Aspiration in Japanese speakers’ English
A study of the acquisition of new phonetic categories in a second language

Martin Ekelund
Aspiration in Japanese speakers’ English
A study of the acquisition of new phonetic categories in a second language
Martin Ekelund

Abstract
This study aims to explore if it is possible to form separate categories of aspirated voiceless stops in a second language, distinct from the equivalent categories in the native language, for native speakers of a language with an intermediate degree of aspiration, and if such category formation is eased by long-term exposure to another language in which aspirated voiceless stops exist. Two groups of adult native Japanese speakers who had lived in Sweden for a long and short time respectively were recorded when reading a list of sentences containing word-initial, utterance-medial /p t k/ in Japanese and English. Both groups produced higher VOT values for the English stops than for the Japanese stops. The results were significant for /t/ and /k/ and for the long-term residents’ /p/, but not for the short-term residents’ /p/, presumably because of a low number of tokens. The results are nevertheless interpreted as evident of the possibility of phonetic category formation even though there is only a small difference from the corresponding category in the native language. Since both groups had established new phonetic categories, degree of exposure to Swedish is interpreted as likely not to be a relevant factor.

Keywords
Japanese, English, second language acquisition, aspiration, voice onset time, phonetic category, long-term exposure effects, voiceless plosives.
## Contents

1. Introduction ................................................................................... 1  
   1.1 Acquiring the sounds of a second language ................................. 1  
       1.2 Voice onset time in voiceless stops ....................................... 3  
       1.2.1 VOT in English and Swedish ........................................... 4  
       1.2.2 VOT in Japanese ............................................................... 5  
2. Research aims ................................................................................ 5  
   2.1 Acquiring English stops: possibilities ........................................ 5  
   2.2 Research questions ................................................................... 6  
3. Method .......................................................................................... 7  
   3.1 Participants .............................................................................. 7  
   3.2 Material .................................................................................. 8  
   3.3 Data analysis ........................................................................... 8  
       3.3.1 Measurements .................................................................... 8  
       3.3.2 Statistics ........................................................................ 10  
4. Results and analyses ...................................................................... 10  
   4.1 Individual Japanese-English comparison .................................... 10  
   4.2 Grouped Japanese-Japanese comparison ..................................... 12  
   4.3 Grouped English-English comparison ....................................... 12  
   4.4 Grouped Japanese-English comparison ..................................... 13  
5. Discussion ..................................................................................... 13  
   5.1 Individual Japanese-English comparison .................................... 14  
   5.2 Grouped Japanese-Japanese comparison ..................................... 14  
   5.3 Grouped English-English comparison ....................................... 15  
   5.4 Grouped Japanese-English comparison ..................................... 15  
       5.4.1 Additional findings ............................................................ 15  
   5.5 Conclusions ............................................................................ 16  
References ........................................................................................ 17  
Appendix .......................................................................................... 19
1. Introduction

This section covers some of the basic notions of second language acquisition and previous studies of the acquisition of aspirated stops in English, along with the theoretical framework used in this study.

1.1 Acquiring the sounds of a second language

Which sounds are acquired by learners of a second language and which are not has long been a matter for intense study. Selinker (1972, 1992) uses the term interlanguage (IL) to account for systematicity of error in learning a second language (L2). While the notion of systematic errors seems to imply that learners make mistakes in a systematic fashion, error is only half of the systematicity. The other half is an equally important concept: The systematicity of variation. When learning a new language, learners will sometimes produce correct or “target-like” utterances in the L2, and sometimes drift back into production flavored by their native language (L1). Variation in IL is attributed to various factors. For instance, a relaxed emotional state might trigger a drift back into L1 phonological/syntactical patterns (see Selinker, 1972:215-216), which are then transferred to the second language, affecting the L2 performance (production).

Transfer from an L1 to an L2 is perhaps the most intuitively conceivable explanation of foreign accent. For instance, when Swedish native speakers speak English, they are known for substituting /z/ for /s/, since /z/ is not part of the Swedish phoneme inventory. Instead, Swedish native speakers use the perceived closest equivalent: /s/. In this way, features from L1 are transferred into L2, resulting in foreign accent. While this is an easily observed phenomenon, a learner’s interlanguage is not only constructed by transfer from L1. Eckman (1981) found that native speakers of both Spanish and Mandarin adhere to the phonological rule “Terminal Devoicing” in final-position voiced obstruents in their L2 English. Since this rule is not part of the phonology of either Spanish or Mandarin, and not of English either, this devoicing is accounted for by a universal phonological rule. Codas are universally marked, and voiced codas are even more marked than voiceless codas (“markedness” is a relative concept). Universal features of interlanguage are understood as the result of the processes involved in learning a second language being similar regardless of L1 or L2. The notion of universal markedness provides a backdrop to the stage on which learners perform their L2, and along with the language-specific features of the specific languages involved, some general predictions about their L2 (e.g. foreign accent) are possible.

IL has received criticism for a monolingual bias (Kachru 1994, Sridhar 1994) since the framework assumes that the learner aims for nativelikeness in the target language, while many learners of English in the outer and expanding circles do not learn English in a context of native speakers (see Melchers & Shaw (2003) for an account of the inner, outer and expanding circles of English). The notion of target competence is criticized for its focus on deficits and on lack of L2 competence (Sridhar 1994:802). This is probably due to the fact that it is hard to draw a line between when learners use interlanguage and when they have reached their full capacity in the target language. The
explanatory aspects of IL are however still applicable and here, they are to be understood and used as a way of explaining the difference in production between native speakers and non-native speakers.

Flege’s Speech Learning Model (SLM) (1995) suggests that one important factor in the ability to learn L2 segments (i.e. sounds) is the ability to perceive them as distinct sounds. In the example of Swedish substitution of /z/ for /s/, it is often the inability to perceive realizations of /z/ and /s/ as members of two different phonetic categories that leads to the production errors. The members of the phonemic categories /z/ and /s/ in English are perceived as equivalent to the single category /s/ and undergo *equivalence classification*, which means that the learner fails to create a new phonetic category for the L2.

