The cross-language acquisition of stops differing in VOT: Historical overview

James Emil Fleget
Professor Emeritus
School of Health Professions
University of Alabama at Birmingham

SICSS 2017: 2017 Seoul International Conference on Speech Sciences

10–11 November 2017
Seoul National University, Korea

aSupported by the National Institute for Deafness & Other Communicative Disorders (USA)

bContact: jeflege@uab.edu  website: www.jimflege.com
In his *Outline of English Phonetics (1922)*, Daniel Jones observed that whereas

- speakers of languages like Danish tend to produce English /p t k/ with *too much* aspiration

- speakers of Romance languages such as French produce these stops with *too little* aspiration

Today we would refer to these differences in terms of voice onset time (VOT)

*This talk can be downloaded from: [http://www.jimflege.com/conferences](http://www.jimflege.com/conferences)*
Professor Jones had little interest in measuring VOT because, in his view

“Experimental phonetics is a highly interesting study in itself, but it must not be regarded as … indispensable … for those who wish to learn to pronounce a foreign language correctly” (1922, p. iv)

Perhaps he was right

However, as I see it, experimental work is essential for an understanding of how new forms of speech are acquired naturalistically, that is, outside the classroom and without instruction
VOT: Historical overview

This talk will provide introductory information regarding the voice onset time (VOT) dimension.

In my keynote address tomorrow I will present research examining the production and perception of /p t k/ and /b d g/ in a second language (L2).

My hope is that this overview will provide a better general understanding of L2 speech learning.
As a preview of things to come: The research I’ll be presenting has convinced me that L2 *input* is a more important determinant of eventual success than is the age at which L2 learning begins.

Above: *Eliza Doolittle receiving practical phonetic training from Professor Henry Higgins in the film “My Fair Lady”*
Undergraduate students often learn about the VOT dimension, which has been popular in L2 research due, at least in part, to the fact that the dimension is easily measured. However, it is also easy to make mistakes when examining VOT.
Consider the comparison of two groups of university students in the Netherlands carried out by Flege & Eefting (1987)

As expected, most of those who studied English in Utrecht had a better overall pronunciation of English than those who studied Engineering in Delft
Introduction

However, some of the “English majors” in Utrecht seem to have figured out the aim of our research and began to exaggerate VOT.

Alas, there is no way to distinguish those who “exaggerated” from those who spoke “normally” and so these data are essentially useless due to an avoidable error in elicitation.
Introduction

The research I will consider tomorrow focuses on voiced (/b d g/) and voiceless (/p t k/) stop consonants occurring in word-initial, pre-stressed position. For example:

Paul ate carrots and peas

\[
\begin{array}{ccc}
\text{p} & \text{h} & \text{h} \\
\text{p} & \text{h} & \text{h} \\
\end{array}
\]

**Why this focus?**

In post-stress position VOT is less prominent and more variable

In word-final position dimensions other than VOT (e.g. preceding vowel duration, stop closure duration, F1 offset frequency) are the crucial determinants of stop voicing
What is VOT?

VOT varies according to the timing of laryngeal and supra-laryngeal gestures. It is defined as: the interval of time between the release of stop closure and the onset of glottal pulsing (voicing) in the larynx.

For /b d g/: the vocal folds must be brought together (adducted) at midline for glottal pulsing (voicing) to occur;

For /p t k/: a very precise abduction-adduction gesture is needed to rapidly open and then re-close the vocal folds at midline.
What is VOT?

Daniel Jones referred to the perceptual effect of VOT differences as differences in aspiration.

Arthur Abramson and Leigh Lisker, working at Haskins Laboratories in the USA, brought the VOT dimension to the attention of a wide range of researchers in phonetics and allied disciplines in the 1960s.

Professors Abramson (left) and Lisker (right) at a meeting of the Acoustical Society of America in 2004 (Source: Haskins Laboratories)
What is VOT?

Use of the VOT dimension has:

• helped systematize our understanding of phonetic differences between languages;

• inspired hundreds of published studies (see, e.g., Abramson & Whalen, 2017)

• provided a serviceable “bridge” between speech production and perception

  VOT functions both as an important acoustic phonetic dimension to be regulated in speech production and as an important auditory cue for the perceptual identification of stop consonants;
Early work (e.g., Lisker & Abramson, 1964) called attention to the existence of three “modal” VOT ranges in human languages.

<table>
<thead>
<tr>
<th>Modal VOT category</th>
<th>Acoustic description</th>
<th>Phonetic name</th>
<th>Phonological status</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lead</td>
<td>Voicing begins before stop release</td>
<td>voiced</td>
<td>voiced</td>
</tr>
<tr>
<td>short-lag</td>
<td>Voicing onset occurs at the time of stop release or very soon after</td>
<td>voiceless unaspirated</td>
<td>voiced or voiceless</td>
</tr>
<tr>
<td>long-lag</td>
<td>Delay between stop release and onset of voicing</td>
<td>voiceless aspirated</td>
<td>voiceless</td>
</tr>
</tbody>
</table>

Note: The glottal pulsing (voicing) that occurs during the closure interval of word-initial tokens of /b d g/ is referred to both as “lead” VOT and as “pre-voicing”. I will use both terms, as in the literature.
What is VOT?

