The Instrumental Study of L2 Speech Production: Some Methodological Considerations*

James Emil Flege
University of Alabama at Birmingham

This article discusses the design and interpretation of instrumental phonetic studies of L2 speech production. The notion phonetic norm plays a crucial role in many recent studies that examine acoustic dimensions such as the frequencies of F₁-F₃, vowel duration, or voice-onset time (VOT). The speech of L2 learners is evaluated to determine to what extent it diverges from the differing phonetic norms of L1 and L2, which are estimated from the speech of a small number of native speakers. It is advisable to use monolinguals to establish the L1 norms because L2 learning has been shown to affect L1 production. In initial stages of learning, L2 learners often produce words in L2 with L1 sounds: they do not modify phonetic dimensions when producing corresponding sounds in L1 and L2 phonetic norm. L2 learners with relatively greater experience in L2, on the other hand, generally approximate the phonetic norm of L2: the values of the phonetic dimensions being measured generally fall between the mean values representing the phonetic norms of L1 and L2.

Research in the area of interlanguage phonology has focused increasingly on how foreign or second language (henceforth L2) are actually pronounced (parole), rather than on theoretical considerations pertaining to the L2 learner's evolving linguistic competence (language). We have learned that differences between native and nonnative speakers arise either because of difficulties at a "categorical" or "subcategorical" level. A "categorical" difficulty exists when L2 learners fail to recognize that two phones in the L2 are realizations of different categories. This may happen because only one of the two L2 phones occurs with regularity on the phonetic surface of the native language (L1), or because both L2 phones closely resemble phones in L1 which are realizations of a single L1 category. Examples of this phenomenon are abundant, such as the confusion of English /r/ and /l/ by native speakers of Japanese, or the Spanish speaker's confusion of English /i/ and /ɪ/.

*This report was supported by NIH grant NS20963. The author would like to thank S.G. Fletcher for commenting on a previous draft of this report.
“Subcategorical” difficulties exist when learners are aware of contrasts that exist in L2, but fail to realize those contrasts effectively due to phonetic interference (Flege 1980, 1981; Flege and Port 1981). Phonetic interference is the transfer into L2 of patterns established during L1 acquisition for speech production and perception (Flege and Davidian 1984; Flege and Hillenbrand 1985, 1987). Segmental articulation has been the most frequent topic of previous studies examining phonetic interference. It has been shown, for example, that the vowel /u/ is realized with significantly higher F2 values in English than French (probably because English has no /y/ category with which /u/ might be perceptually confused). This means that a native English speaker who produces a French word such as *tous* (/tu/) with an “English” /u/ might be perceived as having produced /y/ (in some extreme cases) or to have diverged from the French phonetic norm for /u/ (in more typical situations; see Flege 1984; Flege and Hillenbrand 1984).

It is sometimes difficult to differentiate categorical from subcategorical difficulties. How can one determine with certainty whether a Japanese L2 learner “knows” (in other than a purely metalinguistic sense) that [r] and [l] are not realizations of the same category? Or that a Spanish L2 learner “knows” that the phones [i] and [l] are distinct from one another when they occur in English words such as *beat* and *bit*? The ability to discriminate a contrast in English minimal pair words does not in itself demonstrate convincingly that an L2 learner is aware a categorical distinction exists. The acoustic difference between phones used to realize two categories might be discriminated auditorily but not phonetically.

One simple—and important—test of the existence of a categorical distinction is intelligibility. The consistent production of a distinction between many words in the L2 containing the contrast of interest: “rake” versus “lake”, that is recognizable to native speakers constitutes important evidence that an L2 learner possesses knowledge of a categorical distinction. The failure to produce a perceptually effective distinction between two categories, on the other hand, does not constitute proof that an L2 learner lacks such tacit knowledge. L2 learners might not succeed in producing a perceptually effective distinction even though they are aware the distinction exists. It is generally agreed that in L1 acquisition, perceptual awareness of categorical distinctions precedes the ability to implement those distinctions. A certain amount of time and practice is also needed for adults or children to establish the motor control patterns
needed for producing new phones in an L2 (Flege and Davidian 1984; Flege, McCutcheon, and Smith 1987).

In some instances instrumentally derived measures may provide a better method than transcription for addressing questions related to knowledge of categorical differences. Studies of children acquiring their L1 have shown that they may produce a statistically significant acoustic difference between two categories ([t] versus [d]) that is not auditorily perceptible to adults. Instrumental data provide an objective basis for determining whether a category difference exists that is unbiased by previously learned patterns of speech perception (Flege and Hillenbrand 1985, 1987). Acoustic analyses might reveal, for example, that L2 learners produce a systematic acoustic difference between categories along different dimensions than native speakers, or that they produce a contrast along the same dimensions as native speakers that is smaller in magnitude than that produced by native speakers.