According to the SLM, when learning a second language, learners have to integrate the new phonological system into the L1 system since the two languages occupy the same phonological space in the brain. It is further hypothesized that it is easier to acquire new phonetic categories that are perceptually different from any previously known phones. Sounds that are phonetically similar, “diaphones”, are at higher risk of being subjected to equivalence classification, in which case the L2 learner will use the L1 phonetic category instead of the target L2 category. When this happens, the L1 and L2 diaphones will come to resemble one another in production (1995:239). The likelihood of equivalence classification increases as age of (onset of) learning increases, so adult learners will have greater difficulties perceiving phonetic category differences in an L2 than will children.

When discussing the acquisition of the voiceless aspirated stops /p t k/ in English for native speakers of languages with unaspirated voiceless /p t k/, Flege states that the mechanisms of equivalence classification may limit the accuracy with which L2 phones are produced (1995:258), even though L2 learners may be able to form categories distinct from their corresponding L1 categories: “When category formation is blocked, the phonetic norms of English may be approximated indirectly through a restructuring of the properties specified in a phonetic category used to process perceptually linked L1 and L2 diaphones” (1995:258). Flege and Eefting (1987) found that native Spanish speakers of English produced /p t k/ with voice onset time (see section 1.2 for definition) values at 57 milliseconds, compared to monolingual Spanish speakers’ Spanish /p t k/ at 26 milliseconds and monolingual English speakers’ /p t k/ at 87 milliseconds. The intermediate values are interpreted as “Spanish-like” (1987:78) since they fall between the two groups of native speakers’ categories. Flege and Eefting interpreted this as failure to establish new phonetic categories, since the values were significantly shorter than the native English values. In other words, equivalence classification implies non-nativelike production. Furthermore, another group of Spanish speakers of English who were fully bilingual, had established nativelike /p t k/ categories, but had significantly shorter voice onset time values in their Spanish /p t k/ than had monolingual Spanish speakers. This was interpreted as a result of *polarization*, when the L1 phonetic category is modified to establish a clear contrast between two categories.

Flege assumes nativelikeness to be important when learning a second language. However, the Spanish English “in-between” voice onset time data obtained in Flege and
Eefting (1987) show that the native Spanish speakers were aware of phonetic differences between the two languages. It is arguable that they had created near-nativelike phonetic categories of English /p t k/ with less aspiration than there should be in English, but with enough aspiration to avoid /p t k/ being perceived as /b d g/. In a perception test in the same study, Flege and Eefting found that native English speakers classified /p t k/ as such when the voice onset time was 43 milliseconds or more. Thus, if Spanish English speakers produce /p t k/ at 57 milliseconds, they seem to have become aware of the necessary English-specific category boundaries and produce their L2 within those boundaries. Therefore, the definition of the acquisition of new phonetic categories in a second language differs in this study from the nativelikeness-definition used by Flege and Eefting (1987) and Flege (1995). The assumption is rather that non-native speakers need to establish new categories that are similar enough to L2 norms in order to avoid being misunderstood. Non-native speakers may not need to sound nativelike, and with regard to voice onset time, they may not experience the need to sound nativelike either.

In the light of her study of non-native speakers of English who interact with other non-native speakers, Jenkins (2002) proposes a pragmatic, mutual intelligibility-oriented approach to the key phonological/phonetical features for English learners to focus on. This approach is different from the proficiency goals implied by IL (and to some extent, the SLM) since it focuses on the communicative function of the language rather than the idea of nativelikeness. Nevertheless, she lists aspiration in the word-initial voiceless plosives /p t k/ as a necessary feature to learn, since these phonemes are at risk of being perceived as their voiced counterparts /b d g/ if they are unaspirated (2002:96). Hence, it appears that whether a learner’s goal is nativelikeness or intelligible communication, the aspiration of the voiceless plosives in English is an important feature to master.

When using the L1 phonological system as a template for learning new sounds, some learning tasks are more demanding than others. Japanese learners of English are known for having problems distinguishing English /r/ and /l/, since Japanese has only one liquid, /r/. However, this task is even more difficult than creating a new category, because the Japanese /r/ cannot be used for either of English /r/ or /l/. Hence, they have to split their L1 category into two categories for English. When learning the English aspirated stops /p t k/ from an L1 with unaspirated /p t k/, the properties of the categories need to be redefined for English. The level of difficulty is lower than when the learner needs to split a category, since English has no phonemic contrast between unaspirated and aspirated /p t k/. The main difficulty in learning the English stops /p t k/ is to perceive them as distinct phonetic categories (and then produce them accurately).

1.2 Voice onset time in voiceless stops

Ever since Lisker & Abramson’s (1964) study of the universality of voice onset time (VOT) (the time from the release of the stop to the onset of the vibration of the vocal cords in the following sound) there has been a broad generalization about languages into two groups; short-lag languages and long-lag languages. Short lag means that the VOT is relatively short (0 – 25 milliseconds) for the voiceless stops and long lag means that
the VOT is relatively long, about 60 – 100 milliseconds.\textsuperscript{1} The voice onset time is the most commonly used measurement and definition for aspiration; it was used in Lisker and Abramson (1964, 1967), Flege and Eefting (1987), Cho and Ladefoged (1999), Riney, Takagi, Ota, and Uchida (2007), Helgason and Ringen (2008), and it will be used here as well. Universally, with regard to VOT, there are three possible phonological contrasts to make along the VOT continuum; voiced, voiceless unaspirated, and voiceless aspirated.\textsuperscript{2} Other than the three-way classification, there seems to be a universal tendency for bilabials and alveolars to have shorter VOT than the velars in languages that have those three categories (see Lisker & Abramson (1964) and Cho & Ladefoged (1999) for an account of the possible phonetic reasons for this). If a language has a distinction between aspirated and unaspirated stops, speakers want to keep a safe durational distance between them. Other than this, the variation of the lag in different languages is basically arbitrary.