However, later work revealed the presence of “non-modal” VOT values

- Japanese, for example, has stops whose VOT values fall in between the “modal” short-lag and long-lag VOT values of Romance /b d g/ and English /p t k/ (Raphael et al. 1995; Cho & Ladefoged, 1999; Flege & Port, 1984, Riney et al., 2007)

- Languages like Danish have VOT values exceeding the “modal” long-lag VOT of English (Cho & Ladefoged, 1999; Mortensen & Tøndering, 2013)
What is VOT?

Children learning their L1 establish language-specific **phonetic categories** based on what they hear.

For example, children learning English in a monolingual environment will encounter a distribution of VOT values like this for /t/ in slow, careful speech.

What they hear will define their later production and perception.

*The distribution of VOT values in 60 /t/ initial words and non-words produced by 20 monolingual English adults (Flege et al., 1998)*
Children learning languages having shorter or longer VOT values than English will develop different language-specific phonetic categories.
What is VOT?

VOT in word-initial stops may co-vary or co-exist with other acoustic properties having perceptual cue value:

1. height of the following vowel (e.g., Mortensen & Tøndering, 2013);
2. degree of stress/emphasis (Lisker & Abramson 1967)
3. speaking rate, including duration of the following vowel (e.g., Theodore et al. 2009) which may vary according to the number of syllables in the word and position in utterance;
4. speech clarity (e.g., Kessinger & Blumstein, 1997; Smiljanić & Bradlow, 2005)
5. socio-phonetic factors (e.g., Docherty et al., 2011)
6. burst and aspiration intensity (e.g., Repp, 1979)
7. F0 onset frequency and movement pattern (e.g., Hombert, Ohala & Ewan, 1979; Dmitrieva et al. 2015)
8. F1 onset frequency and movement pattern (e.g, Hillenbrand, 1984)
9. spectral tilt, H1-H2 (e.g., Kong et al., 2012)
English vs. Romance languages

Now let’s consider how VOT is used in English and Romance languages such as Italian.

English is classified as an **aspiration** language, Romance languages as **voicing** languages.

The distinction - aspiration vs voicing - is based on cue reliability.

**Aspiration** is **reliability present** in the long-lag productions of English /p t k/ but is minimal and sometimes absent in short-lag productions of /p t k/ in Romance languages;

**Glottal pulsing** (=voicing) is **reliably present** in the lead VOT productions of Romance /b d g/ but is often absent in English /b d g/.
The difference between /p t k/ in English and Romance languages is straightforward, but not the differences for /b d g/. For example: MacKay et al. (2001) examined production of /b/ by 20 Italian monolinguals; 99.5% of tokens were produced with pre-voicing.

VOT values in word-initial /b/ tokens produced by 20 Italian monolinguals

adapted from MacKay et al. (2001), Fig. 1
English vs. Romance languages

The English monolinguals pre-voiced /b/ just 24% of the time and their pre-voicing, when it did occur, died out before stop release in 72% of instances, something never observed in Italian monolinguals’ productions of Italian /b/

adapted from MacKay et al. (2001), Fig. 1
Flege & Brown (1982) calculated the percentage of /b/ and /p/ closure intervals that were filled with glottal pulsing (voicing)

In the initial position of utterance-medial and final words a large difference existed between /b/ and /p/, even in post-stressed position.

However, in absolute utterance initial position the difference in glottal pulsing between English /b/ and /p/ was minimal.
Dmitrieva et al. (2015), like others before, observed enormous differences in the use of pre-voicing by 30 English monolinguals. One participant always pre-voiced, seven never pre-voiced, and 22 others pre-voiced some of the time.
The language-specific **phonetic realization rules** (PRRs) used to produce phonetic differences between /b d g/ and /p t k/ are straightforward in Romance languages.
Phonetic organization is less clear in English. A consideration of laryngeal timing patterns in English may help clarify the situation.

