Donders Graduate School for Cognitive Neuroscience

Master of Science Programme

MSc Thesis

Do we learn from our mistakes?
The effects of communication disruptions on the production-perception link in L2 sound learning.

By

Lilian Ye

Supervised by:

1. Aurora Troncoso-Ruiz
2. Emily Felker
3. Dr. Mirjam Broersma

Radboud University Nijmegen

Date of final oral examination: 20-04-2020

Student Name: Lilian Ye

Student Number: s1013401

Do we learn from our mistakes? The effects of communication disruptions on the production-perception link in L2 sound learning

Lilian Ye[1]

Supervisors: Aurora Troncoso-Ruiz[2], Emily Felker[2] & Dr. Mirjam Broersma[2]

[1]Radboud University Nijmegen, Donders Institute for Brain, Cognition, and Behaviour, The Netherlands

[2]Centre for Language Studies, Radboud University, Nijmegen, The Netherlands

ABSTRACT

This study investigates the effects of communication disruptions on L2 sound learning. Specifically, it investigates whether implicit negative feedback on the L2 learner's production lead to adaptations in the L2 production and perception of problematic non-native sounds. German speakers of English were tested on their production and perception of two American English sound contrasts known to be difficult for the population: the /æ/-/ɛ/ vowel contrast and the word-final /t/-/d/ contrast. Their production and perception of these four sounds were assessed in a pre-post-test design. In between the pre- and post-test, participants interacted with a confederate, who they thought was a native American English speaker, in a cooperative computer-based task. During this interaction, they received negative feedback on their production of either the vowel or the word-final consonant contrast. Results showed that learning effects do not cross over to the perceptual domain, indicating that interactional production feedback does not lead to adaptations of the perceptual representations of the four difficult sounds. Disruptions in communication can raise the awareness for the difference between two contrasting sounds in the production domain, as Germans showed more native-like productions of these sounds in the post-test. However, whether the improved L2 productions are related to the type of sound contrast addressed in the interlocutor's feedback depends on the degree of difficulty of the respective sound contrast.

Keywords: second language acquisition (SLA), sound learning, speech perception, speech production, ventriloquist paradigm

INTRODUCTION

Learning the sounds of a novel language can be hard for non-native speakers, especially when they have to deal with new sounds that do not exist in their native language (L1) or when new sounds do occur in the L1, but are used differently in the second language (L2). It is therefore not surprising that the mismatch between the L1 and L2 sound inventories can make learners struggle when trying to pronounce L2 sounds. In turn, this could lead to miscommunications between the L2 learner and the native interlocutor. But, are L2 learners aware of the errors they make during an interaction, and more importantly, do they learn from it? The present study investigates the effect of communication disruptions on the L2 sound learning process. More specifically, we aim to increase our understanding of the link between speech production and perception processes in second language acquisition (SLA), as both processes are essential for communication. We are the first to examine the L2 perception-production relationship in an interactive context, in which the ecological validity is preserved while the phonetic input is fully controlled.

L2 sound learning during communication

Conversational interactions between two interlocutors are influenced by many factors. Speakers generally differ in physiology and may have different language backgrounds, for instance. Furthermore, social factors such as age and educational level, but also motivation can influence the nature of an interaction. Yet, speakers manage to overcome the enormous amount of variability in speech signals due to these factors, often by adjusting their pronunciations depending on whom they are conversing with. Pickering & Garrod (2004, 2006) refer to this adjustment as ‘alignment’, and argue that it is the key to successful communication. According to their Interactive-alignment Model of Dialogue, two interlocutors are able to successfully converse with each other by aligning their linguistic representations through automatic priming. For example, speakers may adjust their pronunciation or perception of particular sounds during a conversation in order to match the interlocutor’s phonetic representations. The alignment model is based on conversational interactions between two native speakers; however, it does not predict how alignment is reached between a native and a non-native speaker.

The following studies specifically examined phonetic alignment between native and non-native speakers. Kim, Horton, & Bradlow (2011) investigated phonetic alignment between native and Korean speakers of English. The study demonstrated stronger phonetic alignment between two native Korean speakers and two native English speakers, but weak phonetic alignment between natives of English and Korean speakers of English. Their results suggest that phonetic alignment between two interlocutors is dependent on language distance. However, Hwang, Brennan, & Huffman (2015) reported the opposite, as they found that Korean speakers of English produced more English-like sounds after they heard the native English confederate, while they did not align to the Korean-English confederate. Similarly, Lewandowski & Nygaard (2018) also showed that language background does not necessarily have to affect phonetic alignment, as native American English speakers aligned their productions to both native American English and Spanish-English speakers in the study.

The results mentioned above do not provide a clear account of phonetic alignment between L1 and L2 speakers, as there is both evidence for strong (Hwang et al., 2015; Lewandowski & Nygaard, 2018) and weak L1-L2 phonetic alignment (Kim et al., 2011). Costa, Pickering, & Sorace (2008) offer several reasons for why L1-L2 alignment might be reduced compared to L1-L1 alignment. One reason could be that L2 speakers require more conscious speech processing of the target language than L1 speakers, which could hinder their automatic retrieval of phonological representations. Even when L2 speakers are highly motivated to speak in the L2 and want to align with native speakers, sometimes their cognitive resources might limit them from doing so. On the other hand, L2 speakers may also consciously decide to avoid pronouncing certain sounds, because of their lack of L2 knowledge or confidence when speaking in the L2. Furthermore, Costa et al. (2008) suggest that (phonetic) alignment in L1-L2 conversation could be suboptimal, because the phonological representations in the L2 are strongly affected by the learner's L1 and therefore difficult to alter.

Besides cognitive factors and the L1-L2 relationship, Hwang et al. (2015) argue that communicative factors also play a role in phonetic alignment. In their second experiment, Korean speakers were visually exposed to words with particular sounds that are absent in the Korean language, and had to produce these for their native English partner to click on her screen. Participants showed more English-like productions when both target and phonetically-similar distractor words were visible on the screen. In other words, participants were able to successfully adapt their pronunciations only when they became aware of a potential ambiguous situation that could arise if they would not clearly disambiguate the two phonetically-similar sounds in their pronunciations. The researchers therefore suggest that learners are better able to phonetically align with the interlocutor when it is communicatively relevant to do so.

