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

Jana Thorin, Makiko Sadakata, Peter Desain, and James M. McQueen

Citation: The Journal of the Acoustical Society of America **144**, 92 (2018); doi: 10.1121/1.5044415

View online: https://doi.org/10.1121/1.5044415

View Table of Contents: http://asa.scitation.org/toc/jas/144/1

Published by the Acoustical Society of America

Perception and production in interaction during non-native speech category learning[a)]

Jana Thorin,[b)] Makiko Sadakata,[c)] Peter Desain, and James M. McQueen[d)]

Donders Institute for Brain, Cognition and Behaviour, Centre for Cognition, Radboud University Nijmegen, Montessorilaan 3, 6500 HE Nijmegen, The Netherlands

(Received 27 July 2017; revised 6 June 2018; accepted 9 June 2018; published online 6 July 2018)

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 four-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. © *2018 Acoustical Society of America*.

https://doi.org/10.1121/1.5044415

[TCB] Pages: 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 *et al.*, 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, 2007).

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.*, 1999). 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.*, 1999), 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.*, 2005),

[a)] Preliminary results related to this study were presented during the "Society of Neurobiology of Language 2016" in London, UK, and "Psycholinguistics in Flanders," Leuven, Belgium, 2017.

[b)] Electronic mail: j.thorin@donders.ru.nl

[c)] Also at: Institute for Logic, Language and Computation, University of Amsterdam, Amsterdam, The Netherlands.

[d)] Also at: Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands.

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, U.S. 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, 2004), 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 two-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 and 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.* (2015), 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 investigate the effect of related production practice in a four-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 I). 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.

1. 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*).

2. 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.

3. 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

TABLE I. Factors matched during group assignment. n.s. indicates non-significant results of independent sample t-test comparing groups.

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)$^{n.s.}$

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

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 h 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 × 30 cm; 1920 × 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 (∼25 dB). 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-min 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.

D. Experimental tasks

1. 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.

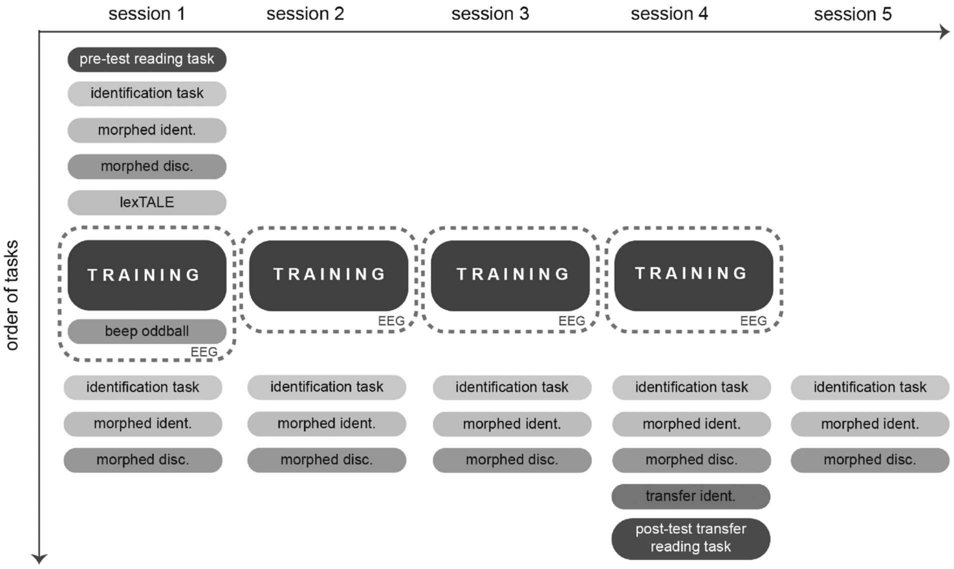

FIG. 1. Schematic timeline of the 5-day training paradigm consisting of several perceptual and production tasks conducted once prior to the full training and four times directly after a training session (post-test I-IV), as well as a delayed post-test and one set of transfer tests. Only type of training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

Before the training, participants were given verbal instructions and a 5-min practice task with unrelated stimuli (i.e., *bout-but*, *heat-height*). A full training session took approximately 50 min 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).

