Can you hear what you cannot say?

The interactions of speech perception and production during non-native phoneme learning

Jana Thorin

Can you hear what you cannot say?
The interactions of speech perception and production during non-native phoneme learning
PhD thesis, Radboud University Nijmegen
Cover Artwork and Layout by Fenna Schaap
Printed by Proefschriftmaken.nl
ISBN 9789462842236

This research was carried out at the Donders Graduate School for Cognitive Neuroscience and the International Max Planck Research School for Language Sciences. It was funded by a research grant of the Language in Interaction Consortium.

Can you hear what you cannot say?

The interactions of speech perception and production during non-native phoneme learning

Proefschrift
ter verkrijging van de graad van doctor
aan de Radboud Universiteit Nijmegen
op gezag van de rector magnificus prof. dr. J.H.J.M. van Krieken,
volgens besluit van het college van decanen
in het openbaar te verdedigen op donderdag 20 februari 2020
om 10.30 uur precies

door
Jana Thorin
geboren op 28 september 1988
te Leverkusen, Duitsland

Promotoren
Prof. dr. James M. McQueen
Prof. dr. Peter W.M. Desain

Copromotor
Dr. Makiko Sadakata

Manuscriptcommissie
Prof. dr. Ton A.F.J. Dijkstra
Prof. dr. Niels O. Schiller (Universiteit Leiden)
Dr. Melissa Baese-Berk (University of Oregon, Verenigde Staten)

Don’t judge each day by the harvest you reap, but by the seeds you plant.
– Robert Louis Stevenson

Failure is the mother of success
- Unknown

Table of Contents

1

Chapter 1

Introduction

Nienke likes learning languages. She received decent grades when finishing her English exams in high school and she has little effort following English-spoken TV shows or movies. Like so many other Dutch natives, she is a proficient speaker of English. What stands in her way of mastering the English language completely, however, are a few sounds which she cannot quite distinguish. If it was not for the context, she would have a hard time hearing the difference between a native speaker's pronunciation of the English words "pan" and "pen". To her, both of them sound like versions of the Dutch word "pen" (which also happens to mean the same as the English word "pen"). Similarly to her difficulties in perceiving the vowels, she also struggles to produce them. She wonders how she should best go about mastering these challenging sounds. How can she learn to pronounce something she cannot hear in the first place? And conversely, how can she hear what she cannot say?

While acquiring our native language in childhood comes naturally to most of us, mastering a second language in all its intricacies later on in life involves some additional challenges. Just like Nienke experiences it, those challenges oftentimes comprise the non-native sound system. As new-borns, we are able to distinctively perceive a wide range of sounds. Already after a few months of exposure to our native language (L1), however, this perceptual spectrum has narrowed as we specialise in distinguishing between phonemes that are relevant to our mother tongue. Werker and Tees (1984) famously revealed this effect when testing the discrimination ability of three groups of infants (aged 6–8 months, 8–10 months, and 10–12 months respectively) on several consonant contrasts (present in either English, Hindi, or a native Indian language). Their results showed that the youngest group was able to distinguish between all of the contrasts, while 10-12 months old infants could only do so between the contrasts, which were relevant in the context of their L1 processing. Similarly, Kuhl and colleagues (1992) showed that already by the age of 6 months, infants' perception of the /i/ and /y/ was distinctively altered by exposure to their native language, here either Swedish or American English. Again, those infants had become specialised in perceiving linguistically relevant phonemes of their L1, while becoming relatively insensitive in distinguishing between less relevant ones.

Given these early influences of native language exposure on sound perception, it is not surprising that (re)learning to perceive phonemes that become relevant in the context of a foreign language (L2) later on in life can be a challenging process. Despite these difficulties, however, findings suggest that adult learners can still – to some degree - succeed in forming non-native phoneme categories, for instance, through targeted phoneme training (see among others, Bradlow et al., 1997; Herd et al., 2013; Lambacher et al., 2005; Wang et al., 2003), or through extensive language immersion (Flege et al., 1996, 1999a, 1999b). The phonemic system thus stays plastic throughout the lifespan, though it is known to decrease in flexibility with age (Flege et al., 1996, 1999b; Piske et al., 2001).

In order to reach sufficiently high levels of proficiency in a foreign language to enable efficient communication in various linguistic contexts and speaker situations, non-native phonemes need to be established both in speech perception and production. This means that, through learning, stable representations need to be formed upon which both speech processes can be based. A non-native language user, like Nienke above, needs to be able to discriminate between relevant phonemes of her L2, while also being able to distinctively pronounce them. During second language acquisition, this might seem like a "chicken-and-egg" problem: what comes first (or rather, what needs to come first), the proficient perception or the correct production of a non-native phoneme? One speech process seems to depend on the respective other making it hard to disentangle where learning could start. For instance, whenever attempting to produce a challenging foreign sound, it seems plausible that a speaker will listen to their own utterance in order to evaluate and potentially adjust the articulatory process. This could happen even before the actual realisation of producing a sound through internal feedback loops. In either case, however, it seems plausible that in order to efficiently evaluate the quality of his or her own utterance, a speaker and "self-listener" needs to have an idea about what the articulation ought to sound like. Conversely, it seems daunting for a listener trying to discriminate between two identical appearing sounds produced by another speaker, while not being able to distinctively pronounce them herself.

This dilemma taps directly into an ongoing debate in the speech sciences. While results from various training and immersion studies, such as the ones mentioned above, have taught us that learning to perceive and produce non-native phonemes in adulthood is still possible (to some degree), how the relationship between the two speech modalities in this learning process can best be characterised is still inconclusive. It is this debate that motivates the main focus of the present dissertation. More concretely, in the following four empirical chapters (Chapters 2-5), I would like to further our understanding of how speech perception and speech production interact during second language sound learning. My aim is to further investigate how this relationship between the speech modalities can be best described by testing under which circumstances they support or potentially even hinder each other in the process of establishing non-native sound categories.

The second aim of this dissertation is to ask what we can learn from a deeper understanding of perception-production interactions with respect to choosing and developing efficient training approaches for adult learners to successfully master L2 sounds in both modalities. Before turning to the empirical chapters, each dealing with different aspects of learning to perceive and/or produce non-native sounds, I will first elaborate on the difficulties associated with non-native sound processing and how they could be accounted for, and then, second, specify and discuss the variety of methods used in this dissertation.

I. THE CHALLENGE OF NON-NATIVE SOUNDS

Which sound categories of a second language are difficult to perceive and pronounce for the learner of that L2 are largely determined by the interplay between the learner's native and the non-native phonetic system. What seems easy for some native speakers represents a major difficulty for others. When learning English as a foreign language, for instance, learners from various L1 backgrounds differ largely with respect to which phonemes pose a challenge for them: for example, Japanese natives generally struggle with producing the two English liquids (as contrasted in English *read* and *lead*; Bradlow et al., 1997, 1999), while Italian natives have difficulties with various English vowel contrasts, such as, differentiating between /ɔː/ and /ʌ/ (as in English *bought* and *but*; Flege et al., 1999a), whereas Dutch native speakers tend to struggle specifically with the discrimination of English /æ/ and /ɛ/ (as in English *bad* and *bed*; Broersma, 2002, 2005).

Plausible explanations for this phenomenon come from widely accepted accounts of cross-language speech perception. The Speech Learning Model (SLM; Flege, 1995) postulates that during the initial phase of encountering a given L2, a new speech category will more likely be established the more it is dissimilar from any existing L1 category. The Perceptual Assimilation Model (PAM; Best, 1995) similarly assumes that formation of novel speech categories depends on perceived (dis)similarity between the native and the non-native phonological system and makes concrete predictions on how a given L2 phone gets perceptually assimilated into an existing L1 sound category. According to the model, this assimilation can happen in three different ways: (1) a non-native phone can get *categorised* if it gets assimilated to an L1 phoneme based on its perceived close similarity to it, (2) it could stay *uncategorised* in case it resembles no similarity with any L2 phoneme (it will then stay in an "untuned region" in between categories), or (3) it could be *non-assimilable*, which means that it is not perceived as speech and will in that case be outside the listener's L1 phonological space (Faris et al., 2018). The term categorisation here relates to the perceptual phenomenon of *categorical perception* (Liberman et al., 1961). In speech perception, it describes the tendency of listeners to perceive distinct categories even when presented with a continuum of sounds. This enables listeners to deal with a certain degree of variability in speech produced by others by perceiving both good and poor examples of a given phonemic category as valid realisations of it.

In the context of L2 speech perception, assimilation of novel speech sounds can both be helpful and misleading. It will be trivial for L2 learners to discriminate two non-native phonemes if they are similar to (and thus directly relate to) two categories in the native sound space (Best and Strange, 1992). Another easy case has been shown to be the perception of non-native sounds that do not resemble any similarity with any of the L1 sound categories. This explains why, for instance, English native speakers have few problems distinguishing between Zulu clicks (Best et al., 1988). According to PAM, however, a difficult situation for the non-native language user will arise whenever two phones of a given L2 contrast are similar to a single L1 category. As both phonemes will then get perceived as examples of a single category, the L2 listener will have difficulties in discriminating the two.

To illustrate this process, we can consider one of the above examples in more detail. The Dutch native phonological system comprises the category /ɛ/ (as in Dutch *het*), which is perceptually similar (though not identical) to the English category /ɛ/ (as in English *pen*). As there is no Dutch phoneme /æ/, PAM would predict that it gets assimilated to the closest similar category, here Dutch /ɛ/, and that discrimination between the two L2 phonemes will thus be poor for Dutch native listeners. This is indeed what has been shown in various experimental settings. The two English vowels were shown to be confused by Dutch native speakers, for instance, in a word recognition task (Escudero et al., 2008), in a lexical decision task (Broersma, 2002), a spoken-word recognition task (Weber and Cutler, 2004), and also led to spurious lexical competition in a priming task (Broersma, 2002). In sum, the English /æ/-/ɛ/ vowel contrast represents a good example of a difficult to perceive (and pronounce) speech contrast for Dutch native speakers (even in the case of intermediate/high levels of English proficiency) and was therefore employed in all empirical chapters of this dissertation as a tool to further our understanding of how speech perception and production interact in the course of learning to perceive a challenging novel speech contrast.

II. METHODOLOGY

This dissertation is built on the use of a variety of scientific methods, which were employed – in their combination - to illuminate various aspects of sound learning and general cognitive and linguistic performance relevant during this process. At the core of this work stands the use of two multi-day training studies. While being relatively time-consuming and costly, multi-day training approaches are a valuable method to induce and examine learning in a controlled experimental setting. As compared to cross-sectional studies, for instance, they thereby also enable researchers to address causality instead of simple correlational effects. Using multiple sessions spread over multiple days came with the additional advantage of enabling us to study the time course of learning, for instance, how it is shaped and to what extent it is influenced by other possible factors, such as, consolidation during sleep. Spreading the training over a longer period of time, during which participants will leave the controlled experimental setting between training sessions, did mean losing some control over confounding variables. For instance, a participant trained in the perception and production of English vowels could have engaged in considerably more conversations with English speakers between sessions, potentially leading to additional gains that would be independent of the targeted training. This general confounding factor, however, can be reduced by using sufficient numbers of participants and by including a valid control condition to any training design, which we did in both training studies.

Performance in both speech perception and production was assessed by a wide range of methods, which were complementing and thus strengthening each other. Firstly, we made use of various behavioural measures. Perceptual learning was assessed both in terms of changes in identification and discrimination ability, and the degree to which training transferred to new linguistic contexts, such as new speakers or new words. More concretely, we quantified the degree to which participants could identify auditorily presented English words containing one of the difficult vowels (such as in the English word *pen*) and also how well they could discriminate between the two challenging vowels when being auditorily presented in a sequence. By employing an eleven-step continuum between the critical /æ/ and /ɛ/ vowels (artificially created sound stimuli with adjusted values of the first two formants) in one of the discrimination tasks, we were also able to quantify the sharpness of the perceptual boundary between the two phonemes for each participant. Additional to these speech perception measures, each participant's level of English proficiency was assessed by means of a computerised English vocabulary test (the LexTALE task; Lemhöfer and Broersma, 2012) and questionnaires collecting data on a participants' native language background and both their proficiency and everyday use of foreign languages. These quantitative measures were complemented by some qualitative measures in the form of open questions concerning, for instance, participants' motivation or potential comments on the perceived efficiency of the training. All of these behavioural measures had the advantage of being relatively easy to administer both in terms of time and costs. It should be kept in mind, however, that they are relatively indirect measures and, as we will see in **Chapter 3**, in some regards potentially less sensitive than more direct neural measurements.

Secondly, we made use of speech signal analysis in order to quantify participants' ability to pronounce the non-native phonemes. The phonetic quality of the English /æ/ and /ɛ/ vowels can be well characterised (and distinguished) in terms of their first two formants, F1 and F2, which refer to the first two prominent frequency bands in a speech signal's spectral representation. We therefore based the quantification of production learning on the F1 and F2 values of the vowel productions made by participants before, during, and after training. Those could be used both to examine the degree to which participants were able to distinctively pronounce the two vowels (how much did the formant values differ between the vowels for a given participant?) and, in the case of the production training presented in **Chapter 4**, how similar their vowel productions were to those of typical native speakers (how close do the formant values come to a native model?). Notably, we did see that values based on automatic formant extraction tend to differ (slightly) depending on the extraction method used (for instance, mean formant value across the entire vowel segment or mean value across 50% centred portion of the vowel) and the exact timing of the defined vowel segment, which varies depending on whether those segments were defined based on an automatic or manual method (and also in between manual segmentations by different evaluators). As we show in **Chapter 4**, however, different formant extraction methods resulted in the same overall patterns suggesting that different methods are similarly valid. In sum, this automatised method enabled us to time-efficiently, consistently, and objectively assess the quality of non-native vowel productions. We therefore chose it above non-automatised alternative evaluation approaches, such as speech ratings by native English speakers as possible.

To further validate vowel evaluations based on formant extractions, the above analysis was complemented by an automatic speech recognition (ASR) approach in **Chapter 2**. Here, we trained a binary classification model with a set of word recordings containing either of the two English target vowels (the same that were produced by the Dutch participants) and then used it to classify if a given word produced by the Dutch participants at pre- and post-test of the training contained either the /æ/ or the /ɛ/ vowel.

Lastly, we also made use of electrophysiological measurements in **Chapters 3** and **5**. In a nutshell, electroencephalography (EEG) is the method of recording the continuous electrical activity produced by synchronously active neurons by means of placing one or more electrodes on the scalp (Kemmerer, 2015). In light of its relatively low spatial resolution, it is not a favourable method to investigate the location of a cognitive process, but EEG has the advantage of having an excellent temporal resolution (in the range of milliseconds). When averaging the activity time-locked to a specific event, such as a response or the presentation of a stimulus, across multiple occurrences of that event, one can reveal so-called event-related potentials (ERPs). This computation is a good way to increase the signal-to-noise ratio (random noise tends to cancel out in the averaging process, while consistent signal does not), which can get compromised by muscle movements (e.g. during speaking) or other sources of electrical noise.

In the context of this dissertation, the most relevant of these ERPs were the auditory mismatch negativity (MMN) and the error-related negativity (ERN). The MMN is typically observed in response to a deviating stimulus in a sequence of repeated standard stimuli and therefore known to be an efficient tool to measure automatic auditory change detection even in the absence of focused attention (Näätänen et al., 1997, 2007). The response has found applications in a wide range of scientific disciplines, including music cognition (Fujioka et al., 2004; Koelsch et al., 1999), consciousness (Fischer et al., 2010) and sleep research (Sculthorpe et al., 2009), and psycholinguistics (Pulvermüller and Shtyrov, 2006). Most relevant for this dissertation, the auditory MMN has been shown to be useful in the context of L2 perception, for instance, to assess individual differences in non-native listeners' ability to discriminate between L2 phonemes (Díaz et al., 2016; Jakoby et al., 2011a), to quantify nativelikeness of discrimination ability of L2 sounds

(Grimaldi et al., 2014; Peltola et al., 2005; Rivera-Gaxiola et al., 2000) and also specifically to complement behavioural measurements of L2 training evaluation (Lu et al., 2015; Tamminen et al., 2015; Ylinen et al., 2010; Zhang et al., 2009). In **Chapter 3,** we used the MMN to evaluate if and to what degree Dutch participants were able to perceive the difference between the trained English vowels during and after targeted phonemic training. We thereby complemented our behavioural measurements of perceptual learning outcomes by an additional and, as it turned out, potentially more sensitive tool in assessing L2 learner's discrimination ability. The above mentioned ERN response became relevant in the design and outcomes presented in **Chapter 5.** The ERN is a widely used potential that is typically observed shortly after an erroneous response action (Gehring et al., 1993; Hohnsbein et al., 1991). Originally employed in the research of action monitoring, it was subsequently shown to be a valuable tool when applied in language production research as well (Ganushchak et al., 2011; Ganushchak and Schiller, 2006; Trewartha and Phillips, 2013).

Although each of the above methods, involving analyses of behavioural, speech and neural data, exhibit their respective benefits and drawbacks when considered separately, they become especially powerful when used in combination with each other, then enabling conclusions based on converging evidence.

III. THESIS OVERVIEW

This dissertation reports and subsequently discusses several experimental studies focussing on the interaction of speech perception and speech production during non-native speech category learning. Language learners in all experiments are native Dutch participants who are intermediate/highly proficient speakers of English. The non-native speech sounds in focus are the English /æ/ and /ɛ/ vowels, which are known to be challenging for Dutch learners both in producing and perceptually discriminating these despite their high proficiency in English (see above). **Chapters 2, 3, and 4** examined the extent to which learning in one modality could be supported by training in the other (i.e., training in perception influences production learning, and vice versa), if such cross-modality transfer is possible in both directions, and by which factors it might be influenced. More specifically, the aim of **Chapters 2 and 3** was to investigate the additional effect of relevant production practice of a trained speech contrast as integrated part of a perceptual training protocol on that contrast. To do so, we evaluated both the perception and production performance of Dutch native speakers who had undergone a four-day perceptual training on English /æ/ and /ɛ/ that was either combined with producing related (related production group) or unrelated speech tokens (unrelated production group). During each training trial, participants had to make a categorical decision based on stimulus words they were presented with auditorily. After visual feedback on their response, participants in the related production group produced a visually prompted word including one of the challenging English vowels, while participants in the unrelated production group produced a word not including any of the trained phonemes. By means of comparing training outcomes in both speech domains between the two groups, we attempted to tap into the influence of producing the trained words as opposed to engaging in speech perception during training more generally. **Chapter 2** reports behavioural findings of this training study, both in speech perception and production, while **Chapter 3** presents outcomes based on electrophysiological measures taken during and after the four-day training.

In **Chapter 4**, the focus turned to investigating effects of production training on both speech production and perception thereby complementing the examination of the cross-modality transfer from the previous two chapters. Here, Dutch natives participated in a two-day production training study, in which they received immediate, trial-by-trial visual feedback on their own productions of English words including either /æ/ or /ɛ/. The feedback consisted of a visual representation of mouth-tongue positioning during articulation and indicated how close a given utterance was with respect to that of a typical native speaker. Participants in a control group received a general indication on how in terms of tongue location the target vowels are pronounced by a typical native speaker but no direct feedback on the quality of their own vowel productions. Both groups received explicit phonological instructions on the challenging English contrast prior to training and their performance in both identifying and reading the vowels was measured before and after the two training sessions. The chapter thus had two aims, the evaluation of the effectiveness of the training tool and the examination of the degree to which learning in production would transfer to the perceptual modality.

Chapter 5 focussed on the verbal self-monitoring system and its role in the context of second language speech acquisition. More specifically, we tested how easily the self-monitoring system could adapt to evaluating newly-learnt non-native phonemes in order to support L2 speech category formation. Both previously trained participants (those reported on in **Chapters 2** and **3)** and participants in an untrained control group were tested in a fast-paced speech production task involving the trained vowels, English /æ/ or /ɛ/. During this phoneme substitution task, participants were visually presented with single English words, such as SAND, of which they had to mentally substitute the vowel by its respective counterpart (thus /æ/ by /ɛ/, and vice versa) followed by verbally producing the result of this substitution (in this case “send”). By means of time pressure and some catch trials not including any of the two vowels (to which participants had to respond by saying “no”), the task was designed to trigger verbal errors. The focus of this chapter was to compare electrophysiological signatures of error monitoring between a previously trained and an untrained control group in order to test the degree to which (trained) participants showed typical indicators of response evaluation of their erroneous verbal responses.

Finally, in **Chapter 6,** I summarise the overall findings of the empirical chapters and relate these to the ongoing scientific debate on perception-production interaction during second language speech learning. I thereby aim to further our understanding of how the relationship between the two speech modalities can best be characterised, which lines of further research will still be needed, and how a deeper understanding of perception-production interaction during L2 learning could inform the development of efficient training methods for non-native category formation.

2

Chapter 2

Perception and production in interaction during non-native speech category learning

ABSTRACT

Establishing non-native phoneme categories can be a notoriously difficult endeavour – in both speech perception and speech production. This study asks how these two domains interact in the course of this learning process. It investigates the effect of perceptual learning and related production practice of a challenging non-native category on the perception and/or production of that category. A 4-day perceptual training protocol on the British English /æ/-/ɛ/ vowel contrast was combined with either related or unrelated production practice. After feedback on perceptual categorisation of the contrast, native Dutch participants in the related production group (N=19) pronounced the trial's correct answer, while participants in the unrelated production group (N=19) pronounced similar but phonologically unrelated words. Comparison of pre- and post-tests showed significant improvement over the course of training in both perception and production, but no differences between the groups were found. The lack of an effect of production practice is discussed in the light of previous, competing results and models of second-language speech perception and production. This study confirms that, even in the context of related production practice, perceptual training boosts production learning.

Note that this chapter is based on:
Thorin, J., Sadakata, M., Desain, P., and McQueen, J. M. (2018). "Perception and production in interaction during non-native speech category learning," *The Journal of the Acoustical Society of America*, 144, 92–103.

I. INTRODUCTION

Mastering the sound system of a second language goes beyond the already non-trivial task of learning a new vocabulary and grammatical system. In many cases, it entails building novel sound categories. Many adult learners will experience this process as a major challenge, especially if the sounds of their native language only partly match those of their respective second language (Best, 1995). It remains to be established where exactly the learner's struggle to differentiate between specific non-native sounds, both in perception and production comes from. Putting it simply: Can they not hear the difference and therefore are unable to produce it, or vice versa? What effect does training one modality have on improving the other? Results in this field are still inconclusive. The goal of the present study is to further our understanding of second language (L2) sound learning and more specifically the nature of the relationship between speech perception and speech production in this process.

Various findings suggest an intimate relationship between the speech perception and production systems. There is extensive neurobiological and neuroimaging evidence showing automatic activation of brain areas related to speech production during numerous aspects of speech perception (reviewed in Skipper, Devlin, & Lametti, 2017). There is also evidence of direct links between an individual's perceptual and production abilities, such as auditory acuity influencing production variability (Brunner et al., 2011; Franken et al., 2017) and a listener's prototype for different speech categories correlating with the production of those categories (Newman, 2003). Well-known models of L2 speech perception and production assume a close link between the two systems, though they make different claims about the exact nature of this relationship. In his Speech Learning Model (SLM), Flege (1995) suggests that production accuracy might directly depend on the precision of someone's perceptual ability. Best and colleagues, however, claim in the context of their Perceptual Assimilation Model (PAM, as well as PAM-L2) that articulatory gestures are direct primitives of speech perception and that perceptual assimilations of speech sounds are thus driven by their articulatory features (Best, 1995; Best and Tyler, 2007a).

Both models predict that new phonemic categories can still be established throughout the lifespan. This prediction is in line with many findings supporting the view of a phonemic system that stays adaptable, though decreasing in flexibility with age (Flege et al., 1999a). Perceptual training of non-native sound categories has repeatedly been shown to successfully enhance both perception and production ability of those sounds for various combinations of L1 and L2. Examples are the frequently cited training of English liquids in Japanese learners (Bradlow et al., 1997), with retention effects after a 3-month period (Bradlow et al., 1999b), but also more recent training studies of French nasal vowels in US-American English learners (Inceoglu, 2016), English vowels in native speakers of Japanese (Lambacher et al., 2005b), English consonants trained in Spanish natives

(Lopez-Soto and Kewley-Port, 2009), and a Hindi voiced-prevoiced contrast in native English speakers (Baese-Berk, 2010). These successful training effects on the segmental level have also been extended to, for instance, non-native learning on the suprasegmental level with respect to Mandarin tones in native US-American learners (Wang et al., 2003), phonotactics (Kittredge and Dell, 2016), and syllable structure (Huensch and Tremblay, 2015). Remarkably, all of these studies show enhanced production without any direct training in this modality.

Outcomes have been more mixed concerning the reversed direction of transfer, that is, enhanced perception due to production training. Several studies showed successful transfer. For example, US American natives significantly improved their identification of a Spanish intervocalic three-way contrast after either production-only or perception-only training (Herd et al., 2013). Similar transfer effects from production training to perception were revealed when training English natives in the production of Japanese liquids (Hattori and Iverson, 2008) and of Japanese pitch and durational contrasts (Hirata, 2004a), and also when teaching French speakers production of four Danish vowels (Kartushina et al., 2015),

In other recent studies, however, potential advantages of production training for perceptual learning are not evident. Lu et al. (2015) compared discrimination ability in English learners of lexical tones after a single-day perception-only versus combined perception-production training and found similar improvement effects in perception for the two groups, thus no additional effect of production training. Herd et al. (2013) also tested a third type of training, in which production and perception training procedures were combined. There was no advantageous effect on perception of the trained Spanish contrast compared to the perception-only or production-only groups. As the authors note, however, this missing effect could be due to differences in amount of training, as the combined group received only half as much perception and production training as each of the one-domain training groups.

Interestingly, another line of research has revealed negative effects of additional production training on perception of non-native sounds. In a 2-day training protocol on a voiced-prevoiced contrast present in Hindi, native English speakers were trained in either a perception-only or combined perception-production paradigm (Baese-Berk, 2010). As mentioned earlier, results showed a clear transfer of perception-only training on production. Participants in the combined group, however, showed no improvement in discrimination ability between pre- and post-test measurements. As the author argues, participants' perceptual learning was thus disrupted by the additional involvement of production training.

More recently, Baese-Berk & Samuel (2016) replicated those results with a group of Spanish natives trained on a Basque fricative-affricate contrast. The design they employed was similar, though with a more active perceptual training regime, that is, a discrimination task with immediate feedback after each trial in contrast to passive exposure to a bimodal distribution of the to-be-trained contrast used in the earlier study. They further investigated potential causes for this disruptive effect and revealed that prior experience could reduce but not remove the negative effect of additional production training. In a separate experiment, in which they tested whether the disadvantageous training effects were due to general engagement of the production system (single letter production) or to specifically producing the to-be-learnt contrasts, they discovered that even unrelated production disrupted learning - though to a much smaller extent – and thus concluded that the disrupted perceptual learning is not simply related to participants listening to their own "bad" utterances.

One alternative explanation offered by Baese-Berk and Samuel (2016) for their findings is a potential difference in cognitive load between the two types of training. In all three experiments, participants in the combined training groups had to pronounce the target sound before making their perceptual judgment, whereas perception-only trained participants could either indicate their choice immediately after auditory presentation of the stimuli or, in the case of the unrelated production condition, simply produce a single letter displayed on the screen before making their choice. In both cases, a difference concerning the perceptual training itself instead of simply adding production practice to the paradigm was introduced, which makes it difficult to interpret the results. This variance could explain the difference in outcomes from the study by Lu et al., in which they found neutral effects of additional production training (though with Mandarin tones instead of Spanish consonants), but requires further investigation. When comparing the above-cited studies, it is also important to keep in mind that production training was implemented in different ways, as unlike for (high-variability) perception training there is as yet no well-established way of implementing production training. In order to test whether there is transfer from production to perception, it appears crucial to keep the task load, especially in the perceptual element of the training, identical across conditions.

In the present study, we investigated the effect of related production practice in a 4-day perceptual training protocol, involving minimal word pairs that contrast the English /æ/-/ɛ/ vowels, on the perception and production abilities of native Dutch speakers who were upper-intermediate/advanced L2 speakers of English. Cognitive load was carefully controlled for between two types of training. In the *related production* group, feedback on a perceptual categorisation task was combined with pronouncing the respective correct word on every trial, whereas in the *unrelated production* group it was combined with pronouncing a similar but phonologically unrelated set of words. The English /æ/-/ɛ/ vowel contrast (as in the words *pan* and *pen* respectively) is known to be challenging for even proficient Dutch speakers of English (Broersma, 2002; Escudero et al., 2008; Wanrooij et al., 2014), as their native vowel space exhibits a single category /ɛ/ (as in the Dutch word *pen*) that lies between the two English ones. Though the /æ/

category may already be weakly established in some (experienced) listeners, the two vowels are often confused (Broersma, 2005; Weber and Cutler, 2004). We sought to use a moderate amount of stimulus variability by employing multiple tokens of five minimal pairs recorded by four native speakers. This degree of variability takes into account the evidence that high stimulus variability is known to be advantageous for generalizability of the trained phonological contrasts (Lively et al., 1993; Logan et al., 1991), but also the finding that high variability can have harmful effects on the improvement in learners with relatively weak perceptual abilities (Perrachione et al., 2011).

We predicted improvement in both identifying and pronouncing the trained contrast due to the perceptual training. Such a finding would extend similar prior findings to another contrast and L1-L2 pairing with proficient L2 speakers. Such a finding would also show that transfer from perception to production can arise even when speakers engage in production practice, as is the case in real-world L2 learning. Predictions concerning the effects of production practice on the target contrast relative to unrelated practice, based on models of sound learning and previous findings, go in opposing directions. Production practice of the target phonemes could either help or hinder (or simply have no effect on) perceptual learning. According to the SLM, someone's perceptual ability limits the quality of their production and there would thus be no advantageous effects of production practice on perceptual learning. The PAM, in contrast, predicts transfer from production to perception. On the one hand, it seems reasonable to expect that production practice will have a positive effect on the quality of a learner's pronunciations, as practice usually improves the trained skill. On the other hand, exposure to potentially suboptimal examples of the vowel contrast (because the learners listen to their own voice) could have a negative effect on production, perception, or both.

II. METHODS

A. Participants

Thirty-eight native speakers of Dutch took part in the experiment (20 females and 18 males, mean age = 22.7 ± 3.7) and were paid or received course credit for their participation. None of them reported any history of neurological or psychiatric diseases, nor abnormal hearing ability. They were upper-intermediate/advanced L2 speakers of English (see TABLE 1). The Ethics Committee of the Faculty of Social Sciences at Radboud University, Nijmegen approved the study and all participants gave their written informed consent prior to the experiment.

B. Stimuli

All speech stimuli used in the experiment were based on recordings of 10 native speakers of British English born and raised in Southern England (5 females, mean age 24.8 ± 4.9).

As specified below, different ways of selecting and processing stimuli were used for each of the experimental tasks. Common preprocessing steps were band-pass filtering (50-8000 Hz) in order to reduce noise, and alignment in loudness by normalising based on root mean square amplitude.

TABLE I. Factors matched during group assignment.

Group	N	Gender (f/m)	Age	LexTALE	Pre-score identification (%)
Related production	19	10/9	23.2 (± 4.7)[n.s.]	80.7[a] (± 9.6)[n.s.]	75.8 (±10.6)[n.s.]
Unrelated production	19	10/9	22.2 (± 2.5)[n.s.]	76.3 (± 13.0)[n.s.]	76.1 (±11.0)[n.s.]

[n.s.] non-significant result of independent sample t-test comparing groups

[a] A LexTALE score of 80 falls at the boundary between upper intermediate and advanced users (Lemhöfer & Broersma, 2012)

Training and identification task

A set of 10 English ConsonantVowelConsonant (CVC) words contrasting the vowels /æ/ and /ɛ/ in five minimal pairs, *fan-fen, ham-hem, jam-gem, man-men, pan-pen*, was used. We restricted the final consonants to nasals in order to enable a transfer test to other phonemes after the training (see transfer conditions I-III). The full dataset, that is, 7 tokens of each of the 10 words pronounced by 4 different speakers (2 females and 2 males), consisted of 280 audio files. As non-native speakers have been found to rely more on durational differences between vowels and sometimes even exaggerate them in production, while English natives are more likely to attend to spectral differences (Flege et al., 1997), the training stimuli used here were duration-equalised in order to encourage learners to focus on more native-like distinguishing features. All recordings were normalised in length using Praat (Boersma and Weenink, 2015). This normalisation was based on average phoneme length across all tokens of the four speakers within one word pair, and resulted in the following durations: 565 ms (*fan-fen*), 504 ms (*jam-gem*), 530 ms (*ham-hem*), 533 ms (*man-men*), and 486 ms (*pan-pen*).

Identification and discrimination on morphed continuum

An eleven-step continuum between the English words /vɛt/ and /væt/ was created using TANDEM STRAIGHT (Kawahara and Morise, 2011) by adjusting both F1 and F2 values of the contrasted vowels. The two endpoints were duration-normalised recordings of one of the female speakers with a total duration of 632 ms.

Transfer identification and reading task

Six transfer categories were established by selecting stimuli which each represent a single new or adapted feature: (1) new starting consonant (C1): *tan-ten*, (2) new final consonant

(C2): *mash-mesh*, (3) new C1&C2: *gas-guess*, (4) length: *cattle-kettle*, (5) 2 new speakers: *pan-pen*, and (6) naturally-timed versions of the training set: *fan-fen, ham-hem, jam-gem, man-men, pan-pen*. Speakers were the same 2 males and 2 females who produced the training and test stimuli, except for the "new speakers" condition for which one new male and female voice was used. Per speaker there were 5 tokens used per word (n=20) resulting in a full set of 200 audio files. Apart from the last category, all stimuli were normalised in duration (again separately for each phoneme based on its average across tokens and speakers) resulting in the following durations for categories 1-5 respectively (in ms): 500, 585, 529, 518, 486. The naturally timed stimuli ranged from 450 to 650 ms.

C. Procedure

The full training paradigm consisted of several behavioural and EEG tasks on five separate sessions, in the order given in FIG. 1 (an additional EEG-based phoneme substitution task completed after all relevant post-tests in session 5 is omitted here; this was part of another study). The present paper presents the behavioural results only. All sessions for one participant were scheduled within 10 days, with maximally 3 days between two consecutive sessions. The duration of the sessions (including the additional task in session 5) differed between 2 and 3 hours with the first one being the longest.

In each session, participants were comfortably seated in a shielded room in front of a BenQ monitor (size 53.2 x 30 cm; 1920 x 1080 pixels; refresh rate of 60 Hz). All auditory stimuli were presented binaurally through in-ear headphones (Etymotic Research ER4P-T) at a comfortable volume for the participant (~25dB). All instructions and conversations during the experiment were held in English.

Group assignment was based on matching a combination of different variables prior to training, namely age, gender, English vocabulary knowledge quantified by LexTALE scores (see below) and pre-test identification scores, all summarised in TABLE I. None of the independent sample t-tests comparing each of these factors revealed any significant differences between the groups ($p > 0.05$).

The LexTALE task is a 2-minute test assessing lexical vocabulary size in English and is known to correlate with proficiency (Lemhöfer and Broersma, 2012). Participants were verbally instructed to read single words on the screen and to indicate by clicking either 'yes' or 'no' whether it is an existing English word or not. If in doubt, they were supposed to choose 'no'. A participant's score of correct answers was displayed on the screen after completion.

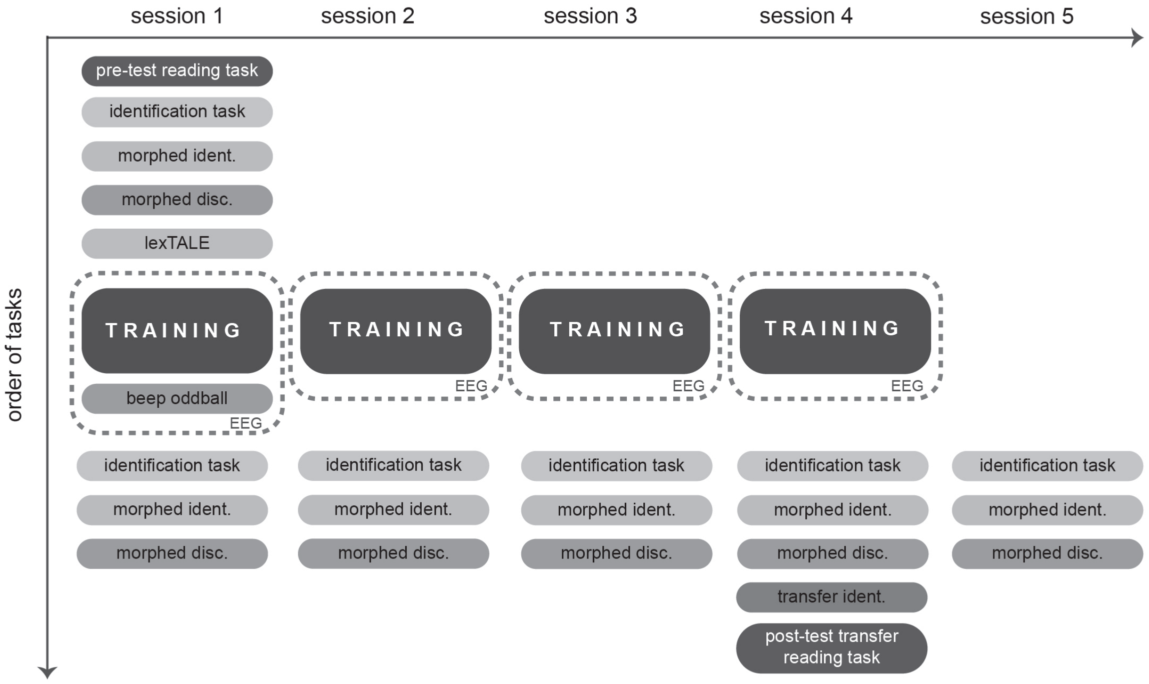

FIG. 1. Schematic timeline of the 5-day training paradigm consisting of several perceptual and production tasks conducted once prior to the full training and four times directly after a training session (post-test I-IV), as well as a delayed post-test and one set of transfer tests. Only type of training differed between the two experimental groups (i.e., related versus unrelated production practice).

D. Experimental Tasks

Training

The participants' task was to listen to sequences of English words, to indicate at the end of each sequence which word they heard last, and to then pronounce a single word shown to them on the screen. Each session consisted of 5 blocks of 40 trials. On each trial, participants listened to a sequence of 4-6 standard stimuli of the same word (multiple speakers and tokens mixed) followed by a final word that was either deviant (i.e. the standard word's minimal pair counterpart, e.g. *pen* for the standard *pan*; 75% of trials), or another version of the standard word (25% of trials). The interstimulus interval (ISI) was 300 ms, while the stimulus onset asynchrony (SOA) differed between trials depending on the duration of the minimal pair.

During auditory presentation, participants saw a fixation cross on the screen, which was then replaced by two words, the two members of the trial's minimal pair. Participants had to choose between the words in order to indicate which one they heard last. The orientation of the alternatives on the screen was counterbalanced between participants keeping the side of the /æ/- and /ɛ/-word constant for individual participants in order to avoid confusion with the button presses. Following a response, the selected word turned either green or red indicating a correct or incorrect response respectively, while the non-chosen word disappeared. After this visual feedback, a blue word appeared in the centre of the screen and had to be read out aloud. Depending on the type of training, this word was either the correct answer from the immediately preceding auditory sequence (for the related production group), or one out of another CVC minimal pair set not containing either of the target vowels (i.e., *shot-shut, hot-hut, cot-cut, dog-dug*, or *hog-hug* for the unrelated production group). After each block, the number of correct answers was displayed on the screen and participants could take a self-paced break.

Before the training, participants were given verbal instructions and a 5-minute practice task with unrelated stimuli (i.e., *bout-but*, *heat-height*). A full training session took approximately 50 minutes and EEG was recorded throughout all four of the training sessions. The task was run using a combination of the Matlab toolbox Brainstream and the Python based software package Psychopy (Peirce, 2007).

Identification task

For this two-alternative forced choice task, participants were instructed to listen carefully to single English words and then indicate by button press which of two visually presented words in a minimal pair they heard. The entire task consisted of a total of 120 randomly presented trials (10 words x 4 speakers x 3 repetitions) and lasted about 5 minutes. The total score of correct answers was presented to participants afterwards.

Identification on morphed continuum

In order to assess steepness and position of participants' categorical boundary between the two target vowels, participants also performed a two-alternative forced-choice identification task on a morphed phonetic continuum. On each trial, participants listened to one of the (morphed) stimuli on the /vɛt-væt/-continuum and then indicated whether they heard either *vat* or *vet* which were visually presented on the screen. The total number of 110 randomly presented trials (11 stimuli x 10 repetitions) took about 4 minutes to complete.

Discrimination on morphed continuum

Participants had to make a two-alternative choice based on auditory-presented words. We employed a 4-interval-2-alternative-forced-choice task (4I2AFC), in which participants heard a sequence of 4 words where either the second or the third stimulus was a deviant (i.e., AABA or ABAA; Gerrits and Schouten, 2004). Participants were asked to indicate the deviant's sequential location (i.e., '2' or '3'), by pressing a button. On each trial, two stimuli from the morphed continuum were presented. The pairings were created with a constant step size of 3 on the morphed continuum resulting in 8 possible pairings. In total, there were 96 randomly presented trials (8 contrasts x 2 orders x 2 deviant positions x 3 repetitions). The task took about 7 minutes to complete.

