

Learning English vowels with different first-language vowel systems: Perception of formant targets, formant movement, and duration

Paul Iverson^{a)} and Bronwen G. Evans

Department of Phonetics and Linguistics, University College London, 4 Stephenson Way, London NW1 2HE, United Kingdom

(Received 5 December 2006; revised 20 August 2007; accepted 20 August 2007)

This study examined whether individuals with a wide range of first-language vowel systems (Spanish, French, German, and Norwegian) differ fundamentally in the cues that they use when they learn the English vowel system (e.g., formant movement and duration). All subjects: (1) identified natural English vowels in quiet; (2) identified English vowels in noise that had been signal processed to flatten formant movement or equate duration; (3) perceptually mapped best exemplars for first- and second-language synthetic vowels in a five-dimensional vowel space that included formant movement and duration; and (4) rated how natural English vowels assimilated into their L1 vowel categories. The results demonstrated that individuals with larger and more complex first-language vowel systems (German and Norwegian) were more accurate at recognizing English vowels than were individuals with smaller first-language systems (Spanish and French). However, there were no fundamental differences in what these individuals learned. That is, all groups used formant movement and duration to recognize English vowels, and learned new aspects of the English vowel system rather than simply assimilating vowels into existing first-language categories. The results suggest that there is a surprising degree of uniformity in the ways that individuals with different language backgrounds perceive second language vowels. © 2007 Acoustical Society of America. [DOI: 10.1121/1.2783198]

PACS number(s): 43.71.Hw, 43.71.Es [ARB]

Pages: 2842–2854

I. INTRODUCTION

It is clear that one's first-language (L1) phonetic categories affect second-language (L2) vowel learning. For example, Spanish listeners have difficulty learning to discern the difference between English /i/ and /ɪ/ (e.g., Escudero and Boersma, 2004; Flege *et al.*, 1997; Morrison, 2002), presumably because they both sound like the same Spanish vowel (/i/). In contrast, German listeners have less difficulty learning English /i/-/ɪ/ (Bohn and Flege, 1990; Flege *et al.*, 1997), presumably because they sound like two different German vowels (/i/ and /ɪ/). These types of L1/L2 interactions have been well established at the level of individual phonetic categories, but there has been little research on whether these interactions have broader implications for how individuals learn entire vowel systems. The present study addresses this issue by comparing how individuals with a range of L1 vowel systems (Spanish, French, German, and Norwegian) learn English vowels.

The task of learning an L2 vowel system may be fundamentally different for individuals whose L1 vowel system is large and complex (e.g., Norwegian) than for individuals whose L1 vowel system is small and simple (e.g., Spanish). One possibility is that individuals could take the cues used in their L1 vowel system and apply them to learning an L2, which could be an advantage to listeners with more complex L1 vowel systems (i.e., those that use more cues). For ex-

ample, L1 English speakers are better than L1 Spanish speakers at learning vowel length contrasts in Swedish, and this may occur because English vowels vary more systematically in duration than do Spanish vowels (MacAllister *et al.*, 2002). English speakers also use duration when learning French /ɔ/-/o/, even more than do L1 French speakers (Gottfried and Beddor, 1988). Likewise, Japanese has a vowel duration contrast, and individuals appear to apply this to learning English such that they primarily use duration to distinguish English /i/-/ɪ/ (Morrison, 2002) as well as using duration to contrast stressed and unstressed vowels within words (Lee *et al.*, 2006). However, the opposite pattern of results is sometimes found; individuals with no L1 vowel duration contrasts (e.g., Spanish and Catalan) often still use duration to distinguish English /i/-/ɪ/ (Cebrian, 2006; Escudero and Boersma, 2004; Morrison, 2002). A reliance on duration when learning L2 vowels may simply be a strategy that is often used when listeners have difficulty discerning a spectral difference (Bohn, 1995; Bohn and Flege, 1990), regardless of whether their L1 contrasts vowel duration (see Flege *et al.*, 1997).

There has been almost no work on whether the use of formant movement (i.e., both diphthongs and intrinsic formant movement for monophthongs) transfers between one's L1 and L2; Italians are better able to learn formant movement for English /e/ if they begin learning at younger ages (Flege *et al.*, 2003), but it is unknown whether the use of formant movement in one's L1 makes learning formant movement in an L2 easier or harder. Even among English

^{a)}Corresponding author.

monophthongs, formant movement has been shown to be an important cue for native listeners; recognition accuracy can decline by 13–23 percentage points when formant movement is flattened in synthesized or signal processed speech (e.g., Assmann and Katz, 2005; Hillenbrand and Nearey, 1999; Iverson *et al.*, 2006). The reliance on such acoustic information has important theoretical implications, because it suggests that listeners may have phonetically detailed category representations for vowels (i.e., exemplars; e.g., Goldinger, 1996, 1998; Hawkins and Smith, 2001; Johnson, 1997; Nygaard *et al.*, 1995, Nygaard and Pisoni, 1998; Pisoni, 1997), rather than having more abstract representations based only on the primary acoustic cues.

In addition to the potential effects of L1 cues, the sheer number of vowels in an L1 may have implications for L2 vowel learning. Novice L2 learners are thought to use their existing L1 categories when listening to the L2 (i.e., they assimilate the L2 vowels into L1 categories; Best, 1995; Best *et al.*, 2001; Flege, 1995, 2003). This L1 assimilation strategy could be problematic for individuals with small L1 vowel systems, because it is more likely that there will be cases of multiple L2 vowels assimilating to the same L1 category (e.g., English /i/ and /ɪ/ assimilating to Spanish /i/), making it harder for listeners to discern differences among these L2 vowels. Despite this initial difficulty, the small L1 inventory may make it easier for individuals to learn. That is, Flege (1995) has argued that new categories are easier to learn when they are far away from existing categories, and one could imagine that individuals with smaller vowel systems would have more unused room in the vowel space to learn new categories (although it is not clear that individuals with smaller vowel systems actually have more unoccupied space; see Meunier *et al.*, 2003).

Individuals with larger L1 vowel systems may be more successful in using assimilation (i.e., less chance of multiple L2 vowels assimilating to the same L1 category), but they may have more difficulty learning new categories. If individuals with large vowel systems have less unoccupied space to learn new vowels, they would need to change their existing L1 category representations to better match the L2 vowels, creating merged or compromise categories (Flege, 2003; MacKay *et al.*, 2001). Changing existing categories in this way is thought to be more difficult than learning entirely new categories (Flege, 1995, 2003; Munro *et al.*, 1996). It is thus possible that individuals with larger L1 vowel systems may rely more on L1 assimilation and less on new learning than do individuals with smaller L1 vowel systems.

The present study investigated whether L1 speakers of Spanish, French, German, and Norwegian fundamentally differ in the cues that they use when listening to English vowels. Spanish has five vowels (/i/, /e/, /a/, /o/, and /u/), and duration is not used contrastively (Stockwell and Bowen, 1965; Flege, 1989). The status of formant movement is less clear; Spanish seems to lack true diphthongs (i.e., single phonemes marked by the movement between two vowel positions), but monophthongal vowels can occur consecutively in Spanish and are sometimes considered to be diphthongs (Stockwell and Bowen, 1965; Delattre, 1965). French has eleven oral monophthongal vowels (/i/, /y/, /e/, /ø/, /ɛ/, /œ/,

/a/, /ɑ/, /ɔ/, /o/, and /u/), and four nasal vowels (\tilde{a} , \tilde{o} , $\tilde{ɛ}$, / $\tilde{œ}$ /; the present study focuses only on the oral vowels). French has no diphthongs and duration contributes negligibly to vowel distinctions (Delattre, 1965). German has 15 monophthongal vowels that form seven long-short (tense-lax) vowel pairs (/i/-/ɪ/, /e/-/ɛ/, /u/-/ʊ/, /o/-/ɔ/, /y/-/ʏ/, /ø/-/œ/, and /a/-/a:/, plus /ɛ:/), and three diphthongal vowels (/ai/, /aʊ/ and /ɔɪ/; Delattre, 1965; Strange *et al.*, 2005). Norwegian has 18 monophthongal vowels that form nine long-short (tense-lax) pairs (/i:/-/i/, /y:/-/y/, /e:/-/ɛ/, /ø:/-/œ/, /æ:/-/æ/, /ɑ:/-/ɑ/, /ɔ:/-/ɔ/, /u:/-/u/, and /ʉ:/-/ʉ/), and four diphthongs (/ai/, /ei/, /æʉ/ and /øʉ/; see Kristoffersen, 2000). These L1s thus vary both in terms of the number of vowels and in terms of the cues that are used.