\subsection*{1.2.1 VOT in English and Swedish}

English and Swedish are long-lag languages. Exact values are hard to provide, since they are partly contextual and idiolectally flavored. Flege and Eefting (1987) reported mean VOT values at 78, 89 and 94 milliseconds for adult monolingual native English speakers’ /p/, /t/ and /k/ in initial position for isolated words. Riney et al. (2007) reported mean VOT values for five monolingual American English speakers at 78.2, 94.8 and 102.8 ms for stressed, word-initial, utterance-medial (samples were obtained using a carrier phrase) /p/, /t/ and /k/ respectively. Lisker & Abramson (1964) reported values at 58, 70 and 80 ms with ranges of 20-120, 30-105 and 50-105 ms respectively for single utterance-initial /p t k/ and 28, 39 and 43 ms in the range of 10-45, 15-70 and 30-85 ms for /p/, /t/ and /k/ respectively in word-initial position in sentences. Lisker & Abramson (1967) reported VOT values at 33, 38 and 48 ms for the three voiceless stops respectively in sentences. There was a small difference between stressed and unstressed syllables in the sentences, while the difference was larger in single word utterances. It is worth noting that English VOT seems to decrease in sentences and that the variation in VOT realization is huge, not only in sentences but also in utterance-initial position. The decrease of VOT in sentences is particularly interesting, since the VOT values in some cases are lower than they would have to be in Flege and Eefting (1987) to be classified as voiceless by native speakers. However, that classification is based on isolated words with no contextual cues. It is quite possible, even probable, that English voiceless stops in sentences do not need to be clearly aspirated because there is much contextual information available. However, while Lisker and Abramson (1967) discuss contextual effects on the sentence values, they do not account for possible aspiration loss due to preceding word coda /s/, nor do they provide the actual sentences used. While VOT is

\textsuperscript{1} However, languages with phonemic aspiration who contrast between voiceless unaspirated and voiceless aspirated stops tend to have higher VOT values for the aspirated stops; see for instance Cho and Ladefoged (1999).

\textsuperscript{2} The observant reader should note that this three-way distinction is based on the VOT continuum and does not account for “aspirated” voiced stops in, for instance, Hindi or Korean.
decreased in sentences, it is not clear exactly how much. For Swedish, Helgason and Ringen (2008) reported means at 49, 65, 78 ms for /p/, /t/, and /k/ respectively in utterance-initial position for six speakers and 14, 23 and 31 ms respectively for word-medial intervocalic position. In both languages, aspiration is lost when the stop is preceded by /s/. Non-initial voiceless stops do not generally have aspiration in either of the two languages. The English VOT means might be slightly higher than the Swedish, but overall, the aspiration patterns of the two languages are similar.

1.2.2 VOT in Japanese

Tsujimura (2007) states that Japanese voiceless stops lack aspiration. Contrarily, Riney et al. (2007) found that Japanese monolinguals produced mean VOT values at 30.0, 28.5 and 56.7 milliseconds for utterance-medial, word-initial /p/, /t/ and /k/ respectively. The data in Riney et al. (2007) finds universal support in Cho & Ladefoged (1999) who present data on several languages where the VOT is somewhere between the short-lag and long-lag categories; hence, both studies discuss the adaptation of a category with an intermediate degree of aspiration. The aspiration patterns for utterance-initial stops in Japanese are unknown, but for utterance-medial, word-initial stressed position, they are assumed to be somewhere between the short-lag high point and the long-lag low point; not aspirated, even though they might be longer than in the average short-lag language. It might be worth noting that the VOT values in Riney et al. (2007) are fairly equivalent to the values for English sentences in Lisker and Abramson (1964, 1967). However, it is possible that the Japanese values, much like the English, decrease in the sentence context.

2. Research aims

This study aims to explore some of the hypotheses in the Speech Learning Model in an attempt to establish where to draw the line between adaptation of new phonetic categories and what Flege would define as equivalence classification. The study tackles Japanese L1 speakers’ production of Japanese and English sentences with regard to the speakers’ rate of exposure to Swedish and their degree of aspiration in the word-initial, utterance-medial voiceless stops /p t k/ in English and Japanese.

2.1 Acquiring English stops: possibilities

According to the Speech Learning Model, in order to acquire nativelike English stops with aspiration, Japanese English learners need to establish new phonetic categories. Flege and Eefting (1987) found that Spanish adult learners of English produced /p t k/ with longer VOT than in their native language, but the VOT values were nevertheless shorter than they should be in English. According to Flege (1995), this is evidence of L2 category formation failure. However, as will be detailed in section 3.3.2, this study will accept statistically significant differences between Japanese and English VOT values as evidence of new category formation. Since the Japanese VOT values are somewhere
between short-lag and long-lag, the main difficulty for Japanese learners of English is assumed to be to perceive the English \( /p\ t\ k/\) as members of different phonetic categories than their Japanese \( /p\ t\ k/\). If the categories have been perceived as similar, it is assumed that there will be no significant difference between their produced \( /p\ t\ k/\) in Japanese and English.

The Japanese L1 speakers should have an additional possibility to perceive and acquire the new phonetic categories of aspirated stops through long time exposure to Swedish. This is based on the assumptions that all phonemes and allophones (L1, L2, L3) occupy the same phonological space in the brain, and that acquisition of new phonetic categories is possible given time and language use. Furthermore, this study assumes that if an aspirated category has been established in Swedish, it can also be perceived and used in English. What is tested here is if the long-term residents in Sweden have a higher degree of aspiration in English than those who have lived in Sweden for a short time. If they do, it could be attributed to the acquisition of new phonetic categories in their L2 English through Swedish. Logically, there are four possibilities:

1. Both groups produce voiceless stops with different VOT values for both Japanese and English (establishment of new phonetic categories regardless of length of stay)
2. Both groups produce voiceless stops with similar VOT values for both Japanese and English (no new category formation)
3. The long-term resident group produce different VOT values for the two languages, but the short-term resident group does not (establishment due to exposure)
4. The short-term resident group produces different VOT values for the two languages, the long-term resident group does not (unknown; possible L3 interference)

If there is a difference between the two groups, it will also be possible to compare the two groups’ Japanese values, to see if category formation has caused polarization of the Japanese and English categories. If so, the Japanese VOT values will be significantly lower for the Japanese speakers who have formed new phonetic categories than for those who have not. Furthermore, it will be possible to compare the two groups’ English VOT values to see if they are significantly higher for the group of speakers that (supposedly) has polarized the two categories, or simply to see if their VOT values for English have increased as a result of long-term exposure to Swedish.

