal group, 3g5 may be less confusable with the other tones because of its rather striking "broken" (midglottalized) nature, which is probably highly salient to SVN listeners despite not being part of their production repertoire.

IV. GENERAL DISCUSSION

While the dimensions of the perceptual space found by the MDS solution are the same for both listener groups, the distribution of tones within that space differs by dialect. In particular, stimulus pairs involving laryngealized tones (22g, 3g2, and 3g5) were much more confusable for SVN listeners than for NVN listeners, suggesting that that SVN listeners were, in fact, perceiving this acoustic-phonetic information. However, the patterns of confusion differed for NVN and SVN listener groups. In particular, the tone pairs 3g2/3g5 and 22g/3g2 clearly stand out from the others. These pairs showed higher error rates among SVN listeners, as well as significantly different reaction times within and across groups. Furthermore, hierarchical clustering found that the laryngealized tones formed a clear node in the SVN perceptual space, whereas NVN listeners seemed to find tones with similar f_0 profiles more confusable than those which shared voice quality.

Taken together, these results suggest that primary dialect experience, in addition to native language experience, affects perceptual tuning: if NVN listeners have come to rely heavily on voice quality cues to distinguish tones (as argued by Brunelle, 2009), then they may be less attuned to subtle differences in f_0, making tones which differ along this dimension more confusable. Similarly, if SVN listeners are more highly attuned to f_0 distinctions, tones distinguished

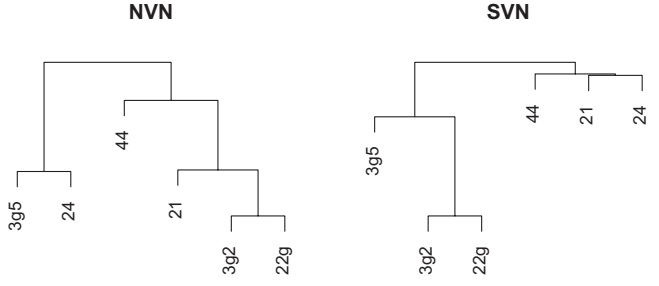

FIG. 8. Hierarchical clustering solutions for NVN and SVN listener groups.

purely along this dimension would be more easily discriminated, whereas tones containing unfamiliar acoustic-phonetic cues might be more easily confused.

It is important to note that, despite only a limited familiarity with cues to voice quality, SVN listeners do show some sensitivity to these cues in both discrimination and identification tasks. However, this sensitivity emerges much more obviously in an on-line, bottom-up task like discrimination, where the presence of an unfamiliar cue like voice quality may actually aid SVN listeners when discriminating between two syllables where one is laryngealized and one is not. Conversely, in an off-line task like identification, this same unfamiliarity can make it difficult to associate the (nevertheless acoustically striking) cue with a particular tone. Thus, while SVN listeners are clearly sensitive to the acoustic properties unique to NVN tones, they may dismiss those cues as irrelevant in top-down tasks such as identification, which would explain the difference between these results and those of Brunelle (2009).

Interestingly, the hierarchical clustering results suggest that NVN listeners do not perceive laryngealization to be particularly salient, despite the fact that it is an important cue distinguishing tones in that dialect. This result also highlights importance differences in the effects of dialect experience given the parameters of the experimental task. Gordon *et al.* (1993) demonstrated that under conditions of comparatively light attentional load, listeners pay less attention to primary cues than to secondary ones. If discrimination constitutes a situation of relatively limited attentional availability for NVN listeners (compared to the relatively unlimited amount of attention available during an identification task), this may in part explain why they perceive primary laryngealization cues as less salient in the results reported here.

For SVN listeners, on the other hand, these cues may stand out in prelinguistic processing precisely because of their unfamiliarity. An unexpected result of this experiment was that in trials where at most one of the stimuli were laryngealized, SVN listeners appear better at discriminating NVN stimuli than NVN listeners (Table II). As noted above, this may be a result of SVN listeners benefiting from unfamiliar acoustic information in a discrimination task. A reviewer suggests that another explanation may lie in the comparatively longer reaction times of SVN respondents, i.e., they may be responding more accurately because they are taking longer to issue a response. However, if both high error rate and large reaction times are taken as inversely proportional to perceptual distance, this is a somewhat paradoxical result. One possible reason for this may lie in the length of the interstimulus interval used in the discrimination task. Previous studies have shown that non-native listeners' performance on discrimination tasks is better in short ISI conditions, while native listeners' performance is better in longer ISI conditions (Werker and Tees, 1984; Burnham and Francis, 1997). Thus, the higher average accuracy rates for SVN listeners may reflect a task-dependent processing advantage more than an accurate measure of perceptual distance.

