Are Phoneme Categories Context-Dependent or Context-Independent?

English Phoneme Discrimination by Dutch-English Early Bilinguals in Dutch and English Pseudo-Word Contexts

by

Andreea Geambașu

Supervisor: Prof. Dr. Paul Boersma

Co-Assessor: Drs. Titia Benders

Research Report for the Completion of MSc in Brain and Cognitive Sciences
Track: Cognitive Science
Amsterdam, 2011

COGNITIVE SCIENCE CENTER AMSTERDAM

AMSTERDAM CENTER FOR LANGUAGE AND COMMUNICATION

ACLC

Acknowledgements

Many wonderful people have helped me over the year that I've been working on developing and executing this project.

First of all, thank you to Titia Benders for helping me with the brainstorming process for this project, for your valuable feedback, for taking the time to help me with the studio recordings and splicing, and for assuming the role of co-assessor.

For their guidance on the topic of EEG research, thank you to Titia van Zuijen and Karin Wanrooij. I appreciate that you took the time to talk with me on multiple occasions, even though you were not directly involved in this project, and even though time did not allow for us to execute the proposed EEG part of this project. I have learned so much from you both nonetheless.

Also thanks to Jan-Willem van Leussen for coordinating the testing rooms and for taking the time to help with stimuli recordings; to Maartje Raijmakers for your help with the ethical approval; to David Weenink for filling the role of UvA representative; and to Dirk Vet for all the technical support.

Finally, a sincere thank you to my supervisor Paul Boersma for all of your guidance along the way, your knowledgeable feedback, your help with the analysis and for pushing me to teach myself basic statistics, for giving me the chance to attend L.O.T., and for your help on my first attempt at my own PhD project proposal. I appreciate having had the opportunity to work with you and everything I have learned from the experience.

Also, special thanks to all my friends in the Netherlands and in the U.S. for your help with the piloting, and to everyone who participated in the experiment. I couldn't have done it without all of you.

The people I have thanked until now have helped me professionally/academically, but I would also like to thank the people who have helped me emotionally and psychologically.

To my friends, scattered all over the world and here in the Netherlands, who have all helped make this sometimes difficult two years a bit easier, with chats (online or live), lunches, dinners, and of course, drinks at Kriterion.

Thank you to my parents for supporting me, pushing me, and believing in me my whole life, and for everything you've done for me. And especially to my mom for making me take that one lower division linguistics class during my first year of college, when I had no idea what linguistics was and thought that if I studied linguistics I would turn into a spelling and grammar freak like her. Without that nudge I may have become a starving artist. Now I am only a slightly hungry scientist.

And, of course, to Erdinç for inspiring me, encouraging me, and for making me laugh every day.

Contents

Abstract

The aim of this study was to investigate the properties of phonetic categories. In particular, we wished to examine whether phonetic categories are rigid and context-independent, or flexible and context-dependent. Language mode effects have previously been shown for phoneme identification but not yet for phoneme discrimination. In order to test language mode effects on discrimination abilities, early bilinguals in Dutch and English (learning English between 6 and 12 years old; M=9.4 years old) were tested on their ability to discriminate two vowels that are phonemic in English but are part of the same category in Dutch: /ɛ/ and /æ/ in English are encompassed by a broader /ɛ/ category in Dutch. As such, this is a contrast that has previously been shown to be difficult for Dutch speakers to perceive. Their discrimination ability was tested in both a Dutch and English pseudo-word context. Results did not show that subjects performed differently in one language context or another. Most subjects who performed well in one condition also performed well in the other, and vice versa, regardless of the language context. This pattern suggests that phonemic categories may not be context-dependent. In addition, the only difference in performance seemed to be the result of a training effect, with subjects significantly improving in discrimination performance and reaction time in the second condition. This effect was independent of which order the languages were presented in. These results taken together showed that there is a range of individual differences within a relatively homogeneous group of speakers in terms of their discrimination ability. The results also suggested that identification abilities may not necessarily be predictive of discrimination abilities, as has usually been assumed. Theoretical implications of these results are discussed in detail.

Introduction

Aim

The aim of this study was to investigate the nature of second language phonetic categories in bilinguals. We asked whether categories that have already been established since childhood can be modified when a speaker learns a second language early in life. This is not a novel question in itself. Much previous work has been done on pronunciation of speakers who have acquired their second language (L2) at various stages in life, and on speakers' ability to identify sounds of their L2. What is novel about the present research is that while much work has been done on phoneme identification, both in and out of language context, there has been little work on bilingual phoneme discrimination abilities and how they are affected by which language mode a speaker is in (for one exception see Winkler et al., 2003, detailed below). While it is assumed that identification performance is a predictor for discrimination abilities (as discriminating two sounds is only possible to the extent that they can be identified as different) (Liberman et al., 1957), the processes involved in each task are quite different. Identification requires the subject to keep a sound in memory and to explicitly categorize it within one of his established phonetic categories. Discrimination, on the other and, requires the subject to keep in memory two minimally different sounds, and, depending on the task, either say which sound is most like another (AXB task), say whether two sounds are alike or different (AX task), or say when the sound changes (Oddball task, typically used in EEG)(Strange and Shafer, 2008). A speaker's discrimination ability can tell us about the phonetic categories he has (or has not) formed in his mental representation; it can also tell us whether these categories are flexible and context-dependent, or if they are the same no matter the context (context-independent). Moreover, discrimination can be performed on the basis of acoustic properties or phonetic properties (Pisoni, 1973). Discrimination on the basis of acoustic properties involves the auditory system picking up on the slightest changes between the two sounds (i.e., formant frequency, duration, voice onset time, voicing, etc.). Discrimination on the basis of phonetic properties involves picking up on these auditory differences but also includes a psychological component of implicitly categorizing the perceived sounds on the basis of established phoneme categories in the speaker's mind (Aaltonen et al., 1987). For example, a difference of 100 Hz in the first formant may constitute a purely acoustic difference in one language but may cross a phoneme boundary in another language. Thus, the principal question in this study was whether highly-proficient, early bilinguals' ability to phonetically discriminate non-native sounds that constitute phoneme categories in their L2 but not in their L1 changes as a function of which of their two languages they are listening to (i.e., as a function of which language mode they are in; Grosjean, 1998, 1999). In a Dutch context they should be discriminating acoustically, and thus find the task more difficult (as it is more difficult to hear changes within phoneme boundaries than across; according to the Perceptual Magnet Effect; Kuhl, 1991); when they are in English language mode, however, they should be discriminating on the basis of phoneme categories, making the task easier. The theoretical and experimental precedents leading up to this work will be described in more detail in the following sections.

Development of phoneme categories

Previous work has shown that native phoneme categories are formed during the first year after birth, a period during which there is a decline in the infant's ability to discriminate non-native contrasts (Werker & Tees, 1984). It has been shown that vowel categories are formed around 6 months of age, while consonant categories are formed around 10 to 12 months of age (Bosch & Sebastián-Gallés, 2003). Behavioral studies, such as Maye et al. (2002, 2008), have shown that, initially, children's ability to create new, contrasting phonetic categories relies on the statistical distribution of variance between the two. For example, the syllables /ta/ and /da/ in English form a bimodal distribution, which allows children to differentiate the two sounds and begin to form two separate phoneme categories for the /t/ and the /d/. These authors have shown that this occurs even before an infant turns 6 months old.

Such behavioral findings have been confirmed in studies using EEG to measure neurophysiological responses to native language categories versus non-native category contrasts (i.e., phonemes in the non-native language that are allophones of the same phoneme in the native language). An EEG component called the Mismatch Negativity (MMN) has been found to be larger for native category contrasts than non-native contrasts and larger for phonemic changes than acoustic changes (Dehaene-Lambertz, 1997; Dehaene-Lambertz & Gliga, 2004; Näätänen et al., 2007). In Cheour et al. (1998), this effect was seen in children as young as 12 months old: Finnish and Estonian monolingual infants produced smaller MMNs for non-native contrasts than native contrasts, indicating a language-specific, experience-driven development of phonemic categories takes place before the child's first birthday. Others, such as Dehaene-Lambertz and Baillet (1998), have even shown that much younger infants (3 months old), show evidence of phonemic category formation, with higher amplitude MMNs for cross-category deviants in an Oddball listening task than for within-category deviants. They showed that this occurred despite the fact that the acoustic distance between each phoneme was identical to the difference that constituted a phonetic change (same difference between frequencies, but within phoneme boundary, as opposed to across phoneme boundary).

Bilingual speech perception

We have seen that phonemic memory traces are established even before a child turns one year old. What happens then when speakers learn a second language years after their phonetic categories have been established? Do they simply add on a few new phoneme categories to their already-existing phonetic space? Or do they create a whole new and separate phonetic space for this L2? Previous work has shown the former option seems to be how learners deal with novel phonemes: they fit their newly-acquired phonemes into the phonetic space that is already in place from their native language (Best et al., 2001; Bosch et al., 2000; Escudero & Boersma, 2002; Flege, 1995; Sebastián-Gallés & Soto-Faraco, 1999), resulting in discrimination and pronunciation abilities that are clearly non-native. Flege proposed a formal explanation for this in his Phonological Translation Hypothesis (Flege, 1981). Flege was the first to hypothesize that pronunciation errors commonly found in L2 speakers had nothing to do with the age of acquisition or a so-called critical period for learning speech sounds. Instead, he claimed that it was because speakers had a "tendency...to interpret sounds occurring in a foreign language in terms of sounds found in their native language." According to Flege, the influence is not one-sided. Indeed, if the L1 modifies the establishment of the L2 categories, the L2 may also modify the categories of the L1. Flege's (1995) Speech Learning Model (SLM) expanded on his earlier Phonological Translation Hypothesis, proposing that category formation for L2 phonemes will not occur if the phonemes can be equated with phonemes in the L1. He clarifies that this is not to say that phonetic learning does not occur at all, but instead that the phonetic learning will take the form of a modified version of the L2 phoneme that can be absorbed by the L1, which in turn modifies the already established L1 categories as well. This has been shown to be true, for example, by MacKay et al. (2001), who tested this theory on early and late Italian-English bilinguals. These speakers had immigrated to Canada from Italy as children, as teenagers, or as adults. The authors argued that early bilinguals' improved ability to perceive and pronounce the L2 stops /b d g/ as compared to late bilinguals speakers was not a result of the early bilinguals establishing new phonetic categories for the English stops, but rather because of the difference in quantity and quality of phonetic input between the two groups. Indeed, early bilinguals (mean age of acquisition = 8 years old), did not acquire native English categories, and still produced non-native exemplars of the target consonants with Italian phonetic properties (i.e., prevoicing) more than natives, although significantly less than late bilinguals. Moreover, the authors also showed that the acquisition of the L2 sounds had modified all of the speakers' established L1 categories: these speakers pronounced Italian stops with less pre-voicing than native Italian monolinguals did. Similarly, Caramazza et al. (1973, 1974) had also earlier showed the pronunciation of French stops by French-English bilinguals as being less native-like, and more English-like (lengthier), that those of monolinguals.

Similar to Flege's SLM, Best's Perceptual Assimilation Model (PAM) (Best, 1995; Best et al., 2001) predicts that L2 sounds will be discriminated better if they are less assimilated into the L1 phonetic space. However, these authors claim that, at least initially, L2 phonemes *are* assimilated into the L1 phonetic categories. In addition, the PAM predicts that two distinct phonemes of the L2 will be better discriminated if they fit into two distinct categories of the L1, whereas two L2 categories that can be equally fit into one category of the L1 will be difficult to discriminate. For example, the English /ε/ and /æ/ both fit into the Dutch /ε/ category, and are thus difficult to discriminate, as their "goodness of fit" into that category may be equal (Cutler et al., 2004; Escudero et al., 2008).

Flege's SLM and Best's PAM share the assumption that L2 learners map the newly acquired sound inventory of the L2 onto their already established phonetic space of their L1. However, this does not mean that no phonetic learning takes place, or that L2 speakers will never learn to pronounce or perceive L2 categories correctly. It also does not mean that L1 categories are unchangeable. The fact that previous research has shown that L1 pronunciation is modified with the acquisition of L2 phonetics indicates that there is an interesting change in the phonemic system going on that is a result of modifications in the perceptual system. With this in mind, we assumed that our Dutch-English bilingual subjects had integrated some kind of meta-knowledge about English phoneme categories into their perceptual system. Testing whether the language mode turns one phoneme inventory off and another on is a direct test of Flege's and Best's models. If not, it would be an indication that the L2 phonemes do not form a new inventory but are indeed mapped onto the L1 space, as Flege and Best hypothesized. However, if language mode does dictate which phoneme inventory is activated, it would pose a challenge to these models.

Context effects on identification

Previous work that has already investigated the role of context in perceptual abilities has so far been mostly limited to identification studies. One of the few examples of studies on discrimination of phonemes in linguistic context comes from Winkler et al. (2003), who studied whether language context would cause different MMN amplitudes depending on language context and on whether the *meaning* of the deviant would have an effect on this change-detection process. In one of their experiments, the standard and deviant varied only acoustically, while in another experiment they differed phonemically (and thus semantically). The authors first conducted an identification experiment in which they tested monolingual Hungarians and Finns on their ability to correctly categorize Finnish /æ/ and /e/, two phonemes that are encompassed by the phoneme /ɛ/ in Hungarian. This was done in order to make sure monolingual Finns could categorize this distinction correctly, and to test whether monolingual Hungarians could categorize these two vowels differently on the basis of acoustic differences (since the differences are not phonemic in their native language). They found that while Finns had a performance rate of around 80% correct, Hungarians performed around chance level, correctly identifying around 50% of the items. Then they conducted a discrimination test in Finnish and Hungarian context on bilinguals who had immigrated to Finland from Hungary at a variety of ages. The authors found that Hungarian subjects identified standard and deviant words as different in both Hungarian and Finnish context, even though in Hungarian the difference did not constitute a semantic difference but would be considered two exemplars of the same word (i.e., the difference was not phonemic). The authors conclude that in L2 speakers, once the distinction has been learned, it is used no matter what the language context and independent of the relevance to the language context in use. This effect was even present in late learners, who showed no significant difference when compared to early learners. The authors find support in the work of Hahn et al. (2002) who showed that auditory deviance detection and syntactic processes act independently of each other. Therefore, if change detection works independently of syntactic and semantic processes, we should investigate whether purely phonological processing can modulate the phonetic activation of other phonemes. In other words, we should investigate whether the simple act of hearing consonants uttered by a native speaker of that language would activate the phonetic space of the target language only, or if, on the other hand, once one learns an L2, there is only one, modified phonetic space that is always activated no matter the phonological context.