### 2.2 Research questions

The research questions are divided into two categories; general questions and specific questions. The general questions are:

1. Is it possible to acquire categories of aspirated stops in a second language if the first language has an intermediate degree of aspiration in the corresponding categories?
2. Is time of exposure to another language which possesses such categories a factor in attaining them?

3. If new categories have been formed, have they undergone polarization, in which case the Japanese VOT values are lower than they should be?

The general questions are operationalized as the following specific questions:

1. Do Japanese speakers of English produce /p t k/ with significantly higher VOT values in English than they do in Japanese?
2. Do Japanese speakers who have lived in Sweden for a long time produce significantly lower VOT values for Japanese /p t k/ than the Japanese speakers who have lived in Sweden for a short time?
3. Do Japanese speakers who have lived in Sweden for a long time produce significantly higher VOT values for English /p t k/ than the Japanese speakers who have lived in Sweden for a short time?
4. Do Japanese speakers of English who have lived in Sweden for a long time produce /p t k/ with significantly higher VOT values in English than in Japanese, and do Japanese speakers who have lived in Sweden for a short time not produce significantly higher VOT values in English than in Japanese?

3. Method

3.1 Participants

8 Japanese native speakers were recorded for this study. 5 of them were students, two of whom had spent about 9 months in Sweden, one had spent 13 months in Sweden, and the two others had spent 21 months in Sweden. These students are grouped together based on their relatively short time as residents in Sweden and compared to three other participants who had lived in Sweden for more than 20 years. Some of the students were not very proficient in Swedish, while all of the 8 participants had had no further education in English than the course in Japanese school. Since the degree of proficiency in Swedish varied among the speakers, for the purposes of this study, the length of exposure is used as the main divisor between the two groups. The participants reported that they spoke no other languages than Japanese, English and (some of them) Swedish. The two exceptions are one of the long-term residents who had some knowledge of Italian (a short-lag language) and another of the long-term residents who had some very limited knowledge of Dutch (another short-lag language). In other words, none of the participants knew any language from which they could have learned aspirated stops other than English and/or Swedish.
3.2 Material

All participants were instructed to read a list of sentences in Japanese and English respectively (see Appendix) containing words with /p t k/ in word-initial, utterance-medial position. The English sentences were designed to disguise the fact that they primarily were included to gather tokens of stops by mixing the stop-initial words with various other sounds. Special attention was paid to avoiding possible assimilation effects over word boundaries, since stop-initial words could potentially lose their aspiration due to an /s/ in the coda of the preceding word. While stress varies to a greater extent in sentences than in isolated words, the difference between stressed and unstressed syllables in relation to degree of aspiration reported in Lisker and Abramson (1967) was small enough to consider stress variation a marginal factor. Another reason for avoiding isolated words was the unknown properties of utterance-initial stops in Japanese (see section 5.4.1). One reason for not using carrier phrases as in Riney et al. (2007) was that sentences were believed to reveal possible interlanguage variation patterns: In carrier phrases, where the same phrase is repeated over and over, speakers are likely to be aware of the word that changes from phrase to phrase. The sentences were deliberately somewhat complex to test the stability of possible L2 category formation and reveal interlanguage variation; for instance, the sentence “he had spent so much money, he couldn’t afford to bring a pen to the test” contains several instances of word-initial voiceless stops, but when reading it, participants would (hopefully) not pay attention to every phonetic detail. The idea was to capture how the speakers actually speak the two languages.

The Japanese sentences did not have any design element, because of the author’s limited knowledge of Japanese. Sentences containing the appropriate sounds were instead acquired from the Denshi Jisho online Japanese dictionary (2010) by searching for example sentences containing words containing the appropriate sounds in the appropriate position, then assembled to form a list. As a result of this method, the Japanese sentences were comparatively shorter than the English sentences. However, the study does not aim to test the stability of Japanese /p t k/, and with regard to contextual factors, the sentences are still sentences and therefore comparable with the English sentences.

The readings of sentences were recorded onto a laptop computer using the recording freeware Audacity, through the external sound device Audio Kontrol 1 using a swan-neck microphone featured in the KORG MicroKORG XL Synthesizer package. The recordings were digitized at 44100 Hz with 32-bit resolution. Sound files (.wav) were then analyzed using the freeware Praat.

3.3 Data analysis

3.3.1 Measurements

In Riney et al. (2007), the VOT is defined by measuring from the start of the release to the zero crossing of the first periodical wave. This is interpreted to refer to the point where the first periodical wave crosses the zero axis. The measurements in this study might differ slightly from that measurement and closer approximate those in Cho &
Ladefoged (1999), for two reasons. Firstly, the start of the release is not always easy to define. In this study, for a release to count as a release, it has to be followed by some kind of turbulence (see image 1). Secondly, the periodical wave cannot always be said to start by crossing zero. In most cases, it is possible to follow such a definition, but sometimes the periodical wave starts from a point above zero (see image 2). For this reason, the measurements are made by viewing both the waveform and the spectrogram.

Image 1. Measurement of a stop with multiple releases. The first release is not followed by turbulence and therefore ignored (the marked area in the middle is measured for a VOT value).

The noise following the release of the plosive consist of two parts: the burst and the aspiration. The acoustic characteristics of the aspiration will be influenced by the following vowel (Ladefoged, 2003), which is why such a division is possible to analyze acoustically. However, both phases occur after the release, so for these VOT measurements, the two phases are not distinguished from one another.