Reading tasks

Two versions of a reading task were employed: one pre-test version containing all 10 English training words and one post-test version, completed after the last training session, containing 8 additional words (the same as used in the transfer identification task: *tan, ten, mash, mesh, gas, guess, cattle* and *kettle*). In both versions, stimulus words were randomly presented individually on the screen and subsequently pronounced by the participants. In total there were 30 trials (10 words x 3 repetitions) or 54 trials (18 words x 3 repetitions) for the two versions respectively. Both versions were self-paced and took about 3-5 minutes.

III. RESULTS

A. Perception results

All responses for the pre-, post-(I-IV), delayed post and transfer-test identification task as well as identification judgments during the training and discrimination on the morphed continuum were transformed to d prime (d') scores based on hit and false alarm rates to /æ/-stimuli: d' = Z(hit rate) – Z(false alarm rate) with effective limits of 0.9999 for hit rates and 0.0001 for false alarm rates resulting in a highest possible d' score of 7.4380. Also the response bias c was calculated: c = - 0.5 *Z(hit rate) + Z(false alarm rate) (Macmillan and Creelman, 1991). For all statistical tests, whenever Mauchly's test of sphericity indicated that the assumption of compound symmetry did not hold, corrected p values according to the Huynh-Feldt approximation are reported.

Identification task (pre-post-test)

Group averages of d' scores for the six measurement times (pre-test, post-test I-IV and delayed post-test) can be found in FIG. 2. Individual participant data for the pre-test and delayed post-test are also shown. Results of a repeated measures analysis of variance (ANOVA) with the between-factor group and within-factor time revealed significant increases of d' in time (main effect time: $F(5, 175) = 24.96$, $p < 0.001$, eta^2 = 0.42), but no differences between the two groups for this change in time (interaction group x time: $F(5, 175) = 1.02$, $p > 0.05$).

A similar ANOVA on the bias term c also revealed a significant change in time, though with small effect size (main effect time: $F(5, 175) = 3.70$, $p < 0.05$, eta^2 = 0.10) and with no difference between the two groups (interaction group x time: $F(5, 175) = 0.50$, $p > 0.05$). Participants' bias changed from a tendency to identify stimuli as /æ/ words before the training (negative values of c) to a tendency towards /ɛ/ words (positive values after first training session).

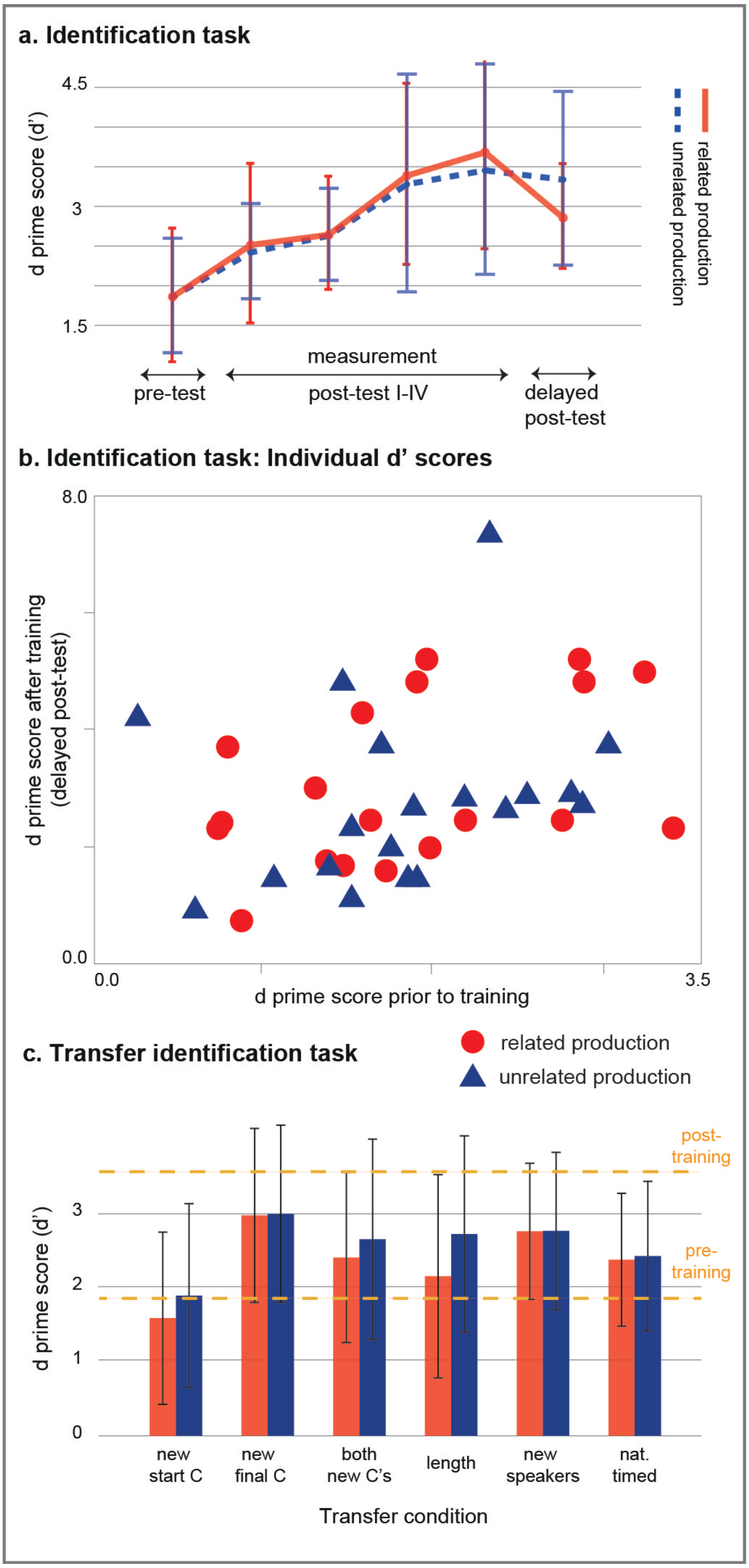

FIG.2. (a) Group average d' scores of the pre-test, post-test I-IV and delayed post-test measurements for the two training groups: related production versus unrelated production. Error bars indicate standard deviations across participants in given group. (b) d' scores of the individual participants during pre-test and delayed post-test (c) Average d' scores for the six transfer conditions. Horizontal, dashed lines indicate average d' scores on the training stimuli prior to training and after the last session respectively.

Identification task (during training)

For the identification judgments during training, a repeated measures ANOVA with between-factor group and within-factor time, showed a significant improvement of d' in the course of training (F(3,108) = 7.33, $p < 0.001$, eta^2 = 0.17) which again did not differ between groups (F(1,108) = 0.12, $p > 0.05$).

Transfer identification task

Testing for perceptual transfer effects of the training, we compared d' scores for each of the six transfer conditions with those in the identification task prior to the training and after the last training session (day 4) respectively (FIG. 2). Results of repeated measures ANOVAs are summarised in Table II. Overall, the training effects transferred to new kinds of stimuli. Participants scored significantly higher during transfer than in the identification task prior to the training in all except from one transfer condition: Identification of words starting with a consonant not included in the training did not improve. Scores on transfer tasks, however, were still significantly lower than post-training identification scores. The two groups did not differ in any of these effects.

TABLE II. Summary of statistical results regarding the transfer of identification ability.

	Post versus transfer				Pre versus transfer			
	Time		Interaction time x group		Time		Interaction time x group	
Condition	F(1,34)	p	F(1,34)	p	F(1,34)	p	F(1,34)	p
New start C	43.20	< 0.001	0.74	n.s.	0.26	n.s.	0.34	n.s.
New final C	6.55	< 0.05	0.19	n.s.	26.18	< 0.001	0.00	n.s.
Both new C's	14.26	< 0.001	0.58	n.s.	9.78	< 0.05	0.32	n.s.
Length	14.58	< 0.001	1.54	n.s.	7.22	< 0.05	1.75	n.s.
Novel speakers	12.42	< 0.05	0.15	n.s.	24.18	< 0.001	0.00	n.s.
Nat. timed	29.31	< 0.001	0.27	n.s.	9.51	< 0.05	0.01	n.s.

n.s. - non-significant result of repeated measures ANOVA.

Identification on morphed continuum

In order to quantify both sharpness and position of the category boundary on the 11-step /vɛt-væt/-continuum, we performed sigmoidal curve fitting using Matlab on the number of classifications per stimulus (see FIG. 3a). Resulting slope (boundary steepness) and 50% crossover point (boundary position) were used for further analyses.

Results of a repeated measures ANOVA on the slope, employing time of measurement as within-subject factor and group as between-factor, revealed no change of boundary steepness in time (F(5, 180) = 1.0, $p > 0.05$), nor any differences between the groups

(F(5, 180) = 1.0, $p > 0.05$). Similar null results were shown for 50% crossover point (main effect time: F(5, 180) = 1.0, $p > 0.05$; group x time interaction: F(5, 180) = 1.0, $p > 0.05$) indicating no shift in boundary position in the course of the training for either of the groups (FIG. 3a).

Discrimination on morphed continuum

A 3-way repeated measures ANOVA with the within-participant factors stimulus contrast pair (8 levels) and time (6 levels), and the between-participant factor group (2 levels) compared the d' scores. It revealed significant main effects for stimulus contrast pair F(7, 84) = 9.88, $p < 0.001$, eta^2 = 0.45) and time (F(5, 60) = 12.37, $p < 0.001$, eta^2 = 0.56). None of these effects differed between the two groups. Post hoc analyses comparing the pre-test score with the final post-test measurement only showed that those effects were driven by a higher percentage correct for stimulus pairs 5, 6, and 7 in the post-test ($p < 0.05$, corrected for multiple comparisons according to the Tukey-Kramer procedure). As higher numbered stimulus pairs were contrasting morphed stimuli closer to the /æ/-stimulus on the continuum, this reflects a shift of categorical boundary towards the /æ/ endpoint after training (FIG. 3b).

B. Production results

The speech data were analysed in two complementary ways, firstly by extracting and analysing the formant and duration patterns of the produced vowels and secondly by classifying the data in an automatic speech recognition (ASR) system. Due to high ratios of noise, some participants' data had to be removed from further analyses (resulting in N=15 and N=16 for the related and unrelated production groups respectively).

For the formant analysis, F1, F2, and vowel duration were automatically extracted using Praat (Boersma and Weenink, 2015). The extractions were based on manually segmented vowels (determined by visual inspection of both spectrogram and oscillogram), and were mean values across the 50% portion of the vowel centred on the vowel midpoint, therefore avoiding the vowels' border areas that could be affected by co-articulation. All formant values (in Hz) were transformed to log values for further processing, as those are known to better match the properties of the auditory system. The speech recordings obtained during training sessions were too noisy to be analysed.

Formant analysis

In order to quantify the distinctiveness between the two vowel categories regarding their first two formants, we used the Mahalanobis distance (Kartushina and Frauenfelder, 2014). This measure expresses the distance between a point and a distribution in a 2D-space, thus here the logF1-logF2 space (FIG. 4). For every participant and measurement time (pre-, post- and transfer-test), we calculated the distance between the centre of one vowel

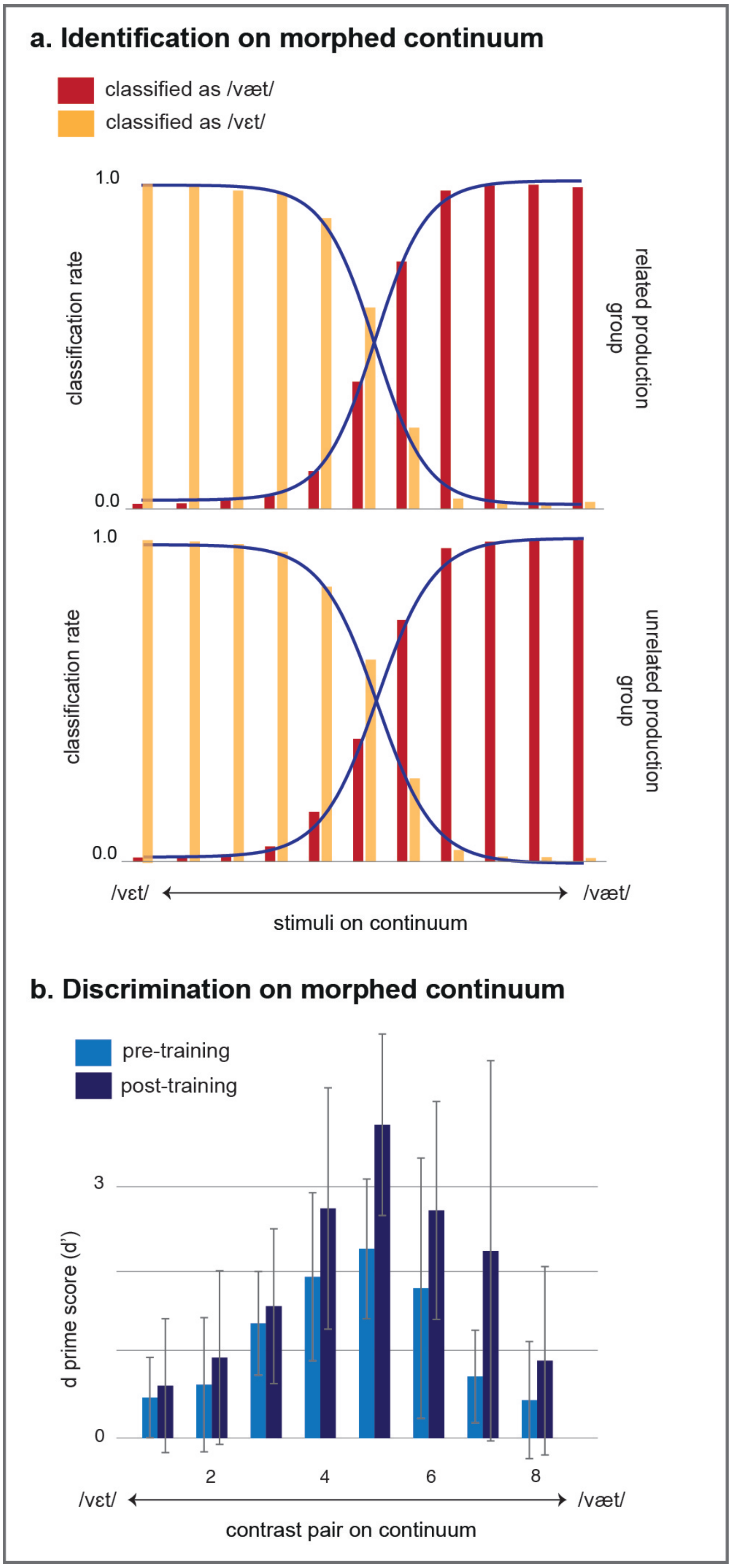

FIG. 3. (a) Grand average percentage correct identifications on the /vɛt/-/væt/ continuum for the two training groups separately (top: related production group, bottom: unrelated production group). Sigmoidal curve fitting of the classifications are indicated as lines. (b) Grand average d' score of discrimination between stimuli of 8 contrast pairs on the /vɛt/-/væt/ continuum (across measurements and training groups; there was no difference between groups). Standard deviations are indicated as error bars.

distribution and the respective other distribution, and vice versa. The mean Mahalanobis distance per participant in those two directions served as the dependent variable in a repeated measures ANOVA with group as between-participant factor and measurement time as within-participant factor. The test revealed a significantly larger distance between the pre- and post-measurement (main effect time: F(1,29) = 24.069, p < 0.001), but no difference between groups regarding this effect of training (interaction group x time: F(1,29) = 1.971, p > 0.05). The two vowel categories thus became more distinct after training in both groups.

Regarding transfer of training, a similar test revealed significant transfer of production learning to novel words: distances between vowels were significantly larger for the productions of the transfer words than for the pre-test words (F(1,29) = 27.227,

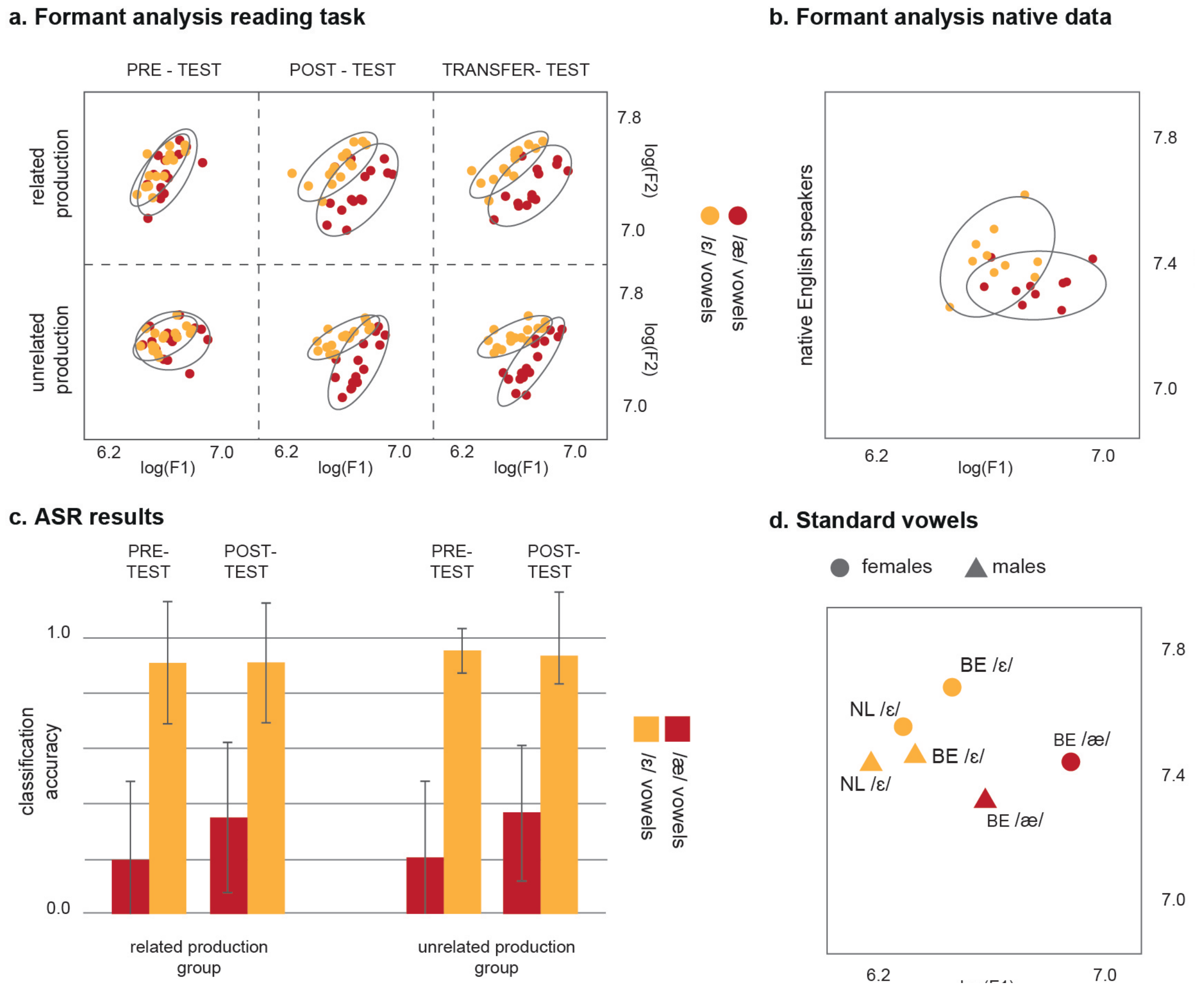

FIG 4. (a) Log(F1) and log(F2) values for the two English vowels pronounced in CVC words of pre-, post- and transfer reading tasks (from left to right column) after either related (top) or unrelated production training (bottom). (b) Log formant data of the two vowels pronounced by 10 British English speakers. (c) Classification results of the two English vowels by the automatic speech recognition system before and after training, separately for the two training groups. (d). Standard formant values log(F1) and log(F2) for the two British English (BE) vowels, /ε/ and /æ/, and the similar Dutch (NL) vowel category /ε/ (based on Deterding (1997) and Adank, Hout, & Smits (2004) respectively).

$p < 0.001$). Even though the mean logF1 and logF2 values per participant seem to show similar patterns for post- and transfer-test, the Mahalanobis distance, taking into account an individual's variability in production, is still significantly smaller compared to the post-test distances ($F(1,29) = 10.82$, $p < 0.01$) indicating that the transfer is incomplete. There were no group differences in either of these effects (pre- versus transfer-test: $F(1,29) = 0.164$, $p>0.05$; post- versus transfer-test: $F(1,29) = 2.41$, $p > 0.05$).

Durational analysis

To check for any potential influences of the duration normalised training stimuli on the durational distinction participants made when producing the two vowels, we compared differences in vowel duration in a repeated measures ANOVA with the between-factor group and the within-factor measurement time (pre versus post). The results showed that the durational distinction was significantly larger after training (main effect time: $F(1,29) = 9.523$, $p < 0.01$, $eta^2 = 0.25$), though with a relatively small effect size. There was no difference between the groups regarding this effect (interaction group x time: $F(1,29) = 0.115$, $p > 0.05$).

Automatic speech recognition

In the second approach to analyse the production data, we employed an automatic speech recognition (ASR) system specifically trained on the ten minimal pairs used in the training and pre-test reading task[1]. The model was created using the Hidden Markov Model Toolkit (Young et al., 2009) and trained on the speech data of all 10 British English native speakers (10 speakers x 10 stimulus words x appr. 10 tokens = appr. 1000 words). In order to identify native-like utterances in the reading tasks, the ASR system was then used to classify the English pronunciations by the Dutch speakers of this study. For this purpose, the model was restricted to two classes per trial (one minimal pair) in order to avoid classification errors due to other aspects than the quality of the vowel itself. The resulting classification accuracy of a 5-fold cross-validation procedure with the English training data was 86% and judged to be sufficiently high to employ the model as a tool for automatically validating the reading task data in this study.

Correct responses for word productions in the following analyses were therefore defined as trials, in which the word that had to be produced by participants was the same as the one classified by the system (FIG. 4c). Results of a 3-way repeated measures ANOVA employing the factors time (2 levels) x vowel (2 levels) x group (2 levels), showed

[1] The acoustic model consisted of a set of single-Gaussian monophone HMMs. The HMM's topology was a 3-state left-right model with no skips, where each data vector contained 13 MFCC coefficients, plus the corresponding delta and acceleration coefficients. The MFCC coefficients were calculated using a frame length of 10ms, a Hamming window, first-order pre-emphasis, and a filter bank of 26 channels.

significant main effects of measurement time ($F(1,29) = 21.89$, $p < 0.001$, $eta^2 = 0.43$) and vowel ($F(1,29) = 84.70$, $p < 0.001$, $eta^2 = 0.75$), as well as an interaction effect for time and vowel ($F(1,29) = 27.89$, $p < 0.001$, $eta^2 = 0.49$). A post hoc analysis revealed that this interaction was driven by a significantly larger percentage of native-like validated word productions containing the /ɛ/-vowel after training ($p < 0.001$).

Correlation analysis: Perception and production data
A two-tailed correlation analysis between the learning effect in perception (differences in d' score between pre-measurement on day 1 and after last training session on day 4) and learning effect in production (difference between Mahalanobis distance before and after training) revealed no significant relationship ($p > 0.05$).

IV. DISCUSSION

This study investigated how the domains of speech perception and speech production interact in the course of learning the British English /æ/-/ɛ/ vowel contrast by native speakers of Dutch. More specifically, it aimed at evaluating the effect of related (as opposed to unrelated) production practice during a 4-day perceptual training on perception and production of this contrast.

The two training groups clearly improved their perceptual abilities in the course of training. This improvement further validates the effectiveness of multiple-day perceptual training paradigms (Bradlow et al., 1997; Rato, 2014). It thereby confirms findings that non-native learners can still establish novel sound categories in adulthood (e.g. Bradlow et al., 1997; Lambacher et al., 2005; Inceoglu, 2016). The perceptual enhancement also transferred to new stimuli and speakers suggesting the formation of phonologically abstract categories (Sadakata and McQueen, 2013). It is noteworthy that participants' performance on the transfer task is still lower than their post-test performance on the trained stimuli. This finding indicates that the learning is not purely abstract in nature but instead is also tied in some way to the specific training words. It can also be argued that the variability in the training stimuli was not sufficiently high for a robust generalisation of the target vowels. The variability was notably lower than many studies with the high-variability paradigm, such as the one by Bradlow et al. (1999) with 68 minimal pairs for two liquids spoken by 5 speakers, or the one by Wong (2013) with 20 minimal pairs produced by six speakers.

Learners of both groups also clearly improved in the production domain showing more distinct and more native-like pronunciations of the two vowels after training. However, neither in production nor in perception did the outcomes of the training differ between the two groups. Related production practice in the current experiment could not be shown to affect learning in either of the two domains. Perceptual learning in both groups, that is improvement independent of related training in production, is in line with similar comparisons of perception-only versus combined perception-production training (Herd et al., 2013; Lu et al., 2015). Although we cannot exclude entirely that production learning is due to engagement of the general articulatory system, as both types of training in the current design involved word production, it seems unlikely that learners improved the pronunciation of the target vowels simply by producing unrelated words. If that were the case, it seems surprising that the trained phoneme contrast was still relatively poorly established prior to training. It is much more likely that the production enhancement is due to transfer from perceptual learning.

This successful transfer from perception to production again replicates earlier findings (e.g. Lopez-Soto & Kewley-Port, 2009; Wang, Jongman, & Sereno, 2003) and extends them to another non-native speech contrast with proficient L2 speakers. Despite the overall transfer from perceptual to production learning, there was no direct correlation between the improvements in the two domains. This finding is in agreement with many earlier approaches investigating the relationship between perception and production (Bradlow et al., 1997; Huensch and Tremblay, 2015; de Jong et al., 2009) and could be interpreted as the absence of a direct link between the two systems. This interpretation would resonate well with Flege's notion (1995) that the production and perception systems might not be brought into perfect alignment, as occurs in L1 speech acquisition.

One of our aims was to add to the discussion on whether related production practice in a perceptual training protocol either helps or hinders perceptual and/or production learning. Because of the current null findings concerning the differential effect of training type, we are not able to draw any final conclusions on this matter. Related production practice could potentially have a negative effect on both perceptual and production learning due to increased cognitive load during training, and on production specifically given the exposure to bad examples of the to-be-learnt phonemes as part of learner's listening to their own speech. In the current study, however, we could not replicate the negative effect of combined perception-production training on perceptual learning of non-native categories shown by Baese-Berk and Samuel (2016). The most crucial difference between their design and the current one (as well as those of Herd et al. and Lu et al.) is that learners had to produce tokens of the target contrast *before* making, or at least indicating, a categorical decision. This additional production of a challenging contrast could have increased cognitive load during the perceptual task. Earlier research indicates that cognitive load can reduce perceptual acuity during different kinds of speech discrimination tasks (Mattys et al., 2014; Mattys and Wiget, 2011) and might result in competition for working memory processes at the encoding stage (Mitterer and Mattys, 2017). Based on those findings, the increased task load in the production practice condition in Baese-Berk and Samuel (2016) is likely to result in suboptimal encoding of the trained contrast.

Baese-Berk and Samuel (2016) show that producing tokens of the to-be-learnt contrast disrupted perceptual learning to a stronger degree than producing unrelated utterances. They interpret this effect as evidence for the production of the contrast itself causing the disruption of perceptual learning. One could again argue, however, that this difference is due to differences in cognitive load, as it is to be expected that producing words containing a challenging non-native sound will disturb ongoing perceptual categorization to a stronger degree than producing single letter strings. Furthermore, prior experience with the to-be-learnt contrast was shown to have an alleviating effect on the disruption of perceptual learning (Baese-Berk & Samuel, 2016). Once again, however, perceptual learning could be hindered more by the production of a challenging and novel contrast than by one that is already known to some degree.

Intuitively, it would make sense to expect improvement of a skill due to practicing it, but we could not find evidence for any additional effect of related production practice within the training period. While some previous studies did not measure the effects of a combined training protocol on production, the two that did do so (Baese-Berk, 2010; Herd et al., 2013) show similar results. This outcome could be explained in different ways. Firstly, production learning could be driven purely by perceptual improvement, as suggested by the SLM (Flege, 1995). Transfer from production to perception without any perceptual training (e.g., Herd et al., 2013), however, speaks against this possibility. There are also various studies in which speech production (of non-native contrasts) is successfully trained, for instance in an efficient computer-based system training Mandarin and Cantonese native speakers in three English vowel contrasts (Wang & Munro, 2004), or in a training system providing trial-by-trial visual feedback on the production accuracy of Danish vowels by native French speakers (Kartushina et al., 2015).

These successful training examples tie directly to a second explanation of the current findings. A crucial aspect of successful production-training studies is that learners receive immediate and informative feedback on their pronunciation. Practicing a skill is only beneficial if the practice itself is efficient. In the current study (and Lu et al., 2015, Baese-Berk & Samuel, 2016; Baese-Berk, 2010), participants did not receive any external feedback on their utterances. Internal feedback on one's own production might simply be insufficient in triggering actual improvement in production learning, as it requires a satisfactory degree of perceptual skills when evaluating the self-produced utterances. Any positive effects of simple production might even be counteracted by increased exposure to bad examples of the to-be-trained contrast, as learners are listening to their own utterances (though there is evidence suggesting that this effect is unlikely, see Kraljic & Samuel, 2005). In the context of investigating effects of combined perception-production training, related productions were followed by feedback only in the study by Herd et al. (2015). After producing the target word, participants had to visually compare their own utterance to that of a native speaker. Despite this feedback, the results did not show any additional benefits of production. In order to disentangle whether absent effects of additional production training are due to either no or insufficient external feedback, it will be important to directly test the effects of explicit and informative feedback in a similar design to that used here.

Another aspect in the interpretation of effects due to related production practice is the factor time. All of the above studies investigating effects of combined perception-production training differ substantially in duration and amount of training. They range from a single session (Lu et al., 2015), over 2-session paradigms on consecutive days (Baese-Berk, 2010) or days separated by 48 hours (Baese-Berk and Samuel, 2016) to 6 training sessions during a period of 2-3 weeks (Herd et al., 2013). Interestingly, an additional day of training in Baese-Berk (2010) did reduce the disadvantage of perceptual learning due to combined perception-production training in their design. This finding could again be accounted for by a reduced effect of cognitive load during perceptual processing, assuming that the training protocol demands less capacity the more experienced learners become with it. Alternatively, production learning might take place on a different, namely slower, time scale than perceptual learning of a non-native contrast. Harmful effects were revealed by short training procedures and might disappear after 3 or more days of training (also depending on the difficulty of the to-be-learnt contrast). A strength of the current study is the relatively long duration of training. Although we do not have data on the exact timecourse of production learning in the course of the present 4-day training protocol, there are no indications for differences between the two groups in terms of their perceptual learning curve. It seems thus unlikely that a potentially harmful effect would be due to differences in timecourse.

The results also have implications for the nature of the perceptual improvement. In particular, learners showed a boundary shift in the discrimination task. This shift is interesting for different reasons. Firstly, it is noteworthy that there is a clear boundary effect detected in the first place. In the 4I2AFC design used in the discrimination task, listeners usually tend to make non-categorical responses based on low-level acoustic differences between the presented stimuli (Gerrits and Schouten, 2004; Sadakata and McQueen, 2013). Use of this task, however, will not entirely prevent listeners from using any (even weakly established) category knowledge. As can be seen in FIG. 3b, the vowel stimuli used here did indeed encourage listeners to make use of their boundary knowledge. This task characteristic compensates for the low sensitivity of the identification task on the morphed continuum, in which neither changes in boundary sharpness nor boundary position were detected in the course of the training. In discrimination, however, both training groups show a peak before training, indicating the existence of /ɛ/ and /æ/ categories, and a boundary shift towards the /æ/-endpoint after training, indicating a perceptual restructuring as the /æ/ category becomes stronger. The relatively high performance on the identification task prior to training also suggests that, at least in perception, L2 learners already had a weak /æ/ category at the start of the experiment.

In the production domain, however, the /æ/ category appears to be less well established (see FIG. 4a). Participants started out with relatively accurate productions of the /ɛ/-vowel prior to training, while its counterpart /æ/ was not clearly distinguished from those productions. Patterns of the production learning reveal that the two non-native categories develop in an asymmetrical fashion. This development makes sense given the location of the relevant English and Dutch categories in vowel space. Though the realisations of the English and Dutch phoneme /ɛ/ are not identical, the Dutch /ɛ/ lies closer to the English /ɛ/ than to English /æ/, as can be seen in FIG. 4d. This tendency of native Dutch speakers to map the non-native /ɛ/ to their similar native one can also be found in, for instance, results from a lexical decision task. Here, Dutch participants showed a tendency to classify non-existing words as real words, when an /ɛ/ vowel in an existing English word was replaced by an /æ/, such as in *dask* (Broersma, 2002). Similarly, in a visual word paradigm initial parts of distractor words containing the /æ/ vowel, such as *pan-* in the word *panda*, activated the word *pencil*, while the opposite, activation of *pencil* by the distractor *panda,* was not the case (Weber and Cutler, 2004). These findings suggest that, while Dutch listeners can hear the difference between /æ/ and /ɛ/ (otherwise the results for *panda* and *pencil* would have to be symmetrical) there are nonetheless strong effects of native categories on perception. In line with PAM predicting that unfamiliar non-native categories are assimilated by close native categories, examples of the English /æ/ vowel tend to be collapsed into the /ɛ/ category, while the reverse assimilation is less likely. This process is reflected in our pre-test production data. But the pre-test identification and discrimination findings suggest that there is already at least a weak perceptual category for /æ/. These findings indicate that perceptual and production learning might follow different time-courses.

The Dutch learners changed their perceptual cue weighting of the English /æ/-/ɛ/ contrast in the course of this training. It is known that non-native listeners of a vowel contrast tend to rely more on durational differences than on the spectral differences that are more important for native listeners (Flege et al., 1997). Any durational cues facilitating the differentiation of the two trained vowels (the English /æ/ is usually longer than its counterpart /ɛ/) were removed from the training stimuli in the current design. Perceptual categorisations made by the learners in this study were thus likely based on spectral differences. Despite being trained on duration-normalised examples, participants did not reduce the durational distinction made in their productions of the vowels; that is, they start out with longer /æ/'s than /ɛ/'s and show a more native-like pattern after the training (i.e., they increased the durational difference). The successful change to (more) native-like phonetic cue weighting due to perceptual training is in line with earlier findings (Hu et al., 2016; Ylinen et al., 2010). Most interestingly, it further confirms that listeners are able to rely on some prior knowledge regarding the distinction between the two vowel categories in perception that goes beyond the spectral differences that they were exposed to. That is, at least in perception, participants start out with some concept of the perceptual categories for both vowels, which is then further strengthened in the course of training and successfully transferred to the production domain.

V. CONCLUSION

The current study confirms that perceptual training boosts production learning. Learners can evidently improve their production of a challenging non-native vowel contrast by training their perceptual categorisation ability, which corroborates the view that perceptual enhancement tends to support and to precede production learning. Related production practice, however, did not lead to additional improvement in either of the two speech domains. In order to further clarify potentially beneficial effects of combined perception-production training protocols, we recommend the study of explicit and informative feedback on participants' productions during a similar training study. Until then, the question remains open whether production training leads to improved category formation in either perception or production. What the current results already indicate, however, is that perceptual training improves production in the context of production practice. This context is the one present in natural L2 learning, where the learner is trying to improve both speaking and listening skills.

VI. ACKNOWLEDGMENTS

This research was supported by a PhD grant of the Language in Interaction Consortium (https://www.languageininteraction.nl/) to the first author. Special thanks go to Philip van den Broek for his technical support, Matthias Franken for his work on the automatic speech recognition system, as well as Malin Mangersnes, Mariia Naumovets, Dennis Uhrig, Inez Wijnands, Josh Ring, and Eliana Garcia-Cossio for their help with data collection. We thank three reviewers for their constructive feedback. Preliminary results related to this study were presented during the Society of Neurobiology of Language 2016 and Psycholinguistics in Flanders 2017.

3

Chapter 3

Perception-production interactions in non-native sound learning: EEG evidence

ABSTRACT

This study explores how speech perception and speech production interact during the learning of non-native phoneme categories. We evaluated neurophysiological signatures during and after a 4-day perceptual training protocol that was complemented by production practice on words that were either related or unrelated to the training materials. Sequential unbalanced bilinguals of Dutch (L1) and English were trained on the British English /æ/-/ɛ/ contrast. Despite no behavioural differences in training-related improvements between the two training groups (earlier presented in Thorin et al. 2018), the related production group showed a mismatch negativity (MMN) response to the English /pæn/-/pɛn/ contrast that was absent in the unrelated production and untrained control groups. This suggests that positive effects of perceptual training combined with related production practice can become apparent if the measurements taken are sufficiently sensitive to identify fine-grained differences in perceptual ability. These differences might not be detectable by conventional behavioural methods.

Please note that this chapter is based on:
Thorin, J., Garcia-Cossio, E., Sadakata, M., Desain, P., and McQueen, J. M. (under revision). "Perception-production interactions in non-native sound learning: EEG evidence.

I. INTRODUCTION

Successfully learning a foreign language in adulthood naturally involves improvements in both speech perception and speech production. A major challenge for many late bilinguals is that they can neither hear the difference between specific non-native phonemes nor pronounce those phonemes reliably, especially when those are assimilated into a single category of the learner's native language (L1) (Best, 1995; Best and Tyler, 2007a). How exactly the two speech modalities interact when solving this difficulty in the process of developing novel sound categories is still unclear. More specifically, it is inconclusive whether training in one modality improves performance in the other modality, and whether it is beneficial (or indeed detrimental) to combine training in the two modalities. The goal of the present study is to further our understanding of these interactions by investigating neurophysiological changes in unbalanced bilinguals who underwent additional training on relevant non-native phonemes.

Although the precise nature of the relationship between the perception and production modalities remains to be further explored, there is sufficient evidence indicating that they interact on various levels rather than being fully independent. The involvement of the speech production system during various stages of auditory speech perception has become evident in a large body of behavioural, neuroimaging, and lesion studies as well as contributions from computational modelling (reviewed in Skipper et al., 2017). Some behavioural examples include adaptations to altered auditory feedback resulting in changes of auditory speech perception (Lametti et al., 2014) and manipulations of listener's facial muscles biasing their perception of words towards sounds that are more aligned with their somatosensory input (Ito et al., 2009). Close links between the two speech modalities were also observed in studies on individual differences revealing correlations, for instance, between a speaker's variability in phoneme productions and their perceptual acuity (Brunner et al., 2011; Franken et al., 2017) and between a listener's perceptual prototypes of a phoneme category and their average production of that category (Newman, 2003). A recent factor analysis including measurements from various linguistic and general sensory tasks has shown a close relationship between the perception and production modalities present in links between phonological processes in L1, second language (L2) and an unknown language but absent in non-linguistic skills, such as audio-visual or sensory-motor processing, suggesting that the nature of this relationship is language-specific (Schmitz et al., 2018).

When learning a second language, the close perception-production link has repeatedly been shown to benefit learners of non-native sounds in the form of transfer from perceptual learning to improvements in production (reviewed in Sakai and Moorman, 2018). Available examples of perceptual training studies comprise various combinations of L1 systems and trained L2 sounds, such as native Japanese learners trained in the perception of English liquids (Bradlow et al., 1997, 1999b) and English vowels (Lambacher et al., 2005b), native English speakers learning a Hindi voiced-prevoiced contrast (Baese-Berk, 2010) and French nasals (Inceoglu, 2016), Spanish natives learning English consonants (Lopez-Soto and Kewley-Port, 2009), Korean natives trained in English vowels (Lee and Lyster, 2017), Mandarin and Cantonese learners of English vowels (Wang and Munro, 2004) and Russian natives trained in various English phonemes (Qian et al., 2018). This already wide range of phonemic contrasts perceptually trained on the segmental level has been further extended to the suprasegmental level in the form of, for instance, Mandarin tones in native US-Americans (Wang et al., 2003), as well as to phonotactics (Kittredge and Dell, 2016), and syllable structure (Huensch and Tremblay, 2015).

Complementing the robust transfer from perceptual to production learning, there are also examples of perceptual gains resulting from isolated production training. This has been observed, for instance, with Danish and Russian vowels in native French speakers (Kartushina et al., 2015, 2016b), Japanese pitch and durational contrasts in English natives (Hirata, 2004b), English liquids in Japanese natives (Hattori and Iverson, 2008), and a Spanish intervocalic three-way contrast in US-Americans (Herd et al., 2013).

It seems plausible, on one hand, that combining perceptual paradigms with some form of production training would be beneficial, as (additional) training is usually a good predictor for improvement, and the complementary training in and transfer from both modalities could strengthen their reciprocal relationship. On the other hand, a more complex training paradigm could lead to less efficient learning, while the exposure to bad examples of the to-be-learnt non-native phoneme when listening to self-produced speech could counteract the evidently positive effects of perceptual training. In keeping with these divergent theoretical accounts, outcomes from different studies investigating combined versions of perception and production training go in opposite directions and thus paint a more inconclusive picture than that derived from single-domain training paradigms.