Escudero and Boersma (2002) investigated such a possibility by studying the role of language context on the identification ability of native Dutch learners of L2 Spanish. In one task, subjects were presented with Spanish /e/ and /i/ within a Spanish word context that they were told was Dutch, embedded within a Dutch carrier phrase (i.e. *luister naar [kes]*). The fact that the Spanish pseudo-words were embedded in a Dutch sentence made the Dutch speakers perceive and identify the two Spanish vowels as three Dutch vowels /ɛ/, /ɪ/, and /i/. In a second task, when the same Spanish pseudo-words were placed in a Spanish carrier sentence, but subjects were asked to "listen with their Dutch ears," they identified the native Dutch categories less. This finding is interesting because it shows that the surrounding environment plays a significant role in modulating perceptual and identification abilities. This study was used as inspiration for the present work, as it leads to the following novel empirical question: if native speakers are able to suppress their native categories when hearing a non-native phonological context, we wondered if the inverse was true. Namely, would speakers perceive non-native English vowel contrasts better when hearing them in an English context than when hearing them in a Dutch context?

Dutch perception of English /ɛ/-/æ/

Dutch and English are both members of the same language family, namely, West Germanic. As such, they share a number of phonetic and prosodic features. In fact, some English infants perceive both of the two languages as native (Christophe & Morton, 1998) at 2 months old, an age at which they are clearly able to use phonological properties to recognize only English as native in English-Japanese or English-French language pairs. When looking at the phonemic space of the two languages, however, we see some very clear distinctions. Whereas American English has 11 vowels (Hagiwara, 1997), Dutch has 15, including diphthongs (Adank et al., 2007). For the purpose of this study, we will focus on two vowels of English that are not found in Dutch: /ɛ/ and /æ/ which are encompassed in Dutch by only one phoneme /ɛ/. In Figure 1 we can see the properties of the two English vowels, while in Figure 2, we can see that Dutch /ɛ/ is similar with respect to the second formant (F2; backness) features of English /ɛ/ and /æ/ (all three are front vowels with an F2 value concetrated between 1500 and 1900 Hz), but that the first formant (F1; height) features of the Dutch /ɛ/ encompass English /ɛ/ and /æ/. Figure 1 shows American English /ɛ/ as having F1 values approximately between 500 and 600 Hz, and /æ/ as having F1 values approximately between 600 and 800 Hz (based on natural speech of Western males, Clopper et al., 2005). Figure 2 shows the Dutch /ɛ/ as encompassing a broad F1 area between approximately 500 and 800 Hz (based on Pols et al., 1973). Thus, where English has a bimodal distribution between these two frequencies, Dutch has a broad unimodal distribution encompassing only one vowel category. This makes the two English vowels interesting to test in Dutch-English bilinguals in tasks that aim to see whether speakers who are trained on unimodal distributions during their formative years are able to eventually modify (specifically to split) these categories later when learning a second language.

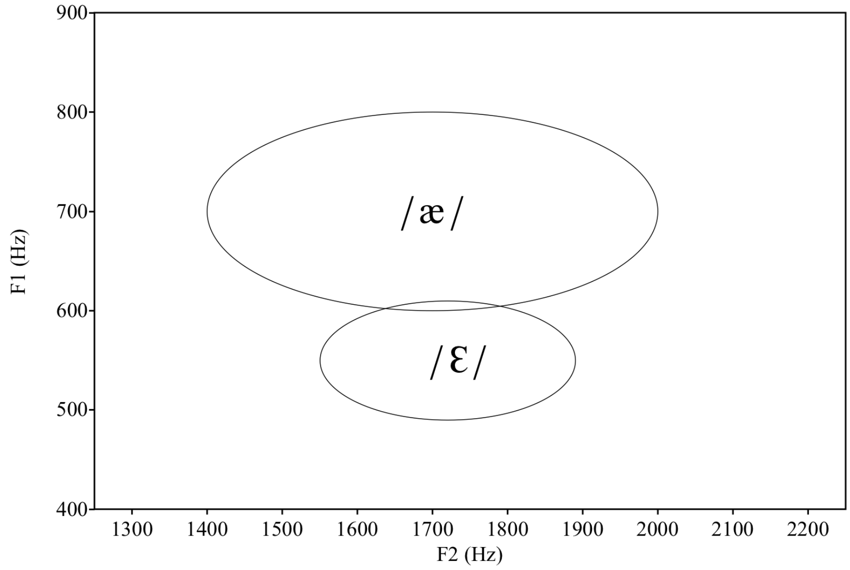

Figure 1. F1 and F2 frequencies of the vowels /ɛ/ and /æ/ in the recorded speech of Western male speakers. The figure has been redrawn from Clopper et al. (2005). Exact values were not given for the limits of the ellipses, and were thus estimated and redrawn from Figure 10. Of interest is the y-axis showing F1 values for /ɛ/ between roughly 500 and 600 Hz, and for /æ/ roughly between 600 and 800 Hz.

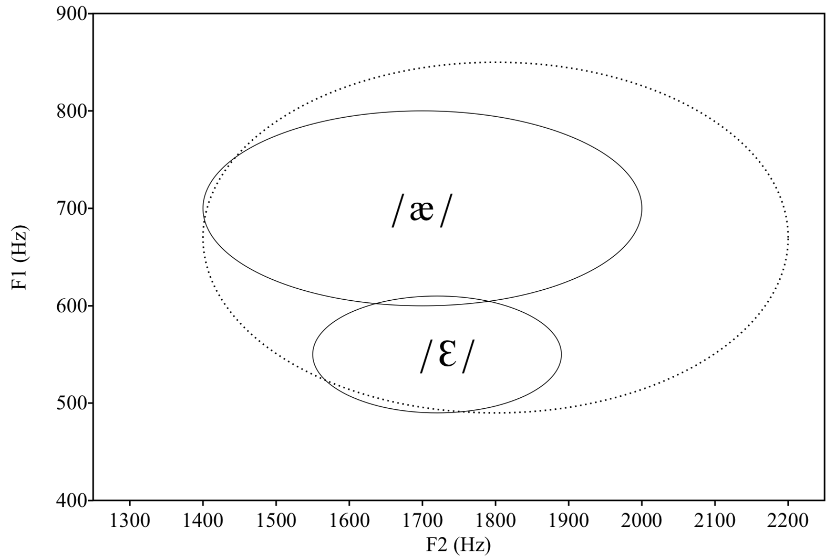

Figure 2. F1 and F2 frequencies of the vowel /ɛ/ in recorded speech of Dutch male speakers is outlined in a dotted line encompassing the values of American English /ɛ/ and /æ/. Dutch values taken from Figure 3 in Pols et al. (1973) and American English values were taken from Figure 10 in Clopper et al. (2005). Notice that the F1 of the Dutch /ɛ/ is between 500 Hz and 800 Hz, showing that where Dutch has a unimodal distribution, English has a bimodal distribution.

Previous identification and discrimination tasks have used these vowels with Dutch speakers before. Dutch speakers have been shown to have difficulties in correctly identifying these vowels in various tasks, whereas native American English speakers have been shown to identify them correctly around 80% of the time, even under noise (Weber & Smits, 2003). A confusion matrix for American English speakers indicated that under 0db signal-to-noise ratio (SNR) (i.e., high noise), these speakers correctly identified /ɛ/ 78.8% of the time (confusing it with /æ/ 5.7% of the time) and /æ/ 80.5% of the time (confusing it with /ɛ/ 8.1% of the time). In Cutler et al. (2004), the authors presented confusion matrices comparing American English and Dutch native speakers based on an identification task. Under the same SNR, Dutch speakers correctly identified /ɛ/ in syllable-initial position 58.3% of the time (confusing it with /æ/ 33.6% of the time) and

/æ/ in syllable-initial position 56.3% of the time (25% of the time). Although confusion matrices for these vowels under low noise are not given, Figure 3 (Cutler et al., 2004, p. 3675) shows that no matter the level of noise or the syllable position of the vowel (initial or final), Dutch speakers had difficulty with vowel height differences that constituted phonetic differences in English. This is particularly relevant to us, as we are testing the discrimination abilities on two vowels that differ mainly in height (F1).

Weber and Cutler (2004) also showed in an eye tracking study that this difficulty that Dutch L2 speakers of English have in distinguishing /ɛ/ and /æ/ led to a confusion when hearing real words as well. As subjects heard the word *panda*, the eye tracker recorded that they initially looked toward the picture of a pencil before selecting the correct picture. However, when they heard the word *pencil*, they did not tend to look towards the picture of a panda. These results indicate that the two English phonemes are not represented as interchangeable homographs by L2 speakers. However, subjects were sensitive to frequency effects, with the word *hedge* causing subjects to look at the picture of a *hat* more frequently. The authors conclude that these results indicate the power of lexical knowledge in disambiguating similar sounds. They claim that lexical knowledge is more powerful in L2 speakers' category learning than phonetic information: "L2 listeners maintain lexical distinctions even when success in mapping spoken input correctly to the two distinct categories is at best no more than a faint future hope." In this way, highly proficient speakers can correctly understand and even sometimes pronounce the L2 with little discernable accent, but still have problems in identifying and discriminating non-native phonemes in contexts that do not provide lexical, semantic, or syntactic cues. This is the reason for which we chose not to include real words as the linguistic context for our stimuli. Pseudo-word contexts were thought to be more appropriate in examining purely phonetic knowledge, as L2 speakers at this level may rely on lexical knowledge to determine which phoneme they hear (Escudero & Boersma, 2004; Weber & Culter, 2004).

Escudero et al. (2008) did exactly this, bypassing the problem of allowing speakers to rely on lexical cues by using non-words *tenser* and *tandek* in an eye-tracking task. This experiment also sought to understand whether subjects would better learn contrasts that were presented explicitly (audio training and spelling provided; similar to real life L2 phonetic learning) or implicitly (audio training only; similar to L1 phonetic learning). They found that the group that was trained implicitly looked equally at both images at the beginning of the syllable; that is, when they heard *ten*- from *tandek*, their eyes looked back and forth from the picture of a *tandek* and that of a *tenzer. T*he same pattern was seen when they heard *tan*- from *tandek*. However, subjects who were trained explicitly showed results similar to Cutler and Weber; that is, *tan*- made subjects look to the *tenzer* picture but not vice versa. This means that implicit training made the subjects represent the two sounds according to the L1, while explicit training made them form a meta-knowledge (as Weber and Cutler, 2004 had hypothesized) of the lexicon that allowed them to discriminate similar-sounding words. For this reason we did not use orthographic knowledge or explicit instruction about the phonemes in question. The intent was to simply understand the nature of the phonetic cues used by the subjects without the use of lexical, semantic, or orthographic cues.

Expected Results

We hypothesized that native Dutch speakers who are highly proficient speakers of English would perform better on a phonetic discrimination task in an English context than in a Dutch context. This hypothesis is based on previous research showing an effect of linguistic context on identification abilities (such as Winkler et al., 2003 and Escudero & Boersma, 2002), and on the assumption that identification performance is a predictor of discrimination performance, as has been the usual hypothesis in categorical perception research since Liberman et al. (1957). If our hypothesis is true, and subjects do change their discrimination ability as a result of context, it would mean that not only could highly proficient L2 speakers eventually modify their L1 categories to fit the phoneme categories of the L2, but also that these categories could be turned on and off depending on the language context. That is, when listening to Dutch, speakers would have no need to have their English phoneme categories "activated" and would perceive both English /ɛ/ and /æ/ as one category (i.e., as allophones of the same phoneme /ɛ/), resulting in chance-level performance. On the other hand, when listening to English, subjects would need to use the sound inventory of English and would thus use all the categories of this language. This would result in above-chance performance in the English context, as the subjects switch from one phoneme inventory to another. If they have formed phonetic categories for this language, they should perform above chance in accordance with the Phonetic Magnet Effect (PME; Kuhl, 1991), which predicts that discrimination across phoneme boundary is easier than discriminating sounds that fall within a phoneme boundary (i.e., two exemplars of /æ/ would be more difficult to discriminate than two prototypical exemplars of /æ/ and /ɛ/).

A second possibility is that subjects have not shifted their native categories at all. This would result in a chance performance in both Dutch and English conditions. This result would mean that, despite years of training and input, subjects have not been able to form a phoneme category for their L2 nor to modify their L1 categories to include L2 phonemes.

The third and final possibility would be that subjects have indeed modified their native categories to include the L2 categories, but that these categories, once modified, are not context-dependent, but always activated no matter the context. Such a result would be compatible with Flege's SLH and Best's PAM as detailed above, which assume that L2 speakers map L2 phonemes onto an L1 phonetic inventory.

Method

Creating the Stimuli

A great deal of effort was put into researching and selecting the stimuli, and into piloting and refining these stimuli on native speakers of both Dutch and English.

A language's vowel space often varies as a function of dialect. As our intent was to measure the discrimination abilities of non-native speakers of two vowels that form two distinct categories in American English, a dialect which distinguishes between /ɛ/ and /æ/ with minimal overlap was necessary. For this reason, Southern, Northern, and Eastern dialects were immediately excluded. These dialects are perceived as non-standard within the US and abroad, and would thus not be taught to foreigners, might potentially be unfamiliar to Dutch speakers, and include a variety of overlaps and movements within the vowel space. Figures of these non-standard dialects' vowel spaces can be found in Clopper et al. (2005) and Hillenbrand (2005).

In addition, as can be seen from Table 1 and from Figure 3, Midland (Midwestern) values for our target vowels were nearly identical in terms of all formant values. The only distinctive variable was duration, which we wanted to keep constant for this experiment. As a result of these findings, the Western American dialect was chosen. Once it was established that Western American English was the most appropriate for use in this experiment, specific values needed to be established for each vowel in order to create synthetic vowels that sounded as natural as possible, and could be discriminated by native speakers of English. We found that Southern Californian American English was the most distinctive in terms of F1 values, as seen in Hagiwara (1997) (see Table 2).

Table 1. Average values of the Midland men. The /ɛ/ and /æ/ are nearly identical in terms of all F values. Their only distinctive variable is duration, which we want to keep equal (values taken from Hillenbrand, 2005).

