University of Amsterdam
Bachelor's thesis Linguistics
June 27, 2012

Supervisors:

Prof. dr. P.P.G. Boersma
K.E. Wanrooij, MA

Does enhanced bimodal distributional training improve perception of English /æ/ and /ɛ/ for adult native speakers of Dutch?

Johanna de Vos
5755387

TABLE OF CONTENTS

1. PREFACE

This thesis was written to conclude my bachelor's degree in Linguistics at the University of Amsterdam. For the last year and a half of my degree, I worked as a research assistant for Karin Wanrooij in prof. dr. Paul Boersma's NWO[1]-funded Vici project *Emergent Categories and Connections*. Karin offered me the possibility of writing my thesis within this project. Under her supervision, I went through all stages that come with conducting an experiment, from recording and creating the stimuli to writing everything down. Karin always took the time to help me and gave me detailed feedback. My other supervisor prof. dr. Paul Boersma was also involved in the whole process. He challenged me to reflect critically on the literature and helped in shaping the thesis. I would like to thank Paul and Karin very much for their help.

[1] Netherlands Organization for Scientific Research.

2. ABSTRACT

The aim of this thesis was to examine whether enhanced bimodal distributional training improves perception of the Standard Southern British English /æ/-/ɛ/ vowel contrast for adult native speakers of Dutch. Distributional training methods make use of the statistical patterns in speech with regard to phonemic contrasts, here the values of the first and second formant of /æ/ and /ɛ/. Our distribution was an enhanced bimodal one, i.e. with two peaks and with more extreme values as the endpoints than average production values.

All participants performed a pre-test and a post-test in an XAB format. In between the tests, half of the participants were exposed to an enhanced bimodal distribution of the /æ/-/ɛ/ contrast for two minutes, whereas the other half listened to classical music. On average, perception improved for all participants, but the improvement was significantly larger for participants who had listened to music as opposed to an enhanced bimodal distribution.

This outcome was unanticipated. However, prior studies on distributional training had already yielded conflicting results, sometimes finding a positive effect of distributional training, and sometimes finding no effect. With this study showing a negative effect, it seems that the effectiveness of distributional training for aiding listeners with perception of difficult non-native vowel contrasts may be doubted.

3. INTRODUCTION

For adult second language learners it is often difficult to acquire phonemic contrasts that are non-phonemic in their native language. A well-known example is the difficulty Japanese learners experience with distinguishing English /r/ from /l/ (e.g. Aoyama, Flege, Guion, Akahane-Yamada & Yamada, 2004). Escudero and Wanrooij (2010) found that similar difficulties hold true for Spanish adults learning Dutch: they usually perceive the Dutch vowels /a:/ and /ɑ/ as the single Spanish phoneme /a/. In turn, native speakers of Dutch experience considerable difficulties with telling the English vowels /æ/ and /ɛ/ apart, perceiving both as /ɛ/ (Cutler, Weber, Smits & Cooper, 2004, for American English; Escudero, Hayes-Harb & Mitterer, 2008, for Standard Southern British English).

At the same time, babies seem to be able to perceive phonemic contrasts which are not present in their native language (e.g. Trehub, 1976; Werker, Gilbert, Humphrey & Tees, 1981; Werker & Tees, 1984). This indicates that linguistic experience influences sensitivity to phonemic contrasts: the ability to discriminate sounds that are phonemic in the native language increases, while the ability to discriminate sounds which make no difference to meaning decreases. However, it has been shown that adults can be trained to strengthen their sensitivity to non-native phonemic contrasts by means of a procedure called distributional training (Maye & Gerken, 2000, 2001; Gulian, Escudero & Boersma, 2007; Hayes-Harb, 2007; Escudero, Benders & Wanrooij, 2011; Wanrooij, Escudero and Raijmakers, submitted). For more information on distributional training, see section 3.1.

The aim of this thesis is to submit the method of distributional training to testing through an experiment in which adult native speakers of Dutch will be trained to better perceive a difference between the Standard Southern British English vowels /æ/ and /ɛ/. This is in line with various research projects conducted at the University of Amsterdam, which will be elaborated in section 3.2.

3.1 Distributional training

The idea of distributional training is based on the belief that statistical patterns affect learning. Concerning the acquisition of speech sounds, distributional training exploits the statistical patterns in speech with regard to phonemic contrasts. This rests upon the fact that different phonemes have distinct distributions on some acoustic dimensions, which has long been known (e.g. Liberman, 1957). For instance, English /æ/ and /ɛ/ mainly differ from each other in terms of their first and second formant (F_1 and F_2).[2] Although the tokens in these categories are produced with a lot of variation and even overlap, the most frequently heard tokens will fall in two clusters, thereby creating a bimodal distribution (Maye & Gerken, 2000). Figure 1 shows such a distribution after the example of English /æ/ and /ɛ/.

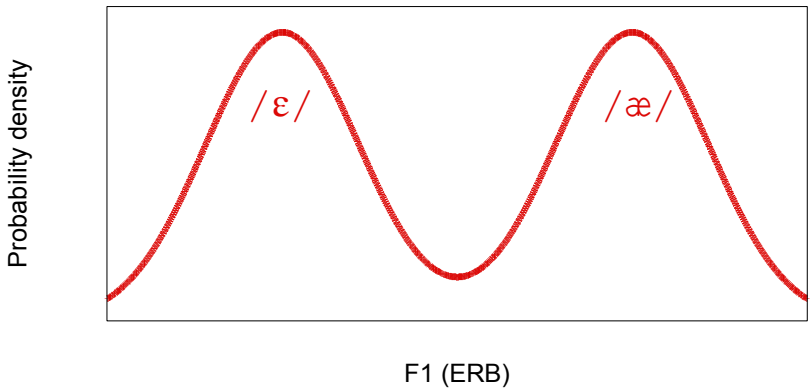

Figure 1. A bimodal distribution of English /æ/ and /ɛ/.

Distributional training methods make use of these statistical patterns by offering listeners a bimodal distribution of speech sounds along the acoustic continuum that comprises the two categories that are to be trained. In the distribution, these are represented as two peaks. Listeners learn to discriminate between these categories because they are most frequently exposed to tokens near the endpoints of the acoustic continuum, and less frequently to tokens in between the two sound categories (Escudero et al., 2011).

Maye & Gerken (2000, 2001) were the first to show that an adult's sensitivity to a speech contrast could be influenced by listening to a bimodal distribution of speech sounds that were not phonemes (only allophones). Similar results have been obtained in studies by Hayes-Harb (2007), Gulian et al. (2007), Escudero et al. (2011), and Wanrooij et al. (submitted), although the speech sounds under investigation and the research designs varied. Participants in these studies either had to indicate whether the sounds that were

[2] Hawkins & Midgley (2005) showed that the average formant frequencies for 35-40 year old male speakers of Received Pronunciation (RP) English are 512Hz (/ɛ/) and 696Hz (/æ/) for F_1, and 1888Hz (/ɛ/) and 1574 Hz (/æ/) for F_2.

played were the same or different (discrimination task), or they had to group the sounds into two different categories (categorisation task).

3.2 Motivation for the current study

Recently, researchers at the University of Amsterdam are among the first to have tried to show an effect of distributional training by means of electroencephalography (EEG). Wanrooij, Van Zuijen & Boersma (2012) trained adult Dutch participants in the English contrast /æ/ - /ɛ/, using four different experimental conditions. The first group of participants was exposed to a bimodal distribution of this contrast, the second group to an enhanced bimodal distribution (see section 4.3.2 for a detailed description of bimodal versus enhanced bimodal distributions), while the third group of participants was exposed to a unimodal distribution of the /æ/ - /ɛ/ contrast (see figure 2). The latter type of distribution consists of only one peak in the middle, thereby grouping instances of /æ/ and /ɛ/ together in one cluster. In this way, the ability to discriminate between both vowels is not facilitated. This leads to the hypothesis that only the bimodal group should improve on the ability to discriminate between /æ/ and /ɛ/. Finally, the fourth group served as a control group and listened to classical music during training.