Comparing outcomes of perception-only, production-only and combined perception-production training, Herd et al. (2013) trained native speakers of American-English on a Spanish intervocalic three-way contrast during 6 sessions in a period of 2-3 weeks. Training only in perception and training only in production both strengthened processing in the trained modality and transferred to the other modality, while the degree to which learning in one modality transferred to the other one strongly depended on the phonological relationship between the trained sounds. Even though combined perception-production training was most efficient in improving production as compared to the two single-modality training conditions, it resulted in no additional gains in perception. The authors stress, however, that the lack of gains in perception could be due to the fact that participants in the combined training received only half of the training in each modality compared to the single-modality training groups. Amount of perceptual training was more balanced in a study by Lu et al. (2015) using Mandarin tones when comparing outcomes of perception-only and combined perception-production training (with imitation as production element) on tone discrimination in a single-day paradigm. Here both groups similarly improved in their perceptual ability (in a discrimination task), thus showing neither positive nor negative effects of additional production training.

Negative effects of additional production practice in the context of perceptual learning were found for a Hindi voiced-prevoiced contrast in native English learners (Baese-Berk, 2010) and for a Basque fricative-affricate contrast in native speakers of Spanish (Baese-Berk and Samuel, 2016). In both studies, multiple-day combined perception-production training led to no improvements in discrimination ability comparing pre- and post-training measurements, despite clear gains after perception-only training. When further investigating the reasons for this disruptive effect, it was revealed that negative effects on perceptual learning could be reduced by prior experience with the trained non-native contrast. The disruptive effect, however, was still present – though to a smaller degree – when perceptual training was complemented with a general production task unrelated to the trained non-native contrast instead of with productions of the trained contrast. This led the authors to conclude that the negative effect was due to more general interference from the production system instead of being caused by more specific interference due to learners' exposure to their own suboptimal utterances. An alternative explanation to these findings was offered by the authors and points to differences in cognitive load between the two training conditions, which could have led to reduced perceptual acuity during training followed by suboptimal encoding, as suggested by other findings (Mattys et al., 2014; Mitterer and Mattys, 2017).

Thorin et al. (2018) therefore balanced cognitive load between training conditions, when similarly investigating the effect of additional production practice in the context of perceptual learning (note that the present study is based on the same dataset). During their 4-day training paradigm, native speakers of Dutch and sequential unbalanced bilinguals were trained in the identification of the English /æ/-/ɛ/ vowel contrast by receiving perceptual training that was complemented with production practice on words that were either related or unrelated to the training materials. There was perceptual learning in the course of training for both groups, which also transferred to production improvements (quantified as increased distance between the formant values of the two vowel categories in F1-F2 space). Interestingly, there were no behavioural differences for the two production practice groups on any of the task outcomes including abilities in identification, discrimination, production and transfer to novel stimuli in both modalities. The findings of both Thorin et al. (2018) and Lu et al. (2015) thus point towards neutral – as opposed to disruptive – effects of additional production training or practice, but have to be taken with caution given that this interpretation is based on null results in both studies.

It is possible that these null results reflect lack of sensitivity in the measures that were taken. There are reasons to assume that EEG measurements might be more sensitive than behavioural outcomes in detecting subtle effects of training or practice. Specifically, the auditory mismatch negativity (MMN) is known to be a useful tool in measuring automatic auditory change detection even in the absence of attention. This event-related potential (ERP) is a negative-going deflection in the difference wave between responses to frequently presented standard stimuli and infrequently presented deviant stimuli, typically peaking around 150-250 ms after stimulus onset (e.g. Näätänen et al., 1997, 2007). It has found wide applications in the evaluation of listener's ability to hear differences between various types of auditory input ranging from complex auditory patterns (Atienza et al., 2002; Näätänen et al., 1993), through music (Fujioka et al., 2004; Koelsch et al., 1999) to phoneme discrimination (Bomba et al., 2012).

The MMN has also repeatedly been used to evaluate native-likeness of L2 learners' ability to discriminate between non-native phonemes, both in children (Peltola et al., 2005) and adults (Grimaldi et al., 2014; Peltola et al., 2003, 2005; Rivera-Gaxiola et al., 2000), and to examine individual differences in non-native phoneme processing (Díaz et al., 2016; Jakoby et al., 2011b). In the same context, it has also been a tool in assessing L2 training outcomes complementing behavioural findings (Lu et al., 2015; Tamminen et al., 2015; Ylinen et al., 2010; Zhang et al., 2009). Interestingly, the ability to discriminate between a challenging non-native phoneme contrast quantified as a stronger MMN response has even been shown to precede behaviourally measured improvements in the course of perceptual L2 training (Tremblay et al., 1998). The MMN thus has the potential to offer a valuable window into the time course of non-native speech learning and represents a potentially more sensitive measurement of the additional effect of production training in the context of perceptual training of a non-native speech contrast.

The present report compares neurophysiological signatures related to the process of learning to perceive (and produce) the English /æ/-/ɛ/ vowel contrast in Dutch native unbalanced bilinguals undergoing perceptual training. Feedback on phoneme categorisations in each trial was either followed by pronouncing words including the to-be-trained vowels (related production group) or by pronouncing a set of similar but irrelevant words (unrelated production group). These neurophysiological measurements were recorded during the study reported in Thorin et al. (2018). The phoneme contrast was chosen as even proficient Dutch speakers of English are known to have difficulties differentiating between the English /æ/ and /ɛ/ as in the words *pan* and *pen* (Broersma, 2002; Escudero et al., 2008; Wanrooij et al., 2014). The reason for this confusion is that the Dutch phonological system has a single vowel category /ɛ/ (as in the Dutch cognate *pen*) that lies in between the two English vowels. Furthermore, because the Dutch /ɛ/ lies closer to but is not identical to the English /ɛ/, the misperception for native Dutch speakers can be asymmetrical in nature, even though the /æ/ category might be already weakly established in some (experienced) learners (Broersma, 2005; Weber and Cutler, 2004). While Thorin et al. (2018) focussed on the behavioural outcomes of the perceptual training paradigm, the current report presents findings based on the EEG measurements that were recorded before, during and after the same perceptual training.

Hypotheses regarding the additional effect of production practice on perceptual discrimination ability quantified as the strength of MMN response go in two directions. On the one hand, learners in the related production group could be hindered by the additional involvement of the production training and be negatively reinforced by the oftentimes suboptimal examples of their own vowel pronunciations. On the other hand, learners in the related production group could benefit from the additional practice in the production modality that could readily transfer to perception and thus strengthen the outcomes of the perceptual learning. In the latter case, we would expect to see a difference in MMN to the trained English stimuli between the two training groups after the training and potentially already emerging during training.

II. METHODS

A. Participants

Fifty-four sequential unbalanced bilinguals who were native speakers of Dutch and upper-intermediate/advanced L2 speakers of English took part (see TABLE 1 for participant details including English proficiency measures). Thirty-eight of those, namely the participants of the two training groups, were the same individuals whose behavioural data were presented earlier (Thorin et al., 2018). The other 16 participants were assigned to the control group. None of the participants reported any history of neurological or psychiatric diseases, nor abnormal hearing ability. All participants were compensated for their participation. The study received approval by the Ethics Committee of the Faculty of Social Sciences at Radboud University, Nijmegen, and all participants gave their written informed consent prior to the experiment.

TABLE 1. Participant information for the three groups. N.S indicates non-significant results of Two-way ANOVA comparing groups.

Group	N	Gender (f/m)	Age	LexTALE*
Related production	19	10/9	23.2 (±4.7)$^{n.s.}$	79.4 (± 9.6)$^{n.s.}$
Unrelated production	19	10/9	22.2 (±2.5)$^{n.s.}$	76.3 (±13.0)$^{n.s.}$
Control	16	9/7	22.8 (±2.7)$^{n.s.}$	82.5 (±13.3)$^{n.s.}$

*LexTALE is a brief computerised task assessing vocabulary knowledge of English and is known to correlate with general English proficiency. A score of 80 marks the boundary between upper intermediate and lower advanced learner (Lemhöfer and Broersma, 2012).

B. Stimuli

Behavioural Stimuli

Materials were constructed for four behavioural tasks: identification (both for a perception test and for training), identification on a morphed continuum, and a reading-aloud task. The two experimental groups performed a fifth task measuring discrimination on a morphed continuum, but since the control group did not do the same task it will not be reported here (see Thorin et al., 2018, for details on the task and the data; note that the two experimental groups did not differ on the task).

For the identification task, we used five minimal pairs with English ConsonantVowelConsonant (CVC) words contrasting the vowels /æ/ and /ɛ/: *fan-fen, ham-hem, jam-gem, man-men,* and *pan-pen*. Each word was spoken by 2 male and 2 female native speakers of British English. Seven tokens per word were used to increase variability of phonemic realisations, though all tokens were duration normalised per word pair. Of importance for the following analyses of the training EEG task are the respective vowel onsets for the respective word pairs: 139 ms, 64 ms, 111 ms, 112 ms and 90 ms (for more details see Thorin et al., 2018). These five pairs of words were also used in the reading-aloud task.

The identification task on morphed stimuli was based on an eleven-step continuum between the English words /vɛt/ and /væt/ (for details see Thorin et al., 2018). Additional stimuli were selected to test for transfer of learning in both identification and production. The stimuli for the transfer identification task can be divided into six categories, each introducing a new feature to the set of trained stimuli: (1) new starting consonant (C1): *tan-ten*, (2) new final consonant (C2): *mash-mesh*, (3) new C1&C2: *gas-guess*, (4) length: *cattle-kettle*, (5) 2 new speakers: *pan-pen*, and (6) naturally-timed versions of the five word pairs in the training set. Stimuli used for the transfer reading task are identical to those of categories 1-4 in the transfer identification task.

EEG task stimuli

EEG stimuli were constructed for three tasks: an active perceptual training task, a pre-test tonal oddball, and a post-test word oddball. Stimulus words used for the training task were identical to those used for the identification task (see above). For the tonal oddball paradigm, we created two pure tones with a frequency of 600 Hz and 500 Hz respectively with a duration of 100 ms and then normalised their amplitude together with the other stimuli used in the training.

The post-test oddball paradigm used the English word pair /pæn/-/pɛn/, and the two Dutch pairs /pɔt/-/pʏt/ (in English: pot - water well) and /pɑn/-/pɛn/ (in English also pan-pen). Stimuli were recorded by one female native speaker of Dutch with a native-like British English accent. Three tokens per word were selected and normalised in length (separately per phoneme) resulting in a stimulus duration of 400 ms.

C. Procedure

The complete training paradigm was composed of multiple pre- and post-test behavioural tasks assessing both participants' perceptual and production performance and the three tasks with EEG measurements (the pre-test tonal oddball paradigm, the active perceptual training, and the post-test passive oddball paradigm (see FIG 1 for a timeline[2]). All 5 sessions per training participant took place within a period of 10 days with not more than 3 days between consecutive sessions. The duration of the sessions ranged from 2 to 3 hours. Participants of the control group were tested in a single session only, in which they completed the post-test passive oddball task, the identification task, the identification on the morphed continuum task, and the reading task.

All experimental tasks were conducted in a shielded room and presented on a BenQ monitor (size 53.2 x 30 cm; 1920 x 1080 pixels; refresh rate of 60 Hz), in front of which participants were comfortably seated. All auditory input was played at a comfortable volume (~25dB) using in-ear headphones of the type Etymotic Research ER4P-T. Interactions between participant and experimenter were held in English. EEG was recorded throughout training sessions and during both the pre-test tonal oddball and the post-test oddball paradigm.

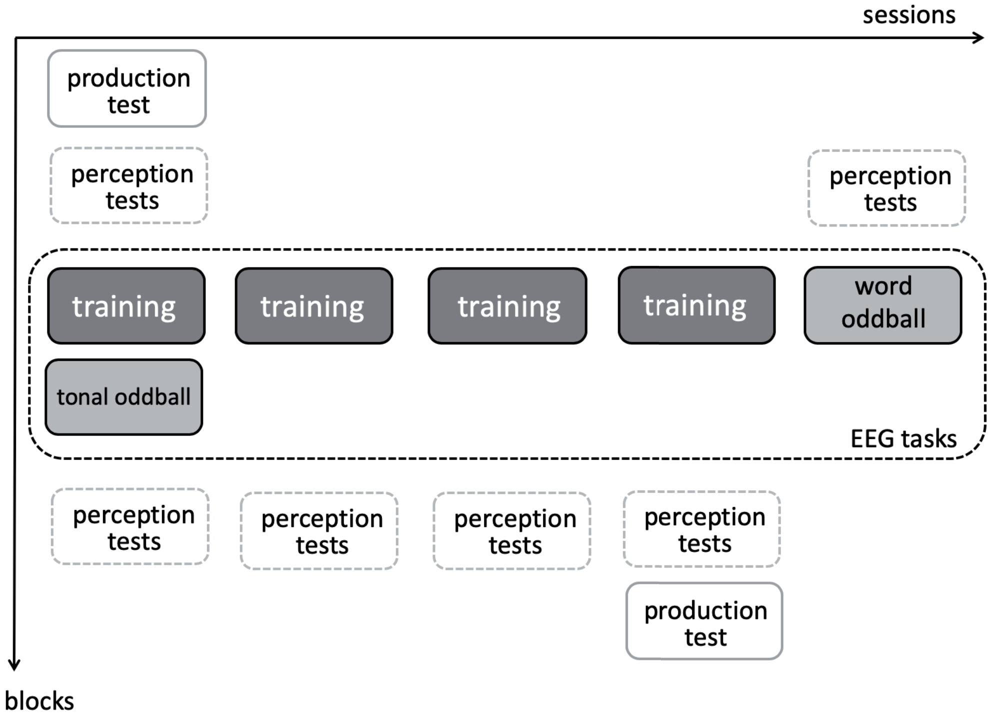

FIG 1. Training timeline. The full paradigm consisted of five separate testing days including four training sessions on each but the last day and five days of testing a battery of several behavioural and EEG tasks.

[2] Note that an additional phoneme substitution task taking place after the completion of the post-test passive oddball task was part of another study and will not be discussed further here

D. Experimental Tasks

EEG: The active perceptual training

During the perceptual training, participants were asked to listen to sequences of English words, classify the final word and then, after having received feedback, to pronounce a single word prompted to them on the screen. The four training sessions each comprised 5 blocks of 40 trials.

Each trial started with a fixation cross on the screen, during which participants listened to a sequence of 4-6 standard stimuli of the same word (varying speakers and tokens), followed by either a deviant final word (the minimal pair's counterpart, such as *ham* following the standard *hem*; 75% of the trials) or another version of the standard word (25% of trials). The stimulus onset asynchrony (SOA) varied between trials depending on the duration of the minimal pair, while the interstimulus interval (ISI) was 300 ms. Another 300 ms after the offset of the sequence's final word, the fixation cross was accompanied by two choice options, the members of the trial's minimal pair. The association between button orientation and word choice (for example, /æ/-words consistently on the left-button side) was held constant for individual participants across trials and sessions to avoid confusion, but was counterbalanced between participants. After the participant categorised the final word by button press, the non-selected alternative disappeared while the selected one turned either green or red to serve as feedback (correct or incorrect response respectively).

The visual feedback was presented for 2 seconds and was replaced by a single word printed in blue in the centre of the screen, which participants were asked to read out aloud. Depending on the production practice condition, this English word either contained one of the target vowels or one of two unrelated vowels. Participants in the related production group saw the final word of the immediately preceding oddball sequence and thus the correct answer of the categorical choice, while participants in the unrelated production group were presented with one of an unrelated set of minimal pairs: *shot-shut, hot-hut, cot-cut, dog-dug*, or *hog-hug.* At the end of each block, a prompt was displayed on the screen summarising the participant's correct answers and encouraging her/him to hold a self-paced break.

Prior to the first training block, participants were verbally instructed and could practice the task with a set of unrelated practice stimuli (i.e., *bout-but*, *heat-height*) taking about 5 minutes. The total duration of a training session was about 50 minutes. The experimental software was a combination of the Matlab toolbox Brainstream (http://www.nici.ru.nl/brainstream/twiki/bin/view/BrainStreamDocs/WebHome) and the Python based, open-source software package Psychopy (Peirce, 2007).

EEG: Post-test word oddball (EN and NL)

Each block contrasted one of the three minimal pairs (English /pæn/ - /pɛn/, Dutch /pɔt/ - /pʏt/ and Dutch /pɑn/ - /pɛn/) in a classical passive oddball paradigm (Näätänen et al., 2007). Each stimulus within one pair served both as standard and deviant in two separate blocks resulting in 6 blocks in total. The SOA (stimulus onset asynchrony) was constant at 700 ms with an ISI of 300 ms resulting in a total duration of 7 minutes per block. The deviance rate was 15% with a total of 90 deviants, which resulted in 600 trials per block. Participants watched a silenced nature movie called "Planet Earth" (BBC, 2006) and were asked to focus on the movie without engaging in any active auditory task.

EEG: Pre-test tonal oddball

The two pure tones (500 Hz and 600 Hz) were contrasted in a classical passive oddball paradigm. Stimuli were played in random order with at least 7 standards occurring before a deviant. The deviance rate was 15% with a total of 90 deviants, which resulted in 600 trials per block. The SOA (stimulus onset asynchrony) was constant at 700 ms (ISI of 600 ms) and each block therefore took 7 minutes. Each stimulus was presented both as deviant and standard in two separate blocks. During the entire task, participants watched the same silenced nature movie as above and were again asked to concentrate on the movie without engaging in any active auditory task.

Behavioural: Perception tests

The two behavioural tasks assessing a participant's perceptual ability to identify and discriminate between the two English target vowels /ɛ/ and /æ/ each took 5-7 minutes. The *identification task* was a two-alternative forced choice task, during which participants listened to single English words and indicated by button press which member of a given minimal pair they heard. Participants similarly categorised randomly played words, here one of the morphed stimuli on the /vɛt-væt/-continuum, in the *identification on morphed continuum task*. Assessing transfer of learning to new words, speakers and acoustic features (see stimuli), participants also did a *transfer identification task* after the final training session. For more details on all behavioural tasks, please refer to Thorin et al. (2018).

Behavioural: Production tests

Assessing participants' pronunciation ability of the target vowels before the start and after completion of the training, participants had to read out all 10 words used in the training during a *reading task*. As noted above, the transfer reading task after the last training session contained an additional set of words (for more details see Thorin et al., 2018).

E. Electrophysiological measurements

EEG was measured with 64 BioSemi active electrodes (BioSemi B.V., Amsterdam, The Netherlands), which were placed on the head according to the 10-20 system. The sampling rate differed between 512-2048 Hz and all data was thus offline resampled to 512 Hz. To detect eye-movements, we used two horizontal EOGs (electrooculograms) placed at the outer canthi of both eyes and one vertical EOG above and below the left eye. Both left and right mastoids were used as references.

F. EEG data processing and analyses

Offline processing of the data was carried out using the Matlab toolbox Fieldtrip (Oostenveld et al., 2011). All EEG recordings were cut into epochs based on trigger values for deviants and final standards before a deviant. While *word onsets* served as zero points for both pre-test tonal oddball and post-test passive oddball EEG data, all data from the active perceptual training was time locked to *vowel onset*. The reason for this was that, contrary to the other two EEG tasks, the stimulus set varied between sequences and thus within blocks in the active perceptual training, resulting in differing vowel onset times. Initial epochs were generously chosen from 10 sec before to 11 sec after stimulus onset to avoid filter artefacts in relevant parts of the epochs. Distortions of the signal due to eye movements were automatically removed based on correlations with the EOG channels (Gratton, 1998). Remaining motor activity caused by, for instance, speech articulation was classified based on typical spectral properties (relatively large power but a very low auto-correlation for motor activity) and also removed. All channels, in which the influence of 50 Hz frequencies deviated more than 3 standard deviations from the average influence were labelled as bad and interpolated based on neighbouring channels. Thereafter, the data was separately low-pass (0.1 Hz cut-off) and high-pass (30 Hz cut-off) filtered, using a two-pass Butterworth filter of 4th order with a Hamming window. Data padding on each side of the epochs was subsequently removed resulting in epoch sizes of -50 ms to 800 ms. After re-referencing to the mastoids, all remaining artefacts were automatically identified as those exceeding a threshold of 50 mV in a given trial and removed before the data was baseline corrected based on a 50 ms window prior to stimulus onset.

Preprocessed data was first averaged across all trials in a given condition for participants separately to compute event related potentials (ERPs) and then across participants' averages (grand average ERPs). Difference curves were computed by subtracting standard ERPs from respective deviant ERPs.

Statistical testing of the EEG data was done in the non-parametric framework employing a cluster-based permutation test that is part of the Matlab toolbox package FieldTrip (Maris and Oostenveld, 2007). We set the number of randomisations to 1000 and used the default Monte Carlo method to calculate significance probabilities. This choice of method gave us a straightforward solution to the multiple-comparison problem typically present in the analysis of multidimensional data structures (Maris and Oostenveld, 2007). All reported permutation tests were based on the entire set of electrodes in the time window specified per respective test and depending on the related research question. In an effort to balance sufficient statistical power and the risk of false alarms between cluster-based permutation tests, we used Bonferroni corrections whenever using multiple tests for a specific comparison within a given dataset.

III. RESULTS

A. Behavioural results

Effects of training (related versus unrelated production training)

The behavioural outcomes concerning the learning patterns in both perception and production for the two training groups have been presented in detail in Thorin et al. (2018). The main results revealed perceptual learning in the course of training reflected by increased identification scores (d prime), which also transferred to production performance. The latter was quantified as increased distance between the position of the two vowel categories in the F1-F2 space in terms of Mahalanobis distance, which is the distance between a point and a distribution in a 2-D space. Interestingly, there were no significant behavioural differences for the two training groups on any of the conducted behavioural test outcomes: identification, identification on morphed continuum, discrimination on morphed continuum, the production task, the transfer production task or the transfer identification task. See TABLE 2 for a summary of the main descriptive statistics (for pre- and post-measurements).

TABLE 2. Summary of main behavioural results comparing type of training between the related production group and the unrelated production group. Results of the main perception test (identification task) are presented as d prime scores before training and after the last training session, while results of the production test (reading task) are quantified as log Mahalanobis distance before training and after the last training session. Standard deviations are presented in brackets.

Task	Group	Time of measurement	
		Pre-training	Post-training
Perception test (ident. as d prime)	Related production	1.86 (±0.84)	3.68 (±1.24)
	Unrelated production	1.84 (±0.74)	3.43 (±1.35)
Production test (log Mahal. dist.)	Related production	5.87 (±10.72)	52.22 (±58.12)
	Unrelated production	2.79 (±6.07)	40.93 (±44.14)

Baseline perception performance

The perceptual performance quantified both in the identification task and identification in morphed continuum task did not differ between the control group and the two training groups prior to training. Concerning the pre-test identification task data, a one-way ANOVA with the between-subjects factor group (control, related production training, unrelated production training) resulted in no effects regarding d prime scores (F(2,51)= .5, p > .05) confirming that the performance between the three groups prior to training did not differ (TABLE 3).

For the pre-test identification on morphed continuum we performed a sigmoidal curve fitting on the number of classifications per stimulus on the 11-step /vɛt-væt/-continuum using Matlab. Thereby we could quantify both the sharpness and the position of each participant's category boundary. Resulting slope (boundary steepness) and 50% crossover point (boundary position) were compared between the three groups in two separate one-way ANOVAs with group as the between-subjects factor (TABLE 3). This revealed no significant difference between the groups for either slope (F(2,50) = 2.25, p > .05) nor crossover point (F(2,50) = .57, p > .05). Note that the data of one control group participant was excluded as the pattern of responses seemed random resulting in very low explained variance by sigmoidal curve fitting.

TABLE 3. Grand average data for the perceptual tests, identification and identification on morphed continuum, comparing the two training groups, related production and unrelated production, prior to active perceptual training, with the untrained control group. None of the scores differed significantly among the three groups.

Task	Related production	Unrelated production	Control
Identification task (d prime)	1.86 (± .84)	1.84 (± .74)	1.71 (± .77)
Ident. on morph. cont. (slope)	6.11 (± .70)	5.92 (± .84)	5.83 (± .79)
Ident. on morph. cont. (crossover)	3.01 (± 4.24)	1.08 (± .62)	1.87 (± 2.17)

B. EEG results

Active perceptual training

We compared the responses to deviant versus standard stimuli in the expected MMN time window (FIG 2) to evaluate whether active evaluation of the oddball sequences during training was accompanied by MMN responses to deviant final stimuli. To keep the number of tests small and thus prevent decrease of statistical power, we restricted this initial question to data from the first and final days of training. The typically observed latency for auditory MMN responses ranges from 150 - 250 ms after stimulus onset (Näätänen et al., 2007) but had to be corrected because the data was time-locked to vowel onset. With an approximate average of 100 ms from word onset to vowel onset across the 5 stimulus pairs, this resulted in an expected window of 50 - 150 ms. Cluster-based permutation tests on the average (per participant) responses to deviant as compared to standard responses (4 tests: 2 groups x 2 days) revealed no significant effects for either of the two training groups on day 1 or day 4 in the typical MMN window (p > .0125, Bonferroni corrected threshold for .05/4).

Based on the distinct response pattern in the later time window (becoming apparent in FIG 2), we decided to run an additional, exploratory analysis directly testing for P300 effects. The P300 is a positive-going peak around 300 ms after stimulus onset that is known to be related to an attentional switch due to decision making processes (Polich, 2007). The active categorisation of auditory stimuli during training could thus well explain the occurrence of such a potential. We tested for P300 effects for any of the two training groups at the start and end of the training, similarly to the MMN analysis above. Cluster-based permutation tests in the typical P300 latency of 250-500 ms (Polich, 2007), which was again shifted by 100 ms to correct for the time-locking to vowel onset, revealed significant differences for deviant and standard stimuli for both the related production group (p = .003) and the unrelated production group (p = .002) on day 4 but not day 1 (p > .0125; see FIG 2). The significant clusters spread over (close to) the entire set of channels. Following up on those P300 effects, we tested for an interaction between group x time in the permutation modality by evaluating the F-statistics of the differences between the two groups regarding their deviant-standard difference responses during the 4 days of training, again focussing on data in the typical time window for P300 effects. No significant clusters were found, indicating that any changes over the course of training did not differ between the two training groups.

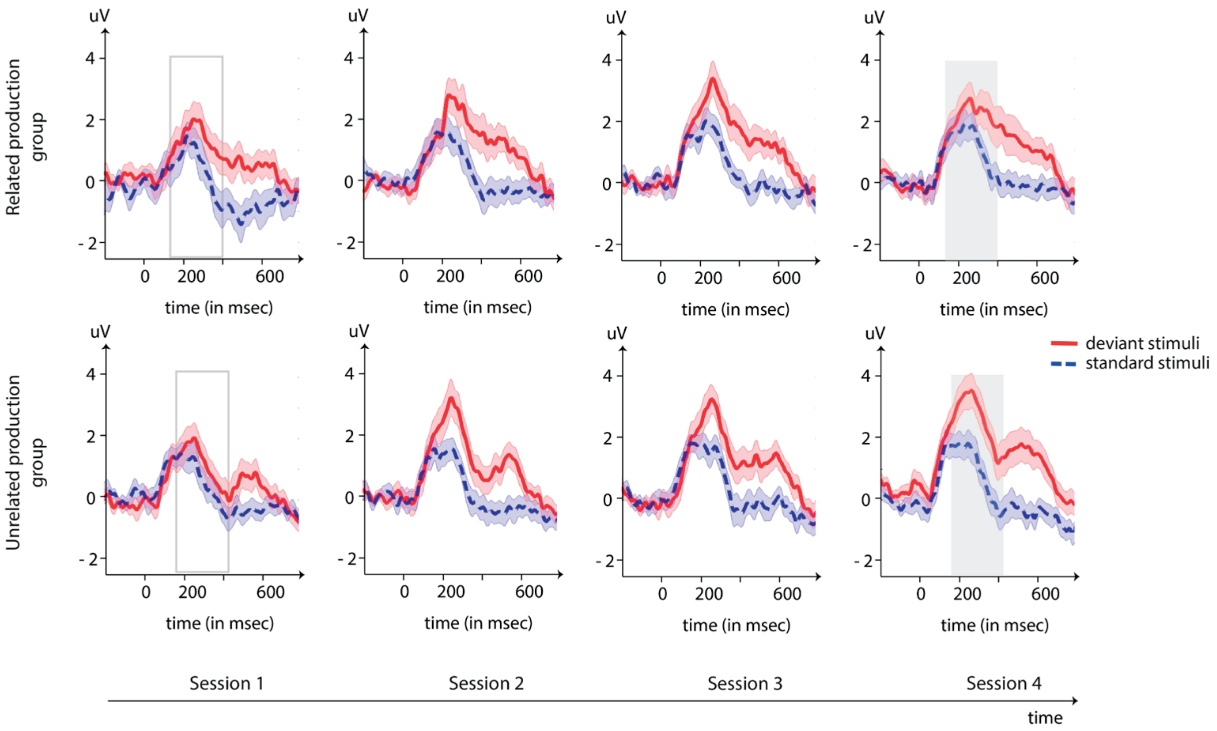

FIG 2. Active perceptual training paradigm results. Grand average ERP responses to standard (blue) versus deviant (red) stimuli time locked to vowel onset for training sessions 1-4 (left to right) and the two training groups: related production group (top) and unrelated production group (bottom). Responses are averages across a fronto-central cluster of electrodes and shaded areas indicate standard errors. Cluster based permutation tests comparing responses to deviant and standard responses were based on the typical P300 windows highlighted by grey frames (250-500 ms shifted to the left by 100 ms to account for timelocking to vowel onset instead of word onset) and further emphasized by filled grey frames whenever the comparison was significantly different (p < .0125 corrected).

Post-test word oddball

We tested for differences between the respective standard and deviant responses across all available electrodes in the typically observed time window for auditory MMN responses to assess whether all three groups showed an MMN response to the two Dutch contrasts (see FIG 3). This again ranged from 150 to 250 ms after stimulus onset (Näätänen et al., 2007). Cluster-based permutation tests revealed significant differences between responses to standards and deviants for all three groups in response to the Dutch /pɔt/ - /pʏt/ contrast. In the Dutch /pɑn/ - /pɛn/ condition, we found significant MMN effects for the related production and control group but after Bonferroni correction only a marginal effect for the unrelated production group (see TABLE 4). All difference effects were spread over a relatively wide cluster of electrodes spanning in some cases almost the entire set of electrodes.

TABLE 4. Post-test word oddball. Result summary of the cluster-based permutation test comparing event-related responses to deviant and standard stimuli in the typical MMN window (150 - 250 ms).

	Related production group	Unrelated production group	Control group
/pæn/ –/pɛn/ (EN)	.008*	.309	.098
/pɔt/ – /pʏt/ (NL)	.001*	.011*	.002*
/pɑn/ – /pɛn/ (NL)	.003*	.020	.004*

*Below significance threshold after Bonferroni correction within stimulus set (p = .05/3)

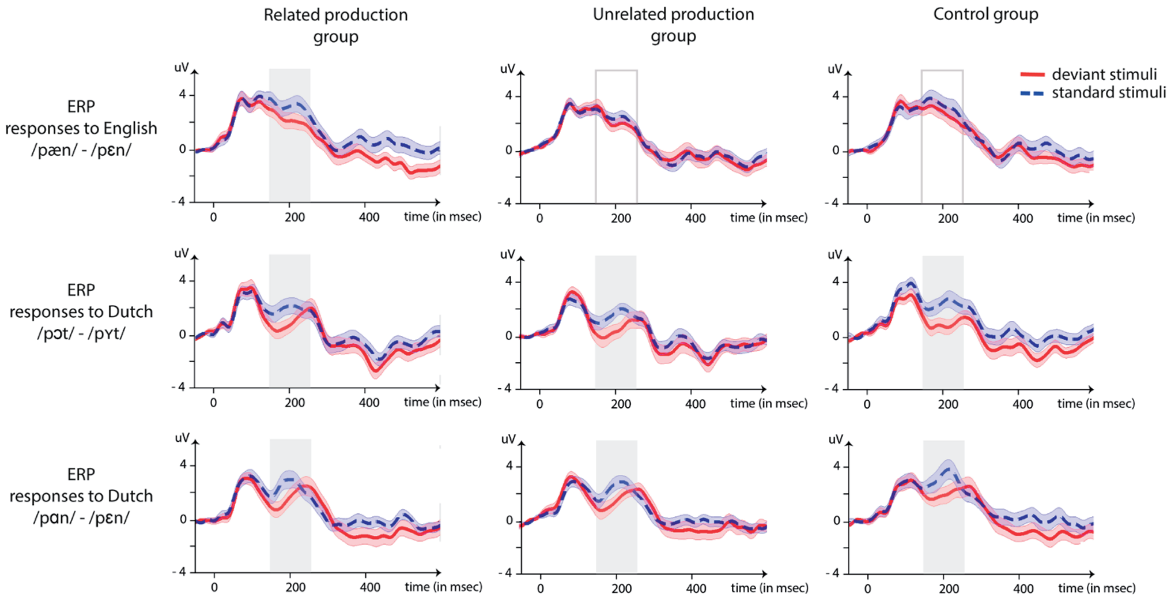

FIG 3. Post-test word oddball. Grand average ERP responses to standard (blue) versus deviant (red) stimuli for three stimulus sets, English /pæn/-/pɛn/ (top), Dutch /pɔt/-/pʏt/ (middle) and Dutch /pɑn/-/pɛn/ (bottom), and also separated for the three groups: related production group (left), unrelated production group (middle) and control group (right). Responses are averaged across a fronto-central cluster of electrodes with shaded areas indicating standard errors. All significant effects in the typical MMN time window are highlighted in filled grey (p < 0.017 corrected).

Similar permutation tests comparing standard and deviant responses to the English stimulus set, revealed a significant difference for the related production group (T-statistics: p = .008) again spread over a wide range of electrodes, but no difference effects for the unrelated production and control group (T-statistics: p > .05).

Answering the question of whether the three groups differed in the size of their MMN difference responses, we compared the difference curves (deviant minus standard response) in the same typical MMN time window of 150-250 ms between the three groups (see FIG 4). Separate permutation tests for the three stimulus sets revealed no significant group differences for any of the datasets (F-statistics: p > .05). Although only the related production group showed a significant difference between deviant and standard responses to the English stimulus contrast, overall the MMN difference responses of the three groups were therefore not shown to differ. Following up on this, we ran an additional permutation test contrasting the two training groups only (in order to increase sensitivity of this relatively conservative test). But again, no significant difference between the two training groups was shown (p >0.05).

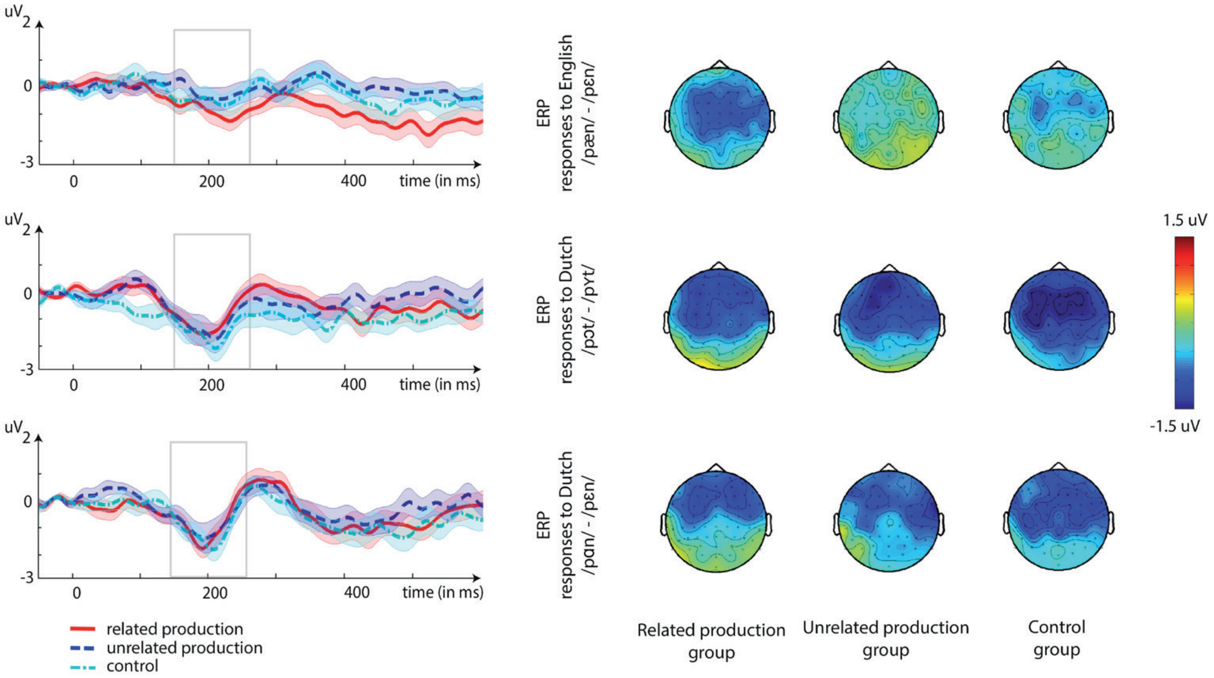

FIG 4. Post-test word oddball. [Left] Difference curves between grand average ERP responses to standard and deviant responses time locked to word onset of three stimulus sets: English /pæn/-/pɛn/ (top), Dutch /pɔt/-/pʏt/ (middle) and Dutch /pɑn/-/pɛn/ (bottom). The three training groups are distinguished by colour. Responses are averaged across a fronto-central cluster of electrodes and shaded areas indicate standard errors. The typical time window for the MMN response, which was also used for the cluster-based permutation test, is highlighted by grey frame. [Right] Corresponding topographic maps averaged across the MMN time window.

Pre-test tonal oddball

Comparing average responses to deviant and standard tonal tones in the expected MMN window (150 – 250 ms), one cluster-based permutation test per training group was performed (as noted above, the control group did not do the tonal oddball task). Results show significant differences between the responses (related production group: p = .001 and unrelated production group: p = .007) spanning over a wide cluster of electrodes including central, frontal and parietal sites. This can be interpreted as a typical MMN effect for both groups (see FIG 5).

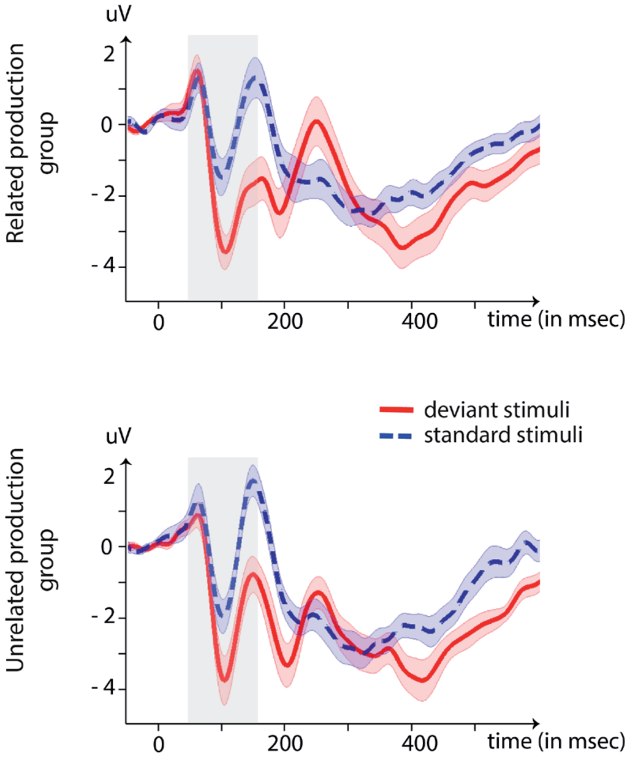

FIG 5. Pre-test tonal oddball task. Grand average ERP responses to deviant (red) and standard (blue) stimuli in a passive oddball sequence contrasting 500 Hz and 600 Hz tones (responses combined) for the two groups prior to training: related production group (top) and unrelated production group (bottom). Shaded curve areas indicate standard errors. All responses are averages across a fronto-central cluster of electrodes. Significant MMN effects are highlighted in grey.

Relation between behavioural performance and MMN responses

Given the significant MMN response in the post-test word oddball task for the related production group, we further investigated a direct relationship between MMN responses and post-training perception and/or production performance for individuals in this training group. To avoid a possible selection bias when quantifying individual MMN responses over an individual peak window, we divided participants into subgroups of good and bad learners first regarding their perception (cutoff at d = 3.5) and then separately regarding their production performance (cutoff at Mahalanobis distance = 30). We then compared overall MMN difference curves for each of those two subgroups using the same permutation-based statistical method as in the earlier analyses.

A cluster-based permutation test comparing difference curves in the typical MMN window (150-250 ms) for the English stimulus contrast between good (N = 8) and bad (N = 11) *perceivers* revealed no significant clusters. A similar test comparing the same data between good (N = 8) and bad *producers* (N = 7) showed no significant cluster after Bonferroni correcting (p = .05/2 = .025) for the two tests within one stimulus set (p = .041).