	/ɛ/	/æ/
Dur	189	278
F0	127	123
F1	580	588
F2	1799	1952
F3	2605	2601
F4	3677	3624

Table 2. Average F1-F3 values (in hertz) of Southern California English vowels /ɛ/ and /æ/ from 15 Southern Californian speakers. When comparing with Table 1, we can observe that the F1 of /ɛ/ and /æ/ in this dialect differ more than in the Midland dialect from Hillenbrand (2005). F2 and F3 values of male speakers are nearly identical. Values taken from Hagiwara (1997).

	/ɛ/	/æ/
F1	529	685
F2	1670	1601
F3	2528	2524

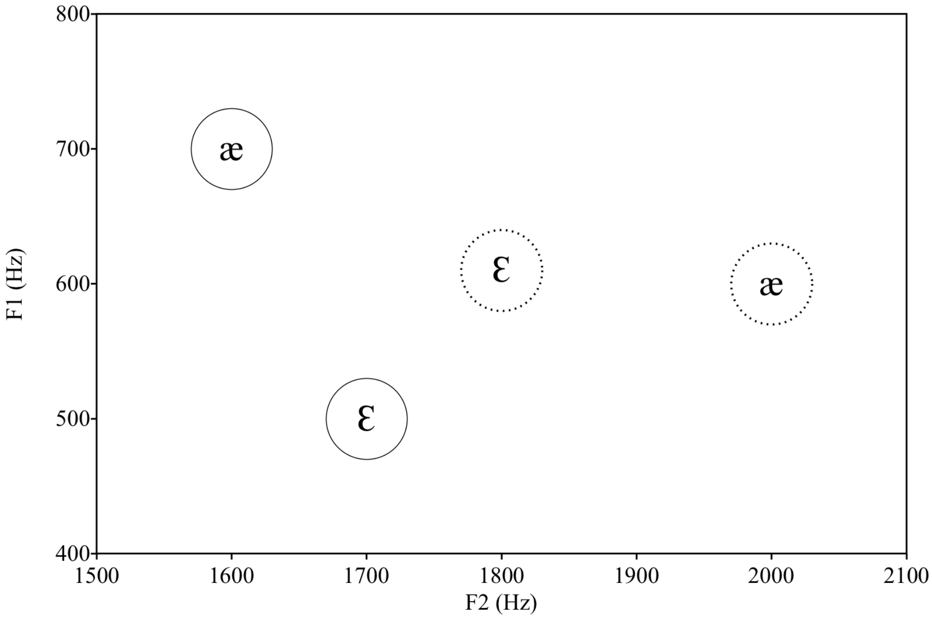

Figure 3. Vowels /ɛ/ and /æ/ of Southern California English (solid-line circle) and Midwestern English (dotted-line circle) adapted from Hagiwara (1997). Southern Californian /ɛ/ and /æ/ vary as a function of F1, while in Midwestern English they seem to vary as a function of F2. As the Dutch /ɛ/ varies from the English vowels in terms of F1, it is this value that we want to be variable, and not the F2.

The stimuli were then created by taking natural speech sounds from recordings made by one native Dutch and one native English male speaker. Recordings were made in a soundproof booth using a Sennheiser HF condenser microphone MKH-105, recorded on a Tascam CD-recorder: CD-RW900, 16 bits, 44100 Hz. The consonant sounds were spliced with synthetic American English male vowels created in Praat based on a programming script by Karin Wanrooij and Paul Boersma. Synthetic speech was used to allow us to control for all formant values besides F1 and for duration, in order to test subjects' ability to discriminate two phonemes that differ acoustically only with respect to height. This control was also done in order to allow for the potential use of the same stimuli in an EEG follow-up, in which the two sounds should not vary by more than one feature in order to know that an elicited MMN component is found as a result of the change in height and not as a result of a combination of changes (F1 frequency and duration, for example).

Piloting the stimuli

Four pilot tasks were performed by native speakers of Dutch and of English. Two were identification tasks and two were discrimination tasks. As Western American English /ɛ/ and /æ/ lie on a continuum from 500 to 800 Hz, the identification tasks were done in order to find the boundary of the F1 frequency between the two English vowels that native English speakers agree on. This was in order to find where the range is split into two phonemes by native speakers so that we do not expect non-native speakers to perceive a phonetic switch across a threshold that is actually not phonemic for native speakers. In addition, the same task was performed with Dutch speakers in order to make sure that they did not identify any point on the continuum as something other than /ɛ/.

Discrimination pilots were then performed as another check as to where the boundary is and in order to find a pair of F1 frequencies that differed enough to be considered two different categories by native speakers, but were minimally different such that they would not allow non-native speakers to perform at ceiling.

These pilots are described in detail below.

Table 3: Results of identification listening task by English native speakers of various backgrounds. Locations of origin in Loc; arrows represent the subjects' movement throughout their lifetime to different states and countries. Letters at the top of the columns refer to the subject's name. R refers to the pseudo-random order in which subjects heard the stimuli. Left column 500-800 represents the continuum from 500 to 800 Hz of the F1.

	P. R1	P. R2	C. R2	C. R1	J. R2	J. R1	J. R3	J. R4	N. R4	N. R3
500	kess	kuss	kiss	kess	kess	kess	kiss	kiss	kiss	kiss
520	kiss	kuss	kiss	kiss	kiss	kess	kess	kess	kess	kiss
540	kess	kess	kiss	kess	kess	kess	kess	kess	kess	kess
560	kess	kess	kess	kess	kess	kess	kess	kess	kess	kess
580	kess	kess	kess	kess	kess	kess	kess	kess	kess	kess
600	kess	kess	kess	kess	kess	kess	kess	kess	kess	kess
620	kess	kess	kess	kess	kess	kess	kess	kess	kess	kess
640	kass	kess	kess	kess	kess	kess	kess	kess	kess	kess
660	kass	kess	kess	kess	kess	kess	kess	kess	kess	kess
680	kass	kess	kess	kass	kess	kess	kess	kess	kess	kess
700	kess	kess	kess	kess	kess	kess	kess	kess	kess	kess
720	kass	kass	kass	kess	kess	kess	kess	kess	kess	kess
740	kass	kass	kass	kass	kess	kass	kess	kess	kass	kass
760	kass	kass	kass	kass	kass	kass	kess	kess	kass	kass
780	kass	kass	kass	kass	kass	kass	kass	kass	kass	kass
800	kass	kass	kass	kass	kass	kass	kass	kass	kass	kass
Loc.	New Zealand -> UK -> NL		Missouri -> Southern CA -> New York		Michigan -> Southern CA		Southern CA		Northern CA -> Southern CA	
Age	29		25		25		25		27	
Sex	F		F		M		F		M	

English Identification

Five native speakers of English (3 women, 2 men) of various dialects (see Table 1) were asked to listen to 92 pseudo-randomized trials of four pseudo-words: targets /kɛs/ and /kæs/ and fillers /kis/ and /kɪs/ (32 targets and 60 fillers). They were asked to write down what they heard, and then to move on to the next word. They were given the option of writing down *kess*, *kass*, *kiss*, and *keess*, as well as an indication of which real English words each of these pseudo-words rhymed with. They were also told that if they heard something else, they could write that instead. All fillers (/kɪs/ and /kis/) were identical across trials. The only thing that changed was the F1 value of the target vowels. Subjects heard two exemplars at each of the 16 F1 frequency intervals between 500 and 800 Hz. The results of the English speakers' performance on the target identification pilot are detailed in Table 3. They show that most speakers shifted from perceiving /ɛ/ to perceiving /æ/ at around 720 or 740 Hz. Thus, on the basis of these results, we decided to temporarily assume that the native boundary is at 730 Hz, and distributed our stimuli around this value for the subsequent discrimination pilots.

Dutch Identification

The Dutch identification pilot procedure was identical to the English task, with the sole difference being the consonant context in which subjects heard the target vowels. In this condition, five subjects (4 women, 1 man) heard words that replaced the English /k/ and /s/ with Dutch /χ/ and /s/. The stimuli were thus 32 targets of /χɛs/ and /χæs/ and 60 fillers of / χis/ and / χɪs/. The consonants were changed in order to make sure that subjects thought that they were listening to Dutch, just as would eventually be the case in the experimental conditions. In addition, the consonant change was done in order to make sure that subjects did not already perceive our target distinction when listening to Dutch. Results (which can be seen in Table 4) indicate that some subjects were able to perceive that values on the higher end of the spectrum were not good exemplars of their native /ɛ/ but nevertheless, most of the time they identified the entire spectrum as an /ɛ/. The sole exception was a subject who had been born and raised in Arnhem, an area of the Netherland that includes a very open /ɛ/ and open /ɪ/ (Paul Boersma, person. commun.). However, this subject did not identify a boundary between /ɛ/ and /ɪ/ in our target frequency range, meaning that this was not the cause for her different performance.

Table 4: Results of identification listening task by Dutch native speakers of various backgrounds (locations of origin in Loc. – i.e., location). Letters at the top of the columns refer to the subject's name. R refers to the pseudo-random order in which subjects heard the stimuli. Left column 500-800 represents the continuum from 500 to 800 Hz of the F1. Arrows represent the subjects' movement throughout their lifetime to different cities.

	N.	N.	M.	M.	L.	L.	M.	M.	V.	V.
	R1	R2	R2	R 1	R4	R3	R4	R3	R4	R3
500	ges	ges	ges	ges	ges	gis	ges	gus	ges	gis
520	gis	ges	ges	ges	ges	gis	gus	gus	ges	gis
540	ges	ges	ges	ges	ges	ges	ges	ges	ges	ges
560	ges	gis	ges	ges	ges	ges	ges	ges	ges	ges
580	ges	ges	ges	ges	ges	ges	ges	ges	ges	ges
600	ges	ges	ges	ges	ges	ges	ges	ges	ges	ges
620	ges	ges	ges	ges	ges	ges	ges	ges	ges	ges
640	ges	ges	ges	ges	ges	ges	ges	ges	ges	ges
660	ges	ges	ges	ges	ges	ges	ges	ges	ges	ges
680	ges	ges	ges	ges	ges	ges	ges	ges	ges	ges
700	ges	ges	ges	ges	ges	ges	ges	ges	ges	ges
720	ges	ges	ges	ges (gas)	ges	ges	ges	ges	ges	ges
740	gas	ges	ges (gas)	ges (gas)	ges	ges	ges	ges	ges	ges
760	ges	ges	ges (gas)	ges (gas)	ges	ges	ges	ges	ges	ges
780	ges	ges	ges (gas)	ges (gas)	ges	ges	ges	ges	ges	ges
800	ges	ges	ges (long e)	gas (ges)	ges	ges	geas	ges	ges	ges
Loc.	Gent (BE)		Arnhem ->Nijmegen -> Amsterdam		Eindhoven ->Amsterdam		Bodegraven		Woerden ->Amsterdam	
Age	30		25		24		27		25	
Sex	F		F		F		M		F	

Overall, however, this pilot suggests that most native Dutch speakers who are highly proficient speakers of English do not perceive a switch from /ɛ/ to /æ/ over the entire spectrum from 500 to 800 Hz. This indicates that although the subjects are highly proficient speakers of English, they do not use the phoneme inventory of English when identifying Dutch pseudo-words. This sets up our research question concerning discrimination, namely, if subjects do not identify English vowels differentially in Dutch context, will they be able to discriminate them in Dutch or in English context?

The identification pilots confirmed that, as expected, Dutch speakers have one phoneme category where English speakers have two. In the subsequent discrimination pilot, we tested where exactly this boundary lies for English speakers and whether Dutch speakers can perceive the difference between the two vowels.

Discrimination Pilot

On the basis of the English identification task we can conclude that the boundary between /ɛ/ and /æ/ is somewhere between 720 Hz and 740 Hz. More specifically, the boundary should be at 730 Hz, which is not only the halfway point between these two values, but also the point where the boundary is most likely to lie as judged from how the subjects performed in identifying sounds with an F1 of 720 and 740 Hz. While 720 Hz was identified as /ɛ/ 70% of the time, 740 Hz was identified as /æ/ 70% of the time (see Table 3). However, since not all subjects agreed on this boundary, we could not be convinced that we had found the definitive cutoff point between the two vowels from the identification pilot. A discrimination pilot was conducted in order to re-confirm this cutoff point, focusing now on the spectrum between 700 Hz and 800 Hz only. Besides this, the main reason for conducting the discrimination pilot was to find how different in frequency the two vowels must be in order to be discriminated by the English population most of the time (75-90%) when listening in English, while at the same time causing the Dutch to perform around chance level when listening in a Dutch context. The frequency pairs that were tested (in an AXB task) can be seen below in Table 5.

Table 5. F1 frequency pairs tested in an AXB discrimination task on both English and Dutch native speakers, in English and Dutch pseudo-word context respectively.

700 – 720				
700 – 740	720 – 740			
700 – 760	**720 – 760**	740 – 760		
700 – 780	**720 – 780**	**740 – 780**	760 – 780	
700 – 800	**720 – 800**	**740 – 800**	**760 – 800**	780 – 800

Each of the 15 pairs seen in Table 5 has four possible order combinations when placed in an AXB task. For example, the pair 700 Hz and 720 Hz can be placed in the following orders: 700-700-720; 700-720-720; 720-700-700; 720-720-700, which results in 60 possible combinations. All of the 60 combinations were heard twice for a total of 120 stimuli, meaning that each discrimination pair was heard eight times.

The pairs in bold in Table 5 were expected to be easiest to discriminate for the English and possibly also for the Dutch (with a difference of 60, 80, or 100 Hz between them), thus potentially non-options for our final stimulus pair, and more comparable to fillers in this pilot phase. The most difficult pairs to discriminate were predicted to be the ones with a difference of 20 or 40 Hz between them, and these were the pairs from which we expected to choose the stimulus pair (if the English speakers were able to discriminate them). Underlined stimuli were predicted to be difficult for both Dutch and English subjects to discriminate (with a difference of 20 Hz between them). Bold and underlined pairs were predicted to potentially be easy for the English and to potentially be difficult for the Dutch, and thus were our predicted choices for potential stimuli pairs. They are all 40 Hz and 60 Hz distance from each other (going diagonally on Table 5, the first diagonal line is 60 Hz difference, and the second diagonal line is 40 Hz difference).