We gave subjects a battery of tests to evaluate their vowel recognition and perceptual category representations. The baseline recognition ability of subjects was evaluated by having them identify natural recordings of English /b/-V-/t/ words. Two tests evaluated the subjects' representation of target (static) formant frequencies, formant movement, and duration. Listeners identified signal-processed natural vowels in noise, in order to examine whether flattening formant movement or equating duration affected recognition for L2 English speakers (see Iverson *et al.*, 2006). Listeners also mapped their perceptual vowel spaces (best exemplar locations) in both their L1 and L2, within a five-dimensional acoustic space that included F1 and F2 at the onset and offset of each vowel, and duration (see Iverson and Evans, 2003; Iverson *et al.*, 2006). Finally, listeners rated how natural English vowels assimilated into L1 categories. Our aims were to: (1) examine to what extent the representation (measured by best exemplars) of target formant frequencies, formant movement, and duration are able to predict the ability of L2 learners to recognize natural English vowels; (2) whether the representation of these cues varies between language groups (e.g., whether individuals more accurately represent L2 duration when their L1 contrasts duration); and (3) whether these representations are due to category assimilation (i.e., using an existing L1 category when listening to the L2) or new learning.

Our L1 subject groups were selected to be relatively heterogeneous (e.g., in terms of English experience) in order to increase individual variation in vowel recognition accuracy. For example, we examined which aspects of English vowel categorization (e.g., representation of duration) correlated with individual differences in English vowel identification accuracy within each L1 group, rather than focusing solely on between-group differences. Although it is more common in the literature to closely control groups for experience, experience is only one of the many factors that can determine whether an individual is good or poor at recognizing L2 vowels (e.g., motivation, aptitude, and type of experience are also important). Our research strategy was to take advantage of this individual variability in order to understand the vowel recognition process, rather than treat it as a confound that should be removed.

II. METHOD

A. Subjects

A total of 114 subjects were tested. Five of these were omitted because of missing data (computer problems or unable to complete). Subjects were additionally screened based on their L1 tests (see Procedure); three were omitted because their L1 signal-to-noise threshold for vowel recognition was more than 2 SD above the average (indicating a possible hearing problem), and six were omitted because their L1 best exemplars were more than 2 SD different from the average (indicating an inability to reliably perform the task). After screening, the numbers of subjects within each L1 group were 25 for Spanish, 19 for French, 21 for German, 18 for Norwegian, and 17 for English. All subjects reported that they had no hearing or language impairments.

Spanish speakers were all tested in London. We tested speakers from a range of countries (nine Spain, five Colombia, three Peru, two Chile, two Cuba, and one each from Argentina, Mexico, Uruguay, and Venezuela), but all had a standard Spanish five-vowel system. The subjects were 21–45 years old (median 29 years). Subjects began learning English when they were 4–24 years old (median 13 years). The subjects had 0–4 years of experience living in English-speaking countries (median 1 year).

French speakers were all tested in London, and all subjects grew up in France. The subjects were 21–55 years old (median 29 years). Subjects began learning English when they were 2–14 years old (median 11 years). The subjects had 0–17 years of experience living in English-speaking countries (median 3 years).

German speakers were tested in Potsdam, Germany (11 subjects) and London (ten subjects); all subjects grew up in Germany. We did not test any subjects who had nonstandard German vowel systems (e.g., Bavarian accent). The subjects were 19–64 years old (median 28 years). Subjects began learning English when they were 6–15 years old (median 11 years). The subjects had 0–30 years of experience living in English-speaking countries (median 2 years).

Norwegian speakers were tested in Trondheim, Norway (15 subjects) and London (three subjects); all subjects grew up in Norway. The subjects were 20–35 years old (median 22 years), and subjects began learning English when they were 8–10 years old (median 9 years). The subjects had 0–6 years of experience living in English-speaking countries (median 0 years).

English speakers were tested in the United Kingdom. All were monolingual and grew up in England. They were 18–49 years old (median 28 years). The English speakers completed only a subset of the tasks to provide normative data (English speech in noise and English vowel-space mapping).

B. Stimuli and apparatus

All subjects were tested in quiet rooms, with stimuli played over headphones at a user-controlled comfortable level. PCs (desktops, laptops, and pocket PCs) were used to play the stimuli and collect responses. Stimulus recordings

were made in an anechoic chamber, with 44,100 16-bit samples/s, and later down sampled to 11,025 samples/s.

1. Natural /b/-V-/t/ recordings

We recorded a single speaker for each language, and used this same speaker for all tasks (except for English speech in noise). This was designed to facilitate across-task comparisons of the results (i.e., eliminate variability due to talker differences). All speakers were male, and were native speakers of their respective L1. A /b/-V-/t/ context was used (/b/-V-/tɑ/ for Spanish) because this was phonologically legal in all of our languages. This context created nonwords in several of the languages, so speakers were also given common real words with that vowel to help illustrate what should be said. The English speaker recorded the words *beat* /i/, *bit* /ɪ/, *bet* /ɛ/, *Burt* /ɜ/, *bat* /ɑ/, *Bart* /ɑ/, *bot* /ɒ/, *but* /ʌ/, *bought* /ɔ/, *boot* /u/, *bait* /eɪ/, *bite* /aɪ/, *bout* /aʊ/, and *boat* /əʊ/, in the sentence *Say___again*; English vowels that would create nonwords in the /b/-V-/t/ context (e.g., /ʊ/) were not included in the study. The Spanish speaker recorded the words and nonwords *bita* /i/, *beta* /e/, *bata* /a/, *bota* /o/, and *buta* /u/ in the sentence *Digo la palabra___de nuevo*. The French speaker recorded the words and nonwords *bit* /i/, *but* /y/, *bête* /ɛ/, *beute* /ø/, *bête* /ɛ/, *boeute* /œ/, *batte* /a/, *bâte* /ɑ/, *botte* /ɔ/, *bôte* /o/, and *bout* /u/, in the sentence *Je dis___encore*. The German speaker recorded the words and nonwords *biet* /i/, *büüt* /y/, *bitt* /ɪ/, *bütt* /y/, *beet* /e/, *bööt* /ø/, *bett* /ɛ/, *bäüt* /ɛ:/, *bött* /œ/, *batt* /a/, *bad* /a:/, *bott* /ɔ/, *boot* /o/, *butt* /ʊ/, *buud* /u/, *beit* /aɪ/, *baut* /aʊ/, and *beut* /ɔɪ/ in the sentence *Sag___nochmal*. The Norwegian speaker recorded the words and nonwords *bit* /i:/, *byt* /y:/, *bitt* /ɪ/, *bytt* /y/, *bet* /e:/, *bøt* /ø:/, *bett* /ɛ/, *bøtt* /œ/, *bætt* /æ/, *bæt* /æ:/, *batt* /a/, *bat* /a:/, *bått* /ɔ/, *båt* /o:/, *bott* /ʊ/, *butt* /ʌ/, *bot* /u:/, *but* /ʌ:/, *bait* /aɪ/, *beit* /ɛɪ/, *baut* /æʌ/, and *bøyt* /øɪ/ in the sentence *Det var___jeg sa*. The speakers read each word individually off of a computer screen, in random order. Each word was recorded four times, and was screened for intelligibility. Speakers also read a short passage (Aesop's *The north wind and the sun*, translated into each language).

In order to facilitate cross-language comparisons (particularly for vowel-space mapping), signal processing was used to equate the formant frequencies and F0 of each talker in the carrier sentences and in the paragraph (e.g., to eliminate differences in vowel quadrilaterals related to vocal tract length). The signal processing followed the Praat *Change Gender* command (Boersma and Weenink, 2005), except that the processing stages were applied individually so that the pitch pulse analyses could be hand corrected. Specifically, a new sample rate was imposed on the stimuli to scale the formants (e.g., the sample rate could be changed from 22,050 to 24,255 if one needed to raise the formant frequencies by 10%), and then pitch synchronous overlap and add (PSOLA) was applied to scale the F0 and duration to correct for changes introduced by the sample rate change, as well as to equate F0 between talkers. The formant frequencies were scaled so that the F2 of the /i/ produced by each talker was 2290 Hz (an average male value; Peterson and Barney, 1952). The F2 of /i/ was selected because it is consistently produced across talkers and is easy to measure accurately.

The F0 was scaled so that the median (as measured from the short passage) was 112 Hz (an average male value; Hazan and Markham, 2004). Across talkers, there were only small changes made to the formant frequencies (maximum 8% change) and F0 (maximum 15 Hz change), so the effects of signal processing were thus subtle.