IV. DISCUSSION

This study aimed at furthering our understanding of bilingual speech processing by investigating how the speech perception and production modalities interact during non-native speech category learning and how behavioural improvements in the course of learning are related to traceable changes in the brain. More specifically, it examined neurophysiological changes in native speakers of Dutch, who were fluent L2 speakers of English, during and after a 4-day perceptual training protocol on the British English /æ/-/ɛ/ contrast. Critically, the perceptual training was complemented by either related or unrelated production practice, making it possible to evaluate the effectiveness of combined perception-production training paradigms.

Interestingly, results from the post-test word oddball indeed revealed differences of training outcomes. Passive listening to the English /pæn/-/pɛn/ contrast in a classical oddball paradigm triggered an MMN response in the related production training group that was not detectable in the unrelated production or untrained control groups. Importantly, the groups did not differ in their electrophysiological responses to native Dutch stimuli and pure tones. All three groups showed an MMN response to the Dutch /pɔt/ - /pʏt/ contrast and, in case of the related production and control group, also to the Dutch /pɑn/-/pɛn/ contrast (the response of the unrelated production group turned out not significant after Bonferroni correction). Similarly, in the baseline test contrasting two pure tones in the pre-test tonal oddball task, the two training groups both exhibited a typically shaped MMN response. Also regarding their behavioural responses before training, the three groups seem sufficiently well matched. They did not differ regarding their identification ability quantified as d prime score in the identification task. Nor did they differ in their boundary steepness and position between the two phonemic categories in the identification on the morphed continuum task. Taken together, these outcomes suggest that the differences in responses to English stimuli are indeed due to differing degrees of perceptual sensitivity to the trained non-native phonemes instead of due to any systematic differences in measurable MMN responses between the groups. On the group level, participants in the related production group were thus able to implicitly discriminate between the critical vowels, suggesting that they had established two non-native categories after training.

Several similar perceptual training studies offer support for the interpretation that the MMN is a valid indicator of enhanced perceptual ability due to training. A group of Finnish natives learning the voicing contrast in fricative sounds in a 3-day (4 sessions) listen-and-repeat training paradigm showed a significantly larger MMN after two days of training while synchronously improving their perceptual performance (Tamminen et al., 2015). Similarly, native Finns trained in more native-like spectral cue weighting in the English /i/-/I/ contrast showed a post-training MMN response that was absent in the pre-test, while also successfully improving in their behavioural ability in the course of 10 training sessions during a 3-week period (Ylinen et al., 2010). Another example comes from native English speakers trained on the novel /mba/-/ba/ contrast during a 10-day perceptual training, which led to both enhanced identification and MMN responses (Tremblay et al., 1998). Also the MEG equivalent to the electrophysiological mismatch response, the mismatch field (MMF), has been shown to increase after training, as was revealed by training American-English liquids in native Japanese learners. Twelve sessions of training here resulted in improved identification ability as well as enhanced MMF with significant correlations between neural and behavioural improvements (Zhang et al., 2009). Additional to this evidence from training non-native phonemes, there are also several examples of increased MMN responses to non-speech, complex auditory patterns trained during multiple-day paradigms (Atienza et al., 2002; Näätänen et al., 1993).

One line of research seemingly contradicts the current findings in showing that perceptual training can lead to smaller MMN responses after as compared to before training, despite improved perception (Kaan et al., 2007; Lu et al., 2015). Those studies, however, focus on Mandarin tone learning (in English speakers) as opposed to phonetic learning on the segmental level and the authors' explanation of their findings relates directly to the nature of the tonal contrast. Their reasoning is that before training the learners were more sensitive to the F0 onset differences between the lexical tones and had then shifted their attention after training more towards F0 direction instead, which in turn likely led to reduced MMN responses after training as detecting differences between the latter is evidently harder for native English speakers than the former feature (Kaan et al., 2008). The outcomes of these tone-learning studies are thus likely due to the specific mapping between native cue weighting with trained non-native tones and hence might not be directly comparable to the current findings. This also exemplifies the importance of considering that all the discussed training studies differ in their exact training methods, types of stimuli and combinations of L1 and L2 sound spaces. These are all crucial factors for determining the degree of difficulty listeners have in establishing novel non-native sound categories and thus all need to be taken into account when comparing and interpreting results across perceptual training studies.

Positive effects of related production practice on perceptual learning in the current study stand in contrast to the findings of Baese-Berk and Samuel (2016), in which combined perception-production training resulted in absent discrimination gains compared to those obtained through perception-only training. Based on this, the authors argued that additional production practice during perceptual training disrupts perceptual learning. Comparing the design of their study with the current one, a crucial difference is that learners in their study were asked to pronounce tokens of the trained phonemic contrast *prior* to indicating, or at least taking, a categorical decision. This could have introduced a difference in cognitive load during the perceptual tasks of the two training conditions in the study by Baese-Berk and Samuel (2016). There is evidence showing that cognitive load can reduce perceptual sensitivity during speech discrimination (Mattys et al., 2014; Mattys and Wiget, 2011) and lead to suboptimal memory encoding (Mitterer and Mattys, 2017). It therefore seems likely that learners in the combined training condition showed disrupted perceptive learning due to suboptimal encoding of the trained phonemes.

Other behavioural studies evaluating the effects of combined perception-production training on the learning of non-native phonemes show – similarly to the current study's behavioural results presented in Thorin et al. (2018) – neutral behavioural effects of additional production practice (Herd et al., 2013; Lu et al., 2015). But none of them have (potentially more sensitive) electrophysiological measurements to complement their findings. To our knowledge, the only other study presenting training relating changes in MMN directly comparing perception-only and combined perception-production training is the previously discussed study on Madnarin tone learning by Lu, Wayland, & Kaan, (2015), which as we have already argued is hard to compare to the current study.

The electrophysiological signature of (increased) perceptual sensitivity in the absence of any behaviourally measured benefits of related as compared to unrelated production practice (presented in Thorin et al., 2018) leads us to the question whether electrophysiological changes are indeed more sensitive in picking up subtly evolving effects of phonetic training than common behavioural measures. Studies investigating the learning process necessary for perceptual discrimination and/or identification of auditory contrasts which have included behavioural and electrophysiological measures have produced findings which go in different directions. On the one hand, synchronous changes in MMN responses and behaviour have been observed in the previously mentioned study by Tamminen et al. (2015), in which native Finns were trained on a voicing contrast in fricative sounds, and also in an MEG study revealing increasing amplitudes of MMF signals to a subtle frequency contrast accompanied by similar behavioural improvements in the course of training (Menning et al., 2000).

On the other hand, a few other studies have observed an MMN effect to auditory contrasts in the absence of any behaviourally measured indicators of perceptual change detection. The auditory presentation of three unknown Hindi phoneme contrasts each in a passive oddball paradigm was shown to elicit MMN responses without any indication of behavioural discrimination ability (Rivera-Gaxiola et al., 2000). Another example comes from the previously mentioned perceptual training study on a non-speech, complex auditory pattern (Atienza et al., 2002). Training here initially led to both behavioural gains in the form of improved discrimination ability of learners and enhanced MMN responses. Interestingly, however, further neural changes became apparent during the day(s) after training completion, while behavioural learning effects stayed relatively stable: The P2 response was further enhanced after 24 hours and also the MMN response was significantly larger after 36 hours. Also Tremblay et al. (1998) revealed a dissociation between (the learning curves of) their EEG and behavioural measurements when training English natives in the perception of the voice onset contrast /ba/-/mba/ during a 10-day paradigm including 4 training sessions. Results varied substantially across participants (N=10) with behavioural improvements only detectable 2-3 sessions thereafter in about half of the group, while all participants showed significant increased MMN responses relative to the first training session.

In the light of those studies, it is hard to disentangle whether the current findings are signs of increased sensitivity of electrophysiological measures as indicators of perceptual learning or, in fact, point towards differential time courses between, on the one hand, neural restructuring that provides the basis for perceptual learning and, on the other, the manifestation of this process in the form of behavioural improvements. Based on the current data and related findings, however, it seems reasonable to expect that a behavioural effect in the form of enhanced perceptual ability in the related production group going beyond the improvements made in the unrelated production group would eventually emerge. We predict that this would be measurable as increased identification and/or discrimination performance either with more training and/or with a longer delay after training. With potentially more sensitive measures, such as EEG, one might also expect to see effects of related production practice on production performance itself. Future research could further clarify these issues by establishing and using valid electrophysiological indicators of improvements in distinctively producing non-native phoneme contrasts, as well as investigating the seemingly differential time-courses of neurophysiological changes and their behavioural counterparts.

We were not able to track phonemic category formation neurophysiologically *during* training: Neither during the first nor last of the four training sessions were there significant differences between responses to deviant and preceding standard stimuli in either of the two training groups. The reason for the absent MMN effects, despite evident improvement in learners' behaviourally measured perceptual ability, is likely to be found in the nature of the training task differing from the classically used passive oddball tasks, such as the post-test word oddball task, in several regards. Firstly, the training task involved active listening and decision-making in order to serve as feedback-based training task. Secondly, this also meant that stimulus presentation had to be discontinuous through the use of short sequences ending on either a standard or deviant stimulus to which participant could actively respond. Lastly, the stimuli during training were substantially higher in variability, including seven tokens of five word pairs recorded by four speakers (two genders), as compared to three tokens of a single word pair with a single speaker used in the post-test oddball task. Taken together, the trade-off we made in our choice of training design with increased stimulus variability, which has been shown to improve learning (Bradlow et al., 1999b; Wong, 2013), as well as integrating the EEG measures in the active training itself was that those benefits came with increased difficulty to detect a mismatch response.

Despite the absence of MMN responses during training, a P300 effect that was not yet present during the first session synchronously developed for the two training groups in the course of training sessions. P300 responses are commonly thought to reflect information processing cascades involving attentional and memory related mechanisms elicited by the process of active decision-making (Polich, 2007). In the context of an active oddball paradigm, attentional resources are thought to be allocated to incoming target stimuli, here the last stimulus word in a given trial sequence, in order to compare it to the model of the standard stimulus in working memory. Whenever this active comparison leads to a mismatch, a P300 response is generated. Learners of both groups were thus increasingly able to consciously differentiate between the two target vowels, which is in line with the earlier presented behavioural data from the same training paradigm similarly showing that the number of correct classifications of final words improved with training (Thorin et al., 2018). Similarly to reported behavioural outcomes, however, no differences between the two training groups became evident. Taken together, the current electrophysiological findings suggest that the P300 effect here seems to be more closely tied to the processes underlying behavioural performance than the MMN, while the MMN (in the post-test word oddball) captures a sensitivity not detectable yet in behaviour. This reasoning ties in with the prediction made above that we would expect to see a behavioural manifestation of the evident neurophysiological changes after either more training or a longer delay without training in the related production group (though it would be crucial to compare any effects of extended training or consolidation in the related production group with potential changes in the unrelated production group, too)

An attempt to find a direct link between behaviourally measurement improvements, both in perception and production, with MMN amplitudes in the post-test word oddball by splitting up the data in terms of "good" and "bad" learners was unsuccessful as no differences between the two subsets were found. A reason could be the relatively conservative analysis approach, which comes with the drawback of reduced power, while in the current case avoided heavy data pre-selection.

V. CONCLUSION

In sum, the current study provides support for positive effects of related (as opposed to unrelated) production practice on perceptual learning in unbalanced bilinguals, even in the context of perceptual training. It did so by showing that training including related production practice led to a mismatch negativity response to the trained L2 vowel contrast after training that was not evident after perceptual training with unrelated production practice or in an untrained control condition. Those outcomes also confirm that neurophysiological measures are sufficiently sensitive to identify fine-grained differences in perceptual ability that might not (yet) be detectable by conventional behavioural methods.

VI. ACKNOWLEDGMENTS

We would like to thank Philip van den Broek for his technical support, as well as Malin Mangersnes, Mariia Naumovets, Inez Wijnands, and Josh Ring for their help with data collection.

4

Chapter 4

The effects of production practice on perceptual learning of non-native vowels

ABSTRACT

Speech perception and production have repeatedly been shown to interact during non-native sound learning in the form of perceptual learning transferring to production improvements. The reverse transfer, namely from production learning to gains in perception has received less attention. The present study had two aims. The first was to evaluate the effectiveness of a two-session computerised pronunciation training protocol on the English /æ/-/ ɛ/ vowel contrast. During training, Dutch native speakers were provided either with trial-by-trial visual feedback on their own vowel productions (experimental group) or with a general indication of how in terms of tongue location the target vowels are pronounced by a typical native speaker (control group). The study's second aim was to further our understanding of the mutual relationship between the two speech modalities by testing the effect of improved L2 pronunciation on perceptual learning. Results of two experiments showed that both groups improved their productions, while there was no overall evidence that the trial-by-trial feedback (further) supported learning. Interestingly, production learning transferred to improvements in participants' perceptual ability despite the lack of direct training in this modality. Taken together with previous findings, this outcome points towards a bidirectional relationship between the perception and production modalities during non-native speech category learning.

I. INTRODUCTION

When expressing oneself in a second language, the most challenging sounds to pronounce will usually be the ones that are not present in one's native language. This is especially the case whenever the second language (L2) has two similar phonemic categories where the native language (L1) possesses only a single one in that part of phonemic space (Best, 1995). The major challenge then consists of learning to distinguish between the two phonemes. In order for the L2 learner to reach a native-like level, this learning process needs to succeed both in speech perception and speech production. Even though it seems intuitive that the two modalities are linked to some extent, the exact interactions between them in the course of establishing non-native sound categories are still unknown. The present study focusses on investigating whether targeted pronunciation training of a challenging L2 contrast will not only improve learners' production performance but also transfer to their ability to perceptually discriminate the two categories.

Perceptual training approaches that target non-native sound categories have been an efficient method to explore the degree to which L2 learners' phonemic representations are still plastic in adulthood. Studies from the last three decades have confirmed that those perceptual training paradigms can lead to successful improvement of identification and discrimination performance including a wide range of combinations between L1 and L2 sound spaces (see Sakai and Moorman, 2018, for a review and meta-analysis). Among those studies, the most frequent example is the training of Japanese native speakers in perceiving the English liquids /ɹ/–/l/ (Bradlow et al., 1997, 1999b; Iverson et al., 2005; Logan et al., 1991; Shinohara and Iverson, 2018), while other studies focused on, for instance, English vowels in Dutch native speakers (Thorin et al., 2018) or in Japanese natives (Lambacher et al., 2005a), a Basque contrast in Spanish natives (Baese-Berk and Samuel, 2016), Japanese vowels and consonants in native speakers of American English (Hirata, 2004b), or Mandarin tones in native English speakers (Wang et al., 2003).

Those perceptual training studies have also served the purpose of investigating the relationship between the perception and production domain during non-native sound learning. A recent meta-analysis has shown that non-native perception training overall led not only to medium-sized improvements in perception but also transferred to gains in the production modality (though small in effect size; Sakai and Moorman, 2018). The reverse relationship, however, namely if and how training to correctly pronounce challenging non-native sounds would affect the perception of those sounds, especially after training production in isolation, has received less attention.

Isolated production training and its effects on the two speech modalities was investigated, for instance, in production-only training targeting a Spanish intervocalic three-way contrast, which led the native English learners to improve both in their production and their perception (Herd et al., 2013). Successful production training similarly transferred to speech perception when training English natives to produce a Japanese pitch and durational contrast (Hirata, 2004b) and also when training native French speakers in the production of four Danish vowels (Kartushina et al., 2015). In contrast to these results, pronunciation training of English liquids in native speakers of Japanese did not lead to any improvements in the perceptual modality despite native-like ceiling effects with respect to their production performance (Hattori, 2009). Production-only training also did not transfer to perceptual improvements when training Cantonese native speakers on an English vowel contrast (Wong, 2013).

Several studies also inspected the effects of combined perception and production training on improvements in either of the two modalities and created a pattern of mixed findings. No additional and thus neutral effects of combined perception-production training as compared to perception-only training was observed, for instance, in English natives learning lexical tones (Lu et al., 2015), and in an additional training group of Herd et al. (2013) in which, as mentioned above, native English speakers were trained on a Spanish three-way contrast. In addition, training Dutch natives both in perception and production of an English vowel contrast did not lead to increased learning in either of the modalities when comparing behavioural outcomes to perceptual training combined with productions of unrelated tokens (Thorin et al., 2018). Additional, more sensitive electrophysiological measurements of the same training groups, however, revealed advantageous effects of combined training on the automatic discrimination of the trained contrast in the form of a mismatch negativity response (Thorin et al., in revision). In contrast to these neutral or even positive effects, however, perceptual learning was negatively affected by combined perception-production training as compared to perceptual single-modality training both in Spanish natives trained in a Basque consonant contrast (Baese-Berk and Samuel, 2016) and in native English speakers trained in a Hindi contrast (Baese-Berk, 2019).

The above-mentioned studies on production training vary tremendously in the type of training and, more specifically, the way they did (or did not) implement feedback. External feedback on verbal responses during production training is likely to play an important role in supporting the process of correctly pronouncing challenging non-native sounds. Complementing the above production training studies in this regard, a line of research has tested different methods for effective computerised pronunciation training. Examples range from different automatic speech recognition systems (Arora et al., 2018; Machovikov et al., 2002; Neri et al., 2006) to technically more extensive approaches, such as the use of electropalatography enabling real-life feedback on tongue movements during articulation (Hacking et al., 2017; Katz and Mehta, 2015). Especially the ASR (automatic speech recognition) approaches, however, typically come with the drawback of giving less informative, often binary feedback which might be less effective in supporting L2 category formation and thus transferrable pronunciation skills. Similar to the approach by Kartushina et al. (2015), other studies have used visual representations of spectral elements of a learner's productions as basis for immediate feedback, such as a pilot study training Dutch natives in their pronunciation of Spanish vowels (Lie-Lahuerta, 2011). But as the focus in those automatised training approaches was typically directed towards the efficiency of the respective computerised pronunciation tool in terms of the degree to which it improved learners' productions, how those improvements affected perceptual performance was often not measured. In other words, though there are multiple examples of computer-assisted pronunciation training studies, only few studies investigate the interaction between the speech perception and production domain during the process of production learning, especially transfer of learning from production to perception.

The present study therefore used a two-session computerised pronunciation training approach providing learners with trial-by-trial visual feedback on the distance between their own productions and those of a typical native speaker's example (similar to the training design used in Kartushina et al, 2015, 2016, and 2019). Learners were chosen to be Dutch native speakers and they were trained on the English /æ/-/ɛ/ vowel contrast, as such speakers have repeatedly been shown to have difficulties with both discriminating and distinctively pronouncing the two vowels despite high levels of proficiency in English (Broersma, 2002; Escudero et al., 2008; Wanrooij et al., 2014). During both training sessions, the Dutch native learners had to read out aloud single English words each containing either /æ/ or /ɛ/, such as in English /pæn/. After each trial, they were presented with a graphical representation of the F1-F2 space (1st and 2nd formant) and as part of that with an indication of how their own pronunciation was located relative to that of a typical native speaker in terms of tongue "frontness" (horizontally on the F1-axis) and openness of mouth (vertically on the F2-axis; see methods for more details). The performance of this experimental group both on pre- and post-test perception and production measurements was compared to a control group, in which participants pronounced the identical words but received no direct feedback on their own pronunciations. The current aim is to, firstly, evaluate the effectiveness of such a training method (combined with pronunciation instruction) and, secondly, to assess to which extent gains in production transfer to improved perception. The observation of transfer from the production to the perception domain would suggest the presence of a reciprocal link between the two modalities in the process of learning novel speech sounds and thus further our understanding of their underlying interactions.

II. EXPERIMENT I

1. Methods

A. Participants

Eighteen native Dutch-speaking females (mean age = 23.6 ± 2.2) who were lower-intermediate/advanced L2 speakers of English participated (TABLE I). The reason for only including female participants was due to the training tool being specifically designed for female voices (see description below). All participants were native speakers of Dutch, born and raised in the Netherlands. Next to the English language, the majority of participants also reported some knowledge (but only rare to no active use) of German and French, which are both commonly taught in Dutch high school education. Some were also familiar with Spanish (16%) or in rare cases with Greek, Latin, Swedish, Norwegian, Russian, Czech, Frisian or Chinese (2-3%). All participants gave their written informed consent prior to participation. They reported to have normal hearing, normal or corrected-to-normal vision and no reading difficulties, such as dyslexia. The study was approved by the Ethics Committee of the Faculty of Social Sciences at Radboud University Nijmegen and all participants were paid or received course credit for their participation.

TABLE I. General information on the experimental and control group of Experiment I: mean values followed by standard deviations in brackets.

Group	N	Age	LexTALE
Experimental	9	23.6 (± 1.9)$^{n.s.}$	69.1 (± 17.0)$^{n.s.}$
Control	9	23.6 (± 2.5)	74.9 (± 15.1)

$^{n.s.}$ indicates non-significant results of independent sample t-tests comparing the two groups.

B. Procedure

The full training study consisted of two sessions each taking about 1.5 hours, during which participants completed perception and production pre- and post-tests, some general language tasks, and the production training itself. The tasks were completed in the order given in FIG 1. During both sessions, which took place within the maximum period of a week, participants were comfortably seated in front of an iMac computer (27" retina display; 5120 x 2880 pixels) with all auditory input given via in-ear headphones (Etymotic 262 Research ER4P-T) at a comfortable volume. All interactions with the participants were held in English. Participants were randomly allocated to either the control or the experimental group prior to the first session.

C. Stimuli

The training stimuli consisted of 5 sets of ConsonantVowelConsonant (CVC) minimal pairs each contrasting the two English target vowels /æ/ and /ɛ/: *fan-fen, ham-hem, jam-gem, man-men, pan-pen*. This stimulus set further described below was identical to the one used in Thorin et al. (2018).

For the perception task, these words were recorded by 2 male and 2 female native speakers of British English who were born and raised in Southern England. For each

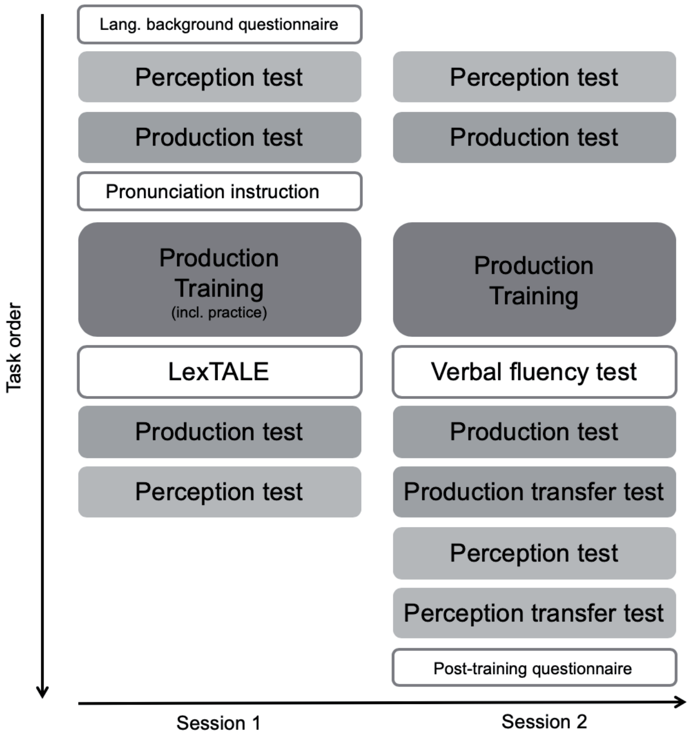

FIG 1. Schematic timeline of the two training sessions consisting of the production training, the perception and production pre- and post-tests, and the other, general language tasks.

word, 7 tokens were selected and normalised in duration to each phoneme's average duration across all tokens and speakers. For the transfer perception task, the training set was extended by six transfer categories each introducing a single new or adapted feature: (1) new starting consonant (C1): *tan-ten*, (2) new final consonant (C2): *mash-mesh*, (3) new C1&C2: *gas-guess*, (4) length: *cattle-kettle*, (5) 2 new speakers: *pan-pen*, and (6) naturally-timed versions of the training set: *fan-fen, ham-hem, jam-gem, man-men,* and *pan-pen* (for details see Thorin et al., 2018). Transfer categories (1)-(4) were also used for the transfer production task.

D. Experimental tasks

Training tool

The training tool[3] was a computerised task implemented in the freely available software Praat (Boersma and Weenink, 2015), which aimed at supporting production learning of the English /æ/-/ɛ/ vowel contrast by providing learners with immediate, trial-by-trial visual feedback in the form of a representation of where their own productions were located relative to those by typical native speakers. On each trial of training, the tool visually presented one of the English training stimuli and recorded the participant's

[3] The training tool is openly available on GitHub: https://github.com/GiselaGovaart/vowel-production-feedback

subsequent pronunciation of it. The tool then automatically segmented and analysed the speech signal before providing immediate feedback on the quality of the response.

Segmentation: The participants' utterances were segmented using Praat's inbuilt segmentation function, which employs a speech synthesizer of the specified language (in this case, English) creating a synthesized version of the target word (based on a text transcription of the target) that can then be aligned with the sound file of the participant's utterance. The boundaries of the segments in the recorded signal were then set based on this alignment. The segmentation procedure was validated prior to the experiment by comparing average formant values across whole vowel duration that was segmented automatically (total of 400 recordings by four different speakers) with average formant values based on the same but manually segmented recordings (N=400) (Govaart, 2016). Based on correlation values of 0.91 and 0.94 for F1 and F2 values respectively, we concluded that the method was sufficiently accurate to use for this online feedback training.

Feature extraction and native model: As the purpose of the tool was to quantify the quality of a given pronunciation in terms of its distance to a typical native utterance, we needed to first determine the most suitable features to distinguish between the two vowel categories and to then use those features when building a model of native speaker pronunciation. For both steps we used recordings of the 10 training words by 10 native speakers of British English (5 female and 5 male; 10-11 repetitions). To establish which features were the most informative in discriminating between the /æ/ and /ɛ/ vowels, we carried out a Linear Discriminant Analysis (LDA) with different formant and non-formant measures[4]. All formant extraction was based on Praat's standard formant extraction method (with gender specific formants ceilings of 5000 Hz and 5500 Hz for male and female voices respectively), which uses Burg's algorithm (Press et al., 1992) to compute the linear predictive coding (LPC) coefficients (see the Praat manual). Based on the outcomes of the LDA, we chose to use (1) average F1 and F2 values extracted from the entire vowel segment to differentiate and visually represent the vowels and (2) to use different models for the 5 word pairs as the information on the start-consonant turned out to be among the most informative features as well. The latter is due to co-articulation differences among the initial consonants. Unexpectedly, adding F0 did not improve the

[4] The LDA model included the following 7 features representing the spectral information of the produced vowels: F1 and F2 values based on (1) whole vowel duration, (2) midpoint, (3) 15 ms period centered around midpoint, (4) 50% centered portion of the segment, (5) 20%, 50% and 80% location of the vowel segment (cf. Hillenbrand et al., 1995), (6) 50% location of the vowel segment plus the difference between the 50% and the 20% point, minus the difference between the 80% and the 50% measuring point, see the *production undershoot model* of Stevens and House (1963), and (7) Mel-Frequency Cepstral coefficients (MFFC) 1 to 12. In addition, gender, start-consonant, final-consonant and fundamental frequency (F0) were taken into account.

performance of the model, and it was therefore not included in our analyses.

The F1-F2 distributions of the 10 native speakers revealed a much clearer separation between the two vowel categories for female than for male speakers (see FIG 2) which is in line with findings showing that females tend to have larger vowel spaces and more separable vowel categories (Escudero et al., 2009; Hillenbrand et al., 2001; Simpson, 2009). The clearer distinction between female vowel distributions was the reason for initially restricting the use of the tool to female learners of English and similarly basing the native comparison model on female recordings only. We furthermore restricted the native model to the F1 and F2 distributions of two female speakers with an evidently clear distinction in vowel realization (speaker W1 and W2 in FIG 2). The final native model thus comprises the average F1 and F2 values (across whole vowel duration) for the two selected female speakers related to the 5 word pairs respectively (see above).

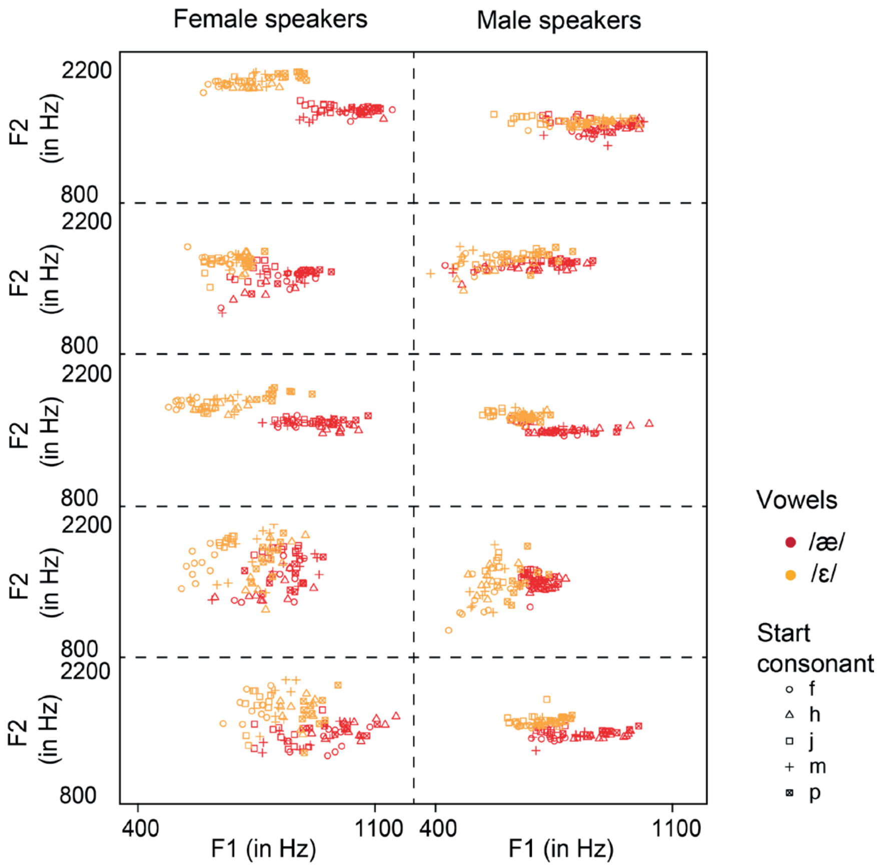

FIG 2. Vowel distributions in the F1-F2 space per speaker for each start consonant (indicated in different shapes): [left column] (W1-W5 from top to bottom), [right column] female speakers male speakers (M1-M5 from top to bottom).

Feedback. Participants received immediate visual feedback after each verbal response. In the experimental group, this feedback consisted of a representation of a trial's pronunciation response in terms of its location in the F1-F2 space[5] together with the location of the target word and its vowel counterpart produced by the native model (see FIG 3 for an example of the feedback screen). Whenever the participant's utterance was too far away from the native model or no sensible formant values could be extracted, the text "Too far from the target vowels" appeared instead. The axes of the F1-F2 spaces were not shown on the feedback screen, because they differed according to the initial and final consonants. In the control group, participants were shown the identical representation of the native model for a trial's word pair in the F1-F2 vowel space, but did not see the data point referring to their own verbal response (blue dot in FIG 3).

Instructions and practice task. Prior to the first training session, participants in both groups received identical pronunciation instructions introducing the common problem for Dutch learners of English when pronouncing the English /æ/ and /ɛ/ and how the two categories can be pronounced distinctively by varying how far to the front the tongue is located and how open the mouth is when producing them. They also received an example picture of the feedback screen and how to interpret the location of the native model on the screen in terms of frontness and openness. Because the exact location of the two native vowels differed depending on the word pair in the current trial, participants in both groups were informed that speakers tend to pronounce vowels differently depending on the consonants preceding and following them. In the experimental group, participants additionally were instructed that they would see the location of their own productions and that their task consisted of coming as close as possible to the native target word indicated in green. The aim for participants in the control group was to practice their pronunciation by repeating the English words shown to them while making use of the information regarding ideal tongue location and mouth openness. Both groups could practice their task during a brief testing period and got the chance to ask additional questions before starting the first training session.

[5] For the visual feedback, the F1 and F2 values were 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.

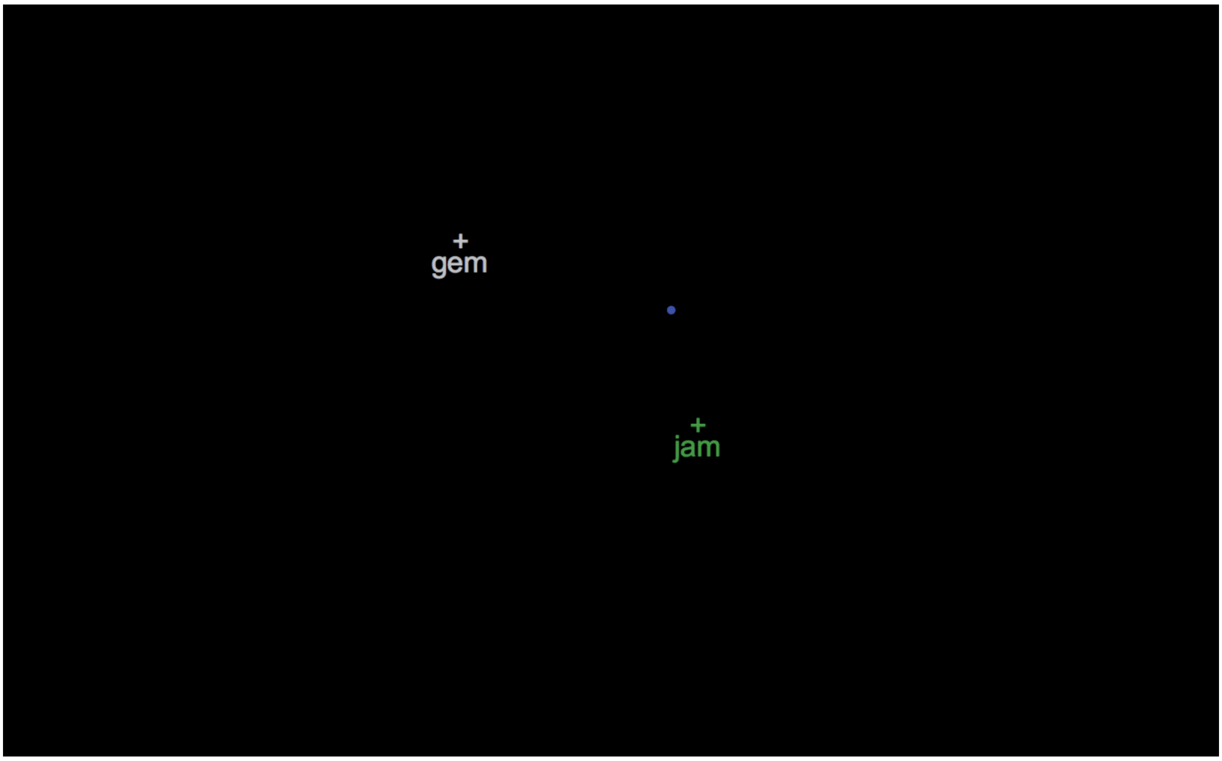

FIG 3. Example of the feedback screen participants in the experimental group saw after each of their verbal responses. The blue dot indicates the position of the vowel pronounced by the participant in the F1-F2 space, while each "+" indicates the location of the word pair in the native model (the target in green and its minimal pair in white). While participants in the control group saw the same information about the native model, they did receive any information on the quality of their own verbal response.

Calibration phase. To accommodate the fact that different speakers have different-sized vocal tracts and therefore varying formant ranges, each training session started with a brief calibration phase, during which participants had to pronounce the corner vowels /i/, /a:/ and /u/. The researcher ensured that the three vowels of a given participant were sufficiently distinct and formed a clear triangle in the F1-F2 space before starting the training. The information on a participant's corner vowels enabled us to then project a participant's vowel space onto an artificial vowel space on which the normalized natives' vowel spaces had also been projected (i.e., the natives' corner vowels were also recorded and used to normalize the input stimuli). As a result, the participants' vowels could be compared with the natives' vowels. Technical details on the z-normalization of F1 and F2 used in this procedure can be found in Lobanov (1971).

Production task and transfer production task

This was a self-paced reading task, during which participants read out single English words that were presented to them on the screen. The 10 training words were repeated 3 times in a shuffled order resulting in a total duration of 3-5 minutes. The transfer production task was identical aside from including the additional transfer words (see above) and consisted thus of 24 trials. Both versions were run using Praat.

Perception task and transfer perception task

In each trial of this two-alternative forced choice task, participants listened to a single English word after which they saw two choice options on the screen, representing the two versions of the trial's minimal pair. They had to indicate which of the two they heard by pressing the corresponding button. A summary of correct responses was presented at the end of the total of 120 trials (10 words x 4 speakers x 3 repetitions). The task took about 3-5 minutes to complete. It was run using the Python-based software Psychopy (Peirce, 2007). The transfer perception task was similar and consisted of 200 trials (10 minimal pairs x 2 speakers x 5 repetitions/tokens) resulting in a duration of 5-7 minutes.

Verbal fluency test

English verbal fluency was assessed by asking participants to produce as many English words of a certain kind within one minute as possible. Both a semantic and phonemic version was used, in which words belonging to the category "animals" or words starting with the letter "S" respectively had to be produced. All valid and unique productions were summed up to produce a total score.

LexTALE

The LexTALE is a brief computerised task assessing English vocabulary size, in which participants have to evaluate whether single words appearing on the screen are either existing English words or nonwords, by selecting either a "yes" or "no" button. The task consists of 63 trials and its score is known to correlate with English proficiency (Lemhöfer and Broersma, 2012).

Questionnaires

We used a questionnaire to be filled out before the training to collect general information on participants' language background (country of origin, L1 of father and mother, other languages they were familiar with and to what extent) and to rule out that they had any problems with their vision or hearing. It also included a Likert-like scale ranging from "No, not at all" to "Yes, very much", on which participants had to indicate their motivation to reach a native-like accent in English.

2. Results

As we tested the first participants, we discovered a software-related runtime error, which occurred and re-occurred in the majority of the sessions. It appeared in the form of a pop-up window, which could easily be removed by clicking "Ok", but it still interrupted the flow of the task and might have reduced the participants' concentration and/or motivation. This problem could not immediately be solved by us as it could be handled by the developers of Praat only. We therefore decided to treat the experiment as a pilot study, continuing to test participants who had already been recruited (9 in each group).

A. Production learning

In order to evaluate the effectiveness of training, we first tested whether participants' pronunciation of the two vowels came closer to the native model during the course of training. This analysis relied on the mean F1 and F2 values across the automatically segmented[6] vowel portions of the produced words before and after training. To circumvent including trials with no or too noisy productions, values based on vowel durations below 20 ms were removed (0.5% of the data). Similarly, we reduced the chance of including errors caused by Praat's automatic formant extraction method by excluding formant values deviating more than 2.5 standard deviations from the mean value of a given participant's data on a given time of measurement (1.9% of the data). The formant values (in Hz) were then converted to log values, as those are known to better resemble the properties of the auditory system. The distance to the native model was expressed in terms of the Mahalanobis distance (Kartushina and Frauenfelder, 2014). This measure quantifies the distance between a point and a distribution in a 2D-space, while taking into account the shape of the distribution. It does so by measuring 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). For each produced vowel, we computed the distance between its position in the F1-F2 space to the distribution of the native model. The average distance per participant and measurement served as dependent variable in a repeated measures ANOVA including the within-subject factor time (pre-test 1, post-test 1, pre-test 2, post-test 2) and between-subject factor group (experimental, control). Note that throughout this paper corrected p values are reported whenever Mauchly's test for sphericity was positive. Results showed that the Mahalanobis distance between /æ/ vowels pronounced by participants and those by the native speakers significantly decreased over the course of training (main effect of time: $F(3,48) = 7.91$, $p < 0.001$, see FIG 4), while there was no indication that this effect differed between the groups (group x time interaction: $p > 0.05$; main effect of group: $p > 0.05$). A similar analysis on distances for the /ɛ/ category showed no effects ($p > 0.05$).

Given the fact that even native speakers' vowel distributions in the F1-F2 space vary

[6] To reassure us that results based on automatic phoneme segmentations by the tool are valid, a subset of the speech recordings was manually segmented and F1 and F2 values extracted from the two processing versions were compared. This comparison showed that although the exact formant values differed slightly (due to influences of co-articulation, the exact borders of the vowel segment will influence the mean formant values across the vowel duration), the overall pattern was similar and led to the same pattern of statistical outcomes as presented below.

widely (as we saw in FIG 2), we complemented the above analyses by also assessing how distinctively the two vowels were pronounced with respect to each other (FIG 5) instead of how close each of them were to the respective native example. To do this, we again used the Mahalanobis distance. For every measurement and participant, we computed the distance between a vowel distribution to the centre of the respective other vowel's distribution. The mean distance across those two directions per participant was then used as the dependent variable in a repeated measures ANOVA including the within-subject factor time (pre-test 1, post-test 1, pre-test 2, post-test 2) and between-subject factor group (experimental, control). Similar to the above results, this test revealed a main effect of time (F(3,48) = 5.86, p = 0.004) but neither a main effect of group, nor group x time interaction (p > 0.05). The distance between the two vowel categories increased over the course of training and the two groups did not differ in this learning effect.