Both Dutch and English subjects performed an AXB discrimination task programmed in Praat (Boersma and Weenink, 2011). Subjects heard three sounds and were to indicate which sound was most like the middle sound: the first or the third sound they heard. This pilot was identical in procedure to the experimental task (described in more detail below). The only difference was that here, each subject performed the task only in his or her native language.

The total results for both Dutch and English subjects can be seen in Table 6. Results show that five Dutch speakers (4 women, 1 man), and four English speakers (1 woman, 3 men) did not differ in their performance on this task. In fact, as can be seen in the table, on some pairs, the Dutch were better able to discriminate than the native English speakers. Only on four pairs were native English speakers better able to discriminate than the Dutch (seen in Green in Table 6). These pairs were all from the underlined group in Table 5, which we had predicted would be difficult for all speakers.

Table 6. Results of the discrimination pilot. Values represent mean scores of five subjects in the Dutch condition and four subjects in the English condition. Red = Dutch subjects performed better than English subjects, Green = English subjects performed better than Dutch subjects; Black = English and Dutch subjects performed virtually equally.

DUTCH				
700-720 = 0.5				
700-740 = 0.725	720-740 = 0.5			
700-760 = 0.9	720-760 = 0.775	740-760 = 0.525		
700-780 = 0.95	720-780 = 0.85	740-780 = 0.725	760-780 = 0.575	
700-800 = 0.95	720-800 = 0.925	740-800 = 0.875	760-800 = 0.7	780-800 = 0.575
ENGLISH				
700-720 = 0.584				
700-740 = 0.542	720-740 = 0.583			
700-760 = 0.792	720-760 = 0.792	740-760 = 0.625		
700-780 = 0.958	720-780 = 0.833	740-780 = 0.75	760-780 = 0.583	
700-800 = 0.875	720-800 = 0.875	740-800 = 0.875	760-800 = 0.542	780-800 = 0.625

As a result of this pilot, the values of F1=710 Hz and 750 Hz were chosen for the experimental task. This was done because it was very near to the values found in the discrimination pilot to be more easily discriminated by English speakers, and because 40 Hz was found to be the difference in F1 values that needed for both Dutch and English to be able to discriminate at all. Note that the values of 710 and 740 Hz are quite high when compared to previous work on American English vowels (Hagiwara, Hillenbrand). This may have been a result of a compensation for the lack of durational cues, causing subjects to shift their category boundary to the higher end of the F1 spectrum. More on durational versus spectral cues will be addressed in the discussion section.

Experiment

Participants

Participants were adults recruited from the student and administrative population of the University of Amsterdam. All participants were native speakers of Dutch, and highly proficient speakers of English who had been learning English from primary school (mean age of acquisition = 9.44 years old; SD = 1.64). After performing the experiment, subjects completed an exit questionnaire involving questions about language proficiency (Appendix D). The language questions were inspired by Li et al.'s (2006) language history questionnaire. Based on this self-report and the instructions on the information brochure (Appendix B), no participants reported suffering from any language or hearing problems. There were 27 participants, divided into two groups: one group ("E-D"; *n*=13) heard the English condition first and the other group ("D-E"; *n*=14) heard the Dutch condition first. The groups were matched for age: the average age of group E-D was 25.9 years old, and the average age of group D-E was 27.8.

According to self-reports, subjects spoke an average of 7.28 hours (SD = 9.3) of English per week, were exposed to an average of 13.12 hours of audio-visual (media) input per week (SD = 13.24), and spent an average of 11.92 hours of reading English per week (SD = 11.21), although each of these numbers varied greatly by subject.

Participants rated themselves on a scale of 1 to 10 in production and comprehension abilities, with 1 being the worst, and 10 being native. Their average scores (not including those who were removed from analysis, described later) were 7.94 (SD = 0.65) in production and 8.66 (SD = 0.8) in comprehension, making them highly proficient bilinguals. Self-report scores for each subject (including those who were excluded for analysis) can be found in Appendix E.

Stimuli

The experimental stimuli were made up of consonants from the natural, recorded speech of male native speakers of Dutch and of English. Along with the synthetic vowels, they created the C-V-C pseudo-words of English (/kɛs/ and /kæs/) and Dutch (χɛs/ and /χæs/) that were also used in the piloting phase. The stimuli were presented to the participants over headphones at a sound pressure level of 70 dB (standardized with an external sound card: Edirol UA-25, 24 bit, 96 KHz). The formant values of the /æ/ and /ɛ/ were the following: F1=710 Hz for /ɛ/ and 750 for /æ/; F2=1650, F3=2525, F4=2700 for both /ɛ/ and /æ/. The fundamental frequency (F0) for both was 75 Hz. The total duration of /kɛs/

and of /kæs/ was 386 ms. The total duration of /χɛs/ and of /χæs/ was 370 ms. A visual representation of each of the four stimuli can be seen in Figures 7-10. The difference in duration between the English and Dutch pseudo-words is due to the difference in duration of the individual consonants, which is due to the quality of the /s/ being different in Dutch and English, and due to the use of a consonant /k/ in English and /χ/ in Dutch. However, the lengths of the target vowels were always 90 ms (100 ms with 5 ms overlap with the surrounding consonants to mimic coarticulation).

During all of the phases of the experiment, the words were presented with an inter-stimulus interval of 1 second. This inter-stimulus interval (ISI) was chosen because longer ISI are shown to measure categorical perception rather than acoustic perception, which is what we intended to measure here (Näätänen et al., 2007; Strange and Shafer, 2008; Werker and Logan, 1985).

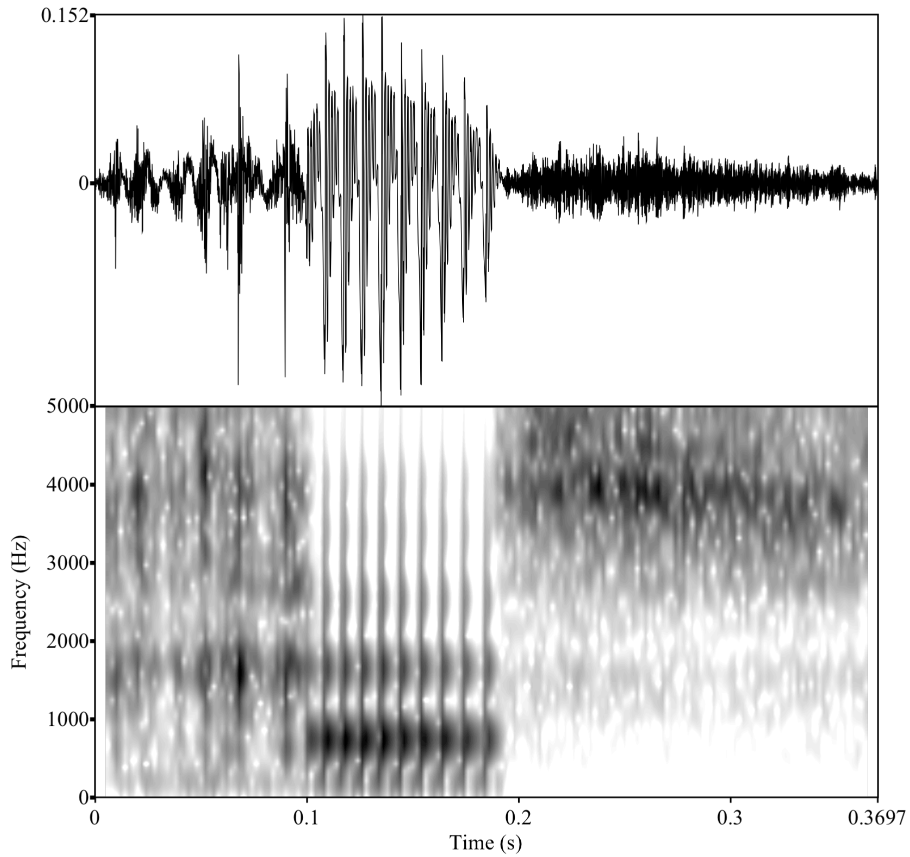

Figure 7: Waveform (top) and spectrogram (bottom) of one of the Dutch stimuli: a synthetic vowel with an F1 of 710 Hz (/ε/) inserted between naturally spoken Dutch consonants / χ / and /s/.

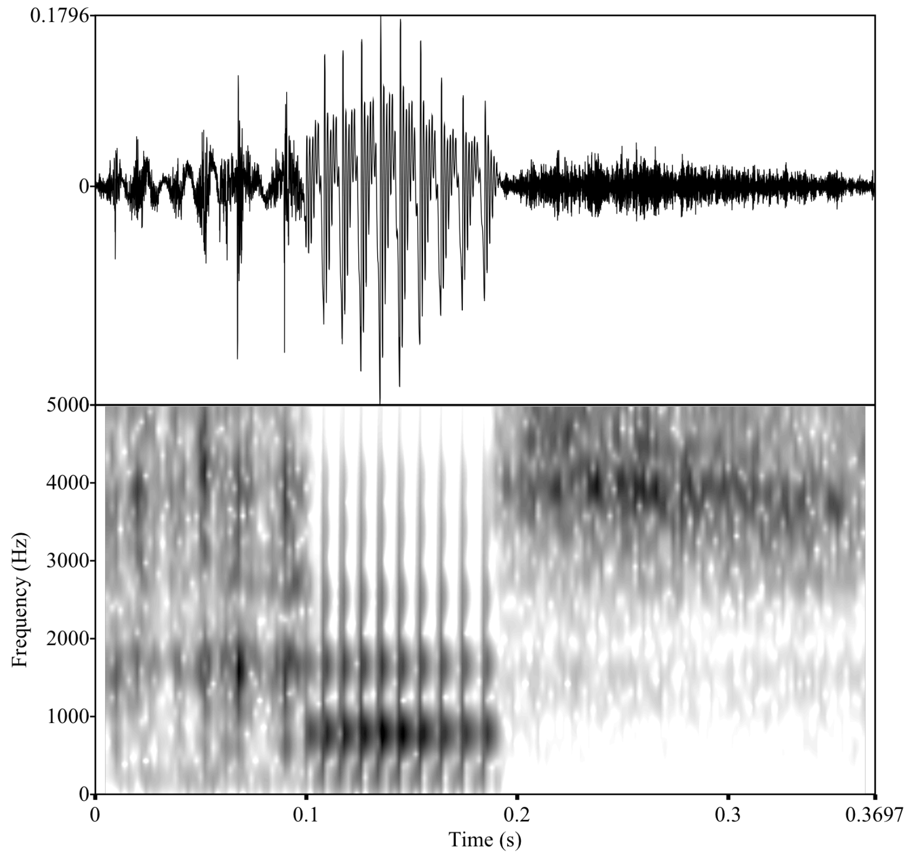

Figure 8: Waveform (top) and spectrogram (bottom) of one of the Dutch stimuli: a synthetic vowel with an F1 of 750 Hz (/ε/) inserted between naturally spoken Dutch consonants /χ/ and /s/.

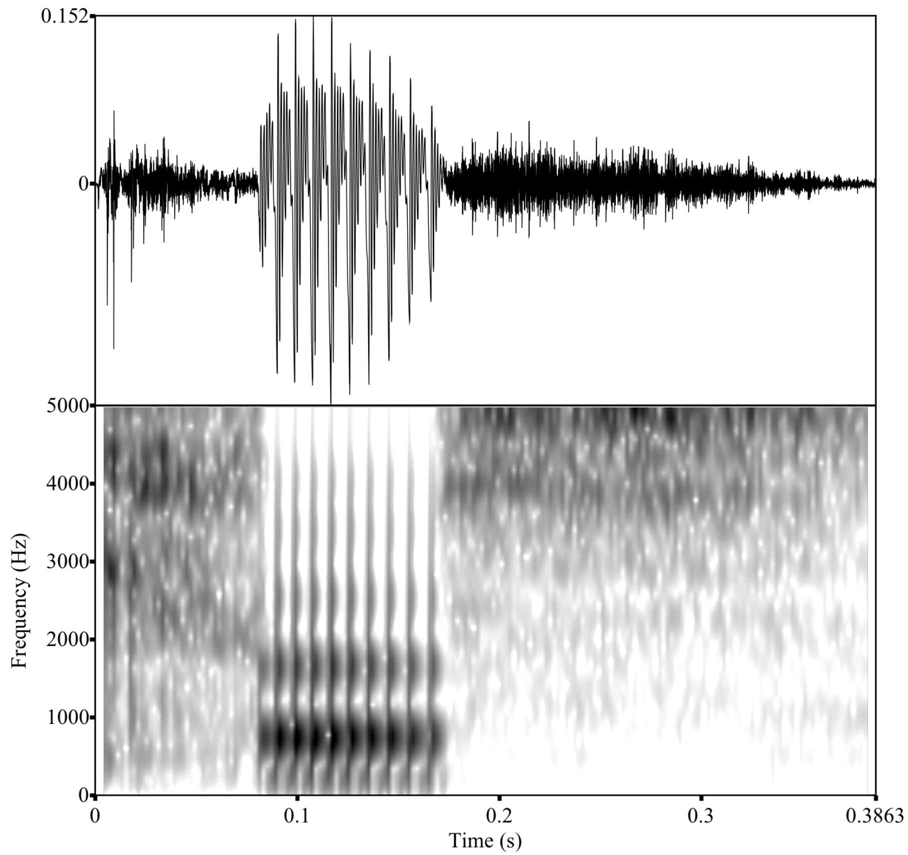

Figure 9: Waveform (top) and spectrogram (bottom) of one of the English stimuli: a synthetic vowel with an F1 of 710 Hz (/ɛ/) spliced between naturally spoken American English consonants /k/ and /s/.