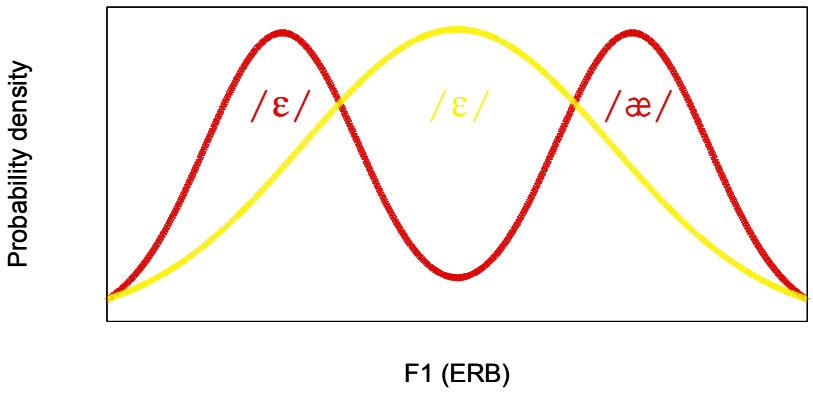

Figure 2. A bimodal distribution of /æ/ and /ɛ/ (in red), as opposed to a unimodal distribution of /ɛ/ (in yellow). The bimodal distribution can be seen as representing English /æ/ and /ɛ/, while the unimodal distribution can be seen as representing Dutch /ɛ/.

Improvement was calculated as the difference in size of mismatch negativity (MMN) between a pre-test and a post-test. MMN is induced by offering a listener a deviant stimulus in a sequence of standard stimuli, and is measured by subtracting the waveform of the event-related potential (ERP) in reaction to the standard stimulus from the waveform of the ERP in reaction to the deviant stimulus. In both pre-test and post-test, some participants listened to

a sequence of instances of /ɛ/ with occasionally a deviant /æ/, while for other participants it was the other way around. MMN should occur each time when participants heard a deviant stimulus to the sequence of stimuli.

The results were surprising: whereas it was expected that participants would show improved discrimination after bimodal training, it was found that the post-test MMN was significantly smaller for all participants combined (but not for any of the groups alone). There were no significant differences between the groups, including the control group. This implies that distributional training does not influence the ability of adult Dutch learners to discriminate between /æ/ and /ɛ/. As discussed, previous research on distributional training has shown positive effects of distributional training, for example Escudero et al.'s (2011) success with training Spanish participants to discriminate between Dutch /a:/ and /ɑ/ on an XAB task[3] using an enhanced bimodal distribution.

The aim of this thesis is to find out what could possibly cause the discrepancy between results as obtained in this EEG experiment and in Escudero et al.'s (2011) XAB task. Therefore, we will use English /æ/ and /ɛ/ stimuli as was done in the EEG experiment, but test Dutch participants in an XAB task according to Escudero et al.'s (2011) design. As a matter of fact, there were a number of differences between the EEG and XAB task test design that could account for the unexpected findings.

During training, participants in the EEG experiment listened to a continuous sequence of vowels (900 different stimuli; participants never heard the same stimulus twice), whereas participants in the XAB task listened to a discontinuous sequence (8 different stimuli; participants heard the same stimulus over and over). Moreover, the length of the training for the EEG experiment was 12 minutes, as opposed to only 2 minutes in the XAB task. Because of this, participants in the EEG experiment could have experienced a greater degree of fatigue or boredom, this influencing the learning process. Prior research (e.g. Lang, Eerola, Korpilahti, Holopainen, Salo & Aaltonen, 2005) has shown that the size of MMN in EEG experiments can decline due to sleepiness.

The testing phase too was subject to differences. The participants in the XAB task were actively involved in testing: they had to click on a computer screen which vowel they had perceived. Participation in the EEG experiment was of a more passive character, because everyone watched a silent film and was told to ignore the stimuli. This could cause participants in the XAB task to learn from the test itself, an effect which would most likely not occur in the EEG experiment because the participants' attention was fixed on the film.

[3] In the cited literature on distributional training, usually the term *categorisation task* is used for what is called *XAB task* in this thesis. We opted for the latter, more literal term because it cannot be shown that actual phonemic categorisation, instead of only a comparison of phonetic differences, takes place in an XAB task test design. This test design will be explained in section 4.2.

Another difference between the EEG experiment and the XAB task was the use of natural and synthetic stimuli in the latter, while the former made use only of synthetic stimuli. The use of both natural and synthetic stimuli makes a test more varied and probably more difficult.

Although we have seen that there are a number of differences between both test designs that could account for the different outcomes, it is also conceivable that these are caused by differences between the Dutch and Spanish vowel systems. The Dutch language has a relatively full vowel space, consisting of fifteen vowels (10-13 monophthongs and 3-6 diphthongs, according to dialect; Booij, 1995). In Spanish, there are only five different vowels (Hammond, 2001). This could influence the ability to acquire new vowel contrasts.

To find out whether this is the case, we will test the effect of Dutch participants being trained by means of distributional training by using Escudero et al.'s (2011) test design. If we now find a positive effect of distributional training, we can exclude the possibility that differences between the Dutch and Spanish vowel systems have led to different outcomes in the EEG experiment as compared to the XAB task. In that case, we will know that either differences in the testing phase or differences in the training phase account for the contradictory findings of Wanrooij et al.'s (2012) EEG experiment and Escudero et al.'s (2011) XAB task.

4. METHODOLOGY

4.1 Participants

A total of 100 participants, 64 females and 36 males, took part in the study, all of them students or recent graduates. Their ages ranged between 18 and 30, with an average age of 22. All participants had Dutch as their native language and were raised monolingually. Some had travelled or lived in a foreign country, but no longer than four weeks in countries where English is the national language. Half of the participants were assigned randomly to the experimental group and the other half to the control group, although gender was controlled for.

4.2 Procedure

The procedure of this experiment is identical to the procedure used by Escudero et al. (2011). Participants performed an XAB task (pre-test), followed by two minutes of training, after which they performed a second XAB task (post-test). As for the training, the experimental group listened to an enhanced bimodal distribution of the Standard Southern British English vowels /æ/ and /ɛ/ (see section 5.3.2), whereas the control group listened to classical piano music by Chopin. From now on, the experimental group will be called *enhanced group* and the control group will be called *music group*.

The pre-test and the post-test were identical two-alternative forced choice tasks in an XAB format. In each test, listeners heard eighty trials containing three vowels with an inter-stimulus interval (ISI) of 1.2s between the end of one sound and the beginning of the next. The ISI was relatively long because this benefits categorisation, whereas discrimination in which only sensory traces are compared decreases with increasing ISI (Van Hessen & Schouten, 1992) (but see footnote 3).

The participants had to indicate on the computer whether the first vowel (X) was more similar to the second vowel (A) or to the third vowel (B). Per test, there were forty trials where the X stimulus was /ɛ/ and forty trials where it was /æ/. The presentation of the A and B stimuli was counterbalanced across trials and trial order was randomized per participant (for both pre-test and post-test).The participants were told to listen to all sounds in a trial before making a decision, and to guess in case they were not sure.

During training, the enhanced group listened to 128 vowel tokens with an ISI of 750ms. They were instructed to listen to the vowels carefully, because they would perform a second XAB task after the training. Participants in the music group were told that they would listen to classical music and could relax, after which they would perform another task similar

to the first one. For both groups, the duration of the training was 1:57 minutes. After finishing the second XAB task, all participants did a listening comprehension test in English (DIALANG, version 0.93.1, Lancaster University). The running time of all tests together was approximately 45 minutes.

4.3 Stimuli

4.3.1 Test stimuli

The eighty X stimuli were naturally produced tokens of English /æ/ and /ɛ/ spoken by six female and five male native speakers of Standard Southern British English. Most recordings were made in the studio of the Language Lab at the University of Amsterdam, except for three of the male speakers, whose recordings had already been made in previous research projects (Escudero et al., 2008; Daniel Williams, University of Sheffield).

Most of the tokens were extracted from a /h-V-d/ context (*head* / *had*) or a /f-V-f/ context (*fef* / *faf*). To add additional variation, some vowels were extracted from other contexts, namely /s-V-s/, /b-V-s/, /h-V-s/, /m-V-s/ and /t-V-s/. Table 1 shows the average fundamental frequency (F_0), first formant (F_1), second formant (F_2) and duration values of /æ/ and /ɛ/ for the female and male speakers separately, per vowel and recording context.

Table 1. Average duration (in ms), F_0, F_1 and F_2 values (in Hz) of the X stimuli.