2. Vowel-space mapping

A large set of synthesized stimuli were created to map best exemplars. The stimuli were embedded in the natural carrier sentences, including the /b/ burst and the /t/ stop gap from the natural recordings. The stimuli were created using the cascade branch of a Klatt synthesizer (Klatt and Klatt, 1990). For each language, the synthesis parameters were chosen so that the synthesized vowel approximated the original vowel in the natural carrier sentence in terms of F0 and amplitude contours. All other synthesis parameters were the same for each language. The upper-formant frequencies (F4=3500 Hz, F5=4500 Hz), formant bandwidths (100, 180, 250, 300, and 550 Hz for F1–F5), tilt (0 dB slope), and open quotient (60%) were the same for all stimuli in all languages. The stimuli primarily varied F1, F2, and duration. F3 varied as a function of F2; F3 was fixed at 2500 Hz whenever F2 was less than 2300 Hz, but otherwise F3 was raised so that it was always 200 Hz greater than F2.

The F1 and F2 formant frequencies changed linearly from the beginning to the end of the vowel, and there were no additional consonantal formant transitions. F1 frequency was restricted so that it had a lower limit of 5 equal rectangular bandwidth (ERB) (Glasberg and Moore, 1990) and an upper limit of 15 ERB. F2 frequency was restricted so that it had a lower limit of 10 ERB, was always at least 1 ERB higher than F1, and had an upper limit defined by the equation $F2 = 25 - (F1 - 5)/2$. The stimuli were synthesized in advance with a 1-ERB spacing of the vowel space, and with seven log-spaced levels of duration (54, 75, 104, 144, 200, 277, and 383 ms), for a total of 109,375 individual stimuli for each language. The ERB and log-duration transforms allowed us to efficiently distribute the stimuli with regard to perception, although the goodness optimization procedure does not require this equal perceptual spacing.

As discussed above, different natural carrier sentences had been signal processed to equate formant frequencies and F0 across the talkers of the different languages (see description above of the natural /b/-V-/t/ recordings), with the goal of making sure that the perceptual vowel space maps would not differ across languages due to physiological differences between the talkers' voices (see Ladefoged and Broadbent, 1957). To test whether this was successful, a pilot experiment was run in which six L1 English speakers chose best exemplars for *beat*, *Burt*, and *bat* in the carrier sentences for every language (subjects were asked to ignore the fact that the carrier sentences were not always in English). The procedure was the same as for the vowel space mapping procedure described below. Repeated-measures analyses of variance (ANOVA) revealed that there was no significant effect of sentence context for any of the acoustic dimensions (F1 and F2 at the onset and offset, and duration), $p > 0.05$. This demonstrates that the acoustic normalization was successful, and

that the best exemplars chosen in different languages in this experiment could thus be directly compared.

3. English speech in noise

The stimuli were from a previous study (Iverson *et al.*, 2006), recorded from a different British English talker (female) than used in the rest of this study. The speaker was recorded saying *Say hVd again* with 11 words: *heed* (/i/), *hid* (/ɪ/), *head* (/ɛ/), *had* (/ɑ/), *heard* (/ɜ/), *hud* (/ʌ/), *hod* (/ɒ/), *hard* (/ɑ/), *hoard* (/ɔ/), *hood* (/u/), and *who'd* (/u/). Diphthongs were not included because they become unintelligible when formant movement is removed. Two additional versions of the vowels were created that: (1) removed all formant movement, and (2) equated duration. The changes to the stimuli were made using Praat (Boersma and Weenink, 2005). Formant movement was removed using linear predictive coding (LPC) analysis and resynthesis. Specifically, LPC analyzed the signal from the start of voicing after the /h/ to the start of the /d/ closure; the signal was inverse filtered to produce an LPC residual; a time slice of the LPC analysis was identified that represented the vowel's target formant frequencies (defined as the point where F1 reached a peak); and this single LPC slice was used to filter the entire LPC residual. This process created stimuli that retained the natural F0 of the original stimuli, but had formant frequencies that remained fixed at each vowel's target values. Duration was equated using PSOLA, such that the durations of the /h/, the /d/ closure, and the vowel were set to the mean values for the talker.

The speech-shaped noise conformed to CCITT Rec. G227 and was produced by a Wandell and Goltermann RG-1 noise generator. The signal-to-noise ratio (SNR) was calculated for each individual stimulus, by comparing the RMS amplitude of the stimulus and noise.

C. Procedure

1. English vowel identification in quiet

Subjects heard natural recordings of the English speaker and gave a closed-set identification response (all 14 words as response options). To give their response, they mouse clicked on a button which listed the stimulus word (e.g., *bot*) as well as a common English word that had the same vowel (e.g., *hot*). Prior to starting the experiment, they heard the speaker read *The North Wind and the Sun* passage. They were shown the word response alternatives and were able to ask questions if they were unsure which vowels were indicated. Subjects identified four repetitions of each vowel, for a total of 56 trials in a random order.

2. English vowel identification in noise

The task was the same as for identification in quiet, except that speech-shaped noise was added to the stimuli adaptively to find the SNR level that yielded 50% correct responses (i.e., 1-up/1-down Levitt procedure; Levitt, 1971). The adaptive series were blocked by the three stimulus conditions (natural speech, no formant movement, equated duration), and there were three blocks for each stimulus condition (i.e., a total of nine blocks). The task began with a

+10 dB SNR level. At the start of the experiment (prior to three reversals) the level was reduced by 6 dB after each correct response and increased by 6 dB after each incorrect response. Afterwards the level was changed in 2 dB steps. The procedure was stopped after seven reversals. The SNR was calculated for each condition by averaging the last four reversals of each block (when the step size was 2 dB), and taking the median of the values for the three blocks. The SNR was classified as undefined when there was more than one adaptive series that was aborted due to repeatedly reaching the upper limit (+10 dB SNR) of the search (i.e., indicating poor recognition even with minimal noise).

3. L1 vowel identification in noise

The procedure was identical to English vowel identification in noise, except that the words were from the listener's L1 and only natural stimuli were used. This task served as a hearing screening for subjects; subjects were omitted when their SNR thresholds were more than 2 SD above the average for their L1 group.

4. L1 assimilation

Subjects heard the English /b/-V-/t/ words, and identified which of their own L1 /b/-V-/t/ words sounded closest to the word that they heard. They were told to imagine that they were listening to an L1 English speaker who was trying to learn to speak their language. After each identification, they mouse clicked on a graphical continuum to rate whether this stimulus was *close* or *far away* from this L1 vowel category. Subjects completed 28 trials (two repetitions of 14 English /b/-V-/t/ words).

5. Vowel-space mapping

In separate experiments, subjects found best exemplars for /b/-V-/t/ words in both English and their L1. Subjects first heard the speaker read *The North Wind and the Sun* in the respective language to familiarize them with the talker. On each trial, subjects saw a /b/-V-/t/ word on the computer screen (e.g., *bot*), as well as a more common word that had the same vowel (e.g., *hot*), and heard a stimulus (synthesized /b/-V-/t/ embedded in a natural carrier sentence). They rated on a continuous scale how far away the /b/-V-/t/ that they heard was from being a good exemplar of the printed word. Their ratings were given by mouse clicking on a continuous bar presented on a computer screen.

A goodness optimization procedure (Evans and Iverson, 2004, 2007; Iverson and Evans, 2003; Iverson et al., 2006) was used to iteratively change the stimuli that subjects heard on each trial, to search through the multidimensional stimulus space for better exemplars. Estimates of best exemplar locations were able to be found after 35 trials per vowel, despite the large stimulus set (109,375 stimuli). The procedure involved searching along individual vectors (i.e., one-dimensional straight-line paths crossing through the five-dimensional stimulus space), and finding the best exemplar on each vector. There were a total of seven search vectors and five trials per vector for each vowel. The vectors were chosen so that Vector 1 would allow most subjects to find a

close approximation of their best exemplar (the search path passed through frequencies measured from natural productions), Vectors 2–6 orthogonally varied the five acoustic dimensions over a wide range, and Vector 7 fine tuned the position of the best exemplar.

Specifically, Vector 1 was a straight-line path that passed through two points: (1) the F1 and F2 frequencies at the beginning and ending of the natural productions of the target vowel, and (2) a neutral stimulus in the middle of the vowel space (F1=500 Hz and F2=1500 Hz, at both the onset and offset); duration was not varied along this vector, so each of the points was thus defined by four frequency values. Vector 2 varied duration, keeping formant frequencies fixed. Vector 3 varied the onset F1 and F2 frequencies (i.e., duration and offset formant frequencies were fixed) along the same basic path as the first vector (i.e., through a straight-line path including a neutral vowel and the onset formant frequencies of the natural production). Vector 4 was orthogonal to Vector 3 in the F1/F2 onset space. Vectors 5 and 6 were analogous to Vectors 3 and 4, except that the offset F1 and F2 frequencies were varied. Vector 7 varied all dimensions, passing through the best value found thus far on each dimension and the neutral vowel.

The endpoints of all vectors were constrained by the boundaries of the vowel space. For example, Vector 1 for *beat* crossed diagonally across the vowel space, starting from the high-front boundary of the space (i.e., low F1 and high F2), passing through the middle of the space, and ending at the low-back boundary of the space (i.e., high F1 and low F2). That is, the search spanned the entire space so that listeners had freedom to choose whatever acoustic values that they thought were the best, rather than being constrained to stimuli near *beat*.

