Place-Dependent VOT in L2 Acquisition

Kuniya Nasukawa
Tohoku Gakuin University

1. Introduction[1]

It has been acknowledged that *voice onset time* (VOT) behaves as an important phonetic cue for voicing categorization typically in (word-initial) prevocalic position. The value of VOT is in general sensitive to place of articulation. For example, bilabial stops occur with shorter VOT, velar stops are reproduced with longer VOT, and alveolar stops lie somewhere in between.

This study investigates the different degrees of improvement that L2 learners of English experience in the production of VOT. Two sets of comparative data were collected, one at the beginning of the training period and another at the end, from six lower-intermediate students of English at a Japanese university. The initial data showed no significant differences in VOT between their L1 (Japanese) and their L2 (English): they produced English voiceless stops with VOT values similar to those used in Japanese. However, after 9 months of auditory and visual training there emerged different results according to place of articulation: alveolar and velar exhibited a clear improvement in the production of VOT for English voiceless stops. Although it remained shorter than the VOT value of voiceless stops in English monolinguals, their VOT value for English voiceless stops clearly became longer than that for Japanese voiceless stops. On the other hand, VOT in the English voiceless bilabial stop remained similar to the bilabial stop in Japanese.

Based on these results, this study discusses why the VOT value of the voiceless bilabial stop is difficult to acquire compared with those of voiceless alveolar and velar stops. The discussion goes beyond the physiological mechanisms of speech production, and assumes that the phonological structures of the sounds in question influence the degree of improvement in acquiring a native-like VOT. According to one version of Element Theory (Backley & Nasukawa, 2009a, 2009b; Nasukawa & Backley, 2005, 2008)—which departs from orthodox distinctive feature theories in terms of its perception-based view of phonetic realization and its use of monovalent features (elements) to represent contrastive properties—the resonance element marking bilabials (known as |rump| and represented by |U|) is structurally headed and therefore dominant while the elements marking alveolars and velars are non-headed. VOT in English (aspiration) is also represented by a structurally-headed element, the noise element (represented by |H|). Having this combination of two headed elements in a single segment such as a voiceless bilabial stop makes it difficult for Japanese L2 learners of English to acquire the VOT value for the English voiceless bilabial stop, since the phonological grammar of their L1 does not employ double headedness of this kind.

The structure of this paper is as follows. Section 2 briefly reviews a cross-linguistic study of two-way laryngeal-source contrasts in VOT and a cross-linguistic tendency for place-dependent VOT values. Section 3 describes the method employed in this study and the following section presents its results. Before examining the data and providing a phonological analysis of the issue within the framework of Element Theory, section 5 introduces some fundamental notions employed by Element Theory in comparison with orthodox feature theory; this is for the benefit of those readers who are unfamiliar with the theory. Section 6 then discusses the structural differences between voiceless stops

[1] This research was supported by a Faculty Research Grant for Improvement in Teaching and Learning Strategies 2006–2008 and also partially supported by the Japanese Ministry of Education, Culture, Sports, Science and Technology under grant 22320090. This paper has benefited from insightful comments by Phillip Backley and two anonymous reviewers. Also I thank Hitoshi Muranoi for helping me with the statistical analysis of the data.

© 2010 Kuniya Nasukawa. Selected Proceedings of the 2008 Second Language Research Forum, ed. Matthew
T. Prior et al., 197-210. Somerville, MA: Cascadilla Proceedings Project.

in Japanese and in English. It goes on to explain why Japanese L2 learners of English show little improvement in their production of VOT for the English voiceless bilabial, and discusses some of the pedagogical implications. The final section offers a short conclusion.

2. VOT in English and Japanese

2.1. An overview

Before discussing the L2 acquisition data, this section presents a cross-linguistic overview of VOT contrasts and shows how these operate in English and Japanese. VOT constitutes one aspect of laryngeal-source contrasts, and refers to the timing relation between the release of stop closure and the beginning of vocal-fold vibration in the following vowel (Abramson & Lisker, 1970; Lisker & Abramson, 1964; cf. Harris, 1994). This is typically observed in prevocalic position.

Stop plus vowel sequence
Glottal trace: unshaded lines = vocal folds abducted; shading = vocal folds adducted
“0” = the release of stop closure

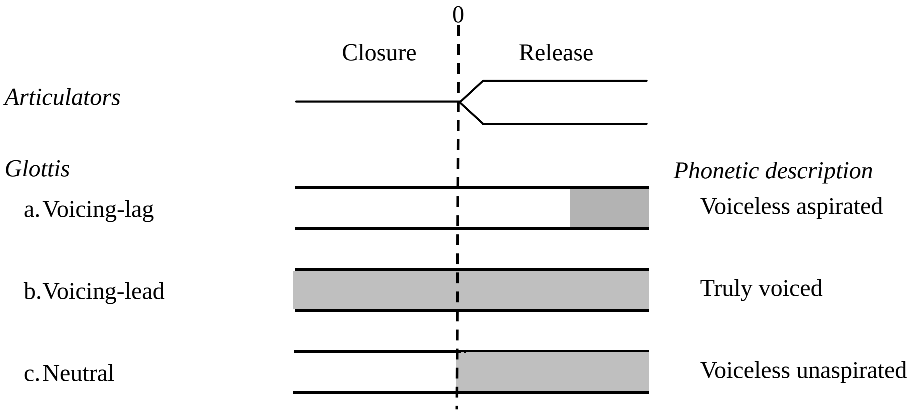

Figure 1. Voice onset time.

Limiting the argument to languages exhibiting a two-way laryngeal-source contrast, studies of VOT reveal that there are two distinct ways of creating this type of contrast. One of the ways is found in languages such as English and Swedish, and bases the contrast on the distinction between long voicing lag (variously referred to as positive VOT, voiceless aspirated or fortis: Figure 1a) and zero/short voicing lag (alternatively called neutral VOT, voiceless unaspirated, neutral or lenis: Figure 1c). The other way is taken up in languages such as Japanese and Spanish, and shows a contrast between long voicing lead (also known as negative VOT, voiced unaspirated or truly voiced: Figure 1b) and zero/short voicing lag (Figure 1c). In both cases, neutral VOT is always involved: no language forms a two-way laryngeal-source contrast without utilizing neutral VOT.