Vowel	Context	Speaker	Number of tokens	Duration	F_0	F_1	F_2
	fæf	Male	8	127.49	108.80	761.42	1356.84
		Female	11	122.40	202.36	983.14	1625.40
	hæd	Male	3	126.11	114.19	767.34	1526.39
		Female	11	130.48	198.25	949.46	1602.14
/æ/	sæs	Male	1	104.33	115.36	824.30	1479.93
		Female	2	95.71	190.32	925.20	1757.71
	bæs	Male	1	95.46	133.58	782.87	1498.84
	hæs	Male	1	67.30	129.82	784.19	1612.49
	mæs	Male	1	77.78	122.76	766.23	1523.69
	tæn	Male	1	74.01	134.07	720.18	1544.27
	Total /æ/	Male	16	113.57	115.53	766.95	1443.32
		Female	24	123.88	199.47	962.88	1625.76
	fɛf	Male	7	113.77	106.17	612.61	1575.00
		Female	11	125.05	204.95	722.79	1926.60
	hɛd	Male	4	110.18	118.51	582.00	1790.44
		Female	11	114.84	213.93	655.46	2128.67
/ɛ/	sɛs	Male	1	80.93	110.18	535.43	1483.08
		Female	2	99.02	184.53	615.00	1946.07
	bɛs	Male	1	73.24	129.79	559.75	1675.77
	hɛs	Male	1	63.15	132.11	557.52	1738.77
	mɛs	Male	1	62.88	126.54	601.52	1723.82
	tɛn	Male	1	41.09	145.60	515.95	1744.61
	Total /ɛ/	Male	16	97.40	116.34	586.65	1659.55
		Female	24	118.20	200.56	682.95	2020.83

All values were calculated in the Praat program (Boersma & Weenink, 2012, version 5.3.14). The analysis window had a duration of 50ms and covered the area between 25ms before and after half of the total vowel duration. One vowel had a duration of only 41ms; for this vowel the analysis window was the total 41ms.

The first and second formant were calculated from the sound with the Burg method, with automated time steps, a maximum number of formants of 5, a maximum formant in Hz of 5500 for female speakers and 5000 for male speakers, a window length of 0.025 s, and a pre-emphasis of 6 dB per octave from 50 Hz. Then, the mean of these formants was queried for the time range in the analysis window.

The fundamental frequency, measured as the pitch level at which the vowels were pronounced, was extracted within a pitch search range between 75 Hz and 300 Hz for male speakers and between 100 Hz and 500 Hz for female speakers. In the same way as for the F_1 and F_2, the mean pitch was calculated over the time stretch of the analysis window.

The A and B stimuli were synthetic tokens created in Praat (Boersma & Weenink, 2012, version 5.3.14), and were similar to the A and B stimuli used in Wanrooij et al.'s EEG experiment (2012). They only varied in their F_1 and F_2 values; duration and the formants 3 to 10 were kept the same. Table 2 shows the duration, F_0, F_1 and F_2 of the A and B stimuli.

Table 2. Average duration (in ms), F_0, F_1 and F_2 values (in ERB and Hz) of the A and B stimuli.

Vowel	Duration	F_0		F_1		F_2	
		ERB	Hz	ERB	Hz	ERB	Hz
/æ/	140	4.25 to 3.01	150-100	11.99	642.20	19.32	1648.35
/ɛ/	140	4.25 to 3.01	150-100	10.95	552.26	20.04	1797.14

Because the research design of Escudero et al. (2011) was followed as closely as possible, the F_1 and F_2 values of the A and B stimulus were chosen to lie on the intersection of the bimodal and unimodal continua,[4] i.e. at the point where the yellow and red lines in figure 2 overlap. Figure 3 shows where the test stimuli are located on the acoustic F_1 continuum.

[4] Actually, in this experiment no unimodal distribution was used, but if it had been, stimuli A and B had lain at its intersection with the bimodal distribution.

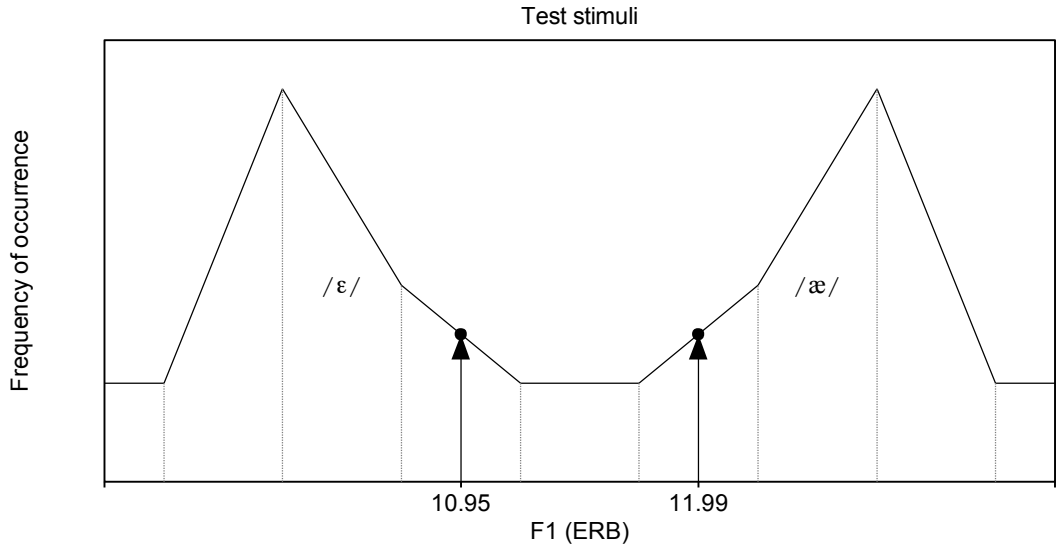

Figure 3. Location of the A and B test stimuli on the acoustic F_1 continuum (in ERB).

4.3.2 Training stimuli

For the training, eight synthetic vowels were used which were generated following the same procedure as for the A and B stimuli. The endpoints (tokens 1 and 8) of the F_1 and F_2 values in the distribution were based on the values measured by Hawkins & Midgley (2005) for five male speakers of RP English in the age range of 35-40 years, see table 3. The Standard Southern British English /æ/ and /ɛ/ were chosen rather than the American English variants, because the American English vowels are diphthongs, and the direction of the formant transition in these sounds seems to be an important cue for perception (Hillenbrand, Getty, Clark & Wheeler, 1995).

Table 3. Average F_1 and F_2 values in ERB and Hz for English /æ/ and /ɛ/ by five male speakers (35-40 years) of RP English (Hawkins & Midgley, 2005).[5]

	/æ/	/ɛ/
F_1 (ERB)	12.50	10.44
F_2 (ERB)	18.94	20.42
F_1 (Hz)	689.88	511.48
F_2 (Hz)	1573.78	1881.74

As opposed to a natural bimodal distribution of vowels, which is based directly on the F_1 and F_2 values of natural tokens, for this study an enhanced bimodal distribution was used. The endpoints of an enhanced distribution exaggerate the natural differences between the vowels in the bimodal continuum, therefore making them possibly easier to distinguish from one another. Our choice for an enhanced distribution was motivated by Escudero et al.'s (2011)

[5] The conversion formula used is: $F_{ERB} = 11.17 \ln ((F_{Hz} + 312) / (F_{Hz} +14680)) + 43$.

finding that only an enhanced bimodal distribution yielded a significant effect of training, and a natural bimodal distribution did not. Figure 4 shows the difference between a natural bimodal distribution (in brown) and an enhanced bimodal distribution (in red).

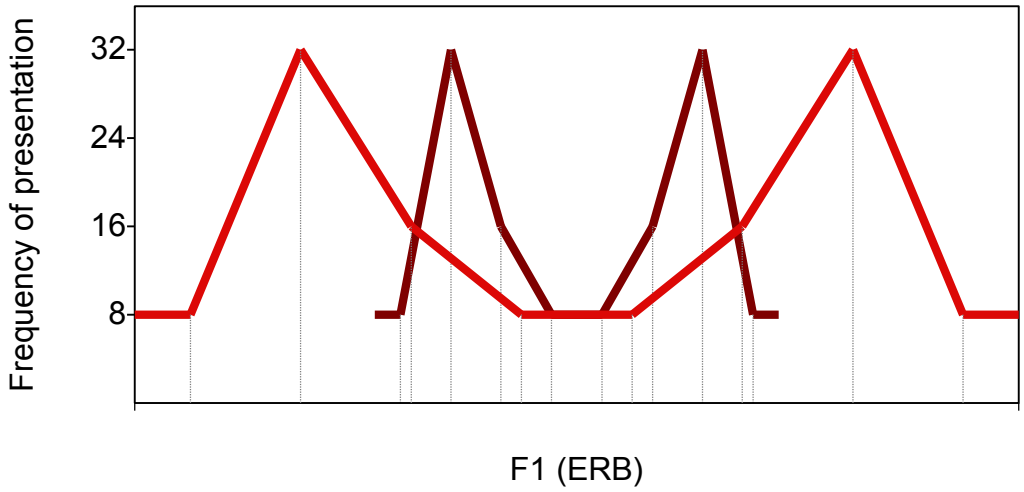

Figure 4. A natural (in brown) and an enhanced (in red) bimodal distribution of a continuum.