2. 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 × 4 speakers × 3 repetitions) and lasted about 5 min. The total score of correct answers was presented to participants afterwards.

3. 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 × 10 repetitions) took about 4 min to complete.

4. 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 × 2 orders × 2 deviant positions × 3 repetitions). The task took about 7 min to complete.

5. Reading tasks

Two versions of a reading task were employed: one pre-test version containing all ten English training words and one post-test version, completed after the last training session, containing eight 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 × 3 repetitions) or 54 trials (18 words × 3 repetitions) for the two versions, respectively. Both versions were self-paced and took about 3–5 min.

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: $d' = Z(\text{hit rate}) - Z(\text{false alarm rate})$ 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: $c = -0.5 \times Z(\text{hit rate}) + Z(\text{false alarm rate})$ (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.

1. 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: $F(5, 175) = 24.96$, $p < 0.001$, $\text{eta}^2 = 0.42$], but no differences between the two groups for this change in time [interaction group × 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: $F(5, 175) = 3.70$, $p < 0.05$, $\text{eta}^2 = 0.10$] and with no difference between the two groups [interaction group × 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).

2. 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 [$F(3,108) = 7.33$, $p < 0.001$, $\text{eta}^2 = 0.17$] which again did not differ between groups [$F(1,108) = 0.12$, $p > 0.05$].

3. 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

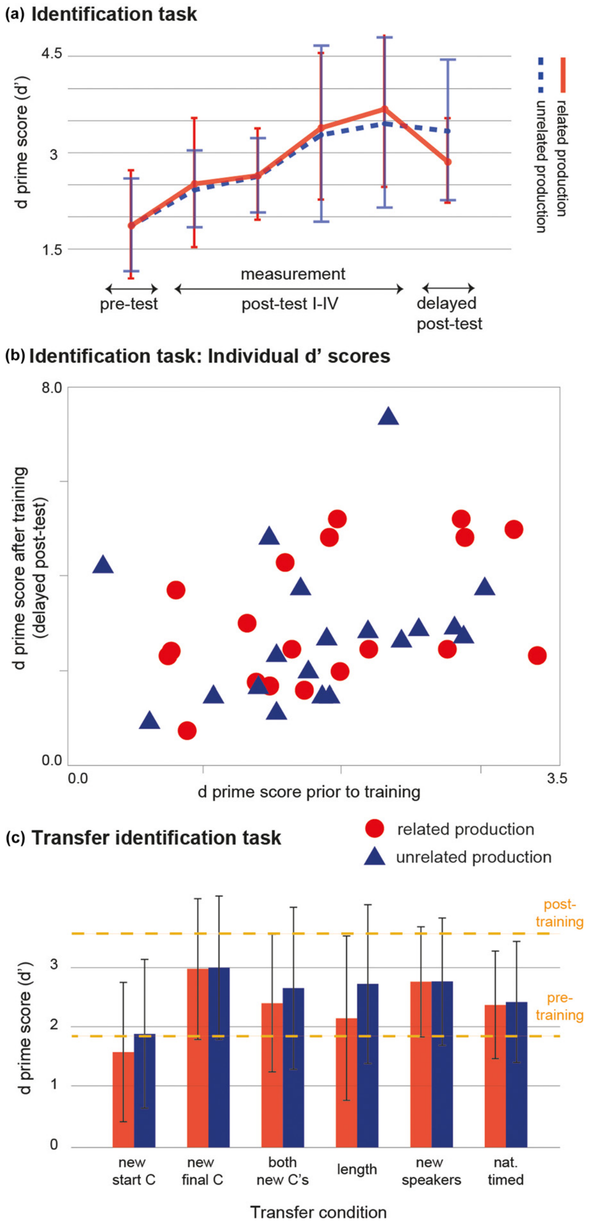

FIG. 2. (Color online) (a) Group average d′ scores of the pre-test, post-test I-IV, and delayed post-test measurements for the two training groups: related production versus unrelated production. Error bars indicate standard deviations across participants in given group. (b) d′ scores of the individual 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.

significantly lower than post-training identification scores. The two groups did not differ in any of these effects.

4. Identification on morphed continuum

In order to quantify both sharpness and position of the category boundary on the 11-step /vεt-væt/-continuum, we performed sigmoidal curve fitting using MATLAB on the number of classifications per stimulus [see Fig. 3(a)]. 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 × 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. 3(a)].