The best exemplars were found for each vector over five trials. On the first two trials, subjects heard the most extreme stimuli that it was possible to synthesize along the vector (e.g., in the case of *beat*, they heard extreme high-front and low-back vowels, with the order of these two trials randomized). The selection of stimuli on the remaining trials was based on the subjects' judgments, using formulas that were designed to find stimuli along the path that would be perceived as better exemplars. On the third trial, subjects heard a stimulus that was selected by a weighted average of the first two stimuli, according to the equation

$$c = a \frac{f(b)}{f(a) + f(b)} + b \frac{f(a)}{f(a) + f(b)}, \quad (1)$$

where a and b are the positions on the search path for the first two trials; $f(a)$ and $f(b)$ are the goodness ratings for the stimuli on those trials (the goodness responses of *close* to *far away* were scaled from 0 to 1); and c is the new path position selected for the third trial. On the fourth and fifth trials, the stimuli were selected by finding the minimum of a parabola that was defined by the equation

$$\min = \frac{b - 0.5 \{ [b - a]^2 [f(b) - f(c)] - [b - c]^2 [f(b) - f(a)] \}}{[b - a]^2 [f(b) - f(c)] - [b - c]^2 [f(b) - f(a)]}, \quad (2)$$

where b is the path position of the best stimulus found thus far; a and c are the most recently tested positions on either

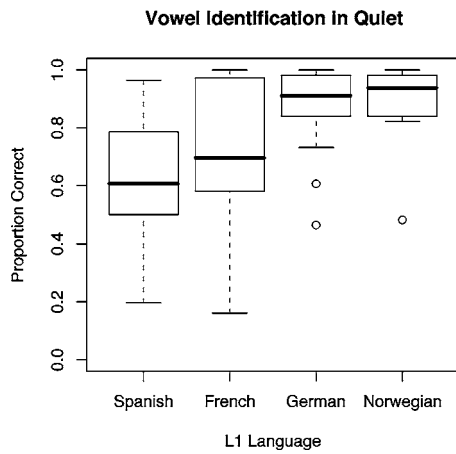


FIG. 1. Accuracy of natural English vowel identification in quiet for the four language groups. The boxplots display the distribution of individual differences (quartile ranges), with outliers indicated by circles (i.e., points that would otherwise make the whiskers longer than 1.5 times the interquartile range).

side of b ; and $f(a)$, $f(b)$, and $f(c)$ are the goodness ratings for those stimuli. At the completion of the fifth trial, subjects were allowed to repeat the search if it had produced a poor exemplar (i.e., they made an explicit judgment about whether the stimulus was or was not close to sounding good). If the best exemplar was judged to be close, the parameters of the best stimulus found thus far were passed onto the next stage of the search algorithm (i.e., to search along the next vector).

The searches for the different words were interleaved. Specifically, the search progressed stage by stage, such that listeners completed Vector 1 for all words, then completed Vector 2 for all words, etc., with the word order randomized for each vector. Listeners thus switched vowel categories relatively frequently (they had five trials in a row with the same vowel) rather than repeatedly making judgments on single vowels.

III. RESULTS AND DISCUSSION

A. English vowel identification in quiet

Figure 1 displays the accuracy with which each L1 group identified English vowels in quiet. There were obvious differences between the groups; Germans and Norwegians were more consistently accurate than Spanish and French listeners. A one-way ANOVA on arcsine-transformed scores demonstrated that this effect of L1 was significant, $F(3, 79)=10.20$, $p < 0.001$.

B. English vowel identification in noise

Figure 2 displays the threshold SNR values for each group of listeners (i.e., the level at which English vowels were identified 50% correct). A repeated-measures ANOVA examined the effect of L1 and stimulus condition (normal, no formant movement, and no duration contrast). There was a significant main effect of L1, $F(4, 91)=23.39$, $p < 0.001$, with Spanish and French speakers having poorer thresholds than German, Norwegian, and English speakers (i.e., the same pattern as for identification in quiet). There was also a significant main effect of condition, $F(2, 184)=12.35$, p

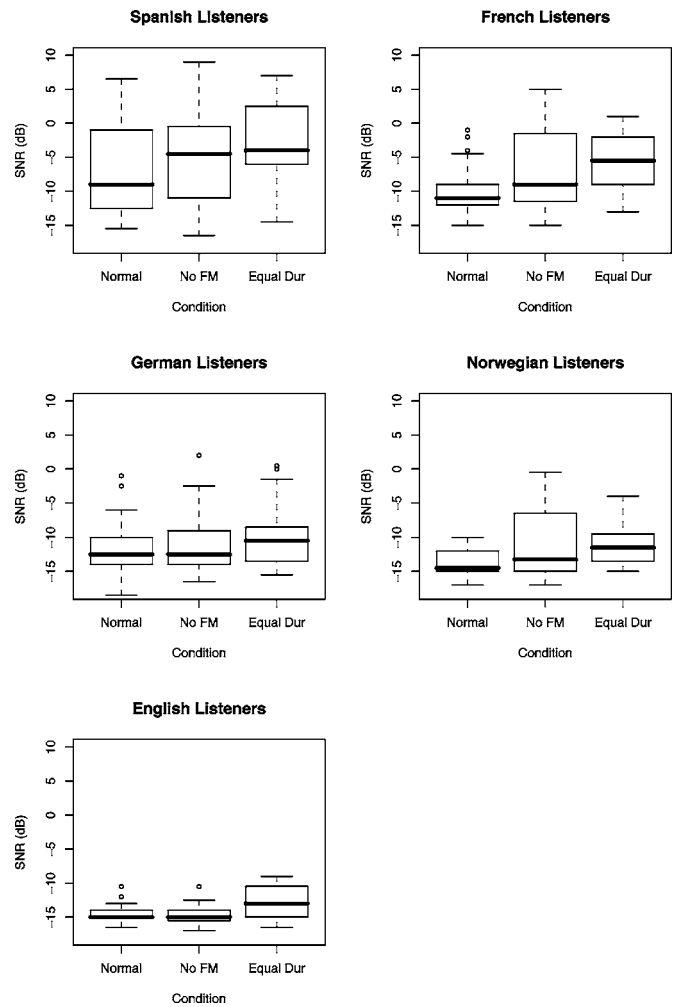


FIG. 2. Boxplots of the SNR thresholds (50% identification accuracy) for the different language groups, for natural vowels (Normal), vowels in which the formant movement had been flattened (No FM), and vowels in which duration had been equated (Equal Dur).

< 0.001 ; across language groups, SNR thresholds were raised by an average of 1.5 dB when formant movement was flattened and 2.6 dB when duration was equated. There was no significant interaction between L1 and condition, $p > 0.05$. Figure 2 makes it appear as if the median change in thresholds may have varied for the different language groups (e.g., Spanish speakers had larger differences in terms of the medians), but the between-language differences were small compared to the individual variability.

The results thus indicate that the L1 groups use formant movement and duration to similar extents when recognizing English vowels. Although the differences may seem small, a 1 dB change in the SNR translates into about a 7 percentage-point change in recognition accuracy (estimated by inspection of the psychometric functions for the present experiment). Thus, flattening formant movement reduced recognition accuracy by about 10 percentage points and equating duration reduced recognition accuracy by about 18 percentage points, which is comparable to the respective 13 and 14 percentage-point reductions that we found previously with L1 English speakers who had cochlear implants (Iverson *et al.*, 2006).

average best exemplars for L1 English subjects. These distances were calculated separately for F1/F2 location, formant movement, and duration. The F1/F2 location accuracy was measured by averaging the beginning and ending frequency of each vowel for F1 and F2, giving a two-dimensional F1/F2 coordinate for that vowel with no formant movement. The Euclidean distance (i.e., root mean square) was then calculated between the F1/F2 locations of each individual's English best exemplars and the L1 English averages. Formant movement accuracy was measured by subtracting the F1/F2 location values above, so that each vowel was represented as a vector representing the direction and magnitude of F1/F2 formant movement, with the center of each line passing through zero (i.e., normalizing the vowel's location in the vowel space). As above, Euclidean distances between these formant movement vectors were measured for each individual's vowels and the L1 English averages. Duration accuracy was quantified by calculating the average absolute-value difference between the durations of each individual's best exemplars and those of the L1 English averages.

The accuracy measures are displayed in Fig. 4. Accuracy in best exemplar locations had the same basic pattern as identification accuracy for English vowels in quiet, with Norwegians and Germans being more accurate than French or Spanish speakers. Separate one-way ANOVAs confirmed that the L1 groups were significantly different in terms of F1/F2 location accuracy, $F(4, 95)=10.85$, $p<0.001$, formant movement accuracy, $F(4, 95)=8.98$, $p<0.001$, and duration accuracy, $F(4, 95)=9.12$, $p<0.001$.