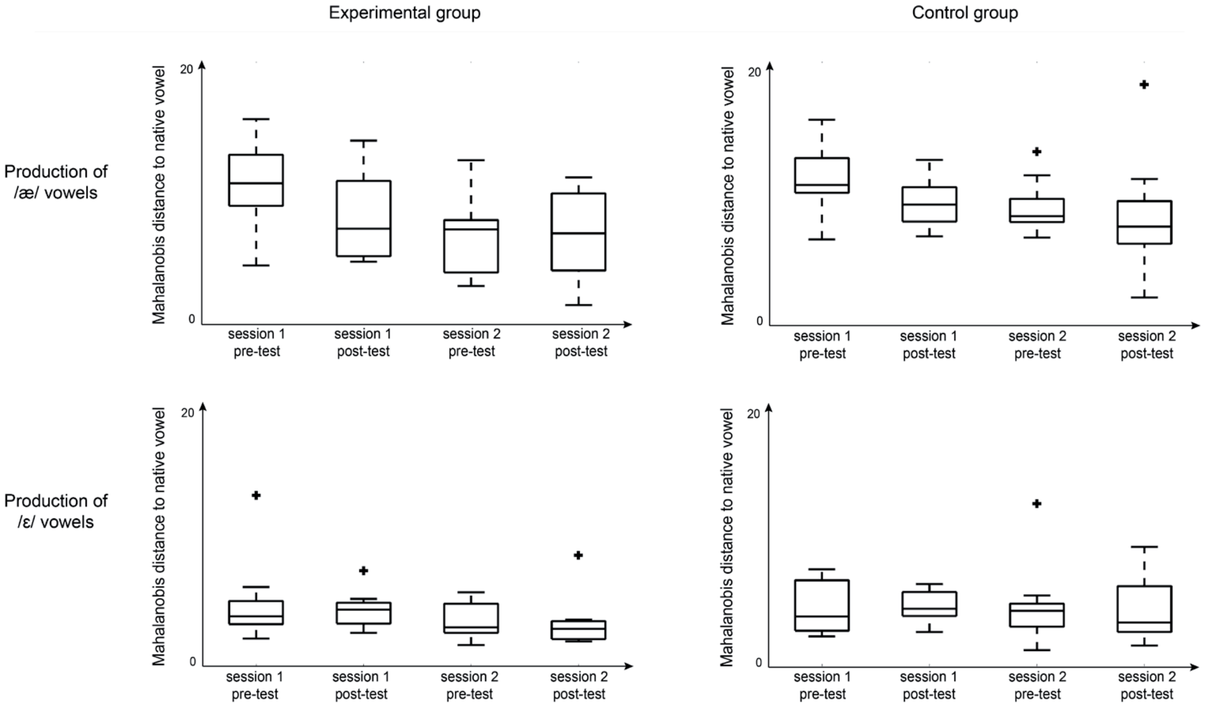

FIG 4. Experiment I. Mahalanobis distance to native model for /æ/ vowel [top row] and /ɛ/ vowel productions [bottom row] comparing performance of experimental group [left column] and control group [right column] for the four times of measurement. Each boxplot's central mark indicates the median of a given measurement, while the outer lines of the box represent the 25th and 75th percentile respectively. Outliers are denoted by '+' and the outer marks indicate the most extreme points of the data excluding any possible outliers.

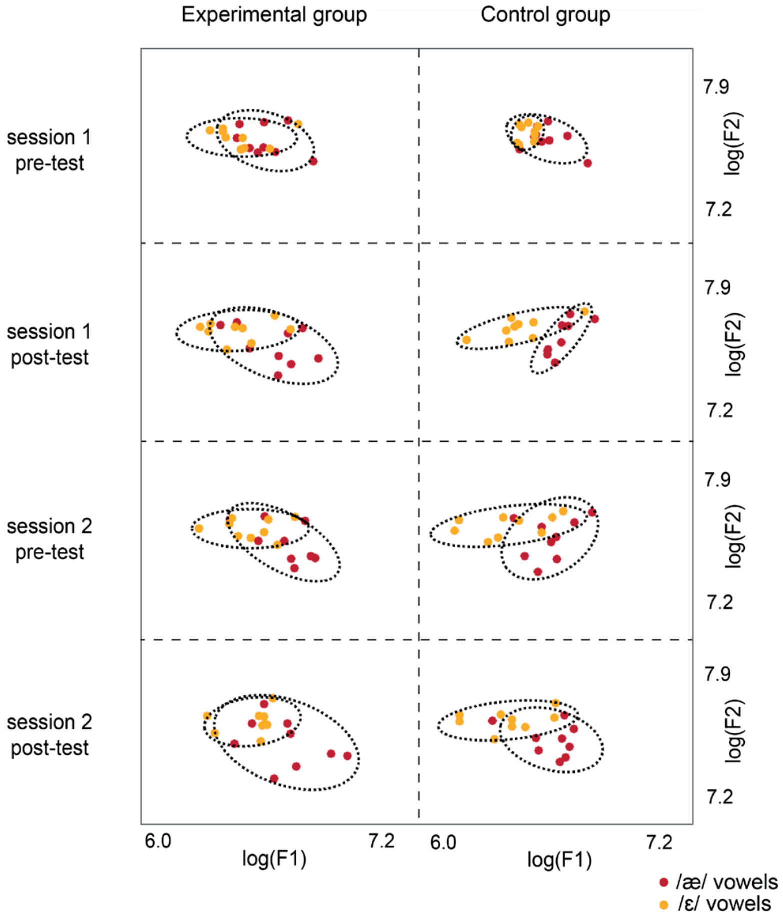

FIG 5. Experiment I. Mean log(F1) and log(F2) values of vowel productions by participants in the experimental [left column] and control group [right column] at the four times of measurement [top to bottom rows] contrasting /æ/ vowel [red] and /ɛ/ vowel [yellow] productions. Dotted lines indicate 95% confidence ellipses.

In order to assess the transfer of production learning to new stimulus words, the verbal responses from the transfer production test were similarly analysed in terms of their distance between the two vowel categories in the logF1-logF2 space. (Note that we did not have a native model for the transfer stimuli and thus did not compare the Mahalanobis distance between non-native and native utterances). Cleaning of the transfer data led to no exclusions due to exceedingly low durations (< 20ms) but 3.2% removal due to deviating and thus likely invalid formant values (> 2.5 standard deviations). A repeated measures ANOVA comparing the pre-training and transfer level (within-subject factor time) Mahalanobis distance for the two groups (between-subject factor group) revealed a main effect of time (F(1,26) = 10.98; p < 0.004), but no main effect of group or time x group interaction (p > 0.05). The distance between the categories for both groups was significantly larger in the productions of the transfer words after training as compared to before the training (mean values: 2.33 and 4.94 before training and 11.55 and 11.30 at transfer for the experimental and control group respectively, see FIG 6).

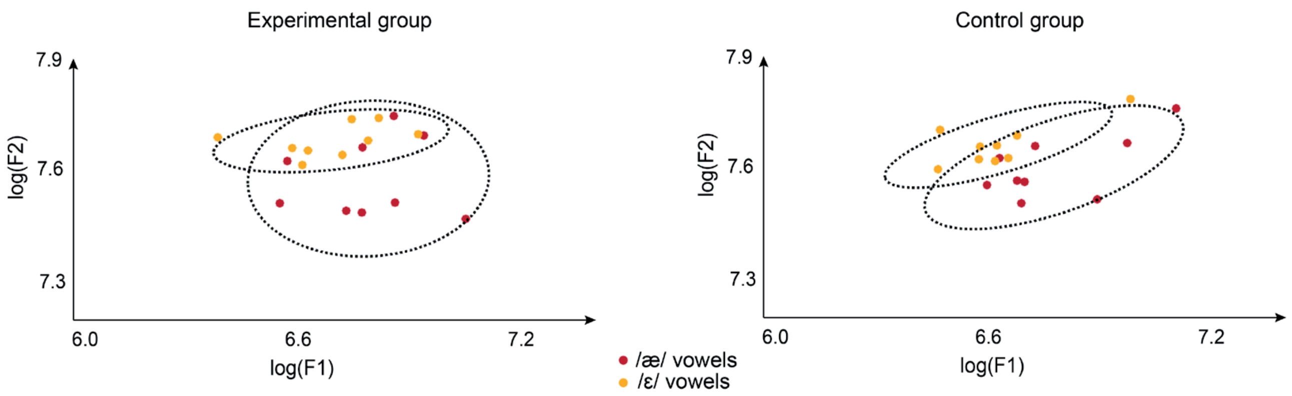

FIG 6. Experiment I. Production transfer. Mean log(F1) and log(F2) values of vowel productions by participants in the experimental [left column] and control group [right column]. Dotted lines indicate 95% confidence ellipses.

B. Perceptual learning

A 2-way repeated measures ANOVA on d prime scores with the within-subject factor time (pre-test 1, post-test 2, pre-test 2, post-test 2) and the between-subject factor group (experimental, control) revealed a main effect of time (F(3,48) = 4.96, p = 0.004), but neither a main effect of group nor group x measurement interaction (p > 0.05). The scores thus significantly changed over the course of training, being larger for both the experimental and control group after (2.43 and 2.87 respectively) as compared to before the training (1.72 and 2.24 respectively, see FIG 7).

Transfer of perceptual learning to new stimuli and speakers was tested by means of a similar repeated measures ANOVA with between-subject factor group (experimental, control) comparing d prime scores between pre-test 1 and at transfer identification task (see FIG 7-B). The d prime scores at transfer test were shown to be significantly larger (mean = 2.79) than at pre-test (mean = 1.98; main effect of time F(1,16) = 23.34, p < 0.001), but did not differ between the groups (no main effect of group, nor time x group interaction, p > 0.05). Taken together, this indicates that the perceptual learning successfully transferred to new stimulus words for both groups.

C. General English language measures

To quantify overall proficiency in English and any differences between the groups related to it, we computed a condensed self-report measure by adding the scores each participant had given herself/himself for the four subcategories (listening, speaking, reading, writing) on the 1-7 Likert-scale (1= very poor, 7 = native-like). For the verbal fluency test, a score was computed by counting all valid and unique words in each of the two categories (animals, start letter ‘S’). The percentage of use score is the answer participants gave to the question on how much they use English in their everyday live. Independent t-tests comparing the two groups (experimental, control) on each of those test results revealed no significant differences on any of the scores (p > 0.05, see TABLE II). There is thus no indication that the two groups differed with respect to their general English proficiency.

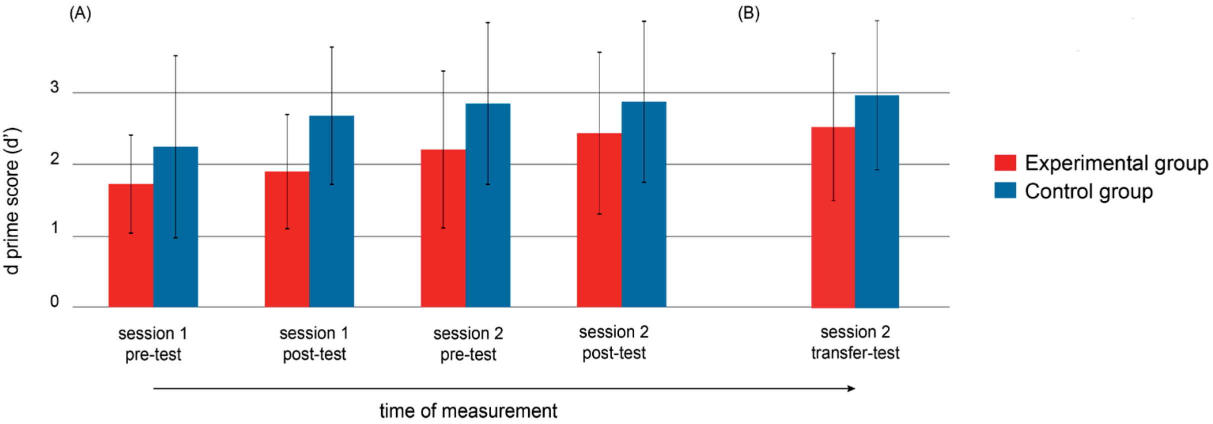

FIG 7. Experiment I. Perceptual performance. **(A)** Grand average d prime scores for the two groups (red = experimental group, blue = control group) during the four measurements: pre-test session 1, post-test session 1, pre-test session 2, post-test session 2. Error bars indicate standard deviations. **(B)** Grand average d prime during transfer identification task. Error bars indicate standard deviations within a given group and measurement.

TABLE II. Experiment I. Behavioural scores (means and SDs) of participants' English language skills measured by verbal fluency tests, condensed self-report measure of their proficiency in speaking, listening, reading and writing, as well as the percentage of their use of English in every day.

Group	Verbal fluency		Proficiency	Use of English (%)
	animals	first letter	(self-reported)	
Experimental	18.1 (± 2.9)$^{n.s.}$	15.9 (± 3.3)$^{n.s.}$	20.0 (±7.1)$^{n.s.}$	18.1 (±12.5)$^{n.s.}$
Control	19.1 (± 3.9)	14.4 (± 4.3)	23.1 (±3.5)	20.6 (±21.3)

$^{n.s.}$ indicates non-significant results of independent sample t-tests comparing the two groups (p > 0.05).

III. EXPERIMENT II

1. Methods

After the error in the feedback tool was fixed and it was thus possible to avoid disruptive effects that might have occurred in Experiment I, we decided to re-run the training with a larger number of participants. The procedure, tasks and stimuli were identical and as described above. The only addition to what was given participants in Experiment I consisted of a post-training questionnaire, which participants were asked to indicate their motivation to gain a native-like accent after training and whether the training was helpful in reaching a native-like accent in English. As in the pre-training questionnaire, they could indicate both of their answers on a Likert-like scale ranging from "No, not at all" to "Yes, very much" (note: in Experiment I, we only measured motivation before but not after training).

A. Participants

Twenty-eight native Dutch-speaking females (mean age = 22.2 ± 3.0) who were lower-intermediate/advanced L2 speakers of English participated (TABLE III). Participant's language background was similar and general requirements were identical to the ones described for Experiment I (see above).

TABLE III. General information on the experimental and control group of Experiment II (mean values followed by SDs).

N	Age	LexTALE
14	21.3 (± 2.5)$^{n.s.}$	72.1 (±13.9)$^{n.s.}$
14	22.2 (± 3.0)	71.6 (± 14.1)

$^{n.s.}$ indicates non-significant results of independent sample t-tests comparing the two groups (p > 0.05).

2. Results

A. Production learning

The data processing and analyses were executed as described in Experiment I. The data cleaning procedure led to 1.8% and 1.7% removal of the data due to exceedingly low vowel durations (< 20 ms), which were likely due to errors in the automatic segmentation method and/or high levels of noise, and deviating formant values (> 2.5 standard deviations from participant's mean for a given vowel on a given measurement) respectively.

We again used a repeated measures ANOVA including within-subject factor time (pre-test 1, post-test 1, pre-test 2, post-test 2) and between-subject factor group (experimental, control) to evaluate changes in Mahalanobis distance to the native model per vowel category (FIG 7). Results for /æ/ vowel productions indicated, similar to Experiment I, a significant main effect of time (F(3,78) = 4.53, p = 0.006) and again no main effect of group (p > 0.05) nor an interaction effect for time x group (p > 0.05). In other words, the Mahalanobis distance to the native model significantly decreased during training in a similar manner for both of the groups (see FIG 8).

An analogous ANOVA on the production of /ɛ/ vowels revealed again no main effect of time and time x group interaction (p > 0.05), but a main effect of group (F(1,26) = 5.96, p = 0.022). This means that while the two groups did not change their productions of the /ɛ/ vowels over the course training, they showed an overall group difference in their pronunciation of these vowels that was likely independent of the training.

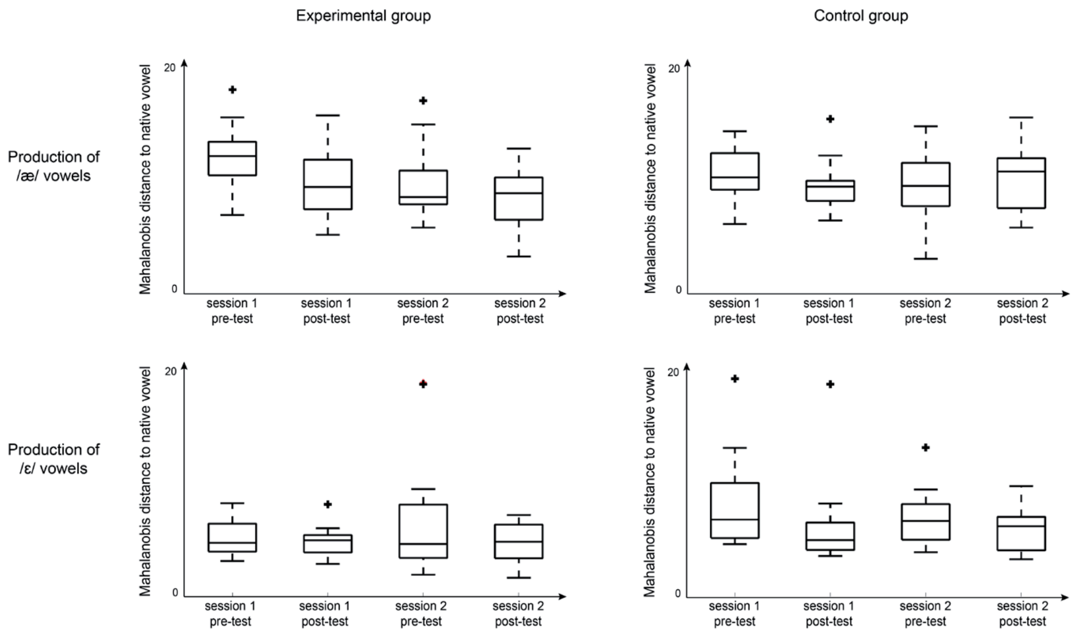

FIG 8. Experiment II. Mahalanobis distance to native model for /æ/ vowel [top row] and /ɛ/ vowel productions [bottom row] comparing performance of experimental group [left column] and control group [right column] for the four times of measurement. Each boxplot's central mark indicates the median of a given measurement, while the outer lines of the box represent the 25th and 75th percentile respectively. Outliers are denoted by '+' and the outer marks indicate the most extreme points of the data excluding any possible outliers.

We again complemented this first analysis by also assessing the distance between the two vowel productions with respect to each other instead of to the native vowel model (FIG 9). To this end, we conducted another repeated measures ANOVA evaluating the influence of within-subject factor time (pre-test 1, post-test 1, pre-test 2, post-test 2) and between-subject factor group (experimental, control) on the Mahalanobis distance between vowel categories. It showed both a significant main effect of time (F(3,78) = 5.68, p = 0.001) and group (F(1,26) = 6.47 p = 0.017), together with a significant time x group interaction (F(3,78) = 5.17, p = 0.003). Both groups develop a larger distance between categories over the course of training. However, the control group shows a less linear improvement and overall larger distance, though also with a notably larger variability in productions across participants (see FIG 10).

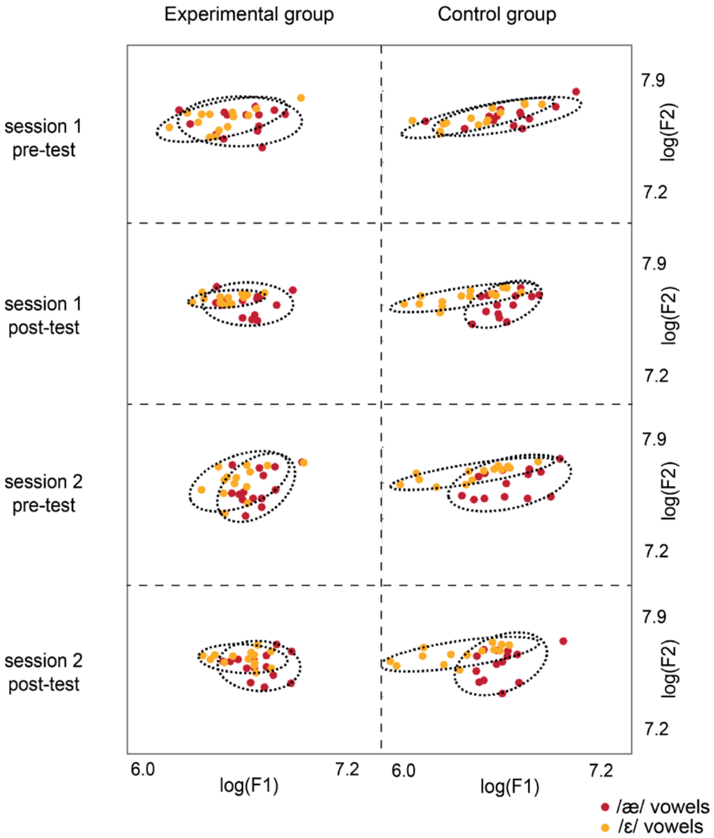

FIG 9. Experiment II. Mean log(F1) and log(F2) values of vowel productions by participants in the experimental [left column] and control group [right column] at the four times of measurement [top to bottom rows] contrasting /æ/ vowel [red] and /ɛ/ vowel [yellow] productions. Dotted lines indicate 95% confidence ellipses.

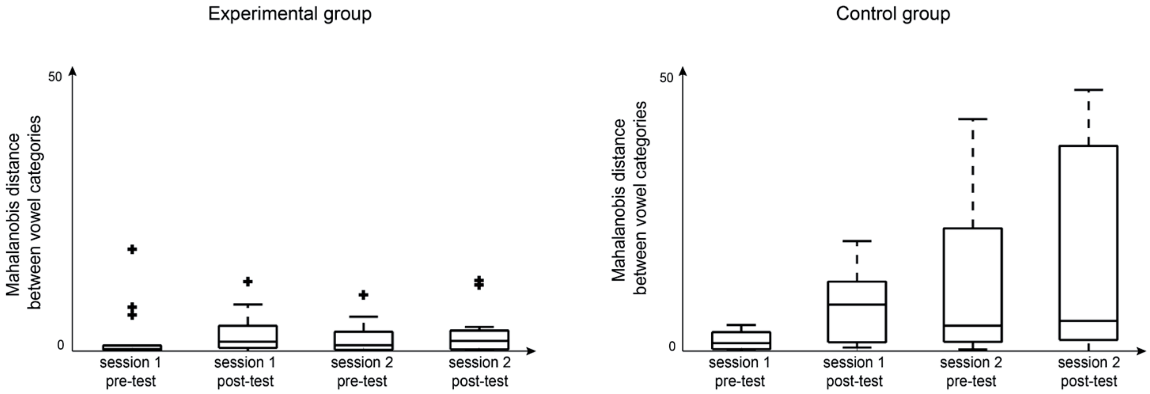

FIG 10. Experiment II. Average Mahalanobis distance between vowel categories for the experimental group [left] and the control group [right] at the four times of measurement. Each boxplot's central mark indicates the median of a given measurement, while the outer lines of the box represent the 25th and 75th percentile respectively. Outliers are denoted by '+' and the outer marks indicate the most extreme points of the data excluding any possible outliers.

After cleaning the transfer data similarly to the above procedures (4.1% and 8.4% exclusions due to low duration and deviating formant values respectively), a repeated measures ANOVA including the between-factor group (experimental, control) and the within-factor time (pre-test, transfer-test) was used to assess differences in Mahalanobis distance between vowel categories before training with those at transfer test. Similar to outcomes of Experiment I, results showed a significant main effect of time F(1,26) = 4.51; p < 0.043, but no main effect of group, nor a time x group interaction effect (p > 0.05). The distance between the two phoneme categories was significantly larger at transfer test (18.21 and 5.28 for experimental and control group respectively) as compared to pre-test (2.73 and 2.00 for experimental and control group respectively) and did not differ between the groups (see FIG 11).

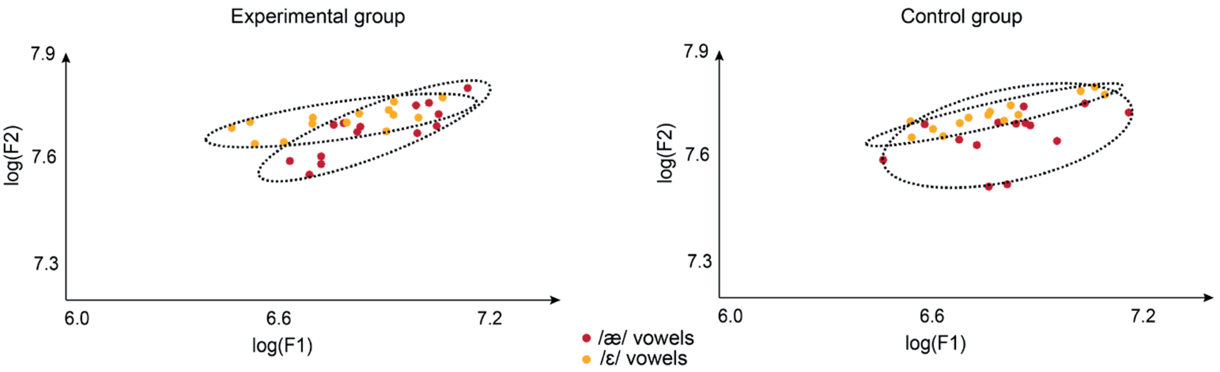

FIG 11. Experiment II. Production transfer. Mean log(F1) and log(F2) values of vowel productions by participants in the experimental [left column] and control group [right column]. Dotted lines indicate 95% confidence ellipses.

B. Perceptual learning

An analogous 2-way repeated measures ANOVA on d prime scores as used in Experiment 1 revealed the same pattern of effects: a main effect of time (F(3,78) = 13.39, p < 0.001), but neither a main effect of group nor a group x time interaction (p > 0.05). Average d prime scores were significantly larger for both the experimental and control group after (2.29 and 2.44 respectively) as compared to before the training (1.81 and 1.74 respectively, see FIG 12-A).

We again also evaluated to what extent perceptual learning transferred to new stimuli and speakers by comparing d prime scores before training (pre-test 1) with those at the transfer identification task. A repeated measures ANOVA with between-subject factor group (experimental, control) and within-subject factor time (pre-test, transfer) showed a significant main effect of time (F(1, 26) = 43.30, p < 0.001), but no main effect of group nor a time x group interaction (p > 0.05). The d prime scores at transfer (mean = 2.65) were significantly larger than at pre-test (mean = 1.78), which indicates a successful transfer in both groups (see FIG 12-B).

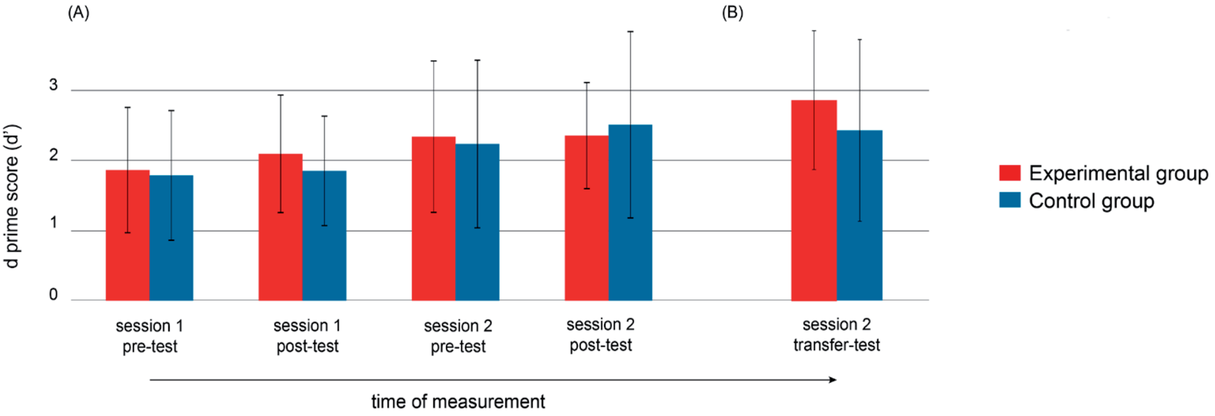

FIG 12. Experiment II. Perception performance. (A) Grand average d prime scores for the two groups (red = experimental group, blue = control group) during the four measurements: pre-test 1, post-test 1, pre-test 2, post-test 2. Error bars indicate standard deviations. (B) Grand average d prime during transfer identification task. Error bars indicate standard deviations within a given group and measurement.

C. General English language measures and motivation

There was no indication that the two groups (experimental, control) differed with respect to their English proficiency as measured by verbal fluency, self-reported proficiency and use of English (see TABLE IV). This was shown by non-significant results of independent sample t-tests comparing the two groups on each of the test scores (p > 0.05). Additionally, we also compared participants' motivation before and after training as well as perceived helpfulness of the training and did not find any differences (neither in time nor between the groups, p > 0.05, see TABLE V).

TABLE IV. Experiment II. Behavioural scores of participants' English language skills measured by verbal fluency tests, condensed self-report measure of their proficiency in speaking, listening, reading and writing, as well as the percentage of their use of English in every day (mean values followed by SDs).

Group	Verbal fluency		English proficiency	
	animals	first letter	(self-reported)	Use of English (%)
Experimental	20.0 (±3.0)$^{n.s.}$	14.6 (± 4.6)$^{n.s.}$	21.0 (±5.8)$^{n.s.}$	16.4 (±15.7)$^{n.s.}$
Control	18.4 (± 6.3)	14.6 (± 5.2)	22.8 (±4.4)	20.9 (±13.1)

$^{n.s.}$ indicates non-significant results of independent sample t-tests comparing the two groups (p > 0.05).

TABLE V. Experiment II. Behavioural scores (mean values followed by SDs) reflecting motivation to reach a native-like accent before and after the training as well as the degree to which the training was perceived as helpful (both: score 1 "No, not at all" to 5 "Yes, very much").

Group	Motivation		Perceived helpfulness of training
	pre	post	
Experimental	4.1 (±0.7)$^{n.s.}$	4.0 (±0.8)$^{n.s.}$	3.6 (± 0.7)$^{n.s.}$
Control	4.0 (±0.9)	3.9 (±0.9)	4.0 (± 0.7)

$^{n.s.}$ indicates non-significant results (p > 0.05).

IV. DISCUSSION

The present study evaluated the effectiveness of a two-session computerised pronunciation training protocol, which provided Dutch native speakers with trial-by-trial visual feedback on their productions of the English /æ/-/ ɛ/ vowel contrast. We evaluated the protocol's effectiveness by comparing participants' vowel productions before and after training in terms of both their distance to those of a typical native speaker and in terms of how distinctively the two vowel categories were pronounced with respect to each other. A second aim of the current study was to further our understanding of the interactions between the speech perception and production modality during second-language sound learning by testing the effect of improved L2 pronunciation on perceptual learning.

Before discussing the result patterns in more detail, we would like to note that the findings of the two experiments will be considered as equally valid. The original reason for re-running the same experimental setup (Experiment II), were technical disturbances in Experiment I in the form of a reoccurring runtime error, which caused short interruptions (in the range of seconds) for participants in both groups. To avoid (and also directly test) reduced efficiency of production training due to those disturbances, we decided to repeat the experiment while employing a fixed version of the tool and larger number of participants. Results of both experiments show improvements in production suggesting that the task disruption during Experiment I was not that severe as to prevent or severely inhibit learning. We therefore treat the two experiments as two independent but equally valid datasets.

Overall, both experiments show improved vowel productions developing over the course of training, both in terms of closer resemblance with typical productions of a native speaker as well as more distinct pronunciation of the two vowels with respect to each other. In addition to this, production learning was also shown to transfer to new stimulus words. Interestingly, the results of both experiments show such production improvements (and transfer of it) developing not only for the experimental group but also for the control group. In fact, measured in terms of similarity with the native model, the two groups were not shown to differ from each other in either of the two experiments. Neither was there any evidence of group differences regarding their distance between vowel categories throughout the training in Experiment I. Although in Experiment II the exact time course of how participants in both groups started to distinctively pronounce the two phonemic categories differed (the control group developing a clearer distance on average), also here both groups were shown to improve over the course of training. Importantly in this context, the experimental and control groups were not shown to differ regarding their general English proficiency or age and were thus sufficiently well matched. Taken together, we could not show any advantage of external feedback on vowel productions in the form it was given here, while all participants improved in their ability to produce the non-native vowel contrast. This finding could be explained in different ways.

While participants in the experimental group were provided with trial-by-trial feedback on their own productions, participants in the control group received solely an indication of how the two vowels should ideally be located to each other (in terms of mouth and tongue location during pronunciation). As the experimental group did not show any benefit of training, it seems straightforward to assume that the visual feedback was overall not helpful for the participants in evaluating their own vowel productions. A possible reason could be the specific type of visual feedback. The graphical mapping of tongue and mouth location during pronunciation might have been too abstract and thus not sufficiently intuitive to be translated into concrete articulation adjustments (though similar types of feedback have successfully supported production learning in previous studies; see again Kartushina et al., 2015; Kartushina and Martin, 2019; Lie-Lahuerta, 2011). We also know that the automatic vowel segmentations and formant tracking of the feedback tool is not flawless in the sense that the feedback in some few trials will incorrectly reflect the actual spectral information of participants' utterances and will therefore be invalid and potentially misleading. Aside from that, the native speaker's model was based on the production data of two English speakers only. Although we did account for variability between speakers by making use of vowel calibration before each training session (see methods) this might not have sufficiently accounted for production variability (see again high variability of native speakers in FIG 2). A possible way to circumvent the issue of highly variable and strongly overlapping vowel categories across speakers as part of external feedback, would be to (also) provide feedback on the distinctiveness between vowel categories within a given participant (provided that the productions are still in reasonable distance to a general native model), instead of distance to a native model.

Aside from questioning the quality of feedback itself, however, the inefficiency of the training tool in further supporting learning could also be explained by the possibility that external feedback might be less relevant in facilitating production improvements than expected and, more specifically, as compared to internal feedback or other relevant factors present in the training protocol. After all, it is interesting to note that production learning took place in both groups and in both experiments. Evidently, the control group learns well in an unsupervised way, while the experimental group might have even been distracted by trying to make sense of the (potentially unintuitive) visual representation. Especially in the control group, production learning must therefore have been supported by (a combination of) other factors than the external feedback, such as the phonetic instructions given before the first training session, awareness gained of the difficult phonemic contrast (some participants were not aware of the fact that they had habitually mispronounced the two English vowels), as well as extensive practice and some exposure to native pronunciation examples (in the form of auditory stimuli in the perception tests).

When considering phonetic instruction and increased awareness of the difficult contrast as potential driving forces for production learning in the present experiment, it is important to consider the time course of production learning. Phonetic instructions were given before the first training session but after the pre-test of session 1. If this factor had indeed strongly facilitated learning, we would expect to see the clearest improvement at the post-test of session 1 and not much further improvement after this. Looking at the relatively continuous improvement (see FIG 4 and FIG 8), however, it seems unlikely that those two factors (alone) could entirely account for the pattern of production learning.

In this context, it is helpful to also consider the outcomes of the motivation scores measured in Experiment II. Participants in both groups felt similarly motivated to reach a native-like accent in English and tended to reply with "yes" when asked if they felt motivated. The desire to acquire a native-like accent has also previously been associated with positive effects of computerised production training protocols (Lie-Lahuerta, 2011). Interestingly, participants' motivation after participation in the training was found to be independent of whether they had previously undergone the experimental or the control version of the training. Overall, a plausible explanation for the production learning present in both groups and thus independent of external feedback by the tool, could be that a combination of explicit pronunciation instructions, focused attention on the challenge and motivation to improve pronunciation led to more efficient internal evaluation, which in turn then played a key role in supporting the process of improving pronunciation while actively practicing the vowel contrast over the course of two training sessions.

Returning to the second aim of the present study, concerning the interactions between the two speech modalities during learning, it becomes relevant that both experiments revealed perceptual learning over the course of production training. The two groups were again not shown to differ in this effect and gains in perception ability also transferred to new stimuli and a new speaker. In other words, production improvements in both groups went hand in hand with perceptual learning despite no direct training in the perception modality. This is in line with results from other production training studies revealing a transfer from production to perceptual learning (Herd et al., 2013; Hirata, 2004a; Kartushina et al., 2015). It stands, however, in contrasts to findings by Hattori (2009) and Wong (2009), in which successful production training did not lead to any gains in perception.

There are various differences in the design of these studies and the present experiments that could account for the contrasting results, though it will, at this stage, be difficult to pinpoint the exact underlying reasons. First of all, all the above studies use different combinations of L1 and L2. We know that the phonemic spaces of both L1 and L2 mutually influence each other during non-native speech perception, production and also during L2 sound learning (Kartushina et al., 2016b, 2016a). The degree to which production learning transfers to perceptual learning is thus likely influenced by the exact contrast that is trained and in which way it relates to or gets assimilated by the native sound space. Another varying factor between studies is the degree to which learners are already familiar with the trained L2 contrast prior to training. The differential findings, however, cannot be solely explained by participants' familiarity. While Kartushina et al. (2015) showed successful transfer from production to perception learning when training French speakers on Danish vowels, an entirely unfamiliar set of sound categories, all of the other discussed studies, including the present one, trained challenging but already known phonemic contrasts and revealed different findings with regard to cross-modality transfer.

Another crucial difference among the above-mentioned studies is that they vary tremendously in the type of training and, more specifically, the way they did (or did not) implement feedback. Especially concerning the studies investigating the effects of combined perception-production training most did not provide learners with any direct, external feedback on their speech productions (Baese-Berk, 2019; Baese-Berk and Samuel, 2016; Lu et al., 2015; Thorin et al., 2018). In the production-only training by Herd et al. (2013) learners were expected to use self-reflective feedback. They could, after each verbal response, compare waveforms and spectrograms of their own pronunciations to those of native speaker examples and were asked to match those as closely as possible. In the training study presented by Hattori (2009), learners were supported by a phonetically trained Japanese-English speaker who served as personal instructor and could make use of real-time spectrograms in order to track a learner's pronunciation quality. Learners in this study additionally listened to signal processed versions of some of their own productions (aimed at removing between-speaker variance) and discussed those with the personal instructor. Again another feedback approach was used in the training by Hirata (2004), in which learners received visual feedback based on pitch contours. Also Kartushina (2015, and also similar training protocols employed in Kartushina et al., 2016; Kartushina and Martin, 2019) used visual feedback based on spectral features of the learners' production in each trial. More specifically, in each trial, learners' phoneme productions were followed by a visual representation of the distance in the F1-F2 space (first and second formant) between a learner's pronunciation and that of a typical native speaker's production. Those large differences in training design and type of training make it difficult to directly compare those studies and infer which factors exactly influence cross-modality transfer.

Although the production training design by Kartushina et al. (2015) is relatively similar to the present one, outcomes between the two studies differ with respect to the efficiency of visual feedback. While the experimental group in the present study did not show stronger production gains in the course of training than the control group, participants in the experimental group in the study by Kartushina et al. did improve significantly more than those in an untrained control group (with same amount of productions but no external feedback), in which production performance was unchanged. In short, a similar type of visual feedback was constructive in the study by Kartushina et al., but was overall not helpful in the present study. The most crucial differences between the two studies, which might explain differing effectiveness of visual feedback, concern (1) familiarity with the trained contrast (unknown Danish vowels as compared to already familiar vowels in the present study), (2) the fact that the L2 sounds were produced in isolation instead of in words as in the present training, and (3) possible differences in L1 assimilation given different combinations of L1 and L2 sound spaces (i.e. native French speakers trained in Danish vowels in Kartushina et al. as compared to native Dutch speakers trained in English vowels here). Interestingly, however, both studies show learning both in production and perception despite of the absence of any direct training in the perceptual modality.

In future research, it would be recommendable to establish and widely use more standardised training procedures for pronunciation training. Ideally, these procedures should enable researchers to distinguish between (production) learning relying on external and internal evaluation. In other words, it is advisable to make use of paradigms that employ a form of immediate, trial-by-trial feedback on the quality of learners' productions and contrast those with more unsupervised procedures in which learner's evaluation is based on self-evaluation (i.e. untrained control groups with similar exposure to their own productions). More controlled comparisons between different studies could then provide clearer insights into the influence of a variety of relevant factors as, for instance, mutual interactions between L1 and L2 phonemic spaces, type of feedback, imitation versus reading and/or picture naming, and familiarity with trained non-native contrast.

In sum, despite no direct effects of the employed visual feedback as production training method, the present study shows production learning that likely relied on internal feedback supported by explicit pronunciation instructions and awareness of the challenging non-native contrast. Most interestingly, we also observed a cross-modality transfer from production learning to simultaneous improvements in the perception of the trained speech contrasts. In combination with the earlier established transfer in the reverse direction, namely perceptual learning improving production performance of novel sounds, these results point towards a bidirectional (though not necessarily balanced) relationship between the speech perception and production modality in the process of establishing non-native speech categories. Further investigations into the exact factors influencing the mutual relationship between the modalities are still needed.