5. Discrimination on morphed continuum

A three-way repeated measures ANOVA with the within-participant factors stimulus contrast pair (eight levels) and time (six levels), and the between-participant factor group (two 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. 3(b)].

B. Production results

The speech data were analysed in two complementary ways, first by extracting and analysing the formant and duration patterns of the produced vowels and second 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.

1. 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

TABLE II. Summary of statistical results regarding the transfer of identification ability. n.s. indicates non-significant results of repeated measures ANOVA.

	Post versus transfer				Pre versus transfer			
	Time		Interaction time × group		Time		Interaction time × 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.

between the centre of one vowel 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 × 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 [$F(1,29) = 27.227$, $p < 0.001$]. 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: $F(1,29) = 2.41$, $p > 0.05$].

2. 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: $F(1,29) = 9.523$, $p < 0.01$, $\text{eta}^2 = 0.25$], though with a relatively small effect size. There was no difference between the groups regarding this effect [interaction group × time: $F(1,29) = 0.115$, $p > 0.05$].

3. ASR

In the second approach to analyse the production data, we employed an 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 × 10 stimulus words × apprimately 10 tokens = approximately 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 fivefold 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. 4(c)]. Results of a three-way repeated measures ANOVA employing the factors time (2 levels) × vowel (2 levels) × group (2 levels), showed significant main effects of measurement time [$F(1,29) = 21.89$, $p < 0.001$, $\text{eta}^2 = 0.43$] and vowel [$F(1,29) = 84.70$, $p < 0.001$, $\text{eta}^2 = 0.75$], as well as an interaction effect for time and vowel [$F(1,29) = 27.89$, $p < 0.001$, $\text{eta}^2 = 0.49$]. 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 ($p < 0.001$).