Pearson correlations compared the individual differences in vowel-space accuracy measures with arcsine-transformed identification accuracy for English vowels in quiet. F1/F2 location was significantly correlated with identification across language groups, $r=-0.63$, $p<0.001$, and within each language group: Spanish, $r=-0.65$, $p<0.001$; French, $r=-0.49$, $p=0.034$; German, $r=-0.62$, $p=0.002$; and Norwegian, $r=-0.54$, $p=0.022$. Formant movement accuracy was significantly correlated with identification across language groups, $r=-0.54$, $p<0.001$, and within language groups for Germans, $r=-0.65$, $p=0.001$, and Norwegians, $r=-0.53$, $p=0.024$; the correlations were not significant within the Spanish, $r=-0.36$, and French, $r=-0.32$, language groups, $p>0.05$. Duration accuracy was significantly correlated, although weakly, with identification across language groups, $r=-0.27$, $p=0.014$, but was not significant within the Spanish, $r=-0.13$, French, $r=-0.39$, German, $r=-0.37$, and Norwegian, $r=0.01$, language groups, $p>0.05$.

Separate ANOVAs tested whether these relationships between vowel-space accuracy and identification differed between the L1 groups; arcsine-transformed identification accuracy was the dependent measure, L1 was an independent factor, and each vowel-space measure was entered in separate analyses as a covariate. There were no significant interactions of L1 group with F1/F2 location accuracy, formant movement accuracy, or duration accuracy, $p>0.05$. This suggests that the relationships between the accuracy measures and identification were similar for each L1. Confirming our previous analyses, in each of these ANOVAs there were

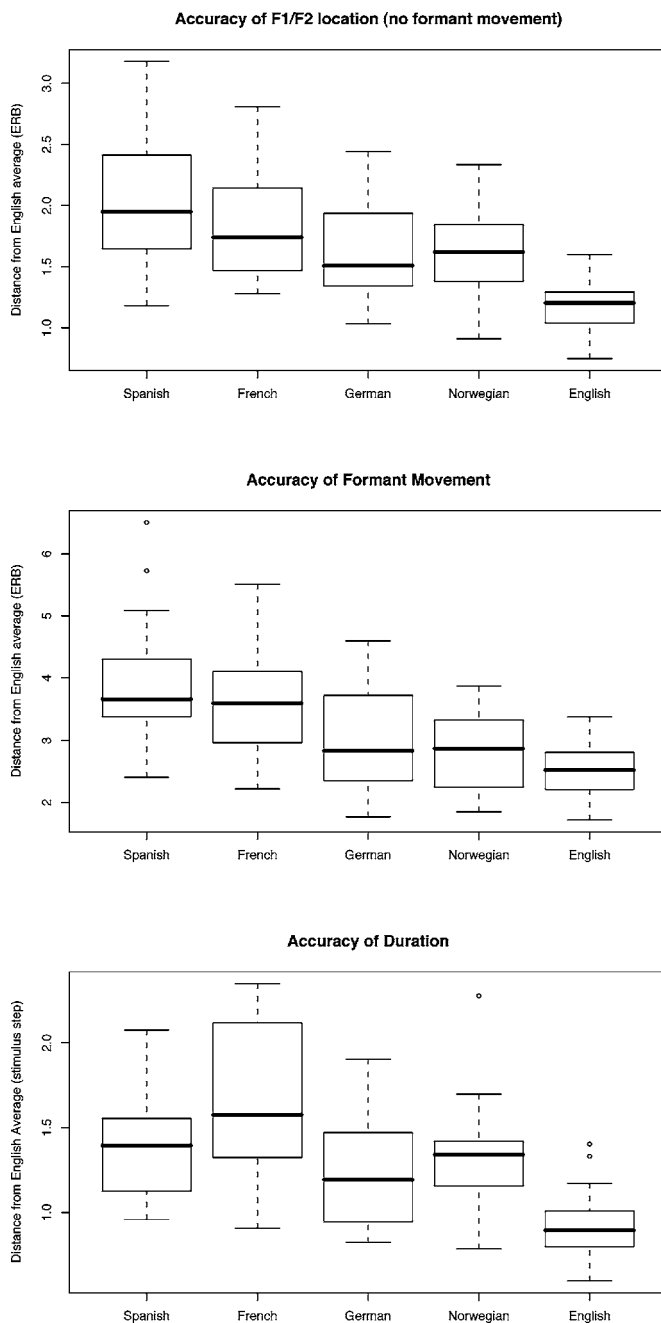


FIG. 4. Boxplots of the accuracy of each language group's English best exemplars (i.e., distance from the averages of native English speakers), in terms of F1/F2 location (i.e., static target frequencies), formant movement (i.e., direction and magnitude of change in F1 and F2 during each vowel), and duration.

significant main effects of L1 [F1/F2 location ANOVA: $F(3, 75)=14.25$, $p<0.001$; formant movement accuracy ANOVA: $F(3, 75)=11.98$, $p<0.001$; and duration accuracy ANOVA: $F(3, 75)=10.72$, $p<0.001$], and each accuracy measure [F1/F2 location: $F(3, 75)=34.78$, $p<0.001$; formant movement accuracy: $F(3, 75)=16.26$, $p<0.001$; and duration accuracy: $F(3, 75)=5.53$, $p=0.021$].

The results from the L2 English vowel spaces are thus in accord with those of English vowel identification in noise. That is, there were large differences in accuracy between L1 groups, but their reliance on F1/F2 location, formant movement, and duration was similar. Despite this commonality, it

should be noted that individual differences in the vowel-space accuracy measures were not always strongly correlated. For example, the correlation between F1/F2 location and duration accuracy across language groups was $r=0.30$, $p=0.004$, and the correlation between F1/F2 location and formant movement accuracy was $r=0.55$, $p<0.001$. The magnitude of these correlations leaves most of the variance unexplained, suggesting that individuals had idiosyncratic patterns of cue use (e.g., being accurate at duration but poor at F1/F2 location, rather than being equally accurate with all cues). That is, individual differences in cue use exist, but they do not appear to be strongly related to L1 background.

D. L1 assimilation and L1 vowel spaces

Table I displays the L1 vowels which were judged, on average, to be the most closely related to each English vowel, as well as listing the average assimilation rating for that vowel (0 for different to 1 for same). The closest vowel was defined by combining identification frequency (i.e., the proportion of trials in which each L1 response category was judged to be the closest) and the average assimilation rating for that vowel (i.e., the two numbers were multiplied and the maximum defined the closest vowel). For Spanish and French listeners, multiple English vowels often assimilated to the same L1 category. For example, Spanish listeners thought that the English vowels /a/, /aɪ/, /aʊ/, and /ɑ/ were all related to the Spanish /a/, with varying degrees of assimilation. However, Germans and Norwegians assimilated most English vowels to a unique L1 vowel. The only exceptions were English /ɔ/ and /əʊ/ which assimilated to German /o/, and English /aʊ/ and /əʊ/ which assimilated to Norwegian /æu/. Thus, there should be less pressure for Germans and Norwegians to learn new vowel categories when listening to English vowels, because they would not make many errors, in theory, if they simply used their existing L1 vowel categories to understand English. In contrast, Spanish and French listeners must learn new categories to avoid confusing English vowels.

Figure 5 displays the average L1 vowel spaces for each listener group. At its most basic level, this illustrates the substantial ways in which the vowel spaces of these listeners differed both in terms of the numbers of vowels and the use of formant movement and duration. Given these large differences in L1 vowel systems, it is notable that the English vowels in Fig. 3 were so similar across groups. That is, it is immediately apparent that subjects did not simply use their existing L1 vowel categories when listening to English (e.g., many of the English best exemplars used by Spanish listeners have no obvious counterpart in Spanish).

Item analyses (Table I) were conducted to determine whether the best exemplars of L2 English vowels were closer to L1 English vowels (i.e., average best exemplars for L1 English speakers) or to the closest vowel in each listener's L1. For example, Spanish listeners judged that English /eɪ/ was closest to Spanish /e/. For each Spanish speaker, we calculated how far their own best exemplar for English /eɪ/ was from the average /eɪ/ chosen by L1 English speakers (Fig. 3) and from the average Spanish /e/ (Fig. 5). The dis-

TABLE I. Degrees of assimilation of English vowels into L1 categories, and item analyses of whether L2 English best exemplars were closer to the L1 English vowel or the closest L1 vowel.