The lack of communicative reason could explain why the Korean speakers from the study of Kim et al. (2011) were unable to phonetically align with the native English speaker. The findings from Hwang et al. (2015) also speak against the notion that adaptations in L2 phonological representations are often constrained by the L1 (Costa et al., 2008), as the Korean speakers successfully learned to produce non-native sounds that do not exist in their L1 sound inventory. More importantly, the results indicate a crucial role for awareness in phonetic alignment and L2 sound learning.

Overall, the exact role of awareness in SLA during interaction is still under debate. The role of awareness in SLA has mainly been studied for other domains of language besides phonology. The Noticing Hypothesis (Schmidt, 1990) is specifically directed at L2 grammar for instance, and suggests that conscious awareness of the input is needed for successful grammar acquisition. It states that learners can only acquire L2 grammar when they consciously 'notice' the grammatical features of the input. However, the hypothesis has been challenged by several researchers (e.g., Truscott, 1998), as it lacks specific detail about the exact definition of conscious ‘noticing’ and how to test this empirically. Another theory suggests that not conscious awareness, but detection is needed for L2 grammar learning (Tomlin & Villa, 1994). Once information is detected and registered in the short-term memory, it can be used by higher levels of cognitive processes for further processing. Importantly, it is claimed that detection does not require awareness, suggesting that (grammar) learning can take place without awareness. Robinson (1995) argues that it is the combination of both detection and awareness that leads to ‘noticing’ and learning. Detection allows the learner to become aware of the registered information and enables learning, subsequently leading to encoding in the long-term memory.

None of the above three theoretical accounts on the role of awareness in SLA pose specific predictions for L2 phonological learning, as most of them were focused on L2 grammar acquisition. Furthermore, none of the accounts have related their hypothesis to previous studies conducted in naturalistic situations of language learning, but only to studies in laboratory or classroom settings. Truscott (1998) has also criticised this, as the Noticing Hypothesis was mainly built on findings from studies involving form-focused instruction for example. However, L2 learners might process non-native speech differently during real-life communication. Besides the fact that interlocutors also have to deal with different social or motivational factors and higher processing loads in conversational interactions, an important distinction between naturalistic and controlled settings is the presence of communicative factors, as this factor is often lacking in controlled settings.

To our knowledge, no studies have looked at awareness and L2 sound learning during interaction so far. Therefore, we discuss two studies on phonological adjustments within native speakers to examine the role of awareness in sound learning during interaction. The first study by Schertz (2013) investigated how native speakers of English adapt their pronunciations during a human-computer interaction to clarify misunderstood speech. Participants read out words while the automatic speech recognition (ASR) system tried to guess which word the participant had produced. When the computer program guessed incorrectly, participants were instructed to repeat the word. Results showed that voice-onset times (VOTs) became longer for word-initial voiceless stops on the second try when the computer perceived the word-initial phoneme as a voiced stop. Vice versa, VOTs became shorter for word-initial voiced stops when the computer had guessed a word that began with the contrasting voiceless stop. This was only the case when the computer had (mis)understood the word as the other member of the voiced-voiceless minimal pair, and not when the computer asked for another repetition. The findings indicate that speakers are able to instantly adapt their pronunciations during interaction, but only when the computer’s guess specifically elicited the phonological contrast.

Secondly, Buz, Tanenhaus, & Jaeger (2016) examined how native speakers clarify misunderstood speech during a more naturalistic type of human-computer interaction, and also under more implicit circumstances. Native speakers of English were told that they would be matched with a human partner through the internet, but unbeknownst to them, they were actually interacting with a simulated partner. Participants saw three words on the screen and were instructed to read out the cued word to their partner, which the (simulated) partner had to click on his screen. The No Feedback group did not see the partner's response, whereas the Positive Feedback and the Mixed Feedback groups did receive information about the partner's choice. Similar to Schertz (2013), the target words from critical trials contained word-initial voiceless stops (e.g., *pill*) and some of the distractor words contained contrastive word-initial voiced stops (e.g., *bill*). The (simulated) partner would always choose the correct target word for the Positive Feedback group, but occasionally chose the voiced competitor word instead of the voiceless target word for the Mixed Feedback group. Results showed that VOTs became longer for the target words beginning with voiceless stops when a voiced competitor word was also present during the trial. Furthermore, the VOTs of word-initial voiceless stops were significantly longer for the Mixed Feedback group, who were occasionally misunderstood by the (simulated) partner. These findings are in line with Schertz (2013), and demonstrate that speakers modify their pronunciations when they notice that they were being misunderstood. Interestingly, the phonological adaptations were not word-specific as participants were not allowed to correct themselves after an incorrect response, which suggests that they may have implicitly learned the phonological contrast due to the negative feedback that they received from the interlocutor.

While Buz et al. (2016) already tried to increase the naturalness of the interaction by employing a simulated human partner, both them and Schertz (2013) used computer-based approaches to study the role of interactional feedback in sound learning. However, earlier research demonstrated that human-computer interactions elicit different speech than human-human interactions (Burnham, Joeffry, & Rice, 2010; Oviatt, Levow, Moreton, & MacEachern, 1998). For example, Burnham et al. (2010) directly compared speech during human-computer to human-human interactions. They found that native speakers of English from both types of interactions overall showed clearer productions when they were clarifying misunderstood speech. More importantly, their results demonstrate that participants hyperarticulated their speech more when they were interacting with the virtual avatar compared to the human partner. Hence, it is not clear to what extent the results from Buz et al. (2016) and Schertz (2013) can be generalised to human-human interactions.