Flege (1982) examined three dimensions in the production of English /b/:

1. the closure and release of labial constriction was defined by variations in intra-oral air pressure;

2. glottal pulsing (voicing) was observed using a sensitive throat microphone;

3. the timing of vocal fold adduction was monitored using a Fourcin laryngograph.
Here is an example of a pre-voiced token of /b/ in which the vocal folds were **adducted** long before **stop release**. Only later – but well before stop release – did voicing (glottal pulsing) begin.
English vs. Romance languages

The Flege (1982) study revealed less variability in laryngeal timing patterns than in acoustically measured VOT:

- 2/9 English monolinguals examined adducted the vocal folds at the moment of stop release, always producing short-lag VOT values in /b/;

- 7/9 adducted the vocal folds about 200 msec before stop release. For these seven English monolinguals there was always a delay between vocal fold adduction and voicing onset; sometimes voicing ceased before stop release.
The one Spanish-English bilingual tested by Flege (1982) showed a different pattern of laryngeal timing, using a soft phonation onset when producing /b/. For him:

- glottal pulsing began as soon as the vocal folds were adducted at midline, about 100 msec before stop release;

- once voicing began, it continued until stop release.

Additional research is needed to determine if a soft phonation onset is typical for native speakers of Romance languages.

If so, it could explain why all of the Italian monolinguals examined by MacKay et al. (2001) pre-voiced, and why their pre-voicing continued without interruption until stop release.
I propose that children learning English as an L1 develop either a “lead” or “short-lag” phonetic categories for /b d g/*

Solution 1
Mostly lead input

Solution 2
Mostly short-lag input

*See also MacLeod & Stoel-Gammon (2009, p. 74) for Canadian English
The kind of category that children learning English in a monolingual environment eventually develop, according to this proposal, will depend on input.

1. A minority of children who learn English as an L1 will hear mostly short-lag realizations of /b/. They will learn to adduct the vocal folds at the moment of stop release and, in later life will always produce short-lag VOT;

2. The majority of children, however, will hear /b/ produced with full or partial pre-voicing more often than with short-lag VOT. Pre-voicing is the dominant pattern for them. They will learn to adduct the vocal folds about 200 msec before stop release.
As adults, the children who establish “lead” phonetic categories for /b d g/ will pre-voice /b d g/ some of the time.

The Phonetic Realization Rules (PRRs) they develop to produce their lead phonetic categories for /b d g/ will not guarantee an immediate onset of glottal pulsing, nor a continuation of voicing until stop release as in Romance languages.

A PRR like the ones presumably found in Romance languages is not necessary in English, an aspiration language.

This is because the primary phonetic goal of English is to guarantee the presence of aspiration in /p t k/, not the presence of glottal pulsing before the release of /b d g/.
I’ve mentioned phonetic categories several times. What are they? A **phonetic category** is a perceptual representation that speaker-hearers develop over time in long-term memory based on the phonetic input they receive.

It is defined by all of the tokens encountered on the phonetic surface that have been identified as being instances of the phonetic category.

According to Flege & Schmidt (1995, pp. 92-93) phonetic categories for word-initial stop consonants specify:

- how the stop consonants “ought” to sound when produced;
- the relative importance (weight) of various acoustic phonetic dimensions as perceptual cues to the distinction between /b d g/ and /p t k/
Phonetic organization

Phonetic categories are inherently multidimensional

• The absolute **normative values** of the dimensions, as well as their relative importance, may vary as a function of “phonetic context … degree of stress or emphasis, and speaking rate”

• Among monolingual speakers of a single language, individual differences may exist in the relative weighting of acoustic phonetic dimensions (“cue weighting”)

• However, such individual difference are normally smaller than those distinguishing speakers of different languages (Flege & Schmidt, 1995, pp. 92-93)
Phonetic organization

As for defining how phonic elements “ought to sound” via self-hearing, consider these results from Flege & Schmidt (1995)

The 17 members of a VOT continuum ranging from /ba/ to /pa/ were randomly presented to English monolinguals, who used a 9-point scale to rate the VOT stimuli for “goodness” as an instance of the English /p/ category.
Phonetic organization

The English monolinguals gave increasingly high ratings as VOT values in the stimuli increased.

As VOT increased further, beyond values typical for English, the listeners gave increasingly lower ratings.

Information stored in their phonetic categories guided the listeners’ responses and defined for them how English /p/ “ought” to sound.

Data from Fégre & Schmidt (1995), Fig. 5(a) for stimuli having a duration of 125 msec.
Several days later the English monolinguals identified the same VOT stimuli using one of three labels: phonologically voiced (/b/), phonologically voiceless (/p/), or as exaggerated /p/*

Their /b/-/p/ phoneme boundary was only slightly more decisive than the crossover between /p/ and “exaggerated /p/*”, a non-category that was not defined or illustrated for the participants.
Here we see both the goodness ratings and the identification data.

The peak in the goodness ratings and the greatest frequency of /p/ identifications coincided at the stimulus having a VOT value of 50 msec.