Assimilation			Item Analysis (t statistic)		
English vowel	Closest L1 vowel	Average rating	F1/F2 location	Formant movement	Duration
Spanish speakers					
i	i	0.78	0.49	-1.93	-2.40
ɪ	i	0.71	-0.89	-3.63 ^a	-1.43
eɪ	e	0.44	0.05	-6.24 ^a	1.89
ɛ	e	0.73	-2.19	-3.96 ^a	1.89
a	a	0.75	-1.06	2.22	-1.43
aɪ	a	0.39	-8.52 ^a	-4.21 ^a	-1.56
aʊ	a	0.37	-4.90 ^a	-4.76 ^a	-1.36
ɑ	a	0.62	-1.80	-2.52	-6.17 ^a
ɒ	o	0.72	2.32	0.78	2.40
ɔ	o	0.64	-4.02 ^a	0.86	-2.03
əʊ	o	0.43	0.85	1.30	-0.55
ɜ	o	0.57	-18.13 ^a	-3.74 ^a	-6.00 ^a
ʌ	o	0.63	-3.04	-1.20	0.77
u	u	0.71	0.79	-0.19	1.89
French speakers					
i	i	0.86	0.32	0.72	1.47
ɪ	i	0.87	-0.78	-4.29 ^a	-0.56
eɪ	ɛ	0.70	-2.57	-43.13 ^a	2.23
ɛ	ɛ	0.86	-0.29	0.20	-0.39
a	a	0.91	0.05	-1.59	-2.28
aɪ	a	0.57	-2.15	-0.37	-1.16
aʊ	a	0.47	-2.71	-3.86 ^a	-2.19
ɑ	ɑ	0.80	-5.16 ^a	-0.47	-1.28
ɒ	ɔ	0.83	-0.30	-4.94 ^a	0.69
ɔ	o	0.84	-1.01	0.05	-2.76
əʊ	o	0.74	0.28	1.01	5.46 ^b
ɜ	ø	0.80	-4.73 ^a	-2.07	-4.47 ^a
ʌ	ø	0.85	-3.75 ^a	0.82	-0.77
u	u	0.82	-0.19	-1.22	2.28
German speakers					
i	i	0.83	-1.72	-6.46 ^a	1.90
ɪ	ɪ	0.87	-0.74	-2.04	2.75
eɪ	e	0.53	-11.14 ^a	-44.80 ^a	-0.94
ɛ	ɛ	0.85	-2.09	-1.24	-1.10
a	a	0.75	-1.36	-1.95	-1.58
aɪ	aɪ	0.81	-6.30 ^a	1.62	-0.59
aʊ	aʊ	0.78	-0.65	4.39 ^b	-1.58
ɑ	aɪ	0.75	-1.61	-1.43	-1.75
ɒ	ɔ	0.84	-2.75	1.56	-0.90
ɔ	o	0.65	-1.35	-2.33	0.21
əʊ	o	0.49	-3.56 ^a	-0.82	1.38
ɜ	ø	0.63	-8.44 ^a	-3.15	1.74
ʌ	ʊ	0.74	-8.22 ^a	0.39	0.04
u	u	0.71	0.25	-6.08 ^a	0.68
Norwegian speakers					
i	i:	0.87	-0.94	-2.25	-0.16
ɪ	i	0.85	-5.16 ^a	-4.13 ^a	0.46
eɪ	aɪ	0.77	-0.46	5.71 ^b	0.87
ɛ	ɛ	0.86	0.90	-0.79	-1.46
a	æ	0.85	-2.60	-0.22	0.00
aɪ	ɛi	0.85	-7.55 ^a	5.47 ^b	-1.13
aʊ	æu	0.68	-4.72 ^a	0.98	-1.09
ɑ	ɑ:	0.84	-1.41	1.80	1.46
ɒ	ɔ	0.85	-0.76	0.54	-0.46
ɔ	o:	0.83	0.03	1.01	0.05

TABLE I. (Continued.)

Assimilation			Item Analysis (t statistic)		
English vowel	Closest L1 vowel	Average rating	F1/F2 location	Formant movement	Duration
əʊ	æʊ	0.59	4.26 ^b	0.76	1.26
ɜ	ø:	0.81	-9.28 ^a	-7.53 ^a	-0.32
ʌ	œ	0.84	-1.05	-1.49	-0.94
u	ʊ:	0.83	-1.82	-1.76	1.12

^a $p < 0.003$ closer to L1 English vowel.

^b $p < 0.003$ closer to L2 vowel.

tance calculations used the same metrics as in the accuracy analyses of Fig. 4. Paired t tests were used to determine whether the L2 English best exemplars were significantly closer to L1 English vowels (indicating learning), significantly closer to the closest L1 vowel of the listener (indicating L1 assimilation), or were nonsignificant (indicating either that the L2 English best exemplars were inbetween these two vowels, or that the variability was higher than the difference between these vowels). The significance level of $p < 0.003$ was chosen to correct for multiple tests.

Spanish speakers clearly learned new aspects of the English vowel system. For diphthongs, they chose formant movement for /eɪ/, /aɪ/, and /aʊ/ that was more English-like than their Spanish patterns of formant movement, and they likewise had more English-like F1/F2 locations for /aɪ/ and /aʊ/. For monophthongs, they chose English-like F1/F2 locations for /ɔ/ and /ɜ/, English-like patterns of formant movement for /i/, /ɛ/, and /ɜ/, and English-like durations (i.e.,

longer than Spanish) for /a/ and /ɜ/. There was no evidence that these listeners preferred significantly more Spanish-like vowels when listening to English.

French speakers also appeared to have acquired new English vowels. For diphthongs, they had English-like formant movement for /eɪ/ and /aʊ/. For monophthongs, they had English-like F1/F2 locations for /a/, /ɜ/, and /ʌ/, formant movement for /i/ and /ɔ/, and durations for /ɜ/. There was evidence that they preferred French durations for /əʊ/; this was a small difference and occurred because they chose slightly longer durations for that vowel than did English speakers.

Despite the fact that German listeners would not need to form many new English categories in order to distinguish English vowels, they showed evidence of learning. Germans chose more English-like F1/F2 locations for /eɪ/, /aɪ/, /əʊ/, /ɜ/, and /ʌ/, and more English-like formant movement for /i/, /eɪ/, and /u/. Their formant movement for /aʊ/ was significantly more like the corresponding vowel in German, indicating a degree of L1 assimilation. None of the durations were significantly more like L1 English or L1 German, because the corresponding English and German vowels had very similar durations.

Norwegians had some L1 assimilation for diphthongs; their English vowels were more significantly like Norwegian in terms of F1/F2 location for /əʊ/, and in terms of formant movement for /eɪ/ and /aɪ/. However, these listeners still were more English-like for many vowels; their English vowels were significantly more like L1 English vowels in terms of F1/F2 location for /i/, /aɪ/, /aʊ/, and /ɜ/, and in terms of formant movement for /i/ and /ɜ/. None of the durations were significantly more like L1 English or L1 Norwegian.

One of the claims of Flege's Speech Learning Model (SLM) (Flege 1995, 2003) is that these patterns of learning ought to be predictable from assimilation. That is, vowels that are weakly assimilated into L1 categories should be easier to learn than vowels that are more strongly assimilated. To test this possibility, an ANOVA was conducted with the average assimilation rating for the closest vowel as the dependent measure (i.e., as listed in Table I), and L1 and learning coded as independent categorical variables. For the learning variable, a vowel was coded as "learned" if it was significantly closer to an L1 English vowel on any of the three accuracy measures, and "not learned" if it was not. There was a main effect of L1, $F(3,48)=7.81$, $p < 0.001$. This occurred because the assimilation ratings of Norwegians were higher than those of Spanish listeners, which makes sense given the differences in the vowel systems (i.e., Norwegians have a more crowded L1 space, and thus have more vowels that are acoustically close to English vowels). However, there was no significant main effect of learning or interaction, $p > 0.05$. Thus there was little evidence that the vowels that were weakly assimilated into the L1 vowel system were easier to learn.

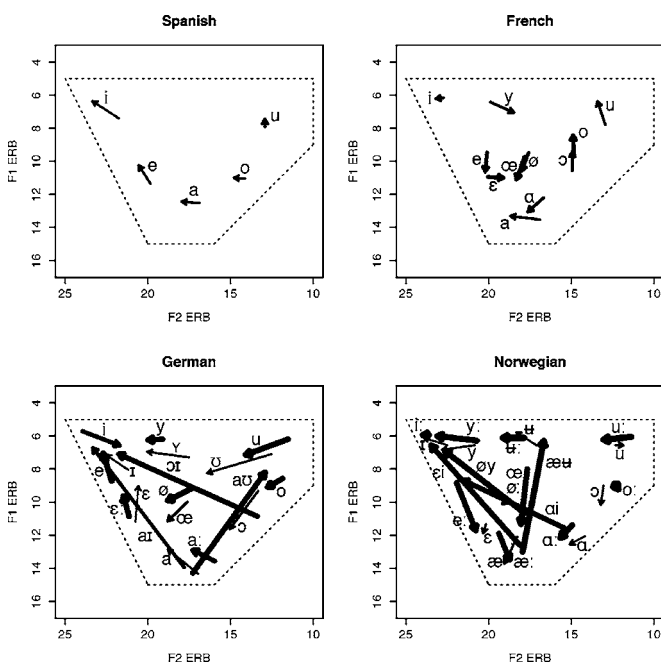


FIG. 5. Average best exemplar locations of L1 vowels for the different language groups. Each vowel is represented as an arrow from the starting F1 and F2 frequencies to the ending F1 and F2 frequencies (i.e., indicating the degree of formant movement). Duration is indicated by the weight of the line (i.e., thicker lines for longer vowels). Dotted lines indicate the boundaries of the vowel space (i.e., limits of the vowels that had been synthesized).