To create this enhanced distribution, we used the same calculations as was done in Escudero et al. (2011). The value of the endpoint of /ɛ/ was calculated by subtracting the standard deviation of F_1 from the vowel's average F_1 value and adding the standard deviation of F_2 to the vowel's average F_2 value. The value of the endpoint of /æ/ was calculated by adding the standard deviation of F_1 to the vowel's average F_1 value and subtracting the standard deviation of F_2 from the vowel's average F_2 value. The steps between the tokens were approximately equal on the psychoacoustic ERB scale (0.44 ERB for the F_1 and 0.30 ERB for the F_2).[6] Table 4 shows the F_1 and F_2 of the eight tokens in our enhanced distribution. Like the A and B stimuli, all tokens in the training had a duration of 140 ms and an F_0 that fell from 150 Hz to 100 Hz.

Table 4. The F_1 and F_2 values of the eight training stimuli.

token number	1	2	3	4	5	6	7	8
token frequency	8	32	16	8	8	16	32	8
F_1 (ERB)	9.93	10.37	10.81	11.25	11.69	12.13	12.57	13.01
F_2 (ERB)	20.74	20.44	20.13	19.83	19.53	19.23	18.92	18.62
F_1 (Hz)	472.72	506.04	540.86	577.25	615.28	681.48	696.62	740.12
F_2 (Hz)	1955.56	1886.28	1817.19	1752.66	1690.32	1593.32	1569.96	1513.76

[6] The steps between the F_1 and F_2 tokens are comparable to the step sizes in Escudero et al. (2011): 0.4 ERB and 0.4 ERB.

The shape of the distribution was identical to distributions used in previous studies of distributional learning with adult participants (e.g. Maye & Gerken, 2000, 2001; Gulian et al., 2007; Hayes-Harb, 2007; Escudero et al., 2011). The near-endpoint tokens 2 and 7 are presented most frequently, and are presented four times as often as the centre tokens 4 and 5. Figure 5 shows a graph of the frequency of occurrence of the eight training tokens on the F_1 continuum.

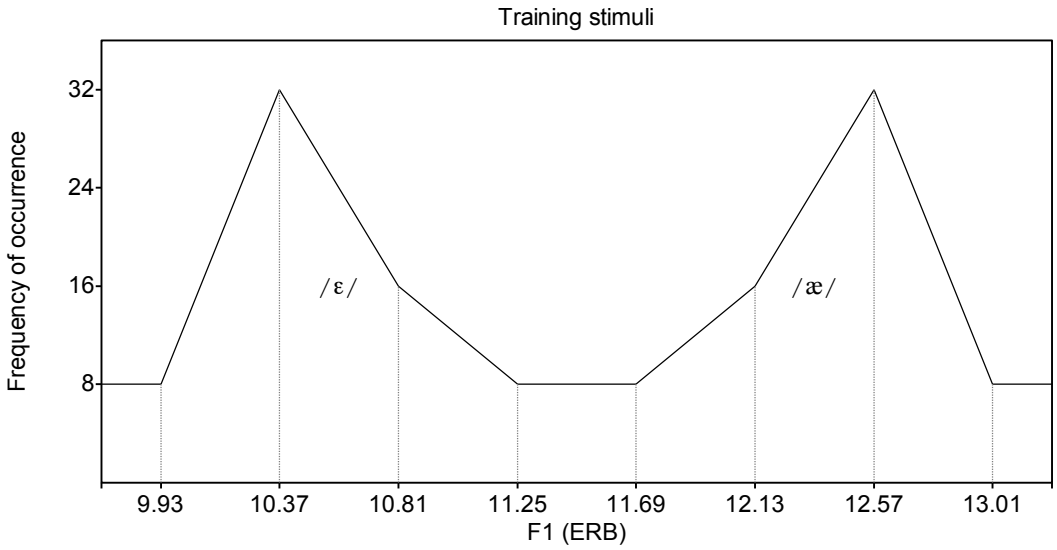

Figure 5. Frequency of occurrence and F_1 values (in ERB) for the eight tokens used in the training.

5. RESULTS

5.1 Descriptives

Table 5 shows the average percentage of correctly answered trials for the pre-test and post-test in the music and enhanced condition, as well as the average difference in percentage points between the pre-test and the post-test. The 95% confidence intervals are shown in table 6.

Table 5. Averaged percentages of correct responses for the pre-tests and post-tests and their average difference score in the enhanced and music condition. Standard deviations between participants are between parentheses.

	Music	Enhanced
Pre-test	64.33 (9.97)	64.98 (12.03)
Post-test	71.50 (12.53)	68.55 (14.29)
Difference	7.18 (7.68)	3.58 (7.51)

Table 6. Lower and upper bounds of the 95% confidence intervals.

	Music	Enhanced
Pre-test	61.49 – 67.16	61.56 – 68.39
Post-test	67.94 – 75.06	64.49 – 72.61
Difference	4.99 – 9.36	1.44 – 5.71

The percentages of correct responses for the pre-test and the post-test per participant are graphically depicted in figure 6. This shows us how the data are spread. Note that one participant obtained scores of only 28.25% (pre-test) and 20% (post-test), having probably somehow misunderstood the purpose of the test or applying a counterproductive strategy. For all participants in general, an upward linear trend is visible. This means that the obtained pre-test scores seem to correlate with the post-test scores (see section 5.3 for a statistical analysis of this correlation).

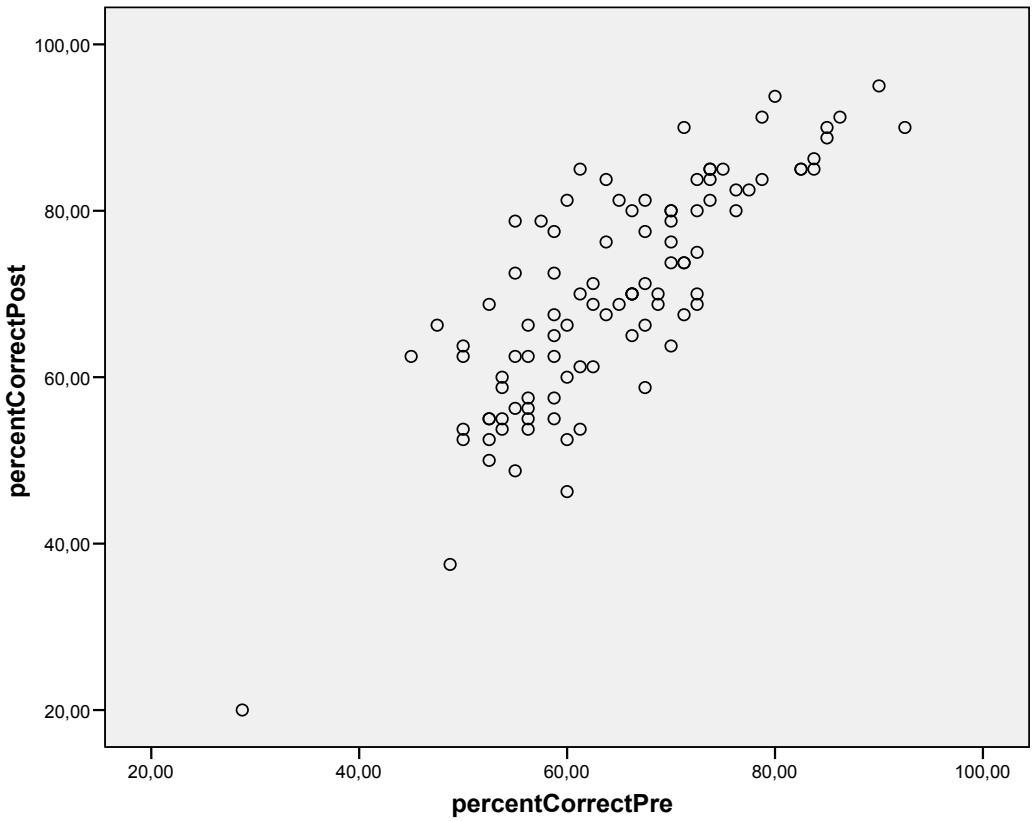

Figure 6. In this scatterplot the x-axis represents the percentage of correct responses per participant in the pre-test, while the y-axis represents the percentage of correct responses per participant in the post-test.

5.2 Distribution of the data

It was examined whether the data were normally distributed. For this, a Kolmogorov-Smirnov test of normality and a Shapiro-Wilk test of normality were conducted. The results are presented in table 7 and are considered significant if $p < 0.05$.