The number of instrumental phonetic studies of L2 learning has increased dramatically in recent years for reasons such as these. Data gathered instrumentally, of course, do not in themselves provide answers to questions of phonological interest. The data must be gathered following accepted observational or experimental techniques and must be interpreted within a theoretical model or framework. The following discussion outlines current practices pertaining to the design and interpretation of instrumental phonetic studies.

Most instrumental studies have examined continuously varying (as opposed to discrete) acoustic phonetic dimensions such as vowel formant frequency (in Hz), voice onset time (VOT) (in ms), or peak intensity (in dB). The focus of these studies is the production of some acoustic or physiological dimension(s) in L2 phones by L2 learners. An instrumental study might examine production of an L2 phone that has no obvious phonological counterpart. In these instances, L1 and L2 production cannot be compared. Other studies, however, compare the production of L2 phones that do have a counterpart in L1. The dimension(s) selected for analysis are those known to differ between languages. For example, several studies have investigated the production of English stops by native speakers of Spanish. The voiceless stop phonemes /p, t, k/ are implemented as short-lag stops in Spanish (with VOT values of about 20 ms), but as long-lag stops in English (with VOT values of about 80 ms; Flege and Eefting 1986).

The failure of native Spanish speakers of English to produce English
/p, t, k/ with long-lag VOT values corresponding closely to the values produced in the same phonetic context by English native speakers may be detected by native speakers of English. If detected, the difference may contribute to the perception of foreign accent (Flege and Hammond 1982; Flege 1984); however, the aim of most instrumental studies has not been to establish which dimension(s) contribute(s) most importantly to foreign accent, but to determine to what extent L2 learners, differing in age and/or linguistic experience, differ from native speakers. The underlying questions are: Can L2 learners produce L2 phones authentically? And, if so, under which circumstances are they most likely to do so?

Instrumental studies are based, either explicitly or implicitly, on the assumption that a phonetic norm exists for the L2 and the L1 and that information pertaining to the phonetic norm of a language importantly influences the perception and production of speech. To accurately gauge the L2 learner's success in modifying previously established patterns of production (or perception), or in establishing new patterns for the production (or perception) of phones in the L2, it is first necessary to document differences between L1 and L2 phones. This is usually done by directly comparing the speech of speakers, who speak the L1 and L2 natively (although some studies have simply made reference to previously published phonetic studies).

Only recently have investigators begun paying close attention to subject selection procedures. An attempt should be made to form subject groups that are as homogeneous as possible. Some method must be used to exclude subjects with hearing problems, or who do not speak their native language normally. Since it may be difficult to recruit subjects with narrowly defined characteristics, both male and female subjects are normally included in a single subject group. Research has not yet determined whether sex influences L2 speech learning. It is therefore advisable to balance the ratio of male to female subjects in each group.

Some thought must be given to the number of subjects to include in each group when employing analysis of variance techniques to test for the significance of intergroup differences. The practice of most investigators is to obtain data under identical conditions from groups of 6-12 subjects. Examining fewer subjects leads to the risk of failing to observe a systematic phonetic difference between groups because of intersubject variability.

Given the difficulty of locating subjects, it is desirable to obtain truly representative data from each subject. Speech production exhibits some
degree of variability. For a single talker, a standard deviation of 5–15 ms is not uncommon for a mean VOT value of 80 ms for /t/. (Even greater variability is to be expected if /t/ is measured in several different phonetic contexts.) Therefore, the values submitted to statistical tests are normally individual subject means based on multiple observations (5–20 measurements) of the dimension being examined.

Estimating the phonetic norm of the L2 poses no special problem for studies conducted in the U.S. that examine English as the target L2. Individuals who speak the same (or a highly similar) dialect of American English can be located readily; however, obtaining a valid estimate of the phonetic norm of the L1 may be more difficult. In many previous studies, the estimate of the L1 phonetic norm has been based on a speech sample taken from the same subjects whose L2 speech was examined. An example of this is Flege and Port's (1981) study of the production of Arabic L1 and English L2. In that study, the L1 and English L2 data were elicited from the same subjects; native speakers of Saudi Arabian Arabic who were living in the U.S. at the time of the study. A repeated measures design, in which subjects' productions of phonetic material in L1 and L2 are directly compared, reduces the number of subjects that needs to be recruited. This design has the added advantage of using each subject to serve as his/her own control. Interlanguage differences that might be obscured (or created) by subject selection biases can, thereby, be minimized.