Table 1. *Language Systems with a Two-Way Laryngeal-Source Contrast*

Type	Language	Voicing-lead	Neutral	Voicing-lag
I	English	▨	✓	✓
II	Japanese	✓	✓	▨

The following figure shows schematic representations of the VOT continuum for English and Japanese monolinguals. “0” indicates the release of stop closure. The numbers in parentheses indicate mean VOT values (ms).

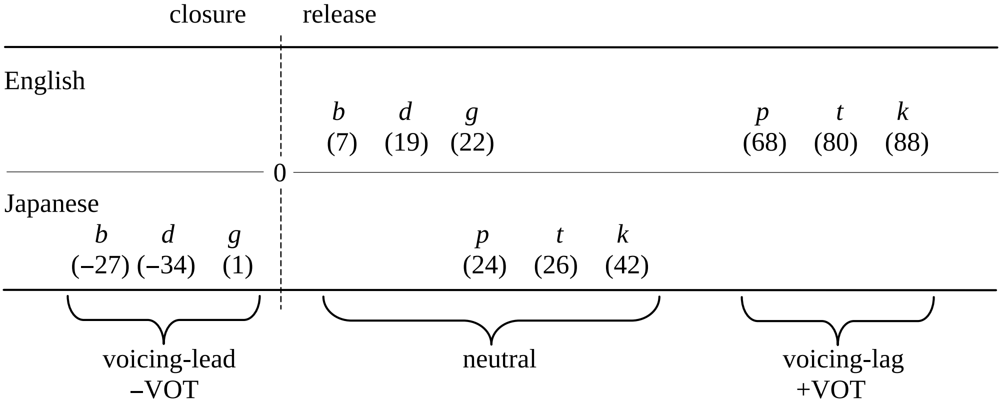

Figure 2. Mean VOT values for English and Japanese stops produced by monolingual adults (adapted from Harada, 2007, p. 372).

As illustrated in Figure 2, the English voiceless stops /p, t, k/ exhibit a relatively long time lag between closure release and the onset of voicing, which is often described as voiceless aspirated. On the other hand, the English ‘voiced’ stops /b, d, g/ are actually voiceless unaspirated and show a relatively short time lag between closure release and the onset of voicing. Japanese voiceless stops /p, t, k/ are produced in a VOT region similar to that used for the English ‘voiced’ stops, while in the case of the Japanese voiced stops /b, d, g/, which are described as truly voiced, there is a relatively long lead time between the onset of voicing and stop release. Although, in precise terms, English voiced stops and Japanese voiceless stops show different time lags, they are often considered to belong in the same region since they behave phonologically in a similar manner. (It is customary to use the labels *long lead* or *–VOT* for the category of truly voiced stops, *long lag* or *+VOT* for the aspirated voiceless stops, and *short lag* or *zero VOT* for plain stops.) Keating (1984) hypothesizes that this polarization of phonetic values within the neutral categories serves to enhance phonetic differences within the opposing category: that is, voicing lead in Japanese; on the other hand, voicing lag in English.

2.2. VOT and place of articulation

It is widely recognized that VOT varies with place of stop articulation (Fant, 1973, p. 64; Jakobson & Waugh, 1979, pp. 102–103; Kent & Read, 1992, p. 114; Shimizu, 1996, chap. 8): bilabials are produced with the shortest VOT values, velars are produced with the longest VOT values, and alveolars lie somewhere in between. These relative differences in phonetic distance among stops have been investigated cross-linguistically. Typical examples are given from English (Harada, 2007, Lisker & Abramson, 1964) and Swedish (Fant, 1973), as illustrated in Table 2.

Table 2. *Place-Dependent VOT Values*

shortest → longest				
	bilabials	alveolars	velars	
	(labials)	(coronals)	(dorsals)	
	p	*t*	*k*	
English	58 ms	70 ms	80 ms	(Lisker & Bramson, 1964)
	68 ms	80 ms	88 ms	(Harada, 2007)
Swedish	40 ms	50 ms	60 ms	(Fant, 1973)

Acoustic explanations are also found. In Kewley-Port (1983), for example, place-dependent VOT values are explained by referring to the acoustic distinction [compact]-[diffuse], which refers to the concentration and spread of energy across the spectrum (cf. Jakobson, Fant, & Halle, 1952) and [flat], [falling], [intermediate], which relate to spectral patterns of the release burst in stops.

As we will see, the L2 acquisition data demonstrate that, among the three types of voiceless stops, the bilabial stop has the lowest degree of improvement. The relative degrees of improvement may be schematized as follows:

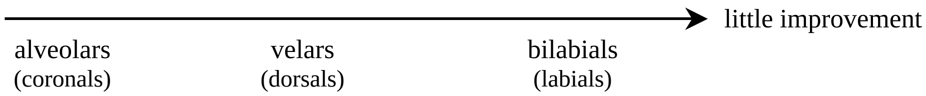

Figure 3. Degree of improvement in acquisition of English VOT values by L2 learners.

This discussion assumes that the difficulty with VOT production in the English bilabial stop is attributed to a difference in the phonological structure of this sound between L1 and L2.

3. Method

3.1. Subjects

Two sets of comparative data were collected for this study, one at the beginning of the training period and another at the end, from six lower-intermediate students of English at a Japanese university. They all grew up in Tohoku, the northeastern region of mainland Japan and were between 20 and 23 years old. At the beginning of the training period it was confirmed that the subjects had no knowledge of VOT.

3.2. Procedure

At the beginning of the training period, first the English data were collected and then the Japanese data were collected from the same group, in order to compare their English and Japanese pronunciations. Each session consisted of a 15-minute reading test administered in a sound-proof recording studio in the English department at Tohoku Gakuin University. After a brief practice reading, each subject read the materials twice at natural speed, having been instructed to place a short pause before each target word. After the data were collected, subjects were given audio-visual aids and told to focus on specific problems they had in pronouncing native-like English VOT. Typically, they received explanations on the following and other points:

(i) The acoustic differences between English and Japanese VOT values (using acoustic analysis software);
(ii) The physiological mechanism for VOT production;
(iii) Vowel quality after VOT.