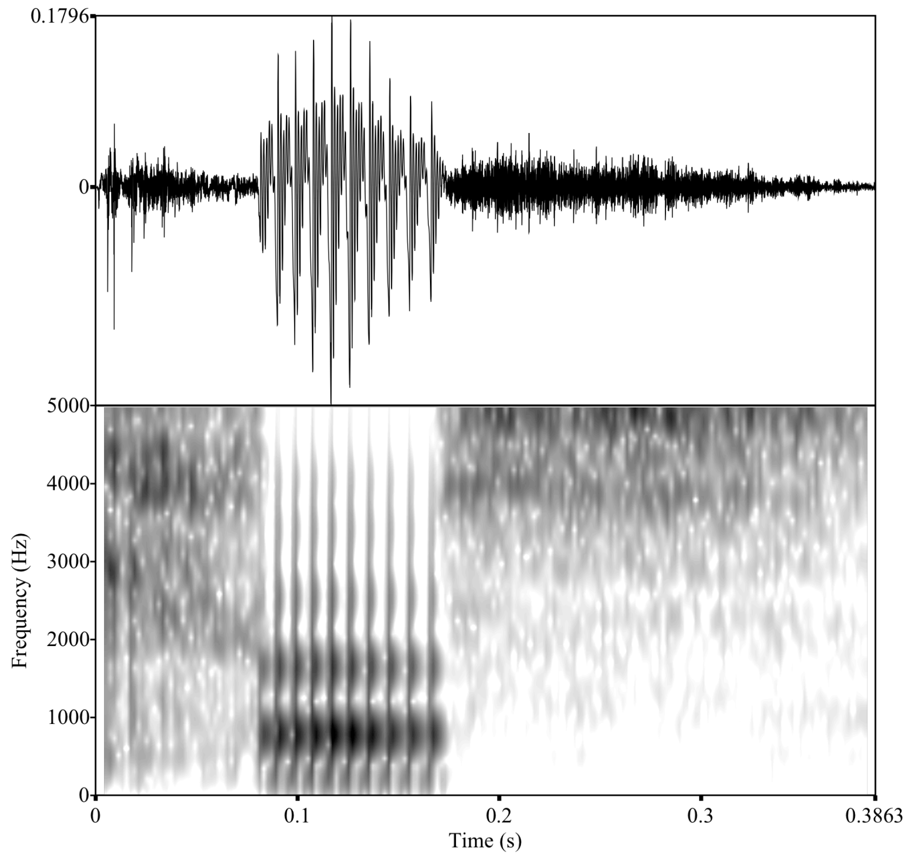

Figure 10: Waveform (top) and spectrogram (bottom) of one of the English stimuli: a synthetic vowel with an F1 of 750 Hz (/æ/) inserted between naturally spoken American English consonants /k/ and /s/.

Design

The task involved listening to pseudo-words (nonwords that follow the rules of each language to sound as if they could be real words) of Dutch and English. Each subject heard stimuli in a Dutch and in an English condition. The order of the conditions was counterbalanced across subjects. The pseudo-word was of the format C-V-C (consonant-vowel-consonant), where the vowels were either /ɛ/ and /æ/. Stimuli were presented in an AXB paradigm. Typical discrimination tasks use AX, AXB/ABX/XAB, or Oddball paradigms, each of which have certain advantages and disadvantages. For the purpose of this experiment, AXB was chosen because as compared to AX, AXB has a higher memory load (Werker and Logan, 1985) and higher stimulus uncertainty (Strange and Shafer, 2008). This means that subjects are not only required to hold more auditory traces in working memory while more stimuli are presented during each trial of an AXB task, but there is less predictability of the sequence order for the participant. In our task there were 4 possible orders in which subjects could hear the sequences: /ɛ/-/ɛ/-/æ/, /ɛ/-/æ/-/æ/, /æ/-/ɛ/-/ɛ/, and /æ/-/æ/-/ɛ/. Furthermore, an AXB task was chosen over an ABX/XAB task because in AXB, A and B are equally distant from the comparison/target stimulus, requiring a medium-level of memory load. Stranger and Shafer (2008, p. 161) state: "As the complexity of the discrimination task increases, performance outcomes begin to reflect not only basic auditory sensory capabilities but increasingly the cognitive processes involved in categorization (including implicit labeling)." In addition, with greater cognitive load, subjects are not only less able to detect acoustic differences, but they tend to fall back on language-specific perceptual influences. Since we wanted to test the subjects' ability to use language-specific perceptual systems based on which language they were listening to, the AXB task was selected over the AX task. And since it is already known that /ɛ/ and /æ/ are difficult contrasts for Dutch speakers to make, an AXB task was selected over an ABX task, as we did not want the task to be overly difficult.

Filler stimuli were purposefully not included so as not to provide any reference point for the speakers to compare the target phonemes to.

Procedure

The experimental duration was around 45 minutes, including a questionnaire that subjects filled in on the computer at the end of the experiment.

The experiment took place in a quiet room at the University of Amsterdam. Participants were instructed that on each trial, they would hear speech-like sounds of English or of Dutch (depending on block), and that they were to indicate which sound the middle sound was most similar to, the first sound or the third sound. They made a choice by pressing either the “z” key (for first) or the “/” key (for third) on the keyboard in front of them. The specific instructions can be found in the information brochure in Appendix B.

The experiment was programmed in Praat (Boersma and Weenink, 2011). The audio was heard over headphones plugged in to an external sound card (Edirol UA-25, 24 bit, 96 kHz). Instructions were always written on the computer screen during the experiment either in English or in Dutch depending on the block. This was the only thing that was written on the screen and was done not only to help the subjects remember the instructions but also to reinforce their language mode for each block by providing additional language cues in either English or Dutch. No feedback was provided at any point during the experiment.

Results

In each group there was one subject that was an outlier in terms of age (subject 20 in group E-D and subject 23 in group D-E). However, these subjects were not outliers with respect to their reaction times, their scores, or their linguistic backgrounds. Thus, they were not excluded from analysis. However, there were two subjects, one in each group, who had to be excluded on the grounds of language background. One of the subjects (subject 26) had lived in the US from the ages of 3 to 6 years old, and the other had grown up with a mother from Guyana, who spoke British English to her from birth (subject 21). These subjects were thus outliers in their backgrounds and as such were excluded in order for the subject group to remain linguistically homogeneous. The average age, average overall scores, and the results of the t-tests did not change with the exclusion of these subjects. Here, only the results excluding these subjects are presented. Tables with the results including these subjects can be found in Appendices F.

We found in the remaining sample group that our subjects performed significantly above chance in both the English condition (24 out of the 25 subjects performed above 50%; $p=0.0000007$ in a binomial test), and in the Dutch condition (23 out of the 25 subjects performed above 50%; $p=0.000009$ in a binomial test). Because we compared the performance of the same population in two different conditions (i.e., Dutch and English), we used a dependent samples (paired) *t*-test, which compared the means of the subjects' scores in each of these conditions. That is, we performed a *t*-test against zero (which is the default mean in a *t* distribution) on the values in column ΔScore E-D (from Table 7 and 8 together). This *t*-test finds no significant difference between the average score on the English condition (*M*=0.69, *SD*=0.16) and the Dutch condition (M=0.68, SD=0.14) [$t(24)=0.31$, one-tailed $p=0.38$]. The difference in the overall means of Score E and Score D was 0.007. The 95% confidence interval for the difference between scores is -0.038 to 0.051, meaning that the true difference in scores lies within this interval, which includes a possibility of a difference of zero. Indeed we see from the *t* and *p* values that the difference between performance on English and Dutch conditions cannot be attributed to more than chance variation, as the difference in means is not statistically significant. As the *t* value is approaching zero, it indicates that the probability of the two scores being the same is quite high, and the *p* value is above .05 also meaning that the probability of obtaining a difference at least as high as we have is quite high and we cannot reject our null hypothesis. A one-tailed *p* value is reported, as we had a clear prediction that English scores should be higher than Dutch scores. These predictions, however, were not confirmed, as we did not find a significant difference between participants' discrimination abilities in the Dutch and English conditions.

Next we sought to find whether there was a training effect, as we noticed that subjects seemed to discriminate slightly better (have higher mean scores) on the second condition with which they were presented (as can be seen in Tables 7 and 8). We again perform a *t*-test against zero, this time on the columns ΔScore2-1 in Tables 7 and 8 together, representing the difference between the scores on the first and second condition. A dependent (paired) samples *t*-test found that there was a significant difference between the average score on the first condition and the second condition [$t(24)=2.61$, one-tailed $p=0.01$]. Our *t* value is far from zero, indicating that the probability of the two means being the same by chance is low, and our *p* value is well below .05, indicating that the probability of obtaining at least such a difference in means is also low. Thus we can say that the difference in performance from the first to the second condition was significant, and can reject our null hypothesis that this difference is due to chance. Again we report a one-tailed *p* value as we had a clear prediction that subjects would perform better on the second trial than on the first as a result of concentrated experience during the experiment. These predictions were confirmed as, overall, the group's performance improved significantly during the second block as compared to the first block. The difference in mean is of 0.05 between the scores with a 95% confidence interval with a lower limit of 0.0011 and an upper limit of 0.089 tells us that the difference in scores of the population lies in this range, which must be at least higher than zero.

Next, we examined whether this visible training effect was only a result of increased experience and focus on the same target (i.e., on the vowels), or whether there was a combination of improvement from both increased experience and from the language mode. That is, did the subjects who performed the English condition second (D-E) perform better than those who heard the Dutch condition second (E-D), or was the training effect independent of the language? By comparing the difference in performance on the first and second condition of D-E against group E-D, we found that both groups performed better on the second condition, independent of the language context. The difference between group E-D and D-E on their mean ΔScore 2-1 was -0.0096, with a 95% confidence interval with a lower limit of -0.09 and an upper limit of 0.071, indicating that the difference in means of the entire population would lie within this range, which may actually be zero. An independent samples *t*-test finds no difference between the groups' mean scores on condition 1 and 2 [$t(23)=0.246$, one-tailed $p=0.4$].

Our *t* value is approaching zero, and our two-tailed *p* value is well above .05, meaning that we cannot attribute this difference in means to more than chance. Our results indicate that subjects' performance did not significantly improve in the second block as a result of language mode but did so independently of language. We may assume that the reason for the training effect is thus a result of concentrated exposure to the target vowel distinction during the course of the experiment.

Table 7. Individual and average scores (out of 1) and reaction times (RT; in seconds) of subjects in the group that heard English first and Dutch second (group E-D). The Δ scores refer to the difference between the score on the English (E) condition and the Dutch (E) condition, as well as the difference in score from the first to the second condition, and the difference between the RT in each of these conditions. Sex and age of participants is also included. Bold numbers indicate where subjects performed better on the first condition than the second condition.

ENGLISH – DUTCH (Group E-D)										
	Score E	**Score D**	**ΔScore E-D**	**ΔScore 2-1**	**RT E (s)**	**RT D (s)**	**ΔRT E-D (s)**	**ΔRT 2-1 (s)**	**Sex**	**Age**
Subject 1	0.57	0.6	-0.03	0.03	4.85	4.08	0.77	-0.77	M	22
Subject 3	0.74	0.91	-0.17	0.17	3.96	3.57	0.39	-0.39	F	21
Subject 5	0.72	0.83	-0.11	0.11	5.14	4.06	1.08	-1.08	M	24
Subject 8	**0.55**	**0.45**	0.1	-0.1	4.51	3.8	0.71	-0.71	F	21
Subject 10	0.56	0.63	-0.07	0.07	4.99	4.53	0.46	-0.46	M	29
Subject 12	0.77	0.83	-0.06	0.06	4.07	4.04	0.03	-0.03	F	23
Subject 14	**0.98**	**0.93**	0.05	-0.05	3.83	3.59	0.24	-0.24	M	21
Subject 16	**0.84**	**0.78**	0.06	-0.06	3.96	3.96	0	0	F	25
Subject 18	0.51	0.58	-0.07	0.07	5.29	4.37	0.92	-0.92	M	21
Subject 20	**0.61**	**0.6**	0.01	-0.01	4.98	4.61	0.37	-0.37	F	46
Subject 22	0.51	0.7	-0.19	0.19	4.28	3.93	0.35	-0.35	M	29
Subject 24	0.56	0.62	-0.06	0.06	4.59	4.29	0.3	-0.3	M	29
Average	**0.66**	**0.71**	**-0.05**	**0.05**	**4.54**	**4.069**	**0.47**	**-0.47**		**25.9**

Table 8. Individual and average scores (out of 1) and reaction times (RTs; in seconds) of subjects in the group that heard Dutch first and English second (group D-E). The Δ scores refer to the difference between the score on the English (E) condition and the Dutch (E) condition, as well as the difference in score from the first to the second condition, and the difference between the RT in each of these conditions. Sex and age of participants is also included. Bold numbers indicate where subjects performed better on the first condition than the second condition.

DUTCH – ENGLISH (Group D-E)										
	Score D	**Score E**	**Δ Score E-D**	**Δ Score 2-1**	**RT D (s)**	**RT E (s)**	**ΔRT E-D (s)**	**ΔRT 2-1 (s)**	**Sex**	**Age**
Subject 2	0.57	0.84	0.27	0.27	4.78	4.3	-0.48	-0.48	F	21
Subject 4	0.75	0.82	0.07	0.07	4.54	4.27	-0.27	-0.27	F	26
Subject 6	0.55	0.61	0.06	0.06	6.11	5.99	-0.12	-0.12	M	19
Subject 7	0.86	0.94	0.08	0.08	4.09	3.95	-0.14	-0.14	F	24
Subject 9	**0.73**	**0.63**	-0.1	-0.1	4.05	3.73	-0.32	-0.32	M	24
Subject 11	0.93	0.94	0.01	0.01	4.44	4.25	-0.19	-0.19	M	30
Subject 13	**0.62**	**0.58**	-0.04	-0.04	4.69	3.76	-0.93	-0.93	F	25
Subject 15	0.78	0.78	0	0	4.35	4.15	-0.2	-0.2	M	26
Subject 17	0.78	0.96	0.18	0.18	3.88	3.72	-0.16	-0.16	F	25
Subject 19	**0.57**	**0.55**	-0.02	-0.02	4.04	3.87	-0.17	-0.17	M	23
Subject 23	**0.51**	**0.48**	-0.3	-0.3	3.57	3.6	0.03	0.03	F	65
Subject 25	0.56	0.75	0.19	0.19	4.24	4.04	-0.2	-0.2	M	22
Subject 27	0.49	0.53	0.04	0.04	4.66	3.45	-1.21	-1.21	M	32
Average	**0.67**	**0.72**	**0.055**	**0.055**	**4.42**	**4.08**	**-0.34**	**-0.34**		**27.8**

We also analyzed whether reaction times of our subjects differed as a result of language mode (i.e., whether subjects performed faster in the English than in Dutch condition). A dependent (paired) samples *t*-test against zero was also conducted on the mean reaction time differences between English condition (M=4.3, SD=0.61) and Dutch condition (M= 4.25, SD=0.52). The values were taken from column ΔRT E-D in Tables 7 and 8 together. The difference in the means was 0.05 with a 95% confidence interval with a lower limit of -0.169 and an upper limit of 0.269, indicating that our population mean is within an interval that includes zero. The means did not differ significantly [$t(24)=0.47$, two-tailed $p=0.64$). As our p value is above .05 and our t value is approaching zero, we could not reject the null hypothesis that the differences varied by chance and thus could not say that there was any meaningful difference between the reaction times by language.

