The /a:/-/ɑ/ contrast for Spanish-speaking learners of Dutch

The effect of speaker gender upon perception

Irene ter Avest
0103004
MA-thesis for the Research Master in Linguistics
First reader: dr. D.J.M. Weenink
Second reader: prof. dr. P. Boersma
Course: 2008/2009
Credits: 30 ECTS

University of Amsterdam

Table of contents

Chapter 1: Introduction

1.1 Abstract

To native speakers of language, it does not matter whether a speaker they listen to is male or female, young or old: they virtually always understand what is being said. This not only holds for words and sentences, but also for the individual sounds of a language, like for example the Dutch vowels /a:/ and /ɑ/ as in *maan* 'moon' and *man* 'man'. For nonnative speakers of a language, however, identifying individual speech sounds becomes much more difficult, especially when the second language makes a distinction between two or more sounds that the first language does not make. To our knowledge, no studies have looked at whether the gender of the speaker does also play a role in the identification process in nonnative speakers.

In the present study, we tested 100 Spanish-speaking subjects, who were learning Dutch and 56 native controls on the /a:/ - /ɑ/ contrast, which is nonexistent in Spanish. Subjects performed two XAB-tasks: they heard tokens of /a:/ and /ɑ/ pronounced by male and female speakers and had to match each token with one of two possible answers. The two possible answers were tokens of /a:/ and /ɑ/ produced by a male computer voice and matched on duration (140 ms). In between the two XAB-tasks, there was a training phase during which subjects listened to either enhanced tokens of /a:/ and /ɑ/ produced by a male computer voice or a piece of classical music. Benders & Escudero (in preparation) have shown that training with enhanced tokens can improve learners' categorization. The results showed that, regardless of the condition, both native and nonnative speakers showed a gender effect on both XAB-tasks: they were more likely to answer /a:/ when the speaker was female and more likely to answer /ɑ/ when the speaker was male. However, this gender effect became stronger for the Spanish-speaking subjects on the second XAB-task, indicating that they were probably less able to adapt themselves to the stimuli used in the tasks than the Dutch subjects. Finally, the results showed that the training with the enhanced tokens improved subjects' categorization of /a:/-stimuli. These results may have consequences for theories on second language acquisition and vowel perception.

1.2 Problem definition

In spoken language, native speakers (or for the purpose of the present study, listeners) of a language seem to be able to understand most of what is said irrespective of the characteristics of the speaker who is saying it. In most cases, it does not matter whether the speaker is male,

female, young or old. This not only holds for complete sentences produced by the speaker, but also for the individual words that are present in these sentences, like the word *tree* in the sentence *The cat is climbing up the tree*. Almost every word is made up out of vowels (e.g. *ee* in *tree*) and consonants (e.g. *t* and *r* in the same word). It is very likely that native speakers are also able to identify these vowels and consonants, even when they are taken from the original context in which they appear and are presented in isolation. However, vowels are much more likely to appear in isolation than are consonants: some consonants, like *b*, are even hard to produce without any surrounding vowel(s). Therefore, most studies have investigated the identification of vowels rather than consonants (e.g. Marean et al., 1992; Kuhl, 1983; Kubaska & Aslin, 1985)[1] and the results show that, in most cases, native speakers are indeed able to identify a vowel, again irrespective of the characteristics of the speaker who is producing it. They seem thus to have some kind of abstract 'prototype' of this vowel, which enables them to identify it every time it is produced, even though it will be produced in a slightly different way by every speaker and even every time it is produced by one and the same speaker. The process of comparing a vowel to an abstract 'prototype' of this vowel and thereby adapting oneself to the speech of a particular speaker is called 'vowel normalization', and it plays an important role in language perception by reducing the influence of between-speaker variation.

If vowel normalization is so important for native speakers, what role does it play in language perception in nonnative speakers of a language? Worldwide there are many people who start learning a second or even a third language at a later age, in most cases after puberty. These learners already master their mother tongue when they start learning the new language and in addition to this, they are beyond what is called 'the Critical Period for language learning' (e.g. Hyltenstam & Abrahamsson, 2003). In theories on second language learning, the term 'Critical Period' is used to define the period during which children are able to reach nativelike proficiency in a second language, and there is an intensive debate going on about the length of this period. Some researchers state that the Critical Period may end as early as age six, at least for the phonology of the second language (e.g. Long, 1990), but most say that it ends around puberty (see Hyltenstam & Abrahamsson, 2003 for a nice overview of current theories). After the Critical Period has ended, it becomes almost impossible for second language learners to reach nativelike proficiency in all aspects of their second language: only

[1] We will summarise these studies in the next section of this chapter.

a few will attain a level of proficiency that makes them indistinguishable from native speakers.

When people start to learn a second language (henceforth L2) after puberty, they not only have established the vowel-prototypes for their mother tongue (henceforth L1), but they also may have passed the Critical Period for establishing them in a nativelike way in the L2. One of the difficulties these "late" L2 learners encounter can be the existence of 'subsets' or 'supersets' of vowels in the L2 (see for example Escudero & Boersma, 2001). A subset means that two or more different vowel categories are mapped onto one vowel category in the L2, whereas a superset means that a single vowel category from the L1 is mapped onto two or more vowel categories in the L2. Both supersets and subsets can cause problems, because L2 learners have to make new distinctions between vowels in the target language or have to "forget about" a distinction they make in the L1. Especially the superset problem appears to be difficult to overcome, because the learner has to learn two or more new categories and the fact that these are not present in the L1 makes their perception and production in the L2 very difficult. On the contrary, when the learner has to form a new category out of two or more categories from the L1, perception and production in the L2 are much less problematic, because in most cases the learner will be able to use a perception mode that is dedicated to the L2 without affecting the L1 (Escudero & Boersma, 2001), even though it will take some time to tune this new perception mode to the sounds of the L2: this is because in the beginning of the learning process the new perception mode is based upon the L1 perception mode and therefore, the learner may make distinctions between vowel categories that are irrelevant in the L2.

As we can see, the subset and the superset are in fact two sides of the same coin: only the "direction" of the problem is determined by the L1 of the learner. This happens for example between Dutch and Spanish and between Dutch and English. The Dutch vowels /a:/ and /ɑ/, as in *maan* 'moon' and *man* 'man' respectively, are mapped onto a single category in Spanish, namely /a/. For Dutch-speaking people learning Spanish this would yield a subset problem, whereas for Spanish-speaking people learning Dutch it would yield a superset problem. On the other hand, in English, the Dutch vowel /ɛ/ is mapped onto two different vowel categories, namely /ae/ and /ɛ/, as in 'bad' and 'bed' respectively. For Dutch-speaking people learning English this is a superset problem, whereas for English-speaking learning Dutch it is a subset problem.

In this study, we will investigate vowel normalization and the creation of vowel categories in L2 learners. We will look at a large group of Spanish-speaking people from various backgrounds[2] who were participating in the project "A longitudinal study of how vowel sounds can either facilitate or impede the acquisition of a third language by immigrant communities" (Escudero, in preparation) at the University of Amsterdam. All were learning Dutch as an L2 and they had attained different proficiency levels in this language[3]. It will be investigated how these L2 learners of Dutch establish the categories for the vowels /a:/ and /ɑ/ (one of the superset problems mentioned above) and how they normalise for speaker gender (male or female speaker). Given the fact that all L2 learners were beyond the Critical Period at the time they started learning Dutch, it is not unlikely that they experience persistent difficulties with the /a:/-/ɑ/ distinction in Dutch. However, the goal of the present study is to investigate whether these learners also have difficulties with vowel normalization for these two vowels. At the moment, very little is known about vowel normalization in L2 learners: to our knowledge, only a few studies have been conducted into the topic. It is still unclear whether these learners start normalising for between-speaker variation as soon as they start learning the L2 or whether they start by using the vowel-prototypes from the L1. However, no matter what the answer will be, it is difficult to imagine nativelike vowel normalization without the existence of (nativelike) vowel categories: for instance, it is difficult to imagine a Spanish-speaking learner of Dutch as an L2 performing a correct normalization for speaker gender for the Dutch vowels /a:/ and /ɑ/ when he or she has not yet established the categories for these vowels (remember that Spanish has only one category here). In other words, if the learner is unable to discriminate /a:/ from /ɑ/, it is unlikely that he or she will be able to normalise these vowels for speaker gender in a nativelike way.

Investigating vowel normalization and, more specifically, normalization for speaker gender in L2 learners can shed a new light on what happens in a learner's head when he or she starts to understand what is being said in the L2, what the role of the L1 and the age of onset (Critical Period) are in the process, and on how the establishment of vowel categories and vowel normalization are related to each other in these learners. However, in order to get a

---

[2] With 'backgrounds', non-linguistic information is meant here (e.g. education, length of residence in the Netherlands, etc.).

[3] Some of the Spanish-speaking subjects had already learned English as an L2 before learning Dutch, but they differed considerably in their proficiency and therefore it will be impossible to pinpoint the influence of the L2 upon the L3 in the present study. Nevertheless, this influence will be investigated in the rest of the project (Escudero, in preparation).

clear idea of what is happening, it will be indispensable to compare the results of the L2 learners to those of a control group of native speakers of Dutch. Investigating normalization for speaker gender in L2 learners and comparing the results to those of native speakers is precisely what we are going to do in the present study.

1.3 Studies that have investigated vowel normalization in native and nonnative speakers

In this section, we will discuss some of the most important studies on vowel normalization and on the establishment of vowel categories in both native and nonnative speakers. We will put emphasis on studies on normalization for speaker gender. The outline of the section will be as follows: in the first paragraph, we will discuss studies that investigated vowel normalization in infants and young children who were acquiring their L1. In the second paragraph, studies that investigated vowel normalization in adult native speakers will be discussed. Finally, in the third paragraph, we will discuss some studies on vowel normalization in nonnative speakers. For the native speakers, we have chosen to discuss the studies conducted with children and adults in separate paragraphs, because of different experimental techniques. For instance, it is impossible to present an infant with two vowels and to ask it whether these two vowels are the same or different, whereas this technique is often used with adults. Most of the techniques that are used with infants and young children are varieties of the 'head turn paradigm'. In this head turn paradigm, the infant is usually trained on two natural or synthesised speech sounds. These sounds are presented in blocks or 'trials' of multiple tokens and the infant is trained to turn its head in the direction of the loudspeaker when there is a change in speech sound within a block. For example, the infant is trained on the vowels /a/ and /e/ and is taught to turn its head when the vowel changes from /a/ to /e/ or from /e/ to /a/. Both sounds are usually presented multiple times to enable the experimenters to observe the infant's response. Again, the infant cannot simply be told what to do and therefore the head turn paradigm makes use of 'visual reinforcers': these are mechanical toys or short movies that attract the infant's attention and which are only shown when a correct response has been made. An experimental session often looks as follows: the infant is tested in a soundproof booth, is held by a parent and faces a TV monitor on which a short movie is shown or a research assistant, who is manipulating various toys. The movie or the toys serve to focus the infant's attention towards the centre of the booth. The experimenter observes the infant's behaviour from outside the booth and cannot be seen by the people inside. The speech sounds are presented to the infant over a loudspeaker located either left or

right of the infant. When there is a change in speech sound and the infant orientates its head towards the loudspeaker over which the sounds are presented, the visual reinforcer is shown and a correct response is scored by the experimenter and the research assistant. The visual reinforcer is always located next to the loudspeaker over which the sounds are presented and is not shown when the infant fails to make a head turn. An error is scored in this case. If the infant makes a head turn when there is no change in speech sound, no visual reinforcer is shown either and an error is scored. If the infant does not make a head turn in this case, a correct response is scored. When the trial is over, the infant's attention is redirected towards the centre of the booth by the movie or the toys manipulated by the research assistant and the next trial can start. The parent who is holding the infant, the research assistant and often also the experimenter wear headphones to prevent them from hearing the speech sounds and (subconsciously) influencing the infant's behaviour and to avoid biases in the scoring of the responses. The assistant and the experimenter both score the infant's responses and a response is only scored correct or incorrect when both have scored the response in the same way. The trials are stopped if the infant starts crying, fuzzing or is not paying attention to the speech sounds anymore. The training phase is often preceded by a conditioning phase, during which only trials that contain a change in speech sound are presented to the infant. In this way, the infant is familiarised with the visual reinforcers. Both the conditioning phase and the training phase last until the infant has reached a given percentage of correct responses (often 90% correct on ten consecutive trials). After the training phase, there is an experimental phase during which the infant is tested on different sounds or on the same sounds presented in different orders using the same procedure as the one used for the training phase.

A technique that is related to the head turn paradigm and which is used in one of the studies we will discuss below, is the 'habituation looking procedure'. This procedure also consists of a conditioning phase, a training phased and a test phase, but now the infant is not taught to make head turns. Just like in the head turn procedure, the infant is held by a parent and is facing a TV monitor on which a picture is shown. The speech sounds the infant is trained and tested on are now also presented via a loudspeaker that is next to the TV monitor. Instead of the numbers of correct and incorrect head turns, the amount of time the infant is looking at the screen (looking time) is determined for each trial. Infants normally look longer at the screen when the speech sounds that are presented are "new" to them. After a few trials, they become habituated to the sounds and looking times become shorter. When new sounds are presented, looking times will usually become longer again, but will also decrease after a

few trials. When a trial has ended, the infant's attention is redirected to the screen by means of a flashing light or a toy. The trials are stopped when the looking times become shorter than a previously established criterion, which is often the average of the two longest looking times on the first three trials, or when the infant starts crying or fuzzing.

The experimental techniques that can be used with adults are much more varied: a common procedure is to present subjects with sequences of two sounds and to ask them whether the sounds are different or identical. This type of task is called an AX-task. Varieties of this task are the XAB-task and the ABX-task. During these tasks, subjects hear sequences of three sounds. When the task is an XAB-task, they have to indicate whether the first sound is identical to the second sound or the third sound and when the task is an ABX-task, they have to indicate whether the last sound is identical to the first or the second, so the only difference between the two tasks is the order in which the sounds are presented. The task are usually presented on a computer and subjects have to press keys in order to make a response. Relatively new techniques that start to be used in research with adults are techniques that measure the blood flow to certain parts of the brain (hemodynamic response), like for example near-infrared spectroscopy, or techniques that measure the amount of electric current present in various parts of the brain, like for example magnetoencephalography (MEG). With near-infrared spectroscopy, an increase in blood flow to certain parts of the brain present over several trials indicates that these parts are very likely to be involved in the process that is being investigated (e.g. vowel discrimination), whereas with MEG, different neural responses in the brain have been associated with different processes and situations, like difficulties with the semantic integration of a word into a sentence or difficulties with establishing the grammatical structure of a sentence. In most cases, the target sentences or sounds have been manipulated by the experimenters. Overall, the techniques that can be used with adults are often more suitable to determine what characteristics of a vowel are the most important ones for categorization and normalization than the techniques that can be used with infants and children.

1.3.1 Vowel normalization in infants and children who are acquiring their L1

As was stated before, native speakers of a language do not seem to experience any difficulties with vowel normalization. This certainly holds for adults, but research has shown that infants who are acquiring their L1 are able to perform vowel normalization and normalization for speaker gender from a very early age onwards. When they are only a day old, these infants

already possess a memory for speech sounds: Swain et al. (1993) found that after habituation, newborn infants retained memory for speech sounds for at least 24 hours. About two months later, infants are able to perceive spectrally different speech sounds as perceptually equivalent: Marean et al. (1992) conducted a study with 2, 3 and 6-month-old infants who were acquiring American English as their L1 using a variety of the head turn paradigm. They tested in four trial types whether these infants were able to perceive synthetic tokens of the vowel /a/ or the vowel /i/ as belonging to the same vowel category (if only one of the vowels was presented to the infant during a trial) or to different vowel categories (if both vowels were presented to the infant and there was a change in vowel during a trial). It was investigated whether the infants were able to categorise these vowels despite differences in speaker gender (male/female) and pitch contour (rising/falling). Vowels were always presented in five pairs and the second member of the pair could either be identical or different. The infants were trained with male tokens of the two vowels just mentioned. Pitch contour was randomised across these tokens. In the test phase, within a block of five pairs there could be either no change or a change in vowel, speaker gender or vowel and speaker gender at the same time. Just as during the training, pitch contour was randomised across tokens. All trials began with a male /a/. The results showed that infants reached a proportion correct of at least 70% on all four trial types and there was no significant effect of age: the 2, 3 and 6-months old did equally well. Marean et al. explain that this does not mean that there is little developmental change in vowel categorization (and normalization) between two and six months of age: by making the task more complex (e.g. by introducing a third speaker) age differences could possibly be brought to light. It also remains unknown whether infants categorise vowels in the same way as adults do: for instance, adults may pay attention to different characteristics of a vowel than infants do.

Kuhl (1983) conducted a similar study: she tested 6-month-old infants on vowel categorization for the American English vowels /a/ and /ɔ/ using the computer-synthesised voices of men, women and children and again varying the pitch contour of the tokens. There were two experiments: in the first one, the infants were trained to discriminate these two vowels when they were pronounced by the synthesised male voice with a falling pitch-contour. Then, gradually, differences in pitch-contour and speaker (female) were introduced. In the last stage, the differences in pitch-contour (rising/falling) and speaker (male/female) were presented to the infant at the same time and after a while, the vowels produced by the synthesised child's voice were introduced, bringing the total number of stimuli for each of the

two vowels to six (3 different speakers * 2 pitch-contours). The results showed that the infants did not have problems with the differences in pitch-contour and speaker: they were still able to perceive the two vowels as belonging to different categories. In a second experiment, the infants were also trained on the male tokens with falling pitch-contour, but now all possible differences in pitch-contour and speaker were introduced at once after the training. In this experiment, they also performed significantly above chance level for most of the "new" combinations of pitch-contour and speaker, with the exception of the combinations child speaker/rising pitch-contour and male speaker/rising pitch-contour, indicating that pitch dimension can effect infants' vowel categorization, at least when it is introduced together with differences in speaker gender and age.

Grieser & Kuhl (1989) looked at a different aspect of infant vowel categorization: they trained 6-month-old infants to discriminate tokens of the American English vowels /ɛ/ and /i/. The tokens the infants heard during the training phase were all "good" exemplars of their respective vowel categories, which means that they matched prototypes defined by adult speakers. Subsequently, the infants were tested with 32 new tokens from each vowel category, which differed in the degree to which they resembled the adult-defined prototypes: some were "better" exemplars of their respective categories than others. The quality of the vowels was changed by manipulating the first two formants: the frequencies of one or both of these formants could be increased or decreased. It turned out that the infants were able to correctly categorise the test stimuli in more than 90% of the cases. In a second experiment, infants of the same age were trained with either a good or a bad exemplar of the vowel category /i/ and were tested with 16 exemplars of that same category. These 16 exemplars again differed in the degree to which they matched the adult-defined prototype. It turned out that infants who had been trained on the good exemplar of the category were significantly better at generalising the knowledge acquired during the training phase to the novel stimuli than infants who had been trained on the bad exemplar of the category. This seems to indicate that infants group vowels that belong to the same vowel category in their language around 'prototypes' that represent this category. These prototypes are exemplars that resemble other exemplars belonging to the same category to a higher degree: there is more overlap. Garner (1974) defined these prototypes as being more "redundant".

Kuhl (1991) tested the use of prototypes in the categorization of the vowel /i/ in infants, adults and monkeys (rhesus macaques). She found that when the prototype of the category was used as the referent vowel, which means that the other stimuli were compared to

this vowel, humans showed a significantly better generalisation to other members of the category than when a non-prototypical exemplar of the category was used as the referent vowel. This effect was not found in the monkeys. Kuhl calls the effect found in humans the 'perceptual magnet effect': prototypical members of a vowel category assimilate other, less prototypical members of the category and pull these towards the prototype. The effect also makes it more difficult to discriminate between two exemplars of a category with formant frequencies that are close to those of the prototype than between two exemplars of the same category with formant frequencies that are further away from those of the prototype. Kuhl also found that the responses of infants and adults were highly correlated, which indicates that they may use the same kinds of prototypes or at least very similar ones, although more research is needed.