Table 7. Results of the Kolmogorov-Smirnov and Shapiro-Wilk tests of normality.

		Kolmogorov-Smirnov			Shapiro-Wilk		
		Statistic	df	Significance	Statistic	df	Significance
Pre-test	Music	0.14	50	0.01	0.96	50	0.06
	Enhanced	0.09	50	0.20*	0.98	50	0.45
Post-test	Music	0.10	50	0.20*	0.97	50	0.18
	Enhanced	0.08	50	0.20*	0.96	50	0.05
Difference	Music	0.09	50	0.20*	0.98	50	0.66
	Enhanced	0.12	50	0.07	0.97	50	0.31

* This is a lower bound of the true significance.

As can be seen from the table above, all data were normally distributed except for the data from the music condition in the pre-test, which were below the 5% significance level in the Kolmogorov-Smirnov test (i.e. not normally distributed), but not in the Shapiro-Wilk test (i.e.

normally distributed). A normal Q-Q plot (figure 7) for these data shows that the data points are close to the diagonal line, which is an indication of normally distributed data.

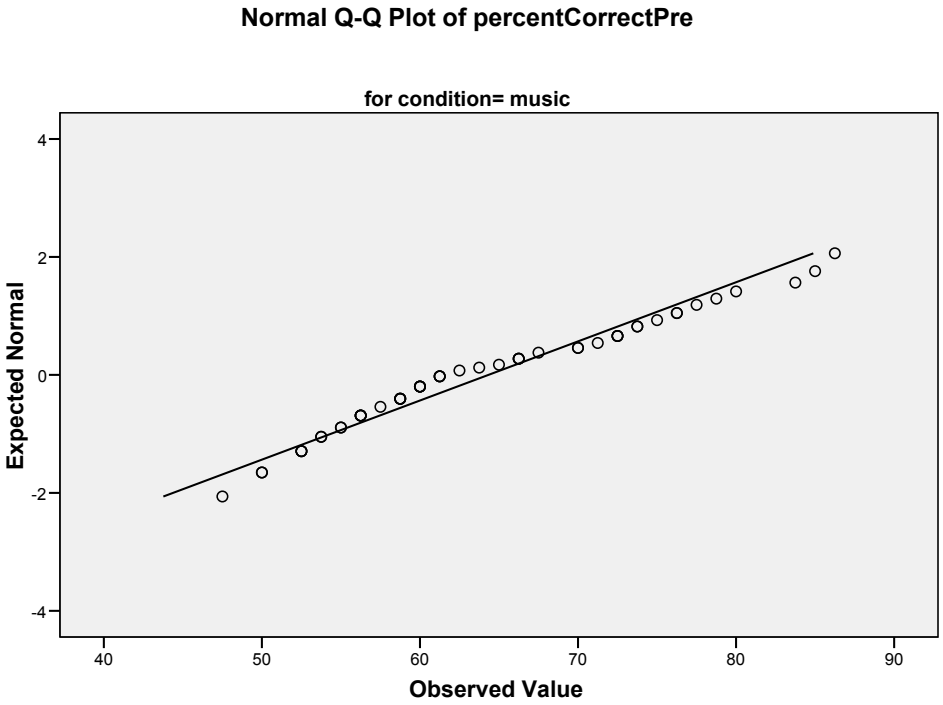

Figure 7. A normal Q-Q plot for the percentages of correct responses in the pre-test for participants in the music condition.

Thus, because it is somewhat unclear if the data from the music condition in the pre-test should be treated as normally distributed or not, in analyses involving these data both parametric and non-parametric statistics will be used, and the results will be compared.

5.3 The effect of training

An independent samples *t*-test revealed no significant difference between the enhanced and music condition in the pre-test ($t(98) = 0.29$, $p = 0.77$, 95% confidence interval = -5.04 – 3.73). The same result was obtained by using a Mann-Whitney U test ($Z = 0.36$, $p = 0.72$). This means that potential differences between both conditions in the post-test or in the difference score cannot be attributed to already existing differences between the groups in the pre-test, and this allows us to compare the effect of training between the groups. Two one-sample *t*-tests also showed that both groups scored significantly above chance level in the pre-test (music: $t(49) = 10.16$, $p = <0.001$, 95% CI = 11.49 – 17.16; enhanced: $t(49) = 8.80$, $p = <0.001$, 95% CI = 11.56 – 18.39).

In order to determine whether training had a beneficial effect on the test scores, a dependent samples *t*-test was used for the music and enhanced conditions to examine the relationship between the scores on the pre-test and the post-test. For both groups, there was a significant increase in achievement (music: $t(49) = 6.60$, $p = <0.001$, 95% CI = 4.99 – 9.36; enhanced: $t(49) = 3.37$, $p = 0.001$, CI = 1.44 – 5.71). Non-parametric testing by means of a Wilcoxon Signed Rank Test for the music condition revealed the same outcome ($Z = 5.06$, $p = <0.001$). Thus, on average participants achieved higher scores in the post-test than in the pre-test.

There were also significant positive Pearson correlations between the obtained scores on the pre-test and the post-test: $r = 0.79$ for the music group ($p = <0.001$) and $r = 0.85$ for the enhanced group ($p = <0.001$). For the music condition, a non-parametric Spearman's rho correlation yielded the exact same outcome ($\rho = 0.79$, $p = <0.001$). As can be seen from figure 6, this indicates that participants who do well on the pre-test, usually do well on the post-test too, and vice-versa.

To assess whether the size of the improvement was significantly different between the music group and the enhanced group, an independent samples *t*-test was conducted on the average difference score between the pre-test and post-test for both groups. Contrary to expectations, it was found that having listened to music during training yielded a significantly larger improvement than having received an enhanced bimodal distributional training ($t(98) = 2.37$, $p = 0.02$, 95% CI = 0.58 – 6.62).

Finally, no significant correlations (Pearson and Spearman's rho) were found between the number of mistakes made in the DIALANG test (version 0.93.1, Lancaster University) on the one hand, which measured general listening comprehension in English, and the pre-test score, the post-test score and the difference score on the other hand for both groups combined and separated. Thus, general listening comprehension and the ability to (learn to) compare differences between sounds do not seem to be correlated.

5.4 Improvement within tests

We also investigated whether the participants not only learned from training, but from the task itself too. This is especially relevant to know for the participants in the music condition, which cannot be regarded as a real training.[7] For this purpose, the data sets for the pre-test

[7] Although long-term musical training has been shown to facilitate the processing of speech sounds (e.g. Besson, Chobert & Marie, 2011; Chobert, Marie, François, Schön & Besson, 2011), as far as we know listening to classical music for two minutes does not have an immediate effect on the perception of non-native vowel contrasts.

and the post-test were split in half (two times forty trials instead of one time eighty trials). Descriptive statistics are shown in table 8 and 9.

Table 8. Averaged percentages of correct responses for the pre-test and post-test in the enhanced and music condition split in two halves. Standard deviations between participants are between parentheses.

	Music	Enhanced
Pre-test trials 1-40	61.60 (9.63)	62.75 (11.46)
Pre-test trials 41-80	67.05 (13.91)	67.20 (15.17)
Difference	5.45 (13.22)	4.45 (12.00)
Post-test trials 1-40	70.90 (14.51)	67.40 (14.65)
Post-test trials 41-80	72.10 (12.17)	69.70 (16.38)
Difference	1.20 (9.41)	2.30 (12.19)

Table 9. Lower and upper bounds of the 95% confidence intervals.

	Music	Enhanced
Pre-test trials 1-40	58.86 – 64.34	59.49 – 66.01
Pre-test trials 41-80	63.10 – 71.00	62.89 – 71.51
Difference	1.69 – 9.21	1.04 – 7.86
Post-test trials 1-40	66.78 – 75.02	63.24 – 71.56
Post-test trials 41-80	68.64 – 75.56	65.05 – 74.35
Difference	-1.47 – 3.87	-1.17 – 5.77

As table 8 reveals, in both tests and for both conditions the average score was higher in the second half of trials than in the first. Before testing whether these increases are significant, Kolmogorov-Smirnov and Shapiro-Wilk tests were used to determine whether the data were normally distributed. Table 10 shows that according to the Kolmogorov-Smirnov test, for the music condition pre-test trials 1-40 and 41-80 were not normally distributed, and for the enhanced condition pre-test trials 1-40. According to the Shapiro-Wilk test, for the music condition pre-test trials 1-40 and post-test trials 1-40 were not normally distributed, and for the enhanced condition this was the case for pre-test trials 1-40 and 41-80, and for post-test trials 41-80. Therefore, non-parametric testing was used.

Table 10. Kolmogorov-Smirnov and Shapiro-Wilk tests of normality.