We are therefore left with two gaps in the literature on awareness and SLA. Firstly, the role of awareness in *L2* sound learning remains unclear, and secondly, it is not known how awareness is involved in more *naturalistic* settings of language learning such as conversational interactions. As speakers from both Buz et al. (2016) and Schertz (2013) differentiated phonological contrasts best when they received negative feedback from the interlocutor that highlighted the difference between the two target sounds, we could assume from the two studies conducted with native speakers that error detection might be an important mechanism underlying sound learning. Whereas the meaning behind positive feedback can be somewhat ambiguous, negative feedback clearly indicates some deficiency in a speech process (Ellis, 2009). Therefore, the interlocutor's negative feedback could help speakers to detect phonetic details of the language and shift their attention towards it, eventually leading to adaptations in their phonological representations.

On that account, error detection might also be an important mechanism for L2 sound learning, as L2 learners have yet to become familiarised with the sounds and phonological rules from the novel language, making them more susceptible to making pronunciation errors than native speakers. Therefore, one could say that it is especially important for L2 learners to become aware of their pronunciation errors in order to learn non-native sounds. One way to clearly 'notice' or become aware of one's mistake is when the pronunciation error results in a disruption of communication. When a speaker is being misunderstood, the flow of the conversation is interrupted. Interlocutors are then required to re-evaluate their speech perception or production in order to find the cause of the disruption and solve it, before they can resume the conversation. This whole process could make interlocutors, and especially L2 learners, aware of phonological details in the language that otherwise may have never come to their attention, ultimately leading to adaptations in their (L2) production.

If it turns out that L2 speakers improve their pronunciation of non-native sounds due to interactional feedback, perhaps it may also be the case that they also enhance their perception of these sounds. We will discuss the relationship between L2 speech perception and production processes more detailed in the sections below.

Learning new phonological categories

Investigating the relationship between speech perception and production is important in order to understand L2 learning, as both processes are needed for successful communication. Common L2 perception and production mistakes include segmental errors, such as confusing acoustically-similar consonants or vowels. A typical example includes the English /r-l/ contrast that Japanese speakers often find difficult to perceive and produce, as Japanese only has a phoneme that falls somewhere in between the English /r/ and /l/ sounds (Goto, 1971; Lotto, Sato, & Diehl, 2004). Dutch and German speakers of English typically struggle with the English /æ/-/ɛ/ vowel contrast, namely because the /æ/-vowel does not exist in both languages' phonological inventories (Bohn & Flege, 1990, 1992; Broersma, 2002; Weber & Cutler, 2004). This population also seems to experience difficulty with the English distinction between voiced and voiceless word-final sounds such as /t/ and /d/. Whereas native speakers of English already tend to devoice the voiced /d/ sound when it is placed in word-final position (Ladefoged, 1982), the distinction of voiced and voiceless word-final obstruents is even harder for Dutch and German speakers of English (Broersma, 2002; Port & O'Dell, 1985; Smith, Hayes-Harb, Bruss, & Harker, 2009).

The examples mentioned above suggest that the type of non-native sounds that L2 learners find difficult to discriminate seems to be dependent on the amount of phonetic overlap between the L1 and the L2. However, different models have different predictions about these mismatches between L1-L2 inventories. The Perceptual Assimilation Model (PAM) hypothesises that the perception of non-native sounds is influenced by the learners phonological experience in the native language (Best, McRoberts, & Sithole, 1988; Best & Tyler, 2007). The model has several predictions about how the learner will perceive unfamiliar non-native sounds, depending on the L1-L2 relationship. For example, unfamiliar non-native sounds can be perceptually mapped onto the phoneme in the L1 that comes closest in terms of articulatory features. This type of perceptual assimilation is called Single Category (SC) assimilation. It is predicted that L2 learners poorly discriminate SC assimilated sounds, as the contrasting phonemes are mapped onto the same native phoneme. Broersma & Cutler (2008) demonstrate how L2 speech processing can be constrained because of SC assimilation by studying the phenomena of phantom word activation in Dutch listeners of English. Their results showed that Dutch listeners perceived more non-words as real words that contain the English contrast of voiced and voiceless phonemes in word-final position, compared to British native speakers. Furthermore, these non-words also primed the lexical activation of real English words. In a follow-up study, Broersma & Cutler (2011) demonstrate a similar effect for words containing the British English /æ/-/ɛ/ vowel contrast, as non-words including these vowels were indeed co-activated in the lexical competition process for the Dutch listeners of English.

Likewise, The Speech Learning Model (SLM; Flege, 1995) also predicts that L2 speech processing is based on the similarity between novel sounds and already existing phonetic categories from the L1. An important difference between the two models is that SLM posits a central role for mental representations of phonological categories stored in the long-term memory in speech processing, whereas PAM assumes that articulatory features of the speech signal and how accurately they are perceived underlie speech processing. Both theoretical accounts do agree that L2 learners are able to overcome perceptual assimilation of non-native sounds, but propose different mechanisms underlying the learning process. PAM postulates that perceptual learning is driven by the articulatory gestures of non-native sounds, suggesting that representations are shared between speech perception and production. SLM states that L2 production is dependent on how well the learner can perceptually discriminate contrasting sounds, meaning that phonological representations are at the basis of L2 sound learning.

Whether representations are shared between perception and production processes or not, both theories predict that L2 learners are able to successfully learn problematic L2 sounds over time. Empirical studies mainly use phonetic training paradigms to illustrate L2 sound learning, many finding that L2 learners can successfully enhance their speech processing of the problematic sound contrast through extensive training. Japanese listeners’ ability to identify the English /r/ and /l/ phonemes improved after they received high-variability perceptual training on the two sounds (Lively, Logan, & Pisoni, 1993; Logan, Lively, & Pisoni, 1991). Likewise, high-variability training also enhanced both German and Spanish listeners identification of English vowels (Iverson & Evans, 2009), and Dutch listeners identification of Japanese geminates (Sadakata & McQueen, 2013). L2 learners production skills have also been shown to improve as a result of pronunciation training or form-focused instruction (Hirata, 2004; Saito & Lyster, 2012).