C. 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

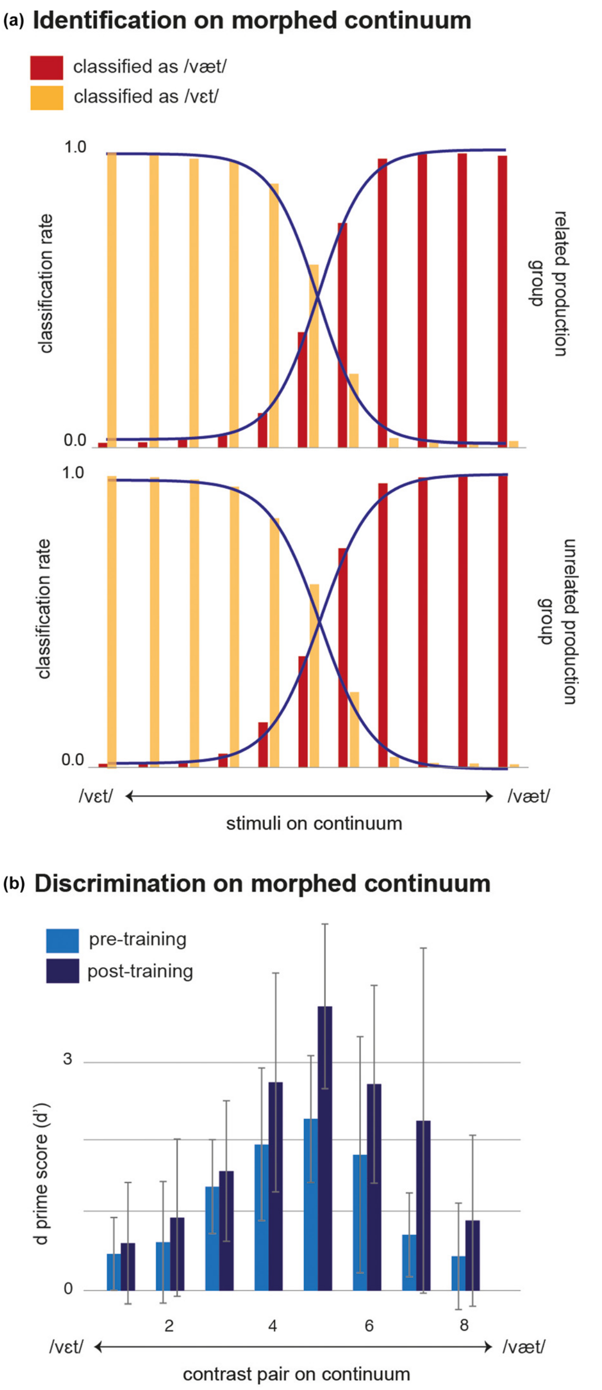

FIG. 3. (Color online) (a) Grand average percentage correct identifications on the /vɛt/-/væt/ continuum for the two training groups separately (top: related production group, bottom: unrelated production group). Sigmoidal curve fitting of the classifications are indicated as lines. (b) Grand average d′ score of discrimination between stimuli of eight 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.

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 five 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 and Kewley-Port, 2009; Wang *et al.*, 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; de Jong *et al.*, 2009; Huensch and Tremblay, 2015) and could be interpreted as the absence of a direct link between the two systems. This interpretation would resonate well with the notion of Flege (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

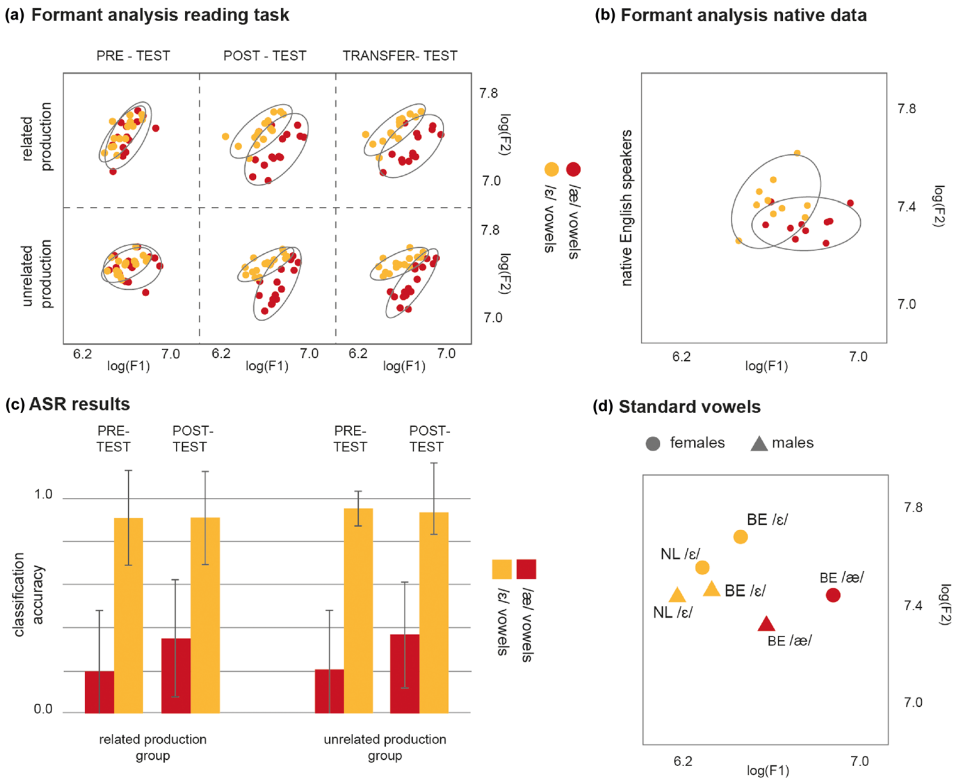

FIG. 4. (Color online) (a) log(F1) and log(F2) values for the two English vowels pronounced in CVC words of pre-, post-, and transfer reading tasks (from left to right column) after either related (top) or unrelated production training (bottom). (b) Log formant data of the two vowels pronounced by 10 British English speakers. (c) Classification results of the two English vowels by the ASR 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, /ɛ/ and /æ/, and the similar Dutch (NL) vowel category /ɛ/ [based on Deterding (1997) and Adank *et al.* (2004), respectively].

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.* (2013) and Lu *et al.* (2015)] 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 and 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. 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. First, 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 and 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 and 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 and 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.* (2013). 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 two-session paradigms on consecutive days (Baese-Berk, 2010) or days separated by 48 h (Baese-Berk and Samuel, 2016) to six 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 four-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. First, 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. 3(b), 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. 4(a)]. 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. 4(d). 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.*, 2009). 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.