In a more recent study, Lively & Pisoni (1997) tried to replicate Kuhl's findings with adults, but found no perceptual magnet effect. They discovered that the acoustic context in which a vowel appeared had an important influence upon its perception and that discriminability of specific exemplars of /i/ was not affected by the goodness of the category members. Furthermore, they let subjects label the stimuli used by Kuhl (1991) and found that many of the non-prototypical /i/'s were not labelled as /i/'s but as members of different vowel categories. Their findings indicated that the perceptual magnet effect may not be very robust and that the tokens of /i/ used by Kuhl (1991) may have spanned more than one vowel category.

Kubaska & Aslin (1985) tested vowel categorization and normalization in older children, namely 3-year-olds. In their first experiment, the children were trained with isolated tokens of /a/ and /i/ pronounced by an adult male speaker. During the subsequent test phase, they presented the children with isolated tokens of /a/ and /i/ pronounced by an adult male speaker, an adult female speaker, a male child speaker and a female child speaker. The results showed that the children were able to generalise their responses for the adult male vowel tokens to the tokens pronounced by the other speakers and thus showed perceptual constancy of natural tokens of /a/ and /i/ across speakers who differ in sex and age. The second experiment of the study was identical to the first one, but now subjects were presented with natural tokens of /æ/ and /ʌ/, which showed more overlap in the formant frequencies F1 and F2 than the /a/ and /i/ tokens. The results showed that again the 3-year-olds were able to generalise their responses to the new stimuli presented to them during the test phase, indicating that they were able to correctly normalise for speaker gender and age. However, it

remains unknown what mechanism underlies vowel normalization in children and adults and how this mechanism develops in early infancy and childhood.

These studies show that children who are acquiring their L1 are able to categorise vowels and to normalise for speaker age and gender from a very early age onwards. However, it is also important to look at what adult native speakers do. In the next paragraph, we will discuss some of the studies that were conducted with adult native speakers.

1.3.2 Vowel normalization in adult native speakers

With respect to vowel normalization in adults, it has been known for years that the first two formants are important acoustic cues to a vowel's phonetic identity (see for example Delattre et al., 1952), despite their variability across different speakers and contexts. Ladefoged (2005a & 2005b) argued that the third formant may also be important, at least for some languages, given the fact that it depends much on the position of the lips. In French, for example, the third formant is important for making the distinction between the words *lit* 'bed' (pronounced as [li]) and *lu* 'read' (pronounced as [ly]): when one pronounces these words, the tongue is virtually in the same position, but the position of the lips differs. Halberstam & Lawrence (2004) investigated the role of F0 and F3 information in the process of vowel normalization in blocked speaker and mixed-speaker conditions. In a blocked speaker condition, all stimuli within a block are produced by the same speaker, whereas in a mixed-speaker condition, the speaker may vary randomly from one stimulus to the other. They used whispered (no F0 information) and phonated tokens of natural vowels. For some stimuli, formants above F2 were filtered out and for other stimuli they were left intact. The results indicated that F0 played a role in vowel normalization, but the results for F3 were inconclusive: this formant seemed to be more important for the discrimination of whispered vowels than for the discrimination of phonated vowels. Error rates were higher in the mixed-speaker condition than in the blocked-speaker condition, a finding that was also reported in other studies (see the dissertation of Adank (2003) p. 49 for an extensive overview of these studies).

Roberts et al. (2004) made use of the magnetoencephalography (MEG) technique to investigate the involvement in vowel categorization of certain neuronal structures in the brains of human adults. The neuromagnetic component they were interested in was the so-called M100: previous studies (Roberts et al. refer to Roberts & Poeppel, 1996) had shown that this neuromagnetic component, which peaks around 100 ms post-stimulus onset, is

sensitive to stimulus attributes, among which intensity and frequency. They first presented subjects with a continuum of sinusoidal tones. These sinusoidal tones had F1 frequencies that matched F1 frequencies for the two vowel categories that would be tested later on in the experiment, /u/ and /a/. Two extreme tokens were added at 100 Hz and 1kHz. Each sound had a duration of 400 ms. They found that the latency of the M100-response to the sinusoidal tones showed a decrease along a 1/f distribution. The latency ranged from about 123 ms for the sound with the lowest F1 to about 100 ms for the sound with the highest F1. The 1/f curve served as a baseline to which the results for the vowels were compared. Roberts et al. made a continuum of eleven vowel-like stimuli. The F1 of the tokens on the continuum varied from 250 to 750 Hz in steps of 50 Hz. To increase the naturalness of these vowel-like stimuli, a second and a third formant were added with bandwidths of 300 Hz. Subjects had to indicate for each token whether they thought it was /u/ or /a/. The three tokens with the highest and lowest values of F1 were almost always assigned to the categories of /a/ and /u/ respectively. The remaining five tokens were ambiguous. Roberts et al. looked at M100 latencies along the continuum and found that for the tokens that were clear members of one of the two categories, M100 latencies clustered and did not show a 1/f distribution related to F1-frequency: the latencies for /u/-tokens clustered around 121 ms and the latencies for /a/-tokens clustered around 95 ms. However, for the ambiguous tokens, M100 latencies did show a 1/f distribution identical to the one found for the sinusoidal tones and which would be predicted if latencies depended on the mere acoustic properties of the stimuli. The results seem to indicate that, at least in native speakers of a language, M100 latencies map onto vowel categories showing clusters for tokens that are unequivocally identified as belonging to a given category. It is quite well possible that these clusters of latencies or “latency plateaus” reflect the activity of neuronal cohorts in the brain that belong to the vowel category in question. For the ambiguous stimuli, however, no single neuronal cohort is activated and the M100 latencies show more or less the same distribution as that found for sinusoidal tones which have identical F1-frequencies. In this way, the clusters of M100 latencies found for prototypical members of a vowel category may reflect a perceptual magnet effect as the one described by Kuhl (1991), with prototypical members of a category attracting M100 latencies. Another interesting finding was that reaction times for /u/-tokens were significantly longer than those for /a/-tokens. The M100 latencies were also longer for /u/-tokens than for /a/-tokens.

There are various factors that seem to influence vowel normalization in native speakers, like speaker gender, length of the vocal tract and the context in which a vowel

appears. Van Bergem et al. (1987) investigated vowel normalization in native speakers of Dutch. For their experiment, they used carrier sentences produced by a man and a child. The man imitated the child's pitch to the best of his abilities. Each carrier sentence had the structure *Matroos pVt eet kaas* 'Sailor pVt eats cheese'. Both speakers were asked to stress the word *kaas*. In this way, the target word pVt would be unstressed. The following vowels could be inserted in the pVt context: /a/, /ɔ/, /ɛ/, /ɪ/, /ʏ/, /u/, /y/ and /i/[4]. Each subject was presented with the same five vowel categorization tasks or "conditions" in the following order: 1) target words produced by the man and the child presented in isolation, 2) target words produced by the man and the child embedded in either their own carrier sentences or in carrier sentences produced by the other speaker, 3) filtered versions of the sentences from condition 2 (reduced timbre information: especially the higher formants were filtered out by employing cut-off frequencies of 5500 Hz, 1650 Hz, 1150 Hz and 650 Hz), 4) the beginning of the carrier sentences *Matroos pVt* in the four versions from condition 2 and finally 5) the end of the carrier sentences *pVt eet kaas*, again in the four versions from condition 2. Results showed that the recognition of the target vowel in pVt could be strongly influenced by the carrier sentence: subjects committed many more errors when the target word was embedded in a carrier sentence produced by the other speaker. However, acoustic context that preceded the target vowel turned out to be more influential than acoustic context that followed the target vowel. For the filtered sentences Van Bergem et al. found that the error rates were higher when the target word was embedded in a carrier sentence produced by the other speaker than when the whole sentence was produced by the same speaker for all cut-off frequencies, except for the lowest cut-off frequency (650 Hz) when the target word was produced by the child. Van Bergem et al. argued that listeners choose an appropriate template ("male", "female" or "child") on the basis of the speaker's pitch and timbre conveyed by the acoustic context provided by the carrier sentence.

Mitterer (2006) found that the context in which a vowel appears influences its identification: target vowels that appeared in carrier sentences in which the other vowels fell within a low F2-range, were more likely to be perceived as front vowels: these vowels have a high F2. Although previous experiments had shown that lexical status of the items in the carrier sentence had an effect on vowel perception, this effect was not found here. In contrast,

[4] For the vowel /ʏ/, Van Bergem et al. use the symbol /œ/, which was the way in which this Dutch vowel was transcribed at the time the experiments were conducted.

it was found that when the carrier sentence contained only mid-to-high front vowels, vowel categories shifted only for these vowels.

Johnson et al (1999) found that native speakers, when asked to identify the vowels /ʊ/ and /ʌ/ as in 'hood' and 'hud' along an F1-continuum, set the boundary between the two vowels differently for male speakers than for female speakers when they were shown videoclips of either a male or a female speaker: for the former, this boundary was located at a lower point of the continuum than for the latter. The "stereotypicality" of the voice also played a role here: when a voice had been qualified as stereotypically male or female by a previous group of subjects in the study, the differences between the phoneme boundaries between the two genders became larger. Interestingly, the difference was still visible when subjects were presented with a set of gender-ambiguous phonemes within the same continuum, but were asked to imagine either a man or a woman pronouncing them: they did not see any videoclips of male or female speakers. Again the boundary between the two phonemes was lower for imagined male speakers than for imagined female speakers.

1.3.3 Vowel normalization in nonnative speakers

As was pointed out in section 1.2, people who start learning an L2 after puberty already fully master their L1. They have all vowel categories relevant to this L1 "in place" and are able to normalise for differences between speakers for each of these vowels. However, these fully established vowel categories can make the acquisition of a new language more difficult. For instance, it is possible that the new language contains phonemes and rules for combining these phonemes within syllables that the L1 does not have. This not only makes the pronunciation of the L2 more difficult, it may also hinder acquisition: Polivanov (1931) gives some nice (and sometimes anecdotical) examples of people who, in one way or another, repeated what they thought they had just heard in a language which was not their native language. These repetitions were often very different from what the native speaker thought he or she had pronounced. Even after various repetitions of the word by the native speaker, the nonnative speaker was still unable to hear the "correct" pronunciation. According to Polivanov, this is caused by the fact that people, when listening to a language which is not their native language, use the "linguistic consciousness" that is appropriate for their L1. They map the sounds they hear onto the phonemes present in their mother tongue and often even use their first language's rules for syllabic structure: for example, a Japanese, when hearing the Russian word *tam* 'there', is very likely to perceive it as *tamu*. This is because in

Japanese, syllables may not end in a consonant. This phenomenon often occurs when a nonnative speaker is not very familiar with the L2, but it is nonetheless very difficult to overcome. There is a lively debate going on about whether adults who start learning an L2 will ever be able to overcome these difficulties (see Cucchiarini, 1993 for an overview) and what might be causing them: a loss of neural structures (Cucchiarini refers to Eimas (1975)) or a shift in attentional focus. Results have been inconclusive so far, but it may be the case that the influence of the L1 is stronger for consonants than it is for vowels (Werker & Polka, 1993). Nevertheless, correctly perceiving the vowels of the L2 can be extremely problematic, especially for the superset problem (Escudero & Boersma, 2001).

Polka & Werker (1994) found that the sensitivity to phonemic contrast that are not present in the L1 decreases already in the first year of life. They tested the discrimination between the German vowel pairs /ʏ/-/y/ and /ʊ/-/u/ in a dVt context by 6-8 and 10-12-month-old English-learning infants using the head turn paradigm. The control stimuli were the English vowels /i/ and /a/, which also appeared in a dVt context. The results showed that the younger infants were better at discriminating the vowels from two vowel pairs than the older infants and that the behaviour of the former was compatible with a perceptual magnet effect: they performed better when the vowel to which the other vowels were compared was the vowel that most resembled the English phoneme in both contrasts. The 10 to 12-month-olds did not show this behaviour. However, because the scores of the 6-8 -month-olds were below levels that had been reported for nonnative consonant contrasts (Polka & Werker refer to Werker & Tees (1984) and to Werker & Lalonde (1988)), a second experiment was carried out with two groups of younger infants: a group of 4-month-olds and a group of 6-month-olds. The same stimuli were used, but now the procedure was a habituation looking procedure. The results showed that the 4-month-olds were able to discriminate the vowels from both German vowel pairs, but that the 6-month-olds were not. Neither group showed evidence of a perceptual magnet effect. This may indicate that sensitivity to nonnative phonemic contrasts starts to decrease earlier for vowels than for consonants. The role of the perceptual magnet effect is still unclear.

Minagawa-Kawai et al. (2004) used near-infrared spectroscopy to measure the hemodynamic responses of highly proficient Korean late L2 learners of Japanese. Subjects were presented with the /a/-/e/ contrast, which also exists in Korean (its phonetic space is relatively similar to that of the Japanese contrast) and with long and short versions of these vowels: this durational contrast is not present in Korean. Subjects listened to 16 repetitions of

the stimuli and had to identify whether the second vowel was phonologically long or not. The results showed that the Koreans did not differ behaviourally from a group of native Japanese speakers that had been tested earlier. They also showed neural activity that was comparable to that of the native Japanese speakers for the /a/-/e/ contrast. Nevertheless, the Koreans did show a different neural activity than the native speakers for the durational contrasts /a/-/a:/ and /e/-/e:/: contrarily to the /a/-/e/ contrast, these contrasts did not evoke responses specific to phonemic discrimination. In addition to this, reaction times of the Koreans were significantly longer than those of the native speakers, which may indicate that the L2 learners employed a different strategy than the native speakers for the contrast in duration. Minagawa-Kawai et al. concluded that the neural networks used for the L1 and the L2 may be either shared or language-specific: when a given phonemic contrast is present in both languages occupying the same or similar positions in the phonetic space, the same neural networks may be used in the L1 and the L2 for that contrast. However, if a given contrast is present in the L2 but not in the L1, neural networks that are not specific to phonemic discrimination may be employed.

Cebrian (2006) looked at the categorization of nonnative vowel contrast in Catalan late L2 learners of English, who differed in the amount of experience with the L2, which was operationalised as length of residence in an English-speaking country. He conducted two experiments. The first experiment was a perceptual assimilation task, during which subjects were presented with English and Catalan vowels and had to say in an alternative forced choice task to which Catalan (L1) phonetic category the vowel they heard was most similar. They also had to rate the goodness of the vowel as an exemplar of the Catalan category. Results showed that there was no effect of experience upon the ability to discriminate L1 and L2 vowels, but that L2 learning seemed to have affected L1 vowel identification: the experienced L2 learners identified the L1 vowels significantly worse than the non-English-speaking group. In a second experiment, Cebrian investigated the use of spectral and durational cues in the English tense-lax contrast by native speakers and two groups of Catalan late L2 learners of English: one group had more experience with the L2 than the other. They were presented with synthesised tokens of /i/, /ɪ/ and /ɛ/ that had been manipulated for duration and vowel quality. They had to indicate which vowel they thought they had heard by clicking on one of three words on a computer screen. The words on the computer screen were ‘beat’, ‘bit’ and ‘bet’. The results showed that for the vowels /i/ and /ɪ/ the Catalan L2 learners of English relied more on duration than native speakers. There was again no effect found for

the amount of experience with the L2. The results of this experiment show that even learners whose L1 has no durational contrast can use these duration for categorising vowels in the L2. The results of the two experiments taken together indicate that there may be no strong effect of experience in L2 vowel categorization.

Escudero (2001) investigated to what extent native speakers of English who spoke the Scottish or Southern English dialect of the language and Spanish late L2 learners of both dialects relied on spectral and durational cues when discriminating the tense/lax vowels /i/ and /ɪ/. In Scottish English, the discrimination of the two vowels is largely based upon spectral differences, whereas in the Southern English dialect, duration is the most important factor for discriminating the two vowels. This was confirmed by the results for the native speakers: native speakers of Scottish English indeed relied more upon spectral cues and native speakers of Southern English relied more upon duration. Results also showed that Spanish learners of the two dialects had adapted themselves to the cues available in their input: the learners of Southern English relied exclusively or primarily on duration, whereas the learners of Scottish English relied exclusively on spectral information. Nevertheless, some L2 learners showed non-nativelike reliance on the durational cue: they relied much more on this cue than any of the native speakers.

Werker & Logan (1985) discovered that the nature of the task nonnative speakers have to perform can influence their results and the strategies they employ. Subjects, who were native speakers of English, had to tell whether two sounds they heard were the same or different (an AX-task). Werker & Logan made a distinction between three types of processing strategies: "phonemic perception" (the sounds of the L2 are perceived according to the phonological categories of the L1), "phonetic perception" (the sounds of the L2 are perceived according to the phonological distinction present in a language other than the L1) and "psychoacoustic" or "auditory" processing (subjects make use of acoustic differences that do not correspond to phonetic boundaries in any language). The stimuli used for the experiments were tokens from the Hindi phonemic categories /t/ and /ʈ/. The distinction between these categories is subphonemic in English and both are usually perceived as the alveolar phone /t/ by native speakers of this language. In Hindi, however, the distinction is phonemic. The English subjects in this study had no knowledge of Hindi. The tokens of /t/ and /ʈ/ were followed by the neutral vowel /a/: in this way, the tokens would always appear within the same syllable. There were four tokens from each phonemic category and these four tokens differed acoustically. In this way, there could be three kinds of 'AX-combinations':

1) a token from one of the categories paired with exactly the same token ("physically identical pairing"), 2) a token from one of the categories paired with a different token from that same category ("name-identical pairing") and 3) a token from one of the categories paired with a token from the other category ("different pairing"). The time-interval between the first and the second sound from each pair, the Inter Stimulus Interval (ISI), was manipulated and could be 250 ms, 500 ms or 1500 ms. The processing strategy a subject had used would be determined by looking at how many times this subject indicated that the two sounds were identical: 1) with phonemic processing, the numbers of "same"-responses would be almost identical for all three types of pairings, but the number would be a bit higher for physically identical pairings than for name-identical pairings and also higher for name-identical pairings than for different pairings. 2) With phonetic processing, the number of "same"-responses would be higher for physically identical pairings than for name-identical pairings and these two types of pairings would have a much higher number of "same"-responses than the different pairings. 3) Finally, with auditory processing, the number of "same"-responses would be much higher for the physically identical pairings than for the other two types of pairings, but the number of "same"-responses would still be somewhat higher for the name-identical pairings than for the different pairings. There were two experiments, which consisted of five blocks of stimuli. In the first experiment, every subject was tested in all ISI conditions, which makes ISI a within-subjects variable. The order of the ISI conditions was randomised across subjects. Results showed that subjects probably made use of a phonetic processing strategy in the two shortest ISI conditions, but that they made use of both phonetic and auditory processing strategies in the longest ISI condition. It was also shown that practice (by presenting the subject with AX-pairs) might enhance performance in all ISI conditions, especially when the tokens were from different phonemic categories. However, it turned out that the order of the ISI conditions had a significant effect on performance: performance in an ISI condition tended to be better when this condition had been preceded by the 250 ms condition, which suggests that the subjects found it difficult to switch to a different processing strategy when they were presented with a different ISI condition. Therefore, a second experiment was carried out with ISI condition as between-subjects variable. The results now showed that each ISI condition differentially affected performance: in the 250 ms condition, subjects made use of auditory processing, whereas in the 500 ms condition they made use of both phonemic and auditory processing in the first two blocks of stimuli but shifted to phonetic processing in the last three blocks. For the 1500 ms condition, the results showed that subjects made use of phonemic processing, at

least in the first three blocks. This indicates that ISI can have an important influence upon subjects' performance. Nevertheless, Werker & Logan also state that when the memory demands of a task increase, access to the auditory processing may become limited. This happens for the XAB-task when compared to the 41-AX task (Pisoni & Lazarus, 1974).