The link between L2 speech perception and production

As communication involves both speech perception and production, researchers have also specifically studied whether training in one domain additionally leads to learning effects in the other domain. There seems to be a tight link between perception and production representations for the L1 (Meyer, Huettig, & Levelt, 2016), but this relationship is less clear for the L2. Earlier research on the effects of perceptual training on L2 production overall demonstrate that perception training positively affects L2 production abilities as well. For example, Japanese learners of English improved both their perception and production of the /r/-/l/ contrast after perceptual training, and learning effects were retained 3-months after the training session (Bradlow, Akahane-Yamada, Pisoni, & Tohkura, 1999; Bradlow, Pisoni, Akahane-Yamada, & Tohkura, 1997). Furthermore, native American English speakers' production of Mandarin tones also improved significantly after they had been perceptually trained on the tone contrasts (Wang, Jongman, & Sereno, 2003). However, previous literature about the relationship between production training and L2 perception shows less consistent results.

The following three studies demonstrated a transfer of learning from the production to the perceptual domain. Arabic learners of English improved their perception of the problematic English /p/-/b/ contrast (Linebaugh & Roche, 2013) and three types of vowel contrasts (Linebaugh & Roche, 2015) after they had received production training. Another study showed that native French speakers' production of Danish vowels became more native-like after receiving production training, and they were also able to perceptually discriminate the vowels better compared to those without production training (Kartushina, Hervais-Adelman, Frauenfelder, & Golestani, 2015). The third study by Wong (2014) investigated the effects of both production and perception training on L2 speech processing in Cantonese speakers of English. Results showed that both types of training led to improvements in the other domain; however, the degree of improvement differed between the two groups. Only moderate improvements in the perception of the English /æ/ and /ɛ/ vowels for the explicit articulatory production training group were found compared to the robust enhancement of vowel productions found in the high-variability perceptual training group. Besides a positive transfer effect, the findings also suggest that perception training could be more beneficial than production training for both perceptual and production learning.

Interestingly, all studies mentioned above report transfer of learning effects using explicit production training methods in which trainees receive specific information about the acoustic features of the non-native sounds, for instance through teacher instructions (Linebaugh & Roche, 2013, 2015), live visual feedback on the subjects' production after each trial (Kartushina et al., 2015), or instructions through video recordings and pictures illustrating the acoustic properties of the non-native sounds (Wong, 2014). However, findings from explicit training studies are difficult to relate to L2 learning in more realistic settings, as learners do not have access to these explicit measures during conversations in real-life. As a matter of fact, even native speakers often cannot describe the exact acoustic differences between phonetically-similar sounds or explain phonological rules of their language, despite the fact that they can ‘correctly’ pronounce those sounds. Therefore, implicit learning methods could tell us more about how L2 learners learn non-native sounds in more naturalistic settings.

Thorin, Sadakata, Desain, & McQueen (2018) used an implicit training method to investigate the L2 perception-production link and showed no direct link between speech perception and production processes when learners receive implicit production training, thereby contradicting the results from studies that explicitly inform the learners about the acoustic properties of the non-native sounds (Kartushina et al., 2015; Linebaugh & Roche, 2013, 2015; Wong, 2014). The researchers studied the effect of adding implicit production practice to a perceptual training paradigm on the perception (and production) of the British English /æ/ and /ɛ/ vowels in native Dutch speakers. Participants were divided into a related-production or an unrelated-production group. Training was spread across 5 days and during each training session participants listened to /æ/-/ɛ/ minimal pairs and indicated which word they had heard last in the auditory sequence. They received visual feedback on their perceptual judgement, and thereafter, a word appeared on the screen which they had to produce. The word for the related-production group was always the correct answer to the previous sequence, while the unrelated-production group had to pronounce a word that did not contain the target vowel contrast. Importantly, the participants were never informed about the acoustic properties of the British English vowels, nor did they receive feedback on their pronunciations during the training sessions. The results demonstrated that both training groups showed perceptual learning over time, but no differences were found between the groups, indicating that related production training does not affect perceptual learning differently than unrelated production training. Their findings illustrate that implicit and explicit production training studies may have different effects on the relationship between L2 production and perception processes, partially explaining the inconsistent results in the literature.

Another reason for inconsistent results on the L2 perception-production relationship could be that the link between production training and perception abilities might be more complicated to study and understand than the reverse direction, because most production training paradigms also seem to require speech perception. Whereas learners can easily take part in a perceptual training paradigm without having to produce anything themselves, it is rather difficult to completely eliminate speech perception from production training methods. Perhaps it is for this reason that studies investigate L2 sound learning by combining speech perception and production processes in one task. Llompart & Reinisch (2018) used imitation to study the L2 perception-production link for instance, as both speech perception and production processes are involved in imitation, while Baese-Berk & Samuel (2016) and Herd, Jongman, & Sereno (2013) studied the link by adding production practice to a perceptual training paradigm (similar to Thorin et al., 2018).

Llompart & Reinisch (2018) investigated the relation between imitation ability and the two speech processes separately in German speakers of English. Participants were tested on their perception, imitation and production of the /i/-/ɪ/ contrast, considered relatively easy for German speakers of English, and the more difficult /æ/-/ɛ/ contrast. Their perception of these vowels was assessed in a perceptual categorization task while their production was investigated through a word reading task. In the imitation task, participants heard a sequence of two English words and were instructed to imitate the second word they heard. Results showed a stronger correlation between the perceptual categorization and imitation slopes for the 'difficult' /æ/-/ɛ/ contrast than the 'easy' /i-ɪ/ contrast. This finding indicates that imitation of the two contrasts is influenced by how the vowels are phonologically encoded, suggesting that the phonological representations for /i-ɪ/ are stronger than for /æ/-/ɛ/. The correlation between the productions from the imitation and the word reading task was large for the /i-ɪ/ contrast, but weak for the /æ/-/ɛ/ contrast, suggesting that L2 phonological representations do not necessarily relate to the production of L2 sounds.