We also tested whether reaction times varied significantly from the first to the second condition, comparing the difference in means found in Tables 7 and 8 in column ΔRT2-1. If the difference in reaction times were to be significant, it might help support the finding of a training effect, indicating that with more experience, the test became increasingly easy for the subjects. Again a dependent (paired) samples *t*-test was performed against zero (default mean). It was found that the difference in means was -0.39 (with a 95% confidence interval with a lower limit of -0.54 and an upper limit of -0.25, a range which crucially does not include zero). The *t*-test shows that the differences in mean scores on the first and second condition were significant [$t(24)=-5.8$, $p=.0000028$]. As our *t* value was far from zero, and our *p* value is far under $p<.01$, we can safely assume that the difference between the reaction times on the first and second conditions were not a result of chance, and can thus, reject the null hypothesis. This adds to the finding that subjects performed better on the second condition than on the first, showing that they also performed faster on the second condition. This supports the notion that a training effect took place throughout the course of the experiment as a result of concentrated exposure to our target sounds.

Finally, we analyzed whether the difference in reaction time from the first to the second condition was related to the language order. An independent samples *t*-test was conducted to compare the reaction times of group E-D versus group D-E. The difference in means was of -0.13, with a 95% confidence interval with lower limit of -0.42 and an upper limit of 0.15, a range which includes zero. We found that the groups did not differ significantly in their reaction times [$t(23)=-0.96$, two-tailed $p=0.34$], as our t approached zero and our p value was well above .05. We report a two-tailed p value, as we had no clear predictions with respect to the outcome. As was the case above when comparing groups E-D and D-E with respect to their discrimination ability, we could not say that the difference found between these groups was more than a result of chance differences, and could thus not reject the null hypothesis. As a result, we cannot draw any conclusions with respect to one sub-group of subjects performing faster than the other.

Discussion

Our results indicated that Dutch native speakers who are highly proficient, early learners of English were able to perform above chance on the discrimination task, but did not show a difference in their ability to discriminate between English phonemes depending on the linguistic context. These results are contradictory to the predictions we had set up as a result of the work of Escudero and Boersma (2002), although they are in accordance with the work of Winkler et al. (2003), who did not find context effects in identification (both of these studies were described in Introduction section entitled "Context effects on identification").

In addition, in the identification pilot Dutch subjects identified the entire spectrum between 500 and 800 Hz as /ɛ/ according to their Dutch phoneme categories, yet in the discrimination pilot, they performed similarly to native speakers. These results are compatible with the work of Gerrits and Schouten (2004) who claimed that identification was not necessarily predictive of discrimination performance as has usually been assumed.

We also found a significant training effect, with subjects performing better on the second condition than on the first. This did not depend on which language they were listening to first or second. This indicates that, as researchers such as Maye and Gerken (2000, 2001) have shown, subjects exposed concentrated stimuli can be trained to perceive difficult target contrasts in a very short period of time.

However, questions remain with respect to certain details of our results. According to the studies outline in the introduction, we would not expect subjects to perform like natives in this task. However, almost all of our subjects performed above chance. Furthermore, the performance range for almost half of our subjects was under 65% in both conditions (n = 6 in E-D condition; n = 5 in D-E condition), leaving us with the majority performing above 65% on both conditions. How then can we explain that all subjects performed above chance and that half of them even performed near ceiling?

One hypothesis for the variation between subjects may be that some subjects may have spoken non-standard varieties of Dutch or additional languages with different vowel categories, and that these speakers could transfer their knowledge of these dialects when learning the English vowel inventory. However, our subjects were from all over the Netherlands, with 12 being born in the Western area between Amsterdam and Leiden (including Sassenheim, Heemstede, Leidschendam, Aalsmeer) area and additional 2 others being born in the Northwest (Hoogkarspel, Alkmaar). Only one subject was born in the south (Boxtel), and 4 subjects were born in the east (Hellendoorn, Volkel, Nijmegen, Arnhem). An additional 4 subjects did not specify where in the Netherlands they born. All but 4 subjects were living in Amsterdam (the remaining 4 lived in nearby places: Hoogkarspel, Utrecht, Haarlem, and Egmond aan Zee). Thus, all of our subjects fell in the category of Northern Standard Dutch as defined by Adank et al. (2007). In addition, most subjects spoke at least one other language, as can be seen in the language questionnaire results included in Appendix E. However, none of the subjects listed any language that discriminates between /æ/ and /ɛ/ as one of their additional languages. Furthermore, even if they had indicated some knowledge of such a language, it has been shown that simply having knowledge of an additional language does not make one more sensitive to novel speech contrasts (Werker, 1986), meaning that knowledge of phonemes of an L3, L4, etc. will not necessarily help in discriminating phonemes in the L2. Therefore, these explanations concerning linguistic background are unsatisfactory and do not account for the overall results or the differences between subjects.

A second explanation has to do with the learning method. Dutch children begin learning English at an early age in elementary school. Most subjects that we tested began learning English at school between the ages of 8 and 11 years old (*M*=9.44, *SD*=1.64). However, many previous studies have shown that early bilinguals may not be able to form native-like phonetic categories despite their many years of experience with the L2. Peltola et al. (2003) showed that Finnish adults who had learned English at school and who were considered highly proficient did not show the ability to discriminate at a native-like level. The authors argue that learning in a classroom environment may not allow speakers to form long-term native-like memory traces in the way learning from immersion can. However, Sundara and Polka (2008), using a similar AXB-style task as in our study, have recently shown that adults who were native speakers of Canadian English and had learned French at 5-6 years old in school immersion programs still performed like monolingual English speakers on a discrimination task involving coronal stops native of French and English (in C-V syllables: French /d/+V and English /d/+V). In addition, they performed significantly worse than simultaneous bilinguals. In other words, early bilinguals were not able to modify their native categories, which included only one of the phoneme categories in question, despite many years of input and use on a daily basis in an immersion context. In addition, Gulian et al. (2007) found that while subjects benefitted from both bimodal distribution training and explicit instruction that there are two phonemes to be discriminated, the two methods together actually hampered subjects' ability to learn the contrasts. Our subjects learned primarily by explicit means: their elementary school teacher presumably instructed them there was a sound /ε/ and a sound /æ/ in English. However, presumably the primary input for the students was not from live, native speakers or immersion. Instead it was from a non-native speaker of English (teacher) and from ubiquitous English media from various dialectal sources. According to the studies presented here, a combination of non-native experience and explicit learning should have had a detrimental effect in the phonetic learning of our subjects. However, this was evidently not the case for our subjects, who performed above chance, especially for the half of our subjects who scored as well as 75% or higher. Indeed our subjects were homogeneous in the English background (the way they were taught).

One final explanation for the variation between subjects that may account for some scoring only slightly above 50% and some score nearer to ceiling is based on Escudero and Boersma (2004)'s claim that speakers learning different dialects of English will focus on different acoustic features, and thus cause them to assimilate categories differently. Escudero and Boersma (2004) found that native speakers of Scottish English relied more on spectral cues to discriminate /i/ and /I/, while Southern English speakers relied on both duration and spectral cues to discriminate the same vowels. They also showed that native Spanish speakers who were learning one of these two dialects performed identification task differently as a result of which dialect they had been learning. Those learning Scottish were able to assimilate the /i/ and /ɪ/ into two categories (Best, 1995), mapping the /i/ onto the Spanish /i/ and the /ɪ/ onto the Spanish /e/. On the other hand, those learning Southern English showed a pattern of single-category assimilation, in which they mapped both the /i/ and the /ɪ/ onto the Spanish category /i/. This caused semantic confusion for Southern English as a result of the use of the wrong acoustic cue. In American English, /æ/ and /ɛ/ differ not only with respect to the first formant but also with respect to duration. However, in our experiment, duration was kept equal, eliminating one auditory cue that subjects may use to categorize the vowel sounds. As a result, a similar phenomenon to that shown in Escudero and Boersma (2004) may have occurred here in subjects who performed on the lower end of the spectrum. It may have been the case that for some reason (possibly less exposure to native English, or more exposure to a different dialect of English), the subjects who performed only slightly over 50% were relying on the durational cue, which was unavailable to them in this experiment. It is an empirical question whether another subject group that is given more natural sound samples with the durational cue added would perform like native speaker.

Thus, at this point, the most basic explanation for the variability across subjects can be attributed most basically to individual differences. Some people are better at perceiving certain sounds, at acquiring various aspects of a second language, at being motivated to learn to sound and to perceive the L2 like a native. Some may have had more experience with native speakers than others (although none had lived in an English speaking country for more than 12 months), and the majority (n = 21) of the subjects who were included in the analysis had not lived in an English speaking country at all (Appendix E). More work can be done in the future in understanding the individual differences between subjects in their ability to learn various aspects of language.

The finding that our subjects performed consistently above chance is interesting and surprising considering the previous research. However, we cannot reliably say that these results are not skewed by significant training effect that was found. In the future, a shorter paradigm performed on different subject groups may limit the highly concentrated exposure to the target sound within the experimental setting in order to more clearly understand the nature of the subjects' phoneme categories.

The next question is how can we explain our main finding: the fact that phonetic categories do not seem to be context-dependent? One explanation is that the given linguistic contexts were not enough to cause a switch in language mode. The pseudo-words may not have triggered the desired language to become activated. In this case, the stimuli might be embedded in real sentences; for example: "now listen to the words /kɛs/ and /kæs/" (similar to Escudero & Boersma, 2002). This would allow the correct language mode to be activated and at the same time allow for pseudo-words to be used, without involving semantic cues about which phoneme is necessary (as was the case in Winkler et al., 2003). Similarly, it could be that our experimental setup was not conducive to activating the appropriate language modes (Grosjean, 1998). It is possible that in a future rendition of such a study, subjects might perform differently if all instructions and communications were done by an English native speaker for one condition, and by a Dutch native speaker for the other condition (as was done in Winkler et al., 2003). A different experimental paradigm might compare the scores of two larger groups of speakers who each participate in only one of the conditions.

Another explanation for the lack of difference in performance between the two conditions may be more similar to Flege's interpretation (as detailed in the Introduction section entitled "Bilingual speech perception"). It may be the case that when these subjects learned the phoneme inventory of their L2, they mapped it on to their L1 space, and as a result, also changed their L1 categories slightly in order to fit this new category. This would mean that the subject has formed a hybrid of Dutch and English phoneme inventories that is always activated, no matter the language mode. The only way to be sure that this hypothesis applies to our results would be to compare discrimination abilities with monolingual English and monolingual Dutch subjects. In such a control experiment, we would expect these bilinguals to perform differently from monolinguals of each of the languages. A control experiment with monolingual Dutch subjects was not conducted within the frame of this paper as finding monolingual Dutch subjects would have posed a rather difficult task and may would have raised the control group's mean age significantly in comparison to the experimental group. In future experiments investigating language mode on phonemic discrimination abilities, choosing different languages may resolve this issue.

Future work

Besides the future work that has already been suggested above, a larger extension of the study can be proposed at this time. The present work was conducted with the intent to eventually expand to an electrophysiological study. The justification for doing this is that with behavioral measures you can only assume what the brain is doing as evidenced by the subjects' behavior. However, in order to truly understand whether subjects are discriminating phonemically (rather than purely acoustically), or whether they are discriminating at all, EEG measures should be taken. As was mentioned before, a large MMN component would indicate a phonemic change-detection, while a smaller MMN component would indicate an acoustic change-detection. No MMN would indicate a lack of ability to discriminate at all, acoustically or phonetically. Such an extension would give us more solid ground to stand on when discussing the nature of bilinguals' phoneme categories.

Conclusion

The present study investigated the nature of phonetic categories in highly proficient, early bilinguals. Subjects performed an AXB discrimination task in both Dutch and English context. A difference in performance as a result of language mode was not found. However a training effect from the first to the second condition was found, with subjects showing significantly higher scores and faster reaction times in the second condition as compared to the first condition (regardless of which language mode came first or second). Theoretical implications have been discussed, and future work to clarify our results have been proposed, most notably the addition of control groups, taking measures for accounting for language mode, and the addition of an EEG extension experiment.

References

Aaltonen, O., Niemi, P., Nyrke, T. and Tuhkanen, M. (1987). Event-related brain potentials and the perception of a phonetic continuum. *Biological Psychology, 24*, 197-207.

Adank, P., van Hout, R. and van de Velde, H. (2007) An acoustic description of the vowels of northern and southern standard Dutch II: Regional varieties. *Journal of the Acoustical Society of America*, *121* (2), 1130-1141.

Best, C. (1995). A direct realist view of cross-language speech perception. In W. Strange (Ed.) *Speech perception and linguistic experience: Theoretical and methodological issues* (pp. 171–204). Timonium, MD: York Press.

Best, C., McRoberts, G. and Goodell, E. (2001). Discrimination of non-native consonant contrasts varying in perceptual assimilation to the native phonological system. *Journal of the Acoustic Society of America*, *109*, 775-794.

Boersma, P. and Weenink, D. (2011). Praat: doing phonetics by computer [Computer program]. Version 5.2.29, retrieved 12 July 2011 from http://www.praat.org/

Bosch, L., Costa, A. and Sebastián-Gallés, N. (2000). First and second language vowel perception in early bilinguals. *European Journal of Cognitive Psychology*, *12*, 189-221.

Bosch, L. and Sebastián-Gallés, N. (2003). Simultaneous bilingualism and the perception of language-specific vowel contrast in the first year of life. *Language and Speech*, *46*, 217-243.

Caramazza, A., Yeni-Komshian, G., Zurif, E. and Carbone, E. (1973). The acquisition of a new phonological contrast: The case of stop consonants in French-English bilinguals. *Journal of the Acoustical Society of America*, *54*, 421-428.

Caramazza, A. and Yeni-Komshian, G. (1974). Voice onset time in two French dialects. *Journal of Phonetics*, *2,* 239-245.

