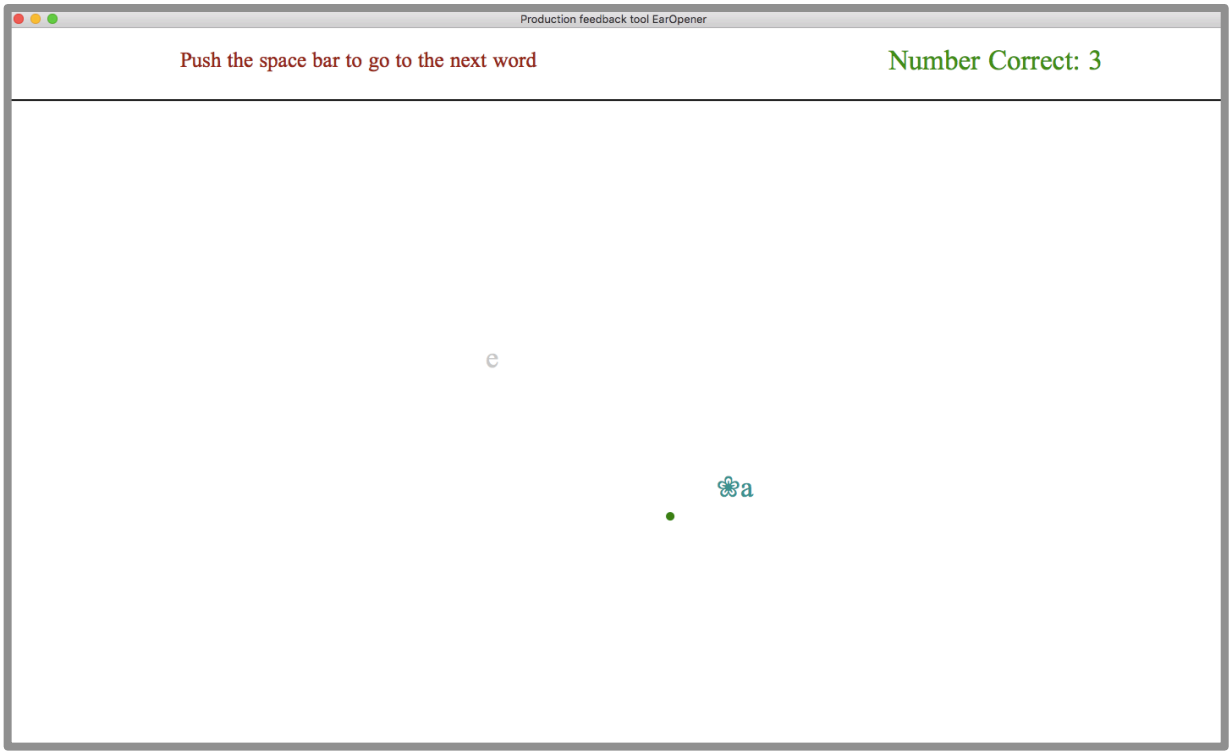
A *pan* is not for writing

Making and testing a tool for online feedback on vowel production of the English /ɛ/–/æ/ contrast



Gisela Govaart, 10089004
June 28th, 2016

Msc Brain & Cognitive Sciences
Track: Cognitive Science
University of Amsterdam
Research Project 1, 26 EC
February 1st - July 1st, 2016

Institutes:
Donders Institute, Nijmegen
ACLC: Phonetic Sciences, Amsterdam

Supervisors:
Dr. Makiko Sadakata (Donders Institute)
Prof. Dr. Paul Boersma (ACLC: Phonetic Sciences)

Co-assessor & UvA-representative:
Dr. David Weenink (ACLC: Phonetic Sciences)

ABSTRACT

Giving reliable feedback on speech production is difficult: many different factors have to be taken into account. In this project, a tool was made to give feedback on the English /ɛ/–/æ/ contrast. This resulted in an extensive (but by no means exhaustive) overview of the different factors that should be taken into account for making a speech production feedback system. First, an LDA model was used to find the most important features for categorization of /ɛ/ and /æ/. These features, the mean F1 and F2 over the whole duration of the vowel, were then used to create the tool. The tool was tested for its accuracy by comparing its feedback on productions of words with /ɛ/ and /æ/ by Dutch natives to ratings on those productions by native English listeners. The results were ambiguous: correlations between the tool's feedback and the native English listeners' ratings were moderate for /æ/, but they were low or even negative for /ɛ/. However, the native English listeners also rated productions by English natives, and unexpected results were found. Native English listeners rated productions of /ɛ/ by English natives often as wrong, i.e. they perceived intended /ɛ/'s as /æ/. This raises the question whether the listeners might have adapted their identification boundary between /ɛ/ and /æ/ according to the Dutch speakers' boundary. It definitely calls into question the suitability of native ratings as a form of feedback, or even as a measure of acoustic correctness. In the discussion section, some recommendations for improvement of the current tool as well as suggestions for further research are given.

Contents

I. INTRODUCTION

In this research project, a tool for online feedback on vowel production was developed in the computer program *Praat* (Boersma & Weenink, 2016), and its functionality was tested. The tool will be used in a later project to analyze and give feedback on productions of the English /ɛ/–/æ/ contrast by Dutch natives, who are being trained to perceive and produce this contrast.

A. Context of the research

This research is part of the *EarOpener* project at the *Donders Institute* in Nijmegen. One of the aims of this project is to investigate the interaction between speech production and perception for L2 learners. These two domains are often studied independently, although they seem to be used and developed in interaction (Franken et al., 2015; Baese-Berk, 2010, McQueen, 2005). However, the exact nature of this interaction is not clear yet. It is sometimes claimed that one needs perceptual discrimination abilities, to guide the "sensorimotor learning of L2 sounds" (Flege, 1995: 238). Others claim that there is no or only moderate correlation between perception and production and therefore reject the idea that perception comes first (Kartushina, 2015). The project addresses the question of how the development of perceptual discrimination abilities for a non-native sound contrast affects the pronunciation of this contrast, and the other way around.

It is known that participants can be trained to learn to perceive a non-native vowel contrast, by exposure to high-variability stimuli (e.g. Logan et al., 1994). Previous research has shown that giving feedback on performance during these training phases helps perceptual learning. Moreover, it is known that feedback on production can improve performance (Neri et al., 2008; Lie-Lahuerta 2011). It has been suggested that exposing participants to these high variability trainings supports the development of abstract phonological categories (Sadakata & McQueen, 2013). The question remains, however, whether these abstract categories are shared for perception and production. This question could be addressed by investigating learning in three different cases: (1) participants receive feedback on perception; (2) participants receive feedback on production; (3) participants receive feedback on both perception and production (EarOpener Project description). For this, a tool is needed that can do an online analysis of the produced utterances, that can give immediate feedback to the participants as to whether or not their pronunciation was correct.

B. Previous studies on speech production feedback and their limits

In the last years, there have been some attempts to give feedback on non-native productions through a speech recognition tool. The most recent paper on this topic is by Kartushina and colleagues (2015). This paper gives an extensive overview of the research on speech production feedback. Most of the studies that are mentioned provide some sort of visual feedback. This feedback can be either based on information about the position and dynamics of the articulators (direct feedback), or based on acoustic analysis of the produced sound (indirect feedback). In the current project, the focus was on indirect feedback. Another example is the Fix Your Vowels (FYV) method (Lie-Lahuerta, 2011), which is a method especially designed with a training purpose: it is designed to teach students to pronounce Spanish vowels.

Most studies that deal with speech production feedback aim at finding out whether the feedback has a significant effect on the quality of non-natives' productions. As long as the tool has the outcome that students/participants indeed show improvement in production, the tool succeeds, at least for practical application (e.g. FYV). However, the exact nature of the feedback is not tested. This means that, even though the tool might improve non-natives' performance, it is possible that the feedback is given on dimensions that are not optimally relevant for the perception of the contrast. On the other hand, it is the question how the tools should be tested: native speakers are notoriously known for their inconsistency in rating productions (Kartushina et al., 2015).

The studies mentioned in Kartushina et al. (2015) differ greatly in terms of behavioral improvements as well as in terms of the methods of the studies. Feedback systems are different, and different acoustic measures are used. There thus seems to be inconsistency in the features that are used to base the feedback on, which means that some studies might give feedback based on features that are not so relevant in speech perception, or they leave out important features. For example, most methods do not take into account the effect of coarticulation. Coarticulation is the phenomenon that the spectral quality of a sound is influenced by surrounding sounds (Stevens & House,1963). It is also known that, regardless of whether the actual sound quality changes, vowel perception is influenced by neighboring spectral content (Holt et al., 2000), and that computer recognition of vowels works better if

it takes coarticulation into account (Nearey, 1989). In previous studies, coarticulation is either not taken into account (e.g. FYV), or vowels in isolation are used to simply avoid the issue (e.g. Kartushina et al. 2015). In the current project, the effect of coarticulation on the categorization of the vowels by the computer will be assessed, to see whether it would be beneficial to add a measure of coarticulation to the tool.

Therefore, the focus of this research project was not to find behavioral effects; instead, it aims at creating a tool of which the performance is known and tested.

D. The English /ɛ/–/æ/ contrast for Dutch speakers

Learning to speak a second language phonologically fluently, i.e. without a foreign accent, is notoriously difficult for late (after puberty) language learners (Escudero, 2005). This is because infants from 6-8 months on learn to specialize to the sound system of their native language(s), which means that they loose the ability to discriminate sound contrasts that are not relevant in their native language (Kuhl, 2004).

The English /ɛ/–/æ/ contrast is known to be a difficult contrast for Dutch speakers (e.g. Flege et al., 1997). This is because Dutch does not have the /ɛ/–/æ/ contrast, and both vowels are close to the Dutch /ɛ/ (Deterding, 1997; Adank et al., 2004). The L2LP model of Escudero (2005) describes this relationship as *new*: the second language (in this case English) has a contrast that the first language (Dutch) does not have, and both members of this contrast are perceptually close to one sound in the first language (in this case the Dutch /ɛ/). Therefore, both sounds will be perceived and produced as this native sound.

E. Current research: aim and overview

The aim of this project was to develop and test a tool that can do reliable online vowel analysis and give visual feedback on the productions of the English /ɛ/–/æ/ contrast by Dutch natives. This paper gives a description of this process, and therefore gives an overview of the different factors that should be taken into account while developing a tool that is meant to give non-native speakers feedback on their production.

Since there are many possible way to calculate the acoustic quality of a vowel, one of the first steps is to find the most accurate way of formant analysis for English /ɛ/–/æ/ contrast. For this, a dataset of productions of the two vowels by English natives was analyzed, and an extensive linear discriminant analysis (LDA) was carried out on a set of utterances produced by 10 native speakers. Finding out which type of formant analysis is the most effective should give future researchers a guideline as to which analysis to use in experiments where feedback on speech production is needed. In Section II, the LDA analysis and the making of the tool are described. Section III discusses the experiment that was carried out to test the functionality of the tool. In Section IV, the sections II and III are discussed, and some suggestions for further research are given. Finally, Section V summarizes the findings.

II. CREATING THE TOOL

This section is structured in the following way. First, the desired functionality of the tool and the procedure for making it will be described. Second, the stimuli that were used for the feature analysis will be described and discussed. Then, the LDA analysis that was carried out to find the most informative features for vowel categorization will be presented and interpreted. Subsequently, the segmentation procedure and the analysis of its performance will be discussed. Finally, the working of the tool will be described and illustrated.

A. Aim and procedure

The aim of the tool is to give feedback on the productions by Dutch natives of a test set of target words, which consists of five minimal pairs: *fan-fen*, *ham-hem*, *jam-gem*, *man-men* and *pan-pen*. In order to do so, first the target word has to be presented, then the utterance has to be recorded, segmented and analyzed, and finally, a form of visual feedback has to be presented.

To make the tool, the following steps were taken. First, productions of the target words by ten English natives were recorded[1]. These utterances were analyzed in order to find the most meaningful features for discrimination of the vowels. For this, different LDA models were compared, each of which had a different combination of formant measurements as its predictive features. Then, *Praat*'s inbuilt segmentation function was tested against hand-segmented utterances. Finally, the feedback system was programmed and designed.

B. Input stimuli

Recordings were made of ten native British English speakers (five female and five male), producing the target words: *fan*, *fen*, *ham*, *hem*, *jam*, *gem*, *man*, *men*, *pan* and *pen*. Every speaker pronounced the words *fan*, *fen*, *ham*, *hem*, *gem* and *men* 10 times, *jam*, *man*, *pan* and *pen* were produced 11 times, and one extra time by speaker 1. This resulted in a total number of 1044 utterances.

For all utterances the raw sound files of the ten speakers were automatically segmented (see Section II.C). Al formants were measured with *Praat'*s standard formant measuring algorithm, which uses Burg's algorithm (Childers, 1978; Press et al., 1992) to compute the LPC coefficients (*Praat* manual: Boersma, 2010). *Praat*'s standard gender specific formants ceilings were used: 5000 Hz for male voices and 5500 Hz for female voices. F0 was measured with the standard *Praat* pitch function *Sound: To Pitch*, which uses auto-correlation (Boersma, 1993); small time steps (0.001 seconds), a pitch floor of 75 Hz (standard) and a pitch ceiling of 600 Hz (standard) were used. F0, F1, F2 and F3 were measured for (1) the mean of the whole duration of the vowel, (2) the 20%, 50% and 80% points of the vowel duration, (3) the mean of the 0.015 seconds around the 50% point of the vowel duration, (4) the mean of 50% of the total vowel duration centered around the 50% point. The measurements were checked for obvious formant miscalculations[2]. Of the 1044 utterances, there were 150 utterances that fell outside the 'normal' range for F0, 39 utterances for F1, and 31 utterances for F2. These miscalculations occurred mainly at the 20% and 80% measuring points. For the whole-duration measurements, no miscalculations were found for F1 and F2, nor for F0. Since the whole-duration formant measurement was used to create the target stimuli (see Section II.D.3), no utterances had to be removed because of obvious measurement errors.