Phonetic organization
As already mentioned, phonetic categories are motorically output using **phonetic realization rules** (PRRs)

PRRs continue to develop until they yield phonetic segments that coincide with inner representations of how the phonetic segments “ought” to sound

> This is a crucial aspect of speech learning inasmuch as we “*speak to be heard in order to be understood*” (Jakobson et al., 1952: p. 13)

> The PRRs are gradually adjusted using feedback provided by self-hearing and oro-sensory sensations. As for the establishment of phonetic categories, the development of PRRs *takes time*

*As reviewed by Flege (1999), production-perception correlations in L2 research average about \( r = 0.50 \) rather than \( r = 1.0 \) because of the time needed for alignment, the inherent incommensurability of dimension in the two domains, and to the fact that speaker-hearers can voluntarily alter production to a much greater extent than perception*
Cross-language differences

Now let’s consider the identification of /p t k/ by speakers of English and Romance languages.

Williams (1977) tested both Spanish and English monolinguals on a synthetic /bi/-/pi/ continuum. For both languages, the percentage of /p/ judgments increased systematically as VOT increased. The phoneme boundary (50% crossover) occurred at longer VOT values for English than Spanish monolinguals.
Cross-language differences

Flege & Eefting (1986) replicated Williams’ identification results using a /da/-/ta/ continuum, whose members were identified by Spanish and English monolinguals, both children and adults.

![Graph showing mean % identification as /t/ across different VOT in stimuli (msec)](adapted from Flege & Eefting (1986) Fig. 2)
Cross-language differences

The findings obtained for English and Spanish suggest that

• VOT functions similarly in English and Romance languages;
• the most important difference between English and Romance languages are the absolute values of VOT needed to shift identification from predominantly voiced to voiceless

This is not always the case, however, as we can see in identification data obtained for monolingual speakers of Canadian French by Caramazza (Caramazza et al. 1973; Caramazza & Yeni-Komshian, 1974)
Cross-language differences

Caramazza expected to obtain identification functions for French like those obtained earlier for Spanish and English monolinguals (adapted from Fig. 3 of Caramazza & Yeni-Komshian, 1974)

adapted from Caramazza & Yeni-Komshian (1974), Fig. 3
Cross-language differences

Here’s what he got instead. The lack of monotonicity in the identification functions suggested to Caramazza that the VOT dimension was insufficient to permit French Canadians to perceptually distinguish /b d g/ from /p t k/.
Cross-language differences

Caramazza’s findings may have reflected a difference in the relative perceptual weight of voicing and aspiration in Canadian and European French (see Serniclaes & Beyster, 1979).

Another possibility: a sound change in progress. Caramazza found that French speakers in France rarely produced /b d g/ with short-lag VOT while those in Quebec did so over half the time, presumably resulting from exposure to short-lag renditions of /b d g/ in English and English-accented French.*

<table>
<thead>
<tr>
<th>Location</th>
<th>% lead</th>
<th>% short-lag</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quebec (Canada)</td>
<td>41%</td>
<td>59%</td>
</tr>
<tr>
<td>Nantes (France)</td>
<td>94%</td>
<td>6%</td>
</tr>
</tbody>
</table>

*MacLeod & Stoel-Gammon (2009) noted a smaller percentage of short-lag stops for French monolinguals living in the area between Ottawa and Montreal.
Cross-language differences

Perhaps to compensate for a change in how French /b d g/ were being produced by people around them, the French speakers in Quebec produced /p t k/ with significantly longer VOT values than those in France.

Doing so would help the French Canadians maintain phonetic contrast between /b d g/ and /p t k/ in their native language.

![Graph showing mean VOT in /p t k/ (msec)](adapted from Caramazza & Yeni-Komshian (1974), Fig. 5)
The kind of cross-language differences just mentioned arise as the result the learning of language specific properties of the native language (L1).

Children acquiring their native language gradually become mature speaker-hearers of their L1 because they have the capacity to make optimal use of the input they receive.

Some believe that not all learners of an L2 have the capacities possessed by L1-learning children.

The Critical Period hypothesis assumes – at least implicitly – that some of the core capacities used by L1 learning children are lost or attenuated somewhat later in life.

The research presented in this section leads me to challenge this untested assumption.
The speech learning capacities I have in mind are the abilities to:

1. auditorily detect phonetic differences between contrastive sets of speech sounds found in the L1 and L2;

2. use statistical information derived from input received to construct auditory equivalence classes (e.g., Maye et al., 2002; Anderson et al. 2003; Kuhl, 1983);

3. develop long-term memory representations – phonetic categories - deriving from the equivalence classes;

4. create stable motor plans that can be used to reproduce the information specified in the phonetic categories
Capacity for speech learning

Children learning an L1 must sort ambient-language phones into sound types

They are forced to depend on messy input

... because they can not know beforehand how many categories their L1 possess
Capacity for speech learning

Cognitive mechanisms permit sounds encountered on the phonetic surface to be aggregated into inchoate equivalence classes based on multiple dimensions.