None of the studies on L2 learning we have discussed so far has looked at normalization for speaker gender. To our knowledge, only one study has indirectly investigated the topic. This study was conducted by Amin (2003) to investigate the influence of speaker gender upon L2 listening comprehension, which is of course a bit different from vowel categorization. The subjects were Iranian L2 learners of English. There were two speakers: a man and a woman. Both were near-native speakers of English as judged by a team of four experienced instructors who had studied in the US for at least five years. The results showed that listening comprehension was significantly better for the male speaker, irrespective of the gender of the listener. However, because there were only two speakers (one of each gender), the results are not generalizable: maybe they were caused by factors other than the gender of the speakers. In the present study, we will take a closer look at normalization for speaker gender in late L2 learners.

1.4 The present study

Numerous studies have investigated vowel categorization in late L2 learners, but to our knowledge, no study has looked at normalization for speaker gender in these learners. Nevertheless, investigating this normalization for speaker gender may provide us with more insight into the strategies that L2 learners employ when listening to the L2. Furthermore, various studies have shown that normalization for speaker gender is present from very early on in infants that are acquiring their L1 (e.g. Marean et al. 1992; Kuhl, 1983) and it would be interesting to see whether it is also acquired early by late L2 learners.

To investigate the topic, we were given access to the results of two entire participant groups (100 subjects in total) on the third session of the project "A longitudinal study of how vowel sounds can either facilitate or impede the acquisition of a third language by immigrant communities" (Escudero, in preparation) at the University of Amsterdam. The project is a longitudinal one and consists of a total of four sessions. One of the goals is to investigate the ability of Spanish-speaking late L2 learners of Dutch to discriminate various phonemic contrast that are not present in their L1. A second goal of the project is to investigate the

influence of proficiency in English (the L2 of many of the participants) upon the ability to discriminate Dutch vowels. On the basis of the results, participants in the project will be provided with feedback that will help them to learn Dutch better and faster. At the beginning of the project, every new participant was randomly assigned to one of seven groups: these groups would differ in the type of training they would receive in each of the sessions and also in the tasks they had to perform at the beginning of the third session. At the beginning of every session, subjects have to indicate their proficiency levels in Dutch and English. They also regularly perform a Dutch listening comprehension task (Dialang). In each of the four sessions of the project, subjects are tested on different contrasts: in the first session, subjects were tested on all five phonemic contrasts that are investigated in the project, namely /a:/-/ɑ/, /i/-/ɪ/, /i/-/y/, /y/-/ʏ/ and /ɪ/-/ʏ/. In order to test subjects' ability to discriminate the vowels, they were given five XAB-tasks, one for each vowel contrasts. In addition to the XAB-tasks, every subject received training on three of the five contrasts, namely /a:/-/ɑ/, /i/-/ɪ/ or /y/-/ʏ/. In the next chapter, we will describe this training in more detail.

In the second session, subjects were only tested on the contrasts /a:/-/ɑ/, /i/-/ɪ/. They were tested twice on every contrast. Just like in the first session, the contrasts were presented in XAB-tasks and subjects received training on each contrast. In addition to the XAB-tasks, subjects had to perform a word recognition task involving all five contrasts.

In the third session, which is the session we will be looking at in the present study, subjects were only tested on the /a:/-/ɑ/ contrast. At the beginning of the session, five of the seven participant groups were given two XAB-tasks. In between these two tasks, they received training on the /a:/-/ɑ/ contrast. The other two groups performed only one XAB-task, in which they were presented with 240 synthetic tokens of /a:/ and /ɑ/, which had been manipulated on F1 and F2. This group received no training. After the second XAB-task or the XAB-tasks with the 240 synthetic stimuli, the session would look the same for all subjects. First, they performed the Dutch Dialang. After that, all subjects performed the two XAB-tasks, that had already been performed by five of the groups, and received training in between. In the last part of the session, subjects were tested in English to assess their listening and discrimination abilities in English: they had to answer the first ten questions of the English Dialang and after that, they had to perform a forced choice task with twelve synthesised English vowels. A complete overview of the third session and the time it took subjects to perform each of the tasks can be found in Appendix 1. For the present study, we only looked at the results on the second set of XAB-tasks (the tasks subjects had to perform

after the Dutch listening comprehension task). In the next chapter, these tasks will be described in more detail.

In the fourth session of the project, which has not finished yet, subjects are being tested on all five contrasts. First they have to perform a word recognition task identical to the one from the second session. After that, they have to perform five XAB-tasks, one for each contrast. Subjects do not receive any training.

The /a:/-/ɑ/ contrast is the only contrast that was investigated in the third session. This contrast is extremely difficult to learn for Spanish-speaking learners of Dutch, but it tends to be one of the easiest for native speakers (Escudero et al., in preparation). In addition to this, background variables like proficiency in Dutch, age of arrival and length of residence in The Netherlands did not explain the large variation that was found between learners (Escudero et al., 2009). Therefore, this contrast seemed ideal to investigate normalization for speaker gender in late L2 learners: possible difficulties with vowel normalization are unlikely to be obscured by ceiling effects and in addition to this, the stimuli for the XAB-tasks the subjects had to perform during the session were produced by both male and female speakers, whereas the two possible answers with which subjects had to match a stimulus were synthesised tokens of /a:/ and /ɑ/ produced by a male computer voice. If subjects have difficulties with normalization for speaker gender, they may have difficulties matching a stimulus produced by a female speaker with a possible answer produced by a male computer voice: /a:/ has higher frequencies of F1 and F2 than /ɑ/, but women also tend to have higher formant frequencies than men. This may lead to confusion in the Spanish-speaking subjects, because vowels produced by female speakers may be more likely to be identified as /a:/. The effects of different types of training upon subjects' discrimination abilities were also investigated in the project. Just like the possible answers, the training stimuli were produced by the male computer voice. There were three types of training: 1) training with "bimodal" tokens of /a:/ and /ɑ/ that were matched on duration (140 ms), 2) training with "enhanced" tokens of these vowels that were matched on duration (again 140 ms) and 3) "music", which served as a control group. The tokens in the first two training conditions were matched on duration to force the subjects to pay attention to spectral cues. Normally, /a:/ has a much longer duration than /ɑ/. When this durational cue is removed, it becomes extremely difficult for Spanish-speaking L2 and L3 learners of Dutch to discriminate /a:/ from /ɑ/, whereas native speakers are still able to do so (Escudero et al., 2009). The training was always given in between the

two XAB-tasks. For the present study, we only looked at subjects who were in the "enhanced" or the "music" training condition. Benders & Escudero (in preparation) have shown that training with enhanced tokens can improve learners' categorization. In the next chapter, we will give a more detailed overview of the stimuli and the procedure.

To be able to compare the results of the Spanish-speaking subjects to those of native speakers, we also tested two groups of native speakers of Dutch (56 subjects in total), who received the same training as the Spanish-speaking subjects: one group was assigned to the "enhanced" condition an the other one to the "music" condition.

Before conducting the experiments, we had the following hypotheses regarding vowel normalization in native and nonnative speakers:

- If the Spanish-speaking subjects have difficulties with vowel normalization, we expect them to perform worse on the female stimuli than on the male stimuli: /a:/ has higher F1 and F2 frequencies than /ɑ/ and since women have overall higher formant frequencies than men, a female /ɑ/ might be confused with a male /a:/.
- Given the fact that in the enhanced condition subjects receive training with male tokens of /a:/ and /ɑ/, we expect the performance of the Spanish-speaking subjects on the second XAB-task to improve more for the male stimuli than for the female stimuli.
- We expect no important differences between both tasks in the music condition for both Spanish-speaking and native Dutch subjects.
- We expect the native Dutch subjects to perform almost at ceiling on all tasks, regardless of the condition.

Chapter 2: Method and procedure

2.1 The subjects

In total, 100 Spanish-speaking persons who were learning Dutch as an L2 or L3 and 56 native speakers of Dutch participated in the study. The Spanish-speaking subjects all participated in the project "A longitudinal study of how vowel sounds can either facilitate or impede the acquisition of a third language by immigrant communities" (Escudero, in preparation) at the University of Amsterdam, which was briefly described in the previous chapter. They had all been living in The Netherlands for at least one year when the project started, but differed in their Dutch proficiency as well as in their backgrounds (e.g. educational). Subjects had been randomly assigned to one of the seven groups at the beginning of Escudero's project; two of these groups were selected for the present study and given the random assignment, it is reasonable to assume that both groups do not differ significantly with respect to background variables that might play a role in vowel normalization, like education, proficiency in English (a variable taken into account in the project) and length of residence in The Netherlands.

The group of native speakers consisted of students from the University of Amsterdam and friends or relatives of the experimenter. They reported not to have studied phonetics and were assigned to one of the two training conditions on a semi-randomised basis: the first 28 were assigned to the "enhanced" condition and the other 28 to the "music" condition.

2.2 The experiments

The experiments in which the subjects participated were especially designed for the project of Escudero (in preparation) to test subjects' ability to discriminate the vowels from five vowel contrasts that are not present in Spanish. The experiments that were used for the present study were part of the third session of the project and consisted of two XAB-tasks, which were separated by a training phase. In both XAB-tasks, subjects' ability to discriminate /a:/ from /ɑ/ was tested. The native Dutch subjects also filled in a questionnaire about their language background.

2.2.1 The XAB-tasks

The XAB-tasks the subjects had to perform before and after the training phase were virtually identical: the only difference was the order in which the stimuli were presented. This order was randomised for every subject and XAB-task. The stimuli used for the tasks were natural

tokens of /a:/ and /ɑ/ and subjects had to match every stimulus with one of two possible answers. The stimuli were a subset of the vowels recorded by Adank et al. (2004). Adank et al. obtained these vowels as follows[5]: the recorded vowels were produced by 20 male and 20 female speakers of so-called 'Standard Dutch'. All speakers were teachers of Dutch at schools for secondary education in Belgium and The Netherlands at the time the recordings were made. The 15 vowels of Standard Dutch, among which /a:/ and /ɑ/ were elicited in a wide variety of tasks during a so-called 'sociolinguistic interview'. The target vowels were put into carrier sentences, which were presented to the participants on a computer screen, with a three-second interval between sentences. The carrier sentences were slightly different for /a:/ and /ɑ/ due to phonological properties of Dutch.

For the vowel /a:/ , the carrier sentence was:

In sVs en in sVze zit de V (In sVs and in sVze is the V)

For the vowel /ɑ/, the carrier sentence was:

In sVs en in sVsse zit de V (In sVs and in sVsse is the V)

For each vowel, two tokens were recorded. The neutral context sentences were down-sampled to 16 kHz. Some of the participants were interviewed in an empty classroom and others at their own home. Due to the differences in recording conditions, background noises were present in some of the recordings. These recordings were excluded from further analyses. For the analyses only the SVS-contexts were selected.

For the present study, only the speakers that came from The Netherlands were selected: 10 men and 10 women. For every vowel, only one of the tokens produced by a speaker was selected. The averages of the characteristics of the tokens of /a:/ and /ɑ/ that were used in the present study can be found in Table 2.1. Some of the selected vowels presented an offglide of formant frequencies or sounded as if they had been extracted from a different context than the one used by Adank et al. (2004): for example, two of the /a:/-stimuli sounded as if they had been preceded by a *t* or a *d*, whereas in the original carrier sentences, they were

[5] For the exact characteristics of the vowels used in the present study as well as the acoustic measurements carried out on the raw data, we refer the reader to the article by Adank et al. (2004).

preceded by an *s*. This is probably caused by the method Adank et al. used for extracting the vowels from the context of the carrier sentence. Nevertheless, all stimuli were clearly recognisable as tokens of /a:/ or /ɑ/ and we decided not to carry out any further analyses. Every stimulus was presented twice during one XAB-task, making a total of 80 trials per task.

Table 2.1: Averages of the characteristics of the natural tokens of /a:/ and /ɑ/ used in the present study. All formant frequencies are in Hz.

	a:					ɑ				
	Duration (ms)	**F0**	**F1**	**F2**	**F3**	**Duration (ms)**	**F0**	**F1**	**F2**	**F3**
Fem.	216	183	923	1552	2845	93	223	719	1239	2957
Male	204	132	652	1424	2448	94	154	584	1156	2455

The two possible answers used in the XAB-tasks were synthesised tokens of /a:/ and /ɑ/ that were generated in Praat with a script. They all sounded as if they had been produced by a male speaker of Dutch. However, contrarily to natural tokens of /a:/ and /ɑ/, the synthesised tokens were matched on duration, with both having a duration of exactly 140 ms. This would make it impossible for subjects to rely on the durational cue when deciding whether a stimulus they heard was /a:/ or /ɑ/. As was mentioned in the previous chapter, when the durational cue is removed, native speakers of Dutch are still able to discriminate /a:/ from /ɑ/, whereas this becomes much more difficult for Spanish-speaking L2 and L3 learners of Dutch (Escudero et al, 2009). The other characteristics of the two synthetic vowels were the same as those mentioned in the article by Pols et al. (1973) and can be found in table 2.2 below. The order in which the two possible answers were presented to a subject was randomised across trials.

The XAB-tasks were presented on a Dell AMD Athlon64 X2 computer using Praat version 4.6.40. The stimuli were presented to the subjects over AKG headphones via an Edirol USB Audio Capture box (model UA-25) with 24 bit/96 kHz filter. During the task, subjects saw two yellow boxes on the screen. The left box had the number '2' in it and the right box had the number '3' in it. These boxes corresponded to the two possible answers and the subject had to click with the mouse on the box he or she thought represented the correct answer to a trial. The interval between the stimulus and the first possible answer (i.e. the interval between 'X' and 'A') was always 1.2 s to prevent subjects from listening acoustically

(cf. Werker & Logan, 1985). As soon as the subject clicked on one of the boxes, the next trial would start 1 s later. There was no time limit for answering. After every 20 trials, the subject could take a short break.

Table 2.2: Characteristics of the synthetic tokens of /a:/ and /ɑ/ used as the two possible answers in the present study. Formant frequency values were taken from Pols et al. (1973: 1094).

		Average formant frequency in Hz		
	Duration (ms)	F1	F2	F3
/a:/	140	795	1301	2565
/ɑ/	140	679	1051	2619

2.2.2 The training phase

The two XAB-tasks were separated by a training phase. During this phase, which lasted about three minutes, subjects would listen to enhanced tokens of /a:/ and /ɑ/ or a piece of classical music, depending on the condition they had been assigned to. The enhanced tokens of /a:/ and /ɑ/ were synthesised tokens, which were generated in Praat by a script and sounded as if they had been produced by the same male speaker as the one who was giving the possible answers during the XAB-tasks. They were obtained as follows: as starting points, the average frequencies of F1 and F2 for /a:/ and /ɑ/ as shown in Table 2.2 above were taken. The frequencies of these two formants were then artificially increased or reduced by the Praat script. In this way, a continuum of /a:/ and /ɑ/-like tokens was obtained. However, the frequencies of F1 and F2 of the newly obtained tokens could not differ more than one Standard Deviation from the average frequencies shown in Table 2.2[6]. The Standard Deviations were the Standard Deviations found for male speakers by Pols et al. (1973: 1094) and were 95 Hz for F1 and 113 Hz for F2 for /a:/ and 80 Hz for F1 and 89 Hz for F2 for /ɑ/. The frequencies of F1 and F2 were manipulated in a total of eight frequency-steps, yielding a total of eight tokens. The frequencies belonging to these tokens are found in Table 2.3. The most extreme tokens had the following frequencies of F1 and F2 respectively: 600 and 1000 Hz for the token with the lowest formant frequencies and 885 and 1430 Hz for the token with the highest formant frequencies. Tokens that were close to the endpoints of the continuum

[6] The only exception was the F2 of the most extreme /a:/-like token (token 8), which was 1430 Hz (see Table 2.3).

would be presented to the subjects more often than the less extreme ones: the tokens 2 and 7 would be presented sixteen times, the tokens 6 and 3 would be presented eight times and the other ones (the most extreme ones and the ones from the middle of the continuum) would be presented only four times. This would make the difference between the two vowels more salient. The F1 for the /ɑ/-like tokens was not made lower than 600 Hz, because then some listeners started hearing /o/ instead of a more /ɑ/-like token (Paola Escudero, personal communication). The enhanced tokens were chosen because a study by Benders & Escudero (in preparation) has shown that training with enhanced tokens of vowels can improve subjects' categorization of these vowels. The order in which the tokens were presented was randomised for every subject and the Inter Stimulus Interval was always 500 ms. The subjects who listened to the piece of classical music listened to a part of Bach's Violin Concert No. 2. Both the enhanced tokens and the piece of classical music were presented over the same headphones as the ones used for the XAB-tasks.

Table 2.3: Frequencies of F1 and F2 for the eight enhanced tokens of /a:/ and /ɑ/ and the number of times each token was presented during the training phase. The duration of a token was always 140 ms.

Token	F1	F2	Nr. of times presented during training phase
1	600	1000	4
2	637	1055	16
3	675	1112	8
4	714	1171	4
5	755	1233	4
6	797	1296	8
7	840	1362	16
8	885	1430	4

2.2.3 The language background questionnaire

The Dutch subjects also filled in a questionnaire about their language background. The Spanish-speaking subjects did not have to fill in such a questionnaire because they had already answered the questions when enrolling themselves in the project. Subjects were asked if they knew any languages other than their mother tongue and if so, how well they spoke and

understood these languages. They were also asked where they had learned the languages and how long they had been learning them. A copy of the questionnaire can be found in Appendix 2.

2.3 Procedure

All subjects were tested at the Department of Phonetic Sciences at the University of Amsterdam. They were tested in the office of the experimenter or in a soundproof booth next to this office. First, the subjects were told how to perform the XAB-tasks. If a subject had understood the instructions, he or she was given four practice trials, consisting of synthesised tokens of /i/ and /y/. The order of the practice trials was always randomised. Before starting with the real XAB-task, subjects were told that the first sounds could be very short sometimes, because they had been extracted from spoken language. They were also encouraged to respond as quickly as possible. When a subject had finished the first XAB-task, the experimenter told the subject that he or she was going to hear a number of sounds or a piece of music (depending on the training condition the subject had been assigned to). When the subject was going to hear the enhanced tokens of /a:/ and /ɑ/, he or she was told to listen very carefully. When the subject was going to hear the piece of classical music, he or she was told to relax and listen to the music. After the training phase, subjects directly started with the second XAB-task. They were told that this task would be the same as the first one and the instructions were repeated. Only subjects who had had difficulties performing the first XAB-task were given the four practice trials again. Subjects were paid when they had completed the whole session. Most of the Dutch subjects filled in the questionnaire at the beginning of the session.

Chapter 3: Results

First, we checked the data to see whether there were any anomalies. The data of 6 Dutch subjects were excluded, because they had had difficulties performing the XAB-tasks or because it turned out that they had some knowledge of phonetics after all. In addition to this, for the Dutch subjects, we decided to exclude subjects who had scored below 70% correct on the first XAB-task, because in previous studies that used the similar tasks and the same stimuli (Escudero et al., 2009; Escudero et al., in preparation), native Dutch listeners scored a very high percentage correct on the /a:/-/ɑ/ contrast and therefore a score below 70% may indicate that the subject had difficulties performing the task. On the basis of this criterion, another 14 subjects were excluded. None of the Spanish-speaking subjects was excluded. The answers on the questionnaire did not reveal anything that could be problematic and were therefore not taken into account when analysing the data.

The data of the remaining subjects were used for the analyses: 100 Spanish subjects (50 in the enhanced condition and 50 in the music condition) and 36 Dutch subjects (22 in the enhanced condition and 14 in the music condition).

3.1 Confusion matrices

To get an indication of how well subjects had been able to discriminate /a:/ from /ɑ/, the results of every group of subjects on every test were accumulated in confusion matrices. A confusion matrix shows how many tokens of each vowel were offered to the group of subjects as a whole on a given XAB-task, how many of these tokens were identified correctly and how many were erroneously identified as being the other vowel. The matrix also shows the total numbers of /a:/ and /ɑ/ answered by the subjects on the test. The confusion matrices we will discuss here only show the results for all stimuli taken together; more detailed confusion matrices, among which the ones that show the results split up for speaker gender, can be found in Appendix 3.