Image 2. The periodical wave starts from a point above the zero crossing. The middle field marks the start and endpoint of the VOT measurement.
These measurement choices are partly motivated by convention and partly by ease. It would simply be harder to draw a clear line between the end of the burst and the start of the aspiration than it is to distinguish a release from silence.

3.3.2 Statistics

This study uses statistical significance calculations to determine whether the speakers use different phonetic categories for their Japanese and English /p t k/ or not. Specifically, two different kinds of t-tests are used, because of the low number of samples (see below). To calculate the standard deviations and the t-value, the formulae in Butler (1985:85, 91) were used. The variables in the formulae (squared sums, sum of squares, degrees of freedom etc.) were calculated individually using simple arithmetics functions in Excel. From the t-value, the p-value was read from the table in Butler (1985:172). The p-value is equal to the probability that the null hypothesis is correct, so if \( p < 0.05 \), there is less than a five per cent probability for the null hypothesis to be true, in which case the results are significant. This significance level is commonly used, and it will be used in this study as well.\(^4\)

4. Results and analyses

In this section, the results of the significance tests are presented, along with brief analyses of the results. The tests are presented separately and preceded by brief explanations.

4.1 Individual Japanese-English comparison

By pairing comparable Japanese and English tokens, it is possible to see if the speakers use similar phones differently in the two languages. For instance, Japanese teni and kare are paired with English test and can respectively, since they have comparable onset consonants and roughly similar following vowels. Because of a low number of tokens for the individual phonemes in the Japanese data (see discussion below for the reasons for this), /p t k/ are grouped together when comparing the mean VOT in Japanese and English for every individual separately. In total, there are 13 pairs: 4 for /p/, 4 for /t/ and 5 for /k/. Since there were more English than Japanese tokens of /p/, /t/ and /k/ available, some of the tokens from the English data need be excluded. The tokens are included or excluded via two criteria: Rough equivalence and order of appearance. The rough equivalence means that the following vowels should be comparable (e.g. the /o/

\(^3\) The null hypothesis is typically a hypothesis stating that there will be no difference between samples, so if there is a five per cent probability that the null hypothesis is true, there is a ninety-five per cent probability that it is not and that the compared samples are samples of two different populations (there is a difference).

\(^4\) Although it is not systematically calculated, the difference between the variances of the samples in the tests are in many cases significant, which weakens the validity of the calculations. In practice, we might need to require \( p \)-values at \(< 0.02 \) or \(< 0.01 \) to balance out the unequal variance.
in *tori* and the /o/ in *tore*). For /t/ specifically, some tokens with /o/ were absent and replaced with /u/ (in Japanese *to* and English *to* respectively) since they are both back vowels. The order of appearance refers to the appearance in the recorded data (the order was the same for all participants). Since the Japanese tokens were fewer, the English counterparts were chosen to match the Japanese (and not vice versa). A one-tailed paired *t*-test for significance compares the Japanese data with the English for every individual separately to see if there is a significant increase in VOT from Japanese to English. The results are shown in table 1.

### Table 1. The participants are represented by an initial and a number corresponding to their time (years:months) in Sweden. The mean VOT values, standard deviation, range and N (number of samples) are shown for every individual’s Japanese and English stops respectively, while the values for *t* and *p* are shown separately (since those values are based on the difference between the two language samples).

<table>
<thead>
<tr>
<th>Speaker</th>
<th>mean VOT (ms)</th>
<th>standard deviation (ms)</th>
<th>range (ms)</th>
<th>N</th>
<th>t-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>S20&lt;</td>
<td>Japanese 45.6</td>
<td>17.4</td>
<td>24.7:88.9</td>
<td>13</td>
<td>-3.737</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td></td>
<td>English 66.4</td>
<td>19.7</td>
<td>21.4:87.7</td>
<td>13</td>
<td></td>
<td></td>
</tr>
<tr>
<td>M20&lt;</td>
<td>Japanese 29.3</td>
<td>14.8</td>
<td>12.6:53.5</td>
<td>13</td>
<td>-4.343</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td></td>
<td>English 57.8</td>
<td>28</td>
<td>22.3:124.1</td>
<td>13</td>
<td></td>
<td></td>
</tr>
<tr>
<td>H20&lt;</td>
<td>Japanese 28.4</td>
<td>11.3</td>
<td>13.5:48.2</td>
<td>13</td>
<td>-2.69</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td></td>
<td>English 44.2</td>
<td>20.5</td>
<td>14.2:86.5</td>
<td>13</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T1:9</td>
<td>Japanese 24.8</td>
<td>12.4</td>
<td>9.5:45.8</td>
<td>13</td>
<td>-5.518</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td></td>
<td>English 66.3</td>
<td>26.4</td>
<td>30.0:116.5</td>
<td>13</td>
<td></td>
<td></td>
</tr>
<tr>
<td>M1:9</td>
<td>Japanese 33.8</td>
<td>20.1</td>
<td>12.2:70.6</td>
<td>13</td>
<td>-1.498</td>
<td>&gt;0.05</td>
</tr>
<tr>
<td></td>
<td>English 41.4</td>
<td>18.3</td>
<td>15.2:77.6</td>
<td>13</td>
<td></td>
<td></td>
</tr>
<tr>
<td>K1:1</td>
<td>Japanese 23.2</td>
<td>10.8</td>
<td>10.8:39.7</td>
<td>13</td>
<td>-5.694</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td></td>
<td>English 53.1</td>
<td>21.5</td>
<td>32.2:94.7</td>
<td>13</td>
<td></td>
<td></td>
</tr>
<tr>
<td>K0:9</td>
<td>Japanese 25.7</td>
<td>13.3</td>
<td>10.1:49.4</td>
<td>13</td>
<td>-3.558</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td></td>
<td>English 41</td>
<td>17.2</td>
<td>16.4:71.8</td>
<td>13</td>
<td></td>
<td></td>
</tr>
<tr>
<td>M0:9</td>
<td>Japanese 23.5</td>
<td>10.3</td>
<td>7.8:44.0</td>
<td>13</td>
<td>-3.078</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td></td>
<td>English 33.2</td>
<td>11.9</td>
<td>14.5:52.5</td>
<td>13</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The paired *t*-test shows that for 7 of the speakers, there is a probability of more than 99% that the Japanese sample and the English sample are drawn from two different populations. The difference between Japanese stops and their English counterparts was highly significant for 7 of the 8 participants and not significant for one speaker, M1:9 (*p* > 0.05). For some of the speakers (M20<, T1:9 and K1:1), the categorical division is as expected in a contrastive sense: an unaspirated or mildly aspirated category for the Japanese stops and one clearly aspirated category for the English stops. Interestingly enough, the difference between the Japanese and the English means is rather small for some speakers (H20<, K0:9 and M0:9). The high Japanese mean VOT value for S20< could possibly be explained by influence from the English category, but the categories are nevertheless kept separate. The difference between the stops produced by M1:9
were not significant, but the difference would likely have been significant with a larger sample.