Finally, it is worth noting that while the pair 22g/3g2 was most confusable for both groups of listeners, there may be distinct reasons for this confusability. As mentioned ear-

lier, the citation forms of the NVN 22g tone are often of significantly shorter duration compared to other tones, due to the effect of the strong final glottalization. Although it has been claimed that this durational difference may be lost in running speech (Vũ, 1982; Brunelle, 2009), it is not clear to what extent listeners of either dialect are sensitive to citation-form differences in duration. If NVN speakers do use temporal information as a means of disambiguating this tone, the lack of that information may have contributed to its confusion with the spectrally similar 3g2 tone. For SVN listeners, on the other hand, there is no evidence (anecdotal or otherwise) that temporal cues play a role in dialect-specific tone processing, and since the SVN equivalent of 3g2 has a much more pronounced f_0 rise than the NVN stimuli used here (see Fig. 1), the surfeit of f_0 information, combined with the presence of unfamiliar voice quality cues, may have contributed to the high confusability of the 22g/3g2 tone pair. Further research will be necessary to explore the role of duration in tonal processing as well as the precise perceptual consequences of the differences in the pitch excursions between tones in the two dialects.

V. CONCLUSIONS

When speakers of a tone language are presented with speech from a non-native dialect, speakers might ignore cues not relevant in their own dialect or they may show some sensitivity to this input. The results of the experiment reported here show that Southern Vietnamese listeners are sensitive to acoustic cues to tone not present in their native dialect, but that their sensitivity differs from that of Northern Vietnamese speakers, for whom the cues are present natively. While the structure of the perceptual tone space is similar across dialects, the relative saliency of the dimensions was found to vary by dialect. This confirms a role for primary dialect experience as well as primary language experience at prelinguistic levels of suprasegmental processing.

ACKNOWLEDGMENTS

This work was carried out while the author was a visiting researcher at the Multimedia, Information, Communication and Applications (MICA) Centre, Hanoi University of Technology, August–December 2008. Thanks to Dr. Eric Castelli and TS Phạm Thị Ngọc Yến for making this visit possible, and to Marc Brunelle, Trần Đỗ Đạt, John Ingram, Mathias Rossignol, and Nguyễn Việt Sơn for valuable discussion. The author also wishes to acknowledge Vũ Thị Thanh Huyền and Nicolas Boffo, who helped recruit subjects in Hanoi, and Nguyễn Văn Huệ, Nguyễn Duy Doài, and Đinh Lư Giang of the Vietnam National University of Social Sciences and Humanities as well as Văn Ly for their assistance and hospitality in Hồ Chí Minh City.

Abramson, A. S., Nye, P. W., and Luangthongkum, T. (**2007**). "Voice register in Khmu': Experiments in production and perception," Phonetica **64**, 80–104.

Andreev, N. D., and Gordina, M. V. (**1957**). "The system of tones in the Vietnamese language (on experimental data)," Vestnik Leningradskogo gosudarstvennogo Universiteta **8**, 132–148 [in Russian].

Boersma, P., and Weenink, D. (**2008**). "Praat: Doing phonetics by computer (version 5.0.24) [computer program]," http://www.praat.org (Last viewed 9/23/2008).

Brunelle, M. (**2003**). "Tonal coarticulation effects in Northern Vietnamese," in Proceedings of the 15th International Congress of Phonetic Sciences, pp. 2673–2676.

Brunelle, M. (**2009**). "Tone perception in Northern and Southern Vietnamese," J. Phonetics **37**, 79–96.

Burnham, D., and Francis, E. (**1997**). "The role of linguistic experience in the perception of Thai tones," in *Southeast Asian linguistic Studies in Honour of Vichin Panupong*, Science of Language Vol. **8**, edited by A. Abramson (Chulalongkorn University Press, Bangkok).

Carroll, J. D., and Chang, J.-J. (**1970**). "Analysis of individual differences in multidimensional scaling via an N-way generalization of 'Eckart-Young' decomposition," Psychometrika **35**, 283–319.

Cutler, A. and Otake, T. (**1999**). "Pitch accent in spoken-word recognition in Japanese," J. Acoust. Soc. Am. **105**, 1877–1888.

Đoàn, T. T. (**1977**). *Ngữ âm tiếng Việt (Vietnamese Phonology)* (Nhà Xuất Bản Đại Học Quốc Gia, Hanoi).

Gandour, J. T. (**1983**). "Tone perception in far eastern languages," J. Phonetics **11**, 149–175.

Gandour, J. T., and Harshman, R. (**1978**). "Crosslanguage differences in tone perception: A multidimensional scaling investigation," Lang. Speech **21**, 1–33.

Gordon, P. C., Eberhardt, J. L., and Rueckl, J. G. (**1993**). "Attentional modulation of the phonetic significance of acoustic cues," Cognit. Psychol. **25**, 1–42.

Gottfried, T. L., and Beddor, P. S. (**1988**). "Perception of temporal and spectral information in French vowels," Lang. Speech **31**, 57–75.