We will start by showing the confusion matrices for the Spanish-speaking subjects in both conditions. After that, we will compare these matrices to those of the Dutch subjects.

3.1.1 Confusion matrices for the Spanish-speaking subjects

In this section, we will show the results of the Spanish-speaking subjects. For the pretest, we decided to take the results of the subjects in the enhanced condition and those in the music condition together, because subjects made this test before they received any training and a close inspection of the data showed that both groups had indeed performed more or less equally. To get an indication of the differences in numbers of answers between both groups, (which are most likely caused by chance), the total numbers of answers were divided by two to obtain the average numbers of both groups. The amount of difference between the groups is indicated between brackets with the '±' sign, because for every number of answers, one group scored above average and the other one below average. The results of the Spanish-speaking subjects on the pretest can be found in Confusion matrix 3.1. It becomes clear that the Spanish-speaking subjects make many errors, irrespective of whether the vowel presented to them is a token of /a:/ or a token of /ɑ/. Interestingly, both groups of subjects seem to differ more on their responses to /a:/-stimuli than on their responses to /ɑ/-stimuli.

Confusion matrix 3.1: Results of the Spanish-speaking subjects on the pretest.

Recognised → Offered ↓	/ɑ/	/a:/	Totals
/ɑ/	1209.5(± 10.5)	790.5 (± 10.5)	2000
/a:/	766.5 (± 36.5)	1233.5 (± 36.5)	2000
Totals	1976 (± 47)	2024 (± 47)	4000

One of the goals of the present study was to investigate whether training with enhanced tokens of /a:/ and /ɑ/ would improve subjects' categorization of these vowels. Therefore, we will now look at the results of the Spanish-speaking subjects in the enhanced condition on the posttest. These results can be found in Confusion matrix 3.2. It is evident that, for subjects in the enhanced condition, the results on the posttest differ from those on the pretest: on the posttest, subjects make less errors (although the error percentage remains high), but remarkably, they have improved much more on the /a:/-stimuli than on the /ɑ/-stimuli. The training with the enhanced tokens seems to have influenced subjects' discrimination of the two vowels, but the influence has been bigger for /a:/-stimuli than for

Confusion matrix 3.2: Results on the posttest of the Spanish-speaking subjects who received enhanced training.

Recognised → Offered ↓	/ɑ/	/a:/	Totals
/ɑ/	1270	730	2000
/a:/	588	1412	2000
Totals	1858	2142	4000

/ɑ/-stimuli. However, practice effects can still not be ruled out: subjects were presented with the same stimuli in both tasks and maybe they "got used to" the stimuli and the procedure. Being more familiar with the stimuli and the procedure can also cause subjects to perform better. Therefore, we will now look at the results on the posttest of the Spanish-speaking subjects who were in the music condition and did not receive any training. These results can be found in Confusion matrix 3.3. When we compare the results on the posttest with those on the pretest, we see that subjects in the music condition also improve on the posttest. However, this improvement is much smaller than the improvement found for subjects in the enhanced condition and, contrarily to what we found for subjects in the enhanced condition, there is no important difference between /a:/-stimuli and /ɑ/-stimuli. This makes it likely that the improvement in the music condition is caused by practice effects, whereas the improvement in the enhanced condition is caused by both practice effects and training.

Confusion matrix 3.3: Results on the posttest of the Spanish-speaking subjects who listened to the piece of classical music.

Recognised → Offered ↓	/ɑ/	/a:/	Totals
/ɑ/	1253	747	2000
/a:/	704	1296	2000
Totals	1957	2043	4000

3.1.2 Confusion matrices for the Dutch subjects

In this section, we will show the results of the Dutch subjects. Unfortunately, the two groups of Dutch subjects did not show the same behaviour on the pretest: 11 subjects from the music condition had to be excluded due to extremely low scores on the pretest, compared to 3 subjects from the enhanced condition. Even though it would be possible to weigh the results for every group of subjects and present them in one confusion matrix, we decided not to do so, because this would make it difficult to compare the results on the pretest with those on the posttest. Therefore, the results of both groups of on the pretest will be shown in separate matrices. The results on the pretest of the Dutch subjects in the enhanced condition are found in Confusion matrix 3.4 and the results on the pretest of the Dutch subjects in the music condition are found in Confusion matrix 3.5 For both conditions, the results show us that the Dutch subjects make less errors than the Spanish-speaking subjects, which is in line with our expectations. For subjects in the enhanced condition, the results show us that the performance on the /a:/-stimuli is the same as the performance on the /ɑ/-stimuli. For subjects in the music condition, the results show us that the performance on the /ɑ/-stimuli was slightly better than the performance on the /a:/-stimuli. This is mainly caused by a large percentage of confusions (30%) on the male /a:/-stimuli (see the matrices in Appendix 3).

Confusion matrix 3.4: Results on the pretest of the Dutch subjects who received enhanced training.

Recognised → Offered ↓	/ɑ/	/a:/	Totals
/ɑ/	748	132	880
/a:/	127	753	880
Totals	875	885	1760

Confusion matrix 3.5: Results on the pretest of the Dutch subjects who listened to the piece of classical music.

Recognised → Offered ↓	/ɑ/	/a:/	Totals
/ɑ/	474	86	560
/a:/	112	448	560
Totals	586	534	1120

The results on the posttest of the Dutch subjects in the enhanced condition are found in Confusion matrix 3.6. Just as for the Spanish-speaking subjects, the results for this condition seem to indicate that the training with the enhanced tokens has improved subjects' ability to discriminate /a:/. Subjects' performance on the /ɑ/-stimuli has not changed after the training with the enhanced tokens.

Confusion matrix 3.6: Results on the posttest of the Dutch subjects who received enhanced training.

Recognised → Offered ↓	/ɑ/	/a:/	Totals
/ɑ/	746	134	880
/a:/	48	832	880
Totals	794	966	1760

Confusion matrix 3.7 shows that the performance of Dutch subjects in the music condition has changed after listening to the piece of classical music: they have become better on the /a:/-stimuli, whereas their performance on the /ɑ/-stimuli has remained almost the same. However, the improvement on the /a:/-stimuli seems to be be smaller than the improvement found in subjects from the enhanced condition. This seems to indicate that, also for the Dutch subjects, the enhanced training has improved subject's discrimination of /a:/. The practice effects found in the music condition are quite large for /a:/-stimuli. This may be caused by the fact that Dutch subjects in this condition made many errors with male /a:/'s on the pretest. They may have been aware of these errors and may have paid more attention to these stimuli on the posttest. We will come back to this in the next chapter. In any case, the

improvement on the /a:/-stimuli of the Dutch subjects in the music condition and their "tiny" improvement on the /ɑ/-stimuli is an important difference with the Spanish-speaking subjects in the same condition, who improved on both types of stimuli and made a smaller improvement on the /a:/-stimuli than the Dutch subjects.

Confusion matrix 3.7: Results on the posttest of the Dutch subjects who listened to the piece of classical music.

Recognised → Offered ↓	/ɑ/	/a:/	Totals
/ɑ/	487	73	560
/a:/	59	501	560
Totals	546	574	1120

The seven confusion matrices show us that the enhanced training improves subjects' discrimination of /a:/. In addition to this, subjects in the music condition show practice effects on the posttest: the performance of the Spanish-speaking subjects improves on both types of stimuli, whereas the performance of the Dutch subjects only improves on the /a:/-stimuli.

3.2 Logistic regression analysis

The next step in analysing the data was to model the /ɑ/-/a:/-responses with a logistic regression analysis (LRA). As an example, we model the responses with two factors: "Gender" and "Vowel". The LR model now is:

$\log(p/(1-p)) = c_0 + c_1$ Gender $+ c_2$ Vowel

If *p* is p(ɑ), because of the identity of p(ɑ) + p(a:) = 1, we can also write the argument of the logarithm as p(ɑ)/(1-p(ɑ)) or as p(ɑ)/p(a:). The LRA then tries to find the coefficients c_0, c_1 and c_2 that best fit the data. In this example, the two factors both happen to be categorical with only two values each.

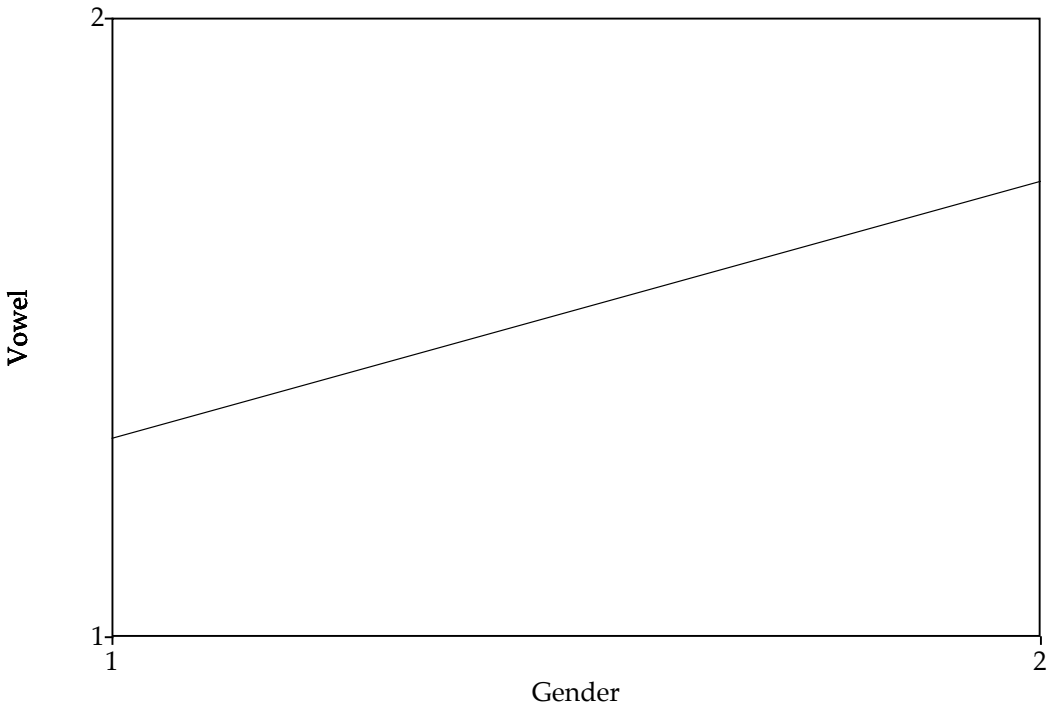
The boundary between the region where p(ɑ) > p(a:) and the region where p(ɑ) < p(a:), is where p(ɑ) = p(a:). On this boundary, the logarithm on the left side of the equation above equals zero. The equation for this boundary in our two-factor example is:

$0 = c_0 + c_1$ Gender $+ c_2$ Vowel

Because we only have two factors, this happens to be the equation of a straight line. For the present example, we let Praat compute the model for the data of the pretest of the Spanish speaking subjects in the enhanced condition. The LRA in Praat resulted in the following model:

log(p(ɑ)/p(a:)) = 0.772 + 0.354Gender - 0.853Vowel. In Figure 3.1 we have drawn the boundary line in the Gender x Vowel space.

Figure 3.1: Regression line obtained for the data of the present example (Spanish-speaking subjects, enhanced condition, pretest). Gender (1 = female, 2 = male), Vowel (1 = /ɑ/, 2 = /a:/).



Because the two-factor model above with Vowel and Gender as factors has only a limited possibility to model the variability that exists in the different /ɑ/ and /a:/ variations of the male and female stimuli, we have to increase the number of factors. Natural candidates are the formant frequencies, the duration and the fundamental frequency.

All analyses were carried out using SPSS for Mac version 16.0. Before conducting the main analyses, we carried out an analysis “per stimulus” and an analysis “per subject” to verify whether there were any anomalies in the data. On the basis of these analysis, we decided not to exclude any of the stimuli or any of the subjects: for the stimuli we did not find anything that could be problematic, but for the subjects we found that the majority differed significantly in their behaviour, which made it impossible for us to exclude any of them.

3.2.1 Analysis

For the analysis of the pretest, the following factors were entered in the LR model: Language, logduration, logf1, logf2, logf0, logf3, Gender and Gender∗Language. The results of the analysis of the pretest can be found in Table 3.1 on the next page. In this analysis and in all subsequent ones, the log of the duration, the fundamental frequency and the formant frequencies was always log10. The factor used for calculating the odds ratio was always *e*: with every one-unit increase of a factor in the model, the odds of a subject answering /a:/ would increase or decrease by a factor that was the outcome of the equation "e elevated to the power of the coefficient B". However, because we used log10 for duration, fundamental frequency and the formant frequencies, a one-unit increase of these factors meant an increase by a factor 10. In Table 3.1 below, for logduration, the increase in odds of a subject answering /a:/ is e^2.77, which means that when the duration of a stimulus becomes ten times longer, a subject is 15.89 times more likely to answer /a:/. Due to rounding of the coefficients, the actual numbers found under "odds ratio" may differ a little from the exact outcomes of the equations.

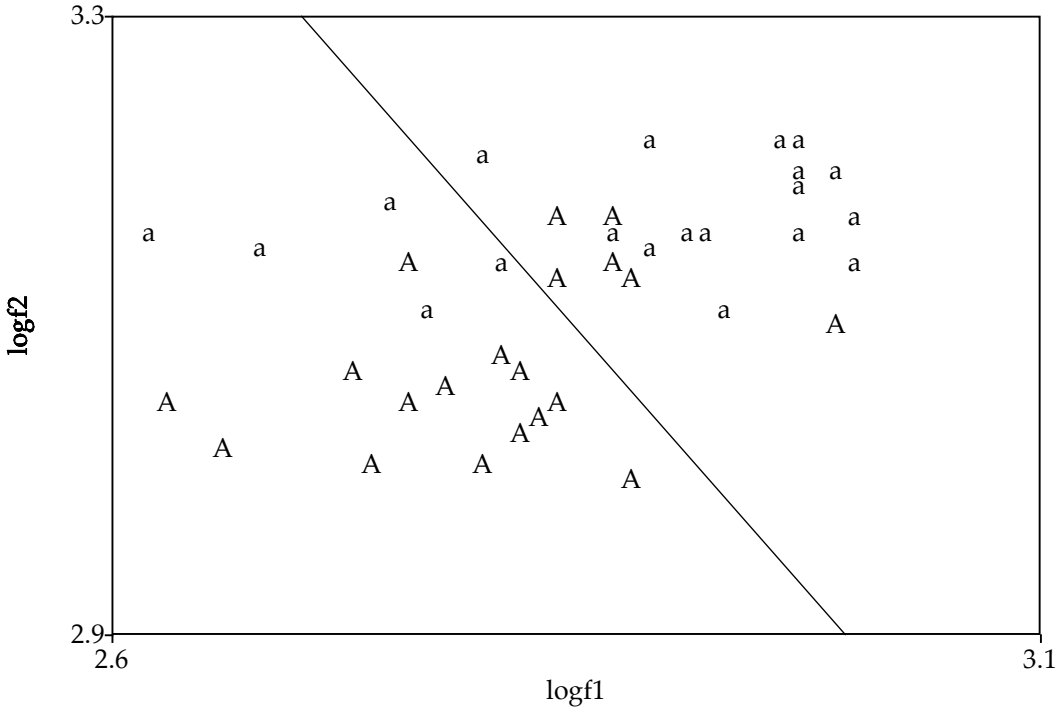
For the pretest, the Wald-statistic showed that the constant, logduration, logf1, logf2, logf3 and Gender were significantly different from zero. The (positive signs of) the coefficients show that the longer the duration or the higher the F1, F2 or F3, the more likely an /a:/ response is. Figure 3.2 shows the regression boundary together with the positions of the stimuli with respect to logf1 and logf2. The results were the same for Spanish-speaking and Dutch subjects, given the fact that Language did not make a significant contribution to the model. For Gender, the analysis confirms that, when the speaker was male, subjects were less likely to answer /a:/ than when the speaker was female. Again, the effect of this factor was the same for both groups of subjects. We also checked the Cook's Distances, Leverage values, Standardised Residuals and DFBetas and found no cases that had an extremely large influence upon the model.

Table 3.1: Results of the logistic regression analysis carried out on the data of the pretest. Language (1 = Spanish, 0 = Dutch), Gender (0 = female, 1 = male), log = log10.

			95% CI for Odds Ratio		
	B	SE	Lower	Odds Ratio	Upper
Constant	-26.48***	2.55			
Language	0.10	0.07	0.97	1.10	1.26
logduration	2.77***	0.17	11.34	15.89	22.27
logf1	2.89***	0.30	9.91	17.91	32.37
logf2	2.47***	0.50	4.46	11.86	31.54
logf0	-0.07	0.30	0.52	0.93	1.68
logf3	1.36*	0.60	1.21	3.88	12.48
Gender*	0.21*	0.11	1.00	1.24	1.52
Gender * Language	-0.07	0.09	0.78	0.93	1.12

R^2= .11 (Hosmer & Lemeshow), .14 (Cox & Snell), .18 (Nagelkerke). Model $\chi^2(8) = 1598.11$, $p < .001$. *p < .05, **p < .01, ***p < .001

Figure 3.2: Regression boundary obtained for the data of the pretest, together with the positions of the stimuli with respect to logf1 and logf2. (A= /ɑ/ and a=/a:/).



In a second analysis, we analysed the data of the posttest, Now, the following factors were entered into the model: Language, Condition, logduration, logf1, logf2, logf0, logf3, Gender, Gender∗Language and Gender∗Condition. The results of this analysis can be found in Table 3.2. It turned out that, apart from the constant, the following factors made a significant contribution to the model: Condition, logduration, logf1, logf2, Gender and the interaction Gender∗Language. The effect of Condition indicated that subjects in the enhanced condition were more likely to answer /a:/ than subjects in the music condition. This was also shown by the confusion matrices in section 3.1. The effects of logduration, logf1 and logf2 were the same as on the pretest: the longer the duration and the higher F1 and F2, the more likely was a subject to answer /a:/. Logf3 did not make a significant contribution on the posttest. The effect of Gender was also the same as on the pretest: subjects were more likely to answer /a:/ when the speaker was female than when the speaker was male. It turned out that the interaction Gender∗Language was also significant: Spanish-speaking subjects were even less likely to answer /a:/ when the speaker was male then the Dutch subjects. There was no main effect of Language. We again checked the Cook's Distances, Leverage values, Standardised Residuals and DFBetas for the constant and the values showed that there were no cases that had an extremely large influence upon the model.

The last part of the logistic regression analysis consisted of checking whether there was multicollinearity between factors, which means that different factors explain the same variance: this makes it difficult to assess which factors are important. We looked at the Tolerance values, the Variance Inflation Factor (VIF) and the eigen values for each factor. The analysis was run separately for the pretest and the posttest. The results of both analyses indicated that there were no serious multicollinearity problems. We found a small amount of multicollinearity between Gender and F3, but this was to be expected, given the fact that the height of this formant frequency tends to correlate with the gender of the speaker: if the speaker is female, F3 is usually higher than when the speaker is male. The exact outcomes can be found in Appendix 5.

Table 3.2: Results of the logistic regression analysis carried out on the data of the posttest. Language (1 = Spanish, 0 = Dutch), Condition (0 = music, 1 = enhanced), Gender (0 = female, 1 = male), log = log10.