Baese-Berk & Samuel (2016) studied the effect of combined perception and production training vs. perception training alone on the perceptual learning of Basque sound contrasts in native Spanish speakers. The perception-only group performed an ABX discrimination task on the Basque /s̺a/–/ʃa/ contrast during training. The combined training group also did the ABX discrimination task during training, but they were also instructed to produce the last heard token before they made the judgment. Their findings showed that the perception-only group learned to discriminate the Basque sound contrast after training, while perceptual learning for the combined group (which included production training) was disrupted. In another experiment, the researchers showed that the disruption of perceptual learning was decreased in Spanish speakers that did have prior knowledge of Basque; however, the disruption could not be totally eliminated by prior language experience. These results suggest that adding production training can disrupt the L2 speaker's perceptual learning abilities, speaking against the positive transfer effect found in other studies. Herd, Jongman, & Sereno (2013) also investigated the role of combined perception and production training on L2 speech perception and production, and did not find a transfer effect either. Their combined training paradigm led to improvements in the production of Spanish intervocalic /d/, /ɾ/ and /r/ by native American English speakers, but no improvements were demonstrated for the perception of these sounds.

Crucially, Herd et al. (2013) additionally examined whether production training *alone* leads to improved L2 perception and compared to the perception-only training. In the perception-only training, participants received high-variability perceptual training of the three Spanish sounds. In the production-only training, subjects read out minimal pairs containing the Spanish sounds. During training, they were presented with the waveforms and spectrograms of both their own and a native speaker's pronunciation of a word with Praat (Boersma & Weenink, 2015), so they could see how close their pronunciation was to that from the native speaker. Importantly, they were only provided with their own and the native speaker's spectrographic information, but never heard the native speaker's actual pronunciation of these words, thereby ruling out any learning effects due to auditory perception. Results showed that perception-only and production-only trainings equally improved Americans English speakers' perception of the three Spanish intervocalic sounds, suggesting that production training alone can positively influence perceptual learning.

Since Herd et al. (2013) also took an explicit approach in their training paradigm, no studies to date have investigated the effect of implicit production-only training on L2 speech perception processes. In addition, the link between production training and L2 perception has mainly been studied in laboratory studies or in the classroom, but no studies investigated the link in more naturalistic settings. Studying L2 sound learning in an interactive context is essential in order to take factors, such as social or communicative reasons, into consideration that are often lacking in controlled settings.

The present study

The goal of the present study was to investigate the effect of communication disruptions on error-based L2 sound learning. The first aim of the study is to examine the effect of implicit production feedback on the L2 learner's pronunciation of non-native sounds. We specifically investigated whether the interlocutor's implicit negative feedback on the learner's production of non-native vowel and consonant contrasts, thereby inducing a disruption in communication, leads to adaptations in the pronunciation of these non-native sounds. Secondly, we also attempted to gain more insight into the L2 speech perception-production link by studying the effect of interactional feedback addressing the L2 learners' production abilities, again inducing a communication disruption, on their perception of the problematic non-native sounds.

Importantly, we study the two research questions during human-human interaction, making us the first study to investigate the L2 speech perception-production link in an interactive context. A big challenge in studying communication between two speakers is the control of phonological input while also keeping the experiment ecologically valid, explaining why many researchers decide to study L2 sound learning through human-computer interactions. The novel ventriloquist paradigm (Felker, Troncoso-Ruiz, Ernestus, & Broersma, 2018) resolves this dilemma as it allows researchers to have full control over the speech input while maximising the ecological validity of the study. In the ventriloquist paradigm, participants interact face-to-face with a confederate who, unbeknownst to them, only interacts with them through pre-recorded speech. Whenever the confederate 'speaks' to the participant, she plays pre-recorded audio files instead. In this way, participants are led to believe that they are having a real conversation with the confederate, which gives researchers the opportunity to expose each participant to the exact same phonetic input while preserving the authenticity of the interaction.

We therefore used the newly developed ventriloquist paradigm (Felker et al., 2018) to study the effect of communication disruptions on L2 speech production and perception in an interactive context. German learners of English were assigned to either a production group or a perception group and interacted with a confederate during the experiment. The German participants were tested on their production or perception of two American English sound contrasts: the /æ/-/ɛ/ vowel and the word-final /t/-/d/ contrasts. The vowels differ in height (which is lower for /æ/), frontness (/æ/ is pronounced less fronted), and duration (/æ/ has longer duration), while the stop consonants vary in terms of voicing (/d/ is more voiced than /t/). These two sound contrasts are known to be problematic for native German speakers (e.g., Bohn & Flege, 1992; Broersma, 2002; Eger & Reinisch, 2019).

The study was divided into a pre- and post-test with in between an interaction phase in which participants engaged in interaction with a confederate. Participants completed an interactive version of the Lexical Decision (LD) task with the confederate in the pre- and post-tests, the so called ‘Word or Not’ Games. One person had to read out words containing the critical sounds while the other had to decide for each item whether it was an existing English word or not. The pre- and post-test were designed to obtain baseline and post-interaction production or perception abilities of the critical sounds. Many studies have investigated the speech processes during the interaction phase itself; however, we decided to specifically compare the production and perception of critical sounds from the pre- and post-tests in order to investigate L2 sound learning over time.

In the interaction phase, participants interacted with ‘a native speaker of English’ in an information-gap task, the so called ‘Code Breaker Game’. In the Code Breaker Game, the confederate mentioned which shape was missing in her puzzle sequence, while the participant read out the word that was linked to this missing shape for the confederate to click on her screen. Importantly, participants were able to see which word the confederate had eventually chosen. We systematically disrupted the communication during this interaction phase by giving participants implicit negative feedback on their production of either the vowel or the consonant contrast, as the confederate always pretended to misunderstand critical target words for the other member of the respective minimal pair.

We propose that the communication disruptions, by means of the interlocutor’s implicit negative feedback, raises L2 learners’ awareness of phonetic details in non-native speech through error detection. As mentioned earlier, we believe that the interlocutor’s feedback could force learners to reflect on the situation, which may lead to reconsiderations of their current phonological representations of the non-native sounds. Learning is especially expected when there is a communicative relevant reason for learners to disambiguate two contrasting sounds in their pronunciation (as demonstrated by Hwang et al., 2015), like solving a puzzle together.