IV. GENERAL DISCUSSION

The results demonstrated that there were large and consistent differences between the language groups in terms of

their identification accuracy in quiet and in noise, as well as the degree to which their English best exemplars matched those of L1 English speakers. Although these group differences seem to be obvious effects of L1 vowel systems, with individuals who have larger L1 vowel systems being more accurate with L2 English vowels, this interpretation is qualified by the fact that our subject populations were not selected to be matched based on English experience, and many relevant factors could not have been matched across L1 groups even if we had tried (e.g., the fact that English television programs are dubbed in Germany but subtitled in Norway). That being said, there remains a clear effect of L1; it is difficult to imagine that other factors could explain, for example, the fact that Norwegians were much more accurate at identifying English vowels than were Spanish speakers.

Beyond these large overall differences, there was surprisingly little evidence that the language groups perceived English vowels in fundamentally different ways, despite the large differences in L1 vowel systems and the heterogeneity of the subject groups. For example, SNR thresholds for vowels increased when formant movement was flattened, with no reliable differences between language groups. This result demonstrates that even the subtle patterns of formant movement among English monophthongs were important to L2 listeners, as has been found previously for native English speakers (e.g., Assmann and Katz, 2005; Hillenbrand and Nearey, 1999; Iverson *et al.*, 2006). Moreover, the accuracy with which listeners represented formant movement in their best exemplars was correlated with identification accuracy, and this relationship did not differ reliably between groups. That is, even though Germans and Norwegians were more accurate with regard to formant movement than were Spanish and French speakers, there is evidence that formant movement was an important part of vowel perception for all groups.

Identification accuracy in noise was likewise reduced when duration was equated, and this reduction was not significantly different between the language groups. The accuracy with which duration was represented in the best exemplars was relatively weakly correlated with identification accuracy in quiet, reaching significance when all subjects were included but not when calculated within each language group. This seems to confirm the status of duration as a more secondary cue in English (e.g., Hillenbrand *et al.*, 2000). That is, the representation of F1/F2 target frequencies is a more significant cause of individual differences in vowel recognition, but the use of duration can have value when the formant information is less clear (e.g., noisy conditions).

The L1-related differences in F1/F2 targets, formant movement, and duration are particularly notable given that we previously found few such differences when testing cochlear implant users (Iverson *et al.*, 2006). That is, postlingually deafened cochlear implant users were nearly as accurate in their best exemplar locations as were normal-hearing individuals, despite the fact that the cochlear implant users averaged only 74% correct when recognizing natural vowels. The vowel-space mapping task thus appears to be relatively unaffected by peripheral distortions like these, probably because the task demands are low (e.g., listeners can repeatedly

listen to the stimuli). These low task demands, as well as the fact that the vowel judgments were made with reference to a particular talker's voice, may have contributed to the general similarity of the mean best exemplars among L1 groups in the present study (i.e., Fig. 3). However, the individual differences in these L2 English best exemplars and their correlation with identification accuracy suggest that the task was still sensitive to variation in the underlying representations for these phonemes.

Although the L2 speech perception literature has emphasized L1 assimilation (e.g., Best, 1995; Best *et al.*, 1988; Best *et al.*, 2001; Flege, 1995, 2003), the present results are mixed with regard to the role of assimilation. On one hand, the large overall differences between L1 groups could be viewed as being in accord with L1 assimilation. That is, the assimilation patterns listed in Table I should have offered an advantage to German and Norwegian speakers at the start of their English studies because they had few instances where multiple English vowels mapped onto a single L1 category; the present results suggest that this advantage may persist even after individuals have years of English experience. However, these L1 assimilation patterns also imply that Germans and Norwegians would have been under less pressure to learn English vowel categories; our item analyses suggest instead that they used many new vowels in English, rather than simply transferring their L1 vowels to English. This learning by Germans and Norwegians was particularly surprising given that their L1 vowel inventories were already quite crowded; these individuals were expected to have difficulty learning new vowel categories.

Degree of assimilation was poor at predicting which individual vowels were learned or not learned. For example, Spanish and Norwegian speakers did not demonstrate learning for English /əʊ/ despite having relatively low assimilation ratings, while Germans demonstrated learning for English /aɪ/ despite having relatively high L1 assimilation ratings. Listeners were able to learn many of the biggest and most obvious differences between the L1 and L2 vowels (e.g., Spanish and French speakers learning to add formant movement to English diphthongs), but were also able to learn subtle aspects, such as Spanish, French, and Norwegian listeners learning more English-like formant movement for /i/.

Although this relatively poor correspondence of assimilation and L2 learning is contrary to SLM (Flege, 1995, 2003), it is worth noting that the weight of the evidence for SLM has been from L2 production data (e.g., Flege, 1987, 1995, 2003; Bohn and Flege, 1992). Moreover, the perceptual evidence for SLM has mostly involved having L2 learners identify pairwise contrasts (e.g., /i/ vs /ɪ/; Flege, *et al.*, 1997; 1999); such restricted identification tasks can reflect perceptual sensitivity as much as categorization. The present study is unique in its examination of the perception of entire vowel systems, and in the use of a task that allows for direct comparisons of the underlying L1 and L2 perceptual representations in multiple dimensions. It may simply be the case that assimilation has a stronger role in constraining the learning of new productions than in the learning of perceptual representation for new vowels.

The view that emerges from the present study is that L2 vowel learning is quite pervasive, with individuals learning even when L1 assimilation is sufficient to distinguish L2 vowels, and individuals learning secondary cues for vowels (e.g., formant movement in monophthongs, and duration) rather than simply learning more primary cues (e.g., static F1/F2 targets). This more holistic pattern of learning (i.e., learning primary and secondary cues together) is compatible with the notion that the underlying categories for vowels are phonetically detailed (Goldinger, 1996, 1998; Hawkins and Smith, 2001; Johnson, 1997; Nygaard *et al.*, 1995, Nygaard and Pisoni, 1998; Pisoni, 1997). This kind of exemplar learning implies that listeners would learn the details of a vowel all at once, rather than only learning whichever individual cues seem best for distinguishing categories. That being said, the evidence for this kind of holistic learning in the present study is mixed. In support of this conclusion, the between-group differences in accuracy of F1/F2 location, formant movement, and duration all follow the same basic pattern (i.e., Spanish and French speakers being less accurate), suggesting that language groups who have poor representations of primary cues also have poor representations of the more secondary cues. However, the individual differences in F1/F2 location and duration accuracy were only weakly correlated, demonstrating that individuals who were accurate at one cue were not necessarily accurate at all others. These idiosyncratic patterns of cue weighting (e.g., individuals representing duration more accurately than target formant frequencies) were not apparent in the cross-language comparisons simply because they were not strongly related to L1 background. It is thus possible that L2 learners may engage in cue-based learning, even though learning may appear to be holistic when looking at entire vowel systems and across L1 groups.

Although vowel category learning seems pervasive across L1 groups in the present study, learning L2 vowels is not always easy. For example, even highly experienced bilingual Spanish-Catalan speakers have difficulty with the Catalan /e/-/ɛ/ distinction if their first language was Spanish (Pallier *et al.*, 2001). In the present study, Spanish and French speakers made many errors recognizing English vowels despite having years of experience. Assimilation models may be able to explain some of the learning problems for individual vowels, but it is clear from the present results that assimilation alone does not fully explain the difficulties that individuals have when learning an L2 vowel system.

ACKNOWLEDGMENTS

We are grateful to Anke Sennema for subject recruitment and hosting the data collection at University of Potsdam, Germany, Dawn Behne for hosting the data collection at Norwegian University of Science and Technology, and Eivind Torgersen for advice on the Norwegian stimuli. This research was funded by Grant No. RES-000-23-0838 from the Economic and Social Research Council of the UK.

Assmann, P. F., and Katz, W. F. (2005). "Synthesis fidelity and vowel identification," *J. Acoust. Soc. Am.* **117**, 886–895.