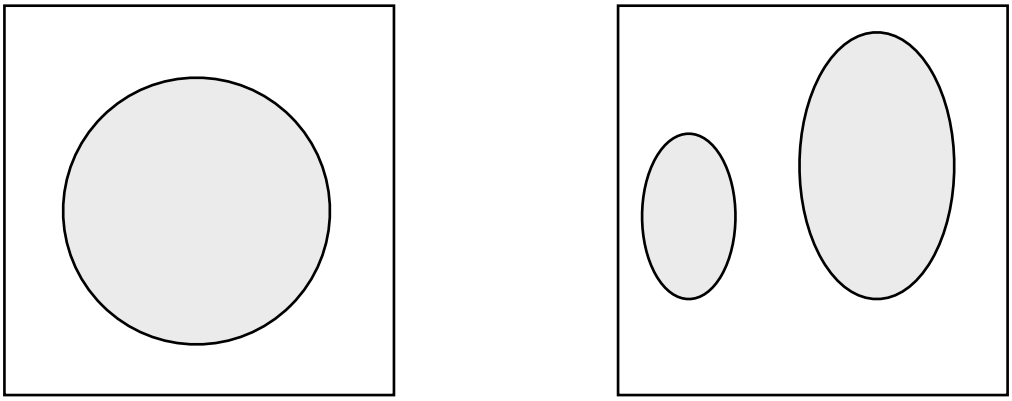
			95% CI for Odds Ratio		
	B	SE	Lower	Odds Ratio	Upper
Constant	-26.98***	2.65			
Language	0.08	0.07	0.95	1.09	1.25
Condition	0.19**	0.06	1.07	1.21	1.36
logduration	3.63***	0.18	26.72	37.82	53.54
logf1	2.64***	0.31	7.62	13.99	25.69
logf2	3.75***	0.52	15.40	42.29	116.17
logf0	0.30	0.31	0.73	1.35	2.50
logf3	-0.23	0.62	0.24	0.80	2.72
Gender	0.36**	0.12	1.13	1.44	1.83
Gender * Language	-0.25*	0.10	0.65	0.78	0.95
Gender * Condition	-0.10	0.09	0.77	0.91	1.08

R^2= .16 (Hosmer & Lemeshow), .19 (Cox & Snell), .26 (Nagelkerke). Model $\chi^2(10)$ =2337.037 , $p < .001$.
*p < .05, **p < .01, ***p < .001

3.3 Discriminant analysis and vocal tract normalization

In the previous section we modelled discrimination on the basis of the subjects' responses. In this section, we will use a discriminant analysis to see how well a machine can classify the 20 tokens of /ɑ/ and the 20 tokens of /a:/. In a discriminant analysis, a mathematical algorithm calculates the weight of each factor when discriminating between the categories. The weight of each factor is then presented as a coefficient in a mathematical equation. This equation consists of an intercept and two or more coefficients that belong to the factors the model takes into account. It can be used for predicting log(p(/ɑ/)/p(/a:/)) for every value of the factors in the model. Normally, the set of stimuli on which the model is trained is different from the set for which the predictions have to be made, but in the present model this is not the case. However, we decided not to use any corrections. This is because we only use the discriminant analysis as an indication of how discriminable the stimuli are. When the equation has been obtained, the different categories are projected in a multidimensional space: the number of dimensions is equal to the number of factors in the model. However, only two of the dimensions are shown when Praat displays the space: it is displayed in a way that, when one looks at it, one sees the maximum degree of separability between the categories. The angle from with one looks at the categories is calculated by the programme's algorithm. Praat also indicates the percentage of correct discriminations for every combination of factors. To make a comparison: when one takes a telescope and looks at distant galaxies, one will be able to see large clusters of stars. However, the clusters one sees from earth would look very different if one observed them from a different angle. See for example Figure 3.3. Imagine that the circle in the first picture is a cluster of stars. When observed from earth, it looks like the stars form one large cluster. However, when observed from a different angle (e.g. from another galaxy), one sees that there are in fact two clusters of stars with a large space in between. This would be an angle from which the two clusters can be perfectly discriminated. The percentage of correct discriminations would thus be 100%. However, an important difference with a "real" discriminant analysis is that Praat computes the average for each category and chooses the category with the smallest distance to a new stimulus for classification (not every factor is assigned the same weight: the weight depends on the amount of variance within each factor). The picture with the clusters of stars thus only serves to illustrate the example.

Figure 3.3: Two imaginary clusters of stars when observed from earth (left) and from another galaxy (right).



In the model used for the present study, we wanted to know how well the algorithm in Praat could separate the categories /a:/ and /ɑ/ from each other. Therefore, the categories in the model are /a:/ and /ɑ/. Each category consists of 20 data points (the tokens of /a:/ and /ɑ/ used in the study). A scatterplot of the categories is shown in Figure 3.4. The factors on the basis of which we wanted the model to perform the analysis were the following: Duration, F0, F1, F2, and F3. The space on which the two categories are projected will thus be a fivedimensional one. Before entering the factors into the model, we took the log of each of them. The script for the discriminant analysis can be found in Appendix 4.

It turned out that the model was able to discriminate /a:/ from /ɑ/ in 100% of the cases on the basis of logduration, logf0, logf1, logf2 and logf3. The percentage of correct discriminations for every combination of duration, fundamental frequency and first three formant frequencies can be found in Table 3.3. Figure 3.5 shows a scatterplot in which the categories are discriminated on the basis of logduration, which turned out to be the most important factor in the model with 97,5% correct discriminations. In fact, the data points are located along a line (because there are only two categories that have to separated from each other), but because Praat has assigned the upper part of the Figure to the vowel /a:/ and the lower part to the vowel /ɑ/, half of the datapoints are located in the upper right part of the Figure and the other half in the lower left part.

Discriminating the two categories on the basis of duration, fundamental frequency and the formant frequencies is in fact what the subjects in the present study had to do, and

normally people are far better at categorising stimuli than are algorithms. However, none of the subjects in the study reached a score of 100% correct on both tasks. Many Spanish-speaking subjects had even less than 65% correct. This means that there is probably a factor external to the stimuli themselves which confuses the subjects. Given the fact that a significantly larger effect of speaker gender was found on the posttest for the Spanish-speaking subjects, it may be the case that they are less able to perform a correct vocal tract normalization than the Dutch subjects. The vocal tracts of men are longer than those of women, and native speakers of a language take this into account (be it subconsciously) when they listen to speech produced by male and female speakers. We included vocal tract normalization in our model and looked whether this indeed made the vowels /a:/ and /ɑ/ more discriminable from each other. However, a simple scaling for vocal tract length differences, i.e. multiplying the female formant frequencies by a factor of 15/17, the approximate ratio of their average tract lengths, was not effective. We also investigated whether the male and female stimuli differed significantly in duration: this was not the case.

Given that the male and female stimuli do not differ significantly in duration and that vocal tract normalization does not really help subjects when discriminating /a:/ from /ɑ/ and cannot explain the presence of the clear gender effect found for the set of stimuli used in this study, it may be the case that the subjects' responses have been influenced by the two possible answers in the XAB-task: the synthesised tokens of /a:/ and /ɑ/, which each have a duration of 140 ms. This hypothesis is supported by the fact that the model used for the discriminant analysis was still able to discriminate /a:/ from /ɑ/ in 97,5% of the cases when duration was the only factor taken into account. Duration seems thus to be a very important cue for deciding whether a stimulus is a token of /a:/ or /ɑ/ and nonnative speakers may rely more heavily on this cue than native speakers do (e.g. Cebrian, 2006; Escudero, 2001). In the next section, we will look whether the possible answers have had any influence by computing the correlations between various factors.

Figure 3.4: Scatterplot of the tokens of /a:/ and /ɑ/ specified for gender (F=female speaker, M=male speaker) for the dimensions F1 and F2.

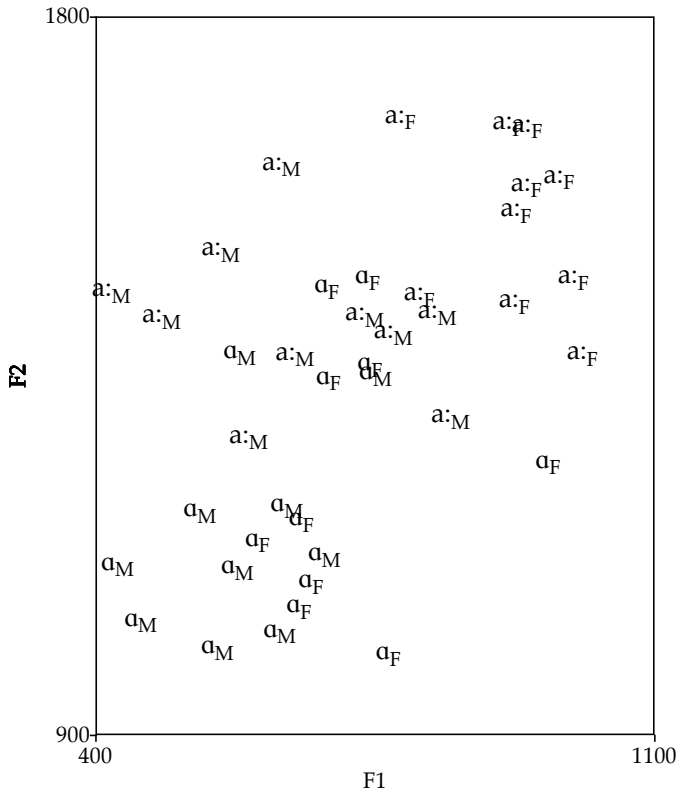


Figure 3.5: Discrimination of /a:/ and /ɑ/ on the basis of logduration. Vowel (1 = /ɑ/, 2 = /a:/)

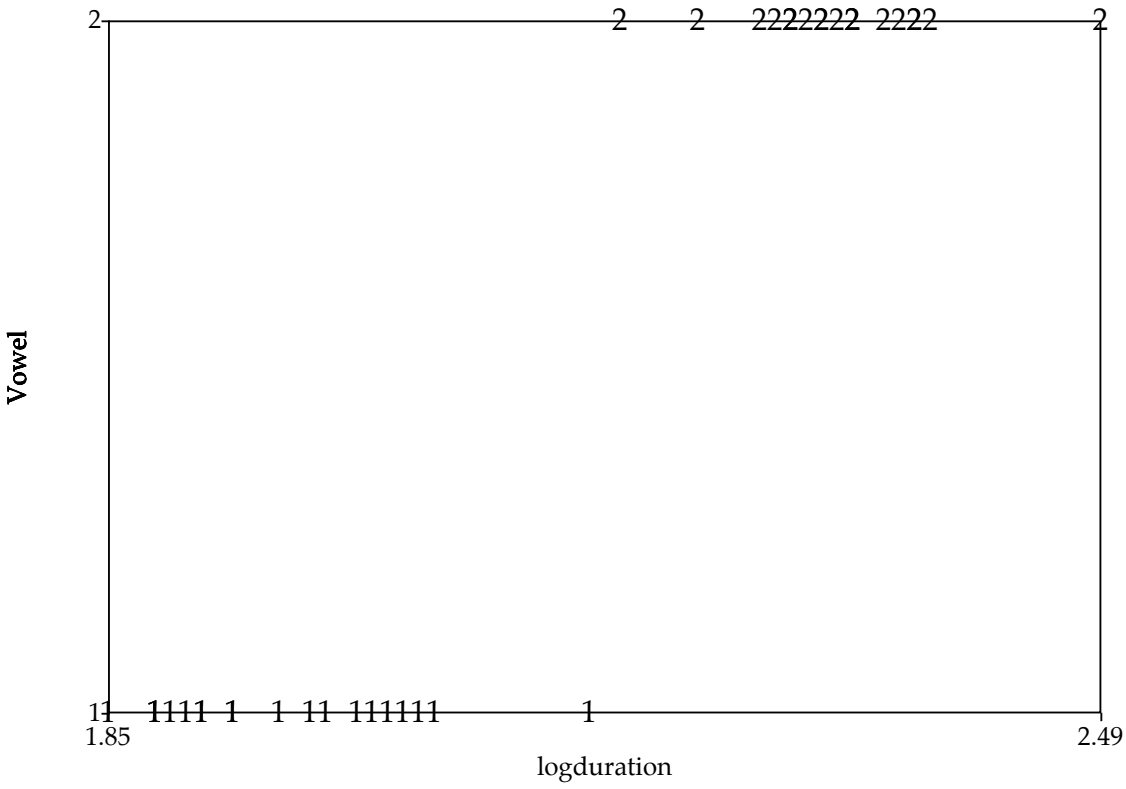


Table 3.3: Percentage of correct discriminations the model for the discriminant analysis made for every combination of factors taken into account.

Factors taken into account by the model	Percentage correct discrimination
logduration logf0 logf1 logf2 logf3	100
logduration logf0	97.5
logduration logf1	97.5
logduration logf2	100
logduration logf3	97.5
logduration logf0 logf1	97.5
logduration logf0 logf2	97.5
logduration logf0 logf3	97.5
logduration logf1 logf2	100
logduration logf1 logf3	97.5
logduration logf2 logf3	100
logduration	97.5
logf0	60
logf1	72.5
logf2	80
logf3	50
logf0 logf1 logf2 logf3	92.5
logf1 logf2 logf3	90
logf1 logf2	80
logf1 logf3	72.5
logf2 logf3	87.5
logf0 logf1	77.5
logf0 logf2	85
logf0 logf3	67.5

3.4 The role of the possible answers: correlations

As was mentioned in section 3.3, we wanted to know whether the characteristics of the two possible answers, the synthesised tokens of /a:/ and /ɑ/ pronounced by a "male" speaker, played a role in the discrimination process of the subjects of the present study. To investigate this, we computed the distance from these answers for every stimulus. We computed these distances as follows: we first computed the square of the outcome of "the log of a formant frequency minus the log of that same formant frequency of one of the possible answers". We did this for F1, F2 and F3. We then added these numbers and after that, we took the square root of the outcome. This number would be the distance of a stimulus to this possible answer. The procedure was then repeated for the other possible answer to compute the distance of the stimulus to this answer. Both distances were included as factors in the computation of the correlation-coefficients. We then computed the correlation-coefficient *r* for each combination of two factors in the model. The correlation-coefficients were computed for every group of subjects (Spanish/Dutch), condition (enhanced/music) and test (pretest/posttest). In addition to this, they were computed separately for all stimuli taken together, the male stimuli and the female stimuli. We did this because possible differences between the male and female stimuli may be obscured when all stimuli are taken together. The script we used for computing the correlation-coefficients can be found in Appendix 4.

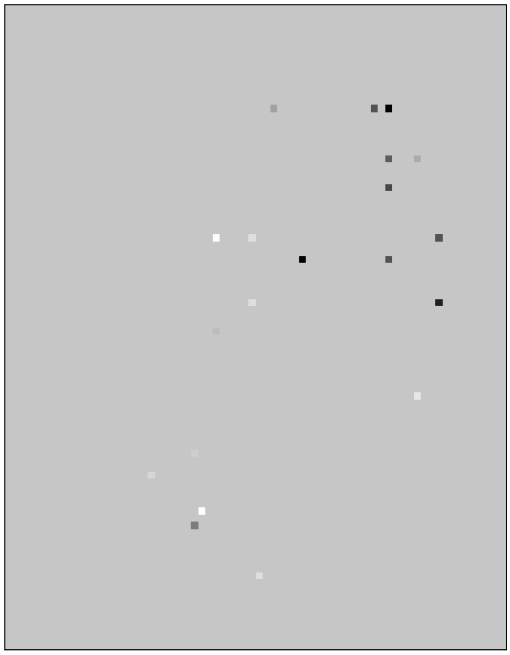
It turned out that, irrespective of subject group, condition and test, the correlation-coefficients were virtually the same for all stimuli taken together, the male stimuli and the female stimuli respectively. The results also indicated that in none of the cases, the distance of a stimulus to the possible answers had played an important role in the discrimination process, with the only exception of the extremely large correlation we found between the distance of a stimulus to the possible answer /ɑ/ and the number of /a:/'s answered to this stimulus for female stimuli. This is probably caused by the fact that the female /a:/'s are all "extreme": they are located far away from the possible answer /ɑ/ in the formant space (see Figure 3.4). The fact that there was no extremely large correlation between the distance of a stimulus to the possible answer /a:/ and the number of /a:/'s answered to this stimulus for female stimuli, supports this hypothesis. The correlations confirmed the outcomes of the LRA in section 3.2.1 for logduration, logf1 and logf2.

3.5 The effect of training upon the attention subjects pay to the frequencies of F1/F2 and the duration of a stimulus

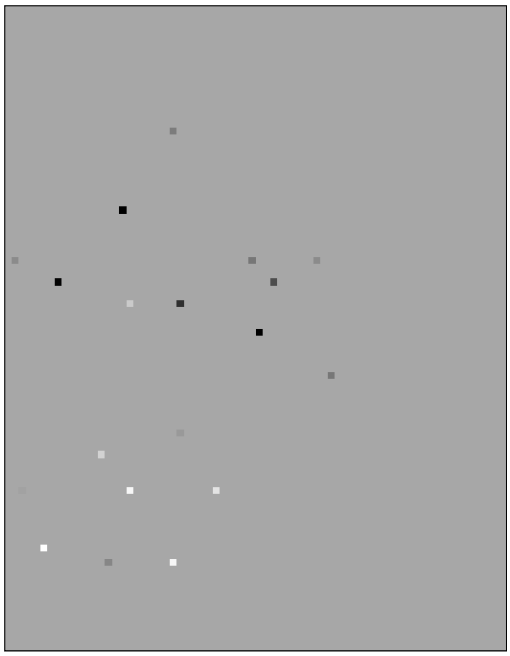
In this section, we will illustrate the effect of training upon the attention subjects pay to the F1, F2 and duration of the stimuli. The outcomes of the LRA in section 3.2.1 and the discriminant analysis in section 3.3 indicated that these three factors are the most important ones in the discrimination process. We will first illustrate the difference between the number of /a:/ responses to a particular stimulus on the pretest and the posttest in relationship with the height of the formant frequencies F1 and F2 of this stimulus. After that, we will analyse the difference between the percentage of /a:/ responses to a particular stimulus on the pretest and the posttest in relationship with its duration.

The effect of training upon the attention subjects pay to the F1 and F2 of a stimulus was visualised as follows: we made matrices that contained the difference between the total number of /a:/ responses per stimulus on the pretest and the posttest for each group of subjects, training condition and speaker gender. All calculations were done in Praat with a script that can be found in Appendix 6. The matrices indicated that, after the training with the enhanced tokens of /a:/ and /ɑ/, both Spanish-speaking and Dutch subjects indeed started to pay more attention to the F1 and F2 of the stimuli. For the female stimuli, the Spanish-speaking subjects had set a boundary based on the F1 of the stimuli, whereas for the Dutch subjects there was no important difference. This is probably due to a ceiling effect on the pretest, which left little room for improvement on the posttest. For the male stimuli, both groups of subjects had set a clear boundary based on the F2 of the stimuli. Spanish-speaking subjects in the music condition did not show any important differences on the posttest, so listening to the piece of classical music did not alter their strategies. The Dutch subjects, however, showed important improvements on the /a:/-stimuli. For female stimuli, it was unclear whether they had started to pay more attention to the F1, the F2 or both formant frequencies, but for the male stimuli, they had started to pay more attention to F2. This is a surprising finding, which can only be explained by an increased familiarity with the stimuli. To illustrate this discussion, we will now show the matrices for the female and the male stimuli for the Spanish-speaking subjects in the enhanced condition. The female stimuli are found in Matrix 3.1 and the male stimuli in Matrix 3.2. If the colour of a particular stimulus is darker than that of the background, this indicates that more /a:/ responses have been made after the training, whereas a lighter colour indicates that less /a:/ responses have been made.

Matrix 3.1: Differences between the numbers of /a:/ answered on the pretest and the posttest on the female stimuli by the group of Spanish-speaking subjects who received enhanced training.



Matrix 3.2: Differences between the numbers of /a:/ answered on the pretest and the posttest on the male stimuli by the group of Spanish-speaking subjects who received enhanced training.



In order to visualise the effect of training upon relationship between a stimulus' duration and the percentage of /a:/responses out of the total number of responses given to this stimulus, we made scatterplots that contained both factors. We used the log of the duration of every stimulus, because in this way, the spaces between the stimuli in the Figures were larger. All calculations were done in Praat with a script that can be found in Appendix 7.