V. ACKNOWLEDGMENTS

This project would not have been possible without Gisela Govaart who played a major role in developing the production feedback tool used in this study. I would also like to thank Laura Pijnacker and Jessie Waaning for their support with the data collection, as well as Paul Boersma and Dirk de Vet for their technical advice during the development of the tool.

5

Chapter 5

Error monitoring of the production of newly-learned non-native vowels

ABSTRACT

The verbal self-monitor enables language users to detect and correct their errors in everyday language use. The present study investigates how easily this system can adapt to newly-learned non-native elements and thereby support second language speech acquisition. Dutch natives who were previously trained on the perception and production of a challenging speech contrast, the English /æ/ and /ɛ/ vowels, as well as an untrained control group, engaged in a phoneme substitution task. In this fast-paced verbal response task, participants had to substitute the vowel of visually presented English words by its counterpart (æ/ or /ɛ/ respectively) whenever it included one of the trained English vowels (for example, SAND should be replaced by responding "send"). Both groups made a substantial number of phoneme substitution errors (overall 26%). Results from electrophysiological measurements, however, revealed no differential neural responses following erroneous and correct responses in the typically observed latency for ERN effects for either of the two groups. Different reasons for these null findings are discussed. Overall, there was no evidence that verbal self-monitoring of non-native vowels differed as a function of whether they have been trained or not.

I. INTRODUCTION

Speech errors, such as mispronunciations, shifts or substitutions at various levels of the articulatory process (Fromkin, 1971), are usually an unintentional by-product of normal speech production. While they might be inconvenient or in some contexts even embarrassing for the language user herself, they provide language researchers of various fields with a valuable tool to investigate the nature of the verbal self-monitoring system (Hartsuiker and Kolk, 2001; Levelt, 1983; Nozari et al., 2011). It is this system that enables language users to realise their mistakes quickly after they occur, but also to correct or, in some cases, even avoid them before they get articulated. It does so not only while using one's native language (L1) but also during the use of a second language (L2). In fact, speakers tend to make more errors when speaking a non-native as compared to their native language (Poulisse, 1999, 2000). This makes it likely that the verbal self-monitor plays a crucial role in the context of second language use and, more specifically, when mastering challenging elements of it, such as the pronunciation of novel sounds. But how easily can the monitoring system adapt so as to be able to evaluate accurately newly-learned L2 elements and hence support further acquisition? The current study addressed this question by examining verbal self-monitoring of L2 learners who were previously trained on a challenging non-native vowel contrast. More specifically, we investigated whether those learners would show typical electrophysiological (EEG) signatures of error monitoring and response conflict related to self-produced speech errors involving the trained non-native sounds.

A key factor in establishing a native-like accent in an L2 acquired in adulthood is to successfully differentiate between its sound categories, both in perception and production. Developing this ability can be a major challenge for language learners, especially if the non-native vowel space exhibits two or more phoneme categories that get assimilated into a single category in their L1 (Best, 1995; Best and Tyler, 2007a). An example for such an assimilation process is the Dutch vowel category /ɛ/ (as in Dutch *pen*) that lies in between the vowels /æ/ and /ɛ/ in the English phonological system (as in English *pan* and *pen*). Although the /æ/ category might already be weakly established in experienced Dutch speakers of English (Weber and Cutler, 2004), even proficient Dutch learners of English tend to have difficulty in both accurately perceiving and pronouncing the two vowels (Broersma, 2005; Escudero et al., 2008; Thorin et al., 2018; Wanrooij et al., 2014).

A natural question arising here is how learners of a second language can still establish a novel phonemic category despite their difficulties given L1 assimilation. A body of studies with various combinations of L1 and L2 sound systems showed that targeted perception training of a novel speech contrast can yield positive results in the form of both improved perception and production (e.g. Herd et al., 2013; Lambacher et al., 2005; Lee and Lyster, 2017; Lopez-Soto and Kewley-Port, 2009) . There are also examples of different pronunciation training schemes involving visual feedback in the form of spectral features of learners' utterances (Hattori, 2009; Hirata, 2004a; Kartushina et al., 2015; Lie-Lahuerta, 2011), speakers' tongue movement (Katz and Mehta, 2015), or outcomes of automatic speech recognition (Arora et al., 2018), which were all shown to successfully improve production performance. In some cases, improvements in production also transferred to advances in perceptual ability (Kartushina et al., 2015), though the transfer from perceptual training to production learning tends to be larger than vice versa (Sakai and Moorman, 2018). Training protocols combining both perception and production practice, however, have resulted in mixed results ranging from disrupted improvement in both modalities (Baese-Berk, 2019; Baese-Berk and Samuel, 2016) to greater gains in perceptual ability (Thorin et al., in revision) or production performance (Herd et al., 2013) as compared to single-modality training, in part depending on the learner's familiarity with the trained non-native contrasts (Baese-Berk, 2019).

Although the exact interaction between the perception and production modality in the process of non-native phoneme learning is still inconclusive, what the above studies have shown is that the vowel space stays adaptive in adulthood (though it is likely and has been shown to decrease in plasticity with age Flege et al., 1999). Substantive improvement towards a native-like accent can be achieved by engaging in targeted training. The present study focusses on the degree to which error monitoring is involved in this process. It seems intuitive that accurate perception is a pre-requisite for successful verbal self-monitoring (at least concerning the external route). But does that also mean that improvements in perception go hand in hand with successful error monitoring? In other words, will an L2 learner get a response conflict whenever mispronouncing a newly established phonemic category, or is non-native, though highly-proficient, perception of a novel L2 sound category not sufficient to enable (native-like) error monitoring?

A way to investigate this question is to employ the error-related negativity (ERN), a widely used event-related potential (ERP) among others observed in the context of an erroneous response action (Falkenstein et al., 1991; Gehring et al., 1993). The potential is known to peak around 80-100 ms after the erroneous action and is likely produced by sources in anterior cingulate cortex (ACC; Miltner et al., 2003), which have been related to action monitoring. After primarily being used in research on performance monitoring, the first ERN in speech production was observed following vocal slips in the Stroop colour word task (Masaki et al., 2001). Since then, it has also been found related to word production errors during other experimental tasks as, for instance, a phoneme substitution task (Trewartha and Phillips, 2013) or a phoneme monitoring task (Ganushchak and Schiller, 2006).

Recently, also a few studies demonstrated an ERN response in the context of erroneous responses in L2 speech production. German learners of Dutch engaging in a word-gender training paradigm with immediate trial-by-trial feedback developed an ERN response to their incorrect gender assignments in the course of training, while also improving their behavioural performance (Bultena et al., 2017). Also Dutch-English bilinguals showed an ERN response whenever incorrectly switching between their L2 and L1 and vice versa (Zheng et al., 2018). More directly related to phonological self-monitoring in L2, Ganushchak and Schiller (2009) investigated the effect of time pressure on the performance of phoneme-monitoring in German-Dutch bilinguals. Their results showed that while errors both with and without time pressure triggered an ERN response, this response was enlarged under time pressure. The authors argue that this effect is due to stronger interference with L1 under time pressure which leads to increased response conflict and thus a larger amplitude ERN response.

Though the above studies have shown that the ERN can be observed in the context of monitoring L2 speech production, to our knowledge there is no study investigating error-monitoring of newly learned non-native phonemes during second language speech production. In order to do so, we employed a phoneme substitution task, which has proven to be a suitable tool in investigating verbal error monitoring in L1 (Trewartha & Philips, 2013). In this fast-paced task, participants were visually presented with single English words that either contained one of the trained English vowels (æ/ or /ɛ/) or not. Whenever it did so, participants had to mentally replace the vowel by its counterpart and quickly respond by verbally producing the substituted word (for example, SAND should be replaced by "send"). In case the word in a given trial did not include one of the targeted vowels, the correct verbal response was to say "no".

All participants included in the present study were part of a training study presented in Thorin et al. (2018, in revision). There Dutch learners of English engaged in a 4-day perceptual training protocol on the English /æ/-/ɛ/ vowel contrast that was either combined with also producing words containing the trained vowels (related production group) or combined with the production of unrelated tokens (unrelated production group), while a group of similar but untrained participants served as control. Results showed that participants in both training groups successfully improved in both their perception and production of the challenging non-native contrast (Thorin et al., 2018), though only the group which had undergone combined perception-production practice had developed an electrophysiological signature of change detection in the form of a mismatch negativity (MMN) after training (Thorin et al., in revision). In the present study we evaluated verbal self-monitoring (in terms of ERN responses) of the related production group and a control group in a phoneme substitution task involving the trained English /æ/-/ɛ/ vowel contrast. Though measuring EEG in the context of speech production is challenging due to extensive muscle activity during speech articulation, previous studies have shown that decent levels of signal-to-noise ratio can be achieved after suitable signal processing (Bultena et al., 2017; Zheng et al., 2018).

If improvement in the perception of novel speech categories indeed goes hand in hand with established speech error-monitoring related to those phonemes, we would expect to see ERN responses to substitution errors in the trained group (response-locked analysis). If L2 learners actively use their perception-based knowledge when monitoring their own productions this would also be further evidence for a tight interaction between perceptual and motor processes. Based on previous findings we would expect that if participants are not able to properly hear the difference between the two English phoneme categories (as evident in the control group), nor distinctively produce it, they would not get a response conflict when evaluating their own productions and therefore no ERN response would be detectable. As control analysis, we also compared stimulus-locked responses to check if both groups show an expected N1, response, typically triggered by an attended stimulus, and did not differ with respect to it.

II. METHODS

A. Participants

Thirty-two native speakers of Dutch (16 females, 16 males; mean age = 23.1 ± 4.0) took part in the experiment in two groups. Participants in the trained group had participated in the combined perception-production training on the British English /æ/-/ɛ/ contrast described in Thorin et al. (2018), while participants in the other group were untrained controls with similar perceptual identification performance on the critical phonemes as the trained group prior to training (TABLE I). Note that another four participants were originally trained, but they had to be excluded from the current study due to technical problems leading to incomplete datasets. All participants were upper intermediate to lower advanced speakers of English (see LexTALE (Lemhöfer and Broersma, 2012) results in TABLE I) with normal hearing as well as normal or corrected-to-normal vision and without any history of neurological or psychiatric disorders. The Ethics Committee of The Faculty of Social Sciences, Radboud University, approved the study and all participants gave their written informed consent prior to participation.

B. Design and Procedure

The experiment consisted of a single, approximately 2-hour session, during which participants were comfortably seated in front of a BenQ monitor (size 53.2 x 30 cm; 1920 x 1080 pixels; refresh rate of 60 Hz) in a shielded room. All auditory input was presented binaurally and at a comfortably chosen volume through in-ear headphones (Etymotic Research ER4P-T). All communication during the experiment, including verbal and written instructions, was in English.

The session started with a battery of short behavioural tasks in the following order: LexTALE, an identification task, an identification on morphed continuum task and a discrimination on morphed continuum task (see detailed descriptions below). Note that the discrimination task was not identical across the trained and control groups and will therefore not be considered further. After EEG cap fitting, participants first performed a passive word oddball task which was part of another study and will not be further described here, and then, most importantly for the present study, the phoneme substitution paradigm consisting of 10 blocks in total. Stimuli and further details of all relevant tasks (identification task, identification on morphed continuum and phoneme substitution task) will be specified below.

TABLE I. General information on the groups regarding number of participants, gender and age, as well as English vocabulary knowledge as quantified by the LexTALE.

Group	N	Gender (f/m)	Age	LexTALE
Training	15	7	23.7 (± 5.0)$^{n.s.}$	80.6 (± 9.6)$^{n.s.}$
Control	17	9	22.6 (± 2.8)	81.1 (± 14.2)

$^{n.s.}$ non-significant outcomes of an independent sample t-test comparing the two groups.

C. Stimuli

Behavioural tasks

The identification task stimuli consisted of five Consonant-Vowel-Consonant (CVC) words contrasting the target vowels /æ/ and /ɛ/ in minimal pairs: *fan-fen, ham-hem, jam-gem, man-men,* and *pan-pen*. For each word, seven tokens recorded by four native speakers of British English (2 male, 2 female). All recordings were duration normalised within word pair.

For the identification on morphed continuum task, two recordings of the words /væt/ and /vɛt/ (female speaker) were first normalised in duration and then adjusted regarding their F1 and F2 values using the software TANDEM STRAIGHT (Kawahara and Morise, 2011) to form an 11-step /væt/-/vɛt/ continuum.

Phoneme Substitution Paradigm

All stimulus words selected for the phoneme substitution task were monosyllabic English words or English pseudowords matched on mean word length and, in the case of word stimuli, also matched on frequency of occurrence as well as orthographic and phonological neighbours. Twelve monosyllabic minimal pair word sets contrasting the English /æ/-/ɛ/ vowels were selected, which had the advantage that all vowel substitutions resulted in other existing English words. Twenty four catch trial words were used as no-substitution trials (see TABLE II for an overview). For the practice substitution task, 27 pseudowords containing the /æ/-/ɛ/ vowels, which resulted in new pseudowords after vowel substitution, were used. The word list for the additional reading task consisted of 25 unrelated monosyllabic English words not containing relevant vowels.

TABLE II. Overview of all word stimuli underlying the analyses.

Word list /æ/-/ɛ/ substitutions	Word list catch trials
fan	big
ham	bin
land	bowl
man	bring
mash	brown
mass	bus
pan	chick
sand	chin
shall	cold
tan	cup
than	duck
vat	flip
fen	fold
hem	four
lend	gross
men	hug
mesh	inch
mess	kid
pen	lip
send	miss
shell	must
ten	plug
then	sold
vet	two

D. Experimental tasks

Behavioural tasks

The LexTALE task is a brief 2-minute test assessing lexical vocabulary size in English by presenting single words on the screen, for which participants have to press a button for “yes” or “no” to indicate whether they see an existing English word or not (Lemhöfer and Broersma, 2012). The final score of correctly classified words is displayed on the screen at the end and is known to correlate well with general English proficiency.

The identification task was a brief 2-alternative-forced choice (2AFC) task taking about 5 minutes to complete. In each trial, a single English word was played and participants subsequently had to indicate which of two words in a visually presented minimal pair they heard. In total, the task consisted of 120 randomly presented trials (10 stimuli x 4 speakers x 3 tokens each). The number of correct trials was presented to the participant as a score after the final trial.

The identification on morphed continuum task took about 4 minutes to complete per phoneme contrast and measured the boundary sharpness and position of boundary between the two given categories, such as English /æ/-/ɛ/ vowels. Similar to the previous task, participants were asked to carefully listen to single stimuli played to them in each trial, here one of the 11-step continuum, and then decide whether they heard either the word /væt/ or /vɛt/. Ten repetitions per stimulus resulted in a total number of 110 randomly presented trials. Note that all participants performed both the identification task and identification on morphed continuum task also on /b/-/p/ and /d/-/t/ morphed continua. These data are reported elsewhere (Garcia-Cossio et al., in revision).

EEG task: Phoneme substitution paradigm

The full phoneme substitution paradigm comprised 10 blocks differing in their task instructions and stimuli used. Each block started with the display of the instructions on how to perform the respective task. Once participants felt ready, they could start the block by pressing a button. All visually presented words were displayed in black on grey and in lowercase font. EEG was recorded throughout all blocks employing the customised MATLAB application BRAINSTREAM (http://www.brainstream.nu/) for stimulus presentation and data recording.

The *first* block was a reading-only task consisting of 50 trials (each stimulus word repeated twice) that had the purpose to familiarise participants with the speed of the (substitution) task. On each trial, a single English word appeared on the screen for 80 ms and was followed by an interstimulus interval (ITI) of 1 sec before the next trial started. Participants had to read out aloud the word presented to them while making sure to finish before the next word appeared. They took a self-paced break after half of the trials. Both during the break and at the end of the block, they received feedback on the average speed of their vocal reactions and how this supposedly related to the performance of native speakers (this comparison was not real but was used to create a feeling of time pressure). The feedback message also included encouraging messages regarding the participants' speed, such as "You are getting there, keep improving!"

The *second* block consisted of the first practice phoneme substitution task presenting, in each of its 27 trials, a single pseudoword containing either an /æ/-/ɛ/ vowel, which had to be substituted by the respective other vowel as fast as possible. For example, participants saw the pseudoword FENT and were supposed to verbally respond by saying "fant". Stimulus presentation was again 80 ms with a constant ITI of 1sec (see FIG 1 for a timeline). Feedback on the speed of verbal responses was given once at the end of the block.

In the *third* block, participants were similarly asked to substitute any /æ/-/ɛ/ vowels, this time in existing English words, by again verbally producing the respective counterpart of a given stimulus. Next to those substitution trials, there were now also catch trials, in which a word not containing either of the target vowels was presented and to which participants should respond by saying “no”. For example, participants saw the word CUP and would have to say “no”. Another difference to the previous block was the dynamic timing of trials in order to create time pressure for the individual participant. While each visual stimulus was presented for 80 ms as before, the ITI in each trial depended on a participant’s verbal response and was set to 400 ms after the automatically detected voice offset of the response (unless no response was given, which led to an ITI of 2000 ms). Feedback on participants’ speed together with an encouraging message (see above) was again given both during a break after half of the trials and after the end of the block. The total number of trials was 96 (24 word stimuli and 24 catch trials both repeated twice).

The *fourth, fifth, sixth* and *seventh* block were, like the second and third block, a pseudoword substitution task for the English phoneme contrasts /d/-/t/ and /p/-/b/ respectively. These were part of another study focusing on consonants instead (Garcia-Cossio et al., in revision) and will not be further discussed here.

The *eighth* block was a reading-only task employing the same word list used for the word substitution task on the /æ/-/ɛ/ contrast, which served as a reference of participants’ pronunciation of the relevant phonemes without time pressure and was relevant for the rating of the stimuli described below. The *ninth* and *tenth* block were similar reading-only tasks used for the consonant study.

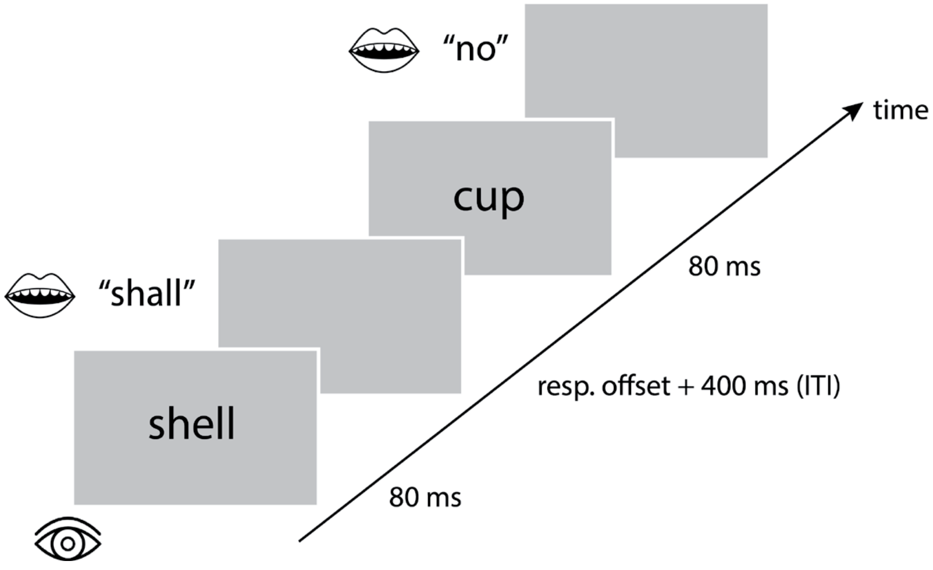

FIG 1. Timeline of the phoneme substitution task here presenting an example of a substitution trial with correct verbal response and followed by a catch trial with a correct catch response. The onset of each trial depended on the automatically detected offset of the verbal response given in the previous (with a maximum of 2 sec).

E. Error ratings

All verbal responses given during the vowel word substitution task (Block 3) were checked for errors by offline classifying each trial's response into one of five categories. To this end, seven native speakers of British English used a self-developed GUI running in MATLAB (see description below). Each dataset was evaluated by a unique combination of 3 raters resulting in about 5 hours of work for each rater, which was split up between multiple sessions to ensure a sufficient level of concentration.

Before starting to rate, but also anytime it seemed informative, raters could listen to the vowel word recordings of a given participant from the respective reading-only task in order to familiarise themselves with a participant's unique voice and way of pronouncing the stimuli, especially the /æ/ and /ɛ/ vowels. When using the rating GUI, raters could play (and re-play) a trial's response and then answer the question "What did you hear?" by selecting on of the following categories (for the example stimulus *pan*):

(1) "pan" (option showed the respective a-word of a trial's minimal pair)
(2) "pen" (option showed the respective e-word of a trial's minimal pair)
(3) "either option 1 or 2, it is difficult to determine"
(4) "no"
(5) "something other than the above"

F. Response onsets

To determine response latencies, the onset of the verbal response given in the vowel substitution task (block 3) was manually marked and extracted per trial using a self-developed GUI running in MATLAB. Trials with exceedingly fast (< 200 ms) or slow (< 1500 ms) responses were excluded from further processing (similary done by Ganushchak and Schiller, 2009), resulting in exclusions of 2.2% and 0.6% of the trials due to the two criteria respectively. Remaining response onsets were used to both compute response-locked EEG responses in the offline processing procedure described below and to compare reaction times (time between visual stimulus presentation and voice onset) between correct and error trials.

G. Electrophysiological measurements

Electroencephalography was measured using a Biosemi Active 2 system with 64 Ag/AgCl active electrodes placed on the scalp according to the International 10/20 System (BioSemi, The Netherlands). The sampling rate was 2048 Hz and impedance of the electrodes was kept below 25KΩ. Electrooculography (EOG) recordings were used to measure eye movements and blinks. For detecting vertical eye movements and blinks, two bipolar electrodes were placed just above and below the right eye, while another pair of electrodes was placed to the outer sides of the left and right eye for detection of horizontal eye movements.

H. EEG Data preprocessing & ERP analysis

All EEG recordings were analysed offline in the open source toolbox Fieldtrip (Oostenveld et al., 2011) running in MATLAB (R2014a, The Mathworks, Inc.). First, the continuous signal was segmented into stimulus-locked epochs by selecting 100 ms before and 1000 ms after stimulus onset plus an additional period of 10,000 ms on each side, which served as data padding to avoid filter artefacts in relevant parts of the epochs during later high- and low-pass filtering. After reducing the sampling rate to 512 Hz, bad channels were identified (criterion: presence of power in the 50 Hz frequencies deviating by more than 3 deviations from the average influence), and then interpolated based on neighbouring channels. Subsequently, a low-pass (0.1 Hz cut-off) and then high-pass filter (30 Hz cut-off) was applied to the data-padded epochs using a two-pass Butterworth filter (Hamming window) of order 2 and 4 respectively.

Epochs were reduced to a length of -100 ms and 1000 ms relative to stimulus onset. The identical filtered data was also used to create an additional dataset including epochs of the length -100 ms to 2500 ms which would be treated similarly in the following preprocessing steps as the stimulus-locked dataset but would eventually be time-locked to verbal responses instead. The epochs were thus larger and due to that more of them were removed from the data as there were more motor artefacts in the later parts of the epochs. In both datasets, artefacts in the signal caused by eye movements were identified and removed based on correlations with the EOG channels (Gratton, 1998). Further distortions produced by motor activity, such as speech articulation, were automatically detected (by making use of the property that EMG has relatively low power in the low frequencies compared to its total power) and were then also removed from further processing. After re-referencing the signal to the mastoids, remaining artefacts were removed by excluding all trials, which exceeded a threshold of 50 mV. The signal was then baseline corrected using a 50 ms window prior to stimulus onset and, in the case of the response-locked dataset, it was time-locked to onset of verbal responses. In both sets, data was split into correct and error trials, averaged across trials within participants for each type, and then averaged across participants to reveal grand averages. EEG recordings from the passive word oddball task were processed following the same analysis pipeline as the stimulus locked data above.

All EEG data were statistically analysed using cluster-based permutation tests, a non-parametrical testing procedure available in the Fieldtrip toolbox, which offers a straightforward solution to the multiple-comparison problem (Maris and Oostenveld, 2007). All reported outcomes were based on 1000 randomisations and the default Monte Carlo method to calculate significance probabilities. In each test, we used the entire set of electrodes in a time window that depended on the respective research question (see specified per test). In an effort to balance sufficient statistical power and the risk of false alarms between cluster-based permutation tests, we used Bonferroni corrections whenever using multiple tests for a specific comparison within a given dataset.

III. RESULTS

A. Behavioural results

Perceptual ability

Independent sample t-tests comparing the two groups (trained, control) at baseline level did not reveal any differences in d prime score (identification task), boundary sharpness or position of boundary on the /æ/- /ɛ/ morphed continuum (identification on morphed continuum; see TABLE III). There is thus no indication that the groups differed in their perceptual performance related to the target contrast before the training. The trained group did, however, significantly improve both their perception and production of the non-native contrast in the course of training (see d prime results in Thorin et al, 2018).

TABLE III. Perceptual scores for the two groups (trained, control) resulting from the identification task (d prime) and identification on morphed continuum task (boundary sharpness and position of category boundary on the continuum).

Group	d prime		boundary sharpness		category boundary	
	pre	post	pre	post	pre	post
Trained	1.99 (±0.9)	3.85 (±1.2)	2.96 (±4.4)	2.50 (±3.5)	6.22 (±0.7)	6.28 (±0.6)
Control	1.52 (±0.9)$^{\text{n.s.}}$	n.a.	1.8 (±2.1)$^{\text{n.s.}}$	n.a.	5.90 (±0.8)$^{\text{n.s.}}$	n.a.

$^{\text{n.s.}}$ non-significant outcomes of an independent sample t-test comparing the two groups.

Phoneme substitution task: Rating of responses

Three independent ratings were used to categorise each response trial of the vowel substitution task (block 3). Whenever at least two raters agreed, their rating was the "rated response" and could have the value 1-5 referring to the respective option in the rating GUI (see methods). If all three of them rated differently, the trial was labelled as "no consensus" and excluded from all further analyses.

The ratio of trials per participant's dataset for which at least 2 raters agreed differed between the groups, as a t-test comparing the percentage of agreed responses revealed (p < 0.002). Raters agreed significantly more often when rating responses produced by the training than by the control group.

Across both groups, two and three raters agreed in 31.05% (954 trials) and 60.31% (1853 trials) of the time respectively, while there was no consensus between raters in 8.63% of trials (265 trials). We used Fleiss' Kappa in order to quantify between-rater reliability. This is a measure taking into account the chance level of agreement given the number of raters and number of possible rating categories (Warrens, 2010). The outcome value can range from -1 to 1 with 0 indicating a rater agreement at chance level and 1.0 indicating perfect agreement. Rater agreement in the current response evaluation resulted in a Fleiss' Kappa of 0.61, which can be categorised as "intermediate to good" agreement above chance level.

TABLE IV. Types of errors that were rated and their description. The (*) marks error type 2, which did not occur and was thus not part of any analyses. It is listed here as theoretically possible case for the sake of completeness only.

		Rating example	
Error type	Response description	Stimulus	Rated response
Correct (type 1)	Correct substitution	HAM	“hem”
	Correct substitution	HEM	“ham”
Correct (type 2)	Correct catch response	BIG	“no”
Error (type 1)	Incorrect catch response	HAM	“no”
	Incorrect catch response	HEM	“no”
Error (type 2)*	Incorrect vowel word response	BIG	“ham”
	Incorrect vowel word response	BIG	“hem”
	Incorrect vowel word response	BIG	“Either option 1 [a-word] or 2 [e-word], difficult to determine”
Error (type 3)	Missed substitution	HAM	“ham”
	Missed substitution	HEM	“hem”
Error (type 4)	Unrelated response	HAM	“none of the above”
	Unrelated response	HEM	“none of the above”
	Unrelated response	BIG	“none of the above”
Error (type 5)	Ambiguous a/e vowel pronunciation	HAM	“Either option 1 [a-word] or 2 [e-word], difficult to determine”
	Ambiguous a/e vowel pronunciation	HEM	“Either option 1 [a-word] or 2 [e-word], difficult to determine”

Phoneme substitution task: Response classification

All responses with a consensus rating were categorised as one of two types of correct responses or one of five types of error responses (see TABLE IV for an overview and examples). Originally, we thought these different response categories would allow us to distinguish between responses which are relatively unambiguous, and can thus be easily classified as erroneous or correct response (catch trials, such as, did the participant not respond “no” even though a stimulus word did not contain one of the target vowels?), and those that are inherently difficult to classify as they are largely dependent on a participant’s pronunciation and the rating of it (for instance, did the participant correctly substitute an /æ/ for an /ɛ/?). In spite of this fine-grained classification method, however, there was not sufficient data per type to look at them separately (see TABLE V) and we hence decided to combine them into two larger categories for the behavioural and EEG analyses below: “error” versus “correct” trials.

TABLE V. Types of responses. Total trial count and ratio per type of error across participants within the trained group and control group respectively.

Groups	Correct subst.		Correct catch resp.		Incorrect catch resp.		Incorrect vowel word resp.		Missed subst.		Unrel. resp.		Ambig. vowel pron.		No consensus	
	N	%	N	%	N	%	N	%	N	%	N	%	N	%	N	%
Trained	752	52	324	23	9	0,6	0	0	127	8,8	113	7,8	50	3,5	65	4,5
Control	631	39	371	23	15	0,9	0	0	206	13	97	5,9	112	6,9	200	12

Phoneme substitution task: Error responses

Out of all the consensus-rated trials, participants in the control and training group produced erroneous responses in 22.0% (±14.5) and 30.4% (±11.8) of the time. A two-way independent t-test revealed that this difference was not significant ($p < 0.08$).

Phoneme substitution task: Response latencies

Exceedingly high (< 200 ms) and low (> 1500 ms) response latencies were excluded from further processing, which resulted in the removal of 2% and 0.6% of the data due to the two criteria respectively. A 2-factor mixed-design ANOVA on the response latencies with group as between-subject factor and correctness as within-subject factor revealed a significant main effect of both correctness ($F(1,30) = 13.20$; $p = 0.001$) and group ($F(1,30) = 4.45$; $p = 0.043$), but no group x correctness interaction ($p > 0.05$). Verbal responses were overall 65 ms faster in the trained than in the control group, while erroneous responses were overall 38 ms slower than correct responses (TABLE VI).

TABLE VI. Grand average response latencies (in ms) and standard deviations (SD) in correct and erroneous trials separated for the two groups: trained and control.

Group	Correct		Error	
	ms	SD	ms	SD
Trained	644	±63	723	±82
Control	620	±79	671	±127

B. EEG results

Response-locked data (ERN)

In the following analyses, only datasets of participants were included, in which more than 5 trials were available after pre-processing (incl. artefact removal as described above) for each of the two response types, namely correct and error responses. This threshold was based on recent evidence suggesting that the minimum number of trials needed for a stable ERN is six to eight (Pontifex et al., 2010). This procedure led to 9 and 12 datasets included in the analysis of the training and control groups respectively.

Firstly, we established whether the two groups showed significant ERN effects. To this end, we used a cluster-based permutation test for dependent samples, one for each group, to compare erroneous and correct responses within groups. All channels and a relatively unrestricted time window (0-600 ms after response onset) were included. Results revealed a significant positive cluster (Bonferroni correction 0.05/2 = 0.025) for the training group (100 - 372 ms; p = 0.012) and similarly for the control group (194 – 403 ms; p = 0.025; see FIG 2). Both clusters span across nearly the entire coverage of channels[7].

Secondly, an additional cluster-based permutation test for independent samples helped us in investigating whether the two groups differed in their ERN response. The test compared the difference curves between erroneous and correct responses between the two groups, again involving the entire set of channels and a relatively unrestricted time window (0-600 ms after response onset). It revealed no significant clusters (p > 0.05).

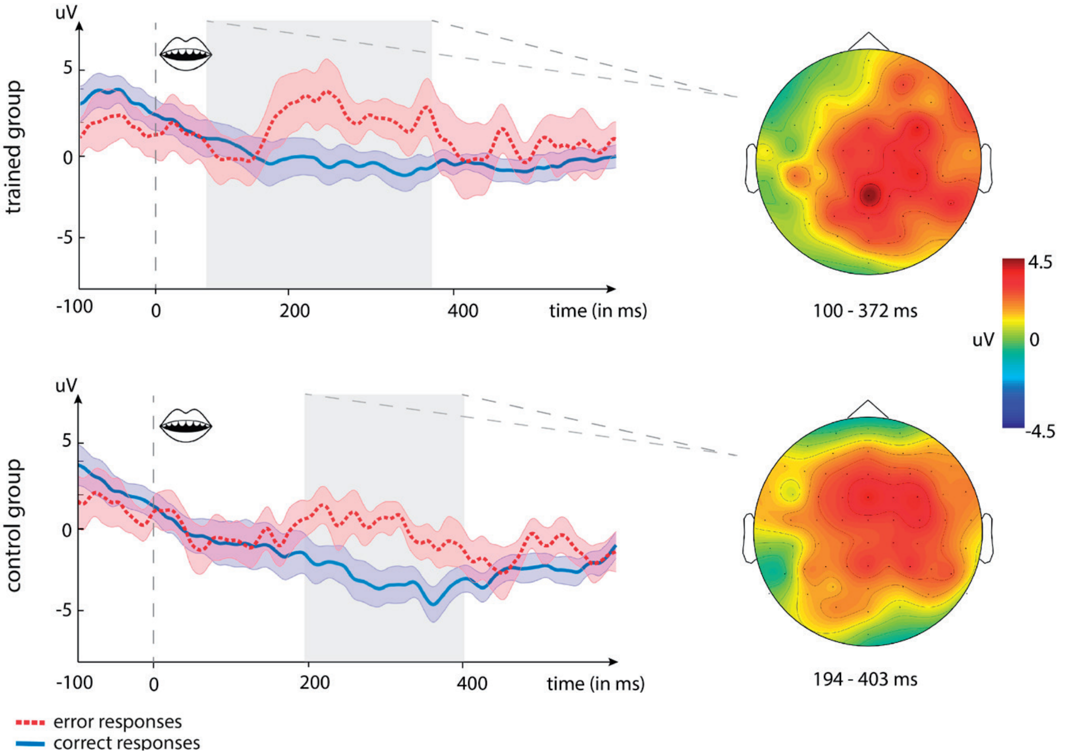

FIG 2. Response-locked analysis. [Left] Grand average ERP responses with zero indicating the onset of the verbal responses for the trained group (top) and control group (bottom) contrasting trials with correct (blue) and error responses (red). The signals are averages across electrodes Fz, FCz and Cz with shaded areas indicating standard error across individual participants' responses. Significant clusters in the comparison between correct and error responses are highlighted in grey. **[Right]** Topographic maps averaged in time across the significant clusters.

[7] Trained group [50 out of 64 channels]: AF3, F1, F3, FC3, FC1, C1, C3, CP5, CP3, CP1, P1, P3, P5, PO3, O1, Oz, POz, Pz, CPz, Fp2, AF8, AF4, AFz, Fz, F2, F4, F6, F8, FT8, FC6, FC4, FC2, FCz, Cz, C2, C4, C6, T8, TP8, CP6, CP4, CP2, P2, P4, P6, P8, P10, PO8, PO4, O2; Control group [54 out of 64 channels]: AF7, F1, F3, F5, F7, FT7, FC5, FC3, FC1, C1, C3, C5, T7, TP7, CP3, CP1, P1, P3, P5, PO7, PO3, O1, Iz, Oz, POz, Pz, CPz, AF8, AF4, AFz, Fz, F2, F4, F6, F8, FT8, FC6, FC4, FC2, FCz, Cz, C2, C4, C6, T8, CP6, CP4, CP2, P2, P4, P6, P8, PO8, PO4, O2

Stimulus-locked data

In addition to the response-locked data, we also performed a control analysis on stimulus-locked data in order to check for the presence of typical N1 effects in response to the visual stimulus and if this was present to an equivalent extent in both groups. Again, only datasets of participants were included in the following analyses, in which more than 5 trials per correct and error type of response could be included in the average per participant. In the case of the stimulus-locked data this lead to 13 and 14 participant's datasets included in the analyses for the trained and control groups respectively. Note that the relevant time window for the stimulus-locked data (0-600 ms) lies before most of the verbal responses, which meant that the data was less influenced by motor artefacts than the response-locked data and thus more trials (and thus participant's datasets) could be included.

Similar statistical tests as presented above for the response-locked ERPs revealed no differences between error and correct responses for either of the two groups (p > 0.025), nor between the two groups regarding their response patterns in the two conditions (p > 0.05, see FIG 3). Both groups, however, showed a typical N1 response-locked to stimulus onset (see FIG 3).

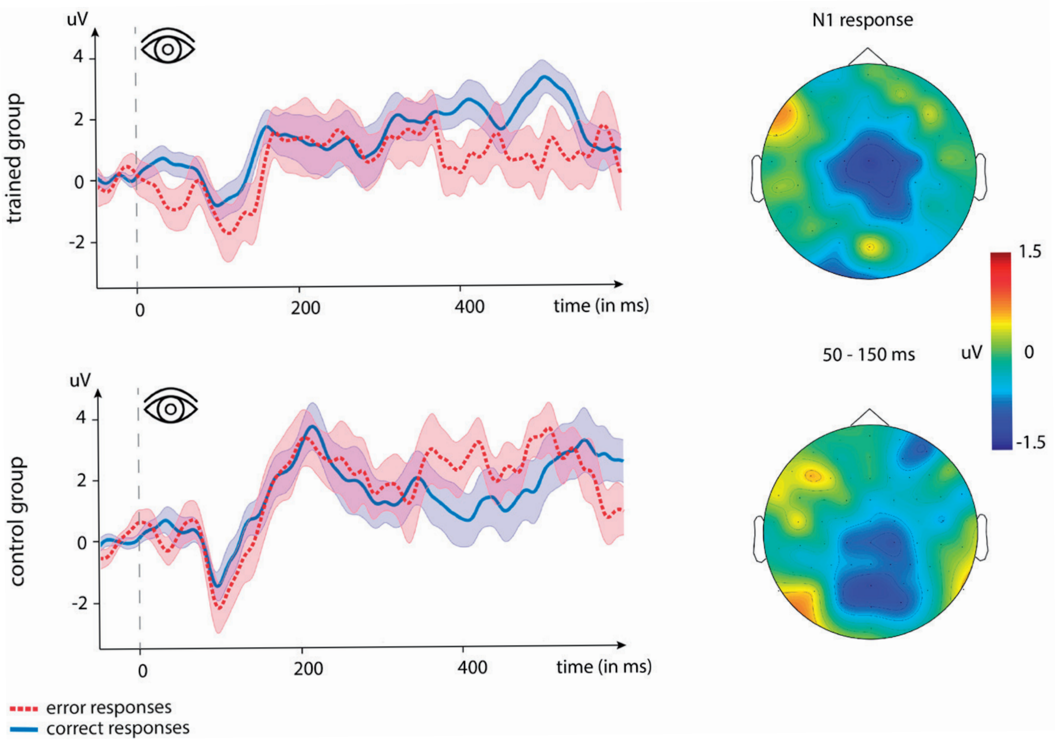

FIG 3. Stimulus-locked analysis. [Left] Grand average ERP responses with zero indicating the onset of stimulus presentation for the trained group (top) and control group (bottom) contrasting trials with correct (blue) and error responses (red). The signals are averages across electrodes Fz, FCz and Cz with shaded areas indicating standard error across individual participants' responses. **[Right]**. Topographic maps of the typical N1 response (50 – 150 ms) obtained by subtracting the average correct response curve from average error response curve.

IV. DISCUSSION

The present study focusses on verbal self-monitoring in the context of second language speech production. Dutch natives who were previously trained on the perception and production of a challenging speech contrast, the English /æ/ and /ɛ/ vowels, engaged in a phoneme substitution task involving the trained categories. Results from electrophysiological measurements of both the trained group and an untrained control group revealed no differential neural responses following erroneous and correct responses in the typically observed latency for ERN effects for either of the two groups. Those findings have to be considered in the light of relevant behavioural measurements showing that the trained group was significantly better at perceiving the non-native vowels than the untrained control group and had also evidently improved their production of the novel sounds in the course of the preceding training (Thorin et al., 2018). This difference arose despite well-matched performance (no difference in identification nor discrimination ability) between the trained and the control group at baseline. Taken together, relatively proficient non-native speech perception and production of the novel sound categories did here not suffice in triggering a detectable ERN effect related to errors made while producing the challenging novel phonemes.

Despite the absence of any differences between correct and erroneous neural responses in the ERN time window, there was a later difference effect arising between the two response types. This difference resulted from a positive-going response related to erroneous responses, which peaked around 200-400 ms after verbal response onset and showed a centro-parietal topography (see FIG 2). This effect, however, also did not differ between the two experimental groups. Both concerning its timing and topography, this response can be related to earlier observations of a “slow wave” response, which has been reported in the context of speech error monitoring following an ERN response (Falkenstein et al., 1991; Masaki et al., 2001). It was earlier interpreted as more conscious or thorough evaluation of a self-detected error and related to subsequent response adjustment.

Stimulus-locked ERP data showed a typical posterior N1 response that did not differ between groups. A visual N1 effect is typically triggered by an attended stimulus and has been shown to be enlarged whenever related to a discriminatory process (e.g. Vogel and Luck, 2000). This explanation seems valid in the context of the current phoneme substitution task, during which participants had to attend the centre of a screen in order to respond to the word stimulus appearing in this location at the beginning of each trial. The discriminatory process here can be thought of as the participants’ classification of a given trial as catch or substitution trial. The nature of the task was identical for both groups and there is no obvious reason to assume that it would differ as a result of phoneme training, which is what the data indeed indicate.