As more input is received, the equivalence classes evolve into phonetic categories that can be used to identify words.
The phonetic categories continue to develop slowly over time via the cognitive mechanisms of

- **Acquired distinctiveness**: which augments sensitivity to differences between categories localized at the boundary between categories;

- **Acquired similarity**: which reduces sensitivity to differences between members of the same category within the perceptual space occupied by the category (see Heeren, 2006)
Capacity for speech learning

As phonetic categories mature, “category centers” emerge.

The category centers, which are derived from frequency patterns in the input received, permit speaker-hearers to:

- identify an array of L1 phones as instances of a single category with increasing speed and accuracy
- recognize speech sounds in non-ideal listening conditions
Capacity for speech learning

The category centers also permit listeners to notice and report differences in the “goodness” of an array of VOT stimuli as instances of a particular category, as was mentioned earlier.

![Graph showing mean goodness ratings vs. Stimulus VOT (msec)]
Capacity for speech learning

If the capacities needed to learn the L1 are lost or diminished after the end of a “critical period” we must ask:

which specific capacities are affected?

Perhaps phonetic information not relevant to phonemic contrasts in the L1 will be discarded and so be unavailable for use in L2 speech learning.

As an example: Native English learners of French might ignore pre-voicing in French if they consider the presence vs. absence of pre-voicing to be a bothersome allophonic detail, especially if they themselves typically produce English /b d g/ with both lead and short-lag VOT values.
Another possibility is that cross-language phonetic differences get filtered out in early stage of auditory processing if not already used to specify L1 phonetic categories.

Flege (1984) evaluated the “filtering” hypothesis by examining the ability of monolingual English adults to detect small VOT differences.
Capacity for speech learning

The stimuli were English /ti/ and /tu/ tokens, half produced by English monolinguals, the other half by native speakers of French who spoke English with mild foreign accents.

VOT in the two sets of stimuli differed by $M = 24$ msec, much less than the VOT difference between English and French monolinguals.

<table>
<thead>
<tr>
<th></th>
<th>Mean VOT in /i/ context</th>
<th>Mean VOT in /u/ context</th>
</tr>
</thead>
<tbody>
<tr>
<td>English stimuli produced by</td>
<td></td>
<td></td>
</tr>
<tr>
<td>native English (n=8)</td>
<td>79</td>
<td>78</td>
</tr>
<tr>
<td>native French (n=8)</td>
<td>63</td>
<td>46</td>
</tr>
</tbody>
</table>

To ensure that vowel quality would not affect judgments, the variable burst + VOT intervals in the natural /ti/ and /tu/ syllables were cross-spliced onto a single native-produced vowel (/i/ or /u/)
Ten native English listeners heard two English stimuli per trial, one produced by a native English speaker, the other by a native speaker of French.

The listeners’ task was to decide which member of each pair of stimuli had been produced by a non-native (French) speaker.
Capacity for speech learning

The listeners were able to select the “foreign” member of the stimulus pairs at significantly above-chance rates.

They were able to detect cross-language VOT differences averaging just 24 msec.

If in the future these listeners want to learn French the much larger French vs. English VOT difference should be readily detectable for them.

We can therefore reject the “filtering” hypothesis.
Capacity for speech learning

In Flege (1984) two stimuli were presented on each trial. The findings did not, therefore, demonstrate the ability to construct auditory equivalence classes for stimuli differing in VOT (see e.g., Kuhl, 1983).

The next study to be presented overcame this limitation.
Flege & Hammond (1982) tested native English students taking a 1st year Spanish class at the University of Florida. The class was taught by a native speaker of Spanish who spoke English with a Spanish accent.

Fifty participants (half male) were selected from the original sample of 125. All of them:

- had lived in Florida for at least the previous 5 years;
- were personally acquainted with native speakers of Spanish.
The participants were recorded in a language laboratory following their mid term oral examination in Spanish. They were asked to read a list of English sentences with a “typical Spanish accent”. No explanation was provided as to how to speak with a Spanish accent. The participants found the task entertaining, likely due to the presence of their Spanish teacher.

Each of the sentences contained two test words. For example: *The vase is on the pig*

The two variable test word in each sentence began with a sound sometimes produced incorrectly in Spanish-accented English. For example, “pig” is sometimes produced with [i] in Spanish accented English rather than as [ɪ].
Phonetic transcription of the test words yielded 508 “Spanish accent” substitutions

We never heard a /d/ for /t/ substitution. However VOT measurements showed that the initial /t/s were indeed modified.