For (i) the subjects were shown Figure 1. In addition, in a supporting tutorial they were asked to use teaching materials (such as voicing basics, plosives basics and VOT) accessible from the Web site (http://www.phon.ucl.ac.uk/resource/tutorials.html#phon) maintained by the research department of Speech, Hearing and Phonetic Sciences, division of Psychology and Language Sciences, University College London (UCL). As for (ii), phonation processes were explained chiefly by showing subjects illustrations of different glottal states of the kind which can be found in most introductory-level phonetics course books. Additionally, subjects accessed information about physiological mechanisms from UCL web tutorials in order to observe the visual movements relating to different glottal states. With respect to (iii), I first gave subjects training in the pronunciation of vowels preceded by the English voiceless aspirated stops. Then I used acoustic analysis software to explain the differences in length, amplitude and pitch between the English and Japanese vowels: when English L1 speakers put stress on the following vowel they make the vowel longer, louder and higher in pitch than other vowels in the word, whereas Japanese L1 speakers tend to make the following vowel shorter and also lower in amplitude.

The same instructions were given once every three weeks over a period of 9 months. Then, following the same procedures, another set of data were collected from subjects at the end of the training period.

3.3. Materials

The selected test words were all names that often appear in English textbooks. This made it easier for the subjects to pronounce the words without any overt prior instruction. In principle, the Japanese versions of the same names were used for recording the Japanese data. Only ‘teddo,’ which in Japanese corresponds to the English name ‘Ted,’ is replaced by ‘tetsuya’ (a Japanese male name) since ‘teddo’ contains an unnatural sequence in Japanese. The subjects had to say a word in the carrier phrase ‘He said, …______’ for English and in the carrier phrase ‘kare ga yuu ni wa …______’ (He said, …) for Japanese. The materials are given below.[2]

English VOT materials
/p/ He said, … Pete is a puppy.
/b/ He said, … Bill is a boxer.
/t/ He said, … Ted is a teacher.
/d/ He said, … David is a diver.
/k/ He said, … Kate is a cook.
/g/ He said, … Guy is a ghost.

Japanese VOT materials
/p/ kare ga yuu ni wa, … piitaa wa poteto ga suki desu. (… wa = topic marker)
/b/ kare ga yuu ni wa, … biru wa beekon ga suki desu. (… ga suki desu = ‘… likes …’)
/t/ kare ga yuu ni wa, … tetsuya wa tomato ga suki desu.
/d/ kare ga yuu ni wa, … deibitto wa doonatsu ga suki desu.
/k/ kare ga yuu ni wa, … keeto wa keeki ga suki desu.
/g/ kare ga yuu ni wa, … gai wa gintara ga suki desu.

3.4. Data measurement

Each of the experimental words appeared twice on each list. This means that the total number of tokens for each corpus is 84 (6 words × 2 repetitions × 7 subjects). The SFS (Speech Filing System) speech analysis program[3] was used for data measurement. The VOT of stops was measured by finding the nearest millisecond from the beginning of the release burst to the onset of voicing energy in F2 formants with the occasional use of the waveform.

4. Results

The initial data collected before the training period show no clear differences in the VOT value of voiceless stops between L1 (Japanese) and L2 (English): subjects produced English stops with VOT values similar those used in Japanese, as illustrated in Figure 4.

[2] In order to investigate loanword phonology, the study originally aimed to compare differences in VOT between Japanese and English stops in correlated words such as English names and borrowed Japanese versions of the same names. The reason why only the pair *Ted*/*Tetsuya* does not conform to this original-borrowed relation comes down to degree of familiarity: the Japanese form corresponding to *Ted* should be *Teddo* or *Tetto* (owing to a ban on voiced geminates: Ito & Mester, 2003, pp. 49–50, et passim), but these forms would have been unfamiliar and unnatural to our subjects and were therefore avoided. Instead, the Japanese name *Tetsuya* was used. Admittedly, this choice may not be entirely suited to my initial aim. If the present study were to be extended in the future, a different English name with a natural (borrowed) equivalent in Japanese should be used.

[3] SFS is a free software tool created at the phonetics laboratory of the Department of Phonetics and Linguistics, University College London. It can be downloaded at [http://www.phon.ucl.ac.uk/resource/software.html].

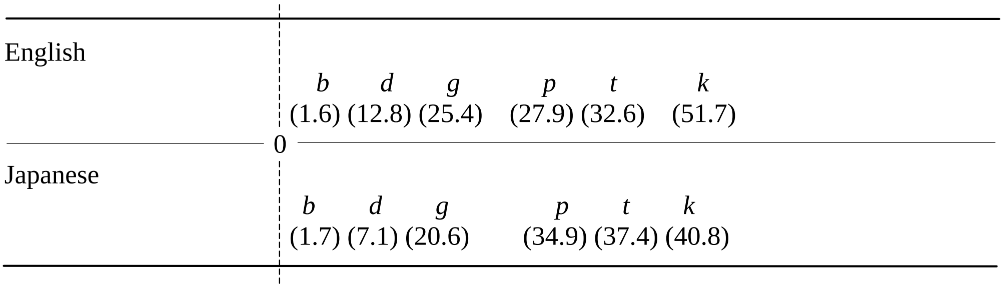

Figure 4. Mean VOT values for English and Japanese stops by L2 learners of English.

In comparison with the mean VOT values of English and Japanese monolinguals in Figure 2, one point to be noted is that the English voiceless stops produced by the L2 learners at this stage show much shorter VOT than those produced by English monolinguals. It is obvious that the L2 learners used the VOT values of their L1 voiceless stops.[4]

As shown in Figure 5, this study clearly suggests that the mean VOT values for English alveolar and velar stops became longer than those collected before the training period, although the values are still shorter than those of English monolinguals.

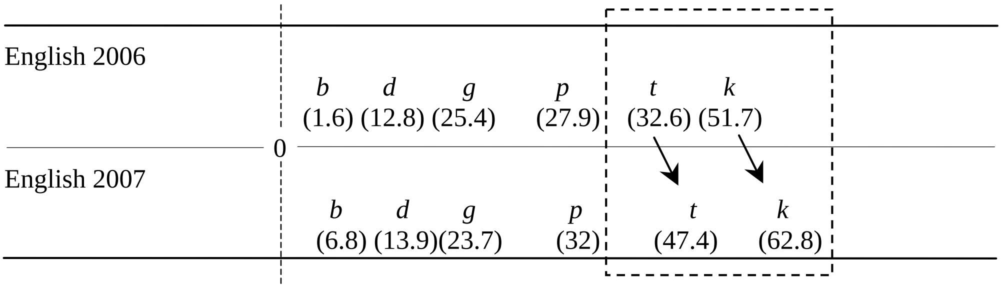

Figure 5. Comparative mean VOT values for English stops by L2 learners of English.