A reasonable assumption is that, if L2 learners produce a phonetic dimension, (the VOT in /t/) with significantly different values between L2 and L1, they must have noted the phonetic difference between L1 and L2 (at some level). Individual learners, who are experienced in L2, usually produce phones in L2 with acoustic parameter values that fall between the mean group values estimating the phonetic norms for LI and L2. For example, Flege and Eefting (1987) observed that subjects in three native Spanish groups produced English /p, t, k/ with VOT values that were greater than those observed for age-matched native Spanish subjects, who did not speak English, but less than the mean values observed for age-matched English monolingual subjects. Such a finding is interpreted to mean that the L2 learners “approximated” the L2 phonetic norm (Flege 1980, 1981). It suggests that, not only have the L2 learners noted phonetic differences between L1 and L2 at a sensory level, they have also learned—at least partially—to effect those differences in speech production at a motoric level.
There is a serious disadvantage that derives from using L2 learners to estimate the phonetic norms of L1. Two recent studies showed that L2 learning may influence L1 production. Flege (1987) examined the production of English by native French speakers and French by native English speakers. Both groups produced stop consonants in L1 which differed significantly (in VOT) from stops produced by monolingual native speakers of the L1, French or English. Their L1 stops resembled the corresponding stops in L2 along the VOT dimension. Flege and Eefting (1987) examined the production of /p, t, k/ with significantly shorter VOT values than native Spanish subjects who did not speak an L2.3 These studies also showed, consistent with previous results (Suomi 1976; Flege 1980), that experienced L2 learners produced L2 stops with VOT values between the phonetic norms of L1 and L2.

The importance of appropriate experimental design and procedures is essential to reaching valid conclusions. This can be illustrated by briefly reviewing an article published recently in this journal. This article appeared to show that L2 learners do not approximate the phonetic norm of L2 produced stops with VOT values between the L1 and L2 norms; however, this conclusion can be challenged on methodological grounds.

Fourakis and Iverson (1985) examined timing patterns produced in Arabic and English L2 by four native speakers of the Egyptian Arabic spoken in Cairo. The speech of four native English subjects was also used to estimate the phonetic norm of English. The language background of the subjects was not described. VOT in prevocalic stops was one of the acoustic dimensions measured. The mean values reported by Fourakis and Iverson (1985) are summarized:4

<table>
<thead>
<tr>
<th></th>
<th>Arabic</th>
<th>English</th>
<th>English L2</th>
</tr>
</thead>
<tbody>
<tr>
<td>/p/</td>
<td></td>
<td>44</td>
<td>27</td>
</tr>
<tr>
<td>/t/</td>
<td>39</td>
<td>59</td>
<td>30</td>
</tr>
<tr>
<td>/k/</td>
<td>65</td>
<td>63</td>
<td>46</td>
</tr>
</tbody>
</table>

The Arabic VOT values shown here are slightly longer than those reported previously by Flege and Port (1981) for tokens of /t/ (37 ms) and /k/ (52 ms) produced by six native speakers of Saudi Arabian Arabic, and are about 20 ms longer than the values reported for Lebanese Arabic by Yeni-Komshian, Caramazza and Preston (1977). It is not certain whether the apparent differences between studies are due to differences in the native dialect of the speakers, their experience in English L2,
the key words examined, the carrier phrase in which the key word was embedded (not reported by Fourakis and Iverson), speaking rate, or some other more subtle factor (such as the language background of the experimenter eliciting the data).

Fourakis and Iverson (1985) did not test the statistical significance of the VOT difference between Arabic and English. It is, of course, difficult to make valid, cross-language, comparisons of parameter values for a phonetic dimension. Many phonetic dimensions are not directly comparable between languages owing to differences in phonological inventories and phonotactic constraints. Flege and Port (1981) attempted to find speech material in Arabic and English that was as comparable as possible in order to compare the VOT in voiceless stops found in those two languages; however, the presence of confounding phonetic factors was inevitable because it was not possible to find a full list of minimal pairs that were real words in Arabic and English. As a result, they compared the VOT in Arabic and English words which had vowels that differed slightly in spectral quality and duration (Arabic [thɑːs] and English [tʰæs]).