Gsell, R. (**1980**). "Remarques sur la structure l'espace tonal en Vietnamien du Sud (parler de Saigon) [Remarks on the structure of the tonal space of Southern Vietnamese (Saigonese)]," Cahiers d'Études Vietnamiennes **4**, 1–26.

Hoàng, C. C. (**1986**). "Suy nghĩ thêm về thanh điệu tiếng Việt (Further thoughts on Vietnamese tones)," Ngôn ngữ **3**, 19–38.

Hoàng, T. C. (**1989**). *Tiếng Việt trên các miền đất nước (Vietnamese in All Regions of the Country)* (Nhà Xuất Bản Khoa Học Xã Hội, Hanoi).

Huang, T. (**2007**). "Perception of Mandarin tones by Chinese-and English-speaking listeners," in Proceedings of the 16th International Congress of the Phonetic Sciences, pp. 1797–2000.

Lee, L., and Nusbaum, H. (**1993**). "Processing interactions between segmental and suprasegmental information in native speakers of English and Mandarin Chinese," Percept. Psychophys. **53**, 157–165.

Lisker, L., and Abramson, A. (**1970**). "The voicing dimension: some experiments in comparative phonetics," in Proceedings of the 6th International Congress of Phonetic Sciences, Prague, pp. 563–567.

MacKain, K. S., Best, C. T., and Strange, W. (**1980**). "Native language effects on the perception of liquids," J. Acoust. Soc. Am. **27**, 527.

MacWhinney, B., Cohen, J. D., and Provost, J. (**1997**). "The PsyScope experiment-building system," Spatial Vis. **11**, 99–101.

Michaud, A. (**2004**). "Final consonants and glottalization: New perspectives from Hanoi Vietnamese," Phonetica **61**, 119–146.

Miller, J. L., and Grosjean, F. (**1997**). "Dialect effects in vowel perception: The role of temporal information in French," Lang. Speech **40**, 277–288.

Minematsu, N., and Hirose, K. (**1995**). "Role of prosodic features in the human process of perceiving spoken words and sentences in Japanese," J. Acoust. Soc. Jpn. **16**, 311–320.

Nguyễn, V. L., and Edmonson, J. A. (**1998**). "Tones and voice quality in modern Northern Vietnamese: Instrumental case studies," Mon-Khmer Studies **28**, 1–18.

Nosofsky, R. M. (**1992**). "Similarity scaling and cognitive process models," Annu. Rev. Psychol. **43**, 25–53.

Otake, T., and Cutler, A. (**1999**). "Perception of suprasegmental structure in a non-native dialect," J. Phonetics **27**, 229–253.

Phạm, A. (**2003**). *Vietnamese Tone: A New Analysis* (Routledge, New York).

Shepard, R. (**1978**). "The circumplex and related topological manifolds in the study of perception," in *Theory Construction and Data Analysis in the Social Sciences*, edited by S. Shye (Jossey-Bass, San Francisco).

Strange, W. (**1995**). *Speech Perception and Linguistic Experience: Issues in Cross-Language Research* (York Press, Baltimore).

Sumner, M., and Samuel, A. G. (**2009**). "The effect of experience on the perception and representation of dialect variants," J. Mem. Lang. **60**, 487–501.

Suwilai, P. (**2001**). "Tonogenesis in Khmu dialects of SEA," Mon-Khmer Studies **31**, 47–56.

Svantesson, J.-O., and House, D. (**2006**). "Tone production, tone perception and Kammu tonogenesis," Phonology **23**, 309–443.
Thompson, L. C. (**1959**). "Saigon phonemics," Language **35**, 454–476.
Thompson, L. C. (**1965**). *A Vietnamese Grammar* (University of Washington, Seattle).
Trần, H. M. (**1967**). "Tones and intonation in south Vietnamese," in *Series A—Occasional Paper No. 9, Papers in Southeast Asian Linguistics No. 1*, edited by D. L. Nguyễn, H. M. Trần, and D. Dellinger (Linguistics Circle of Canberra, Canberra).
Vũ, T. P. (**1981**). "The acoustic and perceptual nature of tone in Vietnamese," Ph.D. thesis, Australian National University, Sydney.
Vũ, T. P. (**1982**). "Phonetic properties of Vietnamese tones across dialects," in *Papers in Southeast Asian Linguistics*, Tonation Vol. **8**, edited by D. Bradley (Australian National University, Sydney), pp. 55–75.
Wayland, R., and Guion, S. (**2004**). "Training English and Chinese listeners to perceive Thai tones: A preliminary report," Lang. Learn. **54**, 681–712.
Werker, J. F., and Logan, J. S. (**1985**). "Cross-language evidence for three factors in speech perception," Percept. Psychophys. **37**, 24–44.
Werker, J. F., and Tees, R. C. (**1984**). "Phonemic and phonetic factors in adult cross-language speech perception," J. Acoust. Soc. Am. **75**, 1866–1878.