For the production group, we hypothesised that German speakers of English achieve more native-like pronunciations of the critical sounds in the post-test due to the interlocutor’s implicit negative feedback during the interaction phase. More specifically, we predicted that participants mainly improve their pronunciation of the sound contrast that was addressed in the disruption trials of the interaction phase, as these instances require participants to reflect on their pronunciation of the respective sound contrast more than trials without a disruption in the communication. To test our predictions, we analysed the pronunciations of words containing the critical sounds during the pre- and post-tests in terms of four acoustic measures: F1, F2 and vowel duration for the vowel contrast, and the amount of voicing for the word-final consonant contrast.

If speech perception and production processes share representations (suggested by PAM; Best & Tyler, 2007), then the interlocutor's implicit negative feedback should also enhance German speakers' perception of the critical sounds. For the perception group, we analysed how accurate participants were in their acceptance of real words and rejection of non-words containing the critical sounds in the pre- and post-tests. Following PAM's hypothesis, we predicted higher accuracy scores in the post-test for words and non-words containing the sound contrast on which participants received negative feedback during the interaction phase. Alternatively, if we follow SLM's hypothesis that perception precedes production and not vice versa (Flege, 1995), then no difference in accuracy scores between the pre- and post-test is expected. SLM states that adaptations in the perceptual phonological representations form the basis for L2 learning. However, the interactional feedback from the Code Breaker Game is primarily aimed at inducing changes in the production domain, hence not directly addressing learners' perception of the critical sounds.

METHODS

Participants

Forty native German speakers participated in this study (age: *M* = 20.8 years, *SD* = 2.24; 15 males). Most of them were students recruited through the Radboud University's Research Participation System. We also recruited German native speakers who were following the crash course Dutch from Radboud in'to Languages in Nijmegen during the summer of 2019. All participants were raised monolingually and the majority (95%) reported having learned English as their first foreign language. The participants had received 9.67 years on average of formal English training, and their average English proficiency was 73.3%, as measured by the LexTALE (Lemhöfer & Broersma, 2012) conducted at the end of the experiment. All participants reported having no learning, hearing or reading deficits.

The study was approved by the Ethics Assessment Committee Humanities at the Radboud University in Nijmegen. Participants gave written informed consent and received 10 euros in the form of a gift card or 1 participation credit after the experiment.

Materials

Word or Not Games. Each pre- and post-test consisted of 192 items: 96 monosyllabic English words and 96 non-words. The critical tokens were selected from a list of 96 monosyllabic English words, of which 24 words contained the /æ/-vowel, 24 words the /ɛ/-vowel, 24 words ended with /t/ and 24 words ended with /d/. We avoided selecting English-German cognates that were similar in orthography *and* phonology, and words that contained more than one of the critical sounds.

Furthermore, we selected high-frequency words taken from the SUBTLEX-US corpus (van Heuven, Mandera, Keuleers, & Brysbaert, 2014) with a mean Zipf value (i.e., the log10(frequency per billion words)) of 4.39.

For each of these real English words, a non-word counterpart was created by replacing the critical sound with the other sound of the respective sound contrast. For example, the non-word '*dask'* was created by replacing the /ɛ/-vowel in '*desk'* with the /æ/-vowel, and the non-word '*golt'* was formed by replacing the /d/ consonant in '*gold*' with the /t/ consonant, thereby creating 96 monosyllabic English non-words. Ultimately, only one token from each word–non-word pair was chosen to be included in the experiment. Hence, the final 96 critical tokens for the Word or Not Games consisted of 48 monosyllabic English words and 48 monosyllabic non-words, both having 12 tokens per sound (/æ/, /ɛ/, final /t/ and final /d/; see Appendix A). The other 96 tokens contained phonemes taken from different sorts of sound contrasts that are easier for German speakers to perceive and produce, and functioned as filler items. To make the presence of the critical sounds less evident, we included different filler items in the pre- and post-test. Importantly, the critical tokens were the same between the pre- and post-test.

Each test presented the items in two blocks of 96 trials, separated by a slide indicating that participants were halfway through the 'Word or Not' Game. Items were presented in pseudorandom order, as we wanted critical items from the same sound contrast to be separated by at least two fillers or critical items from the other sound contrasts.

Code Breaker Game. The Code Breaker Game consisted of 64 trials that were presented in one block. For the game, we selected 64 English minimal pairs. Twelve minimal pairs contained the /æ/-/ɛ/ vowel contrast and 12 minimal pairs contained the word-final /t/-/d/ contrast. These minimal pairs were considered our critical items (see Appendix B), and were different from the critical items of the Word or Not Game. Again, we avoided to include English-German cognates that had similar orthography *and* phonology. We also controlled for words containing more than one of our target sounds. The mean Zipf value (van Heuven et al., 2014) between /æ/ and /ɛ/ words from the /æ/-/ɛ/ minimal pairs did not differ significantly ($t(22) = 0.39$, $p = .70$). The same applied to /t/ and /d/ words from the /t/-/d/ minimal pairs ($t(22) = 0.17$, $p = .86$). The mean Zipf values between the two sound contrasts did not differ significantly either ($t(46) = 0.50$, $p = .62$). Furthermore, we selected 40 minimal pairs containing easier types of English sound contrasts (e.g., /i-ɪ/ and /r-l/) that acted as our fillers.

In addition, we designed 64 puzzle sequences with their corresponding answer and alternatives for the Code Breaker Game. Each puzzle consisted of five coloured shapes presented in a sequence followed by a question mark, and four coloured shapes that functioned as answer options. The puzzle sequences would be presented to the confederate, while the answer options would appear on the participant's computer screen. Importantly, we created the puzzles by mainly using shapes and colours that did not contain one of the four critical sounds, as we want these sounds to be excluded from the ventriloquist's speech recordings (see 'Ventriloquist speech recordings' section). Each puzzle was paired with two minimal pairs (i.e., four words), and each minimal pair appeared twice in the game: once as target sound contrast (meaning that one member of the minimal pair is the correct answer to the puzzle sequence) and once as background sound contrast (meaning that none of the minimal pair members were the correct answer). Moreover, critical minimal pairs were always combined with filler minimal pairs to ensure that the participant would not have to deal with two 'difficult' sound contrasts in one trial.