Figure 1 contains the formants (measured over the whole duration of the vowels, in Hertz) of all utterances for all ten speakers, grouped for vowel and gender. Figure 1 shows that F1 and F2 are slightly lower for males than for females (mean F1 female = 820.11 Hz; mean F1 male = 724.34 Hz; mean F2 female = 1705.66 Hz; mean F2 male = 1573.06 Hz). It has been suggested that this effect is due to the fact that the vocal tract of men is bigger than the vocal tract of women (Simpson, 2009), which makes the resonation space bigger and therefore the formants lower. To assess whether these differences are significant, and whether there is a greater effect for F1 or for F2, the values were transformed to the ERBs scale (see Section II.F). T-tests showed that the difference for female vs. male voices (in ERBs) is significant for F1 ($t(1027.1) = -11.54$, $p < 0.001$) as well as for F2 ($t(841.4) = -11.996$, $p < 0.001$). The effect seems to be slightly bigger for F1 than for F2: the female-male ratio of the mean F1 is 1.07 (mean F1 female = 13.68 ERBs; mean F1 male = 12.77 ERBs) and female-male ratio of the mean F2 is 1.03 (mean F2 female = 19.54 ERBs; mean F2 male = 18.91 ERBs). This effect is lower as for example the effect of gender F1 and F2 in Portuguese as found by Escudero and

[1] This data-collection was done by Jana Krutwig.
[2] The following 'normal' values ranges were used: F0: 75-300 Hz; F1: 300-1200 Hz; F2: 1100-1900 Hz.

colleagues (2009). Moreover, Escudero et al. (2009) found a higher ratio for F2 (1.183) than for F1 (1.170), whereas we found a higher ratio for F1.

Figure 1. Vowels for women and men, plotted for F1 and F2 (in Hertz).

In Figure 2, the vowel categories for all ten speakers are shown: the left column shows the vowel categories for the women; the right column contains the vowel categories for the men. In general, the categories of the women seem to be a bit better separable than those of the men. This is consistent with previous findings that women speak more clearly than men (Simpson, 2009). Additionally, the distributions of the females seem to be spread over a bigger space than the male distributions. The finding that females have a larger vowel space than males is known in the literature (Simpson, 2009; Hillenbrand et al., 2001 (Figure 5); Escudero et al., 2009; all studies use the Hertz scale). Additionally, Figure 2 shows that some speakers mainly use F1 to discriminate the vowels (e.g. speaker 5 and speaker 9), whereas other speakers mainly use F2 (e.g. speaker 4 and speaker 7).

Figure 2. Vowels categories, plotted for F1 and F2 (in Hertz).
The left column contains the female speakers, the right column the male speakers.

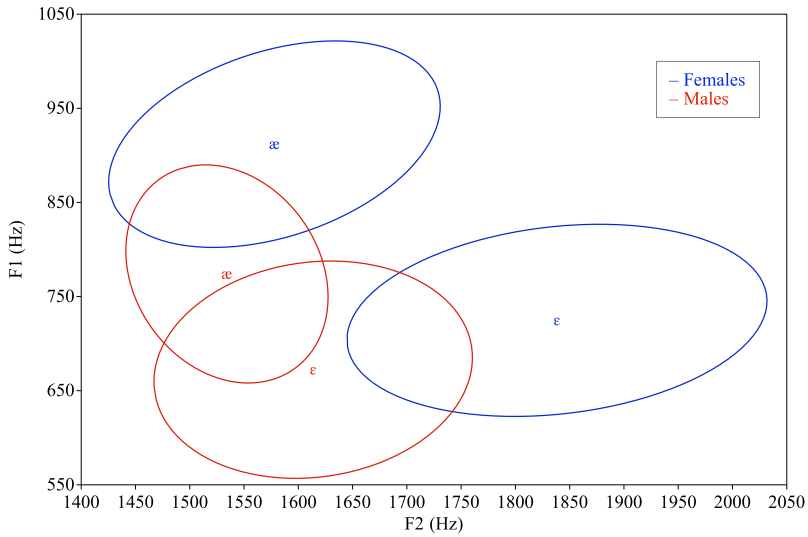


Figure 3. The distributions of the two vowels based on one standard error (sigma = 1), for the female speakers (blue) and the male speakers (red).

Figure 3 visualizes the distributions of the vowels for males and females for F1 and F2, based on one standard error (sigma). In this figure, we can see both above-mentioned observations very clearly: the female vowel categories are more separable than the male categories, and take up a larger part of the vowel space.

Finally, the relationship between the start-consonant and F1 and F2 was assessed. Figure 4 visualizes F1 and F2 of the different speakers, based on the start-consonant and the vowel. For some speakers (e.g. speaker 2, speaker 8, speaker 4), there is a clear influence of start-consonant on F1 and F2, whereas for some other speakers (e.g. speaker 5) this influence is less visible. Over all, it seems as if there is an effect from coarticulation on F1 and F2.

Figure 4. The vowels per speaker for each consonant.
The left column contains the female speakers, the right column the male speakers.

C. Segmentation

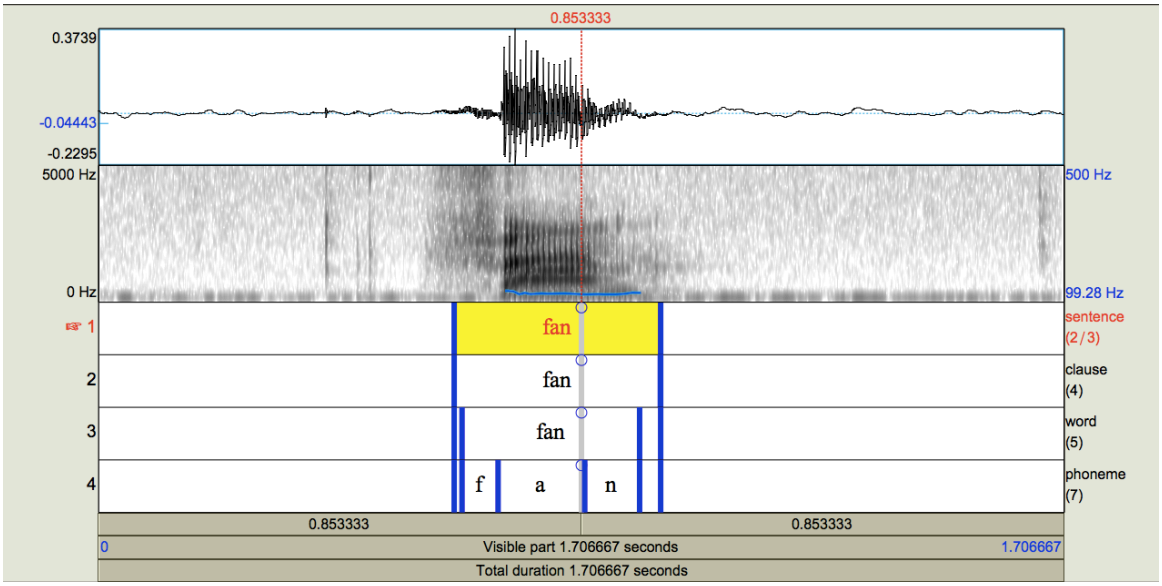
The segmentation was done with *Praat*'s inbuilt segmentation function. This function uses the sound file and a textgrid with the text that is uttered in the sound file.. *Praat* then uses a speech synthesizer of the specified language[3] to create a synthesized version of the provided text, which is then lined up with the provided sound file. Based on the lining up, the word and phoneme boundaries are placed. Because we only use CVC words in this project, segmentation was relatively easy To test how well this function works, the raw sound files of the utterances of speaker 5, 6, 7 and 10 were automatically segmented, and the results were compared to the results of the hand segmentation. The Pearson product-moment correlation coefficient between the duration of the vowels for the automatic segmentation and for the hand segmentation was 0.61. The correlation was also computed for the separate vowels, i.e. the correlation of the duration of /ɛ/ for the automatic and the hand segmentation, and the correlation of the duration of /æ/ for the automatic and the hand segmentation. The correlation for /ɛ/ was 0.55, and for /æ/ it was 0.42. Then, the correlation was assessed for all the different combinations of the start-consonants (/f/, /h/, /j/, /m/, /p/) and both the two vowels separately and for both vowels together. The results of this can be found in Table I. We conclude that there are no great differences between the consonants, only /p/ seems to be a bit easier than the other vowels. This was expected, because /p/ is a stop sound, and therefore easily separable from its neighboring sounds. Also, we see that /m/ and /h/ are the most difficult to separate the vowels from, and that this was more difficult for /ɛ/ than for /æ/. In Figure 5, two automatic segmentations are shown: Figure 5a. shows a good automatic segmentation, whereas Figure 5b. shows a failed automatic segmentation.

Table I. The correlations for the duration of the vowels for the automatic and the hand segmentation for the different combinations of start-consonant and vowels.

	/ɛ/	/æ/	both
/f/	0.692	0.784	0.758
/h/	0.472	0.634	0.778
/j/	0.832	0.816	0.774
/m/	0.568	0.648	0.792
/p/	0.793	0.674	0.834

[3] The default "English" was used.

a.



b.

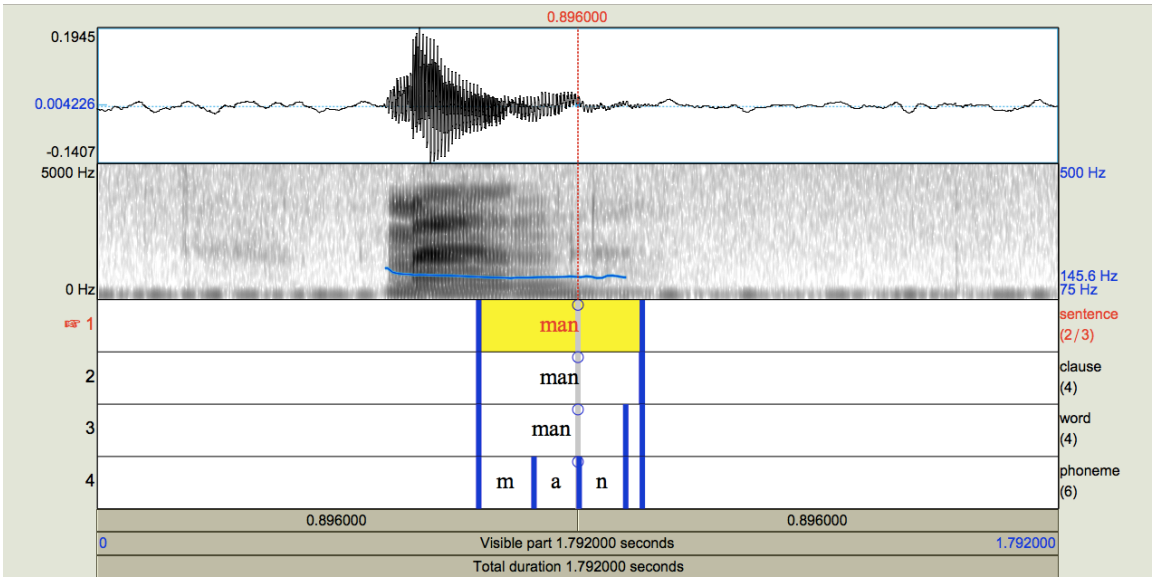


Figure 5. Examples of a good segmentation (a) and a failed segmentation (b).
The upper part of the pictures shows the sound wave and the spectrogram,
the lower part shows the segmentation.

However, it is possible that the duration of the vowels in the automatic and the hand segmentation are very similar but the segmentation was still incorrect. Therefore, it is more meaningful to look at the correlation for F1 and F2 for the automatic and hand segmentation. The formants were measured for the whole-duration measurement, with the same method as described in Section II.B. For F1, the correlation is 0.91; for F2, it is 0.94. The absolute differences in Hertz were also computed. For F1, the mean absolute difference between the hand and automatic segmentation was 17.15 Hz (sd = 57.87 Hz), and for F2 this was 9.26 Hz (sd = 60.93 Hz). We thus conclude that the segmentation function in *Praat* does sufficiently well to be used for our purpose, at least in the case that the formants are averaged over the whole duration of the vowel.

D. Features

To determine which features are the most informative for deciding whether an utterance is an /ɛ/ or an /æ/, a Linear Discriminant Analysis (LDA) was carried out with different sets of features.

1. Procedure

Many different ways of analyzing formants have been proposed in the literature, which could all potentially be implemented. The measurements that were tested are the following.
(1) Whole-duration: them mean formant values over the entire duration of the vowel. This is a common way of measuring vowel quality, and was used for example by Kartushina and colleagues (2015). It

takes into account information of both the middle of the vowel and its edges. It does not differentiate between these different moments, though: it simply takes the average.
(2) Measurements at 20%, 50% and 80% of the vowel duration. Hillenbrand et al. (1995) showed that their pattern classifier[4] was significantly more accurate when it used 20% and 80% measurements, as opposed to a single sample from the steadiest part of the vowel. However, it should be noted that they did not use cross-validation to test their results; these results could thus well be a consequence of over-fitting. Hillenbrand et al. (1995) conclude that the 20%-80% method takes into account spectral changes, and is therefore more informative. Nearey and Assman (1986) also find that the classification of their pattern recognizer improves for models that take into account spectral change as opposed to models that use only measurements from a fixed steady-state portion of the vowel; in this study, cross-validation is used. It is known that vowel identification by humans based solely on the vowel onset and the vowel offset (with a silent nucleus) is very good (e.g. Jenkins et al., 1983; Parker & Diehl, 1984). Moreover, the identification of 'gated vowels', i.e. vowels of which the onset and the offset is silenced, is poor (Assman et al., 1982). There thus seems to be a large amount of information in the vowel onset and offset; this argues for taking into account dynamic information of the vowel. Jenkins and Strange (1999) even argue that vowel identification cannot be seen as simply detecting a certain acoustic target, but is instead a process of "apprehending acoustic changes that specify the style of articulatory change that produced the specific vowel" (Jenkins & Strange, 1999). They argued already in the 80s (Jenkins et al., 1983) that speech perception research puts too much emphasis on vowels produced in isolation in a sustained manner; instead, vowels should be studied in their natural context. This is indeed how research in vowel perception has developed. However, in vowel production research, it is still common to use vowels in isolation (e.g. Kartushina, 2015). Even though it remains the question how much of the findings in speech perception can be generalized to speech production, it still seems likely that in vowel production research the emphasis should also be on vowels in their natural context.
(3) Measurement at 50% of the vowel duration, plus the difference between the 50% and the 20% measuring point, minus the difference between the 80% and the 50% measuring point. This method was chosen according to the *production undershoot model* of Stevens and House (1963). They investigated the effect of different consonant surroundings on different vowels, and found that they effect of coarticulation was the greatest on F2: F2 moved towards more centralized vowels (Stevens & House, 1963; Hillenbrand & Nearey, 1999). The *production undershoot model* hypothesized that in vowel production, people try to reach an articulatory target, namely the vowel frequency of the vowel as it would be produced in isolation. However, because of the articulatory constraints that are posed by the surrounding consonants, these targets are mostly not reached. To model production undershoot, we measured the 50% point, corresponding to which degree the target was reached, the difference between 20% and 50%, because the 20% point still has the information of the preceding vowel, and the difference between 80% and 50%, because that point already contains coarticulatory information about the succeeding vowel. Therefore, adding the 50%-20% point to the target 50% point, and subtracting the 80-50%[5] point from the target 50% point would give us the optimal information about the intended vowel with regard to its context.
(4) Measurement at 50% of the vowel duration. This method was chosen because it is also often used in formant measuring. However, this method is error-prone, because it measures only one point, and this point might be accidentally measured wrongly.
(5) The mean measurement of the 0.015 seconds around the 50% point of the vowel duration (0.0075 seconds on each side). This was taken to account for the error-proneness of the measurement at 50% of the vowel duration: it still stays very close around the mid point of the vowel, but is slightly less prone to measurement errors.
(6) 50% of the total vowel duration, centered around the 50% point. This method takes into account as little formant information of the neighboring consonants as possible, while taking as much as possible of the vowel. This method thus tries to rule out the context information, to get a 'clean' representation of the vowel.
(7) Mel-Frequency Cepstral coefficients (MFCC), 1 to 12. An MFCC analysis gives a number of coefficients, which represent the spectrum of a sound without making use of formant analysis. This analysis first transforms the spectrogram in a Mel Spectrogram, representing an "acoustic time-frequency (on a Mel frequency scale) representation of the sound" (*Praat* manual: Weenink, 2014). Then, this spectrogram is divided into (increasingly bigger) windows, and a Discrete Cosine Transform (Davis & Mermelstein, 1980) is computed for the spectral values in the windows. This method was chosen because it has been suggested that a more complete representation of the spectral slope

[4] They used a quadratic discriminant analysis (Johnson & Winchern, 1982).
[5] Which is the same as adding 50%-80%.

leads to better discrimination than a representation based solely on formants (Zahorian & Jagharghi, 1993). In the 60s and 70s, Pols and colleagues showed that a Principal Component Analysis of the spectral shape of vowel could be plotted such that it resembles formant plots (Pols et al., 1967). They also showed that this spectral shape representation yielded similar results for automatic classification with a vowel-identification algorithm as a formant representation (Klein et al., 1970). Zahorian and Jagharghi (1993) used a further developed model: Discrete Cosine Transform Coefficients, which is very similar to MFCCs, and shows that this representation gives better results for automatic classification than formant representation. However, they did not use cross-validation.