Even when comparable phonetic material can be obtained, as in Barry's (1974) study of German and English, potential confounding factors exist. These include factors such as culturally-based differences in how read and spontaneous speech may differ, differences in degree of stress or emphasis that exist cross-linguistically (or are due to anxiety in producing L2), or differences in speaking rate. Any of these factors might importantly influence phonetic parameters, especially a temporal parameter such as VOT. Perhaps the best solution to this problem is to obtain multiple productions of many words in L1 and L2 that are as comparable as possible. Care should be taken to provide a period of acclimation before recording the speech data that is actually to be analyzed.

Despite the absence of a clear difference between Arabic and English, Fourakis and Iverson (1985) concluded that native speakers of Arabic must learn to increase the duration of VOT in word-initial English stops. Based on previous research, one would expect the native Arabic subjects to have produced /p, t, k/ with slightly longer VOT values in English L2 than Arabic; however, the data reported by Fourakis and Iverson (1985) suggest that they produced voiceless stops with significantly shorter VOT values in English than Arabic.

In two other previous studies the VOT values produced in L2 words by L2 learners did not fall between the L1 and L2 phonetic norms. Suomi
(1976) examined the English spoken by Finns; and Flege and Eefting
(1987 in press) examined the English spoken by native speakers of Dutch.
The voiceless stops /p, t, k/ in the L1 of the subjects examined in those
two studies are implemented as short-lag stops. Some subjects in both
studies were observed to “overshoot” the VOT norm of English. That is,
they produced English stops with even longer VOT values than did
English native speakers. In this regard, they seem to have resembled
some children acquiring English as their L1. “Overshooting” may occur
in L1 acquisition shortly after children first discover how to implement a
contrast based on laryngeal timing between /b, d, g/ and /p, t, k/.

One might speculate that the Arabic subjects examined by Fourakis
and Iverson (1985) produced L2 stops like certain children learning
English as their L1. Some young English-learning children have been
observed to produce a much smaller, albeit statistically significant, con­
trast between /b, d, g/ and /p, t, k/ than do native English adults. At the
same time, it has been observed that stops with short-lag VOT values
may be easier to produce physiologically than long-lag stops; however,
the notion that L1 strategies were “reawakened” during the process of L2
learning by the adult Arabic subjects seems implausible in light of the
fact that they were likely to have been proficient speakers of English. (All
four were enrolled as undergraduate students at the University of Iowa.)
Moreover, Fokes, Bond and Steinberg (1985) recently found that native
Arabic children learning English in the U.S. produced English stops with
VOT values (about 85 ms) that were likely to have equaled or exceeded
the English phonetic norm. One might reasonably expect child L2 learn­
ers to be more likely than adults to show a “reawakening” of L1 strate­
gies.

Fourakis and Iverson (1985) interpreted their data to mean that the
Arabic subjects shortened the VOT interval in voiceless stops when
switching from Arabic to English. They claimed that a shortening
occurred because the learners drew an analogy between voiceless stops
with long-lag VOT values found in English CV syllables and voiceless
stops with short-lag VOT values in word-initial clusters following /s/.
For example, the Arabic subjects produced /t/ in English words like
team with shorter VOT values than is typical for Arabic /t/ because /t/ is
realized in English with short-lag VOT values in words like steam.

However, a difference in phonetic context in the Arabic and English
speech material examined by Fourakis and Iverson (1985) may account
for the apparent shortening effect that was observed. Whereas the two
Arabic words examined (/tuːb/, /kiːd/) contained high vowels, the three English words (/pok/, /tok/, /kok/) all contained the non-high vowel /o/. At least three previous studies have shown that VOT is longer in English /p, t, k/ when these voiceless stops precede high, compared to non-high, vowels. Port and Rotunno (1980) found that VOT was about 12 ms longer preceding /i, u/ than /ʌ. In Weismer's (1979) data, VOT values averaged about 4 ms longer in the context of /i/ than /e/. Klatt (1975) found VOT to be longer by about 12 ms in the context of /i, u/ than /æə/. Since the vowel context effect of VOT is thought to derive from aerodynamic factors inherent in the human speech production mechanism, one would expect to see much the same effect in Arabic. In fact, Yeni-Komshian et al. (1977) reported that the VOT in /t, k/ in Arabic (52 ms) versus English (51 ms) is likely to have underestimated the real VOT difference between Arabic and English because of the difference in vowel context, and, because the subjects used to estimate the Arabic phonetic norm were speakers of English L2 (and may therefore have produced Arabic stops that resembled English stops in terms of VOT). Fortunately, neither the confounding factor of vowel context nor the L2 experience of the Arabic subjects undermines the assumption that voiceless stops in Arabic and English differ in terms of VOT, and that native speakers of Arabic must, therefore, learn to produce voiceless stops with longer VOT values in English than Arabic. The same conclusion was also reached by Flege and Port (1981); however, the confounding of vowel context does undermine the validity of the finding that Fourakis and Iverson's Arabic subjects shortened VOT when switching from Arabic to English. Most, or all, of the observed "shortening" of VOT, which averaged about 14 ms, was likely due to the difference in vowel context in the Arabic and English speech material.