<table>
<thead>
<tr>
<th>Substitution</th>
<th>Total (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>z → s</td>
<td>141 (47%)</td>
</tr>
<tr>
<td>v → b</td>
<td>129 (43%)</td>
</tr>
<tr>
<td>i → i</td>
<td>127 (42%)</td>
</tr>
<tr>
<td>u → u</td>
<td>61 (20%)</td>
</tr>
<tr>
<td>j → č</td>
<td>49 (16%)</td>
</tr>
<tr>
<td>n → ŋ</td>
<td>1 (&gt;1%)</td>
</tr>
<tr>
<td>t → d</td>
<td>0 (0%)</td>
</tr>
</tbody>
</table>
Two subgroups of 10 participants each were selected for acoustic analysis:

- we’ll call those who produced the fewest number of Spanish accent substitutions \((range = 0 \text{ to } 4)\) the “Least Knowledge” group
- those who produced the largest number \((range = 17 \text{ to } 21)\) the “Most Knowledge” group

It is likely that members of the “Most Knowledge” group had received more exposure to Spanish-accented English than those in the “Least Knowledge” group

Flege & Hammond (1982) also recorded production of the same sentences by a Control group of native English speakers who simply read the utterances with no special instruction
For /t/ tokens in both utterance-initial and -final words, the “Most Knowledge” but not the “Least Knowledge” group produced significantly shorter VOT values than the Control group ($p < 0.01$)
Capacity for speech learning

In this frequency histogram we see that members of the **Control group** never produced /t/ with Spanish-like short-lag VOT.
Members of the “Most Knowledge” group did so frequently, Members of the “Least Knowledge” group also produced Spanish-like short-lag VOT, but somewhat less frequently.
In summary: the adult English monolinguals shortened VOT in word-initial /t/ tokens by an average of 30 msec.* They were able to:

- detect VOT differences between their L1 and an accented version of their L1;
- store the detected information in long-term memory;
- later use that information in production

From these results I conclude that young adults who have passed the end of a putative Critical Period maintain the core capacities needed for L2 speech learning.

*A recent study by Neuhauser (2011) indicated that Germans who imitate a French accent in English significantly shorted VOT in phonologically voiceless stops.
Capacity for speech learning

But what about the ability to **aggregate** sets of diverse speech sounds into an equivalence class and, later, into a phonetic category? Two studies bear on this issue.

Pisoni et al. (1982) asked English monolinguals to use two labels to identify the members of VOT continuum. The results were unsurprising.

![Graph showing forced-choice identification on two days by English monolinguals using either 2 or 3 category labels.](adapted from Pisoni et al. (1982) Fig. 2)
On a separate day of testing they were also asked to use three category labels. Nearly all (18/20) spontaneously divided the continuum into “lead”, “short-lag” and “long-lag” VOT categories even though a lead vs. short-lag distinction is not phonemic in English.

Forced-choice identification on two days by English monolinguals using either 2 or 3 category labels

VOT of stimuli in synthetic /ba/-/pa/ continuum (msec)

adapted from Pisoni et al. (1982) Fig. 2
Hillenbrand et al. (1990) recruited three groups of participants at Northwestern University (NWU)

- **Phonetically trained** – Speech researchers and Speech-Language Pathologists recruited in the NWU Communications Disorders Department

- **Musically trained** – highly skilled musicians recruited at the NWU School of Music

- **Untrained** – undergraduates who had just begun an introductory phonetic class taught by the first author
Individual differences

The participants were asked to identify randomly presented members of a VOT continuum as “b” or “p”.

The results obtained for the three groups were quite similar, and unsurprising.

adapted from Hilendland et al. (1990), Fig. 1
Individual differences

However, between-group differences emerged when participants were later asked to use three labels to identify the same stimuli.

As in the Pisoni et al. (1982) study, no training or explanation was provided on how to use three labels.
Individual differences

Here we see just the percentage of “indefinite” (intra-phonemic) judgments.

The “clear b” and “clear “p” judgments are not shown.

The “Phoneticians” were better able to use the non-phonemic “indefinite” category than members of the other two groups.

(adapted from Hillenbrand et al. (1990) Fig. 2)
Individual differences

We don’t know if the Phoneticians’ ability to aggregate a set of stimuli “on the fly” was due to an innate ability, or if the work they did augmented this capacity. Whatever the source, I suspect that differences like the one just shown are relevant to L2 speech learning.
Individual differences

Flege et al. (1998) observed a huge amount of variability among native Spanish Late learners in the production of English /t/. Variability like this has generated intense interest in individual differences (e.g., Markham, 1997)

![Graph showing mean VOT in English /t/ (msec)]

- **English monolinguals**
- **Spanish Late learners** of English (AOA = 18-50)

*Data from Flege et al. (1998)*

*Means are based on 60 observations*

*Error bars +/- 1 SE*
Individual differences

Cognitive tests that identity individual differences in measures of paired associate learning, selective memory, attentional switching and so on usually prove to be only weak predictors of the outcomes of laboratory training involving young adults with normal hearing (Lengeris & Hazan, 2010, p. 3766)

However, the kind of inter-subject variability I just illustrated was the result of input obtained over years, not hours

Learning speech naturalistically through immersion is very different from pushing buttons in a lab and so the results of laboratory studies may not generalize to long-term L2 speech learning
Individual differences

In my view, the best way to identify the source of individual differences in long-term L2 speech learning is to focus on core capacities.