Then, the Wilcoxon signed-rank test, a non-parametric statistical hypothesis test, was conducted to measure the degree of VOT improvement during the training period. This study reveals an interesting finding: there was a *significant* improvement in the VOT value of the English voiceless alveolar stop (Z = 1.992, p = .046) and a *marginally significant* improvement in English voiceless velar stop VOT (Z = 1.69, p = .091); on the other hand, there was no significant improvement in the production of VOT for English voiceless bilabial stops (Z = .943, p = .345 ns).[5] This implies that the VOT of the voiceless bilabial stop is the most difficult to acquire compared with other types of voiceless stop. As we will see, the present discussion assumes that the phonological structure of the sounds in question influences the degree of difficulty in acquiring a native-like VOT.

The next section provides a phonological analysis of the issue. In order to characterize the voiceless stops phonologically, this paper employs a set of features called *elements* (Harris, 2005; Harris & Lindsey, 1995, 2000; Nasukawa & Backley, 2008; et passim).

[4] A further point is that the mean VOT values of Japanese voiced stops are in the neutral region like those of English monolinguals rather than in the region for prevoicing by typical Japanese monolinguals. An explanation for this would take us beyond the scope of the present discussion; briefly, however, it might be suggested that the neutral VOT in /b, d, g/ produced by Japanese subjects may reflect (a) some influence from L2, (b) recent tendencies observed in their regional (Tohoku) accent, or (c) an age-related (early 20s) variable. In order to clarify the motivation for this effect, further investigation is required.

[5] As shown in Figure 5, the mean VOT value for the bilabial stop remained similar to the corresponding sound in Japanese.

5. Element-based melodic theory

5.1. Basic differences between elements and distinctive features

Phonological theories agree that sounds (transcribed by phonemic-alphabetic symbols) are decomposable into smaller units, where these units are regarded as universal primes in melodic representation and are referred to by such terms as *distinctive features*, *elements*, *components*, *gestures* and *particles*. Note my use of the word ‘melodic’ in place of the more traditional term ‘segmental,’ which I avoid on the grounds that it is reminiscent of the now disfavoured notion of ‘segmentation’. This paper employs a set of primes called *elements* (Backley, 1993, 1998; Backley & Nasukawa, 2009a, 2009b; Backley & Takahashi, 1998; Botma, 2004; Brockhaus, 1995; Cabrera-Abreu, 2000; Charette, 1991; Cyran, 1997; Harris, 1990, 1994; Harris & Lindsey, 1995, 2000; Ingleby & Brockhaus, 2000; Kaye, Lowenstamm & Vergnaud, 1985; Kula, 2002; Nasukawa, 1995, 1998, 2005; Nasukawa & Backley, 2005, 2008; Rennison, 1986, 1990, 1998; Scheer, 2004; van de Weijer, 1994; and others). Elements differ from orthodox features (Chomsky & Halle, 1968 and its offshoots) in a number of respects. The most obvious difference is the number of primes involved. Distinctive feature theory typically employs more than twenty, and sometimes over thirty features while Element Theory describes melodic (segmental) structure using a significantly smaller number of elements. For example, the version of Element Theory employed here recognizes a set of six elements |A I U H N ʔ| these are abbreviations for |mass|, |dip|, |rump|, |noise|, |murmur|, and |edge|, respectively (see below for a detailed discussion). It is apparent that an (overly) large set of categories (features) has the capacity to describe a vast set of phonological phenomena, many of which will be unattested. Excessive descriptive power of this kind, and the vast array of unattested feature combinations which may potentially be generated, is now deemed undesirable according to any criterion. By contrast, Element Theory uses a relatively small number of categories (elements), which reduce this excess of descriptive power, and is thus able to explain why certain phenomena are unattested.

All elements are assumed to be present in all languages; their presence in a language is not parametrically controlled. Furthermore, they are strictly phonological in nature since they emerge through the observation of phonological phenomena and form the basis of lexical contrasts: they are identified through the traditional method of modeling how sounds are organized into systems and natural classes. As such, elements may be viewed as mental or internal objects containing linguistic information, which serve to distinguish one morpheme from another. At the same time, however, language serves primarily as a communicative tool, and therefore involves the transfer of linguistically significant information through some physical medium such as speech. It follows therefore that these mental objects must also make some reference to the external world.

Another difference between elements and distinctive features is the way these grammar-internal units refer to the physical world. As their labels indicate, most features have their origins in speech production: for instance [±back] refers to tongue position and [±anterior] to place of articulation. This articulation-based view of features relies on the assumption that melodic structure should be described with a bias towards the speaker: according to this view, phonology is concerned primarily with speech production, not perception (cf. the Motor Theory of Speech Perception: Liberman & Mattingly, 1985; Direct Realist Theory: Fowler, 1986). In contrast, Element Theory rejects the speaker-oriented view in favour of an alternative approach in which phonological objects are associated with properties of the acoustic signal.

This perception-oriented view is supported by evidence from language acquisition and from some widely-attested phonological processes in which language users associate sounds primarily with their acoustic attributes, not with articulation. As discussed in Nasukawa and Backley (2008, p. 4), studies of early language acquisition demonstrate that speech perception apparently exists independently of speech production. It is generally assumed that infants begin to acquire language by first perceiving adult input forms, and then, on the basis of this input they build mental representations which serve as the beginning of their native lexicon; and only later do they start to reproduce these stored forms as spoken language. This implies that speech perception is an indispensable stage on the acquisition path, and that it is a prerequisite for successful acquisition. Besides, early language acquisition, studies of language disorder and impairment also support this perception-oriented view: mutes can acquire a native phonology whereas the profoundly deaf rarely acquire native-like speech production. For all these reasons, the element-based approach takes the view that speech perception is primary and speech production secondary (cf. Jakobson, Fant, & Halle, 1952; Jakobson & Halle, 1956). Accordingly, it employs elements—representational units associated with speech perception—as the link between grammar-internal (cognitive) and grammar-external (physical) properties.