Evaluation of the rating data confirmed that the three native English raters overall showed an intermediate to good rating agreement, which could be used as reliable rating outcome whenever at least two of them agreed (91.36% of verbal responses). Interestingly, raters agreed significantly more often when rating responses produced by the training than by the control group. The explanation here seems straightforward. The trained group evidently produced clearer differences between the two vowels in their pronunciation (see above) and it can thus be expected that it is easier for raters to recognise which vowel was produced leading to higher conformity across different raters. The difference between the rater agreements for the two groups separately does not pose an issue for the interpretation of the current results given, firstly, that the overall inter-rater reliability is still reasonably high and, secondly, because we are focussing on *percentages* of erroneous and correct responses out of all trials that were clearly rated (at least two raters agreeing) in the later steps instead of total numbers. The differential effect of rating agreement can thus be seen as additional evidence that the trained group has improved their production ability of the novel non-native contrast.

The two groups did not differ in their ratio of erroneous responses. This seems intuitive as the nature of the phoneme substitution task primarily involved executive functioning by relying on a fast discrimination between catch and substitution trials followed by rapid verbal response execution. It did only indirectly depend on the production of the challenging non-native vowel contrast. Regarding the number of errors made overall, it can be noted, however, that generally more errors were made by L2 speech users in the present study (22.0% and 30.4% for the trained and control group respectively) as compared to participants in the previously mentioned L1 study employing the same phoneme substitution paradigm by Trewartha and Philips (2013), in which the error rate was about 10%. This is accordance with previous findings by Poulisse (1999, 2000, see above) reporting more speech errors made during L2 than during L1 speech production.

Verbal responses were found to be overall faster in the trained than in the control group. Although it is reasonable to expect that the preceding perceptual training would not influence the rate of errors made (see above), it is a likely explanation that the degree of fluency with the non-native phoneme contrast did also influence the fluency of producing those target words in the present experimental task. In other words, participants who had repeatedly produced words containing one of the target vowels during the 4-day perceptual training, could more quickly respond by producing substitution words in the present study than control participants.

Returning to the electrophysiological data and main focus of the present study, there could be different reasons for the null findings in the typical ERN window. Investigating verbal self-monitoring of difficult-to-produce non-native phonemes in the present context comes with the inherent difficulty of differentiating between a participant's intention and their potential bad pronunciation. For example, if a given response was consistently rated as "mess" even though the correct substitution would have been to respond "mass", is the reason for this (now rated as) erroneous response an actual error (namely a missed substitution) or is the outcome of the rating due to unclear pronunciation of the participant? He or she might have intended to correctly say “mass” but was unable to produce the /æ/ phoneme correctly. With the current task design, we had hoped to being able to differentiate between those two categories for each type of error and for correct responses (see again TABLE IV). But given an insufficient amount of trials in each category, we were not able to directly compare them and therefore had to collapse the subcategories (see Results section). Despite the above mentioned difficulty, however, we did find the “slow wave” response clearly differentiating between correct and erroneous responses, which indicates that the overall classification of error and correct trials was sufficiently good.

Another reason for not finding a difference between the groups could appear to lie in the nature of the task. A possible argument could be that participants’ fast substitution of vowels and the (automatic) evaluation of erroneous responses was experienced more like a cognitive game rather than a natural process of verbal self-monitoring. Evidence against this stance, however, is that the use of the same substitution task did lead to the observation of ERN effects in response to erroneous verbal responses in the study by Trewartha and Philips (2013). That means that if processing of this substitution task did not reflect verbal error-monitoring as opposed to more general task monitoring, we would still or even especially then expect to see ERN responses (for both groups). The fact that an ERN effect could not be detected in either of the two groups in the present experiment and that seemed thus independent of increased perceptual ability to differentiate (and self-evaluate) the to-be-substituted vowels, actually speaks in favour of the notion that the task relied on verbal error-monitoring processes. It is important to keep in mind, however, that this tentative interpretation is based on null results. Stronger conclusions could be drawn based on an additional comparison with a group of English native speakers engaging in the presently used phoneme substitution task. This comparison is currently already indirectly available by considering the results by Trewartha and Philips (2013), showing that substitution errors in the same task (with different stimuli) led to ERN responses during L1 processing.

Together with the findings by Trewartha and Philips (2013), the present results could suggest that improved levels of perception in the training group were not sufficient to enhance the self-monitoring system to a degree that would produce ERN responses, which are typically observed in native speakers. The newly learnt categories might still be too weakly established as to enable efficient verbal self-evaluation. This could be explained by considering the Perceptual Loop Theory (Levelt, 1983). According to this central account on speech monitoring, language users rely on the same mechanisms for evaluating their own speech as when listening to speech of others. In other words, verbal error monitoring is primarily perception-based. In the present context, this would mean that in order to engage in efficient L2 speech monitoring, all necessary non-native phoneme categories have to be fully established. Future research could clarify if this cautious interpretation holds and if more extensive training of non-native phonemes and/ or more time for consolidation would be needed in order to reach native-like levels of verbal error monitoring.

Although this interpretation of the current findings has to be considered with caution as it is based on null findings, it would be in line with findings from L2 perception of challenging non-native phonemes (Sebastián-Gallés et al., 2006). Here, Spanish-Catalan early bilinguals who were dominant in Spanish did not show an ERN in response to their own misperceptions of words containing a difficult-to-perceive, Catalan-specific vowel contrast. This was the case despite their high levels of proficiency in Catalan and although similar Spanish-Catalan bilinguals who were dominant in Catalan did show the expected ERN response.

Less efficient error monitoring during (some aspects of) L2 use could also explain why speakers make overall more errors in L2 than in L1 (see again Poulisse, 1999, 2000). This might be the case while other linguistic aspects can be successfully monitored during L2 use, such as in previously mentioned examples of an ERN in response to L2 word-gender violations (Bultena et al., 2017) or in response to self-produced verbal errors in a fast-paced L2 phoneme monitoring task (Ganushchak and Schiller, 2009).

In sum, the present study investigated verbal self-monitoring of newly-learnt non-native phonemes. Typical electrophysiological signatures of error-monitoring, namely ERN effects, did not emerge in a previously trained group and neither in an untrained control group. Reasons for the absent group difference could be (1) insufficient numbers of errors in the critical category, (2) inherent problems with coding this kind of L2 pronunciation errors. Overall, there was no evidence that verbal self-monitoring of non-native vowels differed as a function of whether they have been trained or not. These null findings could potentially suggest that newly-learnt phonemic categories are insufficient to create native-like patterns of (electrophysiological) error monitoring, but further investigations are needed to verify this tentative interpretation.

V. ACKNOWLEDGMENTS

This study would not have been possible without Eliana Garcia-Cossio who was involved in the design and data collection of the study and developed part of the experimental tasks. Her research contribution as part of a postdoc position was supported by the Dutch Technology Foundation STW (STW11161), which is part of the Netherlands Organization for Scientific Research (NWO), and partly funded by Ministry of Economic Affairs. We would like to thank Philip van den Broek for his technical support, as well as Malin Mangersnes, Mariia Naumovets, Inez Wijnands, and Josh Ring for their help with data collection.

6

Chapter 6

General Discussion

I. SUMMARY OF FINDINGS

The aim of this dissertation was to further our understanding of how speech perception and speech production interact in the course of learning novel phonemic categories. More concretely, it was examined how this learning process in one of the speech modalities would transfer to similar improvements in the other one and if second language learners could benefit from combined training methods involving both modalities. It was also tested to what extent the verbal self-monitoring system could adapt to newly-learnt non-native elements and thereby support second language speech acquisition. To this end, a variety of methods was employed including two multi-day training paradigms as well as the analysis of behavioural, speech and electrophysiological measurements. All experiments in this dissertation are based on a population of Dutch native speakers with intermediate/ high levels of English proficiency and use the British English /æ/ and /ɛ/ vowels. This non-native phonemic contrast is known to be challenging for the chosen population in both perceiving it accurately and producing it distinctively (Broersma, 2002; Wanrooij et al., 2014; Weber and Cutler, 2004).

Chapter 2 focusses on shedding light on the additional effect of production practice of this difficult non-native contrast in the context of perceptual training. This was done by investigating to what extent both non-native perceptual and production learning were influenced by a 4-day perceptual training scheme that was intertwined with either related or unrelated production practice of the targeted /æ/-/ɛ/ vowel contrast. In each trial, this training scheme involved a categorisation of an auditorily presented word, which was followed by visual feedback and then the production of a visually-prompted, single English word that either included one of the target vowels, for instance the English word "pan" (related production group), or a word of similar length and structure not including any of the relevant vowels, such as "dog" (unrelated production group). Behavioural results in both modalities showed that learning took place over the course of training independently of whether the trained contrast was pronounced during training. In other words, perceptual learning transferred to production learning irrespective of any production practice of the relevant speech contrast. Interestingly, in **Chapter 3**, which was based on the same training study, these behavioural results were complemented with more sensitive electrophysiological measurements revealing advantageous effects of related production practice. The group of participants that had engaged in the combined training method, the related production group, exhibited a neural signature of change detection in the form of a mismatch negativity (MMN) in response to the English / pæn/-/pɛn/ contrast after training, which was absent in the unrelated production group and an untrained group. Perceptual learning thus benefitted from additional production practice during perceptual training but these positive effects could only become apparent when measurements were sufficiently sensitive to identify fine-grained differences in perceptual ability.

To further investigate the mutual relationship between the speech modalities, the reverse direction of cross-modality transfer, namely from production learning to perceptual improvements, was examined in **Chapter 4.** Here it was tested how a two-day production training protocol on the British English vowel contrast affected both the production and perception of it. After explicit pronunciation instructions, participants in the experimental group received trial-by-trial visual feedback on their single word productions in terms of how their respective vowel pronunciation related to that of a typical native speaker. The feedback consisted of a visual representation of the tongue and mouth position during articulation (based on F1 and F2 values, the first two formants) and as part of this visualisation the location of a typical native speaker's utterance together with a participant's own vowel production of a given trial. In the control group, participants engaged in a similar paradigm producing the same number of critical words, but instead of direct feedback on their own pronunciation they were merely presented with the general indication of a typical native speaker's utterance for the two vowels. Although there was no detectable effect of this trial-by-trial visual feedback, both groups improved their pronunciation over the course of training. These gains in production could be explained by an interaction of various factors, including explicit pronunciation instructions, focussed attention and motivation to improve, that might have led to more efficient internal evaluation during active practice of the challenging vowel contrast over the course of the two-day training study. Interestingly, despite no direct training in the perceptual modality, participants of both groups also improved their perception of the non-native contrast, which points towards cross-modality transfer from production to perception. Results of **Chapter 4** thereby complement those of **Chapter 2** in suggesting a bi-directional (though not necessarily balanced) relationship between the speech modalities.

Chapter 5 considered the role of verbal self-monitoring in the context of second language use and, more specifically, how easily the self-monitoring system could adapt to evaluating newly-learnt sound categories in order to support the acquisition of non-native phonemes. To this end, previously trained participants (those tested in **Chapters 2** and **3**) engaged in a fast-paced task, during which they had to verbally respond to visually presented English words by substituting the vowel (either /æ/ or /ɛ/) by its respective counterpart. This phoneme substitution task led to a substantial amount of verbal substitution errors in both the trained and an untrained control group. Despite these speech errors, however, electrophysiological measurements did not show typical indicators of error monitoring (during L1 use) in the form of an error related negativity (ERN) for either of the groups. There was thus no evidence for any differences in self-monitoring of L2 vowels as a function of whether they were previously trained or not. Though any interpretation of these null results needs to be taken with caution and further investigations are needed to verify it, this might indicate that newly-learnt phonemic categories are insufficient to create native-like patterns of (electrophysiological) error monitoring.

II. THE LINK BETWEEN PERCEPTION AND PRODUCTION

The main research question of this dissertation was how speech perception and speech production interact in the course of learning non-native sound categories and how their relationship in this process could best be described. To start with, the empirical work of this dissertation could confirm the assumption that the two speech modalities do not function in isolation of each other and is - in this regard - in line with outcomes of previous research related to this question. More specifically, the mutual workings of perception and production processes became evident in the form of cross-modality transfer shown in both possible directions. Perceptual learning resulted in production gains independently of any direct training in this modality in **Chapter 2**, whereas directed production training led to simultaneous improvements in the perception modality in **Chapter 4**. Interestingly, these results do not only illustrate an existing link between the speech modalities but also point towards a bidirectional nature of that link.

Despite this bidirectional transfer, however, we did not find a linear relationship in the form of a correlation between learning in the two modalities. Improvements in perception did not predict those in production in **Chapter 2** or **3**. This means that though the modalities must be linked to some extent, they are not proportionally dependent on each other. In other words, changes in the one modality do not necessarily go hand in hand with (similar) changes in the respective other, while there are important factors influencing their interactions. The relationship between speech perception and production is thus multifaceted as the two processes are evidently linked but the nature of their link seems to involve a complex interplay marked by dynamic changes depending on various factors. I would like to briefly address the most central factors influencing sound learning in the following section and then discuss consequences for the characterisation of perception-production interactions as well as requirements for models describing their mutual relationship.

The first critical factor that should be considered in this context is *time*. This can be thought of on different timescales. On a short-term scale (in the order of milliseconds to seconds), when training both perception and production in combination, the exact timing of the training protocol will be crucial in determining whether the interactions between the modalities are beneficial or rather hindering. To give a short example, perceptual training combined with related production practice turned out to support perceptual learning in **Chapter 3**, while combined perception-production training was earlier shown to hinder perceptual learning (Baese-Berk, 2019; Baese-Berk and Samuel, 2016). Interestingly though, this difference in outcomes could be reasonably accounted for by differences in timing of productions in the training designs used (see below for a more detailed discussion on efficient training design).

On an intermediate time scale (i.e. days to months), it is likely that the mutual relationship between the modalities is dynamically changing over time with L2 learner's increasing proficiency and/or familiarity with a non-native speech contrast. As has also been earlier proposed (Baese-Berk, 2019; Nagle, 2018), the relationship on that scale could be asymptotic or time-lagged. That would be the case, for instance, if improvements in production could only be achieved once a certain perceptual proficiency has been reached, or if perception was leading production only at later stages of category formation. The empirical studies of this dissertation provide insights exclusively focussing on intermediate/highly proficient learners of a challenging L2 contrast and can thus not describe the entire picture. But when comparing the current findings with those coming from the training of entirely novel contrasts, differences can be observed, for instance, those related to cross-modality transfer and the effectiveness of combined training approaches (Baese-Berk and Samuel, 2016; Kartushina et al., 2015). More research, however, needs to be conducted testing the factor familiarity in order to reveal any systematic patterns (Nagle, 2018).

Finally, on a long-term scale (i.e., years), we know that ageing is accompanied by a steady decrease in neural plasticity and thus in the ability to learn new speech contrasts (Flege et al., 1996, 1999b; Piske et al., 2001). Though the dynamics between the modalities are least researched in older age, it is possible that also the relationship between the modalities changes with decreasing cognitive capacity. One could speculate that L2 learner's (both active and subconscious) strategies for speech sound learning change with age. For instance, age-related reductions of hearing capacity could lead to a shift towards focusing more heavily on production learning.

Crucially when reflecting on these different timescales, however, it is important to be aware of the conceptual difference between the perception-production link during real-time cognitive processing (such as during the active process of learning in an ongoing training task) and the way the speech modalities might influence each other during more long-term linguistic development (as also noted by Sakai and Moorman, 2018). More concretely, the fact that the two speech processes might hinder each other in learning under some circumstances does not mean that their relationship will be antagonistic in the course of longer term category formation (the difference is here the one between milliseconds and days).

Other important factors when characterising the way speech perception and production influence each other in the process of learning new phonemic categories are the *type of training* method and, as already presented at the beginning of this chapter, the *mapping* between L1 and L2 phonological spaces. The type of training method refers to design decisions, such as direct or indirect engagement of both speech modalities, timing, and specifics of feedback (when, how and on which aspect of the learner's performance is it given?). All of these decisions will influence the way speech perception and production interact during L2 sound learning. Similarly, the way the to-be-learnt L2 phonemes relate to the existing sound categories of a learner's native phonological space will also have consequences for the L2 learning process (see above).

Another important factor when characterising the way speech perception and production influence each other in the process of learning new phonemic categories is that of *individual differences*. We know and reconfirmed in the two training studies conducted that learners vary considerably in their ability to discriminate and pronounce non-native phonemes. Some learners clearly benefit from targeted training, while some make minor improvements only. Especially listeners' ability to distinctively produce the English vowels was shown to differ and to result in differently shaped learning paths in **Chapter 4** (though this was not the focus of the conducted analyses). Language learners will obviously differ in many regards, among others, in their cognitive ability and age, their motivation, their aptitude for listening and for detecting fine-grained acoustical differences, their experience with (other) foreign languages, and also their talent to imitate and produce novel sounds. This could lead to a scenario in which some L2 learners are "good producers" primarily focussing on production which will then support their perception, while others are "good perceivers" who will initially focus on achieving accurate perception of a non-native phoneme before they shift their focus to producing the sound correctly. In the second case, it seems also likely that their relatively accurate percept of the L2 phoneme will be beneficial in the process of self-monitoring their speech productions (even before articulation). Though we presented an approach to investigate error-monitoring of newly-learnt phonemes in **Chapter 5**, the mechanisms underlying effective self-evaluation of novel (and challenging) L2 sounds have been scarcely researched and hence are relatively unknown. However, it seems worth the methodological challenge to focus more attention on understanding this monitoring process in future research, as this could result in some key answers to the question if and by which modality second language phoneme learning is mainly driven. Can production of a novel (and challenging) non-native phoneme only succeed once relatively accurate perception is established as this percept is a pre-requisite for efficient monitoring, which in turn is necessary for any kind of production learning? Successful production training of novel speech sounds speaks against this notion (e.g. Kartushina et al., 2015), but more direct evidence is still needed.

An additional form of individual differences comes in the form of production variability. Phoneme productions of some speakers result in more widespread distributions of phonemic realisations, while others have a more compact distribution and thus a more consistent way of pronunciation. In L1, production variability was shown to be correlated with perceptual acuity (Franken et al., 2017) and it seems likely that this is also the case in the context of L2 sound learning. It can be predicted that L2 learners with high perceptual acuity for a new contrast also show more compact productions of that contrast. Experimental approaches tailored towards a direct investigation of these links on the level of individual L2 learners could hence provide new insights on the perception-production relationship.

Taken together, relevant models describing second language phoneme acquisition need to be able to account for dynamic changes in the interaction between speech perception and production in the course of learning both during real-time neurological processing (i.e. during the execution of an experimental training task) and longer term linguistic development (i.e. category formation). In addition to this, they need to consider the critical influence of individual differences and the way a learner's L1 and L2 phonological spaces relate to each other. Previously introduced models of cross-language speech perception, the SLM (Speech Learning Model; Flege, 1995) and PAM (Perceptual Assimilation Model; Best, 1995) do so to some degree only.

Concerning the last of these aspects, both PAM and SLM predict that novel speech categories are formed based on their perceived (dis)similarity with the native phonological system. This prediction is in line with the data of this dissertation. To illustrate this, remember that, out of the two trained non-native vowels, the English /ɛ/ phoneme lies close to the existing Dutch category /ɛ/, while there is no direct counterpart to English /æ/. According to PAM, the English /æ/ is thus expected to be assimilated by the Dutch category /ɛ/. This is indeed how the production learning presented in **Chapter 2** could be interpreted. Here Dutch native speakers showed largely overlapping productions of the English /ɛ/ and /æ/ vowels at the start of the training, which were located relatively close to the native category /ɛ/. After training, participants overall showed more distinct productions of the two trained English vowels, which was due to the fact that /æ/ productions tended to be more native-like. In other words, participants started out with a single category at the beginning of the training and developed – to some degree - a novel one for the earlier assimilated vowel category as a consequence of phonemic training.

A relevant and important difference between PAM and SLM concerns the way they characterise the relationship between the perception-production link during non-native language processing. Neither of the two accounts provides a comprehensive explanation of the empirical findings presented in **Chapter 2-4**. Flege's Speech Learning Model sees a learner's perceptual ability as central to their production performance. If perceptual learning was indeed driving someone's production ability (even in experienced L2 users), we would not expect to see any transfer of learning from production training to perceptual improvement. But this effect was indeed revealed in **Chapter 4**. Contrary to the notion by SLM, Best's Perceptual Assimilation Model postulates that perceptual assimilations of novel speech sounds is driven by their articulatory gestures. If simply put, speech perception would thus be dependent on production, which could be extended to the experimental prediction that L2 perception and thus also perceptual learning is driven by the accuracy of non-native production. While this account would explain the transfer from production to perception in **Chapter 4**, it cannot explain the opposite cross-modality transfer revealed in **Chapters 2 and 3**. Taken together, neither of the two models can account for a bidirectional relationship between speech perception and production during non-native phonemic learning.

When evaluating the predictive value of both the PAM and SLM model in the context of this dissertation, however, it is important to keep in mind that neither of them was originally designed to capture non-native phoneme learning in its entirety. In fact, an often neglected difference between the two models is that PAM was originally developed to characterise the initial contact of naïve listeners to novel L2 sounds, while SLM describes production (and perception) by L2 speech learners and thus experienced listeners. Best and Tyler (Best and Tyler, 2007b) later on assessed the commonalities between these two phases and, based on this, formulated an extension of the original model, the so-called PAM-L2 (Best and Tyler, 2007b). Despite this extension, the model still cannot be expected to account for the dynamically changing learning process of L2 acquisition in its entirety.

A more recent attempt to consider the learning path of L2 phoneme acquisition as well as the influence of individual differences (and dynamic changes) is Escudero's Second Language Linguistic Perception model (L2LP model; Escudero, 2005; Mayr and Escudero, 2010). It can, in fact, be seen as a meaningful synthesis of both the SLM and PAM frameworks. It combines clear predictions on how the perception of novel phonemes is influenced by the native phonological space through assimilation and on how category formation in a developmental learning process takes place (like experienced listeners in SLM). Interestingly, it also acknowledges (some form of) individual variation in this process. According to L2LP, learners' initial encounters with the non-native sound space will not only be determined by their native language but also by their accent, including regional, social and idiosyncratic features of their L1. Though further work is still needed to experimentally test all of the model's predictions, empirical support for various aspects of it has already become available (Escudero and Boersma, 2004; Evans and Iverson, 2004) and it is, to my knowledge, the most comprehensive account of L2 sound acquisition.

Considering the above presented outcomes of the empirical work together with previous findings in the scientific field, it becomes clear that the relationship between the speech modalities cannot satisfactorily be described as linear or static. A linear or balanced relationship between the speech modalities would entail that improvements in one of the modalities predicted improvements in the respective other. As the above discussion has shown, however, there are other crucial factors dynamically influencing this mutual relationship in the course of learning. The most important of these factors are: cross-mapping between L1 and L2 phonological space, the way of training L2 perceptual and/or production ability, and individual differences, such as, motivation, age, aptitude for speech perception and/or production, and production variability.

So far I have almost exclusively discussed the relationship between the speech modalities during L2 sound *learning*. Naturally, the question on their interactions during novel sound category formation can be set into a wider context, namely the debate on how the two speech modalities more generally relate to each other during communication. The scope of this dissertation, especially in the experimental parts of it, had to be restricted. Traditionally, the two speech modalities have been studied in isolation assuming there were separate systems for the two speech processes. For good reasons researchers have begun to consider the two processes as more interrelated and thus increasingly studied their interactive nature (e.g. Baese-Berk, 2019; Broos et al., 2016; Franken, 2018). The present investigation of non-native sound learning can be seen as part of this more general development in the speech sciences. Moreover, the present consideration of the processes underlying novel category formation can be seen as a specific example of the way in which speech perception and production work together.

Throughout this thesis, I have been referring to speech perception and speech production as two *modalities*. This terminology is in compliance with common use, as perception and production have typically been considered two separate systems. The implication of this terminology (two systems) has, in fact, been fundamental to the way the main research question of this dissertation was posed (how do speech perception and production interact?). As recently proposed by McQueen and Meyer, however, speech perception and production could instead be seen as two *aspects* of a single common system (McQueen and Meyer, 2019). In their proposal they describe language as single knowledge base comprising a set of processing mechanisms that are recruited depending on the specific language task performed. Speech perception and production would thus be a set of tasks rather than separate modalities or (sub)systems. Though this characterisation of speech perception and production was not contemplated in the process of describing the mutual relationship between perception and production in this dissertation, it could provide a potentially explanatory framework for describing the processes underlying non-native speech learning. In fact, a basic prediction following from the account by McQueen and Meyer was confirmed by the results presented in this dissertation: If there is a single knowledge base, there should be bidirectional influences between perception and production to be observed (during L2 learning), which is indeed what we showed in form of a cross-modality transfer (see **Chapters 2** and **4**). The challenge for future research will be to see if the account can explain the unbalanced nature of these bidirectional influences.

II. HOW TO EFFICIENTLY TRAIN NON-NATIVE CATEGORIES?

The secondary research aim of this dissertation was concerned with how a closer understanding of the speech perception-production interactions during L2 sound learning could inform the choice and development of efficient training methods to improve such learning. When developing such a training method there are countless elements to consider and details to adjust which results in literally an infinite number of possible training protocols. The following discussion will not provide a complete overview of all possible choices but will focus on the key factors that are thought to influence training efficiency most strongly.

A prominent question in the recent debate on perception-production interactions during L2 learning has concerned the question on whether it is beneficial to train both speech modalities in combination or whether to focus on isolated training protocols. As the results of the first training study presented in **Chapter 2** and **3** have shown, it is good to combine training under certain conditions. We have shown positive effects of perceptual training combined with related production practice for a familiar contrast in intermediate/ proficient L2 learners. When training a novel contrast it might be better to concentrate training efforts on the perceptual modality first, as there are findings suggesting an antagonistic influence on learning when speech perception is directly intertwined with production practice under some experimental conditions (Baese-Berk, 2019; Baese-Berk and Samuel, 2016). However, it remains to be verified if this is also the case in a more natural speech learning setting and not mostly due to specifics of timing in laboratory training protocol (i.e. when combining speech perception with production practice in each trial, it was shown that a delay before the production task removed the earlier shown disruptive effect (Baese-Berk and Samuel, 2018; Baese-Berk & Samuel, submitted).

The task used during training is also important. For training the perception of novel sounds, (high-variability) phonetic training, typically in the form of a two-alternative choice task, has become a relatively standardised and widely employed method (see also **Chapter 2**). Especially the use of variable stimuli (e.g. multiple tokens of the same word and/or produced by different speakers) has been shown to be an efficient method in training novel sounds (e.g. Leong et al., 2018; Logan et al., 1991; Sadakata and McQueen, 2013). Concerning the targeted enhancement of L2 production, training protocols seem less standardised and there is a need for greater standardisation of methods in order to enable direct and more controlled comparisons between L2 production training approaches. Two key elements are how to induce the production of to-be-trained sounds (the task) and how to evaluate them (the feedback). The most common options for the first are the use of picture naming, imitation or a reading task. Although L2 learners were shown to be better at imitating difficult non-native phonemes as compared to producing them in a reading task, imitation turned out to not necessarily reflect an L2 learner's productive usage of a non-native phoneme (Llompart and Reinisch, 2018). This makes imitation a less favourable task for production training, while both picture naming and word reading seem valid choices in this context.

Also choosing the type of feedback in the context of production training is a crucial element of any efficient training protocol. In this dissertation, we have presented two different methods for this: no external feedback (**Chapter 2** and **3**) and direct, trial-by-trial informative feedback (**Chapter 4**). Simple practice without any external feedback (though likely internal self-evaluation) was shown to be beneficial in the context of perceptual training in **Chapter 3** indicating that external evaluation is not strictly necessary to support novel category formation. As there are multiple examples of efficient external feedback approaches (e.g. Arora et al., 2018; Hacking et al., 2017; Kartushina et al., 2015; Lie-Lahuerta, 2011; Machovikov et al., 2002; Neri et al., 2006), it can still be assumed that (some forms of) external feedback are helpful during non-native production learning. In the present dissertation, we presented a production training approach, in which participants received trial-by-trial informative feedback on the distance between their own productions to those of a typical native speaker in terms of articulatory features (as locations in the F1-F2 space that were achieved by the tongue and mouth positions see **Chapter 4**). While learners were shown to benefit from similar feedback when learning entirely novel vowel categories in a study by Kartushina et al. (2015), this feedback approach did not turn out to support production improvements in Dutch native speakers who were already fairly familiar with the trained English vowels (the factor familiarity/proficiency might indeed be the crucial difference explaining the different outcomes, but this hypothesis would have to be tested empirically first).

Though not used for the purpose of giving online feedback during production training but for offline speech evaluation instead, this dissertation also presented an automatic speech recognition (ASR) approach that could potentially be used as an online feedback tool (see **Chapter 2)**. Here, we used a Hidden Markov model trained on word recordings by native English speakers to classify the speech data produced before and after phonetic training by Dutch learners of English. By means of this, binary classifications of the minimal pair stimuli, such as English pan-pen, could effectively be used to evaluate whether a given vowel (embedded in a word) can be classified as native-like. Similarly, this method (or a similar one based on powerful new approaches, such as, deep neural networks) could be used to provide learners with immediate feedback on their speech productions, though such approaches will be relatively uninformative as to how to improve productions.

In the light of the increasing use of methods from artificial intelligence in an ever-expanding field of applications (together with the emergence of consumer-market, wearable EEG hardware), it is worthwhile to also consider the use of *neurofeedback* in the search for efficient training methods here. Neurofeedback is a type of brain-computer interface (BCI), during which a person receives feedback on his or her neural activity that is related to a specific cognitive or behavioural state. The basic idea behind the potential efficiency of neurofeedback is that feedback on neural responses is more direct, whereas feedback on behaviour involves a more indirect loop where decision processes and other aspects of cognition may intervene. In principle, any language-related response could be used for neurofeedback provided that it can be condensed to a single measure of neural activity. For EEG, we showed in **Chapter 3** that the MMN was a more sensitive measure in detecting increased perceptual sensitivity than common behavioural methods. The MMN has also previously been shown to precede behaviourally measurable perceptual improvements (Tremblay et al., 1998) and by providing a straightforward measure of a learner's ability to discriminate between a challenging non-native sound contrast, it could serve as an ideal measure for neurofeedback. On-line neurofeedback on frequency discrimination sensitivity (in the form of MMN responses) was already shown to enhance participants' MMN responses relative to those of a sham feedback group (Brandmeyer, 2014) and Chang et al. recently presented preliminary work indicating that L2 learners could improve their discrimination ability of a challenging L2 contrast, here English liquids by Japanese learners, after neurofeedback on their MMN responses (Chang et al., 2017). It seems promising (though technically challenging[8]) for future research to further investigate these possibilities. Closing the feedback loop during speech category learning in this sense would also have important implications for the understanding of mechanisms underlying the process of improving speech perception (and production).

III. CONCLUSION

In this dissertation, I presented empirical evidence aimed at furthering our understanding of the relationship between speech perception and production during second language speech category learning. The empirical work is built on the use of two multiple-day training studies investigating the conditions and outcomes for learning in both speech modalities that results from targeted perception and/or production training. Based on the discussion of these outcomes and their relation to the broader scientific context, it became evident that speech perception and production indeed mutually influence each other during novel sound learning. Their relationship is bidirectional as we could see in cross-modality transfer observed in both directions. Furthermore, L2 sound learning and the way perception and production interact in the process of learning is likely to be

[8] A reliable signal detection (from within noise) is a requirement for effective feedback but still technically challenging. For example, signal-to-noise levels can be enhanced by aggregating responses over time (several trials) thereby boosting the robustness of the detection. As a result, however, feedback presentation based on such a "sliding window" of aggregated neural responses is slowed down, which is likely to reduce the training's efficiency. Though we feel these directions are promising, and in fact were part of our original research plan, we have realised how brittle current single-trial detection approaches of these signatures in EEG still are in practice, see e.g. Dijkstra et al. (2018) on the use of N400 in the context of semantic probing BCI and other applications.

influenced by a number of critical factors, including individual differences (such as, age of acquisition, motivation to reach a native-like accent, production variability, aptitude for listening) and cross-language phonological mapping. Importantly, the relationship between the two speech modalities is likely to change more dynamically over the course of learning than previously thought. More experimental investigations designed to shed light on these (long-term) temporal dynamics are needed, as well as those on individual differences in L2 sound learning that may influence those dynamics.

Let us return to the introductory example of Nienke for an illustration of the astonishing complexity arising between speech perception and production during the process of L2 sound learning. Can Nienke hear what she cannot say? I suggest that the answer to this question is that it depends. While training to both correctly perceive and produce the English contrast, she might very well reach a level at which she could reasonably well tell apart the two sounds when listening to another speaker while not yet being able to properly pronounce them herself. But this state will likely only be temporary, since learning to perceive the sounds will also help her in producing them. The same question, however, could also refer to Nienke's ability to evaluate her own productions by being able to effectively monitor her own utterances. Can she tell if she can or cannot say it correctly? That very much depends on the mechanisms underlying the verbal self-monitor (are they perception-based, production-based or rather depending on forward modelling...?) and will have to remain an open question to be addressed by future research.

Nienke is happy that she has just read a dissertation on why other Dutch speakers struggle to properly hear and pronounce the English words "pan" and "pen" just like her. It also included a section on how they can best go about improving their language skills. The simplified message she takes home from all this scientific research on her everyday problem is the following: There is hope for her to still get better at both hearing and pronouncing the difficult English sounds. Regarding how to overcome her difficulties, she learnt that her awareness of the challenging contrast and the motivation to get better already gives her a head-start. What could further help her is to ask her English classmates to explain to her what it is they are doing with their mouth and tongue while pronouncing the two sounds. Then it is all about practicing, though both for speaking and for listening it holds that practice is even more efficient when combined with constructive feedback. Why not also asking her English classmates to tell her if and in what way her pronunciation is off? And why not using the subtitles when watching her favourite British series that can tell her right away if it was right that she heard the word "bad"?

Learning to hear and to pronounce a difficult foreign sound turns out to be much more closely connected to each other than researchers initially thought. The good news for her about this is that regardless of whether she focusses on first getting better at hearing the *difference between the difficult English sounds or rather at pronouncing them, either way this will probably mean that she gets better at the respective other task, too. So that's a win-win situation. And speaking of winning, what the researchers have gained (or won) are new findings and many more questions based on them. Nienke actually got intrigued to learn more. There is still so much more to find out about perceiving and producing speech that she started to consider whether to become a language researcher herself.*

References

Appendix 1

Adank, P., Hout, R. Van, and Smits, R. (**2004**). "An acoustic description of the vowels of Northern and Southern," Journal of the Acoustical Society of America, **116**, 1729–1738.

Arora, V., Lahiri, A., and Reetz, H. (**2018**). "Phonological feature-based speech recognition system for pronunciation training in non-native language learning," The Journal of the Acoustical Society of America, **143**, 98–108.

Atienza, M., Cantero, J. L., and Dominguez-Marin, E. (**2002**). "The time course of neural changes underlying auditory perceptual learning," Learning and Memory, **9**, 138–150.

Baese-Berk, M. M. (**2010**). *An examination of the relationship between speech perception and production* Unpublished doctoral dissertation, Northwestern University.,1–201 pages.

Baese-Berk, M. M. (**2019**). "Interactions between speech perception and production during learning of novel phonemic categories," Attention, Perception, and Psychophysics, **April**, 1–25.

Baese-Berk, M. M., and Samuel, A. G. (**2016**). "Listeners beware: Speech production may be bad for learning speech sounds," Journal of Memory and Language, **89**, 23–36.

Baese-Berk, M. M., and Samuel, A. G. (**2018**). "The Role of Timing in Perceptual Learning of Non-Native Speech Sounds," Psychonomics,.

Best, C. T. (**1995**). "A direct realist view of cross-language speech perception," Speech Perception and Linguistic Experience: Issues in Cross Language Research, Strange, Winifred, Baltimore, MD, pp. 171–204.

Best, C. T., McRoberts, G. W., and Sithole, N. M. (**1988**). "Examination of perceptual reorganization for nonnative speech contrasts: Zulu click discriminiation by English-speaking adults and infants," Journal of Experimental Psychology, **14**, 345–360.

Best, C. T., and Strange, W. (**1992**). "Effects of Phonological and Phonetic Factors on Cross- Language Perception of Approximants," Haskins laboratories Status Rport on Speech Rearch, **110**, 109–89.

Best, C. T., and Tyler, M. D. (**2007**). "Nonnative and second-language speech perception," In M. J. Munro and O. S. Bohn (Eds.), Second language speech learning: The role of language experience in speech perception and production, Benjamins, John, Amsterdam, pp. 13–34.

Best, C. T., and Tyler, M. D. (**2007**). "Nonnative and second-language speech perception: Commonalities and complementarities," In M. J. Munro and O. S. Bohn (Eds.), Second language speech learning: The role of language experience in speech perception and production, John Benjamins Publishing, Amsterdam, pp. 13–34. doi:10.1121/1.1332378

Boersma, P., and Weenink, D. (**2015**). "Praat: doing phonetics by computer.," Retrieved from http://www.praat.org/

Bomba, M. D., Choly, D., and Pang, E. W. (**2012**). "Phoneme discrimination and mismatch negativity in English and Japanese speakers," **22**, 479–483.

Bradlow, A. R., Akahane-Yamada, R., Pisoni, D. B., and Tohkura, Y. (**1999**). "Training Japanese listeners to identify English /r/and /l/: Long-term retention of learning in perception and production," Perception & Psychophysics, **61**, 977–985. doi:10.3758/BF03206911

Bradlow, A. R., Akahane-Yamada, R., Pisoni, D. B., and Tohkura, Y. (**1999**). "Training Japanese listeners to identify English /r/and /l/: Long-term retention of learning in perception and production," Perception & Psychophysics, **61**, 977–985.

Bradlow, A. R., Pisoni, D. B., Akahane-Yamada, R., and Tohkura, Y. (**1997**). "Some effects of perceptual learning on speech production," Journal of Acoustical Society of America, **101**, 2299–2310.

Brandmeyer, A. (**2014**). *Auditory perceptual learning via decoded-EEG neurofeedback: a novel paradigm* Radboud University Nijmegen, The Netherlands.

Broersma, M. (**2002**). "Comprehension of non-native speech: inaccurate phoneme processing and activation of lexical competitors," Proceedings of the 7th international confer- ence on spoken language processing, Center for Spoken Language Research, University of Colorado, Boulder, 261–264.

Broersma, M. (**2005**). *Phonetic and lexical processing in a second language* Radboud University Nijmegen, The Netherlands.

Broos, W. P. J., Duyck, W., and Hartsuiker, R. J. (**2016**). "Verbal Self-Monitoring in the Second Language," Language Learning, **66**, 132–154. doi:10.1111/lang.12189

Brunner, J., Ghosh, S. S., Hoole, P., Matthies, M., Tiede, M., and Perkell, J. S. (**2011**). "The Influence of Auditory Acuity on Acoustic Variability and the Use of Motor Equivalence During Adaptation to a Perturbation," Journal of Speech Language and Hearing Research, **54**, 727–739.

Bultena, S., Danielmeier, C., Bekkering, H., and Lemhöfer, K. (**2017**). "Electrophysiological Correlates of Error Monitoring and Feedback Processing in Second Language Learning," **11**, 1–18.

Chang, M., Iizuka, H., Kashioka, H., Naruse, Y., Furukawa, M., Ando, H., and Maeda, T. (**2017**). "Unconscious improvement in foreign language learning using mismatch negativity neurofeedback: A preliminary study," (A. J. Newman, Ed.) PLOS ONE, **12**, e0178694. doi:10.1371/journal.pone.0178694

Deterding, D. (**1997**). "The Formants of Monophthong Vowels in Standard Southern British English Pronunciation," Journal of the International Phonetic Association, **27**, 47–55.

Díaz, B., Mitterer, H., Broersma, M., Escera, C., and Sebastián-Gallés, N. (**2016**). "Variability in L2 phonemic learning originates from speech-specific capabilities: An MMN study on late bilinguals," Bilingualism, **19**, 955–970.

Dijkstra, K., Farquhar, J., and Desain, P. (**2018**). *Semantic Probing : Feasibility of using sequential probes to decode what is on a user ' s mind* 1–11 pages. Retrieved from https://www.biorxiv.org/content/10.1101/496844v1

Escudero, P. (**2005**). *Linguistic Perception and second language acqusition - explaining the attainment of optimal phonological categorization* University of Utrecht, The Netherlands, 1–362 pages.

Escudero, P., and Boersma, P. (**2004**). "L2 Speech Perception Research and Phonological Theory," Studies in Second Language Acquisition, **26**, 551–585.

Escudero, P., Boersma, P., Rauber, A. S., and 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**, 1379–1393.

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

Evans, B. G., and Iverson, P. (**2004**). "Vowel normalization for accent: An investigation of best exemplar locations in northern and southern British English sentences," The Journal of the Acoustical Society of America, **115**, 352–361. doi:10.1121/1.1635413

Falkenstein, M., Hohnsbein, J., Hoormann, J., and Blanke, L. (**1991**). "Effects of crossmodal divided attention on late ERP components II Error processing in choice reaction tasks," Electroencephalography and Clinical Neurophysiology, **78**, 447–455.