In addition, Nearey (1989) suggests that speaker-extrinsic information, i.e. relating the vowel utterance to the entire vowel system of the speaker, is important for vowel identification. Moreover, Ménard and colleagues (2002) suggest that F0 is used for perceptual normalization and for the disambiguation of vowels with similar F1 and F2, and that the F0–F1 distance predicts perceived vowel height (Ménard et al., 2002; Kartushina et al., 2015). In Kartushina et al. (2015), feedback therefore consisted of F1–F0 and F2–F0. In the LDA analysis as performed in this project, this is tested through adding F0 as a feature. The formants and pitch were measured in the same way as described in Section II.B.

The LDA model was run and tested with several different sets of features, to determine for which feature set it would make the best separation between the two vowels. In addition to the seven different ways of representing the spectrogram, some other features were taken into account: gender, start-consonant, and end-consonant; these will be called 'non-formant measures'. The features were tested in different combinations:

1) The formant measures were tested by combining the different measures for F1 and F2:
    a. With non-formant measures, with F0 and with F3
    b. Without non-formant measures, with F0 and with F3
    c. With non-formant measures, without F0 and with F3
    d. Without non-formant measures, without F0 and with F3
    e. With non-formant measures, with F0 and without F3
    f. Without non-formant measures, with F0 and without F3
    g. With non-formant measures, without F0 and without F3
    h. Without non-formant measures, without F0 and without F3 (i.e. only F1 and F2)
2) The MFCC analysis was tested by combining:
    a. Coefficients 1-12 with non-formant measures
    b. Coefficients 1-12 without non-formant measures
    c. Coefficients 2-12 without non-formant measures

Then, based on the results of the LDA performance for the above feature sets, some other sets were tested:

3) Whole-duration method for F1 and F2, plus start-consonant
4) Whole-duration method for F1 and F2, plus end-consonant
5) Whole-duration method for F1 and F2, plus gender
6) Whole-duration method for F1 and F2, plus start-consonant and end-consonant
7) Whole-duration method for F1 and F2, plus start-consonant, and consonant, 50-20% and 80-50% for F2
8) Whole-duration method for F1 and F2, plus start-consonant, and consonant, 50-20% and 80-50% for F2 and for F1.

2. Results

First, a correlation plot was made (Figure 6), to see what the correlations between the different formant measures are. It shows that for all speakers together, the different formant measures all correlate quite highly, as can be seen by the dark blue big dots. The correlations are high for all formants, but highest for F0. The correlations are especially high for *whole-duration* with *50%-duration-around-the-middle*, for *whole-duration* with *0.015-sec.-around-the-middle*, and for *0.015-sec-around-the-middle* with *50%-duration-around-the-middle*.

In addition, there is a high correlation between the F1 and F2 measures for *whole-duration* and *50%-duration-around-the-middle* (F1: 0.96; F2: 0.94). This high correlation was also found in the hand-segmented data (F1: 0.96; F2: 0.89). This argues for good automatic segmentation, because if the vowels are poorly segmented, the *whole-duration* methods will take the surrounding consonant information into account, and therefore the formant values will change. The *50%-duration-around-the-middle*, however, is less likely to take consonant information into account in case of bad segmentation, because it only uses 50% of the duration, and will therefore be less close to surrounding consonant information.

In Appendix A are the correlation plots for the ten individual speakers. It was found that for some of the individual speakers there were negative correlations for some formants, e.g. speakers 2, 4 and 8 have a very strong negative correlation for F1 and F2. These negative correlations did not show up in the correlation plot that averages over all speakers.

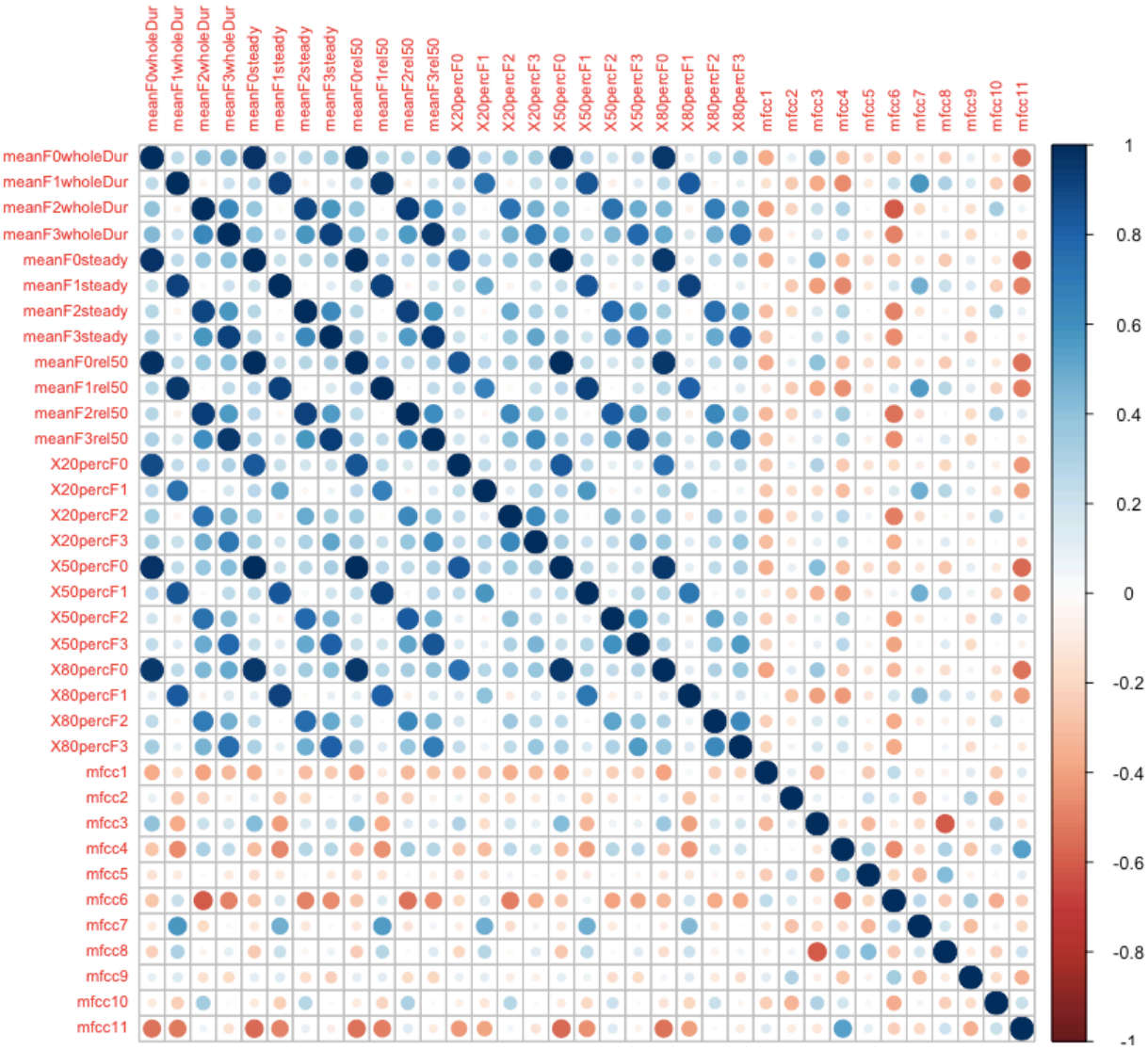


Figure 6. Correlation plot of the different measure methods[6]

In Tables II–IV, the percentages correct for the LDA's with the different sets of features are shown. These percentages are computed with tenfold *cross-validation* (after Boersma, 2016), in which the model was trained on nine out of the ten speakers, and tested on the tenth speaker. This was done for all ten speakers, which gave ten percentage correct scores. These scores were then averaged, which resulted in the scores that are listed in Tables II–IV. The best model from Table II (*whole-duration, g.*: 81.82%) has only F1 and F2 measured with the *whole-duration* method. This model was therefore extended with each of the non-formant measures (Table IV). As can be expected with a correlation plot like the one in Figure 6, for many of the models *R*[7] gave the warning that 'some variables are collinear'. Collinear variables are explanatory variables with a linear relationship, meaning that their explanation of the variance on the data correlates. There was no collinearity for the models in table IV (except for model 6).

[6] meanF*x*steady = 0.015-sec-around-the-middle; meanF*x*rel50 = 50%-duration-around-the-middle
[7] The LDA analysis was run in R.

Table II. Mean percentage correct of the LDA on the different test sets for the formant measures, for the different sets of features.

Formant measures	Combination of features							
	a. With non-formant measures, with F0 and with F3	b. Without non-formant measures, with F0 and with F3	c. With non-formant measures, without F0 and with F3	d. Without non-formant measures, without F0 and with F3	e. With non-formant measures, with F0 and without F3	f. Without non-formant measures, with F0 and without F3	g. With non-formant measures, without F0, without F3	h. Without non-formant measures, without F0, without F3
Whole-duration	78.44%	78.72%	80.09%	77.96%	81.14%	78.72%	81.82%	79.88%
20%, 50%, 80%	76.5%	75.44%	77.29%	76.22%	79.47%	77.23%	78.64%	78.34%
50% + 50%-20% + 80%-50%	76.5%	75.44%	77.29%	76.22%	77.87%	75.79%	77.28%	78.90%
50%	71.24%	70.31%	70.55%	72.65%	74.58%	74.01%	74.20%	75.45%
0.015 sec around 50%	76.12%	76.96%	76.14%	76.62%	79.39%	79.56%	76.14%	78.64%
50% of duration, around 50%	75.38%	73.05%	75.25%	75.26%	78.57%	78.08%	75.25%	77.76%

Table III. Mean percentage correct of the LDA on the different test sets for the MFCC measures.

MFCC	% Correct
a. Coefficients 1-12 with non-formant measures	79.02%
b. Coefficients 1-12 without non-formant measures	79.86%
c. Coefficients 2-12 without non-formant measures	76.70%

Table IV. Mean percentage of the LDA on the additional models.

Additional models	% Correct
3. Whole-duration method for F1 and F2, plus start-consonant	82.77%
4. Whole-duration method for F1 and F2, plus end-consonant	80.55%
5. Whole-duration method for F1 and F2, plus gender	79.40%
6. Whole-duration method for F1 and F2, plus start-consonant and end-consonant	82.77%
7. Whole-duration method for F1 and F2, plus start-consonant, and 50-20% and 80-50% for F2	83.25%
8. Whole-duration method for F1 and F2, plus start-consonant, and 50-20% and 80-50% for F2 and for F1.	82.10%

From Table IV, model 3 and model 7 score the highest (model 6 was eliminated due to collinearity); therefore these models were compared to see whether they differ significantly. Since *R* does not allow model comparison for LDA models, three Generalized Linear Mixed-Effects Models (GLMER) were fitted, with the same features as model 3, model 7 and model 8 from table IV. An LDA takes into account singular interactions, whereas a GLMER does not do this by default; this was therefore specified. Then, the models were compared with a Chi Squared test in *R*'s *anova* function. The results from this comparison were inconclusive. The comparison of model 3 and model 7 has a lower BIC value for model 3 (model 3: 581.40; model 7: 624.22), but a lower AIC value for model 7 (3: 497.23; 7: 465.79). Models 3 and model 1.g (whole-duration) from Table II were also compared, to see whether the addition of the start-consonant was significant. This comparison did give conclusive results: for both AIC (model 3: 710.86; model 1.g (whole-duration): 497.23) and BIC (model 3: 735.62; model 1.g (whole-duration): 581.40), model 3 was preferred.

Upon further examination, the loadings of the LDA for model 7 showed an unexpected result: the loading for 50%-20% had a negative sign and the loading for 80%-50% had a positive sign, indicating that the 50%-20% would have to be subtracted, and the 80%-50% would have to be added to the 50% value. This is the opposite of the prediction that the *production undershoot model* (see Section II.D.1) makes. It is therefore not entirely clear what this model does.

3. Interpretation

Based on the results described above, model 3 (Table IV) was chosen to be the most reliable model, because there was no good indicator to choose between model 3 and model 7, thus the simplest model is preferred. This means that the tool will use the whole-duration measurement of F1 and F2 to compare the natives' productions with the non-natives' productions in order to give feedback.

The fact that model 7 also performs very well, strengthens the idea that the onset and the offset do play an important role in vowel perception. In how far this should be taken into account for vowel production, remains an interesting question; this is discussed in Section V.

It was unexpected that adding F0 did not improve the performance of the model. It could be the case that the mean F0 for the women and the men did not differ so much in out dataset. Also, we saw a correlation between F0 and F1/F2: possibly, F0 did not add information to F1/F2 anymore.