As argued in Nasukawa and Backley (2008, p. 37), Element Theory assumes that listeners naturally seek out linguistic information: when decoding speech, they ignore most of the incoming acoustic stream and focus only on the specifically linguistic information contained within the speech signal. It is recognized that humans have the ability to extract from running speech only those acoustic patterns that are linguistically significant or information-bearing (Harris & Lindsey, 2000). The theory further assumes that the mental phonological categories represented by elements are mapped directly on to those same patterns: elements are recognized as mental constructs in the internalized grammar and they manifest themselves physically as such acoustic patterns. In order to explain these functions of elements, Harris and Lindsey (2000) describe elements as *auditory images*, and define an element as being primarily a mental image of some linguistically significant information, and secondarily a physical pattern in the speech signal which listeners use to cue that mental image. This position makes Element Theory unique in its approach; by comparison, in distinctive feature theories the features themselves are either defined in terms of articulation, as discussed above, or raw acoustics (Chomsky & Halle, 1968; Clements & Hertz, 1991), or otherwise in terms of coexisting articulatory and acoustic specifications (Flemming, 1995).

Another basic difference between elements and distinctive features concerns the way they express phonological oppositions. In Element Theory, contrasts are represented through the presence or absence of a given privative (monovalent) element. For example, ‘nasal’ contrasts are encoded by the presence versus absence of the nasality-related element |N|. In contrast, the competing notion of equipollent (bivalent) oppositions is exploited by orthodox distinctive feature theory, where an opposition is derived by assigning plus and minus values to a given feature. When it comes to predicting phonological behaviour, equipollence yields at least three possibilities: [+nasal] is active in processes; [–nasal] is active; and both [+nasal] and [–nasal] are simultaneously active. However, in the case of nasalisation, only the first prediction is attested; the others fail to be observed in any natural language. To make matters worse, the equipollent format substantially over-generates the number of unattested processes when coupled to a multistratal rule-based system of phonological derivation.

Below, I will outline some further points which are fundamental for understanding the use of elements in melodic representations.

5.2. Fundamentals in Element Theory

It is assumed that, in principle, all six elements |A I U H N ʔ| are employed in all languages. In addition, unlike orthodox distinctive features, any element may occupy any syllabic position. For example, when the element |I| appears alone in a syllable nucleus it is pronounced as the vowel [I], whereas in a consonantal position the same element is pronounced as [j]. As discussed in Nasukawa and Backley (2005, 2008), however, the intrinsic properties of an element influence the likelihood of that element appearing in certain syllabic positions. In general, |A I U| tend to appear in nuclear positions, while |H N ʔ| are typically found in non-nuclear positions. This distributional tendency is reflected in the division of the element set into two groups: |A I U| constitute the ‘resonance’ (‘core’) group, while |H N ʔ| form the ‘edge’ (‘peripheral’) group. Both groups are illustrated below with their physical manifestations (Nasukawa & Backley, 2008, p. 38).

Table 3a. *The Resonance Elements in Nuclei*

Elements		*Typical acoustic correlates*
|A|	mass	central spectral energy mass (convergence of F1 and F2)
|I|	dip	low F1 coupled with high spectral peak (convergence of F2 and F3)
|U|	rump	low spectral peak (convergence of F1 and F2)

Table 3b. *The Edge Elements in Non-Nuclei*

Elements		*Typical acoustic correlates*
|ʔ|	edge	abrupt and sustained drop in overall amplitude
|N|	murmur	broad resonance peak at lower end of the frequency range
|H|	noise	aperiodic energy

Unlike traditional distinctive features such as [+high] or [–son], Element Theory assumes that elements can be interpreted without support from other elements at any level of representation. In nuclear positions, for example, |A|, |I| and |U| are manifest as the low vowel [a], the front vowel [i], and the back rounded vowel [u], respectively, without needing the support of any additional properties from other elements. On the other hand, when |ʔ|, |N| and |H| appear alone in non-nuclear positions, they are pronounced as the glottal stop [ʔ], the uvular nasal [ɴ], and the glottal fricative [h], respectively.

As already mentioned, elements are not tied to particular syllabic positions, but the same element will be subject to a different phonetic manifestation according to the position where it does appear. The resonance elements have the physical values shown in Table 3a when they occupy a nuclear position, while in non-nuclear position they contribute consonant 'place of articulation' properties. As discussed in Nasukawa and Backley (2008), |A| represents the category 'guttural' and is typically interpreted as pharyngeal; |I| represents the category 'coronal' and is typically interpreted as dental; and |U| represents the category 'dorsal' and is interpreted as velar. For a discussion of the phonetic realization of |H N ʔ| in nuclear positions, see Nasukawa and Backley (2005).

Although single elements are pronounceable, most segments are in fact represented by compound expressions containing a combination of elements interpreted simultaneously. When two or more elements are interpreted together, the result is an acoustic signal which is often richer (than an expression consisting of a single element) in linguistic information since it can contain multiple acoustic patterns. For example, the mid vowel compound |I A| is interpreted phonetically as [ɛ], its acoustic profile being a combination of two the patterns shown in Table 3a: |A| contributes 'mass' (central spectral energy mass) and |I| provides 'dip' (F1 coupled with high spectral peak).

In addition to the simple combination of elements, some languages add a further level of complexity to the way in which elements combine. A given compound expression may exhibit an asymmetric relation expressed by the predominance of one of its constituent elements over the other(s). For example, some languages contrast [ɛ] versus [e], which are both represented by the same elements |I A|. In this case, |I| and |A| may enter into a predominance-dependency relation such that, when |I| is dominant over |A|, the whole expression is phonetically mapped onto the close mid vowel [e], since this vowel approximates more closely to the signal pattern for |I| than for |A|. Without any dependency relation, |I| and |A| make equal contributions to the compound and the resulting interpretation is [ɛ]. In contrast, when |A| is dominant over |I|, the expression is interpreted as the open vowel [æ] (For a similar analysis of other elements, see Backley & Nasukawa, 2009a, Nasukawa & Backley, 2005).

Employing the basic Element Theory system outlined above, the three pretonic voiceless stops [p], [t], [k], which are relevant to the present discussion of VOT acquisition, are often described as in Table 4 for English and Japanese (Backley & Nasukawa, 2009a, 2009b; cf. van de Weijer, 1994).