The scatterplots indicated that the relationship between a stimulus' duration and the percentage of /a:/ responses was much stronger for the Dutch subjects than for the Spanish-speaking ones. For subjects in the enhanced condition, this relationship became stronger after the training. For Spanish-speaking subjects in the music condition this was not the case, but for Dutch subjects in this condition, the relationship became stronger for /ɑ/-stimuli, which again can only be explained by an increased familiarity with these stimuli. The scatterplots also indicated that, in general, the stimuli that are problematic for the Spanish-speaking subjects are also problematic for the native Dutch subjects. Nevertheless, the former appear to be much more affected by the relative difficulty of a stimulus than the latter. Both groups of subjects tend to obtain the highest scores on the more extreme tokens of each vowel with respect to F1 and F2 frequencies. This also holds for the stimuli 12aa and 80a, which show virtually no difference in duration. Still 12aa is mostly correctly identified as /a:/ and 80a as /ɑ/. A closer inspection of these stimuli showed that stimulus 80a is situated in the middle of the F1-spectrum and towards the lower part of the F2-spectrum. Stimulus 12aa has high values of both F1 and F2, which are indeed a characteristic of most /a:/-stimuli when compared to /ɑ/-stimuli. This may have enabled subjects to identify 80a as /ɑ/ and 12aa as /a:/. This discussion can best be illustrated by looking at the results on the pretest of both groups of subjects. Just as in the confusion matrices, we took the results of subjects from both conditions together. The results for the Spanish-speaking subjects can be found in Figure 3.6 and those of the Dutch subjects can be found in Figure 3.7.

Figure 3.6: Percentage of /a:/'s answered per stimulus on the pretest for both groups of Spanish-speaking subjects taken together. The two digits indicate the ID of the speaker as assigned by Adank et al. (2004), "a" means that the stimulus is a token of /ɑ/ and "aa" means that the stimulus is a token of /a:/.

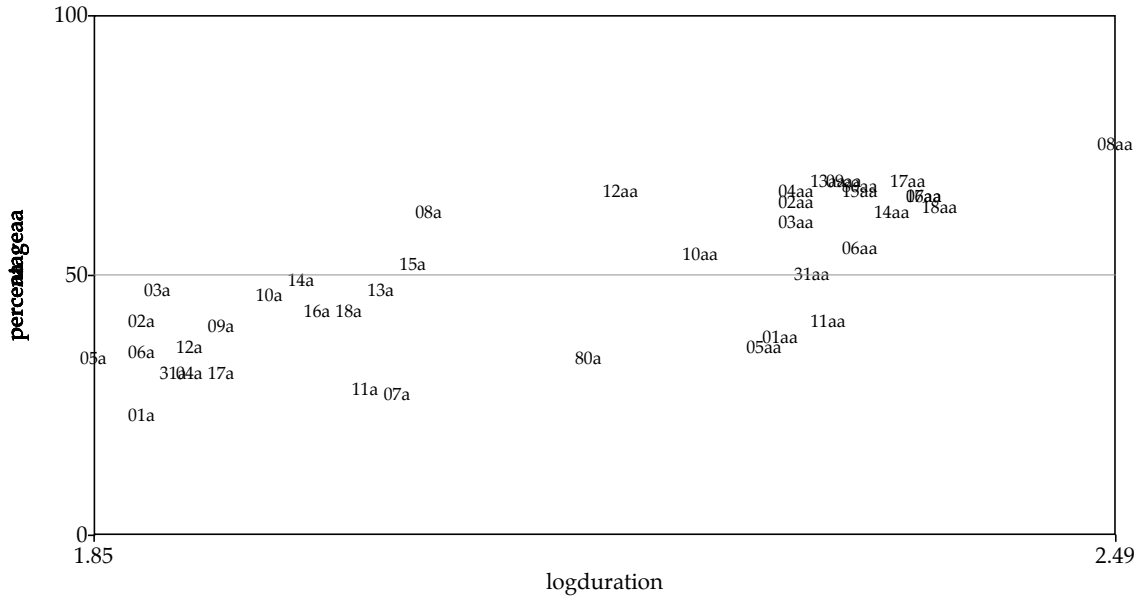
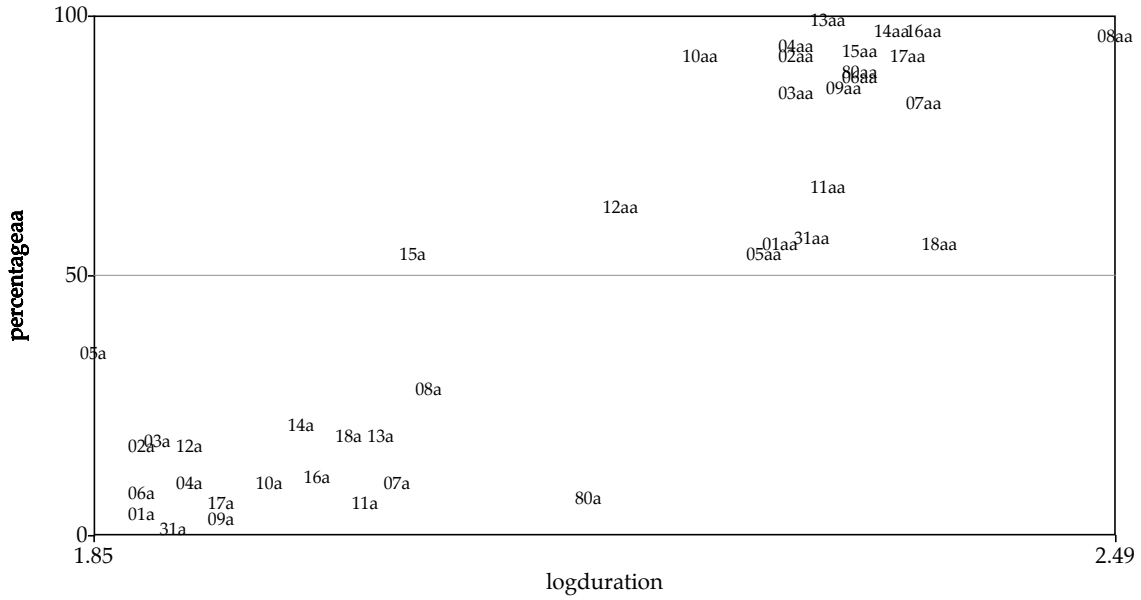


Figure 3.7: Percentage of /a:/'s answered per stimulus on the pretest for both groups of Dutch subjects taken together. The two digits indicate the ID of the speaker as assigned by Adank et al. (2004), "a" means that the stimulus is a token of /ɑ/ and "aa" means that the stimulus is a token of /a:/.



Chapter 4: Discussion and conclusions

4.1 Summary of the results

We will now summarise the results found in the previous chapter, thereby focusing on the role of speaker gender in the perception and categorization of vowels by both groups of subjects in the present study. The confusion matrices showed that in the enhanced condition, both groups of subjects improved remarkably after the training, especially on the /a:/-stimuli. In the music condition, both groups of subjects showed practice effects on the posttest: the Spanish-speaking subjects improved slightly on both types of stimuli, whereas the Dutch subjects only improved on the /a:/-stimuli.

The results of the logistic regression analysis (LRA) indicated that there was a significant effect of speaker gender for both groups of subjects on the pretest as well as on the posttest: subjects were significantly less likely to answer /a:/ when the speaker was male and significantly more likely to do so when the speaker was female. On the posttest, this gender effect became even stronger for the Spanish-speaking subjects, no matter whether they were in the enhanced condition or the music condition.

The results of the discriminant analysis and the vocal tract normalization showed that Praat was able to discriminate /a:/ from /ɑ/ in 100% of the cases on the basis of duration, F0, F1, F2 and F3. Vocal tract normalization did not improve the model's discrimination of the two vowels. This means that the errors the subjects made were probably caused by some factor external to the stimuli themselves. We also investigated whether the male and the female stimuli differed significantly in duration: this was not the case.

The analysis of correlation-coefficients we carried out to investigate whether the two possible answers had played a role in the discrimination process indicated that this was not the case. In addition, the results confirmed the outcomes of the LRA.

In order to visualise the effect of training upon the attention subjects pay to the F1, F2 a stimulus, we made matrices that contained the number of /a:/'s answered to every stimulus before and after the training phase in relationship with the F1 and F2 of this stimulus. The results indicated that the training with the enhanced tokens of /a:/ and /ɑ/ indeed improved subjects' discrimination of /a:/-stimuli. The Spanish-speaking subjects improved more than the Dutch subjects, but this may be due to a ceiling effect found for the latter for many of the stimuli. After the training with the enhanced tokens, subjects started paying more attention to the F1 for male stimuli and the F2 for female stimuli. Spanish-speaking subjects in the music

condition did not pay more attention to the F1 and the F2 of a stimulus after the training phase, but Dutch subjects in this condition seemed to pay more attention to the F2 for male stimuli after listening to the piece of classical music.

In the scatterplots at the end of section 3.5, we showed the relationship between the duration of a stimulus and the number of /a:/'s answered to this stimulus and the effect of training upon this relationship. The results showed that the relationship between duration and number of /a:/'s answered was much stronger for the Dutch subjects than for the Spanish-speaking ones. Furthermore, the training with enhanced tokens made the relationship stronger in both groups of subjects, whereas the piece of classical music did much less so. Again, the effect of the training with the enhanced tokens was more important for /a:/-stimuli than for /ɑ/-stimuli.

4.2 Discussion

For convenience, we will repeat the hypotheses we had with respect to vowel normalization before conducting the experiments:

- If the Spanish-speaking subjects have difficulties with vowel normalization, we expect them to perform worse on the female stimuli than on the male stimuli: /a:/ has higher F1 and F2 frequencies than /ɑ/ and since women have overall higher formant frequencies than men, a female /ɑ/ might be confused with a male /a:/.
- Given the fact that in the enhanced condition subjects receive training with male tokens of /a:/ and /ɑ/, we expect the performance of the Spanish-speaking subjects on the second XAB-task to improve more for the male stimuli than for the female stimuli.
- We expect no important differences between both tasks in the music condition for both Spanish-speaking and native Dutch subjects.
- We expect the native Dutch subjects to perform almost at ceiling on all tasks, regardless of the condition.

The first hypothesis was borne out: the Spanish-speaking subjects indeed showed the expected gender effect. They were more likely to answer /a:/ when the speaker was female than when the speaker was male, but given the confusion matrices, they were also more likely to answer /ɑ/ when the speaker was male than when the speaker was female. Another unexpected finding was that the gender effect was also present in the native Dutch subjects.

The second hypothesis was not borne out: after the training with the enhanced tokens, subjects improved on both the male and the female stimuli to a similar degree. Surprisingly, this improvement was much larger for /a:/-stimuli than for /ɑ/-stimuli. We also found that the gender effect had become larger for the Spanish-speaking subjects after the training phase. This was not caused by the training with the enhanced tokens, given the fact that the interaction Gender∗Condition was non-significant.

The third hypothesis was only partly borne out: the Spanish-speaking subjects only improved slightly after listening to the piece of classical music, as was expected, but the Dutch-speaking subjects made an important improvement on the /a:/-stimuli.

The fourth hypothesis was not borne out: the Dutch subjects did not perform at ceiling, especially on the pretest. We even had to exclude 14 subjects because of extremely low scores (less than 70% correct) on the pretest.

How can these results be explained? The Spanish-speaking subjects showed the expected gender effect, which means that they probably have difficulties with normalization for speaker gender in the L2. It was very surprising, however, to find the same gender effect in the native Dutch subjects. Does this mean that native listeners also have difficulties with this type of vowel normalization? It is unlikely that they experience this difficulty in everyday live, given the fact that, most of the time, they understand what is being said, regardless of the gender of the speaker. Therefore, it is not unlikely that the nature of the task has played a role here. In spoken language the context in which a vowel appears often disambiguates. In XAB-tasks like the ones used in the present study, this is clearly not the case. In addition to this, subjects are asked to compare the first sound they hear to two possible answers, which is not an everyday task either. The fact that subjects had to compare isolated vowels produced by different speakers, which appeared in a randomised order, to the answers given by one particular “male” speaker made it impossible for them to build expectations about the speech of the next speaker they were going to hear. Previous studies (e.g. Halberstam & Lawrence, 2004) already showed that native speakers performed less well in a mixed speaker condition than in a blocked speaker condition. Subjects were also unable to listen acoustically, given the large inter stimulus interval of 1.2 s (cf. Werker & Logan, 1985): this was also shown by the results of the analyses of correlation-coefficients. All this means that they first had to identify a stimulus as being a token of /a:/ or /ɑ/ and had to compare it to the possible answers, which also had to be identified as tokens of /a:/ and /ɑ/, especially in the first trials: it is likely that,

after a few trials, subjects became more familiar with them. However, because subjects never knew whether the next speaker they would hear would be male or female (as was the case in the study conducted by Johnson et al., 1999), it is not unlikely that they used 'prototypes' of the vowels /a:/ and /ɑ/ and compared the vowels produced by every speaker to these prototypes. It is very likely that prototypes of vowels are not specified for speaker gender: previous studies have shown that infants and young children are already able to identify vowels produced by male, female and even child speakers as belonging to the same vowel category (e.g. Marean et al., 1992; Kuhl, 1983; Kubaska & Aslin, 1985), even if they have only been trained on vowels produced by one of the speaker genders. Nevertheless, native speakers are able to "adapt" these prototypes to the speech and gender of a particular speaker, even after a short exposure to this speech (see for example Van Bergem et al., 1987). The use of prototypes by the subjects in the present study would also explain the gender effect: the vowel prototypes used by the subjects in the present study may indeed be more or less the "average" of the vowels produced by male and female speakers. However, as was already explained in previous chapters, the formant frequencies of men are generally lower than those of women and also for /ɑ/ when compared to /a:/. If the "average" protoypes are used, vowels produced by male speakers may be more likely to be identified as /ɑ/ and vowels produced by female speakers as /a:/, especially when these vowels are somewhat extreme with respect to formant frequencies (i.e. low formant frequencies for male speakers and high formant frequencies for female speakers). This is exactly what we found in the present study. However, the use of "average" prototypes does not explain the larger gender effect found for the Spanish-speaking subjects on the posttest. This means that the gender effect found in the Spanish-speaking subjects may have a (partly) different origin than the gender effect found in the native Dutch subjects. The Spanish-speaking subjects seem to get more confused after the training phase, regardless of the condition they are in. It is difficult to imagine them using "narrower" prototypes on the posttest than they did on the pretest, which means that vowels with more extreme frequencies of F1 or F2 are more likely to be discriminated incorrectly. However, if we explain it the other way around by saying that the native speakers start using broader prototypes on the posttest, which also comprise more extreme vowels, the explanation makes much more sense. Native speakers of a language quickly adapt themselves to the speech of speakers they have never heard before and even to speakers who speak a different dialect. In this way, the native speakers in the present study may have also been able to quickly adapt themselves to the stimuli: already on the posttest, they may have used vowel

prototypes that were more adapted to the characteristics of the stimuli used in the experiment. This enabled them to discriminate more of the extreme vowels correctly, even though the gender effect was still visible. If this is the case, nonnative speakers may be less flexible than native speakers with respect to vowel prototypes, either because these prototypes have (still) not been established completely, or because the prototypes from the L1 have an influence upon those of the L2. It may even be the case that nonnative speakers use different neural networks than native speakers do (e.g. Minagawa-Kawai et al., 2004). However, more research is needed.

Another surprising finding were the low scores obtained by many of the Dutch subjects, especially those in the music condition. We had to exclude 11 subjects from the music condition and 3 from the enhanced condition due to a score of less than 70% correct on the pretest. This finding was the more surprising because previous studies showed that native speakers usually reached a very high percentage correct on the same stimuli presented in similar tasks as the ones used in the present study (e.g. Escudero et al., 2009; Escudero et al., in preparation). However, the native speakers who participated in those two studies were mainly students, whereas the subjects who participated in the present study, particularly the Dutch subjects in the music condition, had more varied backgrounds. Their mean age was also a little higher than that of the subjects in the other two studies. It is very well possible that the different backgrounds of the subjects in the present study have caused the differences with the other two studies: the XAB-task is a task with relatively high memory demands (e.g. Pisoni & Lazarus, 1974) and this may have made the task too difficult for people who were unfamiliar with performing such a task on a computer or with the idea of being "tested", as was the case with many of the "older" subjects in the study. The origin of the "problem" may also lie in the characteristics of the two possible answers: both answers had the same characteristics as those mentioned in the article by Pols et al. (1973). However, the present study was conducted in 2009, which means that at the time the subjects had to perform the XAB-tasks, 36 years had passed since Pols et al. established the average formant frequencies for Dutch vowels. Like every living language, Dutch has changed in that time period and this change has affected the pronunciation the vowels /a:/ and /ɑ/. A few subjects indicated that the possible answer /ɑ/ sounded like /o/. After the training with enhanced tokens, more subjects reported having heard /o/ instead of /ɑ/ sometimes. This effect was already found by Paola Escudero (personal communication), but for lower formant frequencies than the ones used for the enhanced tokens of the present study. The possible answer /a:/ may have sounded

strange to some subjects due to the fact that the two possible answers were matched on duration, which made the /a:/ much shorter than it would normally be. Some of the subjects also indicated that they had to get used to the fact that the two possible answers were produced by a computer voice. It is not unlikely that some of the subjects who found the possible answers the "strange-sounding", particularly the ones who were not familiar with performing a task like the XAB-task, became confused by them: if the possible answer /a:/ lies somewhere in between /a:/ and /ɑ/ (due to the match in duration) and the possible answer /ɑ/ sounds more like /o/, subjects may get "lost" and indeed, we found that relatively many of the Dutch subjects performed at or near chance level. This may also explain the fact that many of the subjects in the music condition performed very poorly on the male /a:/-stimuli: none of the two possible answers did match subjects' expectations for a male /a:/ and this made them confused. However, because the speaker was male, they were more likely to choose /ɑ/ as the final answer. The female /a:/-stimuli were all more extreme and could not directly be matched with the two possible answers anyway. Therefore, they were very likely to be matched with the possible answer with the highest formant frequencies: /a:/. Nevertheless, the results on the posttest indicated that subjects' performance on the male /a:/-stimuli had improved remarkably: they had probably had become "aware" of the problem with these stimuli and paid more attention to them. However, they may also have adapted their prototypes to the male /a:/-stimuli and/or the "strange-sounding" possible answers. This may also explain that they started paying more attention to the F2 for male stimuli after the training phase.

The characteristics of the two possible answers bring us to another finding of the present study: the fact that subjects in the enhanced condition improve much more on the /a:/-stimuli than on the /ɑ/-stimuli. We just mentioned that some of the Dutch subjects indicated that the possible answer /ɑ/ sounded like /o/ to them. In the enhanced tokens of /ɑ/, the frequencies of F1 and F2 of this vowel were lowered to make the difference with /a:/ more salient. Especially the lowest frequency steps sounded even more like /o/ than the possible answer, as was also indicated by various subjects. If subjects identify (some of) the enhanced tokens of /ɑ/ as tokens of /o/, it is unlikely that these tokens help them to improve their categorization of /ɑ/. Lively & Pisoni (1997) already found that non-prototypical vowels may be labelled by the subjects as belonging to different vowel categories than the one intended by the experimenter. On the contrary, the enhanced tokens of /a:/ always sounded like tokens of

/a:/. This made them more “effective” than the enhanced tokens of /ɑ/.