E. Type of feedback

Additionally, the question arises as to which sort of visual feedback would be most effective. From the motor learning literature we know that precise, quantitative feedback is more helpful than more general feedback (Schmidt-Lee, 1999). However, previous studies have shown that skilled musicians benefited more from general feedback than from detailed feedback (Brandmeyer, 2011). Something comparable might also be the case for feedback on speech production. However, due to limited time, this was not tested in the current project. Therefore, the recommendations from Öster (1997: 145) on 'Auditory and visual feedback in spoken L2 learning' were taken into account [sic]:

- *The visual pattern must be natural, logical and easily understandable*.
- *The aid should provide a contrastive training, that is, the correct model of the teacher and the deviant production of the learner are shown simultaneously and compared with each other.*
- *The aid should provide a flexible, individual, and structural speech and voice training and give an objective evaluation of training results.*
- *The visual feedback of the voice and the articulation should be shown without delay.*
- *The aid must be acceptable to the teacher as well as to the learner, which means that the aid must be attractive, interesting, easily comprehensible, easy to handle, and motivating.*

We thus hypothesize that a simple, graded feedback system works the best. It will probably be the most encouraging if you do not only see where you were wrong, but also how close you are to the actual utterance. Because of technical restrictions, the visual feedback is shown with a short delay (about 1 second).

F. Working of the tool

In this section, the working of the tool is described. The tool starts with an information form, in which the experimenter fills out (amongst other things) the gender of the participant. If a participant does not identify with either gender (or belongs to another category), the experimenter can make a choice for either the female or the male model based on the perceived quality of the participant's voice. The participant is then presented with a short explanation of how the tool works. Then, one of the ten target words is randomly picked[8] and presented orthographically, and the participant pronounces this word. The utterance is recorded, segmented with *Praat*'s inbuilt segmentation function, and the mean F1 and F2 over the whole duration of the vowel is calculated with *Praat's* standard formant measuring algorithm. The F1 and F2 values are then converted into ERBs. This scale takes into account the working of the human cochlea. Because the distance between hair cells in the cochlea increases from higher to lower frequency ranges, frequencies that have the same distance in Hertz can be perceived as more similar in one frequency range than in another. In the ERB frequency scale equal distances correspond to perceptually equal distances.

To compute the accuracy of the rater's utterance, the Mahalanobis distance between the utterance and the relevant native distribution is measured in the F1/F2 space (in ERBs). For both genders, there are ten possible distributions: two vowels times five start-consonants. The Mahalanobis distance takes into account the shape of the distribution, because it measures how many standard deviations the production is away from the mean of the native distribution along each of its principal component axes (Kartushina et al, 2015). It thus differs from simply taking the Euclidean distance: this method would not be able to distinguish between two points that are equally distant from the mean of the distribution, but one of which would fall nicely into the distribution, and the other one would be quite far away (due to the shape of the distribution).

[8] A within-blocks randomization strategy was used, i.e. every one of the ten words has to be presented once before a word is repeated.

The feedback consists of a screen, as in Figure 7, in which the target is shown in blue, together with a flower, and the target of the other vowel category is shown in silver for reference. The participant's utterance is plotted in green if it was correct, and in red if it was incorrect. The score is kept. The utterance is considered correct if its Mahalanobis distance is not larger than 1 standard deviation (in the case of the females) or 0.5 standard deviations (for the males). This difference was made because of the different shapes of the distributions: the male distributions overlap considerably more, therefore if we set the same threshold for women and men, the men would have an easier task and therefore get less constructive feedback. The feedback thus consists of two parts: gradual feedback – the participant's utterance is presented relative to the target vowel, and binary feedback – the dot is either green or red. This combination was chosen because both ways of feedback have their benefits. Graded feedback gives the participant more information than simply correct/incorrect, and is therefore more stimulating; however, it does not take into account the shape of the distribution. The binary feedback, however, does take into account the shape of the distribution. Moreover, it does not add too much information (the feedback stays simple and intuitively interpretable), and it adds a motivating element. The tool thus seems to meet the criteria posed by Öster (1997).

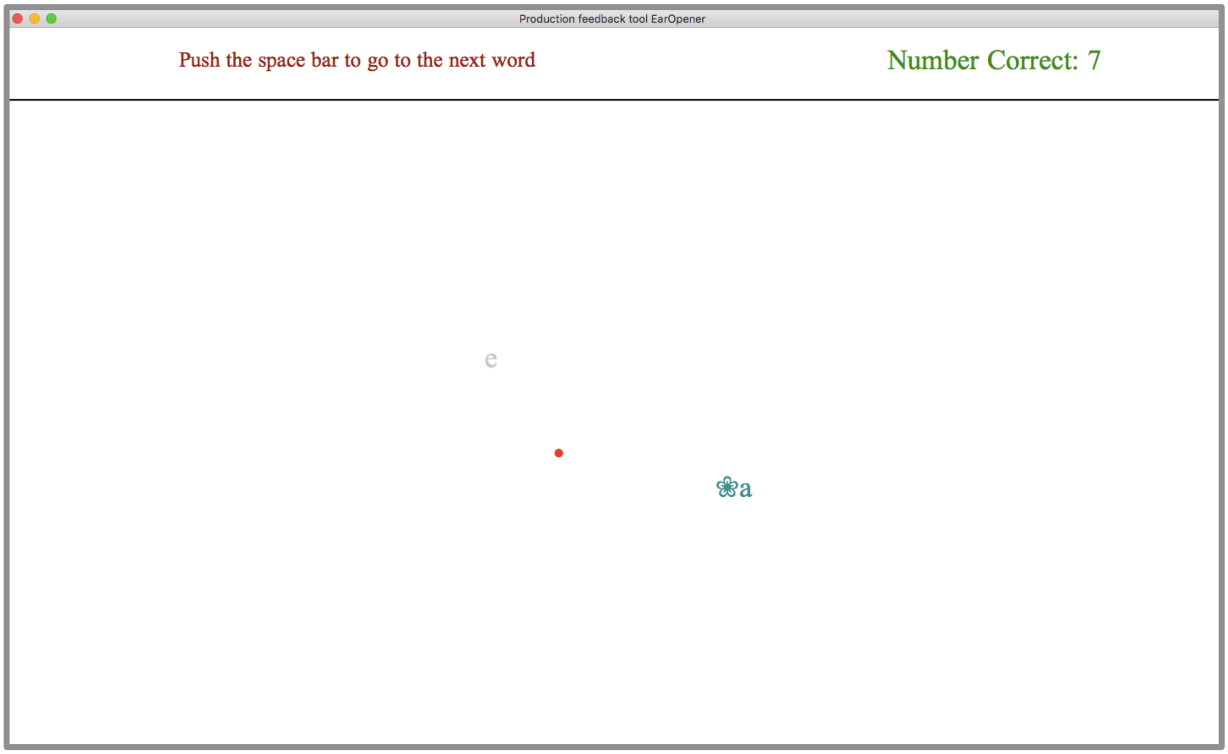


Figure 7. An example of the feedback screen.

The axes of the F1/F2 spaces are not shown in the feedback screen, because they differ according to the consonant. Because we take into account coarticulation, a different distribution is used for each consonant. This means that the F1 and F2 values of the target vowel change per consonant (Table V). To prevent the target vowels from showing up at a different point in the visual field for every consonant, the axes were adapted in such a way that /ɛ/ is always presented at the upper left 25% of the screen, and /æ/ is always presented at the lower right corner of the screen. The axes are thus dependent on the mean values for F1 and F2 of the target vowels; the formulae used for this can be found in table VI. As discussed in Section II.B, the vowel space for males is smaller than for females. Because axes of the feedback are relative to the distance between the target vowels, this would cause the feedback screen of the men to be smaller than that of the women. To prevent the productions of the participants from falling outside the plotting area – which means that no feedback can be given, the axes for the male model were made larger according to the ratios between female-male F1 and F2 as given in Section II.B (1.05 for F1 and 1.04 for F2). These values can be changed according to the experimenter's experience: if (s)he observes that the produced utterance often cannot be plotted, the

values can be set higher. Participants should be instructed that it could be the case that their utterance is not plotted, and that this means that their production was not analyzable.

Table V. The means of the target vowels per consonant for both genders

	Females				Males			
	/ɛ/		/æ/		/ɛ/		/æ/	
	F1 (ERBs)	F2 (ERBs)	F1 (ERBs)	F2 (ERBs)	F1 (ERBs)	F2 (ERBs)	F1 (ERBs)	F2 (ERBs)
/f/	11.83	20.10	14.57	18.73	11.95	18.76	13.01	18.51
/h/	13.17	19.92	14.82	18.63	12.38	19.02	13.74	18.59
/j/	12.28	20.25	14.04	19.33	11.66	19.43	12.65	19.11
/m/	12.87	20.33	14.34	18.88	12.14	19.13	13.19	18.62
/p/	13.77	20.29	14.94	19.02	12.98	19.21	13.81	18.73

Table VI. Formulae used to compute the values of the axes

Axis	**Females**	**Males**
Xmin	Mean F1 of /ɛ/ - 2 * (mean F1 of /æ/ - mean F1 of /ɛ/)	Mean F1 of /ɛ/ - 2.1 * (mean F1 of /æ/ - mean F1 of /ɛ/)
Xmax	Mean F1 of /æ/ - 2 * (mean F1 of /æ/ - mean F1 of /ɛ/)	Mean F1 of /æ/ - 2.1 * (mean F1 of /æ/ - mean F1 of /ɛ/)
Ymin	Mean F2 of /æ/ - 2 * (mean F2 of /ɛ/ - mean F1 of /æ/)	Mean F2 of /æ/ - 2.08 * (mean F2 of /ɛ/ - mean F1 of /æ/)
Ymax	Mean F2 of /ɛ/ - 2 * (mean F2 of /ɛ/ - mean F1 of /æ/)	Mean F2 of /ɛ/ - 2.08 * (mean F2 of /ɛ/ - mean F1 of /æ/)

III. TESTING THE TOOL

To test the tool, an experiment was carried out in which ratings of the tool for pronunciations of the target words by Dutch natives and English natives were compared with ratings by English natives.

A. Participants

Three participants (from now on: raters) were tested. They were all British English native speakers, who are studying in Amsterdam. All 3 of them moved to the Netherlands eight months before the experiment. Rater 1 grew up partly in Southern England (Surrey, from 1–10 years) and partly in Scotland (Inverness, from 10–18 years; Edinburgh, from 18–22). Rater 2 grew up in Northern England (Preston), and lived in Manchester for the last years. Rater 3 grew up in Portsmouth (South England), and then lived in London for the last four years before moving to the Netherlands.

B. Stimuli

The sounds for the experiment were recorded during a perception experiment, in which Dutch native speakers were trained on the /ɛ/–/æ/contrast for 4 days. On the first day they did a production pre-test, in which they produced the ten target words of the tool three times. On the fourth day, an identical post-test was performed. The data of 28 participants of this experiment were used for the current experiment. From the pre-test, the second and the third utterance were used, and from the post-test, the first and the second utterance were used. 42 sounds were removed because they were silent: participants had to press the record button themselves, and they sometimes sneezed during the recording, or were simply too late. This resulted in 1076 utterances by Dutch speakers. The sounds were recorded at 48.000 Hz, except for the pre-tests of two speakers, which were recorded at 44.100 Hz, and were therefore resampled. The intensity was scaled: the new maximum mean intensity was 55.71 dB and the new minimum mean intensity was 34.09 dB.

To prevent habituation to the Dutch accent, which could result in a shifted identification boundary between the two vowels, 112 sounds of native English speakers were added. These sounds came from the database of ten speakers that was used to create the tool with. For each of the target words, one utterance of each of the ten speakers was chosen (100 words), then every target word was added once (each word by a different speaker), and finally two random words were added (*jam* by speaker 9, and *man* by speaker 2). This resulted in a total of 1188 sounds: 1076 by Dutch natives and 112 by English natives; a bit more than 10 percent of the utterances were produced by English natives. The stimuli were not adapted for duration.

The raters were randomly presented 1188 times with occurrences of the 1188 sounds. This means that each rater heard a subset of the total set of words: rater 1 heard different 749 words, rater 2 heard 768 words, and rater 3 heard 742 words[9]. Some of these words they heard only once, some of them were repeated (max. 7 times). This allowed assessing the consistency of the ratings.

After the rating task, the raters did a short task to test their identification boundary between /æ/–/ɛ/. They were presented with a continuum between /ɛ/–/æ/: a morphed spectrum of 11 instances between the recordings of *vat* to *vet* by speaker 6. The utterances were normalized for duration (632 ms) and amplitude (RMS) for the 3 intervals (CVC) separately. Each of the 11 stimuli was presented ten times.

C. Procedure

The raters were instructed that they would hear utterances of English words produced by Dutch natives, and that they were supposed to rate the pronunciations. They were not told that some of the utterances were produced by English native speakers. The raters had to choose between 7 categories: *poor a, okay a, good a, good e, okay e, poor e* and *another vowel.* The categories were presented according to the word that was pronounced, i.e. if they heard the word *fan*, the categories were *poor fan*, *okay fan, good fan, good fen, okay fen, poor fen* and *another vowel*. The raters were asked to rate critically, and to try to pay attention only to the vowel quality and not to duration. As mentioned above, they rated 1188 instances; there was a break after every 50 instances. For rater 1 and 2, *Sennheiser HD 419* headphones were used; and for rater 3, *Sony Dynamic Stereo Headphones MDR-7506* were used.

After the rating task, there was a short break, followed by the morphed continuum task. The raters were instructed to press a key associated to the word they heard. The raters completed the

[9] The expected number of words would be 1188*(1-1/exp(1)) = 750.96.

tasks in different paces: rater 1 needed 55 minutes, rater 2 65 minutes, and rater 3 80 minutes. Each rater received a voucher of €10,- for their participation.

Rater 1 noticed that some of the utterances were produced by native English speakers. Rater 3 reported that sometimes the vowels were quite long, and then the vowel shifted from one category into the other.

D. Results

The rating data was recoded in the following way. If the intended vowel was /æ/, *good a* was coded as 1, *okay a* as 2, *poor a* as 3, and the other four categories[10] as 4, since these four categories all meant that another vowel was perceived. If the intended vowel was /ɛ/, *good e* was coded as 1, *okay e* as 2, *poor e* as 3, and the other four categories[11] as 4. All correlations in this section were computed with a Pearson product-moment correlation coefficient, i.e. the 1–2–3–4 coding was regarded as linearly ordered.

1. Inter Rater Reliability

To assess the inter-rater reliability (IRR), the *Intraclass correlation coefficient (ICC)* was computed. This method is more reliable than computing the percentage agreement between raters (Hallgren, 2012), because it takes into account the possibility that raters' agreement was due to chance. The IRR was computed with a two-way agreement average ICC[12]. The ICC was computed to assess both the consistency among the three speakers and the agreement of the different ratings within one speaker, both for the dataset of the English natives and the dataset of the Dutch natives. ICC values range from -1 to +1: high positive ICC values indicate high agreement, whereas high negative values indicate systematic disagreement.