Ventriloquist speech recordings. ***Voice of the ventriloquist.*** All of the confederate's speech materials were pre-recorded by a female native speaker of English, who spoke with a Midwestern American English accent. The recordings were made in a soundproof booth using a head-mounted microphone that was connected to Audacity, with a sampling rate of 44.1 kHz. Crucially, the female native speaker never produced the four critical sounds as these were excluded from all the speech materials (apart from the critical tokens for the Word or Not Game). This was necessary to make sure that any learning effects would be due to communication disruptions by means of negative feedback, ruling out the influence of auditory exposure of these sounds.

Trial-Linked utterances. The ventriloquist's speech can be divided into two types of utterances: Trial-Linked and Spontaneous utterances (see Felker et al., 2018 for a detailed description of the exact implementation). Trial-Linked utterances refer to speech that is associated with a specific moment in the experiment. For this study, the Trial-Linked utterances consisted of the items for the Word or Not games, the puzzle descriptions for the Code Breaker Game, and sentences that were linked to the introduction, instruction and end screens of the experiment. We will discuss each group of Trial-Linked utterances in the sections below.

Word or Not Game. The items of the 'Word or Not' Games were pre-recorded, so that the confederate could play audio recordings during the perception version of the experiment. For the production version of the pre- and post-tests, it was not necessary for the confederate to play the audio recordings, as participants themselves would read out the items instead. The items were recorded in isolation by the female native speaker. Half of the items contained one of the critical sounds and the other half did not.

Code Breaker Game. In the Code Breaker Game, the Trial-Linked utterances consisted of sentences that describe which coloured shape is missing in the puzzle sequence. We scripted and pre-recorded two sentences per puzzle. The first sentence would always be played by the confederate, whereas the second sentence served as a 'follow-up' description that put more emphasis on the missing shape, and would only be played when necessary (e.g., when the participant would ask for repetition or clarification). As mentioned before, the puzzles were designed so that the puzzle descriptions never required words that contain one of the four critical sounds. For example, we tried to refrain from using triangular and squared shapes in our puzzles as these contain the /æ/-/ɛ/ vowel, the same applied to using colours such as red or white. Whenever a coloured shape that contained one of the critical sounds was included in a puzzle, it was always placed in a non-crucial position so that the confederate never had to mention it while describing the puzzle.

The puzzle descriptions were designed using the same methods that were applied in the original version of the ventriloquist paradigm (Felker et al., 2018). However, former participants that suspected the pre-recorded speech often mentioned that the confederate's speech sounded 'too perfect'. Therefore, new approaches were taken into the scripting process for the current adaptation of the ventriloquist paradigm in order to improve the naturalness of the speech recordings. We successfully improved the quality of the ventriloquist speech recordings, as 79.2% of the participants from the original version reported no suspicion of the pre-recorded speech while the success rate for the current version was 90.0%.

We included improvements in four aspects: naturalness, spontaneity, continuity and interactivity. Firstly, the *naturalness* of the speech was improved by purposely including filler words such as '*Uhm'* and '*Okay'*, or by repeating some words in the scripted sentences, to add some hesitation in the confederate's speech:

(1) Trial 01: "*Okay, so I'm looking for the label under an… an orange sun."*

Secondly, *Spontaneity* was added to the speech recordings by purposely reducing the length of the confederate's puzzle descriptions, as former participants mentioned that their partner always used full sentences to describe the puzzles, which felt unrealistic to them. For the current study, we scripted full sentences particularly for the first few trials of the Code Breaker Game to make sure the participants would understand the essence of the game, but reduced the length of sentences as trials progressed, with the thought of both confederate and participant becoming more familiar with the game at some point:

(2) Trial 10: *"This one is a blue circle"*
Trial 15: "*A purple moon"*

The third improvement about the c*ontinuity* of the interaction included anaphoric elements in order to create direct relations between utterances of consecutive trials:

(3) Trial 19: *"Uh, a brown shape"*
Trial 20: *"Now it's a shape in blue"*

Lastly, the most innovative measure we took was to deliberately leave out crucial information in some of the initial puzzle descriptions to increase the *interactivity* of the conversation between the participant and the confederate. By leaving out information, participants would have to ask more questions about the missing shape, which in turn the confederate could answer by playing the follow-up sentence:

(4) Trial 31, sentence 1: *"This one is a smiley face"*

Trial 31, sentence 2: *“It’s an orange smiley face”*

Other Trial-Linked speech. Beside the speech materials that were necessary to play the ‘Word or Not’ and Code Breaker Games, we also scripted and pre-recorded sentences the confederate could play during specific instruction slides in the experiment. Sentences that explained the goal of each game were included in case the participant would not completely understand the respective instruction slide. For instance, the sentence *“So, I’ll be the one reading the whole time, while you’ll be making the decisions”* was scripted for the perception pre-test instruction slide. Furthermore, we also scripted and pre-recorded the confederate’s introduction and goodbye speech.

Spontaneous utterances. Whereas Trial-Linked utterances could only be played by the confederate during fixed time points in the experiment, Spontaneous utterances could be played throughout the whole experiment at any given time. These utterances served as responses to any situations that could not be handled with the Trial-Linked utterances and also helped maintaining the illusion of a spontaneous conversation. We scripted and pre-recorded spontaneous utterances that can be divided into the following nine categories: *Affirmative, Negative, Listen, Don’t Know, Reassurance, Repeat, Next, Confirming,* and *Rules*. Each category included different variations of the utterances (varying from 5 to 24 unique recordings depending on how frequent the category could be needed during the interaction), so the confederate had enough ‘authentic’ audio files to play. Examples of the Spontaneous utterances can be found in Table 1.