Cheer, M., Ceponiene, R., Lehtokoski, A., Luuk, A., Allik, J., Alho, K. and Näätänen, R. (1998). Development of language-specific phoneme representations in the infant brain. *Nature Neuroscience,* 1 (5), 351-353.

Christophe, A. and Morton, J. (1998). Is Dutch native English? Linguistic analysis by 2-month-olds. *Developmental Science, 1,* 215-219.

Clopper, C.G., Pisoni, D.B. and de Jong, K. (2005). Acoustic characteristics of the vowel system of six regional varieties of American English. *Journal of the Acoustic Society of America*, 118 (3), 1661-1676.

Cutler, A., Weber, A., Smits, R. and Cooper, N. (2004). Patterns of English phoneme confusion by native and non-native listeners. *Journal of the Acoustic Society of America*, *116* (6), 3668-3678.

Dehaene-Lambertz, G. (1997). Electrophysiological correlates of categorical phoneme perception in adults. *NeuroReport*, *8* (4), 919-924.

Dehaene-Lambertz, G. and Baillet, S. (1998). A phonological representation in the infant brain. *NeuroReport, 9,* 1885-1888.

Dehaene-Lambertz, G. and Gliga, T. (2004). Common neural basis for phoneme processing in infants and adults. *Journal of Cognitive Neuroscience*, *16* (8), 1375-1387.

Escudero, P. and Boersma, P. (2002). The subset problem in L2 perceptual development: multiple-category assimilation by Dutch learners of Spanish. *Proceedings of the 26th Annual Boston University Conference on Language Development*, Cascadilla Press, 208-219.

Escudero, P. and Boersma, P. (2004). Bridging the gap between L2 speech perception research and phonological theory. *Studies in Second Language Acquisition*, 26, 551-585.

Escudero, P., Hayes-Harb, R. and Mitterer, H. (2008). Novel second-language words and asymmetric lexical access. *Journal of Phonetics,* 36, 345-360.

Fennell, C.T., Byers-Heinlein, K. and Werker, J.F. (2007). Using speech sounds to guide word learning: the case of bilingual infants. *Child Development*, *78*, 1510-1525.

Flege, J.E. (1981). The phonological basis of foreign accent: a hypothesis. *TESOL Quarterly*, 15 (4), 443-455.

Flege, J.E. (1995). Second-language speech learning: theory, findings, and problems. In W. Strange (Ed.), *Speech Perception and Linguistic Experience: Issues in Cross-language Research* (pp. 229-273). Timonium, MD: York Press.

Flege, J.E. (2003). Assessing constraints on second-language segment production and perception. In A. Meyer and N. Schiller (Eds.) *Phonetics and Phonology in Language Comprehension and Production, Differences and Similarities* (pp.319-355). Berlin: Mouton de Gruyter.

Gerrits, E. and Schouten, M.E.H. (2004). Categorical perception depends on the discrimination task. *Perception & Psychophysics*, *66* (3), 363-376.

Grosjean, J. (1998). Studying bilinguals: methodological and conceptual issues. *Bilingualism: Language and Cognition*, *1*, 131-149.

Grosjean, J. (1999). The bilingual's language modes. In J.L. Nicol (Ed.). *One Mind, Two Languages: Bilingual Language Processing.* Oxford: Blackwell.

Gulian, M., Escudero, P. and Boersma, P. (2007). Supervision hampers distributional learning of vowel contrasts. *International Congress of Phonetic Sciences, XVI*, Saarbrucken, Germany.

Hagiwara, R. (1997). Dialect variation and formant frequency: the American English vowels revisited. *Journal of the Acoustic Society of America*, *102* (1), 655-658.

Hahn, A., Schroger, E. and Friederici, A.D. (2002). Segregating early physical and syntactic processes in auditory sentence comprehension. *NeuroReport*, *13* (3), 305-309.

Hayes-Harb, R. (2007). Lexical and statistical evidence in the acquisition of second-language phonemes. *Second Language Research*, *23*, 61-90.

Hillenbrand, J., Getty, L.A., Clark, M.J. and Wheeler, K. (2005). Acoustic characteristics of American English vowels. *Journal of the Acoustic Society of America*, *97* (5), 3099-3111.

Kuhl, P.K. (1991). Human adults and human infants show a "perceptual magnet effect" for the prototypes of speech categories, monkeys do not. *Attention, Perception & Psychophysics, 50* (2), 93-107.

Li, P., Sepanski, S. and Zhao, X. (2006). Language history questionnaires: A web-based interface for bilingual research. Behavior Research Methods, 38, 202-210. http://www.personal.psu.edu/pul8/questionnaire/L2_questionnaire.html

Liberman, A.M., Harris, K.S., Hoffman, H.S., and Griffith, B.C. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology*, *54*, 358–368.

MacKay, I.R.A., Flege, J.E., Piske, T. and Schirru, C. (2001). Category restructuring during second-language speech acquisition. *Journal of the Acoustic Society of America*, *110* (1), 516-528.

Maye, J. and Gerken, L. (2000). Learning phonemes without minimal pairs. In S. C. Howell, S. A. Fish & T. Keith-Lucas, *Proceedings of the 24th Boston University Conference on Language Development* (pp. 522–533). Somerville, MA: Cascadilla Press.

Maye, J. and Gerken, L. (2001). Learning phonemes: how far can the input take us? In A.H-J. Do, L. Domínguez, & A. Johansen (Eds.), *Proceedings of the 25th Annual Boston University Conference on Language Development* (pp. 480–490). Somerville, MA: Cascadilla Press.

Maye, J., Weiss, D.J. and Aslin, R.N. (2008). Statistical phonetic learning in infants: facilitation and feature generalization. *Developmental Science*, 11*,* 122-134.

Maye, J., Werker, J.F. and Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, *82*, B101-B111.

Näätänen, R., Paavilainen, P., Rinne, T. and Alho, K. (2007). The mismatch negativity in basic research of central auditory processing: A review. *Clinical Neurophysiology*, *118*, 2544-2590.

Pallier, C., Bosch, L. and Sebastián-Gallés, N. (1997). A limit on behavioral plasticity in speech perception. *Cognition*, *64*, B9-B17.

Peltola, M.S., Kujala, T., Tuomainen, J., Ek. M., Aaltonen, O. and Näätänen, R. (2003). Native and foreign vowel discrimination as indexed by the mismatch negativity (MMN) response. *Neuroscience Letters*, *352*, 25-28.

Psion, D.B. (1973). Auditory and phonetic memory codes in the discrimination of consonants and vowels. *Perception and Psychophysics, 13*, 253-260.

Pols, L.C.W., Tromp, H.R.C. and Plomp, R. (1973). Frequency analysis of Dutch vowels from 50 male speakers. *Journal of the Acoustic Society of America*, 53 (4), 1093-1101.

Sebastián-Gallés, N. and Soto-Faraco (1999). Online processing of native and non-native phonemic contrasts in early bilinguals. *Cognition, 72,* 111-123.

Strange, W. and Shafer, V.L. (2008). in J.G. Hansen-Edwards and M.L. Zampini (Eds.) *Phonology and Second Language Acquisition* (pp. 153-185). Amsterdam: John Benjamins Publishing Company.

Sundara, M. and Polka, L. (2008). Discrimination of coronal stops by bilingual adults: the timing and nature of language interaction. *Cognition*, *106*, 234-258.

Weber, A. and Cutler, A. (2004). Lexical competition in non-native spoken-word recognition. *Journal of Memory and Language*, *50,* 1-25.

Weber, A., and Smits, R. 2003 . ‘‘Consonant and vowel confusion patterns by American English listeners,’’ in *Proceedings of the 15th International Congress of Phonetic Sciences (*pp. 1437–1440). Palau de Congressos, Barcelona, Spain.

Werker, J.F. (1986). The effect of multilingualism on phonetic perceptual flexibility. *Applied Psycholinguistics, 7,* 141-156.

Werker, J.F. and Logan, J.S. (1985). Cross-language evidence for three factors in speech perception. *Perception & Psychophysics*, *37* (1), 35-44.

Werker, J.F. and Tees, R.C. (1984). Cross-language speech perception: evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development*, *7*, 49-63.

Winkler, I., Kujala, T., Alku, P. and Näätänen, R. (2003). Language context and phonetic change detection. *Cognitive Brain Research*, *17*, 833-844.

Appendices follow.

Appendices

Appendix A: Recruitment Flyer

The following is a flyer that was placed in various University of Amsterdam buildings for recruitment.

Earn €7*!!*

Native speaker of Dutch?

Is your English good too?

Then please participate in our listening experiment!

For more info or appointment
email attention: Andreea to
SprekenEnVerstaan@gmail.com

Dutch/English listening experiment
7€ / 30 min Email attn:Andreea to
SprekenEnVerstaan@gmail.com

Dutch/English listening experiment
7€ / 30 min Email attn:Andreea to
SprekenEnVerstaan@gmail.com

Dutch/English listening experiment
7€ / 30 min Email attn:Andreea to
SprekenEnVerstaan@gmail.com

Dutch/English listening experiment
7€ / 30 min Email attn:Andreea to
SprekenEnVerstaan@gmail.com

Dutch/English listening experiment
7€ / 30 min Email attn:Andreea to
SprekenEnVerstaan@gmail.com

Dutch/English listening experiment
7€ / 30 min Email attn:Andreea to
SprekenEnVerstaan@gmail.com

Dutch/English listening experiment
7€ / 30 min Email attn:Andreea to
SprekenEnVerstaan@gmail.com

Dutch/English listening experiment
7€ / 30 min Email attn:Andreea to
SprekenEnVerstaan@gmail.com

Dutch/English listening experiment
7€ / 30 min Email attn:Andreea to
SprekenEnVerstaan@gmail.com

Dutch/English listening experiment
7€ / 30 min Email attn:Andreea to
SprekenEnVerstaan@gmail.com

Appendix B: Info brochure

Informational brochure for 'Phonetic discrimination in context'

Dear participant in this study,

You are about to participate in a study investigating language perception. Before you start participating it is important that you understand the procedure being used in this study to the full extend. For this reason, please read the following carefully.

Purpose of this study

You will be participating in an experiment that will investigate the way in which people perceive speech sounds of English and of Dutch.

Instruction

During this experiment you will hear speech sounds, in one block from Dutch and in another block from English. During each trial, you will hear three sounds (non-words). Your task is to indicate which one of the three non-words is most like the second one: the first sound, or the third sound. If the second sound is like the first one, press the 'z' key on the keyboard; if the second sound is like the third one, press the '/' key on the keyboard. These instructions will appear on the screen throughout the experiment for your reference.

Each of the two blocks contains 100 trials. You will be allowed to take a break at an interval of 25 trials, and between the two blocks.

Procedure

Audio will be played to you on headphones. You must listen to the audio and give the appropriate response as indicated above.

Voluntary participation

If you decide to revoke your participation at this point, this will have no consequences for you whatsoever. If at any point during the experiment you decide to withdraw, this will not have any consequences for you either. You are also allowed to withdraw your results in this experiment up to 24 hours after participating. This means you can stop participating in this study at any time. You are free to do so without giving a reason why. If you decide to withdraw from this study during your participation, or up to 24 hours after participating, your results will be removed from our files and destroyed.

Insurance

Based on previous and similar studies, it has been established that this study does not cause any harm or discomfort for the participants. Because participants will not be exposed to any extra health- or safety risks, no special insurance has been taken out.

Confidentiality of research data

The results found in this study will be saved and used in scientific publications. However, your personal information will not be disclosed, ensuring your anonymity.

Participant groups that will be excluded from this study

You will not be allowed to participate in this study if you have hearing problems or language impairments.

Compensation

The participant will receive 7€ for participation.

Further information

If you would like to receive any further information about this study, either now or in the future, I can contact the supervisors of this study, Prof. Dr. Paul Boersma (Spuistraat 210, Room 303, 1012 VT Amsterdam; email: paul.boersma@uva.nl, tel.nr +31-20-5252385) and dr. Maartje Raijmakers (Roetersstraat 15, Room 723, 1018 WB Amsterdam; Email: M.E.J.Raijmakers@uva.nl; tel. nr. +31-20-5257014). If I have any complaints about this study I can contact a member of the Ethical Committee of the Psychology department of the University of Amsterdam, Dr. Wery van den Wildenberg (Roetersstraat 15, 1018 WB Amsterdam, room A7.10, email, W.P.M.vandenWildenberg@uva.nl, tel.nr. +31-20-5256686).

Appendix C: Consent form

A: INFORMED CONSENT PARTICIPATION

I declare that I have been clearly informed about the nature and method of research entitled "Phonetic discrimination in context" as specified in the attached brochure. My questions have been satisfactorily answered.

I am voluntarily participating in this research. I reserve the right to withdraw this consent without giving a reason, and realize that I can, at anytime during the experiment, stop participating. If my research will be used in scientific publications, or otherwise disclosed, this will be done completely anonymously. My personal data will not be viewed by others without my permission.

If I have any information about the investigation, now or in the future, I can contact the researcher Andreea Geambasu (email: a.geambasu@student.uva.nl) or supervisor Prof. Dr. Paul Boersma (tel: +31-20-5252385 or email: paul.boersma@uva.nl, address: Spuistraat 210, 1012 VT Amsterdam Room: 303), or dr. Maartje Raijmakers (Roetersstraat 15, Room 723, 1018 WB Amsterdam; Email: M.E.J.Raijmakers@uva.nl; tel. nr. +31-20-5257014).

For any complaints about this research I can contact the head of the Ethics Committee of the Department of Psychology at the University of Amsterdam, Prof. Dr. V. Lamme (telephone: 020 5256675; e-mail: V.Lamme@uva.nl, Roetersstraat 15, 1018 WB Amsterdam, room 626).

Signed in duplicate:

....
Subject Name Signature Date

B: CONFIRMATION informed consent

I have provided information on the investigation "Phonetic discrimination in context". I declare that I am capable and have answered all questions about the research.

....
Name of researcher Role in research

....
Signature of researcher Date

Appendix D: Language questionnaire

Exit Questionnaire
Please answer all of the following questions thoughtfully and thoroughly.