4.2 Grouped Japanese-Japanese comparison

In order to see if there was a significant difference between the long-term residents’ (group 1) and the short-term residents’ (group 2) Japanese VOT values, a non-directional t-test was performed. Since the number of tokens was much greater than in the paired t-test (42 Japanese tokens for group 1, 70 Japanese tokens for group 2) it was possible to test the three categories /p t k/ separately. The results are presented in table 2.

Table 2. Values from the Japanese data. Group 1 consists of three speakers who had lived in Sweden for more than 20 years, group 2 consists of five speakers who had lived in Sweden for 9-21 months.

<table>
<thead>
<tr>
<th>Japanese</th>
<th>Group 1</th>
<th>Group 2</th>
<th>Group 1</th>
<th>Group 2</th>
<th>Group 1</th>
<th>Group 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>/p/</td>
<td>/p/</td>
<td>/t/</td>
<td>/t/</td>
<td>/k/</td>
<td>/k/</td>
</tr>
<tr>
<td>mean VOT (ms)</td>
<td>30,8</td>
<td>39,1</td>
<td>32,2</td>
<td>21,6</td>
<td>42,6</td>
<td>33,9</td>
</tr>
<tr>
<td>standard deviation (ms)</td>
<td>14,9</td>
<td>14,1</td>
<td>17</td>
<td>12,2</td>
<td>18,3</td>
<td>13,3</td>
</tr>
<tr>
<td>number of samples</td>
<td>12</td>
<td>20</td>
<td>15</td>
<td>25</td>
<td>15</td>
<td>25</td>
</tr>
<tr>
<td>t-value</td>
<td>-1,661</td>
<td>-2,311</td>
<td>1,731</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p-value</td>
<td>&gt; 0,05</td>
<td>&lt; 0,05</td>
<td>&gt; 0,05</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The results show that the long-term residents in Sweden (group 1) produced Japanese /t/ with longer VOT than did the short time residents. The difference was significant (p < 0,05). However, if the higher mean VOT is to be attributed to their residence in Sweden, we should expect a similar pattern in the other two stop categories. Contrarily, we find that the short-term residents (group 2) produced higher values for Japanese /p/ than did group 1, although the difference was not significant (p > 0,05). Group 1 did produce higher values for /k/ than did group 2, but the difference was not significant.

4.3 Grouped English-English comparison

Similarly to the Japanese data, the grouped English data yielded more tokens than the individual data: 93 tokens for group 1 and 155 tokens for group 2. The results presented in table 3 show the mean VOT differences between the two groups of Japanese speakers’ English /p/, /t/ and /k/ separately. A non-directional t-test for significance was carried out comparing their English mean VOT values.

The results presented in table 3 show that there is no significant VOT mean difference between the two groups’ English /p t k/. Since 7 of the 8 speakers had formed new categories (as shown in table 1), this finding is expected.
Table 3. Values from the English data. Group 1 consists of three speakers who had lived in Sweden for more than 20 years, group 2 consists of five speakers who had lived in Sweden for 9-21 months.

<table>
<thead>
<tr>
<th>Group</th>
<th>/p/</th>
<th>/t/</th>
<th>/k/</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Group 1</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>English</td>
<td>56.3</td>
<td>54.4</td>
<td>55.4</td>
</tr>
<tr>
<td>mean VOT (ms)</td>
<td>33.2</td>
<td>17.2</td>
<td>17.2</td>
</tr>
<tr>
<td>standard deviation (ms)</td>
<td>25.0</td>
<td>24.1</td>
<td>22.0</td>
</tr>
<tr>
<td>number of samples</td>
<td>27</td>
<td>33</td>
<td>33</td>
</tr>
<tr>
<td>t-value</td>
<td>1.895</td>
<td>0.994</td>
<td>0.615</td>
</tr>
<tr>
<td>p-value</td>
<td>&gt; 0.05</td>
<td>&gt; 0.2</td>
<td>&gt; 0.2</td>
</tr>
</tbody>
</table>

4.4 Grouped Japanese-English comparison

The grouped data used in 4.2 and 4.3 was also used for a directional t-test of the two groups’ three stop categories separately comparing the VOT means in Japanese and English; the results are shown in table 4.

Table 4. Mean VOT values in milliseconds for group 1 (long-term residents in Sweden) and group 2 (short-term residents) for Japanese and English /p t k/ comparing the groups’ respective difference between the Japanese and English stops.