Table 4. *The Representation of Voiceless Stops: Japanese Versus English*

a.	*Japanese*		b.	*English*	
	t	|I, ʔ, H|		*t* (*t*^{h})	|I, ʔ, $\underline{\text{H}}$|
	k	|U, ʔ, H|		*k* (*k*^{h})	|U, ʔ, $\underline{\text{H}}$|
	p	|$\underline{\text{U}}$, ʔ, H|		*p* (*p*^{h})	|$\underline{\text{U}}$, ʔ, $\underline{\text{H}}$|

Note. Underlining indicates that an element predominates in the expression.

As shown above, stops contain the edge element |ʔ| and the noise element |H|: |ʔ| represents the drop in amplitude which is present in the spectral profile of stops and corresponds to a momentary ‘empty’ slice in the spectral profile; while |H| in non-nuclear position represents aperiodic energy and is typically manifest as voicelessness. Focusing on |H|, the role of this element is different between Japanese and English. Japanese is a *voicing language* and thus employs the voiceless unaspirated series for word-initial pretonic stops (Backley & Nasukawa, 2009b; Jessen, 1998; Lombardi, 1994); in this case, |H| simply represents voicelessness. On the other hand, English is an *aspiration language* which features the voiceless aspirated series in word-initial pretonic stops; as such, the element |H| displays its relative prominence in the overall expression. In perceptual terms, the voicelessness (i.e., voicing lag) associated with |H| is more salient: in other words, voicelessness is exaggerated in aspirated stops owning to the presence of |H̲|. This is consistent with our general understanding of expressions containing a predominant element, where the acoustic properties of a predominant element are expected to be stronger and more prominent than when same element is a dependent. In formal terms, this asymmetric relation between elements refers to the prevailing notion of head-dependency. Predominance is often referred to as ‘headedness,’ which is an additional property that interacts with the elements in a given melodic structure and contributes to structural complexity: an expression containing an element which is headed (i.e., its headedness contributes to relative prominence) is more complex than a melodic expression without headedness (Backley & Takahashi, 1998; Nasukawa, 2005a, 2005b, 2005c). Here, Element Theory describes how the English aspirated stops are structurally more complex than their Japanese unaspirated counterparts. This view is supported not only by acoustic profiles but also by phonological phenomena in both English and Japanese. For detailed discussions, see Harris (1994) for English, Nasukawa (2004, 2005) for Japanese, and Backley and Nasukawa (2009b) for both.

As already mentioned, the resonance elements are responsible for encoding ‘place of articulation,’ in which case they may also be headed. The theory makes the following claims concerning the way headedness affects the interpretation of an element (Nasukawa & Backley, 2008, p. 40).

Table 5. *Interpretation of the Headedness of Resonance Elements in Non-Nuclei*

Elements			*Typical acoustic correlates*
|A|	mass	|A|	pharyngeal
		headed |A̲|	epiglottal
|I|	dip	|I|	dental
		headed |I̲|	palatal
|U|	rump	|U|	velar
		headed |U̲|	labial

Owing to positional markedness differences (Nasukawa & Backley, 2008, p. 37), it is expected that |I| and |U| appear more frequently than |A| in non-nuclear positions. And this is cross-linguistically true: in the representation of voiceless stops in both English and Japanese, |I| and |U| are employed, but not |A|. Additionally, both languages allow only |U| to be headed. The headedness of |U| distinguishes velarity from labiality: when |U| is headed it represents bilabial, and when non-headed it represents velar. This unified approach to labiality and velarity under a single element is supported by the strong phonological correlation between these two properties in terms of their diachronic, synchronic and dialectal behaviour (Backley & Nasukawa, 2009a). A similar argument is also found in Jakobson, Fant and Halle (1952), who posit the feature [grave] for labials and velars which indicates a concentration of acoustic energy at the lower end of the spectrum. However, both languages select a parametric setting which disallows |I| to be headed for stops (although headed |I̲| is employed to represent other types of obstruents in both languages). There are languages such as Icelandic and Burera which allow |I| to be headed for stops, which yields the contrast between dental and palatal: dentals are represented by non-headed |I| and palatals by the headed equivalent (cf. van der Hulst, 1989 on the markedness of structural relations among |A|, |I| and |U|).

Given that headed |$\underline{U}$| represents labials, the degree of structural complexity differs between Japanese and English voiceless stops: English voiceless aspirated stops are structurally more complex than Japanese voiceless unaspirated stops since the latter include only a single headed element |$\underline{U}$|, whereas in the English case two headed elements |$\underline{U}$| and |$\underline{H}$| are present.

6. The relation between L1 and L2 phonologies

In order for Japanese L2 learners of English to acquire the phonology of English, they must learn that some structural properties of the English stops differ from those of the Japanese stops. From the discussion given in the previous section, we have two representational differences between English and Japanese ‘voiceless’ stops, one of which concerns a parametric setting controlling headedness on the noise element |H|: English allows headed |H| to contribute ‘aspiration’ or long-lag VOT in English ‘voiceless’ stops, whereas |H| cannot be headed in Japanese. This is depicted in Table 6.

Table 6. *The Parametric Setting of |H|-Headedness in Japanese and English*

	Japanese	*English*
|H| headed?	OFF	ON

So Japanese L1 speakers need to acquire the English ‘ON’ parameter setting for the headedness of the element |H|. If we recall the initial stage of VOT acquisition by Japanese L2 learners of English shown in Figure 4, it may be supposed that subjects simply used their L1 representation, a non-headed |H|, for the English ‘voiceless’ stops. Thus they realized VOT as short-lag or zero. This is apparently different from the English ‘voiceless’ stops, which show long-lag VOT (i.e., aspiration). Then as a result of training, subjects improved their VOT performance as shown in Figure 5, from which it may be concluded that they have acquired the English setting for |H|-headedness at least for the voiceless alveolar and velar stops, since the mean VOT of these stops exhibit a clear improvement.

At this point, we need to address the question of why the voiceless bilabial stop shows no significant improvement in VOT accuracy. On the face of it, the only difference between the Japanese and English voiceless stops is the headedness of |H|—all other aspects of their representations are identical. In this case, some may claim that the differences in improvement between the bilabial and non-bilabial stops should be attributed to the physiological mechanisms of speech production. However, if so, the degree/rate of improvement of the VOT values among three place-types of stops should be the same, even though there intrinsically exist different place-dependent VOT values. It can therefore be assumed that the issue goes beyond physiological matters concerning speech organs and functions; rather, it appears that the phonological structure of the sounds in question influences the degree/rate of improvement in acquiring a native-like VOT.