ACKNOWLEDGMENTS

This research was supported by a Ph.D. grant of the Language in Interaction Consortium (https://www.languageininteraction.nl/) to J.T. Special thanks go to Philip van den Broek for his technical support, Matthias Franken for his work on the ASR 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.

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

Adank, P., Hout, R. Van, and Smits, R. (**2004**). "An acoustic description of the vowels of Northern and Southern," J. Acoust. Soc. Am. **116**(3), 1729–1738.

Baese-Berk, M. M. (**2010**). "An examination of the relationship between speech perception and production," Ph.D. dissertation, Northwestern University, Evanston, IL.

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

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

Best, C. T. and Tyler, M. D. (**2007**). "Nonnative and second-language speech perception," in *Second Language Speech Learning: The Role of Language Experience in Speech Perception and Production*, edited by M. J. Munro and O. S. Bohn (John Benjamins, Amsterdam), pp. 13–34.

Boersma, P., and Weenink, D. (**2015**). "Praat: Doing phonetics by computer," http://www.praat.org/ (Last viewed July 15, 2017).

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," Percept. Psychophys. **61**(5), 977–985.

Bradlow, A. R., Pisoni, D. B., Akahane-Yamada, R., and Tohkura, Y. (**1997**). "Some effects of perceptual learning on speech production," J. Acoust. Soc. Am. **101**(4), 2299–2310.

Broersma, M. (**2002**). "Comprehension of non-native speech: Inaccurate phoneme processing and activation of lexical competitors," in *Proceedings of the 7th International Conference on Spoken Language Processing* (Center for Spoken Language Research, University of Colorado, Boulder), pp. 261–264.

Broersma, M. (**2005**). *Phonetic and Lexical Processing in a Second Language* (Radboud University Press, Nijmegen, The Netherlands).

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," J. Speech Lang. Hear. Res. **54**(3), 727–739.

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," J. Phon. **37**, 357–373.

Deterding, D. (**1997**). "The Formants of monophthong vowels in standard southern British English pronunciation," J. Int. Phon. Assoc. **27**, 47–55.

Escudero, P., Hayes-Harb, R., and Mitterer, H. (**2008**). "Novel second-language words and asymmetric lexical access," J. Phon. **36**(2), 345–360.

Flege, J. E. (**1995**). "Second language speech learning: Theory, findings, and problems," in *Speech Perception and Linguistic Experience: Issues in Cross-Language Research*, edited by W. Strange (York Press, Timonium, MD), pp. 239.

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

Flege, J. E., MacKay, I. R. A., and Meador, D. (**1999**). "Native Italian speakers' perception and production of English vowels," J. Acoust. Soc. Am. **106**(5), 2973–2987.

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," J. Acoust. Soc. Am. **142**(4), 2007–2018.

Gerrits, E., and Schouten, M. E. H. (**2004**). "Categorical perception depends on the discrimination task," Percept. Psychophys. **66**(3), 363–376.

Hattori, K., and Iverson, P. (**2008**). "English /r/-/l/ pronunciation training for Japanese speakers," J. Acoust. Soc. Am. **123**(5), 3327.

Herd, W., Jongman, A., and Sereno, J. (**2013**). "Perceptual and production training of intervocalic /d, ɾ, r/ in American English learners of Spanish," J. Acoust. Soc. Am. **133**(6), 4247–4255.

Hirata, Y. (**2004**). "Computer assisted pronunciation training for native English speakers learning Japanese pitch and durational contrasts," Comput. Assist. Lang. Learn. **17**(3-4), 357–376.

Hu, W., Mi, L., Yang, Z., Tao, S., Li, M., Wang, W., Dong, Q., and Liu, C. (**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," J. Phon. **52**, 105–120.

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

Kartushina, N., and Frauenfelder, U. H. (**2014**). "On the effects of L2 perception and of individual differences in L1 production on L2 pronunciation," Front. Psychol. **5**, 1246.

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," J. Acoust. Soc. Am. **138**(2), 817–832.