To compute the ICC, a subset of the data was used, namely the subset of the utterances that were rated by all three raters at least once. The ICC for the ratings on the English natives dataset was 0.754. The ICC for the ratings on the Dutch natives dataset was 0.645. The raters thus seem to agree more on the utterances by the English natives. However, this difference could also be due to the different sizes of the datasets (39 utterances for the English natives vs. 259 for the Dutch natives). To assess the consistency within the speakers, the utterances that were rated more than once were taken, and the first two ratings were compared. For rater 1, this gave an ICC of +0.82 for the data of the Dutch natives, and an ICC of +0.387 for the data of the English natives. For rater 2, an ICC of +0.745 was found for the data of the Dutch natives, and an ICC of +0.913 for the English natives. For rater 3, the ICC was +0.834 for the Dutch natives, and +0.46 for the English natives.

From the above, we conclude that in general, the raters seem to agree quite nicely; they agree even more on the data of the English natives, which was expected. However, even with these high ICCs, the correlations between the raters are still only moderate: the correlation between rater 1 and 2 is 0.36 for the Dutch natives and 0.36 for the English natives, the correlation between rater 1 and 3 is 0.47 for the Dutch natives and 0.58 for the English natives, and the correlation between rater 2 and 3 is 0.36 for the Dutch natives and 0.65 for the English natives. Moreover, the raters are quite consistent on their ratings of the Dutch natives, but that rater 1 and 3 are not so consistent on their ratings of the English natives, whereas rater 2 is very consistent on her ratings of the English natives.

2. Correlation for the utterances by Dutch natives

First, the correlation between the raters' ratings and the Mahalanobis distance as measured by the tool was computed for all the stimuli that were rated at least once by one rater. This means that some utterances were rated only once, and some utterances were rated as may as 9 times (by different raters). In the case of multiple ratings by one rater, the mean was taken. The correlation was 0.31, which is quite low. If we only take into account the different genders for all three raters together, we see that for the females, the correlation is 0.37, and for the males it is 0.24. Split up per rater, the correlations were 0.13 (rater 1), 0.39 (rater 2), and 0.34 (rater 3). If we look at the difference of the correlations between the vowels, i.e. how well the tool's Mahalanobis calculation correlates with raters' ratings on the two different vowels, we see a large difference: the correlation is 0.49 for /æ/, but only 0.005 for /ɛ/, which indicates that something surprising happens for /ɛ/.

Because of the big difference between rater 1 and rater 2 and 3, we took a closer look at rater 1. He is a phonetician by training, so it could be the case that he hears the differences in acoustic quality better

[10] *good e, okay e, poor e* and *another vowel*

[11] *good e, okay e, poor e* and *another vowel*

[12] A two-way model was used, because there was only one pool of raters that rated the dataset. The type was 'agreement', because the rating should have the same absolute values in order to be consistent (i.e. consistency is not enough). The average-measure was taken, because the average of the ratings is used for hypothesis testing. (Hallgren, 2012).

than the other raters. To test whether rater 1 does something different, we computed the mean rating on the utterances by the Dutch natives for all three raters. Rater 1 indeed seems to rate quite a bit lower (mean rating pp.1: 2.03; mean rating pp.2: 2.51; mean rating pp.3: 2.27). A one-way ANOVA showed that the raters' means differed significantly: $F(2, 774) = 13.07$, $p < 0.001$. A post-hoc Tukey test was performed, to see which means differed from each other. It was found that all raters differ significantly from one another (2-1: $p < 0.001$; 3-1: $p = 0.031$; 3-2: $p = 0.027$).

Then, the correlation between the Mahalanobis distance as computed by the tool and the ratings that all three raters agreed upon for the first rating was computed. This correlation was 0.39. Again, a comparison between the two vowels was made: the correlation between the Mahalanobis distance and the ratings that all three raters agreed upon in their first rating for /æ/ was +0.59; for /ɛ/ it was -0.02.

The above results are somewhat confusing. The overall correlation of +0.37 (for the ratings of the female Dutch native speakers) is in the same range as the agreement among the different raters. The correlation with rater 2 was even +0.39. However, there was a big difference in the ratings of the two different vowels: correlations for /æ/ were high (+0.49 for all stimuli, and even +0.59 for the stimuli that were agreed on by all three raters on their first rating), whereas correlations for /ɛ/ were very low to negative.

3. Correlation for the utterances by English natives

The correlation between the ratings of the raters and the tool's Mahalanobis distance for the productions by the English natives, on all stimuli that were seen at least once by one rater, is -0.10. Split this up for gender, the correlation for male and female English native speakers are almost the same: for females the correlation is -0.10, for males it is -0.11. The correlations per rater on both female and male English speakers are -0.12 for rater 1, +0.03 for rater 2, and -0.13 for rater 3.
A comparison between the two vowels gave the correlation was -0.04 for /æ/, and -0.14 for /ɛ/.

Again, we had a look at whether rater 1 gave different ratings than raters 2 and 3. In this case, rater 1 seems to rate them a bit higher (mean rating pp.1: 1.68, pp.2: 1.46, pp3: 1.46). A one-way ANOVA, however, showed that these differences were not significant ($F(2, 114) = 0.64$, $p = 0.529$). The difference that was found for the Dutch natives, that led to the speculation that rater 1 might be rating more critically because of his phonetics background, therefore does not seem to hold. However, this rater was the only one to report that he heard that there were also some utterances by English natives: it could be that he therefore rated those higher.

Then, the correlation between the Mahalanobis distance as computed by the tool and the ratings that all three raters agreed upon in the first rating was computed. This correlation was -0.04. Again, a comparison between the two vowels was made: the correlation between the Mahalanobis distance and the ratings that all three raters agreed upon in their first rating for /æ/ could not be computed, because all first ratings that were agreed on for /æ/ were 1, so the standard deviation could not be computed. For /ɛ/, the correlation was -0.11.

These results are again unexpected. Since the raters were native speakers that are supposed to rate productions of other native speakers as 'correct', and the tool computes the Mahalanobis distance to a distribution of native speakers, the ratings and the Mahalanobis distance are expected to correlate. However, the correlations that are found are either very low or even negative. Therefore, we will have a closer look at the ratings of the productions of the English natives.

4. The ratings of the productions of the English natives

In this section, the ratings of the raters on the productions of the English native speakers will be examined more closely. A couple of unexpected results were found. Since the raters were native English speakers, it was expected that they would rate the utterances of the native English speakers as very good. As Figure 8 shows, this was indeed the case for /æ/, but not for /ɛ/. Apparently, the raters often do not perceive an /ɛ/ when this was the intended vowel. To our knowledge, this kind of asymmetry has not been reported in the literature.

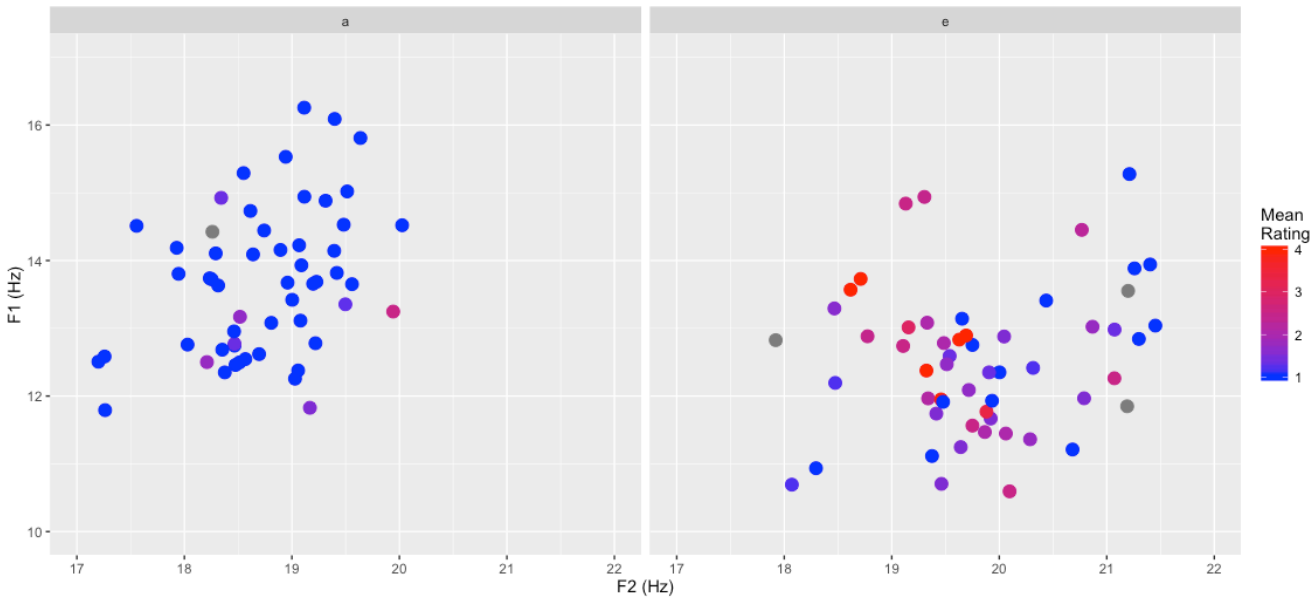


Figure 8. Ratings of all 3 raters for natives utterances of /æ/ (left) and /ɛ/ (right).

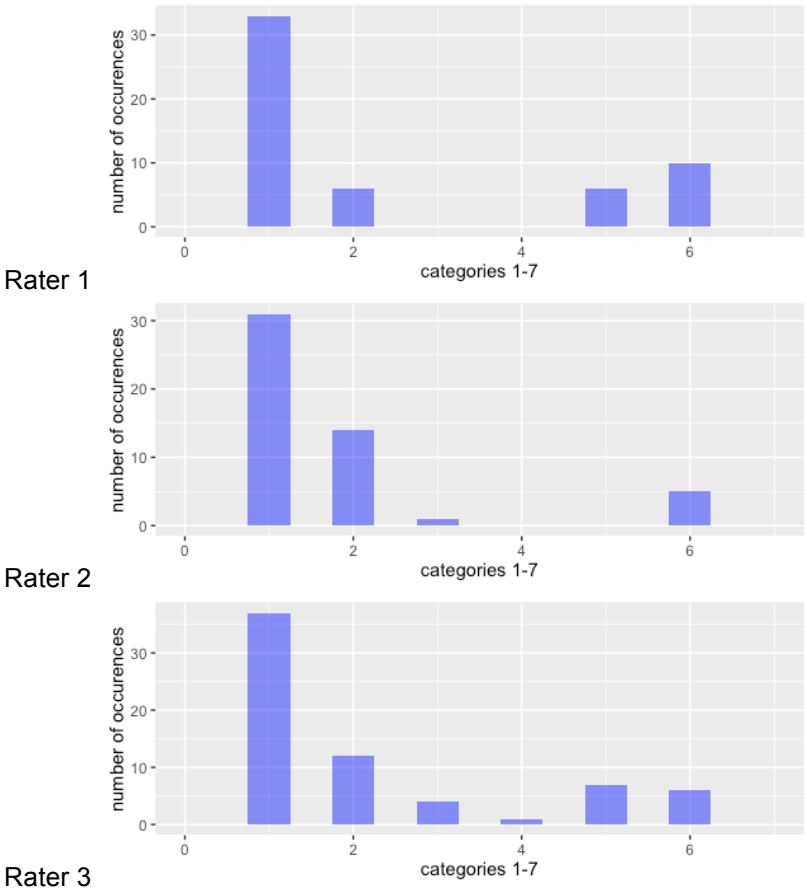
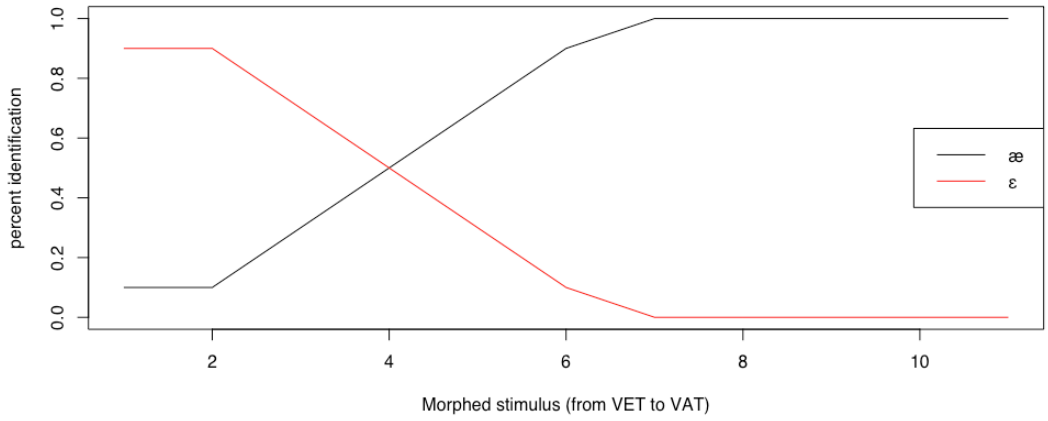
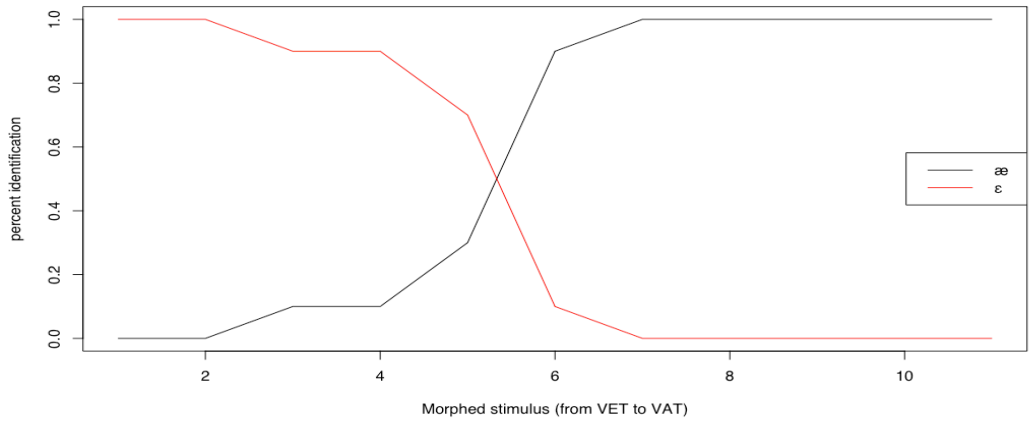


Figure 9. Histogram of the number of ratings per category on the native speakers' productions of /ɛ/ by the three raters.