In order to answer the question, this paper assumes that the notion of double headedness is significant, because, by referring to double headedness in a single melodic expression it is possible to characterize the English bilabial aspirated stop. As Table 4 shows, within (and also beyond) the class of stops, only the English bilabial stop is permitted a double headed structure, which contains |$\underline{U}$| and |$\underline{H}$|. In acoustic terms, this means that not only the ‘dark’ quality of |U| but also the ‘voicelessness’ of |H| exhibit relative prominence in the overall expression. This dual prominence (informally, exaggeration of |$\underline{U}$| and |$\underline{H}$|) results in greater complexity than zero/single prominence. Unlike English, Japanese tends to employ only a single headed element in any consonantal expression, and for vowels, it forbids any vocalic expressions to involve headedness (cf. Backley & Nasukawa, 2009b). As a result, the sound inventory of Japanese exhibits a smaller range of contrastive sounds than we find in English: for example, the Japanese vowel system has five vowels whereas English uses more than twice this number. So, in addition to acquiring the ON setting for |H|-headedness, I assume that Japanese learners of English must acquire two things in order to produce an accurate English VOT for the voiceless bilabial stop: (a) L2 learners, whose L1 grammar does not feature double headedness (e.g., |$\underline{U}$| and |$\underline{H}$| in the same expression), must acquire a structure which allows two elements to be headed; and (b) they also need to learn how the two headed elements are mapped on to the relevant acoustic patterns, which are more prominent than the acoustic profiles of other elements present in the structure. Evidently, the nine-month period of training received by the subjects in this study was sufficient for them to acquire the VOT characteristics of alveolar and velar voiceless stops, but not sufficient for them to acquire the structural characteristics described in (a) and (b) above.

From a pedagogical point of view, Japanese L2 learners of English should first focus on acquiring an appropriate VOT value for the English voiceless bilabial stop, which is phonologically more complex than voiceless stops at other places of articulation. Then, the L2 learners naturally acquire the VOT values for the voiceless alveolar and velar stops, which are phonologically less complex and therefore structurally less marked in English.

7. Conclusion

This paper has discussed a phonological explanation for the link between place of articulation and the success with which Japanese L2 learners of English acquire an accurate VOT production in English stops. The data presented here have shown that Japanese learners of English experience difficulties in acquiring a native-like VOT for the English voiceless bilabial stop in particular. Using an Element Theory approach to melodic representation, this may be explained by appealing to the structural differences between bilabials and other segments—specifically, to the notion of melodic headedness. English bilabial stops are characterized by |H|-headedness and double headedness, neither of which is grammatical in the phonological structure of bilabial stops in Japanese.

This paper is part of longitudinal study of L2 sound acquisition. Therefore further research will be required to validate the arguments given here.

References

Abramson, Arthur S. & Lisker, Leigh. (1970). Discriminability along the voicing continuum: Cross language tests. In Bohuslav Hála, Milan Romportl, & Premysl Janota (Eds.), *Proceedings of the 6th International Congress of Phonetic Sciences* (pp. 569–573). Prague: Academia, Czechoslovak Academy of Sciences.

Benkí, José R. (2001). Place of articulation and first formant transition pattern both affect perception of voicing in English. *Journal of Phonetics*, *29*, 1–22.

Backley, Phillip (1993). Coronal: The undesirable element. *UCL Working Papers in Linguistics*, *5*, 301–323.

Backley, Phillip (1998). *Tier geometry: An explanatory model of vowel structure* (Unpublished doctoral dissertation). University College London, University of London, London, UK.

Backley, Phillip & Takahashi, Toyomi (1998). Element activation. In Eugeniusz Cyran (Ed.), *Structures and interpretation: Studies in phonology* (pp.13–40). Lublin: Folium.

Backley, Phillip & Nasukawa, Kuniya (2009a). Representing labials and velars: A single ‘dark’ element. *Phonological Studies*, *12*, 3–10.

Backley, Phillip & Nasukawa, Kuniya (2009b). Headship as melodic strength. In Kuniya Nasukawa & Phillip Backley (Eds.), *Strength relations in phonology* (pp. 47–77). Berlin: Mouton de Gruyter.

Botma, Bert (2004). *Phonological aspects of nasality: An element-based dependency approach*, Utrecht, The Netherlands: Landelijke Onderzoekschool Taalwetenscap.

Brockhaus, Wiebke (1995). *Final devoicing in the phonology of German*. Tübingen, Germany: Niemeyer.

Cabrera-Abreu, Mercedes (2000). *A phonological model for intonation without low tone*. Bloomington, IN: Indiana University Linguistics Club.

Charette, Monik (1991). *Conditions on phonological government*. Cambridge: Cambridge University Press.

Chomsky, Noam & Halle, Morris (1968). *The sound pattern of English*. New York: Harper and Row.

Clements, George N. & Hertz, Susan R. (1991). Nonlinear phonology and acoustic interpretation. *Proceedings of the XIIth International Congress of Phonetic Sciences* (pp. 364–373). Provence, France: Université de Provence.

Cyran, Eugeniusz (1997). *Resonance elements in phonology: A study in Munster Irish*. Lublin, Poland: Folium.

Fant, Gunnar (1973). *Speech sounds and features*. Cambridge, MA: The MIT Press.

Flemming, Edward S. (1995). *Auditory representations in phonology* (Unpublished doctoral dissertation). University of California, Los Angeles, CA.

Fowler, Carol A. (1986). An event approach to the study of speech perception from a direct-realist perspective. *Journal of Phonetics*, *14*, 3–28.

Harada, Tetsuo (2007). The production of voice onset time (VOT) by English-speaking children in a Japanese immersion program. *International Review of Applied Linguistics in Language Teaching* (*IRAL*), *45*, 353–378.

Harris, John (1990). Segmental complexity and phonological government. *Phonology, 7*, 255–300.

Harris, John (1994). *English sound structure*. Oxford: Blackwell.

Harris, John (2005). Vowel reduction as information loss. In Philip Carr, Jacques Durand & Colin J. Ewen (Eds.), *Headhood, elements, specification and contrastivity* (pp. 119–132). Amsterdam: John Benjamins.

Harris, John & Lindsey, Geoff (1995). The elements of phonological representation. In Jacques Durand, & Francis Katamba (Eds.), *Frontiers of phonology: Atoms, structures, derivations* (pp. 34–79). Harlow, Essex: Longman.