One of these is the ability to aggregate sets of sounds and hold them in memory, as was seen in the Pisoni and Hillenbrand experiments reported earlier.

The “aggregation” ability, in turn, depends on phonological short term memory (PSTM).
Individual differences

MacKay, Meador & Flege (2001) evaluated PSTM using a non-word repetition task. The stimuli were sets of non-words created by splicing together 2 to 5 Italian CV syllables, then digitally modifying the string to ensure the perception of constant penultimate stress.

The five sets of non-word stimuli were presented in blocks of increasing length for repetition, and the number of “correct” (excluding segmental pronunciation errors) counted.

<table>
<thead>
<tr>
<th>Examples (boldface indicates stress)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>5</td>
</tr>
</tbody>
</table>
The non-word repetition test was administered to 72 native Italian participants, all long-time residents of Canada mean LOR = 30 years)

The number of non-words repeated correctly was completely unrelated to language background and L2 experience because PSTM is an individual characteristic

The PSTM scores did, however, partially predict the Italian subjects’ identification of word-initial and word-final English consonants

The consonant identification scores, in turn, partially predicted the recognition of English words (Meador et al., 2000)
L1 development

We can not hope to understand how L2 speech is learned if we do not first understand L1 speech development.

It takes many years for children to become mature speaker-hearers of their L1.

Long after children establish the ability to distinguish the phonemes of their L1 (e.g., “bat” vs. “pat” vs. “cat”), language-specific phonetic categories continue to develop.
L1 development

L1 phonetic development takes many years to complete because

- phonetic categories encode a complex array of information derived from exposure to huge numbers of tokens;

- these long-term memory representations integrate information from multiple dimensions across contexts;

- in addition to defining category centers, phonetic categories must encompass outliers that, although unusual, must still be processed as member of the category

This explains why, for example, adults are better able than children to recognize words produced with a foreign accent (Bent and Atagi, 2017)
L1 development

If an 8-year-old child learning English in a monolingual environment differs from his/her parents, either in production or in perception, we do not say that the child has failed to learn. We say that the child “is still learning”.

Researchers have been less kind to L2 learners, probably due to the influence of the Critical Period Hypothesis (CPH).

As noted by Flege (1987) enthusiasm for the CPH has tended to impede progress in L2 speech research by making more readily testable hypotheses, such as those regarding the influence of the quality and quantity of L2 input received, appear to be “unwarranted” (p. 174).
In my opinion, L2 learners should not be expected to develop phonetic categories and language specific realization rules (PRRs) for L2 sounds more rapidly than so-called “pre-Critical Period” children who are acquiring their native language.

I therefore propose the following benchmark for individuals of all ages who are learning the VOT dimension in an L2:

10 years of native-speaker input

Given time limits, I can only briefly outline the data upon which my proposed benchmark is based.
First, children’s production of VOT is more variable than adults’ until about 13 years of age (e.g., Eguchi & Hirsh, 1969; Koenig, 2001)

Second, children are less able than adults to correctly identify speech sounds in non-ideal listening conditions:

- Children continue to differ from adults when recognizing consonants in reverberation until the age of 13 years (Neuman & Hochberg, 1983)

- And continue to differ from adults in recognizing consonants in noise until the age of 13 to 15 years (Elliott, 1979)
L1 development

Johnson (2000) examined the identification of English consonants in quiet and in noise.

Adults outperformed children aged 6-to-7 and 10-to-11 in both conditions.

When just the voicing feature was considered, even 14-to-15 year-old children differed from adults.
Elliott et al. (1986) found that monolingual English children needed longer intervals than adults in order to discriminate stops differing in VOT.
Finally, Flege & Eefting (1986) found differences in the identification of stops between 8-9 year-old children and adults. In both languages, adults required significantly longer VOT values to shift from predominant /d/ to /t/ judgments than children did.

Mean % identification as /t/

adapted from Flege & Eefting (1986) Fig. 2
A follow-up experiment by Flege & Eefting (1986) revealed that phoneme boundaries of even 17-year-olds differed from those of native English adults.

Mean VOT of /d/-/t/ boundary (msec)

![Bar chart showing mean VOT of /d/-/t/ boundary across different age groups. The chart indicates a significant difference between 17-year-olds and adults.]
Flege & Eefting (1986) also found that Spanish and English adults produced /t/ with longer VOT values than children. (The difference, which averaged 6 msec, narrowly missed reaching significance)
Interlingual identification

I will conclude Part 1 by briefly discussing **inter-lingual identification**, 

This cognitive mechanism serves as a perceptual bridge between the L1 and L2 sound systems
Interlingual identification

Inter-lingual identification operates automatically when listeners encounter forms of speech differing from their own

L2 learners seek and inevitably find correspondences – that is not to say perfect matches – between sounds in their L1 and L2

Example: Spanish monolinguals consistently identify long-lag English stops as “p” despite large cross-language differences in VOT
Interlingual identification

The aim of Bohn & Flege (1993) was to assess the interlingual identification of word-initial English stops by native speakers of Spanish.