		Kolmogorov-Smirnov			Shapiro-Wilk		
		Statistic	df	Significance	Statistic	df	Significance
Pre-test trials 1-40	Music	0.15	50	0.01	0.95	50	0.02
	Enhanced	0.14	50	0.02	0.95	50	0.04
Pre-test trials 41-80	Music	0.13	50	0.04	0.97	50	0.30
	Enhanced	0.11	50	0.20(*)	0.94	50	0.01
Post-test trials 1-40	Music	0.12	50	0.09	0.95	50	0.02
	Enhanced	0.11	50	0.20(*)	0.96	50	0.08
Post-test trials 41-80	Music	0.12	50	0.05	0.97	50	0.18
	Enhanced	0.11	50	0.18	0.95	50	0.04

* This is a lower bound of the true significance.

A Wilcoxon Signed Ranks Test showed that in the pre-test the positive differences between the first and second half of trials were significant for both groups together (Z = 3.90, p = <0.001) and separately (music: Z = 2.73, p = 0.006; enhanced: Z = 2.79, p = 0.005),

indicating that participants did indeed learn from the test itself. A Mann-Whitney U Test revealed no significant differences between the pre-test difference scores from participants in the music and the enhanced conditions.

In the post-test, the differences in correct responses between the first and second half of trials were not significant for participants in both conditions, whereas the average difference between the second half of the pre-test and the first half of the post-test was significant for participants in the music condition only ($Z = 2.21$, $p = 0.03$). However, no significant differences were found between the groups with respect to the first and second half of the post-test trials.

5.5 Differences between [æ] and [ɛ]

To conclude, we wanted to know whether there was a difference in response behaviour regarding /æ/ and /ɛ/. Again, first it was determined whether the data were normally distributed, see table 11. It was found that according to the Kolmogorov-Smirnov test, post-test data of [æ] were not normally distributed, with the Shapiro-Wilk test yielding this same outcome, and additionally that also pre-test data of [æ] were not normally distributed.

Table 11. Kolmogorov-Smirnov and Shapiro-Wilk tests of normality.

		Kolmogorov-Smirnov			Shapiro-Wilk		
		Statistic	df	Significance	Statistic	df	Significance
Pre-test	[æ]	0.09	100	0.07	0.970	100	0.02
	[ɛ]	0.09	100	0.05	0.98	100	0.29
Post-test	[æ]	0.15	100	0.00	0.90	100	0.00
	[ɛ]	0.08	100	0.13	0.99	100	0.37

Then it was analysed whether there was a difference between the music and enhanced condition. Data for both groups are shown in table 12.

Table 12. Averaged percentages of correct responses for [æ] and [ɛ] in the pre-test and post-test for the music and the enhanced condition. Standard deviations are between parentheses.

	Music	Enhanced
Pre-test [æ] trials	73.20 (14.02)	75.15 (15.31)
Pre-test [ɛ] trials	55.45 (11.17)	54.80 (12.35)
Post-test [æ] trials	81.70 (16.88)	80.75 (17.55)
Post-test [ɛ] trials	61.30 (12.21)	56.35 (14.97)

Since not all data were normally distributed, a Mann-Whitney U Test was conducted, and no significant difference between the groups was detected for both [æ] and [ɛ] in the pre-test as well as in the post-test. Therefore, all participants were grouped together for further analysis. Table 13 shows the averaged percentages of correct responses for both vowels in the pre-test and post-test.

Table 13. Averaged percentages of correct responses for [æ] and [ɛ] in the pre-test and post-test. Standard deviations between participants are between parentheses.

	[æ]	[ɛ]	Difference
Pre-test	74.18 (14.64)	55.13 (11.72)	19.05 (14.82)
Post-test	81.23 (17.14)	58.83 (13.81)	22.40 (15.65)
Difference	7.05 (10.97)	3.70 (10.45)	-

This table is quite revealing in several ways. First, there is a huge and significant difference in the scores that were obtained for [æ] and [ɛ], with a mean difference of 19.05 percentage points in favour of [æ] in the pre-test and 22.40 in the post-test (pre-test: $Z = 8.26$, $p = <0.001$; post-test: $Z = 8.14$, $p = <0.001$). Furthermore, the participants' improvement was nearly significantly higher for [æ] than [ɛ]: 7.05 percentage points compared to 3.70 ($Z = 1.96$, $p = 0.05$). Thus, participants perceived [æ] more accurately than [ɛ], and presumably showed a larger improvement in perceiving [æ] than [ɛ].

Since there were only two answer possibilities and the participants more often responded correctly to [æ]-stimuli than to [ɛ]-stimuli, it seems that the participants were biased towards answering [æ] rather than [ɛ]. Table 14 shows the participants' response behaviour.

Table 14. Descriptive statistics of [æ] and [ɛ] responses in percentages.

		[æ]	[ɛ]
Pre-test	Mean	59.53	40.48
	Standard deviation	7.41	7.41
	95% confidence interval	58.06 – 61.00	39.01 – 41.95
Post-test	Mean	61.20	38.80
	Standard deviation	7.83	7.83
	95% confidence interval	59.65 – 62.75	37.25 – 40.35

On average 59.53% of the responses in the pre-test and 61.20% of the responses in the post-test were [æ], whereas only 50% of the trials required an [æ] response. A one sample Wilcoxon signed rank test with a test value of 50 showed that these differences are significant (pre-test: $p < 0.001$; post-test: $p < 0.001$), i.e. that the participants were biased towards answering [æ] rather than [ɛ].

6. DISCUSSION

6.1 Relating the findings to prior research

This study was conducted to shed some light on previously found discrepancies between two different studies on distributional learning. The first study, an XAB task by Escudero et al. (2011), showed that listening to an enhanced bimodal distribution of Dutch /a:/ and /ɑ/ improved Spanish learners' ability to perceive this non-native vowel contrast, whereas this could not be shown for listening to a normal bimodal distribution or listening to classical music. On the contrary, in the other study, an EEG experiment by Wanrooij et al. (2012), it was found that discrimination weakened for Dutch listeners after exposure to a distribution of English /æ/ and /ɛ/ for all participants combined, but not for any of the groups alone (unimodal, bimodal, enhanced bimodal and music). No significant differences between the groups were found.

6.1.1 Differences to the outcome of the present study

The present study used Escudero et al.'s (2011) test design, while maintaining Wanrooij et al.'s (2012) vowel contrast and native language of the participants. It was shown that exposure to classical music yielded more improvement of non-native vowel perception than exposure to an enhanced bimodal distribution. This is contrary to Escudero et al.'s (2011) finding that learners benefited from listening to an enhanced bimodal distribution and less or not at all from listening to classical music, and contrary to Wanrooij et al.'s (2012) finding that there were no significant differences between participants in all conditions. At the same time, the findings from the present study are also contrary to Wanrooij et al.'s (2012) finding that post-test MMN was smaller than pre-test MMN for all participants, indicating that discrimination had weakened after distributional training or having listened to music. As opposed to this, our participants in both conditions on average obtained higher post-test than pre-test scores.

This raises the question as to how these conflicting findings can be explained. First, why did participants in the EEG experiment show weakened discrimination, while participants in the XAB task showed improved performance? Of course, in an EEG experiment and an XAB task a different kind of response is measured. Still, there was no reason to expect that distributional training or listening to music would lead to deterioration of the ability to discriminate speech sounds. In the introduction, various possible explanations for the different outcomes of the EEG experiment and XAB task have been discussed in advance. Although these explanations still stand, for example that participants in the XAB task are more likely to learn from the task itself than participants in the EEG experiment,

most of them do not explain the actual deterioration in perception that participants in the EEG experiment seem to undergo. An explanation that could still be relevant is that fatigue or boredom has influenced the learning process, since the EEG experiment took almost two hours to complete and was conducted in a relatively dark and warm studio. As mentioned in the introduction, the size of MMN in EEG experiments can decline due to sleepiness (e.g. Lang et al., 2005).

Second, the finding that participants in the music condition had a significantly higher difference score than participants in the enhanced condition was very much unanticipated. The enhanced training had especially been designed to aid participants with differentiating between [æ] and [ɛ], and this design had been shown to be successful by Escudero et al. (2011), although Wanrooij et al. (2012) did not find any effect of distributional training. Nevertheless, neither of those studies showed a greater improvement for the music condition in comparison with the enhanced or normal bimodal condition.

Therefore, it is difficult to explain this result, but it might be related to the fact that participants in the music condition could have used the 'training' phase to relax and gain new concentration to answer 80 trials again, whereas participants in the enhanced condition might have gotten tired from nonstop listening to vowels and lost concentration. However, since Escudero et al. (2011) used the exact same test design[8] and still found a greater improvement for participants in the enhanced condition, this explanation is not satisfactory.