<table>
<thead>
<tr>
<th>Group</th>
<th>/p/</th>
<th>/t/</th>
<th>/k/</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Group 1</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Japanese</td>
<td>30.4</td>
<td>42.6</td>
<td></td>
</tr>
<tr>
<td>English</td>
<td>56.3</td>
<td>54.4</td>
<td></td>
</tr>
<tr>
<td>t-value</td>
<td>-2.583</td>
<td>-2.359</td>
<td></td>
</tr>
<tr>
<td>p-value</td>
<td>&lt; 0.01</td>
<td>&lt; 0.025</td>
<td></td>
</tr>
<tr>
<td><strong>Group 2</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Japanese</td>
<td>39.1</td>
<td>33.9</td>
<td></td>
</tr>
<tr>
<td>English</td>
<td>43.2</td>
<td>52.6</td>
<td></td>
</tr>
<tr>
<td>t-value</td>
<td>-0.672</td>
<td>-3.932</td>
<td></td>
</tr>
<tr>
<td>p-value</td>
<td>&gt; 0.1</td>
<td>&lt; 0.005</td>
<td></td>
</tr>
</tbody>
</table>

5. Discussion

The results are interpreted as to answer the specific research questions in the following way: All native Japanese speakers except one did produce significantly higher VOT values in English than in Japanese. There was no significant difference between the two groups’ Japanese VOT values except for /t/, which the long-term residents produced with significantly higher VOT mean values than did the short-term resident group. There was no significant difference between the two groups’ English VOT values. Finally, both groups produced significantly higher VOT values in English than in Japanese for /t/ and /k/, but the higher values for the short-term residents’ /p/ were not significant. The following discussion deals with the interpretation of the results of the different tests separately.
5.1 Individual Japanese-English comparison

In this study, as stated in section 1.1, statistically significant difference between Japanese and English stops is accepted as evidence of new category formation for /p t k/ in English. Since 7 of the 8 speakers had such significant difference, it can be concluded that they had formed new phonetic categories. However, if we assume that all non-nativelike production is a consequence of equivalence classification, we would need to find out whether the VOT values are nativelike or not. Unfortunately, we do not have data on English native speakers reading the same sentences, so it is not possible to make a direct comparison. However, we can make an estimate based on the VOT values obtained in Lisker and Abramson (1964, 1967). According to those two studies, the mean VOT should be at 28-33 ms for /p/, 38-39 ms for /t/ and 43-48 ms for /k/ and at about 40 ms for a combined VOT value, considering that we have one more token of /k/. Only H20<, M1:9, K0:9 and possibly M0:9 are really close to this value. Four of the speakers, S20<, M20<, T1:9 and K1:1 had higher values. It is therefore possible that four of the speakers had established new phonetic categories with the consequence that their VOT values are too long for English. T1:9 had comparatively low VOT values in Japanese, but high values in English. This is quite possibly a consequence of polarization; in order to maintain a contrast, the properties of the L1 category are redefined, as in Flege and Eefting (1987). However, it should be emphasized that the sentence data in this study and the sentence data in Lisker and Abramson (1964, 1967) are likely not to be entirely comparable.

The high standard deviation and range show that VOT varies considerably within the samples. One might be tempted to attribute this variation to interlanguage (which predicts variation in targetlikeness), but for most speakers, the range of VOT values correspond fairly well to those obtained from native speakers of English in Lisker & Abramson (1964) where the /p t k/ VOT values in sentences were between 10 and 85 milliseconds. Only one speaker, M0:9, strayed radically from that range and produced English stops in the range of 14.5-52.5 milliseconds. It is not clear whether this speaker had established an aspirated nativelike category, even though the speaker had a highly significant difference between the Japanese and the English VOT values.

5.2 Grouped Japanese-Japanese comparison

Since 7 of the 8 participants can be said to have established new phonetic categories, it is not possible to compare the Japanese VOT values for those who have established new categories with those who have not. It was expected that the exposure to Swedish would divide the two groups, and since it did not, the comparison, and hence further discussion of polarization effects, is rendered impossible. The significant difference between the two groups’ /t/ is possibly related to long-term exposure effects, but since there was no significant difference between the two groups’ /p/ or /k/, it cannot provide conclusive evidence for long-term L3 exposure effects. Furthermore, since we have no data on the participants’ VOT in Japanese stops before learning English or Swedish, all claims of L2 or L3 effects on L1 are dismissed. However, The Japanese VOT values are nevertheless expected. In Riney et al. (2007) the VOT values were slightly higher overall, but that study utilized a carrier phrase while this study used sentences. It is,
however, possible that this is a non-factor, since the data on the short-lag languages in Lisker & Abramson (1964) showed that the difference between isolated word initial aspiration and word-initial aspiration in sentences was small if even existing. The Japanese data in this study confirms to some extent the findings in Riney et al. (2007) that Japanese is a language with an intermediate degree of aspiration. However, the Japanese speakers in this study were not monolingual, so the results are not entirely comparable.

5.3 Grouped English-English comparison
As with the Japanese data, there was expected to be a difference between long-term residents’ and short-term residents’ English, which turned out not to be the case. Since no significant difference could be observed between the two groups’ English data, long-term exposure to Swedish can be said not to influence the VOT values in these speakers’ English /p t k/.

5.4 Grouped Japanese-English comparison
The results of the test comparing the two groups’ Japanese and English data are fairly consistent with the individual data in table 1. The results were highly significant for the long-term residents /p t k/ and for the short-term residents /t/ and /k/, but not for their /p/. One might want to attribute this to failure of category formation, but a careful look at the data reveals an alternative explanation.

5.4.1 Additional findings
When the sentences used in this study were designed, it was believed that there was no difference between utterance-initial and utterance-medial voiceless stops, since these values were similar in the English data in Flege and Eefting (1987) and Riney et al. (2007) (see section 1.2.1). However, an early observation that the utterance-initial word-initial stops appeared to have consistently higher VOT values than the utterance-medial word-initial stops lead to the exclusion of utterance-initial stops, since they could be part of a distinct allophonic category in Japanese. Unfortunately, this led to very low numbers of tokens per individual.