Harris, John & Lindsey, Geoff (2000). Vowel patterns in mind and sound. In Noel Burton-Roberts, Philip Carr, & Gerry Docherty (Eds.), *Phonological knowledge: Conceptual and empirical issues* (pp. 185–205). Oxford: Oxford University Press.

Hulst, Harry van der (1989). Atoms of segmental structure: Components, gestures and dependency. *Phonology, 6*, 253–284.

Ingleby, Michael & Brockhaus, Wiebke (2000). Phonological primes: cues and acoustic signatures. In Jacques Durand & Bernard Laks (Eds.), *Phonetics, phonology and cognition* (pp. 131–150). Oxford: Oxford University Press.

Ito, Junko & Mester, Armin (2003). *Japanese morphophonemics: Markedness and word structure*. Cambridge, Massachusetts: The MIT Press.

Jakobson, Roman, Fant, Gunnar M., & Halle, Morris (1952). *Preliminaries to speech analysis*. Cambridge, Massachusetts: The MIT Press.

Jakobson, Roman & Halle, Morris (1956). *Fundamentals of language*. The Hague, The Netherlands: Mouton.

Jakobson, Roman & Waugh, L. (1979). *The sound shape of language*. Brighton, Sussex: The Harvester Press.

Jessen, Michael (1998). *Phonetics and phonology of tense and lax obstruents in German*. Amsterdam: John Benjamins.

Kaye, Jonathan, Lowenstamm, Jean, & Vergnaud, Jean-Roger (1985). The internal structure of phonological representations: A theory of charm and government. *Phonology yearbook, 2*, 305–328.

Keating, Patricia A. (1984). Phonetic and phonological representation of stop consonant voicing. *Language, 60*, 286–319.

Kent, Ray D. & Read, Charles (1992). *The acoustic analysis of speech*. San Diego, CA: Whurr Publishing Group.

Kula, Nancy C. (2002). *The phonology of verbal derivation in Bemba*. Utrecht: Landelijke Onderzoekschool Taalwetenscap.

Liberman, Alvin M. & Mattingly, Ignatius G. (1985). The motor theory of speech perception revised. *Cognition, 21*, 1–36.

Lisker, Leigh & Abramson, Arthur S. (1964). A cross-language study of voicing in initial stops: Acoustical measurements. *Word, 20*, 384–422.

Lombardi, Linda (1994). *Laryngeal features and laryngeal neutralization*. New York: Garland.

Nasukawa, Kuniya (1995, September). *Melodic structure and no constraint-ranking in Japanese verbal inflexion*. Paper presented at the autumn meeting of the Linguistic Association of Great Britain, University of Essex, Colchester, UK.

Nasukawa, Kuniya (1998). An integrated approach to nasality and voicing. In Eugeniusz Cyran (Ed.). *Structure and interpretation: Studies in phonology* (pp. 205–225). Lublin, Poland: Folium.

Nasukawa, Kuniya (2004). Word-final consonants: arguments against a coda analysis. *Proceedings of the 58th conference of the Tohoku English Literary Society*, 47–53.

Nasukawa, Kuniya (2005a). *A unified approach to nasality and voicing,* Berlin: Mouton de Gruyter.

Nasukawa, Kuniya (2005b). The representation of laryngeal-source contrasts in Japanese. In Jeroen van de Weijer, Kensuke Nanjo & Tetsuo Nishihara (Eds.), *Voicing in Japanese* (pp. 71–87). Berlin: Mouton de Gruyter.

Nasukawa, Kuniya (2005c). Melodic complexity in infant language development. In Marina Tzakosta, Claartje Levelt & Jeroen van de Weijer (Eds.), *Developmental paths in phonological acquisition, Special issue of Leiden Papers in Linguistics, 2.1*, 53–70.

Nasukawa, Kuniya & Backley, Phillip (2005). Dependency relations in Element Theory. In Nancy C. Kula & Jeroen van de Weijer (Eds.), *Proceedings of the Government Phonology Workshop, Leiden Papers in Linguistics, 2.4*, 77–93.

Nasukawa, Kuniya & Backley, Phillip (2008). Affrication as a performance device. *Phonological Studies, 11*, 35–46.

Rennison, John R. (1986). On tridirectional feature systems for vowels. In Jacques Durand (Ed.), *Dependency and non-linear phonology* (pp. 281–303). London: Croom Helm.

Rennison, John R. (1990). On the elements of phonological representations: The evidence from vowel systems and vowel processes. *Folia Linguistica, 24*, 175–244.

Rennison, John R. (1998). Contour segments without subsegmental structures. In Eugeniusz Cyran (Ed.). *Structure and interpretation: Studies in phonology* (pp. 227–245). Lublin, Poland: Folium.

Scheer, Tobias (2004). *A lateral theory of phonology Vol. 1: What is CVCV, and why should it be?* Berlin: Mouton de Gruyter.

Shimizu, Katsumasa (1996). *A cross-language study of voicing contrasts of stop consonants in Asian languages*. Tokyo: Seibido.

van de Weijer, Jeroen M. (1994). *Segmental structure and complex segments*. The Hague, The Netherlands: Holland Academic Graphics.

Selected Proceedings of the 2008 Second Language Research Forum: Exploring SLA Perspectives, Positions, and Practices

edited by Matthew T. Prior,
Yukiko Watanabe, and Sang-Ki Lee

Cascadilla Proceedings Project   Somerville, MA   2010

Copyright information

ISBN 978-1-57473-439-3 library binding

Ordering information

Orders for the library binding edition are handled by Cascadilla Press.
To place an order, go to www.lingref.com or contact:

Cascadilla Press, P.O. Box 440355, Somerville, MA 02144, USA
phone: 1-617-776-2370, fax: 1-617-776-2271, e-mail: sales@cascadilla.com

Web access and citation information

This entire proceedings can also be viewed on the web at www.lingref.com. Each paper has a unique document # which can be added to citations to facilitate access. The document # should not replace the full citation.

This paper can be cited as:

Nasukawa, Kuniya. 2010. Place-Dependent VOT in L2 Acquisition. In *Selected Proceedings of the 2008 Second Language Research Forum*, ed. Matthew T. Prior et al., 197-210. Somerville, MA: Cascadilla Proceedings Project. www.lingref.com, document #2394.