Another possible explanation could be that the composition of the two groups was unequal in terms of the listening strategies used by the participants. Wanrooij et al. (submitted) showed that changes in learners' improvement in identification of difficult non-native vowel contrasts are dependent on listeners' initial listening strategies, and that group composition in terms of listening strategies can have an important effect on group results. In the present study, possible initial differences in listening strategies were not controlled for and might have caused the unanticipated outcome. However, since statistical analysis yielded no significant difference between the music and the enhanced condition in the pre-test, this explanation does not seem satisfactory either.

[8] The choice of classical music in our study (Chopin's waltz in G flat major, a piece for piano) was different from the music used by Escudero et al. (2011) (Händel's *Water Music*, a piece for orchestra). However, it is not to be expected that this would cause any difference in the ability to differentiate sounds.

6.1.2 Discrepancies between Escudero et al. (2011) and Wanrooij et al. (2012)

Originally, the present study was set up in the hope to be able to explain some of the discussed discrepancies between Escudero et al. (2011) and Wanrooij et al. (2012). On the basis of their results, it was anticipated to either find a positive effect of distributional training or no effect. Finding a positive effect would mean that the discrepancies between Escudero et al. (2011) and Wanrooij et al. (2012) should probably be attributed to differences between the test designs of both studies. In the case of finding no significant effect of distributional training, we could most likely have attributed the discrepancies to differences in the Dutch and Spanish vowel systems.

Yet, rather than finding a positive or no effect of distributional training, we found a relatively negative effect, which is inconsistent with the findings of Escudero et al. (2011) as well as those of Wanrooij et al. (2012). This means that only more questions have arisen, instead of answers. We still cannot say what caused the different outcomes of the studies by Escudero et al. (2011) and Wanrooij et al. (2012).

In the introduction, we mentioned the possibility of vowel system density having influenced the results, the Dutch vowel inventory being much larger than the Spanish. Prior research on this topic by Iverson & Evans (2007, 2009) focused on the effect of vowel system density on learning English vowels via computer-based auditory training. Native speakers of German (which has a dense vowel system comparable to that of Dutch) improved twice as much as native speakers of Spanish. If these results are transferable to the current situation, it is to be expected that distributional training would be more effective for native speakers of Dutch than for native speakers of Spanish, but the opposite proved to be the case.

However, we do not know if the results are transferable to the situation under investigation since distributional training is different from the training used by Iverson & Evans (2007, 2009), and since the speakers of German and Spanish in their study differed from one another in the amount of experience they had had with English.[9] Another important difference between both studies was that Iverson and Evans used a vowel identification task in which participants heard a stimulus word (bVt) and had to select this word from fourteen written response options (bVt with fourteen different vowels), while participants in the present study heard an isolated vowel and had to choose between two audibly presented response options (also isolated vowels). In terms of Boersma's model of Bidirectional Phonology and

[9] The Dutch and Spanish participants in the studies that are discussed in this thesis may also differ from each other in terms of age and experience with the English language. Nevertheless, upon comparing the pre-test scores in the present study and the study by Wanrooij et al. (submitted), which is a replication of Escudero et al. (2011) and showed similar results, an ANOVA with pre-test scores as the dependent variable and condition as factor, yielded no significant differences ($F(4, 245) = 1.79$, $p = 0.13$).

Phonetics (2011), Iverson and Evans seem to have measured accuracy in the recognition of phonological (or possibly lexical) representations, whereas the present study seems to have measured accuracy in the recognition of phonetic (or possibly phonological) representations.

Also mentioned in the introduction were various differences between the EEG and XAB task test design that could have accounted for the finding that enhanced distributional training as compared to music had a positive effect on non-native vowel perception in Escudero et al. (2011), but that there were no significant differences between the experimental conditions in Wanrooij et al. (2012). One of the tentative explanations was the use of a continuous sequence of vowels in the training phase of the EEG versus the use of a discontinuous sequence of vowels in the training phase of the XAB task. In the meantime, Wanrooij & Boersma (personal communication, in preparation) just showed that there are no significant differences in improvement between training with a continuous and a discontinuous sequence of vowels in an XAB task, so this explanation can be ruled out. The other potential explanations discussed in the introduction are still valid.

6.2 Other findings

6.2.1 Improvement within tests

Focusing only on the present study, it should be noted that both groups had significantly higher post-test scores than pre-test scores. Statistical analysis (see section 5.4) showed that this outcome can be explained at least partly from the participants' having learned from the test itself: in the pre-test, on average participants obtained higher scores for the second half of trials compared to the first half, with no significant differences between the music and the enhanced condition. For the music condition but not for the enhanced condition, scores from the first half of the post-test were also significantly higher than those from the second half of the pre-test, although there were no differences between groups on the score of the first and second half of the post-test.

During the pre-test, the participants had not yet received any training. The fact that on average their scores still improved shows us that they learned from testing itself, changing their listening and/or their answering strategies. It is interesting that the positive difference between the second half of the pre-test and the first half of the post-test was significant for the music condition only. Possibly, this was brought about by the different kind of training participants in the music condition received as compared to participants in the enhanced condition, with the music training either having a favourable effect on learning from testing itself, or the enhanced training having a detrimental effect. For both groups, the

scores on the second half of the post-test were a little higher than on the first half, but this difference was not significant. Therefore, it could not be shown that the participants also learned from the test itself in the post-test; probably some kind of plateau was reached.

6.2.2 Bias towards [æ]

Regarding the [æ] and [ɛ] X tokens, it was expected that participants would answer X tokens that were [ɛ] more often correctly than X tokens that were [æ], since Dutch /ɛ/ is much more similar in F_1 and F_2 to English /ɛ/ than to English /æ/ (Adank, Van Hout & Smits, 2004; Hawkins & Midgley, 2005). However, contrary to these expectations, exactly the opposite was the case. Participants answered [æ] trials way more often correctly than [ɛ] trials and presumably showed greater improvement with the former (see section 5.5). It seems that these results are brought about by the fact that participants were biased towards answering [æ] rather than [ɛ].

The most likely explanation for the bias seem to be durational differences between the natural [æ] and [ɛ] stimuli (X stimuli), which were shown in table 1 and repeated here in table 15. The natural stimuli for [æ] on average had a longer duration than the natural stimuli for [ɛ], but for both [æ] and [ɛ] the synthetic stimuli (A & B response options) had a duration of 140 ms. This could have biased the participants to answer [æ].

Table 15. Durational values in ms. for [æ] and [ɛ] in different stimulus conditions and pronounced by male and female speakers.

Vowel	Stimulus	Male	Female
[æ]	X stimuli	114	124
	A & B stimuli	140	-
[ɛ]	X stimuli	97	118
	A & B stimuli	140	-

We can also look for an explanation by comparing the Dutch and English vowel systems. Although the Dutch vowel system lacks /æ/, it does have /a:/ which is quite similar to /æ/ in terms of formants.[10] Table 16 shows the F_1 and F_2 values for Standard Southern British English /æ/ and /ɛ/ as measured in our own X stimuli,[11] and the F_1 and F_2 values for Dutch /a:/ and /ɛ/ as measured by Adank et al. (2004) for /a:/ and /ɛ/ in Northern Standard Dutch.

[10] These vowels do differ in duration, with Dutch /a:/ having an average duration of 203ms for male speakers (Adank et al., 2004).

[11] It should be noted that the values for the A & B response options were based on the values as measured by Hawkins & Midgley (2005).

Table 16. F_1 and F_2 values in Hz and ERB for Standard Southern British English /æ/ and /ɛ/ and Northern Standard Dutch /a:/ and /ɛ/, shown for male and female speakers.

Vowel	Stimulus	Male				Female			
		F_1		F_2		F_1		F_2	
		ERB	Hz	ERB	Hz	ERB	Hz	ERB	Hz
English /æ/	X stimuli	13.27	767	18.23	1443	15.00	963	19.21	1626
	A & B stimuli	11.99	642	19.32	1648	-	-	-	-
Dutch /a:/	Adank et al. (2004)	12.29	670	18.12	1425	14.58	912	18.93	1572
English /ɛ/	X stimuli	11.36	587	19.38	1660	12.43	683	21.01	2021
	A & B stimuli	10.95	552	20.04	1797	-	-	-	-
Dutch [ɛ]	Adank et al. (2004)	09.96	475	19.77	1739	10.74	535	20.89	1990

As can be seen from table 16 (considering male speakers only), with its F_1 value of 12.29 ERB the Dutch /a:/ is actually closer to the [æ] answering possibility (11.99 ERB) than the Dutch /ɛ/ (9.96 ERB) is to the [ɛ] answering possibility (10.95 ERB).[12] Therefore, participants could have perceived English [æ] more easily as Dutch /a:/ (setting aside the durational difference between these vowels) than English [ɛ] as Dutch /ɛ/. This could explain the higher accuracy rate for [æ] compared to [ɛ], but offers no explanation for the observed bias. As for the F_2, potential differences are not so obvious.