An alternative interpretation is provided by the model of phonetic learning presented by Flege (1981, 1987, 1987 in press). He proposed that the phonetic representations used to guide the production of speech sound categories continue to evolve as a function of phonetic input during L2 learning. He hypothesized that "late L2 learners" (individuals who learn English L2 after early childhood) do not establish new phonetic categories for the long-lag stops of English because of the operation of a basic mechanism called equivalence classification. Late L2 learners continue to judge tokens of /p, t, k/ in L1 and L2 to be realizations of.
the same categories. This suggests that the central representations for /p, t, k/ of adult native Arabic speakers of English will ultimately specify stop consonants that merge the acoustic and articulatory properties of corresponding Arabic and English stops.

Flege's model leads to the expectation that Arabic learners of English L2 will, at best, produce English /p, t, k/ with VOT values that are intermediate to the phonetic norm of Arabic and English and will begin producing Arabic stops with English-like VOT values as they gain experience in English (especially, if they are speaking English more often than Arabic at the time of the study). The VOT difference between corresponding Arabic and English stops appears to be relatively small (compared to the difference, say, between Spanish and English stops [Flege and Port 1981; Flege and Eefting 1986, 1987]). Thus, the model leads to the expectation that the Arabic subjects would produce English stops with just slightly longer VOT values in English than Arabic. A study examining just a few Arabic speakers of English would be unlikely to reveal this effect. This is true especially for very experienced adult L2 learners in a repeated measures design, because L2 learning is likely to have resulted in a small increase in the VOT values observed in L1 stops. This would tend to diminish the apparent difference between L1 and L2 stops and make phonetic approximation difficult to observe. Moreover, if it did occur in the Fourakis and Iverson (1985) study, phonetic approximation is likely to have been observed by the confounding influence of vowel context.

NOTES

1 No attempt will be made here to differentiate phonetic and phonemic categories, although it appears that speech is organized at both phonemic and phonetic levels and that both levels significantly influence L2 production (Flege and Eefting, 1987 in press).

2 L2 production studies often examine phonetically constrained speech material, usually key words produced in a constant carrier phrase. This might tend to evoke a style of speaking that is not representative of natural, spontaneous speech. Flege and Hillenbrand (1984) showed that it is possible to obtain a speech sample that appears to resemble spontaneous speech, yet offers the phonetic control needed to make systematic comparisons. The subjects in that study produced a spontaneous story based on a number of phrases, such as "Two little boys". They were required to use the phrases to initiate a sentence in the story. The study focused on the VOT interval in /t/. Thus, the corpus provided a suitable number of tokens of /t/ in the same phonetic context (unstressed, utterance-initial /t/ in the context of /u/).

3 Flege and Eefting (1987 to appear) hypothesized that the opposite pattern of L2 effects on L1 production observed in the two studies just mentioned could be traced to the age at
which subjects began learning their L2. The “early L2 learners” examined by Flege and Eefting (1987) began learning their L2 (Spanish) in early childhood. The “late L2 learners” examined by Flege (1987) began learning their L2 (English or French) in adulthood. It was hypothesized that the basis for the difference between studies was that early, but not late, L2 learners establish phonetic categories for corresponding stops that are realized differently in L1 and L2.

4 The standard deviations presented by Fourakis and Iverson (1985) are not included here because they seem to have been based on the 32 values obtained for the four subjects in each group, rather than on the four mean values obtained for subjects in each group. This may have spuriously reduced intragroup variability, and, thus, led to inappropriately high F-values in the ANOVAs.

5 During production of an initial stop, the tongue tends to anticipate the position required for a following vowel. If a tight tongue-palate constriction exists when a stop consonant is released, as would be expected when the following vowel is /i/ or /u/, it will take somewhat longer for the transglottal pressure difference needed for voicing to be reestablished, than it would if a tight constriction did not exist, as would be expected, if the following vowel were /o/ or /a/.

REFERENCES

Flege, J. E. and J. Hillenbrand 1985. Differential use of temporal cues to the [s-z] contrast