To find out what was going on, a histogram was made for the original 7 ratings[13] on the native utterances that were intended to be an /ɛ/, per rater. The histograms can be found in Figure 9.a–c. Since the utterances were all produced by native speakers, we would expect only ratings of 1 (*good e*) and some ratings of 2 (*okay e*), but this was clearly not the case. According to Figure 9, the raters are mostly misidentifying /ɛ/ because they rate the utterance either as *okay a* (category 5) or as *good a* (category 6); they never heard *another vowel* (cat. 7). This means that the raters often heard /æ/ where /ɛ/ was intended. This might be because the raters still adapted their boundary to the Dutch speakers, even though we tried to prevent this by adding the native English speakers' utterances[14] (see Section III.C). In Figure 10a–b, the identification boundaries for the /ɛ/–/æ/ contrast of rater 1 and 2 are vizualized[15]. Figure 10 shows that the identification boundary of rater 1 and rater 2 are both shifted to the left. This means that they perceive more tokens as /æ/ than as /ɛ/, which is what we would expect, given the observation in Figure 8. The shift to the left is greater for rater 1 than for rater 2, which corresponds with the finding that rater 1 rated more /ɛ/-utterances as /æ/ than rater 2.



Rater 1.



Rater 2.

Figure 10. Identification boundaries for rater 1 (a) and 2 (b).

There are also some other possible explanations for the strange behavior of the raters: it could for example be that the vowels of the native Dutch speaker are in a completely different part of the vowel

[13] In the case that the intended vowel was /ɛ/, the categories were: 1 = *good e*, 2 = *okay e*, 3 = *poor e*, 4 = *poor a*, 5 = *okay a*, 6 = *good a*, 7 = *another vowel*.
[14] However, we did tell them that all the recordings were from Dutch people.
[15] These data are missing for participant 3, because of a technical failure.

space than the vowels of the native English speakers. However, Figure 11 shows that this is not the case: the vowel categories of the Dutch and English natives (based on one standard error[16]) are close together. The difference in categories was expected: native English people show separable categories, whereas Dutch natives show overlapping categories.

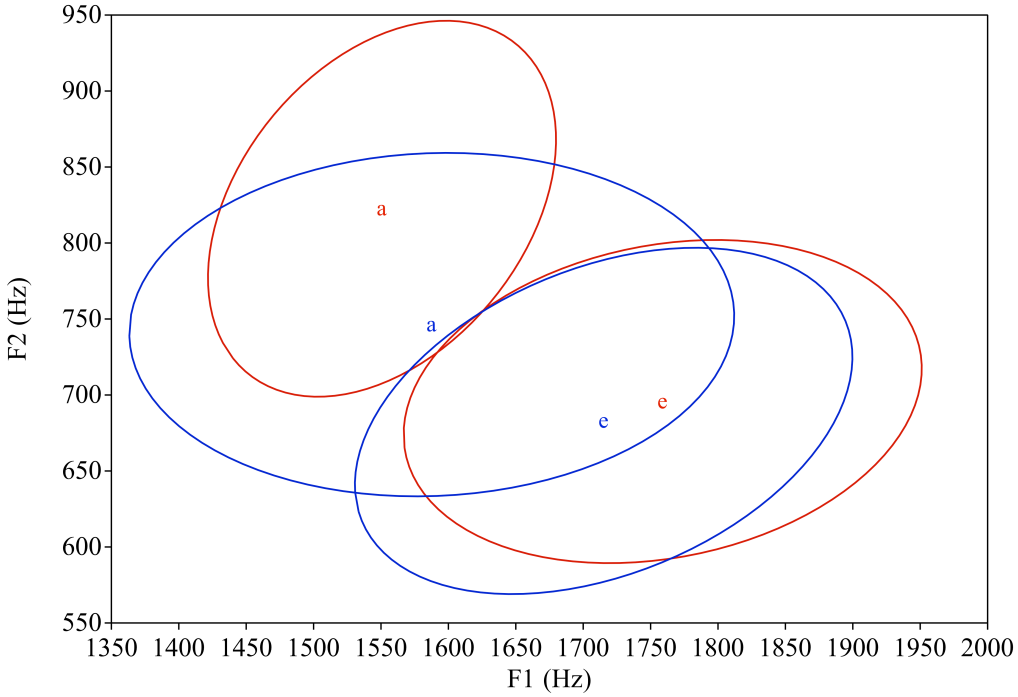


Figure 11. The distributions of the vowels used in the experiment, based on one standard error (sigma = 1), for the English natives (red) and the Dutch natives (blue).

The above results show, as already mentioned in Section I.B, that native raters may not provide the best feedback to non-natives to learn new contrasts. However, then the question arises as to what kind of measure should be authoritative. After all, the goal of speech production training is to attain a native-like accent, and this can only be judged by natives. In other words, there seems to be a rating problem: we cannot use native raters because their feedback is inconsistent, however, the ultimate aim is to get fluent in the ears of native listeners.

[16] The utterances were automatically segmented, as described in Section II.C. The formants were measured over the whole duration of the vowel, as described in Section II.B.

IV. DISCUSSION

The tool that was developed in this project distinguishes itself from existing tools in a few respects. First, it takes into account coarticulation, which was found to be a predictive feature in the LDA model for the categorization of the ɛ/–/æ/ contrast. Second, the tool bases its feedback on a larger set of native data than most previous tools (e.g. in Kartushina et al., 2015, only one speaker of each gender was used). Third, and most importantly, the feedback used in the tool is based on an extensive analysis aimed to find out which features were most predictive for automatic categorization.

Of course, improvements of the tool can be made. First, vowel normalization could be added, for example the calibration procedure described in Lie-Lahuerta (2011) and Lobanov (1971). This procedure could be used to normalize the input stimuli as well as the participants' utterances, by measuring the vowel space of the speakers and mapping them onto each other; feedback is then based on the mapped tokens. This would disentangle the vowel categories for the male voices to a certain degree, and make all the categories less scattered, whilst keeping the advantage basing the feedback on many input speakers. Because calibration can also normalize for gender, only one model would be needed; this would make sure that all raters receive feedback based on the same information. For this method, however, measurements of all natives' and participants' vowel space corners (i.e. /u/, /i/ and /a/) would have to be collected. Alternatively, to reduce the variability in the input data, the tool could also merely take input stimuli from the two most comparable native English speakers per gender (e.g. speaker 2 and 8 for the females, and speaker 4 and 7 for the males).

Second, the target words could be adapted. In this paper, multiple different words were used, to get as highly variable input as possible. This was done because it is known from research in perception that presenting listeners with a variable input improves learning (e.g. Bradlow et al., 1997); it was expected that in production listeners would also benefit from feedback based on variable input. However, for the current tool it might be more important to use a combination of vowels and consonants that are easily separable and distinguishable. This makes the segmentation more reliable, which improves the feedback. Moreover, the question remains whether vowels should be trained in isolation (or in contexts with very little coarticulation, like /t/, /d/ and /h/:) or in a context in which coarticulation takes place. The advantage of using vowels in isolation is that this might create a solid 'target' vowel that is aimed at when the vowel is used in a consonant-context, as suggested in the *production undershoot model*. The advantage of using vowels with consonant-contexts is that this is the way in which vowels are used in daily life, and participants might benefit from training on this. Furthermore, in L1 acquisition, vowels are also presented in their context. This raises a theoretical question about the representation of phonological categories: whether they are more 'prototypical' or more 'exemplar-based'. In the first case, the representation consists of target vowels in combination with rules about how tokens can differ from the target; in the second case representations would be made up many observed tokens of the vowel.

Thirdly, it might be worth the time investment to hand-segment all input utterances. The present paper used automatically segmented data. However, since the automatic segmentation method in *Praat* is not without errors, the feedback is based on target vowel distributions that are partly incorrect. These errors did not show up in the error checking: the formants could have been measured in neighboring consonants that still have similar formants due to coarticulation, or the errors were invisible because the average over the whole duration of the vowel was taken.

The findings of the experiment confirm the suggestion by Kartushina et al. (2015) that objective spectral analysis is more useful than subjective feedback by native listeners: it was found that subjective evaluations are indeed not stable. Since the raters showed strange behavior in the rating of the English natives' utterances with /ɛ/, and it is not entirely clear why, the low correlation between the tool and the ratings of the raters for /ɛ/ should not be too worrisome. In future research, the evaluation method for a tool like this should be very carefully designed.

There are a few suggestions for further research. First, it was suggested that the raters' identification boundary shifted under the influence of the productions of Dutch natives. Unfortunately, the data for the identification test of rater 3 was missing, which makes this interpretation even more speculative. However, the question whether identification boundaries for you native language might change under influence of non-native input would be a good topic for further research. Knowledge on how fast and in which direction categories can move might help teaching people new vowel contrasts, and it could tell something about the phonological representations.

Second, Jenkins and Strange's hypothesis that vowel identification is most importantly a process of attaining to acoustic changes (Jenkins & Strange, 1999; Section II.D.1) raises the question whether vowel production also uses acoustic changes. If this is the case, participant should be trained

on producing acoustic changes instead of on aiming at a certain target (as in the currently existing methods). However, the results from the LDA in Section II.D were not conclusive about whether the 20% and 80% points of the vowel duration improve prediction of the vowel. Therefore, for the current tool uses the average of the whole duration of the vowel. This method does contain information on the 20% and 80% points, but less explicitly; it still trains raters to reach a certain target, and not to produce certain spectral changes. Yet, it could well be that these spectral changes are only used in perception, and are automatically produced through coarticulation. Whether or not people use a representation of acoustic changes in vowel production, or whether this is merely a by-product of coarticulation, could be further researched.

Third, different ways of visualizing the feedback could be compared. Most speech production feedback systems with indirect feedback (based on acoustic measures) use some sort of visualization in the F1/F2 space. It could be tested, for example, whether using the entire vowel space helps training because it gives more reference points, or whether using only the relevant subset of the vowel space is better.

Finally, a general problem with research on feedback on speech production is that training is mostly based on the same principle as the test. For example, a tool that gives feedback based on a steady-state middle part of the vowel will most likely also test the improvement of the production based on the steady-state middle part. This leaves the question whether training then also improves the productions of the raters according to other formant measures, or according to native judgments.

V. CONCLUSION

In this paper, the process of developing a tool for feedback on the English /ɛ/–/æ/ contrast was described. Through this description, an overview of the theoretical as well as the practical factors that should be considered for a speech production feedback tool was given. The result of the project is a tool of which the different elements are made explicit and are thoroughly discussed. The choices for the method of analysis are based on an extensive analysis of the input data.

First, the most important features of the English vowels /ɛ/ and /æ/ were identified, to find which features should be used in feedback on pronunciation by novices. It was found that the mean F1 and F2 over the whole duration of the vowel are the best indicators for the current dataset. Furthermore, it was found that coarticulation is a significant predictor of the vowels, i.e. coarticulation should be taken into account in production feedback devices for the /ɛ/–/æ/ contrast. It is likely that this would hold as well for other vowel contrasts. Another attempt of this project was to base the feedback on the native utterances of many native speakers (5 per gender). This is indeed a theoretical merit; however, it became clear that it is important normalize the differences in the vowel spaces of the natives.

Second, the tool's judgments were tested against native listeners' ratings. It became clear that using native listeners to rate productions is not the most robust method, since they rated many productions of /ɛ/ by English native speakers as /æ/. Therefore, the correlations of the tool's judgment with the natives' ratings, some of which were quite high whereas others were low, were not straightforwardly interpretable.

In sum, we cannot be entirely sure whether the tool gives feedback that is consistent with natives' judgments. Moreover, there are still some possibilities for improvement of the tool, some of which are easier to implement than others. However, the tool in its current form is ready for use, and can already be experimented with.

ACKNOWLEDGEMENTS

I would like to thank my supervisors Makiko Sadakata and Paul Boersma for their great supervision. Thanks to Jana Krutwig for providing the stimuli and the segmentations, and for her input and help. Thanks to Dirkjan Vet for technical assistance and valuable input. Thanks to David Weenink for technical explanation, and for co-assessing this project. Lastly, many thanks to Paul Boersma and David Weenink for on the fly adapting *Praat* to my needs.

REFERENCES

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), 1729–1738.

Assmann, P. F., Nearey, T. M., & Hogan, J. T. (1982). Vowel identification: Orthographic, perceptual, and acoustic aspects. *The Journal of the Acoustical Society of America*, *71*(4), 975–989.

Baese-Berk, M. (2010). An Examination of the Relationship Between Speech Perception and Production. *Linguistics*, *PhD* (December), 201.

Boersma, P. (1993). Accurate Short-Term Analysis of the Fundamental Frequency and the Harmonics-To-Noise Ratio of a Sampled Sound. *Proceedings of the Institute of Phonetic Sciences*, *17*, 97–110.

Boersma, P. (unplublished, version *January 14, 2016*) How to measure emotion from the human voice.

Boersma, P. & Weenink, D. (2016). Praat: doing phonetics by computer [Computer program]. Version 6.0.19, retrieved 10 April 2016 from http://www.praat.org/.

Bradlow, A. R., Pisoni, D. B., Akahane-Yamada, R., & Tohkura, Y. (1997). Training Japanese listeners to identify English /r/ and /l/: IV. Some effects of perceptual learning on speech production. *The Journal of the Acoustical Society of America*, *101*(4), 2299–2310.

Brandmeyer, A., Timmers, R., Sadakata, M., & Desain, P. (2011). Learning expressive percussion performance under different visual feedback conditions. *Psychological Research*, *75*(2), 107–121.

Childers, D. G. (1978). *Modern spectrum analysis*. IEEE Computer Society Press.

Davis, S., & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *28*(4), 357–366.

Deterding, D. (1997). The Formants of Monophthong Vowels in Standard Southern British English Pronunciation. *Journal of the International Phonetic Association*, *27*(1-2), 47.

Escudero, P. (2005). *Linguistic perception and second language acquisition: Explaining the attainment of optimal phonological categorization*. Netherlands Graduate School of Linguistics.

Escudero, P., Boersma, P., Rauber, A. S., & Bion, R. A. H. (2009). A cross-dialect acoustic description of vowels: Brazilian and European Portuguese. *The Journal of the Acoustical Society of America*, *126*(3), 1379–1393.

Flege, J. E. (1995). Second Language Speech Learning: Theory, Findings, and Problems. *Speech Perception and Linguistic Experience: Issues in Cross-Language Research*, 233–277.

Flege, J. E., Bohn, O.-S., & Jang, S. (1997). Effects of experience on non-native speakers' production and perception of English vowels. *Journal of Phonetics*, *25*(4), 437–470.

Franken, M. K., McQueen, J. M., Hagoort, P., & Acheson, D. J. (2015). Assessing the link between speech perception and production through individual differences. In *Proceedings of the 18th International Congress of Phonetic Sciences* (Vol. 2, pp. 1–5).

Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: an overview and tutorial. *Tutorials in Quantitative Methods for Psychology*, *8*(1), 23.

Hillenbrand, J. M., Clark, M. J., & Nearey, T. M. (2001). Effects of consonant environment on vowel formant patterns. *The Journal of the Acoustical Society of America*, *109*(2), 748–763.

Hillenbrand, J., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *The Journal of the Acoustical Society of America*, *97*(5), 3099–3111.

Holt, L. L., Lotto, A. J., & Kluender, K. R. (2000). Neighboring spectral content influences vowel identification, 710–722.

Jenkins, J. J., Strange, W., & Edman, T. R. (1983). Identification of vowels in "voweless" syllables. *Perception and Psychophysics*, *34*(5), 441–450.

Jenkins, J. J., & Strange, W. (1999). Perception of dynamic information for vowels in syllable onsets and offsets. *Perception & Psychophysics*, *61*(6), 1200–1210.

Johnson, R.A., Wichern, D.W. (1982): *Applied Multivariate Statistical Analysis*, Prentice Hall.

Kartushina, N., Hervais-Adelman, A., Frauenfelder, U. H., & Golestani, N. (2015). The effect of phonetic production training with visual feedback on the perception and production of foreign speech sounds. *The Journal of the Acoustical Society of America*, *138*(2), 817–832.

Klein, W., Plomp, R., & Pols, L. C. (1970). Vowel spectra, vowel spaces, and vowel identification. *The Journal of the Acoustical Society of America*, *48*(April), 999–1009.

Kuhl, P. K. (2004). Early language acquisition: cracking the speech code. *Nature Reviews Neuroscience*, *5*(11), 831–843.

Lie-Lahuerta, C. (2011). Fix Your Vowels: Computer-assisted training by Dutch learners of Spanish. *TijdSchrift Voor Skandinavistiek*, *32*(1-2), 69–88.

Lobanov, B. M. (1971). Classification of Russian vowels spoken by different speakers. *The Journal of the Acoustical Society of America*, *49*(2B), 606–608.

Logan, J. S., Lively, S. E., & Pisoni, D. B. (1994). Training Japanese listeners to identify English /r/ and /l/: a first report. *The Journal of the Acoustical Society of America*, *89*(2), 874–886.

McQueen, J. (2005). Spoken-word recognition and production: regular but not inseparable bedfellows. *Twenty-First Century Psycholinguistics: Four Cornerstones*.

Ménard, L., Schwartz, J.-L., Boë, L.-J., Kandel, S., & Vallée, N. (2002). Auditory normalization of French vowels synthesized by an articulatory model simulating growth from birth to adulthood. *The Journal of the Acoustical Society of America*, *111*(4), 1892–1905.

Nearey, T. M. (1989). Static, dynamic, and relational properties in vowel perception. *The Journal of the Acoustical Society of America*, *85*(5).

Nearey, T. M., & Assmann, P. F. (1986). Modeling the role of inherent spectral change in vowel identification. *The Journal of the Acoustical Society of America*, *80*(5), 1297–1308.

Neri, A., Mich, O., Gerosa, M., & Giuliani, D. (2008). The effectiveness of computer assisted pronunciation training for foreign language learning by children. *Computer Assisted Language Learning*, *21*(5), 393–408.

Öster, A.-M. (1997). Auditory and visual feedback in spoken L2 teaching. *Phonum*, *4*, 145–161.

Parker, E. M., & Diehl, R. L. (1984). Identifying vowels in CVC syllables: Effects of inserting silence and noise. *Perception & Psychophysics*, *36*(4), 369–380.

Pols, L. C., van der Kamp, L. J., & Plomp, R. (1969). Perceptual and physical space of vowel sounds. *The Journal of the Acoustical Society of America*, *46*(2), 458–467.

Press, W. H., Flannery, B. P., Teukolsky, S. A., & Vetterling, W. T. (1992). Numerical recipes in C: the art of scientific programming. *Section*, *10*, 408–412.

Sadakata, M., & McQueen, J. M. (2013). High stimulus variability in nonnative speech learning supports formation of abstract categories: Evidence from Japanese geminates. *The Journal of the Acoustical Society of America*, *134*(2), 1324–1335.

Schmidt, R. A., & Lee, T. D. (1999). *Motor control and learning: a behavioral emphasis*. Human Kinetics.

Simpson, A. P. (2009). Phonetic differences between male and female speech. *Linguistics and Language Compass*, *3*(2), 621–640.

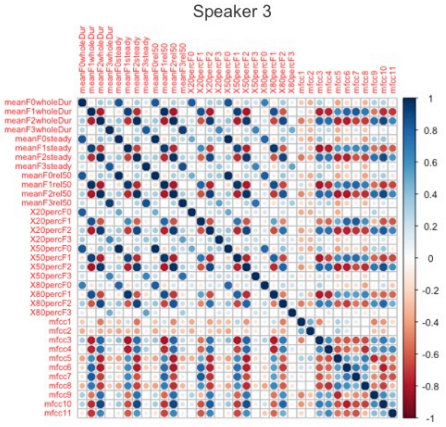
Stevens, K. N., & House, A. S. (1963). Perturbation of vowel articulations by consonantal context: An acoustical study. *Journal of Speech & Hearing Research*.

Zahorian, S. A., & Jagharghi, A. J. (1993). Spectral-shape features versus formants as acoustic correlates for vowels. *The Journal of the Acoustical Society of America*, *94*(4), 1966–1982.