[12] This is also the case for female speakers, but it is more difficult to draw conclusions from this as the A & B response options were only synthetic male voices.

7. CONCLUSION

In short, the present study did not bring us any closer to explaining the different effects of distributional training as described in Escudero et al. (2011) and Wanrooij et al. (2012). Considering Iversons & Evans' (2007, 2009) findings that having a large vowel inventory facilitates non-native vowel learning, it is likely that the cause will lie in the different test designs of Wanrooij et al.'s (2012) EEG experiment and Escudero et al.'s (2011) XAB task. Wanrooij & Boersma (personal communication, in preparation) made a first step towards examining the differences in test design by testing whether (dis)continuity of the vowel sequence during training influences listeners' improvement. However, this did not prove to be the case. Therefore, further work needs to be done in order to establish whether other differences between the EEG experiment and the XAB task can account for the different outcomes. For instance, it would be of interest to assess whether there is an effect of the focus of attention, which was fixed on the test in the XAB task but not in the EEG experiment.

All in all, in view of the findings of the present study and those of Wanrooij et al. (2012), it seems that the effectiveness of distributional training of non-native vowel contrasts may be questionable. Some studies, including Escudero et al. (2011) and studies mentioned in the introduction, have found significant improvement on the perception of non-native vowel contrasts after distributional training. Personal communication with researchers from various other universities, however, informs us that distributional training often seems ineffective, which is also the picture that emerged from our own data.

In spite of this outcome, the present study enhanced our understanding of the perception of non-native vowel contrasts, specifically the perception of the Standard Southern British English /æ/-/ɛ/ contrast for adult native speakers of Dutch. A very interesting finding was that perception improved during testing itself, independent of training. This shows that although distributional training itself may be ineffective, there are other ways to aid second language vowel perception.

8. REFERENCES

Adank, Patti, Roeland van Hout & Roel Smits (2004). An acoustic description of the vowels of Northern and Southern Standard Dutch. *Journal of the Acoustical Society of America* 116 (3), 1729-1738.

Aoyama, Katsura, James Emil Flege, Susan G. Guion, Reiko Akahane-Yamada & Tsuneo Yamada (2004). Perceived phonetic dissimilarity and L2 speech learning: the case of Japanese /r/ and English /l/ and /r/. *Journal of Phonetics* 32, 233-250.

Benz, Anton & Jason Mattausch, eds. (2011). *Bidirectional Optimality Theory*. Amsterdam: John Benjamins.

Besson, Mireille, Julie Chobert & Céline Marie (2011). Transfer of training between music and speech: common processing, attention, and memory. *Frontiers in Psychology*, doi: 10.3389/fpsyg.2011.00094.

Boersma, Paul (2011). A programme for bidirectional phonology and phonetics and their acquisition and evolution. In: Anton Benz & Jason Mattausch, eds. (2011), 33-72.

Boersma, Paul & Weenink, David (2012). *Praat: doing phonetics by computer* [computer program]. Version 5.3.14.

Booij, Geert (1995). *The phonology of Dutch*. Oxford: Clarendon Press.

Chobert, Julie, Céline Marie, Clément François, Daniele Schön & Mireille Besson (2011). Enhanced Passive and Active Processing of Syllables in Musician Children. *Journal of Cognitive Neuroscience* 23 (12), 3874-3887.

Cutler, Anne, Andrea Weber, Roel Smits & Nicole Cooper (2004). Patterns of English phoneme confusions by native and non-native listeners. *Journal of the Acoustical Society of America* 116 (6), 3668-3678.

Do, Anna H.-J., Laura Domínguez & Aimee Johansen, eds. (2001). *BUCLD 25: Proceedings of the 25th annual Boston University Conference on Language Development*. Somerville, MA: Cascadilla Press.

Escudero, Paola, Titia Benders & Karin Wanrooij (2011). Enhanced bimodal distributions facilitate the learning of second language vowels. *Journal of the Acoustical Society of America* 130 (4), EL206-EL212.

Escudero, Paola, Rachel Hayes-Harb & Holger Mitterer (2008). Novel second-language words and asymmetric lexical access. *Journal of Phonetics* 36, 345-360.

Escudero, Paola & Karin Wanrooij (2010). The Effect of L1 Orthography on Non-native Vowel Perception. *Language and Speech* 53 (3), 343-365.

Gulian, Margarita, Paola Escudero & Paul Boersma (2007). Supervision hampers distributional learning of vowel contrasts. In: Jürgen Trouvain & William J. Barry, eds. (2007), 1893-1896.

Hammond, Robert M. (2001). *The sounds of Spanish: analysis and application (with special reference to American English)*. Somerville, MA: Cascadilla Press.

Hawkins, Sarah & Jonathan Midgley (2005). Formant frequencies of RP monophthongs in four age groups of speakers. *Journal of the International Phonetic Association* 35 (2), 183-199.

Hayes-Harb, Rachel (2007). Lexical and statistical evidence in the acquisition of second language phonemes. *Second Language Research* 23 (1), 65-94.

Hessen, A.J. van & M.E.H. Schouten (1992). Modeling phoneme perception. II: A model of stop consonant discrimination. *Journal of the Acoustical Society of America* 92 (4), 1856-1868.

Hillenbrand, James, Laura A. Getty, Michael J. Clark & Kimberlee Wheeler (1995). Acoustic characteristics of American English vowels. *Journal of the Acoustical Society of America* 97 (5), 3099-3111.

Howell, S. Catherine, Sarah A. Fish & Thea Keith-Lucas, eds. (2000). *BUCLD 24: Proceedings of the 24th annual Boston University Conference on Language Development*. Somerville, MA: Cascadilla Press.

Iverson, Paul & Bronwen G. Evans (2007). Auditory training of English vowels for first-language speakers of Spanish and German. In: Jürgen Trouvain & William J. Barry, eds. (2007), 1625-1628.

Iverson, Paul & Bronwen G. Evans (2009). Learning English vowels with different first-language vowel systems II: Auditory training for native Spanish and German speakers. *Journal of the Acoustical Society of America* 126 (2), 866-877.

Lang, A.H., O. Eerola, P. Korpilahti, I. Holopainen, S. Salo & O. Aaltonen (2005). Practical Issues in the Clinical Application of Mismatch Negativity. *Ear & Hearing* 16 (1), 118-130.

Lancaster University (2012). *DIALANG* [computer program]. Version 0.93.1.

Liberman, Alvyn M. (1957). Some results of research on speech perception. *Journal of the Acoustical Society of America* 29, 117-123.

Maye, Jessica & LouAnn Gerken (2000). Learning Phonemes Without Minimal Pairs. BUCLD 24 Proceedings. In: S. Catherine Howell, Sarah A. Fish & Thea Keith-Lucas, eds. (2000), 522-533.

Maye, Jessica & LouAnn Gerken (2001). Learning Phonemes: How Far Can the Input Take Us? In: Anna H.-J. Do, Laura Domínguez & Aimee Johansen (2001), 480-490.

Trehub, Sandra E. (1976). The discrimination of foreign speech contrasts by infants and adults. *Child Development* 47, 466-472.

Trouvain, Jürgen & William J. Barry, eds. (2007). *Proceedings of the 16th International Congress of Phonetic Sciences.* Saarbrücken: University of Saarbrücken.

Wanrooij, Karin, Paola Escudero & Maartje E.J. Raijmakers (submitted). What do listeners learn from exposure to a vowel distribution? An analysis of listening strategies in distributional learning.

Wanrooij, Karin, Titia van Zuijen & Paul Boersma (2012). MMN declines after distributional vowel training. Poster at MMN 2012. *The Sixth Conference on Mismatch Negativity and its Clinical and Scientific Application*, New York, May 4, 2012.

Werker, Janet F., John H.V. Gilbert, Keith Humphrey & Richard C. Tees (1981). Developmental Aspects of Cross-Language Speech Perception. *Child Development* 52, 349-355.

Werker, Janet F. & Richard C. Tees (1984). Cross-language speech perception: Evidence for perceptual reorganization during the First year of life. *Infant Behavior and Development* 7, 49-63.