With respect to the analyses we carried out to investigate the effect of the training with the enhanced tokens of /a:/ and /ɑ/ upon the attention subjects paid to duration, F1 and F2, we saw that this training made subjects pay more attention to the F1 for female stimuli and the F2 for male stimuli. However, this difference may be caused by the characteristics of the stimuli: the female stimuli differed mainly in F1 and the male stimuli in F2. This nevertheless indicates that, after the enhanced training, subjects, even the nonnative ones, are able to “adapt” themselves to these characteristics of the stimuli and do not use the formant frequencies in the same way for both types of stimuli. The enhanced training also made subjects pay more attention to the duration of a stimulus, which is remarkable, because all enhanced tokens were matched on duration (140 ms). The fact that subjects were better able to discriminate /a:/-stimuli as tokens of /a:/ after the training on the basis of F1 and F2 may have left more room for them to pay attention to the duration of a stimulus. However, we still saw that both for the Spanish-speaking subjects and the Dutch ones, the stimuli with the highest scores tended to be the more extreme ones with respect to the frequencies of F1 and F2. We also saw that the relationship between the duration of a stimulus and the number of /a:/’s answered to this stimulus was stronger for the Dutch subjects than for the Spanish-speaking ones. This is most likely caused by the fact that both possible answers had a duration of 140 ms, which confused the Spanish-speaking subjects more than the Dutch ones: the former probably rely more on the durational cue than the latter (Cebrian, 2006; Escudero, 2001) and their discrimination of /a:/ and /ɑ/ becomes much worse if this cue is removed. The Dutch subjects (at least most of them) were still able to tell whether a possible answer was /a:/ or /ɑ/ when the durational cue was removed, making it easier for them to compare a stimulus to these answers. This difference between Dutch subjects and Spanish-speaking ones was also found by Escudero et al., (2009). In the present study, the use of possible answers that were matched on duration in combination with enhanced tokens that had the same duration was very successful in making subjects pay more attention to spectral cues, especially for /a:/-stimuli. The enhanced training also made subjects pay less attention to F3: on the pretest, this formant frequency still played a significant role, whereas on the posttest this was no longer the case. Contrarily to what was found by Halberstam & Lawrence (2004), F0 did not play a significant role.

4.3 Conclusions

On the basis of the findings of the present study, we can conclude the following:

- Both native speakers of Dutch and Spanish-speaking learners of this language experience a gender effect when they have to compare isolated tokens of /a:/ and /ɑ/ produced by male and female speakers to synthesised tokens of these vowels produced by a male computer voice in an XAB-task. This gender effect makes subjects more likely to answer /a:/ when the speaker is female and to answer /ɑ/ when the speaker is male. However, the effect is more persistent in nonnative speakers than in native ones, given the fact that it was larger on the posttest for the Spanish-speaking subjects than for the Dutch ones, irrespective of the condition. Unfortunately, the results of the logistic regression analyses "per subject" indicated that subjects differed significantly in their behaviour. This makes generalization of the results to different groups of subjects problematic.
- The training with the enhanced tokens of /a:/ and /ɑ/ improves subjects' categorization of /a:/ and makes them pay more attention to the spectral cues F1 and F2 and probably also to duration. The training did not have a large effect upon subjects' categorization of /ɑ/, probably due to the fact that some of the enhanced tokens of /ɑ/ sounded like /o/ to various subjects.
- Training with enhanced tokens of /a:/ and /ɑ/ produced by a male computer voice does not affect the gender effect. Rather, this effect seems to be caused by the vowel prototypes used by the subjects and is more persistent in nonnative speakers: it is still unknown whether this is caused by "incomplete" or "defective" prototypes or by a lesser degree of flexibility on behalf of the nonnative speakers. It may even be the case that different neural networks are involved in nonnative speakers than in native ones (e.g. Minagawa-Kawai et al., 2004).
- For native speakers, the XAB-task may be very demanding, especially when they are not familiar with such a task. If the memory demands of the task are too high, this has a negative impact on the results. It is unknown to what extent the Spanish-speaking subjects were affected by the memory demands of the task, because in general, they perform rather poorly on the /a:/-/ɑ/ distinction.

4.4 Suggestions for further research

Based upon our experiences when conducting the present study, we have the following suggestions for further research:

- The study could be replicated with subjects that have more similar backgrounds and proficiency levels in both English and Dutch.
- Subjects' reaction times could be measured. In this way, it will be possible to indicate whether subjects have more difficulties with the female stimuli than with the male ones (longer reaction times) and to exclude answers that were given after extremely long reaction times: these reaction times indicate that the subject was distracted or really did not know the correct answer and "just" clicked on one of the boxes on the computer screen. It should be taken into account, however, that the average reaction time may differ per stimulus category, as was shown by Roberts et al. (2004).
- Given the fact that every stimulus was presented twice to a subject during each XAB-task and that we saw practice effects on the posttest for the subjects in the music condition, it would be interesting to look at practice effects upon the second presentation of the stimulus during an XAB-task.
- It is important to let an additional group of subjects of various backgrounds judge the quality of the enhanced tokens and the possible answers to see whether they are indeed identified as belonging to the same vowel category as the one intended by the experimenter.
- The role of the L1 of the Spanish-speaking subjects should be investigated in more detail.
- The study could be replicated with different vowel contrasts.
- The gender of the listener could also be taken into account.

References

Adank, P.

2003 **Vowel Normalization: a perceptual-acoustic study of Dutch vowels.** [online] Online Resource [April 1st 2009]

Adank, P., Van Hout, R. & Smits, R.

2004 An acoustic description of the vowels of Northern and Southern Standard Dutch. **The Journal of the Acoustical Society of America**, 116, 3 (September 2004), 1729-1738.

Amin, K.

2003 The Effect of Speaker's and Listener's gender on L2 Listening Comprehension. **Indian Journal of Applied Linguistics**, 29, 1, 99-108.

Benders, T. & Escudero, P. R.

Perceptual cue-weighting and stimulus range effect in vowel categorization. In preparation. http://fon.hum.uva.nl/paola/

Boersma, P. & Weenink, D. J. M.

2009 **Praat, doing phonetics by computer.** Version 5.1.08 [online] http://www.praat.org [June 2009]

Cebrian, J.

2006 Experience and the use of non-native duration in L2 vowel categorization. **Journal of Phonetics**, 34, 3, 372-387.

Cucchiarini, C.

1993 "Sources of variability in segmental transcription" **Phonetic transcription: a methodological and empirical study.** S.I.: s.n., 47-62.

Delattre, P., Liberman, A. M., Cooper, F. S. & Gerstman, L. J.

1952 An experimental study of the acoustic determinants of vowel color. Observations on one-and two-formant vowels synthesized from spectographic patterns. **Word**, 8, 195-210.

Eimas, P. D.

1975 Auditory and Phonetic Coding of the Cues for Speech: Discrimination of the [r-l] Distinction by Young Infants. **Perception & Psychophysics**, 18, 5, 341-347.

Escudero, P. R.

2001 The Role of the Input in the Development of L1 and L2 Sound Contrasts: Language-Specific Cue Weighting for Vowels. **Proceedings of the Annual Boston University Conference on Language Development**, 25, 250-261.

Escudero, P. R.

A longitudinal study of how vowel sounds can either facilitate or impede the acquisition of a third language by immigrant communities. In preparation. http://fon.hum.uva.nl/paola/

Escudero, P. R. & Boersma, P.

2002 The Subset Problem in L2 Perceptual Development: Multiple-Category Assimilation by Dutch Learners of Spanish. **Proceedings of the Annual Boston University Conference on Language Development**, 25, 1, 208 - 219. Somerville, Mass.: Cascadilla Press.

Escudero, P. R., Duinmeijer, I., Adank, P. & Van den Velde, H.

2009 **Predicting and explaining problems in L2 vowel perception: The case of Spanish learners of Dutch**. Paper presented at the 7th International Symposium on Bilingualism, Utrecht.

Escudero, P. R., Wanrooij, K. & Clason, K.

The comparative effect of phonology and orthography on L2 vowel perception. In preparation. http://fon.hum.uva.nl/paola/

Field, A.

2009 "Logistic regression" **Discovering Statistics using SPSS (and sex and drugs and rock 'n' roll)** (third edition). London: Sage Publications Ltd: 264-315.

Garner, W.

1974 **The processing of information and structure.** New York: Halsted Press.

Grieser, D. A. & Kuhl, P. K.

1989 Categorization of Speech by Infants: Support for Speech-Sound Prototypes. **Developmental Psychology**, 25, 4, 577-588.

Halberstam, B. & Lawrence, J. R.

2004 Vowel normalization: the role of fundamental frequency and upper formants. **Journal of Phonetics**, 32, 423-434.

Hyltenstam, K. & Abrahamsson, N.

2003 "Maturational constraints in second language acquisition" In: C. Doughty & M.H. Long (Eds.), **Handbook of second language acquisition**. Oxford: Blackwell: 539-588.

Johnson, K., Strand, E. A. & D'Imperio, M.

1999 Auditory-visual integration of talker gender in vowel perception. **Journal of Phonetics**, 27, 359-384.

Kubaska, C. A. & Aslin, R. N.

1985 Categorization and normalization of vowels by 3-year-old children. **Perception & Psychophysics**, 37, 4, 355-362.

Kuhl, P. K.

1983 Perception of Auditory Equivalence Classes for Speech in Early Infancy. **Infant Behavior and Development**, 6, 263-285.

Kuhl, P. K.

1991 Human adults and human infants show a “perceptual magnet effect” for the prototypes of speech categories, monkeys do not. **Perception & Psychophysics**, 50, 2, 93-107.

Ladefoged, P.

2005a “Charting Vowels” **Vowels and Consonants**. **An Introduction to the Sounds of Languages** (2nd edition). Oxford, UK: Blackwell Publishing, 40-48.

Ladefoged, P.

2005b “Vowels Around the World” **Vowels and Consonants**. **An Introduction to the Sounds of Languages** (2nd edition). Oxford, UK: Blackwell Publishing, 154-184.

Lively, S. E. & Pisoni, D. B.

1997 On Prototypes and Phonetic Categories: A Critical Assessment of the Perceptual Magnet Effect in Speech Perception. **Journal of Experimental Psychology: Human Perception and Performance**, 23, 6, 1665-1679.

Long, M. H.

1990 Maturational constraints on language development. **Studies in Second Language Acquisition**, 12, 251-285.

Marean, G. C., Werner, L. A. & Kuhl, P. K.

1992 Vowel Categorization by Very Young Infants. **Developmental Psychology**, 28, 3, 396-405.

Minagawa-Kawai, Y., Mori, K., Sato, Y. & Koizumi, T.

2004 Differential cortical responses in second language learners to different vowel contrasts. **Cognitive neuroscience and neuropsychology**, 15, 5, 899-903.

Mitterer, H.

2006 Is Vowel Normalization Independent of Lexical Processing? **Phonetica**, 63, 209–229.

Pisoni, D. B. & Lazarus, J. H.

1974 Categorical and noncategorical modes of speech perception along the voicing continuum. **Journal of the Acoustical Society of America**, 55, 328-333.

Polivanov, D. E.

1931 “The Subjective Nature of the Perceptions of Language Sounds” In: A.A. Leont’ev & D. Armstrong (Eds.), **Selected works: articles on general linguistics** (Original title: “ÇIzbrannye raboty : stat'i po obésécemu jazykoznanijaÎ” [1968]. English translation by D. Armstrong). The Hague: Mouton [1974], 223-237.

Polka, L. & Werker, J. F.

1994 Developmental Changes in Perception of Nonnative Vowel Contrasts. **Journal of Experimental Psychology: Human Perception and Performance**, 20, 2, 421-435.

Pols, L. C. W., Tromp, H .R. C. & Plomp, R.

1973 Frequency analysis of Dutch vowels from 50 male speakers. **The Journal of the Acoustical Society of America**, 53, 4, 1093-1101.

Roberts, T. P. L., Flagg, E .J. & Gage, N. M.

2004 Vowel categorization induces departure of M100 latency from acoustic prediction. **Cognitive neuroscience and neuropsychology**, 15, 10, 1679-1682.

Roberts, T. P. L. & Poeppel, D.

1996 Latency of auditory evoked M100 as a function of tone frequency. **Neuroreport**, 7, 1138-1140.

Swain, I.U., Zelazo, P. R. & Clifton, R. K.

1993 Newborn Infants' Memory for Speech Sounds Retained Over 24 Hours. **Developmental Psychology**, 29, 2, 312-323.

Van Bergen, D. R., Pols, L. C. W. & Koopmans-Van Beimum, F. J.

1987 Perceptual Normalization of the Vowels of a Man and a Child in Various Contexts. **Speech Communication**, 7, 1, 1-20.

Werker, J. F. & Lalonde, C. E.

1988 Cross-language speech perception: Initial capabilities and developmental change. **Developmental Psychology**, 24, 672-683.

Werker, J. F. & Logan, J. S.

1985 Cross-language evidence for three factors in speech perception. **Perception & Psychophysics**, 37, 1, 35-44.

Werker, J. F. & Polka, L.

1993 Developmental changes in speech perception: new challenges and new directions. **Journal of Phonetics**, 21, 83-101.

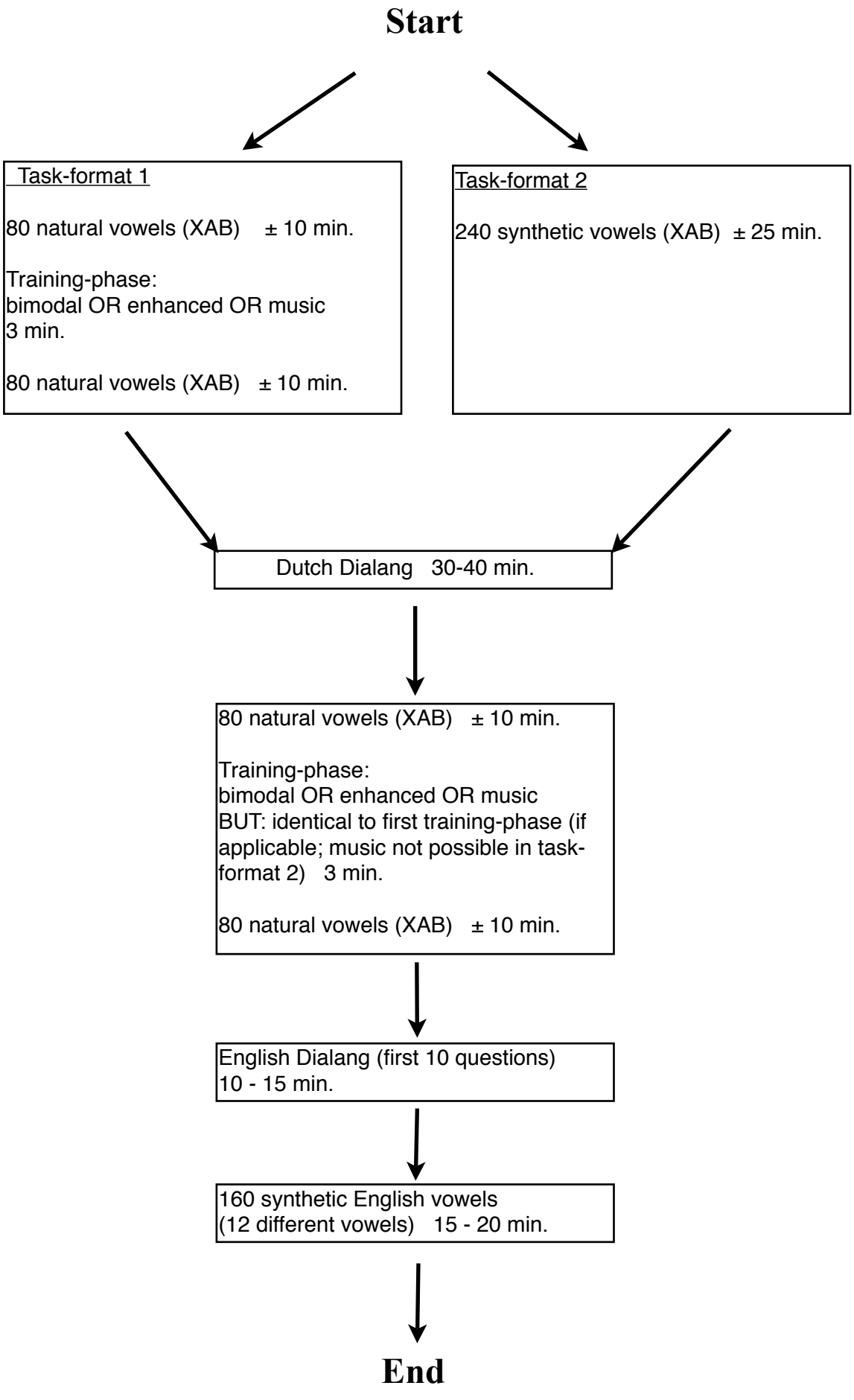
Werker, J. F. & Tees, R. C.

1984 Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. **Infant Behavior and Development**, 7, 49-63.

Appendices

The data that have been included in the appendices are the following:

Appendix 1: Overview of the third session



Appendix 2: Language background questionnaire

Vragenlijst Perceptie-onderzoek

Paola Escudero, Irene ter Avest

Datum: _____

Naam: _____

Telefoonnummer: _____

E-mail: _____

Adres: _____

Leeftijd: _____ Geboortedatum en geboorteplaats: _____

Moedertaal: _____

Beroep: _____

Als je aan de universiteit studeert, in welk jaar/semester zit je nu?:_____

Universiteit: _____Faculteit: _____Hoofdvak(ken): _____

1) Uit welke streek ben je afkomstig?
   Provincie:_____
   Stad / dorp / gemeente: _____

2) In welke streek ging je naar school?
   Lagere school: Provincie:_____ Stad/dorp/gemeente:_____
   Middelbare school: Provincie:_____ Stad/dorp/gemeente:_____

3) Heb je in het Nederlands een accent van een bepaalde streek? Zo ja, welke streek?
   _____

4) Noem steden en landen waar je langer dan twee weken bent geweest sinds je geboren bent.
   Stad en land: _____, Duur van het verblijf: _____

   Stad en land: _____, Duur van het verblijf: _____

   Stad en land: _____, Duur van het verblijf: _____

   Stad en land: _____, Duur van het verblijf: _____

   Stad en land: _____, Duur van het verblijf: _____

5) Waar zijn je ouders geboren? Noem de stad en het land.

   a) Moeder: _____ b) Vader: _____

6) Spreek je, behalve je moedertaal, ook nog andere talen? _____
   Noem welke taal of talen: _____

7) Leer je op dit moment een andere taal of andere talen? _____
Noem de taal of talen en het niveau (bijvoorbeeld: beperkt – matig - goed – heel goed):
Taal: _____, Niveau: _____
Taal: _____, Niveau: _____
Taal:_____, Niveau: _____
Taal:_____, Niveau: _____

8) Waar krijg je les in die taal of talen? (school, taalinstituut, privéles, enz.)
Taal: _____, Plaats: _____
Taal: _____, Plaats: _____
Taal: _____, Plaats: _____
Taal:_____, Niveau: _____

9) Hoeveel uur per week krijg je les?
Taal: _____, Uren per week: _____
Taal: _____, Uren per week: _____
Taal: _____, Uren per week: _____
Taal: _____, Uren per week: _____

10) Hebt je eerder al een andere taal of andere talen geleerd? _____
Noem welke taal of talen: _____

11) Hoe oud was je toen je een andere taal of andere talen begon te leren?
Taal: _____, Leeftijd: _____
Taal: _____, Leeftijd: _____
Taal: _____, Leeftijd: _____
Taal: _____, Leeftijd: _____

12) Waar heb je de andere taal/talen geleerd? (bijvoorbeeld: school, taalinstituut, privéles)
Taal: _____, Plaats: _____
Taal: _____, Plaats: _____
Taal: _____, Plaats: _____
Taal: _____, Plaats: _____

13) Hoeveel uur per week kreeg je les?
Taal: _____, Uur per week: _____
Taal: _____, Uur per week: _____
Taal: _____, Uur per week: _____
Taal: _____, Uur per week: _____

14) Hoe lang leerde je de andere taal/talen?
Taal: _____, Weken/Maanden/Jaren: _____
Taal: _____, Weken/Maanden/Jaren: _____
Taal: _____, Weken/Maanden/Jaren: _____
Taal: _____, Weken/Maanden/Jaren: _____

15) Als je de taal bij een taalinstituut leerde, welk niveau heb je bereikt? _____

16) Omcirkel het nummer dat overeenkomt met de mate waarin je de taal/talen die je hebt geleerd **begrijpt**. (0 betekent dat je niets begrijpt; 7 betekent dat je absoluut alles begrijpt)

Taal: _____,	0	1	2	3	4	5	6	7
Taal: _____,	0	1	2	3	4	5	6	7
Taal: _____,	0	1	2	3	4	5	6	7
Taal: _____,	0	1	2	3	4	5	6	7

17) Omcirkel het nummer dat overeenkomt met de mate waarin je de taal/talen die je hebt geleerd **spreekt**. (0 betekent dat je geen woord spreekt; 7 betekent dat je de taal/talen perfect spreekt, bijna als een moedertaalspreker):

Taal: _____,	0	1	2	3	4	5	6	7
Taal: _____,	0	1	2	3	4	5	6	7
Taal: _____,	0	1	2	3	4	5	6	7
Taal: _____,	0	1	2	3	4	5	6	7

18) Spreek je de taal/talen met andere mensen dan die van je klas? _____
Welke relatie heb je tot deze persoon (bijvoorbeeld: vriend, tante, broer, zus, enz.)
Taal: _____, Persoon: _____
Taal: _____, Persoon: _____
Taal: _____, Persoon: _____
Taal: _____, Persoon: _____

19) Hoeveel uur per week spreek je in de andere taal behalve de lesuren?
Taal: _____, Uur/minuut per week: _____
Taal: _____, Uur/minuut per week: _____
Taal: _____, Uur/minuut per week: _____
Taal: _____, Uur/minuut per week: _____

20) Kijk je televisie in de andere taal/talen? _____
Welke taal/talen? _____

21) Hoeveel uur per week kijk je televisie in de andere taal/talen?
Taal: _____, Uur per week: _____
Taal: _____, Uur per week: _____
Taal: _____, Uur per week: _____
Taal: _____, Uur per week: _____

21) Luister je naar de radio of naar muziek in de andere taal/talen? _____
Welke taal/talen? _____

22) Hoeveel uur per week luister je naar de radio of muziek in de andere taal/talen?
Taal: _____, Uur per week: _____
Taal: _____, Uur per week: _____
Taal: _____, Uur per week: _____
Taal: _____, Uur per week: _____

23) Welke variëteit van het Engels beïnvloedt je, denk je, het meest? (vb. Brits Engels/Amerikaans Engels/Australisch Engels/geen voorkeur/....)