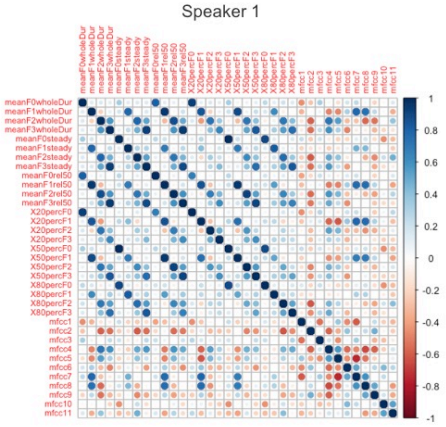
APPENDICES

A. Correlation plots for the separate speakers

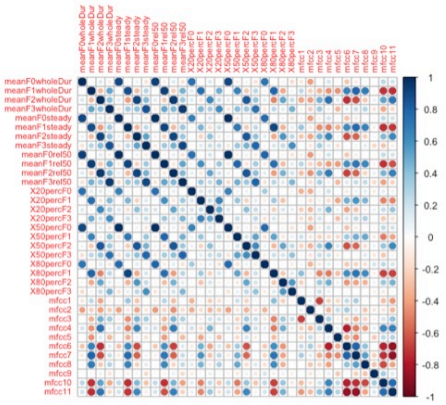
Speaker 3



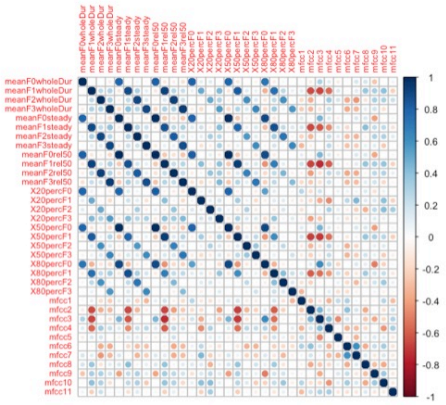
Speaker 1



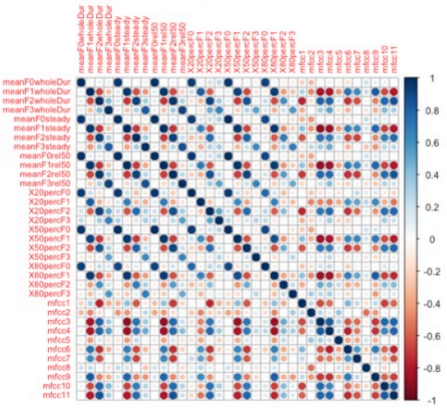
Speaker 6



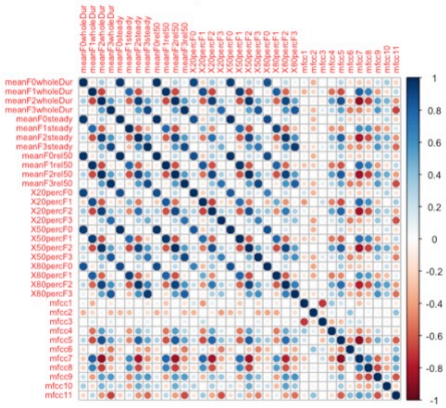
Speaker 3

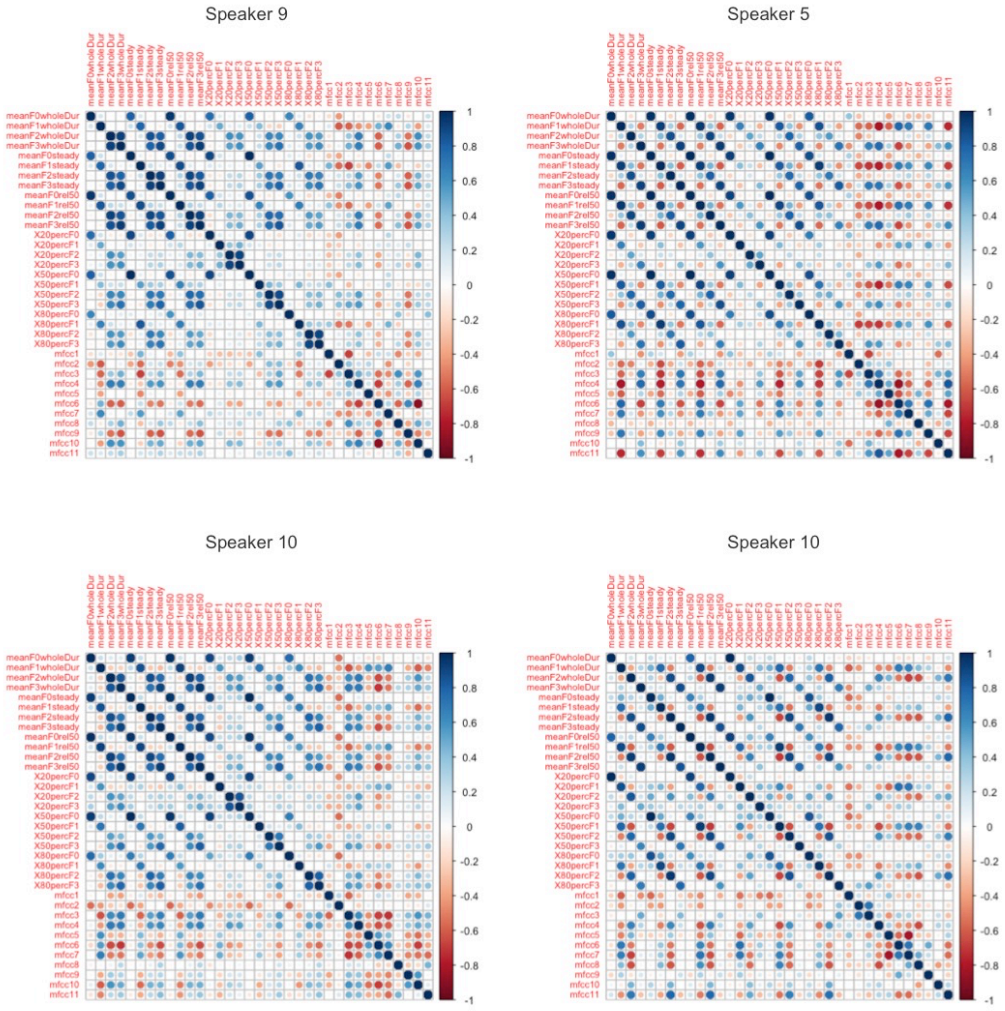


Speaker 8



Speaker 4





B. Script of the tool

```
##############################
# Gisela Govaart
# June 2016
# Research project 1: EarOpener
# Supervisors: Paul Boersma & Makiko Sadakata
##############################

######### Adjustable settings ###########
# How long is the microphone open (sec)
recordTime = 1.5
# How often do you want to repeat each word? (min 2X)
numRep = 2
# After how many words do you want to give the participant a break?
pauseAfter = 20
# In which file are your stimuli
stimfile$ = "stimfile.txt"

#########################################################################

tableMaleF_a = Read from file: "tableForTool_m_f_a.Table"
```

```
tableMaleH_a = Read from file: "tableForTool_m_h_a.Table"
tableMaleJ_a = Read from file: "tableForTool_m_j_a.Table"
tableMaleM_a = Read from file: "tableForTool_m_m_a.Table"
tableMaleP_a = Read from file: "tableForTool_m_p_a.Table"
tableMaleF_e = Read from file: "tableForTool_m_f_e.Table"
tableMaleH_e = Read from file: "tableForTool_m_h_e.Table"
tableMaleJ_e = Read from file: "tableForTool_m_j_e.Table"
tableMaleM_e = Read from file: "tableForTool_m_m_e.Table"
tableMaleP_e = Read from file: "tableForTool_m_p_e.Table"

tableFemaleF_a = Read from file: "tableForTool_f_f_a.Table"
tableFemaleH_a = Read from file: "tableForTool_f_h_a.Table"
tableFemaleJ_a = Read from file: "tableForTool_f_j_a.Table"
tableFemaleM_a = Read from file: "tableForTool_f_m_a.Table"
tableFemaleP_a = Read from file: "tableForTool_f_p_a.Table"
tableFemaleF_e = Read from file: "tableForTool_f_f_e.Table"
tableFemaleH_e = Read from file: "tableForTool_f_h_e.Table"
tableFemaleJ_e = Read from file: "tableForTool_f_j_e.Table"
tableFemaleM_e = Read from file: "tableForTool_f_m_e.Table"
tableFemaleP_e = Read from file: "tableForTool_f_p_e.Table"

meanstableMaleF_a = Read from file: "tableForToolMeans_m_f_a.Table"
meanstableMaleH_a = Read from file: "tableForToolMeans_m_h_a.Table"
meanstableMaleJ_a = Read from file: "tableForToolMeans_m_j_a.Table"
meanstableMaleM_a = Read from file: "tableForToolMeans_m_m_a.Table"
meanstableMaleP_a = Read from file: "tableForToolMeans_m_p_a.Table"
meanstableMaleF_e = Read from file: "tableForToolMeans_m_f_e.Table"
meanstableMaleH_e = Read from file: "tableForToolMeans_m_h_e.Table"
meanstableMaleJ_e = Read from file: "tableForToolMeans_m_j_e.Table"
meanstableMaleM_e = Read from file: "tableForToolMeans_m_m_e.Table"
meanstableMaleP_e = Read from file: "tableForToolMeans_m_p_e.Table"

meanstableFemaleF_a = Read from file: "tableForToolMeans_f_f_a.Table"
meanstableFemaleH_a = Read from file: "tableForToolMeans_f_h_a.Table"
meanstableFemaleJ_a = Read from file: "tableForToolMeans_f_j_a.Table"
meanstableFemaleM_a = Read from file: "tableForToolMeans_f_m_a.Table"
meanstableFemaleP_a = Read from file: "tableForToolMeans_f_p_a.Table"
meanstableFemaleF_e = Read from file: "tableForToolMeans_f_f_e.Table"
meanstableFemaleH_e = Read from file: "tableForToolMeans_f_h_e.Table"
meanstableFemaleJ_e = Read from file: "tableForToolMeans_f_j_e.Table"
meanstableFemaleM_e = Read from file: "tableForToolMeans_f_m_e.Table"
meanstableFemaleP_e = Read from file: "tableForToolMeans_f_p_e.Table"

form Fill in the following information
       word ParticipantID
       word SessionNr
       choice Gender: 1
              button Female
              button Male
endform

fileID$ = "'participantID$'_'sessionNr$'"
createDirectory: "Results/'fileID$'"
outdir$ = "Results/'fileID$'"
createDirectory: "'outdir$'/pdfs"

### Present the word orthografically, save which word it was

strings = Read Strings from raw text file: "stimfile.txt"
for i from 1 to numRep-1
       stringsPart'i' = Extract part: 1, 10
endfor
selectObject: strings
for  i from 1 to numRep-1
       plusObject: stringsPart'i'
endfor
s = Append
Create Permutation: "p", 10*numRep, "yes"
p = Permute randomly (blocks): 0, 0, 10, "yes", "yes"
selectObject: p, s
wordList = Permute strings
nrWords = Get number of strings

demo Erase all
demoWindowTitle: "Production feedback tool EarOpener"
demo Navy
demo 12
@ clearDemoWindow
demo Text: 50, "centre", 50, "half", "This is the production part. You will (instructions). Push the space
bar to start the experiment"
while demoWaitForInput()
       goto SECOND_SCREEN demoInput(" ")
endwhile
label SECOND_SCREEN
@clearDemoWindow
```

```
demo 18
demo Text: 50, "centre", 50, "half", "Press the space bar to record the next word"

counterCorrect = 0
table = Create Table with column names: "tableInfo_'fileID'", 10*numRep, "participantID sessionNr date
gender word wordNr intendedVowel correct? mahalanobis startC F1 F2"

for i from 1 to nrWords

	## the pause ##
	if i mod pauseAfter = 0
		while demoWaitForInput()
		goto PAUSE_SCREEN demoInput(" ")
		endwhile
		label PAUSE_SCREEN
		@clearDemoWindow
		demo 24
		demo Text: 50, "centre", 60, "half", "This is a break. Press the space bar to continue"
		demoShow()
		while demoWaitForInput()
			goto THIRD_SCREEN demoInput(" ")
		endwhile
	endif

	selectObject: wordList
	word$ = Get string: 'i'
	s$ = left$ (word$, 1)
	if s$ = "g"
		s$ = "j"
	endif
	e$ = right$ (word$, 1)
	vowel$ = mid$ (word$, 2,1)
	while demoWaitForInput()
		goto THIRD_SCREEN demoInput(" ")
	endwhile
	label THIRD_SCREEN
	@clearDemoWindow
	demo 24
	demo Text: 50, "centre", 60, "half", word$
	demoShow()

	### Record the participant's utterance

	sound = Record Sound (fixed time)... Microphone 0.99 0.5 44100 2
	j = ceiling((i-1)/10 + 0.05)
	Save as WAV file: "'outdir$'/'fileID$'_'word$''j'.wav"

	### Segment the participant's utterance

	selectObject: sound
	textgrid = noprogress To TextGrid: "CVC", ""
	Set interval text: 1, 1, word$

	# I select m1 and f1, you can also choose one of the other male/female 'voices'.
	selectObject: sound
	if gender = 1
		speechSynt = noprogress Create SpeechSynthesizer: "English", "f1"
	elsif gender = 2
		speechSynt = noprogress Create SpeechSynthesizer: "English", "m1"
	endif

	selectObject: sound, textgrid, speechSynt
	textgridAligned = noprogress To TextGrid (align): 1, 1, 1, -30, 0.1, 0.1
	## -30 dB is the silence threshold.

	selectObject: textgridAligned
	noprogress Save as text file: "'outdir$'/'fileID$'_'word$''j'.TextGrid"

	### Analyze the participant's utterance

	selectObject: textgridAligned
	nrIntervals = noprogress Get number of intervals: 4
	for k from 1 to nrIntervals
		selectObject: textgridAligned
		intervalLabel$ = Get label of interval: 4, k
			if intervalLabel$ = "æ" or intervalLabel$ = "a" or intervalLabel$ = "ɛ" or
intervalLabel$ = "e"
			startVowel = noprogress Get starting point: 4, k
			endVowel = noprogress Get end point: 4, k
			durationVowel = (endVowel - startVowel)
		endif
	endfor

	selectObject: sound
		if gender = 1
```

```
                    formant = noprogress To Formant (burg): 0.001, 5, 5500, 0.025, 50
                    # 0.001, want default is 0.01. dan meet ie dus de formant op iedere 0.01sec.
      elsif gender = 2
                    formant = noprogress To Formant (burg): 0.001, 5, 5000, 0.025, 50
      endif
f1hertz = noprogress Get mean: 1, startVowel, endVowel, "Hertz"
f2hertz = noprogress Get mean: 2, startVowel, endVowel, "Hertz"
f1 = hertzToErb(f1hertz)
f2 = hertzToErb(f2hertz)

### Analyzis on whether the utterance was correct or incorrect:
### MAHALANOBIS Distance
if gender = 1
      if vowel$ = "a"
            if s$ = "f"
                  @mahalanobis: tableFemaleF_a
            elsif s$ = "h"
                  @mahalanobis: tableFemaleH_a
            elsif s$ = "j"
                  @mahalanobis: tableFemaleJ_a
            elsif s$ = "m"
                  @mahalanobis: tableFemaleM_a
            elsif s$ = "p"
                  @mahalanobis: tableFemaleP_a
            endif
      elsif vowel$ = "e"
            if s$ = "f"
                  @mahalanobis: tableFemaleF_e
            elsif s$ = "h"
                  @mahalanobis: tableFemaleH_e
            elsif s$ = "j"
                  @mahalanobis: tableFemaleJ_e
            elsif s$ = "m"
                  @mahalanobis: tableFemaleM_e
            elsif s$ = "p"
                  @mahalanobis: tableFemaleP_e
            endif
      endif
else
      if vowel$ = "a"
            if s$ = "f"
                  @mahalanobis: tableMaleF_a
            elsif s$ = "h"
                  @mahalanobis: tableMaleH_a
            elsif s$ = "j"
                  @mahalanobis: tableMaleJ_a
            elsif s$ = "m"
                  @mahalanobis: tableMaleM_a
            elsif s$ = "p"
                  @mahalanobis: tableMaleP_a
            endif
      elsif vowel$ = "e"
            if s$ = "f"
                  @mahalanobis: tableMaleF_e
            elsif s$ = "h"
                  @mahalanobis: tableMaleH_e
            elsif s$ = "j"
                  @mahalanobis: tableMaleJ_e
            elsif s$ = "m"
                  @mahalanobis: tableMaleM_e
            elsif s$ = "p"
                  @mahalanobis: tableMaleP_e
            endif
      endif
endif

if gender = 1
      if mahalanobis.mahaladist < 1
            correct = 1
            color$ = "green"
            counterCorrect = counterCorrect + 1
      else
            correct = 0
            color$ = "red"
      endif
elsif gender = 2
      if mahalanobis.mahaladist < 0.5
            correct = 1
            color$ = "green"
            counterCorrect = counterCorrect + 1
      else
            correct = 0
            color$ = "red"
      endif
endif
```

```
	### Give the feedback

	demo Erase all
	demo Select inner viewport: 0, 100, 0, 90

	if gender = 1
		if s$ = "f"
			@drawAxes: meanstableFemaleF_a, meanstableFemaleF_e, vowel$
		elsif s$ = "h"
			@drawAxes: meanstableFemaleH_a, meanstableFemaleH_e, vowel$
		elsif s$ = "j"
			@drawAxes: meanstableFemaleJ_a, meanstableFemaleJ_e, vowel$
		elsif s$ = "m"
			@drawAxes: meanstableFemaleM_a, meanstableFemaleM_e, vowel$
		elsif s$ = "p"
			@drawAxes: meanstableFemaleP_a, meanstableFemaleP_e, vowel$
		endif

	else
		if s$ = "f"
			@drawAxes: meanstableMaleF_a, meanstableMaleF_e, vowel$
		elsif s$ = "h"
			@drawAxes: meanstableMaleH_a, meanstableMaleH_e, vowel$
		elsif s$ = "j"
			@drawAxes: meanstableMaleJ_a, meanstableMaleJ_e, vowel$
		elsif s$ = "m"
			@drawAxes: meanstableMaleM_a, meanstableMaleM_e, vowel$
		elsif s$ = "p"
			@drawAxes: meanstableMaleP_a, meanstableMaleP_e, vowel$
		endif
	endif


####################
	# Make sure that the utterance is not plotted in the upper part of the window (where the text is)
	if f2 < drawAxes.upperF2
		demo Paint circle (mm): color$, f1, f2, 2.5
	endif

	demo Axes: 0, 100, 0, 100
	demo Select inner viewport: 0, 100, 0, 100
	demo 18
	demo Maroon
	demo Text special: 30, "centre", 93, "half", "Times", 18, "0", "Push the space bar to go to the next
word"
	demo Green
	demo Text special: 80, "centre", 93, "half","Times", 24, "0", "Number Correct: 'counterCorrect'"
	demo Navy

	### Save the info in a table
	selectObject: table
			Set string value: i, "participantID", participantID$
			Set string value: i, "sessionNr", sessionNr$
			Set string value: i, "date", date$()
			Set numeric value: i, "gender", gender
			Set string value: i, "word", word$
			Set numeric value: i, "wordNr", j
			Set string value: i, "intendedVowel", vowel$
			Set numeric value: i, "correct?", correct
			Set numeric value: i, "mahalanobis", mahalanobis.mahaladist
			Set string value: i, "startC", s$
			Set numeric value: i, "F1", f1
			Set numeric value: i, "F2", f2
	selectObject: sound, textgrid, textgridAligned, speechSynt, formant
	Remove
	demoShow()
endfor

selectObject: table
Save as tab-separated file: "'outdir$'/'fileID$'.Table"

while demoWaitForInput()
	goto END demoInput(" ")
endwhile

label END
@ clearDemoWindow
demo Text: 50, "centre", 50, "half", "This is the end of the experiment. Thanks for participating."

### Drawing pictures to see how the segmentation went
strings = Create Strings as file list: "fileList", "'outdir$'/*.wav"
nrFiles = Get number of strings
for i from 1 to nrFiles
```

```
	selectObject: strings
	currentFile$ = Get string: 'i'
	soundn = Read from file: "'outdir$'/'currentFile$'"
	objectName$ = selected$ ("Sound")
	textgridn = Read from file: "'outdir$'/'objectName$'.TextGrid"
	selectObject: soundn, textgridn
	Erase all
	Select outer viewport: 0, 6, 0, 4
	Draw: 0, 0, "yes", "yes", "yes"
	Save as PDF file: "'outdir$'/pdfs/segmentation_'objectName$'.pdf"
endfor
demoShow()

########################################################################

procedure clearDemoWindow
	demo Erase all
	demo Axes: 0, 100, 0, 100
	demo Select inner viewport: 0, 100, 0, 100
	demo Paint rectangle: "silver", 0, 100, 0, 100
endproc

procedure mahalanobis: .tableForTool
	selectObject: .tableForTool
	Down to TableOfReal: ""
	cov_i = To Covariance
	table_i = Create Table with column names: "table_i", 1, "F1 F2"
	Set numeric value: 1, "F1", f1
	Set numeric value: 1, "F2", f2
	selectObject: table_i
	tor_i = Down to TableOfReal: ""
	selectObject: cov_i, tor_i
	table_mal = To TableOfReal (mahalanobis): "no"
	.mahaladist = Get value: 1,1
	selectObject: cov_i, table_i, tor_i, table_mal
	Remove
endproc

procedure drawAxes: .tableA, .tableE, .vowel$
	selectObject: .tableA
	meanF1a = Get value: 1, "F1"
	meanF2a = Get value: 1, "F2"

	selectObject: .tableE
	meanF1e = Get value: 1, "F1"
	meanF2e = Get value: 1, "F2"
	.f1difference = meanF1a - meanF1e
	.f2difference = meanF2e - meanF2a

	if gender = 1
		if .vowel$ = "a"
			demo Teal
			demo Axes: meanF1e - 2*.f1difference, meanF1a + 2*.f1difference, meanF2a -
2*.f2difference, meanF2e + 2*.f2difference
			demo Text special: meanF1a, "centre", meanF2a, "half", "Times", 25, "0", "\f5a"
			demo Silver
			demo Text special: meanF1e, "centre", meanF2e, "half", "Times", 25, "0", "e"
		else
			demo Teal
			demo Axes: meanF1e - 2*.f1difference, meanF1a + 2*.f1difference, meanF2a -
2*.f2difference, meanF2e + 2*.f2difference
			demo Text special: meanF1e, "centre", meanF2e, "half", "Times", 25, "0", "\f5e"
			demo Silver
			demo Text special: meanF1a, "centre", meanF2a, "half", "Times", 25, "0", "a"
		endif
	.upperF2 =   meanF2e + 2*.f2difference
	endif

	if gender = 2
		if .vowel$ = "a"
			demo Teal
			demo Axes: meanF1e - 2.1*.f1difference, meanF1a + 2.1*.f1difference, meanF2a -
2.08*.f2difference, meanF2e + 2.08*.f2difference
			demo Text special: meanF1a, "centre", meanF2a, "half", "Times", 25, "0", "\f5a"
			demo Silver
			demo Text special: meanF1e, "centre", meanF2e, "half", "Times", 25, "0", "e"
		else
			demo Teal
			demo Axes: meanF1e - 2.1*.f1difference, meanF1a + 2.1*.f1difference, meanF2a -
2.08*.f2difference, meanF2e + 2.08*.f2difference
			demo Text special: meanF1e, "centre", meanF2e, "half", "Times", 25, "0", "\f5e"
			demo Silver
			demo Text special: meanF1a, "centre", meanF2a, "half", "Times", 25, "0", "a"
		endif
	.upperF2 =   meanF2e + 2.08*.f2difference
```

```
        endif

        demo Black
        demo Line width: 1
        demo Draw inner box
endproc
```