In the Japanese data, there were only four tokens of /p/ for every speaker; two words beginning with po and two words beginning with pe. A careful study of the data from group 2 reveals that three of the initial /p/ tokens had ranges of 7,8-30,8 milliseconds (15 tokens in total) and the fourth token had the range of 39,7-50,5 milliseconds (5 tokens in total). While all tokens were utterance-medial, the tokens with higher VOT values were preceded by a short pause. This confirms and extends the possibility of allophonic variation in Japanese between utterance-initial/post-pausal and non-utterance-initial/non-post-pausal voiceless stops. In order to establish whether there is such a difference or not, a t-test was made comparing the Japanese utterance-medial and utterance-initial /k/ of the five speakers in group 2: 8 tokens each (40 in total) of utterance-initial /k/ and 5 tokens each (25 in total) of utterance-medial /k/. The mean
VOT was longer for utterance-initial /k/ than for utterance-medial /k/: 43.5 ms vs. 33.9 ms. The difference was highly significant (t=3.256, p < 0.01). The post-pausal /p/ may have been treated as a member of a more aspirated allophonic category. This could explain why there was no significant difference between group 2’s Japanese and English /p/: One fourth of the Japanese tokens were drawn from a more aspirated allophonic category in Japanese. This means that the mean VOT values for the Japanese stops in the other tests are slightly higher than they would be if this factor would have been controlled for. The low number of samples means that the tests are sensitive to this kind of unforeseen systematicity.

The new finding that the post-pausal/utterance initial /k/ (and probably /t/ and /p/) belongs to a more aspirated category in Japanese raises the question: Have the native Japanese speakers transferred this category into their English and redefined the allophonic properties to encompass all word-initial stops in English? To answer this question, an additional t-test was performed, comparing the utterance-initial /k/ (40 tokens) with the /k/ in word-initial English (55 tokens) for the five speakers in group 2. The English stops had higher mean VOT values: 52.6 ms vs. 43.5 ms. The difference was significant (t=2.444, p < 0.02) for a non-directional test. The English category of /k/ is not constructed by transfer from the Japanese utterance-initial/post-pausal category, since the Japanese category has significantly lower VOT values.

5.5 Conclusions

Given the definitions specified for this study, it is possible to conclude that the native Japanese speakers had acquired new phonetic categories in their English. The findings in this study show that it is possible not only to acquire new phonetic categories, but also how to use them in the sentence context. The participants in this study seem to have at some point become aware of the phonetic differences between Japanese and English /p t k/, established the new categories, and then learned to use them with less aspiration allophonically in sentences.

Since aspiration in itself varies, it is not possible to separate interlanguage variation patterns from “regular” variation patterns in this study. One would have to compare the sentence-context stop realizations of native English speakers and speakers of English as a second language to reveal such patterns.

Since exposure to Swedish turned out to be likely to be a non-factor in acquiring the new categories, but the speakers’ Swedish VOT values were not tested, it is possible that they had formed separate categories for Swedish after establishing the categories for English. To conclude, these are the findings:

1. It is possible to establish an aspirated set of voiceless stops in an L2 if there is a category of intermediately aspirated counterparts in the L1.
2. It is possible to establish such categories after 9 months in Sweden, or time in Sweden is not a factor at all (the categories could have been established beforehand).
3. A small durational distance between the L1 and L2 VOT norms does not necessarily mean that the difference is difficult to perceive and produce.
The Speech Learning Model tries to account for why nativelike category formation fails. Even though this study does not share Flege’s definition of failure, it might provide some understanding of the reasons why L2 English /p t k/ might be produced with lower VOT values than they should in an experimental setting. Flege and Eefting (1987) provided evidence that it is enough to have VOT values at 43 ms to signal to the listener that the stop just produced is part of the voiceless category, not the voiced. Lisker and Abramson (1964, 1967) provided evidence that this is what native speakers do in sentences; they reduce aspiration to the minimum amount needed to be understood. The communicative requirements specified for English /p t k/ by Jenkins (2002) can therefore be interpreted as having enough VOT not to be misunderstood, not that L2 English needs to have /p t k/ VOT values at 80-100 ms. Flege and Eefting (1987) discuss inaccurate L2 input as a possible explanation for the Spanish English speakers’ VOT values at 57 ms. However, if the VOT in English is lowered in sentences, it is quite possible that the heavily aspirated tokens obtained in experiments are rare in natural conversation. One would need to analyze the aspiration patterns of native English /p t k/ in a more conversation-like setting than the experiments with isolated words, carrier phrases or even lists of sentences can provide, and that is beyond the scope of this study.

References


The following sentences are the ones that the participants were asked to read. For some of the Japanese sentences, a square-bracketed romanized version has been added afterwards. The romanization does not necessarily follow romanization conventions: Particles are usually preceded by "-" and some of the word boundaries may not be completely accurate. In the data analysis, only words whose initial boundaries were known were used. Some of the sentences do not include words with initial /p t k/ and there are English sentences with words beginning with various consonant clusters. This is because the sentences were designed to be used in a different project as well.

1. スペインの雨は主に平原に降る。 [supein no ame wa omoni heigen ni huru.]
2. スペインの人口は日本の約三分の一である。 
3. スポーツだけでなく、音楽も好きだ。 
4. 私の好きなスポーツはスキーです。 
5. 私の兄はスポーツはやりません。 
6. ポスト争いは厳しい。 
7. 誘ってもらって嬉しいけど、この次にしておくよ。 [sasotte mo ratte ureshii kedo, kono tsugi nishite oku yo.]
8. 天気は上々だ。 
9. 天気予報が当たった。 
10. 明日、晴れるといいんだけど。 
11. 彼はスペインは言うまでもなく、ポルトガルへも行ったことがある。 [kare wa supein wa yumade mo naku, porutogaru he mo itta koto ga aru.]
12. 七面鳥は鳥より少し大きい。 
13. 彼は私を助けてました。 
14. 厳しい資金不足で計画はストップしている。 
15. 戸には新しくペンキが塗ってあった。 
16. 彼はペンをとり上げて書き始めた。 [kare wa pen o torianite kaki hajimeta.]
17. その町の中心にすてきな公園がある。
18. その大チェーンストアは新しく支店を作った。
19. 彼は助手を非常に信頼している。
20. 少し疲れた。
21. 彼は手にペンを持っていた。
22. このケーキはすてき！
23. ほかほかと暖かい。
24. この夏は雨が少しは降った。
25. これは日本製のストップウォッチ。
26. 私は助手を探している。
27. 私の夢はスペースシャトルで旅することです。
28. 戻ってくると彼は、「すてきな旅でした」と言った。