Best, C. T. (1995). "A direct-realist view of cross-language perception,"

Speech Perception and Linguistic Experience: Issues in Cross-Language Research, edited by W. Strange (York Press, Baltimore, MD), pp. 171–204.

- Best, C. T., McRoberts, G. W., and Goodell, E. (2001). "Discrimination of non-native consonant contrasts varying in perceptual assimilation to the listener's native phonological system," *J. Acoust. Soc. Am.* **109**, 775–794.
- Boersman, P., and Weenink, D. (2005). "Praat: Doing phonetics by computer," [Computer software] (University of Amsterdam, Amsterdam, The Netherlands).
- Bohn, O.-S. (1995). "Cross-language speech perception in adults: First language transfer doesn't tell it all," *Speech Perception and Linguistic Experience: Issues in Cross-Language Research*, edited by W. Strange (York Press, Baltimore, MD).
- Bohn, O.-S., and Flege, J. E. (1990). "Interlingual identification and the role of foreign language experience in L2 vowel perception," *Appl. Psycholinguist.* **11**, 303–328.
- Bohn, O.-S., and Flege, J. E. (1992). "The production of new and similar vowels by adult German learners of English," *Stud. Second Lang. Acquis.* **14**, 131–158.
- Cebrian, J. (2006). "Experience and the use of non-native duration in L2 vowel categorization," *J. Phonetics* **34**, 372–387.
- Delattre, P. (1965). *Comparing the Phonetic Features of English, French, German, and Spanish* (Harrap & Co, London).
- Escudero, P., and Boersma, P. (2004). "Bridging the gap between L2 speech perception research and phonological theory," *Stud. Second Lang. Acquis.* **26**, 551–585.
- Evans, B. G., and Iverson, P. (2004). "Vowel normalization for accent: An investigation of best exemplar locations in northern and southern British English sentences," *J. Acoust. Soc. Am.* **115**, 352–361.
- Evans, B. G., and Iverson, P. (2007). "Plasticity in vowel perception and production: A study of accent in young adults," *J. Acoust. Soc. Am.* **121**, 3814–3826.
- Flege, J. E. (1987). "The production of "new" and "similar" phones in a foreign language: Evidence for the effect of equivalence classification," *J. Phonetics* **15**, 47–65.
- Flege, J. E. (1989). "Differences in inventory size affect the location but not the precision of tongue positioning in vowel production," *Lang Speech* **32**, 123–147.
- Flege, J. E. (1995). "Second Language speech learning: Theory, findings, and problems," *Speech Perception and Linguistic Experience: Issues in Cross-Language Research*, edited by W. Strange (York Press, Baltimore, MD), pp. 233–277.
- Flege, J. E. (2003). "Assessing constraints on second-language segmental production and perception," *Phonetics and Phonology in Language Comprehension and Production: Differences and Similarities*, edited by A. Meyer and N. Schiller (Mouton de Gruyter, Berlin, Germany).
- Flege, J. E., Bohn, O.-S., and Jang, S. (1997). "The effect of experience on nonnative subjects' production and perception of English vowels," *J. Phonetics* **25**, 437–470.
- Flege, J. E., MacKay, I. R. A., and Meador, D. (1999). "Native Italian speakers' production and perception of English vowels," *J. Acoust. Soc. Am.* **106**, 2973–2987.
- Flege, J. E., Schirru, C., and MacKay, I. R. A. (2003). "Interaction between the native and second language phonetic subsystems," *Speech Commun.* **40**, 467–491.
- Glasberg, B. R., and Moore, B. C. J. (1990). "Derivation of auditory filter shapes from notched-noise data," *Hear. Res.* **47**, 103–138.
- Goldinger, S. D. (1996). "Words and voices: Episodic traces in spoken word identification and recognition memory," *J. Exp. Psychol.* **22**, 1166–1183.
- Goldinger, S. D. (1998). "Echoes of echoes? An episodic theory of lexical access," *Psychol. Rev.* **105**, 251–279.
- Gottfried, T., and Beddor, P. S. (1988). "Perception of spectral and temporal information in French vowels," *Lang Speech* **31**, 57–75.
- Hawkins, S., and Smith, R. (2001). "Polysp: a polysystemic, phonetically-rich approach to speech understanding," *Italian Journal of Linguistics* **13**, 99–188.
- Hazan, V., and Markham, D. (2004). "Acoustic-phonetic correlates of talker intelligibility for adults and children," *J. Acoust. Soc. Am.* **116**, 3108–3118.
- Hillenbrand, J. M., Clark, M. J., and Nearey, T. M. (2000). "Some effects of duration on vowel recognition," *J. Acoust. Soc. Am.* **108**, 3013–3022.
- Hillenbrand, J. M., and Nearey, T. M. (1999). "Identification of resynthesized hVd utterances: Effects of formant contour," *J. Acoust. Soc. Am.* **105**, 3509–3523.

- Iverson, P., and Evans, B. G. (2003). "A goodness optimization method for investigating phonetic categorization," Proceedings of the 15th International Conference of Phonetic Sciences, Barcelona, Spain.
- Iverson, P., Smith, C. A., and Evans, B. G. (2006). "Vowel recognition via cochlear implants and noise vocoders: Effects of formant movement," *J. Acoust. Soc. Am.* **120**, 3998–4006.
- Johnson, K. (1997). "Speech perception without speaker normalization: An exemplar model," *Talker Variability in Speech Processing*, edited by K. Johnson and J. W. Mullenix (Academic, San Diego, CA), pp. 145–165.
- Klatt, D. H., and Klatt, L. C. (1990). "Analysis, synthesis, and perception of voice quality variations among female and male talkers," *J. Acoust. Soc. Am.* **87**, 820–857.
- Krisoffersen, G. (2000). *The Phonology of Norwegian* (Oxford University Press, Oxford, UK).
- Ladefoged, P., and Broadbent, D. (1957). "Information conveyed by vowels," *J. Acoust. Soc. Am.* **29**, 98–104.
- Lee, B., Guion, S. G., and Harada, T. (2006). "Acoustic analysis of the production of unstressed English vowels by early and late Korean and Japanese bilinguals," *Stud. Second Lang. Acquis.* **28**, 487–513.
- Levitt, H. (1971). "Transformed up-down methods in psychoacoustics," *J. Acoust. Soc. Am.* **49**, 467–477.
- MacKay, I. R. A., Flege, J. E., Piske, T., and Schirru, C. (2001). "Category restructuring during second-language (L2) speech acquisition," *J. Acoust. Soc. Am.* **110**, 516–528.
- McAllister, R., Flege, J. E., and Piske, T. (2002). "The influence of the L1 on the acquisition of Swedish vowel quantity by native speakers of Spanish, English and Estonian," *J. Phonetics* **30**, 229–258.
- Meunier, C., Frenck-Mestre, C., Lelekov-Boissard, T., and Le Besnerais, M. (2003). "Production and perception of foreign vowels: does the density of the system play a role?" Proceedings of the 15th International Congress of Phonetic Sciences, Barcelona, Spain.
- Morrison, G. (2002). "Perception of English /i/ and /ɪ/ by Japanese and Spanish listeners: Longitudinal results," *Proceedings of the North West Linguistics Conference 2002*, edited by G. S. Morrison and L. Zsoldes (Simon Fraser University Linguistics Graduate Student Association, Burnaby, BC, Canada), pp. 29–48.
- Munro, M. J., Flege, J. E., and MacKay, I. R. A. (1996). "The effects of age of second language learning on the production of English vowels," *Appl. Psycholinguist.* **17**, 313–334.
- Nygaard, L. C., and Pisoni, D. B. (1998). "Talker-specific learning in speech perception," *Percept. Psychophys.* **60**, 355–376.
- Nygaard, L. C., Sommers, M. S., and Pisoni, D. B. (1995). "Effects of stimulus variability on perception and representation of spoken words in memory," *Percept. Psychophys.* **57**, 989–1001.
- Pallier, C., Colome, A., and Sebastian-Galles, N. (2001). "The influence of native-language phonology on lexical access: Exemplar-based versus abstract lexical entries," *Psychol. Sci.* **12**, 445–449.
- Peterson, G. E., and Barney, H. L. (1952). "Control methods used in a study of the vowels," *J. Acoust. Soc. Am.* **24**, 175–184.
- Pisoni, D. B. (1997). "Some thoughts on 'normalization' in speech perception," *Talker Variability in Speech Processing*, edited by K. Johnson and J. W. Mullenix (Academic, San Diego, CA), pp. 9–32.
- Stockwell, R. P., and Bowen, J. D. (1965). *The Sounds of English and Spanish* (University of Chicago Press, Chicago, IL).
- Strange, W., Bohn, O.-S., Nishi, K., and Trent, S. (2005). "Contextual variation in the acoustic and perceptual variation of North German and American English vowels," *J. Acoust. Soc. Am.* **118**, 1751–1762.