Faris, M. M., Best, C. T., and Tyler, M. D. (**2018**). "Discrimination of uncategorised non-native vowel contrasts is modulated by perceived overlap with native phonological categories," Journal of Phonetics, **70**, 1–19. doi:10.1016/j.wocn.2018.05.003

Fischer, C., Luaute, J., and Morlet, D. (**2010**). "Event-related potentials (MMN and novelty P3) in permanent vegetative or minimally conscious states [References]," Clinical Neurophysiology,.

Flege, J. E. (**1995**). "Second Language Speech Learning: Theory, Findings, and Problems," In W. Strange (Ed.), Speech Perception and Linguistic Experience: Issues in Cross-Language Research, York Press, Baltimore, MD, pp. 233–277.

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

Flege, J. E., MacKay, I. R. A., and Meador, D. (**1999**). "Native Italian speakers' perception and production of English vowels," The Journal of the Acoustical Society of America, **106**, 2973–2987.

Flege, J. E., Mackay, I. R. A., and Munro, M. J. (**1996**). "The effects of age of second language learning on the production of English vowels," Applied Psycholinguistics, **17**, 313–334.

Flege, J. E., Yeni-komshian, G. H., and Liu, S. (**1999**). "Age Constraints on Second-Language Acquisition," Journal of Memory and Language, **104**, 78–104.

Franken, M. K. (**2018**). *Listening for speaking* Radboud University Nijmegen, NL, 1–227 pages.

Franken, M. K., Acheson, D. J., McQueen, J. M., Eisner, F., and Hagoort, P. (**2017**). "Individual variability as a window on production-perception interactions in speech motor control," The Journal of the Acoustical Society of America, **142**, 2007–2018.

Fromkin, V. A. . (**1971**). "The Non-Anomalous Nature of Anomalous Utterances," Linguistic Society of America, **47**, 27–52.

Fujioka, T., Trainor, L. J., Ross, B., Kakigi, R., and Pantev, C. (**2004**). "Musical training enhances automatic encoding of melodic contour and interval structure," Journal of Cognitive Neuroscience, **16**, 1010–1021.

Ganushchak, L. Y., Christoffels, I. K., and Schiller, N. O. (**2011**). "The use of electroencephalography in language production research: A review," Frontiers in Psychology, **2**, 1–6.

Ganushchak, L. Y., and Schiller, N. O. (**2006**). "Effects of time pressure on verbal self-monitoring: An ERP study," Brain Research, **1125**, 104–115.

Ganushchak, L. Y., and Schiller, N. O. (**2009**). "Speaking one's second language under time pressure: An ERP study on verbal self-monitoring in German-Dutch bilinguals," Psychophysiology, **46**, 410–419.

Gehring, W. J., Goss, B., Coles, M. G. H., Meyer, D. E., and Donchin, E. (**1993**). "A Neural System for Error Detection and Compensation," Psychological Science, **4**, 385–390.

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

Govaart, G. H. (**2016**). *A pan is not for writing – Making and testing a tool for online feedback on vowel production of the English /ɛ/–/æ/ contrast* University of Amsterdam (UvA). Retrieved from http://www.scriptiesonline.uba.uva.nl/

Gratton, G. (**1998**). "Dealing with artifacts : The EGG contamination of the event-related brain potential," **30**, 44–53.

Grimaldi, M., Sisinni, B., Gili Fivela, B., Invitto, S., Resta, D., Alku, P., and Brattico, E. (**2014**). "Assimilation of L2 vowels to L1 phonemes governs L2 learning in adulthood: a behavioral and ERP study," Frontiers in Human Neuroscience, **8**, 1–14.

Hacking, J. F., Smith, B. L., and Johnson, E. M. (**2017**). "Utilizing electropalatography to train palatalized versus unpalatalized consonant productions by native speakers of American English learning Russian," Journal of Second Language Pronunciation, **3**, 9–33.

Hartsuiker, R. J., and Kolk, H. H. J. (**2001**). "Error Monitoring in Speech Production: A Computational Test of the Perceptual Loop Theory," Cognitive Psychology, **42**, 113–157.

Hattori, K. (**2009**). *Perception and production of English /r/-/l/ by adult Japanese speakers* University College London,1–211 pages.

Hattori, K., and Iverson, P. (**2008**). "English /r/-/l/ pronunciation training for Japanese speakers," The Journal of the Acoustical Society of America, **123**, 3327–3327.

Herd, W., Jongman, A., and Sereno, J. (**2013**). "Perceptual and production training of intervocalic /d, ɾ, r/ in American English learners of Spanish," The Journal of the Acoustical Society of America, **133**, 4247–4255.

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

Hillenbrand, J. M., Clark, M. J., and Nearey, T. M. (**2001**). "Effects of consonant environment on vowel formant patterns," The Journal of the Acoustical Society of America, **102**, 3093–3093.

Hirata, Y. (**2004**). "Computer Assisted Pronunciation Training for Native English Speakers Learning Japanese Pitch and Durational Contrasts," Computer Assisted Language Learning, **17**, 357–376.

Hirata, Y. (**2004**). "Training native English speakers to perceive Japanese length contrasts in word versus sentence contexts," The Journal of the Acoustical Society of America, **116**, 2384–2394.

Hohnsbein, J., Falkenstein, M., Hoormann, J., and Blanke, L. (**1991**). "Effects of crossmodal divided attention on late ERP components I Simple and choice reaction tasks," Electroencephalography and Clinical Neurophysiology, **78**, 438–446.

Hu, W., Mi, L., Yang, Z., Tao, S., Li, M., Wang, W., Dong, Q., et al. (**2016**). "Shifting perceptual weights in L2 vowel identification after training," PLoS ONE, **11**, 1–14.

Huensch, A., and Tremblay, A. (**2015**). "Effects of perceptual phonetic training on the perception and production of second language syllable structure," Journal of Phonetics, **52**, 105–120.

Inceoglu, S. (**2016**). "Effects of perceptual training on second language vowel perception and production," Applied Psycholinguistics, **37**, 1175–1199.

Ito, T., Tiede, M., and Ostry, D. J. (**2009**). "Somatosensory function in speech perception," Proceedings of the National Academy of Sciences, **106**, 1245–1248.

Iverson, P., Hazan, V., and Bannister, K. (**2005**). “Phonetic training with acoustic cue manipulations: A comparison of methods for teaching English /r/-/l/ to Japanese adults,” The Journal of the Acoustical Society of America, **118**, 3267–3278.

Jakoby, H., Goldstein, A., and Faust, M. (**2011**). “Electrophysiological correlates of speech perception mechanisms and individual differences in second language attainment,” Psychophysiology, **48**, 1517–31.

Jakoby, H., Goldstein, A., and Faust, M. (**2011**). “Electrophysiological correlates of speech perception mechanisms and individual differences in second language attainment,” Psychophysiology, **48**, 1517–1531.

de Jong, K., Hao, Y., and Park, H. (**2009**). “Evidence for featural units in the acquisition of speech production skills : Linguistic structure in foreign accent,” Journal of Phonetics, **37**, 357–373.

Kaan, E., Barkley, C. M., Bao, M., and Wayland, R. (**2008**). “Thai lexical tone perception in native speakers of Thai , English and Mandarin Chinese : An event-related potentials training study,” **17**, 1–17.

Kaan, E., Wayland, R., Bao, M., and Barkley, C. M. (**2007**). “Effects of native language and training on lexical tone perception : An event-related potential study,” , doi: 10.1016/j.brainres.2007.02.019. doi:10.1016/j.brainres.2007.02.019

Kartushina, N., and Frauenfelder, U. H. (**2014**). “On the effects of L2 perception and of individual differences in L1 production on L2 pronunciation,” Frontiers in psychology, **5**, 1–17.

Kartushina, N., Frauenfelder, U. H., and Golestani, N. (**2016**). “How and When Does the Second Language Influence the Production of Native Speech Sounds: A Literature Review,” Language Learning, **66**, 155–186.

Kartushina, N., Hervais-Adelman, A., Frauenfelder, U. H., and 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**, 817–832.

Kartushina, N., Hervais-Adelman, A., Frauenfelder, U. H., and Golestani, N. (**2016**). “Mutual influences between native and non-native vowels in production: Evidence from short-term visual articulatory feedback training,” Journal of Phonetics, **57**, 21–39.

Kartushina, N., and Martin, C. D. (**2019**). “Talker and Acoustic Variability in Learning to Produce Nonnative Sounds: Evidence from Articulatory Training,” Language Learning, **69**, 71–105.

Katz, W. F., and Mehta, S. (**2015**). “Visual Feedback of Tongue Movement for Novel Speech Sound Learning,” Frontiers in Human Neuroscience, **9**, 1–13.

Kawahara, H., and Morise, M. (**2011**). “Technical foundations of TANDEM-STRAIGHT, a speech analysis, modification and synthesis framework,” Sadhana - Academy Proceedings in Engineering Sciences, **36**, 713–727.

Kemmerer, D. (**2015**). *Cognitive Neuroscience of Language*, Psychology Press, New York, 1–553 pages.

Kittredge, A. K., and Dell, G. S. (**2016**). “Learning to speak by listening: Transfer of phonotactics from perception to production,” Journal of Memory and Language, **89**, 8–22.

Koelsch, S., Schröger, E., and Tervaniemi, M. (**1999**). “Superior pre-attentive auditory processing in musicians," NeuroReport, **10**, 1309–1313.

Kraljic, T., and Samuel, A. G. (**2005**). "Perceptual learning for speech: Is there a return to normal?," Cognitive Psychology, **51**, 141–178.

Kuhl, P., Williams, K., Lacerda, F., Stevens, K., and Lindblom, B. (**1992**). "Linguistic experience alters phonetic perception in infants by 6 months of age," Science, **255**, 606–608.

Lambacher, S. G., Martens, W. L., Kakehi, K., Marasinghe, C. A., and Molholt, G. (**2005**). "The effects of identification training on the identification and production of American English vowels by native speakers of Japanese," Applied Psycholinguistics, **26**, 227–247.

Lambacher, S. G., Martens, W. L., Kakehi, K., Marasinghe, C. A., and Molholt, G. (**2005**). "The effects of identification training on the identification and production of American English vowels by native speakers of Japanese," Applied Psycholinguistics, **26**, 227–247.

Lametti, X. D. R., Rochet-Capellan, A., Neufeld, E., Shiller, D. M., and Ostry, D. J. (**2014**). "Plasticity in the Human Speech Motor System Drives Changes in Speech Perception," The Journal of Neuroscience, **34**, 10339–10346.

Lee, A. H., and Lyster, R. (**2017**). "Can corrective feedback on second language speech perception errors affect production accuracy?," Applied Psycholinguistics, **38**, 371–393.

Lemhöfer, K., and Broersma, M. (**2012**). "Introducing LexTALE : A quick and valid Lexical Test for Advanced Learners of English," Behavioral research, **44**, 325–343.

Leong, C. X. R., Price, J. M., Pitchford, N. J., and Van Heuven, W. J. B. (**2018**). "High variability phonetic training in adaptive adverse conditions is rapid, effective, and sustained," PLoS ONE,.

Levelt, W. J. M. (**1983**). "Monitoring and self-repair in speech," Cognition, **14**, 41–104.

Liberman, A. M., Harris, K. S., Kinney, J. A., and Lane, H. (**1961**). "The discrimination of relative onset-time of the components of certain speech and nonspeech patterns," Journal of Experimental Psychology, **61**, 379–388.

Lie-Lahuerta, C. (**2011**). "Fix Your Vowels: Computer-assisted training by Dutch learners of Spanish," Tijdschrift voor Skandinavistiek, **32**, 69–88.

Lively, S. E., Logan, J. S., and Pisoni, D. B. (**1993**). "Training Japanese listeners to identify English /r/ and /l/ II: The role of phonetic environment and talker variability in learning new perceptual categories," Journal of Acoustical Society of America, **94**, 1242–1255.

Llompart, M., and Reinisch, E. (**2018**). "Imitation in a Second Language Relies on Phonological Categories but Does Not Reflect the Productive Usage of Difficult Sound Contrasts," Language and Speech, , doi: 10.1177/0023830918803978. doi:10.1177/0023830918803978

Lobanov, B. M. (**1971**). "Classification of Russian Vowels Spoken by Different Speakers," The Journal of the Acoustical Society of America, **49**, 606–608.

Logan, J. S., Lively, S. E., and Pisoni, D. B. (**1991**). "Training Japanese listeners to identify English /r/ and /l/: A first report," Journal of Acoustical Society of America, **89**, 874–886.

Lopez-Soto, T., and Kewley-Port, D. (**2009**). "Relation of perception training to production of codas in English as a Second Language," Proceedings of Meetings on Acoustics Volume, 1–15.

Lu, S., Wayland, R., and Kaan, E. (**2015**). "Effects of production training and perception training on lexical tone perception - A behavioral and ERP study," Brain Research, **1624**, 28–44.

Machovikov, A., Stolyarov, K., Chernov, M., Sinclair, I., and Machovikova, I. (**2002**). "Computer-Based Training System for Russian Word Pronunciation," Computer Assisted Language Learning, **15**, 201–214.

Macmillan, N. A., and Creelman, C. D. (**1991**). *Detection theory: A user's guide.*, Cambridge University Press, Cambridge, UK.

Maris, E., and Oostenveld, R. (**2007**). "Nonparametric statistical testing of EEG- and MEG-data," Journal of neuroscience methods, **164**, 177–90.

Masaki, H., Tanaka, H., Takasawa, N., and Yamazaki, K. (**2001**). "Error-related brain potentials elicited by vocal errors," NeuroReport, **12**, 1851–1855.

Mattys, S. L., Barden, K., and Samuel, A. G. (**2014**). "Extrinsic cognitive load impairs low-level speech perception," Psychonomic Bulletin and Review, **21**, 748–754.

Mattys, S. L., and Wiget, L. (**2011**). "Effects of cognitive load on speech recognition," Journal of Memory and Language, **65**, 145–160.

Mayr, R., and Escudero, P. (**2010**). "Explaining individual variation in L2 perception: Rounded vowels in English learners of German," Bilingualism, **13**, 279–297. doi:10.1017/S1366728909990022

McQueen, J. M., and Meyer, A. S. (**2019**). "Towards a Comprehensive Cognitive Architecture for Language Use," In P. Hagoort (Ed.), Human language: From genes and brain to behavior, MIT Press, Cambridge, UK.

Menning, H., Roberts, L. E., Ca, C. P., and S, Ä. (**2000**). "Plastic changes in the auditory cortex induced by intensive frequency discrimination training," **11**, 817–822.

Miltner, W. H. R., Lemke, U., Weiss, T., Holroyd, C., Scheffers, M. K., and Coles, M. G. H. (**2003**). "Implementation of error-processing in the human anterior cingulate cortex: A source analysis of the magnetic equivalent of the error-related negativity," Biological Psychology, **64**, 157–166.

Mitterer, H., and Mattys, S. L. (**2017**). "How does cognitive load influence speech perception? An encoding hypothesis," Attention, Perception, and Psychophysics, **79**, 344–351.

Näätänen, R., Lehtokoski, A., Lennes, M., Cheour, M., Huotilainen, M., Iivonen, A., Vainio, M., et al. (**1997**). "Language-specific phoneme representations revealed by electric and magnetic brain responses," Nature, **385**, 432–4.

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

Näätänen, R., Schröger, E., Karakas, S., Tervaniemi, M., and Paavilainen, P. (**1993**). "Development of a memory trace for a complex sound in the human brain.pdf," Neuroreport, **4**, 503–506.

Nagle, C. L. (**2018**). "Examining the Temporal Structure of the Perception–Production Link in Second Language Acquisition: A Longitudinal Study," Language Learning, **68**, 234–270. doi:10.1111/lang.12275

Neri, A., Cucchiarini, C., and Strik, H. (**2006**). "ASR-based corrective feedback on pronunciation: does it really work?," Proceedings of Interspeech, Pittsburgh, USA, 1982-1985).

Newman, R. S. (**2003**). "Using links between speech perception and speech production to evaluate different acoustic metrics: A preliminary report," The Journal of the Acoustical Society of America, **113**, 2850–2860.

Nozari, N., Dell, G. S., and Schwartz, M. F. (**2011**). "Is comprehension necessary for error detection? A conflict-based account of monitoring in speech production," Cognitive Psychology, **63**, 1–33. doi:10.1016/j.cogpsych.2011.05.001

Oostenveld, R., Fries, P., Maris, E., and Schoffelen, J. (**2011**). "FieldTrip : Open Source Software for Advanced Analysis of MEG, EEG and Invasive Electrophysiological Data," Computational Intelligence and Neuroscience, **2011**, 1–9.

Peirce, J. W. (**2007**). "PsychoPy-Psychophysics software in Python," Journal of Neuroscience Methods, **162**, 8–13.

Peltola, M. S., Kujala, T., Tuomainen, J., and Ek, M. (**2003**). "Native and foreign vowel discrimination as indexed by the mismatch negativity (MMN) response," Neuroscience letters, **352**, 25–28.

Peltola, M. S., Kuntola, M., Tamminen, H., and Heikki, H. (**2005**). "Early exposure to non-native language alters preattentive vowel discrimination," **388**, 121–125.

Perrachione, T. K., Lee, J., Ha, L. Y. Y., and Wong, P. C. M. (**2011**). "Learning a novel phonological contrast depends on interactions between individual differences and training paradigm design," The Journal of the Acoustical Society of America, **130**, 461.

Piske, T., MacKay, I. R. A., and Flege, J. E. (**2001**). "Factors affecting degree of foreign accent in an L2: a review," Journal of Phonetics, **29**, 191–215. doi:10.1006/jpho.2001.0134

Polich, J. (**2007**). "Updating P300: an integrative theory of P3a and P3b," Clinical neurophysiology : official journal of the International Federation of Clinical Neurophysiology, **118**, 2128–48.

Pontifex, M. B., Scudder, M. R., Brown, M. L., O'Leary, K. C., Wu, C. T., Themanson, J. R., and Hillman, C. H. (**2010**). "On the number of trials necessary for stabilization of error-related brain activity across the life span," Psychophysiology, **47**, 767–773.

Poulisse, N. (**1999**). *Slips of the tongue: Speech errors in first and second language production*, John Benjamins Publishing, 20th ed.

Poulisse, N. (**2000**). "Slips of the Tongue in First and Second Language Production," Studia Linguistica, **54**, 136–149.

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

Pulvermüller, F., and Shtyrov, Y. (**2006**). "Language outside the focus of attention: the mismatch negativity as a tool for studying higher cognitive processes," Progress in neurobiology, **79**, 49–71. doi:10.1016/j.pneurobio.2006.04.004

Qian, M., Chukharev-Hudilainen, E., and Levis, J. (**2018**). "A system for adaptive high variability segmental perceptual training: implementation, effectiveness, and transfer," Language Learning & Technology, **22**, 69–96.

Rato, A. (**2014**). "Effect of Perceptual Training on the Identification of English Vowels by Native Speakers of European Portugese," Proceedings of the International Symposium on the Acquisition of Second

Language Speech Concordia Working Papers in Applied Linguistics, 529–547.

Rivera-Gaxiola, M., Csibra, G., Johnson, M. H., and Karmiloff-Smith, A. (**2000**). “Electrophysiological correlates of cross-linguistic speech perception in native English speakers,” Behavioural Brain Research, **111**, 13–23.

Sadakata, M., and 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**, 1324–1335.

Sakai, M., and Moorman, C. (**2018**). “Can perception training improve the production of second language phonemes? A meta-analytic review of 25 years of perception training research,” Applied Psycholinguistics, **39**, 187–224.

Schmitz, J., Díaz, B., Fernández Rubio, K., and Sebastián-Gallés, N. (**2018**). “Exploring the relationship between speech perception and production across phonological processes, language familiarity, and sensory modalities,” Language, Cognition and Neuroscience, **33**, 527–546.

Sculthorpe, L. D., Ouellet, D. R., and Campbell, K. B. (**2009**). “MMN elicitation during natural sleep to violations of an auditory pattern,” Brain research, **1290**, 52–62. doi:10.1016/j.brainres.2009.06.013

Sebastián-Gallés, N., Rodríguez-Fornells, A., De Diego-Balaguer, R., and Díaz, B. (**2006**). “First- and second-language phonological representations in the mental lexicon,” Journal of Cognitive Neuroscience, **18**, 1277–1291. doi:10.1162/jocn.2006.18.8.1277

Shinohara, Y., and Iverson, P. (**2018**). “High variability identification and discrimination training for Japanese speakers learning English /r/–/l/,” Journal of Phonetics, **66**, 242–251.

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

Skipper, J. I., Devlin, J. T., and Lametti, D. R. (**2017**). “The hearing ear is always found close to the speaking tongue: Review of the role of the motor system in speech perception,” Brain and Language, **164**, 77–105.

Tamminen, H., Peltola, M. S., Kujala, T., and Näätänen, R. (**2015**). “Phonetic training and non-native speech perception - New memory traces evolve in just three days as indexed by the mismatch negativity (MMN) and behavioural measures,” International Journal of Psychophysiology, **97**, 23–29.

Thorin, J., Garcia-Cossio, E., Sadakata, M., Desain, P., and McQueen, J. M. (n.d.). “Perception-production interactions in non-native sound learning: EEG evidence.,”

Thorin, J., Sadakata, M., Desain, P., and McQueen, J. M. (**2018**). “Perception and production in interaction during non-native speech category learning,” The Journal of the Acoustical Society of America, **144**, 92–103.

Tremblay, K., Kraus, N., and McGee, T. (**1998**). “The time course of auditory perceptual learning: neurophysiological changes during speech-sound training,” Neuroreport, **9**, 3557–3560.

Trewartha, K. M., and Phillips, N. A. (**2013**). “Detecting self-produced speech errors before and after articulation : an ERP investigation,” Frontiers in Human Neuroscience, **7**, 1–12.

Vogel, E. K., and Luck, S. J. (**2000**). “The visual N1 component as an index of a discrimination process Request Permissions : Click here The visual N1 component as an index of a discrimination process,” Psychophysiology, **37**, 190–203.

Wang, X., and Munro, M. J. (**2004**). “Computer-based training for learning English vowel contrasts,” System, **32**, 539–552.

Wang, Y., Jongman, A., and Sereno, J. A. (**2003**). “Acoustic and perceptual evaluation of Mandarin tone productions before and after perceptual training,” The Journal of the Acoustical Society of America, **113**, 1033–1043.

Wanrooij, K., Boersma, P., and Zuijen, T. L. Van (**2014**). “Distributional Vowel Training Is Less Effective for Adults than for Infants A Study Using the Mismatch Response,” PLoS ONE, **9**, 1–12.

Warrens, M. J. (**2010**). “Inequalities between multi-rater kappas,” Advances in Data Analysis and Classification, **4**, 271–286.

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

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

Wong, J. W. S. (**2013**). “The Effects of Perceptual and / or Productive Training on the Perception and Production of English Vowels / ɪ / and / i : / by Cantonese ESL learners,” Conference: 14th Annual Conference of the International Speech Communication Association (Interspeech 2013), 1–12.

Ylinen, S., Uther, M., Latvala, A., Vepsäläinen, S., Iverson, P., Akahane-Yamada, R., and Näätänen, R. (**2010**). “Training the Brain to Weight Speech Cues Differently: A Study of Finnish Second-language Users of English,” Journal of Cognitive Neuroscience, **22**, 1319–1332.

Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X. A., Moore, G., et al. (**2009**). “The HTK Book.,”

Zhang, Y., Kuhl, P. K., Imada, T., Iverson, P., Pruitt, J., Stevens, E. B., Kawakatsu, M., et al. (**2009**). “Neural signatures of phonetic learning in adulthood: A magnetoencephalography study,” NeuroImage, **46**, 226–240.

Zheng, X., Roelofs, A., Farquhar, J., and Lemho, K. (**2018**). “Monitoring of language selection errors in switching : Not all about conflict,” PLOS ONE, **13**, 1–20.

Nederlandse samenvatting

Appendix 2

Tijdens het leren van een tweede taal (L2) stuiten veel mensen op moeilijkheden met het correct uitspreken en onderscheiden van bepaalde klanken in hun gekozen L2. Neem bijvoorbeeld Nederlanders die Engels leren. Zij hebben er vaak moeite mee om het verschil te horen tussen de klinkers in de Engelse woorden *pan* en *pen*. De reden hiervoor is dat de Nederlandse taal maar één klinker heeft, namelijk /ɛ/ zoals in het Nederlandse woord *pen*, waar het Engels er twee heeft, /ɛ/ en /æ/ zoals in het eerder voorbeeld. Dit leidt ertoe dat zowel spraakperceptie als spraakproductie van dit soort klanken uitdagend is voor veel Nederlanders. Eerder onderzoek heeft laten zien dat het leren van nieuwe klankcategorieën op volwassen leeftijd nog wel (gedeeltelijk) mogelijk is, maar hoe het proces van leren precies verloopt is nog onduidelijk. Het doel van dit proefschrift was om beter te begrijpen hoe spraakperceptie en spraakproductie op elkaar inwerken gedurende het leren van nieuwe fonemische categorieën. Er is onderzocht in hoeverre dit leerproces in één van de spraakmodaliteiten zou overgaan op vergelijkbare verbeteringen in de andere, en of lerenden van een tweede taal zouden kunnen profiteren van gecombineerde trainingsmethoden die beide modaliteiten omvatten. Er is ook getest in hoeverre het verbale zelfbewakingssysteem (*verbal self-monitor*) zich kon aanpassen aan nieuw aangeleerde *niet-native* elementen en daarmee de spraakverwerving in de tweede taal kon ondersteunen. Hiertoe werd een verscheidenheid aan methoden gebruikt, waaronder twee meerdaagse trainingsparadigma's en de analyse van gedrags-, spraak- en elektrofysiologische metingen. Alle experimenten in dit proefschrift zijn gebaseerd op een populatie van Nederlanders met een gemiddeld/hoog niveau van het Engels en het Brits-Engelse klinker contrast /æ/-/ɛ/ (zie voorbeeld boven).

In **hoofdstuk 2** en **3** staat de vraag centraal welke gevolgen doelgericht trainen van perceptie (wel of niet gecombineerd met het uitspreken van de nieuwe klanken) heeft voor het waarnemen en produceren van de uitdagende Engelse klinkers. De resultaten toonden succesvol leren aan in allebei de modaliteiten als gevolg van perceptietraining, onafhankelijk of taalproductie deel uitmaakte van de training. Er was dus sprake van een transfer van perceptie naar productie. Deze gedragsresultaten werden aangevuld door meer gevoelige elektrofysiologische metingen die gunstige effecten onthulde van perceptie training die gecombineerd werd met productie (**hoofdstuk 3**). Proefpersonen die tijdens de training niet alleen maar naar de Engelse klinkers moesten luisteren maar deze ook uitspraken (in plaats van ongerelateerde woorden) toonden na de vier training sessies een zogenaamd *mismatch negativity effect* (MMN), een elektrofysiologische signatuur die aanduidt dat ze het verschil tussen de twee Engelse klinkers op een automatisch manier konden waarnemen. In de groep die tijdens de training ongerelateerde woorden moest uitspreken was dit effect achteraf niet aanwezig (hoewel ook deze groep een leereffect in de gedragstaken liet zien). Perceptueel leren profiteerde dus van gerelateerde spraakproductie tijdens de perceptuele training, maar deze positieve effecten konden alleen geconstateerd worden wanneer metingen voldoende gevoelig waren om fijnkorrelige verschillen in perceptueel vermogen te identificeren.

Om de wederzijdse relatie tussen de spraakmodaliteiten verder te onderzoeken, werd in **hoofdstuk 4** de omgekeerde richting van transmodaliteitsoverdracht (*cross-modality transfer*) onderzocht, namelijk van productieleren naar perceptuele verbeteringen. Hier werd getest hoe een tweedaags productie-trainingsprotocol op het Brits-Engels klinkercontrast zowel de productie als de perceptie ervan beïnvloedde. Na expliciete uitspraakinstructies ontvingen deelnemers in de experimentele groep *trial-by-trial* visuele feedback op hun woord producties. De feedback bestond uit een visuele weergave van de tong- en mondpositie tijdens articulatie (gebaseerd op F1- en F2-waarden, de eerste twee formanten) en als onderdeel van deze visualisatie de locatie van de uitspraak van een typische native speaker samen met de klinkerproductie van het proefpersoon zelf. In de controlegroep waren proefpersonen met een soortgelijk taak bezig. Ze produceerden hetzelfde aantal kritische woorden, maar in plaats van directe feedback op hun eigen uitspraak, ontvingen ze alleen de algemene visuele indicatie van hoe uitspraak van een typische moedertaalspreker eruit zou zien (ook gebaseerd op F1- en F2-waarden). Hoewel er geen effect van de visuele feedback in de experimentele groep aangetoond kon worden, verbeterden beide groepen hun uitspraak tijdens de training. Deze productiewinst kan worden verklaard door een interactie van verschillende factoren, waaronder de expliciete uitspraakinstructies, gerichte aandacht en motivatie, die tijdens de actieve oefening van het uitdagende klinkercontrast in de loop van de twee sessies tot efficiëntere interne evaluatie zou kunnen hebben geleid. Interessant is dat, ondanks er geen directe training was in de perceptuele modaliteit, proefpersonen van beide groepen ook hun perceptie van het Engelse contrast verbeterden. Dit wijst op transmodaliteitsoverdracht van productie naar perceptie. De resultaten van **hoofdstuk 4** vullen daarmee die van **hoofdstuk 2** aan door een bi-directionele (hoewel niet noodzakelijkerwijs gebalanceerde) relatie tussen de spraakmodaliteiten te suggereren.

Hoofdstuk 5 beschouwde de rol van verbale zelfbewaking (*self-monitoring*) in de context van het gebruik van een tweede taal en, meer specifiek, hoe gemakkelijk het verbale zelfbewakingssysteem zich kan aanpassen aan de evaluatie van nieuw aangeleerde geluidscategorieën om de verwerving van *niet-native* fonemen te ondersteunen. Daartoe namen eerder getrainde deelnemers (in **hoofdstuk 2** en **3** getest) deel aan een experiment met een foneemvervangingstaak (*phoneme substitution task*). Dit is een snelle taak, waarbij proefpersonen mondeling moesten reageren op visueel gepresenteerde Engelse woorden door de klinker te vervangen (ofwel /æ/ of /ɛ/) met zijn respectieve tegenhanger. Deze taak leidde tot een aanzienlijk aantal verbale substitutiefouten in zowel de getrainde als de ongetrainde controlegroep. Ondanks deze spraakfouten vertoonden elektrofysiologische metingen echter geen typische indicatoren van foutmonitoring (aanwezig tijdens L1-gebruik) in de vorm van een *error-related negativity* (ERN), voor geen van beide groepen. Er was dus geen bewijs voor verschillen in zelf-monitoring van L2-klinkers als functie van eerdere foneemtraining. Hoewel elke interpretatie van deze nulresultaten met voorzichtigheid moet worden beschouwd en verder onderzoek nodig is om deze te verifiëren, kan dit erop wijzen dat nieuw aangeleerde fonemische categorieën onvoldoende zijn om native-achtige patronen van (elektrofysiologische) foutmonitoring te creëren.

In **hoofdstuk 6**, de algemene discussie, werden de consequenties van deze resultaten besproken voor een kenschetsing van de link tussen spraakperceptie en spraakproductie tijdens het verwerven van nieuwe foneem categorieën en welke factoren moeten worden benaderd tijdens het ontwikkelen van geschikte trainingsmethoden. Bovendien komen verschillende theoretische modellen aan bod die relevante voorspellingen doen in deze context. Bij elkaar genomen moeten deze modellen rekening kunnen houden met dynamische veranderingen in de interactie tussen spraakperceptie en productie tijdens het leren, zowel tijdens real-time neurologische verwerking (i.e., tijdens de uitvoering van een experimentele trainingstaak) en taalontwikkeling op langere termijn (i.e., categorievorming). Daarnaast moeten ze de kritische invloed kunnen overwegen van individuele verschillen (zoals motivatie, leeftijd, aanleg voor spraakperceptie en/of productie en productievariabiliteit) en de manier waarop de native en *niet-native* fonologische ruimtes van een leerling zich tot elkaar verhouden.

Acknowledgments

Appendix 3

It's been countless times during my PhD journey, especially in those last months close to the finish line, that I caught myself imagining this very moment. Me sitting down behind my computer screen, taking a deep breath (ideally sitting outside) and commencing to write this section: the acknowledgments of my otherwise finished dissertation. *How was that going to feel?*

The answer is: relieved, somewhat proud, and, most of all, grateful. Going through the ups and downs of running a 4-year PhD project and turning the resulting (null) findings into a cohesive scientific story was not always easy (to understate it). The reason that it turned out to be a mostly rewarding, instructive and positive experience is due to the great support I am lucky to have received throughout it - and for that I want to say, *thank you*, both loudly and quietly.

James, Peter and **Makiko**, I feel lucky to have had the wonderful combination of the three of you as my supervision team benefiting from the way you are complimenting each other seamlessly. **Makiko**, you seem to carry an endless source of positivity inside of you. A positivity that floats out of you so naturally and effortlessly that makes it unavoidably contagious. You have been a personal and professional inspiration for me. At work it was a pleasure to experience you both as my daily supervisor and in the role of supervising students together with you. Thank you for your openness, warmth and laughter, for your cleverness, passion and your friendship.

Peter, you're a fountain of ideas! It is hard to imagine a meeting with you not involving a novel perspective, an alternative view or entirely new research idea. But besides your creativity and the healthy push to make me think further than what lies directly in front of me, I have also appreciated your honesty and genuine interest in me as a person.

James, you have been the stable and reliable anchor for me throughout this project and I am not even sure I would have managed to finish it without your support. You have the great ability to see the bigger picture of a project without ever becoming too distanced from it. Thank you for being there for me whenever I needed your input, advice or encouragement, for your consistently quick and detailed feedback on my work and for your faith in me.

All three of you were extraordinary supportive when I was in personal struggle and I would like to express my special thanks for your empathy and warmth.

Karen, Ceci, Eliana and **Marjolein** – you wonderfully smart and whole-hearted colleagues! It was a pleasure to share office space(s) and many good hours of discussing, laughing, concentrating, chatting and eating, learning, teaching and even crying with you guys. I am extremely grateful for having had you guys on my side during (parts of) my project and for having learnt a lot from each of you inspiring beings. **Karen** – thank you for transforming my little mantra "Failure is the mother of success" into a beautiful calligraphic artwork I could put onto my wall ready to regularly remind me of believing in the fruits that would eventually come out of any struggle. You have a razor-sharp mind that has helped me get to the other side of many questions, a kind and loyal spirit and I am grateful for being able to call you my friend. **Ceci**, thank you for your warmth and authenticity, for your smile in the morning, your delicious, baked goods and for always being the first to commit to an invitation. **Eliana**, you are one in a million and you will always be my inspiration for staying true to your own gestures and expressions. May your curls stay bouncy and shiny. **Marjolein**, you might always be my role model for structured work and for producing comprehensible code. Thank you for many inspirations both programming- and sewing-related.

Marpessa and **Andrea,** what a great time we had together as DCC PhD representatives. Thank you ladies for many good meetings, laughter and reflections.

Thanks for many interesting discussions with the members of the **BCI team** and for countless helpful comments from my colleagues in the **Sound Learning** meeting. **Philip**, thank you for your technical support especially during the early stages of my first training study. Thank you, **Lieve** and **Jolanda,** for being both the coolest and sweetest secretaries. Thank you for good chats and your support with all the bureaucratic procedures a university can come up with. Thank you also to all my fellow **IMPRS** and **LiI PhD** colleagues for a good team spirit, advice, inspiration and *gezelligheid*.

I was also blessed with the support of a number of clever and productive assistants and intern students helping me with many hours of data collection. Thank you to **Malin, Inez, Josh, Dennis, Joyce, Mariia, Laura** and **Jessie** for your great work. Thank you also to **Gisela** for your hardworking and ambitious spirit when it came to developing the production feedback tool we used for the second training study. I am sure, you are going to rock your own PhD trajectory.

Next to the support I have received at work, I am also grateful for dear friendships that have accompanied me during (and way beyond) my years as a PhD candidate. Thank you, **Anni, Mira, Katja, Emma, Melanie, Michael**, **Nina** and **Christian**, for being part of my journeys and making me feel rich. I'd like to express my special thanks to you, **Lena**, who might just be the most patient listener and clearest mirror a person can wish for and without whom I might have given up some time along many ways I have walked on.

Mama und **Papa**, auch wenn ihr schon lange keine gemeinsamen Wege mehr geht, habt ihr gemeinsam die Basis dafür gelegt, was ich heute sein und noch immer werden kann. Ihr habt mir beigebracht auf mein Gewissen und meinen Verstand zu hören, andere Menschen zu respektieren und von ihnen zu lernen, ihnen aber nicht gedankenlos zu folgen, sondern Dingen selbst auf den Grund zu gehen. Genau dafür bin ich euch sehr dankbar. Danke auch dir, **Rike**, dafür dass du mich früh und stetig in meinen Interessen gefördert hast, mich mit Lesestoff und immer mal wieder mit einer neuen Perspektive versorgt hast. **Marianne** og **Per**, også I har været med fra starten af min PhD og jeg er taknemmelig for jeres støtte som kom i mange former. Tak for mange gode samtaler, for jeres oprigtige interesse og for jeres hjerterum.

Thomas, du hast von Anfang bis Ende dieser Doktorarbeitsjahre an meiner Seite gestanden und ich möchte dir von Herzen für deine Geduld, deinen Zuspruch, deine kritische Außenperspektive und stetigen Halt danken. Mit dir bin ich dort, wo ich wachsen kann. **Frida**, du wundervoller kleiner Mensch, auch wenn du dies noch nicht selbst lesen kannst, so möchte ich dir dafür danken, dass du mir in Zeiten des akademischen Stresses gezeigt hast, dass sich das Leben um weit mehr dreht als statistische Resultate oder erfolgreiche Publikationen. Ein Blick in deine dunklen, vor Wunder und Lebenskraft lachenden Augen genügt um völlig im Hier und Jetzt anzukommen.

About the author

Jana Thorin was born in Leverkusen, Germany, on the 28th of September 1988 (previous family name Krutwig). She received both her Bachelor's degree in Psychology and her Master's degree in Cognitive Neuroscience from the Radboud University Nijmegen. Motivated by her broad scientific interests, she participated in the Interdisciplinary Honours Programme engaging in a diverse set of courses on quantum physics, literature, water management, and philosophy. Next to her studies, Jana also worked as student assistant in the Neurobiology of Language group led by Prof. Peter Hagoort at the Max Planck Institute for Psycholinguistics and as research assistant supporting the EU-funded STAGES project on Structural Transformation To Achieve Gender Equality in Science. For her Master's thesis entitled "Neural timecourse of language-attention interactions during spoken word processing" she joined the lab of Prof. Yury Shtyrov in Cambridge, UK, and (after the group's relocation) in Aarhus, Denmark. She returned to the Donders Institute for her PhD project investigating the interactions between speech perception and production during second language sound learning under the supervision of Prof. James McQueen, Prof. Peter Desain, and Makiko Sadakata.

Following her interest to gain experience in industry, Jana is currently working as Audiological Systems Developer at the R&D department of Oticon, a hearing aid company based in Copenhagen, Denmark. She is happily married and became proud mother of Frida Thorin in 2017.

Donders Graduate School for Cognitive Neuroscience

For a successful research Institute, it is vital to train the next generation of young scientists. To achieve this goal, the Donders Institute for Brain, Cognition and Behaviour established the Donders Graduate School for Cognitive Neuroscience (DGCN), which was officially recognised as a national graduate school in 2009. The Graduate School covers training at both Master's and PhD level and provides an excellent educational context fully aligned with the research programme of the Donders Institute.

The school successfully attracts highly talented national and international students in biology, physics, psycholinguistics, psychology, behavioral science, medicine and related disciplines. Selective admission and assessment centers guarantee the enrolment of the best and most motivated students.

The DGCN tracks the career of PhD graduates carefully. More than 50% of PhD alumni show a continuation in academia with postdoc positions at top institutes worldwide, e.g. Stanford University, University of Oxford, University of Cambridge, UCL London, MPI Leipzig, Hanyang University in South Korea, NTNU Norway, University of Illinois, North Western University, Northeastern University in Boston, ETH Zürich, University of Vienna etc.. Positions outside academia spread among the following sectors: specialists in a medical environment, mainly in genetics, geriatrics, psychiatry and neurology. Specialists in a psychological environment, e.g. as specialist in neuropsychology, psychological diagnostics or therapy. Positions in higher education as coordinators or lecturers. A smaller percentage enters business as research consultants, analysts or head of research and development. Fewer graduates stay in a research environment as lab coordinators, technical support or policy advisors. Upcoming possibilities are positions in the IT sector and management position in pharmaceutical industry. In general, the PhDs graduates almost invariably continue with high-quality positions that play an important role in our knowledge economy.

For more information on the DGCN as well as past and upcoming defenses please visit: http://www.ru.nl/donders/graduate-school/phd/