In addition to a group of English monolinguals, we recruited three groups of native Spanish speakers in Birmingham, Alabama.

- Monolinguals who had just arrived in Birmingham;
- Early learners of English;
- Late learners of English.

Ocke-Schwen Bohn
Interlingual identification

The stimuli used by Bohn & Flege (1993) to assess inter-lingual identification consisted of four sets of naturally produced CV stimuli (9 each) drawn from Spanish and English.

The task of the four groups of 10 participants each was to identify the syllable-initial consonants as “d” or “t”.

Naturally produced stimuli from the Bohn & Flege (1993) study

Spanish stimuli
lead /d/
short-lag /t/

English stimuli
short-lag /d/
long-lag /t/
Interlingual identification

As we see here, the Spanish and English monolinguals consistently identified stimuli with lead and long-lag VOT as “d” and “t”, respectively. However, the two groups of monolinguals were equally confused in how they identified stimuli that had been produced with short-lag VOT values.
Interlingual identification

English /d/ stimuli with short-lag VOT  If only VOT mattered, the Spanish monolinguals would be expected to identify these stimuli consistently as “t”. In fact, they did so less than half the time. VOT did not by itself, convince the Spanish monolinguals that the short-lag English /d/ stimuli were instances of their short-lag Spanish /t/ category.

![Mean % "t" judgments](image_url)
Interlingual identification

Spanish /t/ stimuli having short-lag VOT If only VOT mattered these stimuli should be consistently identified as “t” by Spanish monolinguals and as “d” by English monolinguals (see for example Elman et al., 1977)
Interlingual identification

**Spanish /t/ stimuli with short-lag VOT** In neither of two “stimulus range” conditions were these expectations met for either Spanish or English monolinguals. Once again, neither group was completely “convinced” by VOT.

![Graph showing % identification of 9 short-lag Spanish [t] tokens as /t/ for English monolinguals, Early learners, Late learners, and Spanish monolinguals. Error bars are present on each bar. The graph is adapted from Bohn & Flege (1993), Fig. 2.]

- **Spanish short-lag [t] tokens presented with:**
  - Spanish lead
  - English long-lag
  - Spanish lead
  - English short-lag
Interlingual identification

**Spanish /t/ stimuli with short-lag VOT** The same held true for Early and Late L2 learners. No one was convinced by VOT, even the bilinguals who were very aware of Spanish vs English VOT differences. Why not?

![Graph showing % identification of 9 short-lag Spanish [t] tokens as /t/ for different groups: English monolingual, Early learner, Late learner, Spanish monolingual.](image)

*adapted from Bohn & Flege (1993), Fig. 2*
Interlingual identification

**Spanish short-lag stops** Bohn & Flege (1993) examined the identification of Spanish short-lag /t/ tokens in two experiments.

Here the data for all 40 participants have been pooled. Participants responded in much the same way in both experiments. Surprisingly, the “token effects” were much larger than the between-group differences!
Interlingual identification

Flege & Bohn (1993) tried to find some acoustic measure(s) that could differentiate the “usually t” from the “ambiguous” stimuli. They measured

- VOT
- Fundamental frequency contours
- Burst intensity
- Burst duration
- Duration of the following vowel

Nothing worked

A large, focused study would probably succeed in identifying acoustic phonetic dimensions that, when taken together, will predict identification of short-lag Spanish and English stops as “t” or “d”.
Interlingual identification

In the meantime, those who do L2 research should remember that

1. VOT is not a unidimensional property of stop consonants

2. the outcome of early stages of L2 learning will be influence by the initial “mapping” of L2 sounds onto L1 categories via inter-lingual identification

The initial mapping between L1 and L2 sounds may, in turn, depend on cue weighting in individual learners’ native language phonetic categories at the time L2 learning begins.

A common error in L2 research is the assumption that all speakers of a particular L1 have a single “starting point” when L2 learning begins

When/if differences exist, they may contribute to differing outcomes in L2 learning that masquerade as difference in speech learning aptitude
References


Dmitrieva, O. et al. 2015. Phonological status, not voice onset time, determines the acoustic realization of onset f0 as a secondary cue in Spanish and English. *J. Phonetics*, 49, 77-95.


References


References


References


References

Neuhauser, S. 2011. Foreign accent imitation and variation of VOT and voicing in plosives. ICPhS XVII, Hong Kong, 17-21 August 2011