Background

1. What is your profession?

2. If student, what subject, level, and year?

3. What is your age?

4. What is your sex? Circle: Female Male

5. Where were you born?

6. Where do you currently reside?

7. Would you like to be included in our database to be contacted for future studies in our department?
   Circle: Yes No

If so, please provide the following information:

First and last name:

Date of Birth:

Telephone number:

Address:

Email:

Language Proficiency

1. Are you a native speaker of Dutch? Circle: Yes No

2. What language(s) was/were spoken in your home by parents/caretakers during your childhood? If not Dutch, please explain.

3. What languages are spoken in your home at present? If not Dutch, please explain.

4. At what age did you begin to learn English?

5. In what context did you learn English? Check one:
   - o At home, please explain
   - o At school, please say during which school years
   - o Other, please explain

6. How many hours per week, on average, do you speak English?

7. In what context(s) do you speak English each week? Check one:

- At work, please explain

- At home, please explain

- Other, please explain

8. Have you ever lived in an English-speaking country? Circle: Yes No
    a. If yes, where?

    b. And for how long?

9. Have you ever lived in a non-English-speaking foreign country? Circle: Yes No
    a. If yes, where?

    b. And for how long?

    c. During your stay in this foreign country, which languages did you speak to other foreigners, and to local population?

10. How many hours per week, on average, are you exposed to English in the following formats:
    - Audio-visual, please explain

    - Reading, please explain (ie, academic, news, websites, literature, etc.)

11. Have you ever had any kind of English pronunciation training (either formal or self-taught?). Please explain.

12. Please fill out the chart below and rate your speaking and comprehension knowledge of English *and* any other languages you may speak on a scale from 1 to 10 (1= worst , 10 = best).

Language	Production	Comprehension

13. If there is anything else you think would interest us about your language background, please write it in the space below:

Thank you!

Appendix E: Results of language questionnaire

All subjects listed Dutch as their native language and learned English as an L2 in school, except Subject 21 and 26, who were excluded because of the answers highlighted in bold. Abbreviations for languages are as follows: AR=Arabic, FR=French, FS=Frisian, GE=German, GR=Greek, HB=Hebrew, IT=Italian, LT=Latin, NL=Dutch, NW=Norwegian, RU=Russian, SP=Spanish, SW=Swedish, TR=Turkish. Other abbreviations: HS=High School, H=Home, S=School, Uni=University.

Subject	Age	Sex	Youth Home Lang	Present Home Lang	Begin English Age	Where?	Hrs/Week Eng	Lived in EN country	Months in EN country	Lived in Non-EN country	Months in non-EN country	Lang most used in non-EN country	Hours Audio-Visual Input	Hours Reading	Pronunciation Training?	EN Production	EN Comprehension	L3 + Self Rating (Production; Comprehension)	L4 + Self Rating (Production; Comprehension)	L5 + Self Rating (Production; Comprehension)	L6 + Self Rating (Production; Comprehension)
1	22	M	NL	NL EN	11	S	24	-	-	Spain	3	SP	10	6	-	8	9	SP 7; 8	-	-	-
2	21	F	NL	NL	8	S	4	-	-	-	-	-	8	12	-	8	7	SW 6.5; 6	GE 7; 6	FR 4;4	-
3	21	F	NL	NL	8	S	-	-	-	-	-	-	10	20	-	7	9	GE 4; 5.5	FR 2;4	-	-
4	25	F	NL	NL	8	S	2	-	-	Israel	12	HB EN	4	15	HS	7	8	HB 5;7	-	-	-
5	24	M	NL	NL	8	S	21	-	-	France	24	EN NL FR SP GE	3	-	Uni	8.5	9	FR 7; 7.5	GE 6;7.5	SP 5;5	-
6	19	M	NL	NL	8	S	0.5	-	-	-	-	-	9	6	-	8	8		-	-	-
7	24	F	NL	NL FR	11	S	5	-	-	France	4	FR	5	20	-	7	9	FR 8; 9	GE 5;6	-	-
8	21	F	NL	NL	11	S	1	-	-	-	-	-	3	3	-	7	8	SP 3;3	FR 2;3	GE 3;4	-
9	25	M	NL	NL EN	6	S	3	-	-	-	-	-	22	-	-	8	10	FR 4;6	GE 6;7	-	-

Subject	Age	Sex	Youth Home Lang	Present Home Lang	Begin English Age	Where?	Hrs/Week Eng	Lived in EN country	Months in EN country	Lived in Non-EN country	Months in non-EN country	Lang most used in non-EN country	Hours Audio-Visual Input	Hours Reading	Pronunciation Training?	EN Production	EN Comprehension	L3 + Self Rating (Production; Comprehension)	L4 + Self Rating (Production; Comprehension)	L5 + Self Rating (Production; Comprehension)	L6 + Self Rating (Production; Comprehension)
10	21	M	NL	NL	8	S	5	-	-	-	-	-	10	5	HS	8	7	GE 7;6	FR 4;3	-	-
11	30	M	NL	NL	10	S	5	-	-	-	-	-	3	8	HS	8	8	FR 5;5	GE 3;5	-	-
12	23	F	NL	NL	9	S	5	-	-	Spain Jordan Turkey	3 3 5	EN, Local langs	2	2	HS	9	9	GE 5;8	FR 7;8	SP 3;5	TR 3;4
13	25	F	NL	NL	10	S	40	UK	9	-	-	-	10	30	-	9	9	FR 4;5	GE 2;3	-	-
14	21	M	NL	NL	8	S	10	-	-	-	-	-	60	40	Self	9	10	FR 4;7	GE 4;8	-	-
15	26	M	NL	NL	11	S	2	-	-	-	-	-	20	5	-	7	9	GE 2;8	FR 2;7	SP 3;7	-
16	25	F	NL	NL	11	S	15	-	-	-	-	-	40	15	HS, Self	9	8	GE 4;4	FR 3;3	-	-
17	25	F	NL	NL EN	11	S	2	-	-	Mexico	6	SP	10	6	-	8	9	SP 6.5;8	-	-	-
18	21	M	NL	NL	7	S	1	-	-	Russia	2	RU	30	20	HS, Self	8	9	RU 6;7	FR 4;6	GE 4;6	LT & GR (written)
19	23	M	NL	NL	8	S	1	-	-	-	-	-	8	4	-	8	9	GE 5;7	NW 4;6	FR 4;6	
20	46	F	NL	NL	10	S	6	-	-	Israel	5	EN, Bit HB	7	14	-	8	9	GE 5;6	FR 5;5	IT 3;5	AR 2;2
21	23	F	**NL, EN**	**NL EN**	**0**	H, S	5	-	-	-	-	-	21	3	-	6	10	SP 2;5	**-**	**-**	**-**
22	27	M	NL	NL	12	S	1	US	12	-	-	-	10	5	HS	8	8	GE 6;8	FR 4;5	SP 2;3	
23	64	F	NL	NL	10	S	5	UK	6	-	-	-	10	10	HS	8	9	GE 7;9	FS 6;8	FR 6;7	SP 5;6

Subject	Age	Sex	Youth Home Lang	Present Home Lang	Begin English Age	Where?	Hrs/Week Eng	Lived in EN country	Months in EN country	Lived in Non-EN country	Months in non-EN country	Lang most used in non-EN country	Hours Audio-Visual Input	Hours Reading	Pronunciation Training?	EN Production	EN Comprehension	L3 + Self Rating (Production; Comprehension)	L4 + Self Rating (Production; Comprehension)	L5 + Self Rating (Production; Comprehension)	L6 + Self Rating (Production; Comprehension)
24	29	M	NL	NL	10	S	2	-	-	Italy	12	IT EN	20	40	-	7	8.5	IT 6;7.5	FR 5;7	GE 5;7	LT 2;5
25	22	M	NL	NL	10	S		-	-	-	-	-	7	10	-	8	10	GE 7;9	FR 5;5	SP 2;2	-
26	22	M	NL	NL	**3**	**US 3-6 y.o.**	-	US	36	-	-	-	4	0	-	8	9	FR 3;4	GE 4;4	**-**	**-**
27	32	M	NL	NL	12	S	7	UK Malta	1 8	Lebanon	few days	didn't peak yet	7	2	Uni	8	8	-	-	-	-

Appendix F: Results with subjects 26 and 21 included

ENGLISH – DUTCH (Group E-D)

	Score E	Score D	∆Score E-D	∆Score 2-1	RT E (s)	RT D (s)	∆RT E-D (s)	∆RT 2-1 (s)	Sex	Age
Subject 1	0.57	0.6	-0.03	0.03	4.85	4.08	0.77	-0.77	M	22
Subject 3	0.74	0.91	-0.17	0.17	3.96	3.57	0.39	-0.39	F	21
Subject 5	0.72	0.83	-0.11	0.11	5.14	4.06	1.08	-1.08	M	24
Subject 8	**0.55**	**0.45**	0.1	-0.1	4.51	3.8	0.71	-0.71	F	21
Subject 10	0.56	0.63	-0.07	0.07	4.99	4.53	0.46	-0.46	M	29
Subject 12	0.77	0.83	-0.06	0.06	4.07	4.04	0.03	-0.03	F	23
Subject 14	**0.98**	**0.93**	0.05	-0.05	3.83	3.59	0.24	-0.24	M	21
Subject 16	**0.84**	**0.78**	0.06	-0.06	3.96	3.96	0	0	F	25
Subject 18	0.51	0.58	-0.07	0.07	5.29	4.37	0.92	-0.92	M	21
Subject 20	**0.61**	**0.6**	0.01	-0.01	4.98	4.61	0.37	-0.37	F	46
Subject 22	0.51	0.7	-0.19	0.19	4.28	3.93	0.35	-0.35	M	29
Subject 24	0.56	0.62	-0.06	0.06	4.59	4.29	0.3	-0.3	M	29
Subject 26	**0.63**	**0.5**	0.13	-0.13	4.45	3.97	0.48	-0.48	M	22
Average	**65.7**	**0.689**	**-3.1**	**3.1**	**4.53**	**4.061**	**0.47**	**-0.47**		**25.6**

DUTCH – ENGLISH (Group D-E)

	Score D	Score E	∆Score E-D	∆ Score 2-1	RT D (s)	RT E (s)	∆RT E-D (s)	∆RT 2-1 (s)	Sex	Age
Subject 2	0.57	0.84	0.27	0.27	4.78	4.3	-0.48	-0.48	F	21
Subject 4	0.75	0.82	0.07	0.07	4.54	4.27	-0.27	-0.27	F	26
Subject 6	0.55	0.61	0.06	0.06	6.11	5.99	-0.12	-0.12	M	19
Subject 7	0.86	0.94	0.08	0.08	4.09	3.95	-0.14	-0.14	F	24
Subject 9	**0.73**	**0.63**	-0.1	-0.1	4.05	3.73	-0.32	-0.32	M	24
Subject 11	0.93	0.94	0.01	0.01	4.44	4.25	-0.19	-0.19	M	30
Subject 13	**0.62**	**0.58**	-0.04	-0.04	4.69	3.76	-0.93	-0.93	F	25
Subject 15	0.78	0.78	0	0	4.35	4.15	-0.2	-0.2	M	26
Subject 17	0.78	0.96	0.18	0.18	3.88	3.72	-0.16	-0.16	F	25
Subject 19	**0.57**	**0.55**	-0.02	-0.02	4.04	3.87	-0.17	-0.17	M	23
Subject 21	0.49	0.50	0.01	0.01	3.89	3.87	-0.02	-0.02	F	23
Subject 23	**0.51**	**0.48**	-0.3	-0.3	3.57	3.6	0.03	0.03	F	65
Subject 25	0.56	0.75	0.19	0.19	4.24	4.04	-0.2	-0.2	M	22
Subject 27	0.49	0.53	0.04	0.04	4.66	3.45	-1.21	-1.21	M	32
Average	**65.6**	**70.7**	**5.1**	**5.1**	**4.38**	**4.067**	**-0.312**	**-0.312**		**27.5**

Appendix G: t-test results

Mean of column "Score-D":
Mean = 0.0067999999999999998
Student's t from zero = 0.3142495447156708
Number of degrees of freedom = 24
Significance from zero = 0.37802348481544923 (one-tailed)
Confidence interval (95%):
Lower limit = -0.03786039952983491 (lowest value that cannot be rejected with α = 0.025)
Upper limit = 0.05146039952983491 (highest value that cannot be rejected with α = 0.025)

Mean of column "Score2-1":
Mean = 0.05000000000000001
Student's t from zero = 2.613541867446584
Number of degrees of freedom = 24
Significance from zero = 0.007615252534350795 (one-tailed)
Confidence interval (95%):
Lower limit = 0.010515295596843749 (lowest value that cannot be rejected with α = 0.025)
Upper limit = 0.08948470440315627 (highest value that cannot be rejected with α = 0.025)

Difference in column "Score2-1" between groups ED and DE of column "Order":
Difference = -0.009615384615384609
Student's t = -0.24613758527294302
Number of degrees of freedom = 23
Significance from zero = 0.4038795375356874 (one-tailed)
Confidence interval (95%):
Lower limit = -0.09042766096750877 (lowest difference that cannot be rejected with α = 0.025)
Upper limit = 0.07119689173673956 (highest difference that cannot be rejected with α = 0.025)

Mean of column "ScoreRTE-D":
Mean = 0.05039999999999992
Student's t from zero = 0.4748910483815995
Number of degrees of freedom = 24
Significance from zero = 0.31957885908762096 (one-tailed)
Confidence interval (95%):
Lower limit = -0.1686407417881387 (lowest value that cannot be rejected with α = 0.025)
Upper limit = 0.26944074178813854 (highest value that cannot be rejected with α = 0.025)

Mean of column "ScoreRT2-1":
Mean = -0.3992
Student's t from zero = -5.804860594405052
Number of degrees of freedom = 24
Significance from zero = 2.7579574144184813e-06 (one-tailed)
Confidence interval (95%):
Lower limit = -0.5411342105469377 (lowest value that cannot be rejected with α = 0.025)
Upper limit = -0.25726578945306233 (highest value that cannot be rejected with α = 0.025)

Difference in column "ScoreRT2-1" between groups ED and DE of column "Order":
Difference = -0.13294871794871777
Student's t = -0.9644389175095982
Number of degrees of freedom = 23
Significance from zero = 0.1724287767885544 (one-tailed)
Confidence interval (95%):
Lower limit = -0.41811491374082543 (lowest difference that cannot be rejected with α = 0.025)
Upper limit = 0.1522174778433899 (highest difference that cannot be rejected with α = 0.025)