Kawahara, H., and Morise, M. (**2011**). "Technical foundations of TANDEM-STRAIGHT, a speech analysis, modification and synthesis framework," in *Sadhana—Academy Proceedings in Engineering Sciences*, Vol. 36(5), pp. 713–727.

Kittredge, A. K., and Dell, G. S. (**2016**). "Learning to speak by listening: Transfer of phonotactics from perception to production," J. Mem. Lang. **89**, 8–22.

Kraljic, T., and Samuel, A. G. (**2005**). "Perceptual learning for speech: Is there a return to normal?," Cogn. Psychol. **51**(2), 141–178.

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," Appl. Psycholing. **26**, 227–247.

Lemhöfer, K., and Broersma, M. (**2012**). "Introducing LexTALE: A quick and valid lexical test for advanced learners of English," Behav. Res. **44**, 325–343.

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," J. Acoust. Soc. Am. **94**(3), 1242–1255.

Logan, J. S., Lively, S. E., and Pisoni, D. B. (**1991**). "Training Japanese listeners to identify English /r/ and /l/: A first report," J. Acoust. Soc. Am. **89**(2), 874–886.

Lopez-Soto, T., and Kewley-Port, D. (**2009**). "Relation of perception training to production of codas in English as a second language," Proc. Mtgs. Acoust. **6**, 062003.

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 Res. **1624**, 28–44.

Macmillan, N. A., and Creelman, C. D. (**1991**). *Detection Theory: A User's Guide* (Cambridge University Press, Cambridge, UK).

Mattys, S. L., Barden, K., and Samuel, A. G. (**2014**). "Extrinsic cognitive load impairs low-level speech perception," Psychon. Bull. Rev. **21**(3), 748–754.

Mattys, S. L., and Wiget, L. (**2011**). "Effects of cognitive load on speech recognition," J. Mem. Lang. **65**(2), 145–160.

Mitterer, H., and Mattys, S. L. (**2017**). "How does cognitive load influence speech perception? An encoding hypothesis," Atten. Percept. Psychophys. **79**(1), 344–351.

Newman, R. S. (**2003**). "Using links between speech perception and speech production to evaluate different acoustic metrics: A preliminary report," J. Acoust. Soc. Am. **113**(5), 2850–2860.

Peirce, J. W. (**2007**). "PsychoPy-Psychophysics software in Python," J. Neurosci. Meth. **162**(1-2), 8–13.

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," J. Acoust. Soc. Am. **130**(1), 461–472.

Rato, A. (**2014**). "Effect of perceptual training on the identification of English vowels by native speakers of European Portugese," in *Proceedings of the International Symposium on the Acquisition of Second Language Speech Concordia Working Papers in Applied Linguistics*, Vol. 5, pp. 529–547.

Sadakata, M., and McQueen, J. M. (**2013**). "High stimulus variability in nonnative speech learning supports formation of abstract categories: Evidence from Japanese geminates," J. Acoust. Soc. Am. **134**(2), 1324–1335.

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 Lang. **164**, 77–105.

Wang, X., and Munro, M. J. (**2004**). "Computer-based training for learning English vowel contrasts," System **32**(4), 539–552.

Wang, Y., Jongman, A., and Sereno, J. A. (**2003**). "Acoustic and perceptual evaluation of Mandarin tone productions before and after perceptual training," J. Acoust. Soc. Am. **113**(2), 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**(10), e109806.

Weber, A., and Cutler, A. (**2004**). "Lexical competition in non-native spoken-word recognition," J. Mem. Lang. **50**(1), 1–25.

Wong, J. W. S. (**2013**). "The effects of perceptual and/or productive training on the perception and production of English vowels /ɪ/ and /iː/ by Cantonese ESL learners," in *Conf. 14th Annu. Conf. Int. Speech Commun. Assoc. (Interspeech 2013)*, pp. 1–12.

Ylinen, S., Uther, M., Latvala, A., Vepsäläinen, S., Iverson, P., Akahane-yamada, R., and Näätänen, R. (**2009**). "Training the brain to weight speech cues differently: A study of Finnish second-language users of English," J. Cogn. Neurosci. **22**(6), 1319–1332.

Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X. A., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., and Woodland, P. (**2009**). *The HTK Book* (Cambridge University Engineering Department, Cambridge).