_____

24) Als je Engels spreekt, welke variëteit streef je dan na? (vb. Brits Engels/Amerikaans Engels/ Australisch Engels/geen voorkeur/....)

_____

De gegevens die in deze vragenlijst en in het experiment verzameld werden worden anoniem verwerkt en uitsluitend gebruikt voor wetenschappelijke doeleinden.

Ik verklaar hierbij dat de verzamelde data gebruikt mogen worden voor academische doeleinden.

Datum: Handtekening:

_____ _____

Appendix 3: Confusion matrices

Confusion matrix pretest, Spanish-speaking subjects, enhanced condition: general

Recognised → Offered ↓	/ɑ/	/a:/	Totals
/ɑ/	1220	780	2000
/a:/	803	1197	2000
Totals	2023	1977	4000

Confusion matrix pretest, Spanish-speaking subjects, enhanced condition: female stimuli

Recognised → Offered ↓	/ɑ/	/a:/	Totals
female /ɑ/	567	433	1000
female /a:/	360	640	1000
Totals	927	1073	2000

Confusion matrix pretest, Spanish-speaking subjects, enhanced condition: male stimuli

Recognised → Offered ↓	/ɑ/	/a:/	Totals
male /ɑ/	653	347	1000
male /a:/	443	557	1000
Totals	1096	904	2000

Confusion matrix posttest, Spanish-speaking subjects, enhanced condition: general

Recognised → Offered ↓	/ɑ/	/a:/	Totals
/ɑ/	1270	730	2000
/a:/	588	1412	2000
Totals	1858	2142	4000

Confusion matrix posttest, Spanish-speaking subjects, enhanced condition: female stimuli

Recognised → Offered ↓	/ɑ/	/a:/	Totals
female /ɑ/	588	412	1000
female /a:/	239	761	1000
Totals	827	1173	2000

Confusion matrix posttest, Spanish-speaking subjects, enhanced condition: male stimuli

Recognised → Offered ↓	/ɑ/	/a:/	Totals
male /ɑ/	682	318	1000
male /a:/	349	651	1000
Totals	1031	969	2000

Confusion matrix pretest, Spanish-speaking subjects, music condition: general

Recognised → Offered ↓	/ɑ/	/a:/	Totals
/ɑ/	1199	801	2000
/a:/	730	1270	2000
Totals	1929	2071	4000

Confusion matrix pretest, Spanish-speaking subjects, music condition: female stimuli

Recognised → Offered ↓	/ɑ/	/a:/	Totals
female /ɑ/	577	423	1000
female /a:/	292	708	1000
Totals	869	1131	2000

Confusion matrix pretest, Spanish-speaking subjects, music condition: male stimuli

Recognised → Offered ↓	/ɑ/	/a:/	Totals
male /ɑ/	622	378	1000
male /a:/	438	562	1000
Totals	1060	940	2000

Confusion matrix posttest, Spanish-speaking subjects, music condition: general

Recognised → Offered ↓	/ɑ/	/a:/	Totals
/ɑ/	1253	747	2000
/a:/	704	1296	2000
Totals	1957	2043	4000

Confusion matrix posttest, Spanish-speaking subjects, music condition: female stimuli

Recognised → Offered ↓	/ɑ/	/a:/	Totals
female /ɑ/	603	397	1000
female /a:/	305	695	1000
Totals	908	1092	2000

Confusion matrix posttest, Spanish-speaking subjects, music condition: male stimuli

Recognised → Offered ↓	/ɑ/	/a:/	Totals
male /ɑ/	650	350	1000
male /a:/	399	601	1000
Totals	1049	951	2000

Confusion matrix pretest, Dutch subjects, enhanced condition: general

Recognised → Offered ↓	/ɑ/	/a:/	Totals
/ɑ/	748	132	880
/a:/	127	753	880
Totals	875	885	1760

Confusion matrix pretest, Dutch subjects, enhanced condition: female stimuli

Recognised → Offered ↓	/ɑ/	/a:/	Totals
female /ɑ/	372	68	440
female /a:/	41	399	440
Totals	413	467	880

Confusion matrix pretest, Dutch subjects, enhanced condition: male stimuli

Recognised → Offered ↓	/ɑ/	/a:/	Totals
male /ɑ/	376	64	440
male /a:/	86	354	440
Totals	462	418	880

Confusion matrix posttest, Dutch subjects, enhanced condition: general

Recognised → Offered ↓	/ɑ/	/a:/	Totals
/ɑ/	746	134	880
/a:/	48	832	880
Totals	794	966	1760

Confusion matrix posttest, Dutch subjects, enhanced condition: female stimuli

Recognised → Offered ↓	/ɑ/	/a:/	Totals
female /ɑ/	374	66	440
female /a:/	10	430	440
Totals	384	496	880

Confusion matrix posttest, Dutch subjects, enhanced condition: male stimuli

Recognised → Offered ↓	/ɑ/	/a:/	Totals
male /ɑ/	372	68	440
male /a:/	38	402	440
Totals	410	470	880

Confusion matrix pretest, Dutch subjects, music condition: general

Recognised → Offered ↓	/ɑ/	/a:/	Totals
/ɑ/	474	86	560
/a:/	112	448	560
Totals	586	534	1120

Confusion matrix pretest, Dutch subjects, music condition: female stimuli

Recognised → Offered ↓	/ɑ/	/a:/	Totals
female /ɑ/	236	44	280
female /a:/	28	252	280
Totals	264	296	560

Confusion matrix pretest, Dutch subjects, music condition: male stimuli

Recognised → Offered ↓	/ɑ/	/a:/	Totals
male /ɑ/	238	42	280
male /a:/	84	196	280
Totals	322	238	560

Confusion matrix posttest, Dutch subjects, music condition: general

Recognised → Offered ↓	/ɑ/	/a:/	Totals
/ɑ/	487	73	560
/a:/	59	501	560
Totals	546	574	1120

Confusion matrix posttest, Dutch subjects, music condition: female stimuli

Recognised → Offered ↓	/ɑ/	/a:/	Totals
female /ɑ/	246	34	280
female /a:/	12	268	280
Totals	258	302	560

Confusion matrix posttest, Dutch subjects, music condition: male stimuli

Recognised→ Offered ↓	/ɑ/	/a:/	Totals
male /ɑ/	241	39	280
male /a:/	47	233	280
Totals	288	272	560

Appendix 4: Script for the logistic regression analysis in Praat, the discriminant analysis, the vocal tract normalization and the computation of the correlation-coefficients

```
# Script_log_dis_voc_and_cors.praat

tab_tmp = Read Table from comma-separated file... ###.csv
tortmp = Down to TableOfReal... Vowel
# logduration logf0 log(f1...f3) aa ah
tor = Extract column ranges... 3 6 8 10 12 14 65 66
tab = To Table... Vowel

call fractions_correct tor

call vtl tor
tor_vtl = selected ("TableOfReal")
call fractions_correct tor_vtl

call correlate tor

procedure tmp
  select tab
  To logistic regression... "Gender Vowel" aa ah
  Rename... gender_vowel

  select tab
  To logistic regression...  "logduration logf0 logf1 logf2 logf3" aa ah
  Rename...  log_data
endproc

procedure get_fraction_correct .tor .column_range$
  select .tor
  .te = Extract column ranges... '.column_range$'
  .dis = To Discriminant
  plus .te
  .ct = To ClassificationTable... yes yes
  .cf = To Confusion
  .fc = Get fraction correct
  printline   '.fc' (Fraction correct for range "'.column_range$'")
  select .cf
  plus .ct
  plus .dis
  plus .te
  Remove
endproc

procedure fractions_correct .tor
  printline all
  call get_fraction_correct .tor 1:7

  printline duration log f0 logf1 logf2 logf3
  call get_fraction_correct .tor 1:6

  printline logduration logf0
  call get_fraction_correct .tor 2 3

  printline logduration  logf1
  call get_fraction_correct .tor 2 4
```

```
printline logduration logf2
call get_fraction_correct .tor 2 5

printline logduration logf3
call get_fraction_correct .tor 2 6

printline logduration logf0 logf1
call get_fraction_correct .tor 2 3 4

printline logduration logf0 logf2
call get_fraction_correct .tor 2 3 5

printline logduration logf0 logf3
call get_fraction_correct .tor 2 3 6

printline logduration logf1 logf2
call get_fraction_correct .tor 2 4 5

printline logduration logf1 logf3
call get_fraction_correct .tor 2 4 6

printline logduration logf2 logf3
call get_fraction_correct .tor 2 5 6

printline alleen logduration
call get_fraction_correct .tor 2

printline alleen logf0
call get_fraction_correct .tor 3

printline alleen logf1
call get_fraction_correct .tor 4

printline alleen logf2
call get_fraction_correct .tor 5

printline alleen logf3
call get_fraction_correct .tor 6

printline logf0 logf1  logf2 logf3
call get_fraction_correct .tor 3:6

printline logf1  logf2 logf3
call get_fraction_correct .tor 4:6

printline logf1  logf2
call get_fraction_correct .tor 4 5

printline logf1  logf3
call get_fraction_correct .tor 4 6

printline logf2 logf3
call get_fraction_correct .tor 5 6

printline logf0 logf1
call get_fraction_correct .tor 3 4

printline logf0 logf2
call get_fraction_correct .tor 3 5
```

```
  printline logf0 logf3
  call get_fraction_correct .tor 3 6

endproc

call tmp

procedure vtl .tor
 select .tor
 .vtl = Copy... vtl
 Formula... if self[1]=1 then if col>=4 and col<=6 then self+log10(15/17) else self fi else
self fi
endproc

procedure correlate .tor
 select .tor
 nrows = Get number of rows
 Insert column (index)... 9
 Set column label (index)... 9 daa
 Insert column (index)... 10
 Set column label (index)... 10 dah
 f1aa =770
 f2aa =1303
 f3aa =2477
 f1ah =687
 f2ah=1104
 f3ah =2490

 Formula... if col= 9 then sqrt((self[4]-log10(f1aa))^2+(self[5]-log10(f2aa))^2+(self[6]-
log10(f3aa))^2) else self fi
 Formula... if col=10 then sqrt((self[4]-log10(f1ah))^2+(self[5]-log10(f2ah))^2+(self[6]-
log10(f3ah))^2) else self fi
 To Correlation

 select .tor
 .ftor = Extract rows where column... 1 "equal to" 1
 Rename... cf
 To Correlation
  select .tor
 .ftor = Extract rows where column... 1 "equal to" 2
 Rename... cm
 To Correlation

endproc
```

Appendix 5: Results of the multicollinearity checks run in SPSS for the logistic regression analyses

Multicollinearity check pretest

Coefficients[a]

Model		Collinearity Statistics	
		Tolerance	VIF
1	language	1.000	1.000
	logdur.	0.392	2.55
	logf1	0.43	2.323
	logf2	0.408	2.449
	logf0	0.381	2.622
	logf3	0.454	2.202
	gender	0.277	3.605

a. Dependent Variable: aa

Collinearity Diagnostics[a]

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions							
				(Constant)	lang.	logdur.	logf1	logf2	logf0	logf3	gender
1	1	7.279	1.000	.00	.00	.00	.00	.00	.00	.00	.00
	2	0.478	3.902	.00	.01	.00	.00	.00	.00	.00	.27
	3	0.234	5.576	.00	.98	.00	.00	.00	.00	.00	.00
	4	0.007	31.823	.00	.00	.32	.00	.00	.03	.00	.01
	5	0.001	98.722	.01	.00	.35	.03	.04	.76	.01	.12
	6	.000	124.918	.01	.00	.07	.92	.03	.00	.02	.12
	7	.000	241.585	.03	.00	.23	.00	.81	.20	.28	.01
	8	4.600E-05	397.802	.96	.00	.04	.04	.13	.01	.69	.47

a. Dependent Variable: aa

Multicollinearity check posttest

Coefficients[a]

Model		Collinearity Statistics	
		Tolerance	VIF
1	language	.990	1.010
	condition	.990	1.010
	logdur.	.392	2.550
	logf1	.430	2.323
	logf2	.408	2.449
	logf0	.381	2.622
	logf3	.454	2.202
	gender	.277	3.605

a. Dependent Variable: aa

Collinearity Diagnostics[a]

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions								
				(Constant)	lang.	cond.	logdur.	logf1	logf2	logf0	logf3	gender
1	1	7.832	1.000	.00	.00	.00	.00	.00	.00	.00	.00	.00
	2	.492	3.990	.00	.00	.25	.00	.00	.00	.00	.00	.20
	3	.446	4.189	.00	.07	.63	.00	.00	.00	.00	.00	.06
	4	.221	5.954	.00	.92	.11	.00	.00	.00	.00	.00	.00
	5	.007	33.015	.00	.00	.00	.32	.00	.00	.03	.00	.01
	6	.001	102.408	.01	.00	.00	.35	.03	.04	.76	.01	.12
	7	.000	129.579	.01	.00	.00	.07	.92	.03	.00	.02	.12
	8	.000	250.595	.03	.00	.00	.23	.00	.81	.20	.28	.01
	9	4.600E-05	412.646	.96	.00	.00	.04	.04	.13	.01	.69	.47

a. Dependent Variable: aa

Appendix 6: Script for the F1/F2 matrices from section 3.5.1

```
# Script F1/F2.praat

file$ = "###.csv"
call get_table .file$
tab = selected ("Table")

# gender = 1 pretest
call get_matrix tab 1
mat1 = selected ("Matrix")

# gender = 2 pretest
call get_matrix tab 2
mat2 = selected ("Matrix")

file2$ = "###2.csv"
call get_2table .file2$
tab2 = selected ("Table")

# gender = 1 posttest
call get_matrix tab2 1
mat3 = selected ("Matrix")

# gender = 2 posttest
call get_matrix tab2 2
mat4 = selected ("Matrix")


# difference female pre- and posttest
   select mat1
   mat5 = Copy... mat5
   Formula... Object_'mat3'[]-Object_'mat1'[]


# difference male pre- and posttest
   select mat2
   mat6 = Copy... mat6
   Formula... Object_'mat4'[]-Object_'mat2'[]


procedure get_table .file$
  .tab_tmp = Read Table from comma-separated file... 'file$'
  .tortmp = Down to TableOfReal... Vowel
  # logduration logf0 log(f1...f3) aa ah
  .tor = Extract column ranges... 3 6 8 10 12 14 65 66
  .tab = To Table... Vowel
  select .tab_tmp
  plus .tortmp
  plus .tor
  Remove
  select .tab
endproc

procedure get_2table .file2$
  .tab2_tmp = Read Table from comma-separated file... 'file2$'
  .tor2tmp = Down to TableOfReal... Vowel
  # logduration logf0 log(f1...f3) aa ah
  .tor2 = Extract column ranges... 3 6 8 10 12 14 65 66
  .tab2 = To Table... Vowel
```

```
  select .tab2_tmp
  plus .tor2tmp
  plus .tor2
  Remove
  select .tab2
endproc

procedure get_matrix .tab .gender
  select .tab
  .nrows = Get number of rows
  .xmin = 400
  .xmax = 1100
  .nx = 70
  .dx = (.xmax-.xmin)/.nx
  .x1 = .xmin+.dx/2
  .ymin = 900
  .ymax = 1800
  .ny = 90
  .dy = (.ymax-.ymin)/.ny
  .y1 = .ymin+.dy/2
  .mat = Create Matrix... mat_g'.gender' .xmin .xmax .nx .dx .x1 .ymin .ymax .ny .dy .y1  50
  for .irow to .nrows
    if Object_'.tab'[.irow, "Gender"] = .gender
      .f1 = 10^(Object_'.tab'[.irow, "logf1"])
      .f2 = 10^(Object_'.tab'[.irow, "logf2"])
      .naa = Object_'.tab'[.irow, "aa"]
      printline '.f1'  '.f2' '.naa'
      .ix = floor((.f1 - .xmin ) / .dx + 1)
      .iy = floor((.f2 - .ymin ) / .dy +1)
      Set value... .iy .ix .naa
    endif
  endfor
endproc

procedure get_matrix .tab2 .gender
  select .tab2
  .nrows = Get number of rows
  .xmin = 400
  .xmax = 1100
  .nx = 70
  .dx = (.xmax-.xmin)/.nx
  .x1 = .xmin+.dx/2
  .ymin = 900
  .ymax = 1800
  .ny = 90
  .dy = (.ymax-.ymin)/.ny
  .y1 = .ymin+.dy/2
  .mat = Create Matrix... mat_g'.gender' .xmin .xmax .nx .dx .x1 .ymin .ymax .ny .dy .y1  50
  for .irow to .nrows
    if Object_'.tab'[.irow, "Gender"] = .gender
      .f1 = 10^(Object_'.tab'[.irow, "logf1"])
      .f2 = 10^(Object_'.tab'[.irow, "logf2"])
      .naa = Object_'.tab'[.irow, "aa"]
      printline '.f1'  '.f2' '.naa'
      .ix = floor((.f1 - .xmin ) / .dx + 1)
      .iy = floor((.f2 - .ymin ) / .dy +1)
      Set value... .iy .ix .naa
    endif
  endfor
endproc
```

Appendix 7: Script for making the scatterplots for the relationship between duration and percentage of /a:/'s answered

```
#Script scatterplot duration

Read Table from comma-separated file... ###.csv
Select outer viewport... 0 7.5 4 8
Scatter plot... logduration 0 0 aa 0 100 Stimulus 8 yes
One mark left... 50 yes yes yes
Text left... yes percentageaa

Read Table from comma-separated file... ###2.csv
Select outer viewport... 0 7.5 8 12
Scatter plot... logduration 0 0 aa 0 100 Stimulus 8 yes
One mark left... 50 yes yes yes
Text left... yes percentageaa

select Table ###
Select outer viewport... 0 7.5 0 4
Scatter plot... logf1 0 0 logf2 0 0 Stimulus 8 yes
```