Table 1.

The nine Spontaneous Categories used in the current adaptation of the ventriloquist paradigm and example utterances to illustrate the purpose of each audio category.

Spontaneous Category	**Example**
Affirmative	*“Yup”*
Negative	*“Nope”*
Listen	*“Mm-hmm”*
Don’t Know	*“Uhm… I dunno”*
Reassurance	*“Oh, it’s okay”*
Repeat	*“Sorry?”*
Next	*“Oh, here’s a new puzzle”*
Confirming	*“Uh, seems fine to me”*
Rules	*“Uh, I think the rules say your job is only to say the label under the shape?”*

Procedure

It is of extreme importance that the participant and confederate only communicate with each other during the experiment in order for the ventriloquist paradigm to succeed, since the confederate's speech is restricted to the pre-recorded audio files. So, besides the procedures of the different experimental phases in the study, we also describe the basics of the ventriloquist paradigm in this section as well as how the entire experimental session was arranged to demonstrate the preconditions necessary for the ventriloquist paradigm.

Before the interaction. Upon arrival, participants were led from the control room (Fig. 1A) to the small room behind the studio (Fig. 1B). This had to be done rather quickly, to prevent them from glancing through the window to inspect the sound-proof studio, as the confederate had already taken place in there (Fig. 1C). In the small room, participants were given the information document and the consent form. After they gave consent, the experiment leader shortly explained the outline of the experiment. During that explanation, it was made clear that participants and their partner ('a native speaker of English') had to communicate with each other during the experiment using the microphones and the noise-cancelling headphones, 'in order to improve the quality of the audio recordings' that were going to be made during their interaction. They were also told that the partner was already in the studio doing another task. Then, participants were quietly brought into the studio and seated in front of a computer screen diagonally across the 'other participant', who was still pretending to finish her computer task. The participant was reminded once more to speak into the microphone on the table during the experiment, while the partner would speak into the 'microphone' that was placed in the corner on her right side. If there were no questions left, the participant could put on the noise-cancelling headphones and was instructed to wait for the 'other participant' to finish her task. The experiment leader then left the studio and took place in the control room.

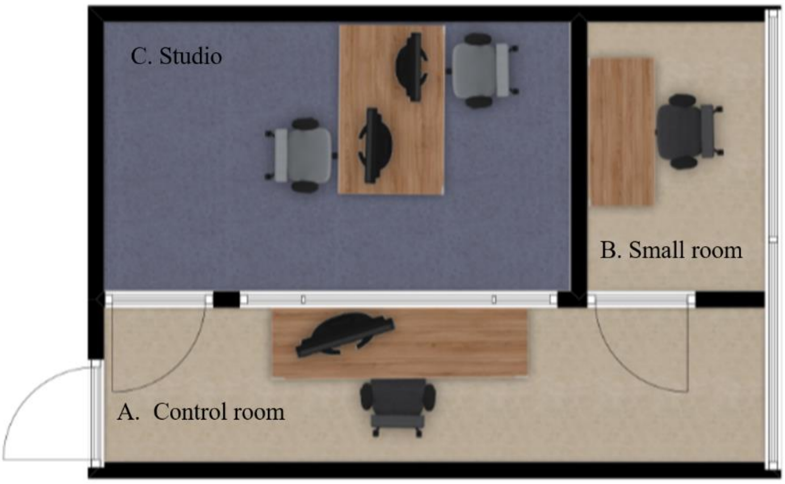

Fig. 1. Schematic overview of the lab. **A.** The control room, where the experiment leader monitors the experiment in the studio through the window. **B.** The small room where the participant is first brought into upon arrival, and later on the room for the confederate to hide in after the experiment had finished. **C.** The soundproof studio where the interaction takes place.

During the interaction. The experiment was built and run in MATLAB (*version* R2018b) and audio recordings were made during the whole interaction using a Roland R-09 Wave/MP3 Recorder. Before we explain the procedure of the experiment, we briefly describe the basic procedure of the ventriloquist paradigm (see Felker et al. (2018) for a more detailed description).

Participants interact face-to-face with a confederate, and together they take part in a cooperative computer-based task. The participant and the confederate are seated across each other diagonally, each facing their own computer screen, and use microphones and noise-cancelling headphones to communicate with each other. While the participant believes that the confederate is just another participant, the confederate actually interacts with the participant only through pre-recorded speech. Each time the confederate 'speaks', she leans forward to 'speak into her microphone' and thereby hides her face behind her computer screen. Then, she presses a key on a hidden numeric pad and a pre-recorded utterance is played through the headphones of the participant. In this way, it seems like the confederate is engaging in the task normally, while in fact her words have already been determined in advance.

Pre-test. The experiment started when the confederate pretended to be finished with her other task and played her introductory audio recording to introduce herself to the participant. After the participant and confederate had introduced themselves, the instructions for the first round of the Word or Not Game (i.e., the pre-test) appeared on the screen. The Word or Not Game was actually an interactive version of the Lexical Decision (LD) task, in which one person had to read out the words while the other person had to decide whether that word existed in English or not.

Participants were pseudo-randomly assigned to the production or the perception version of the Word or Not Game, controlling for gender to be balanced between the two groups. If participants were assigned to the *production* group (N = 20), then participants had to read out the words that appeared on the screen, while the confederate made the decisions. The items were presented one by one in the middle of participants' computer screen with an inter-trial interval (ITI) of 230 ms using MATLAB. Even though participants only had to read out the words, we decided to explain their role in the context of an interactive LD task anyway in order to make it sound more interesting and maintain the conversational goal of the task (that is usually stripped away from shadowing tasks). If participants were assigned to the *perception* group (N=20), then the roles were switched and the confederate 'read out' the words that appeared on her screen (i.e., played the pre-recorded audio files). The participant had to press the 'Y' button on the button box for real English words, and 'N' for non-existing English words. The pre-recorded audio files of the words and non-words were played automatically through the headphones after the participant pressed a button on the button box, with random ITI's ranging from 500 to 1400 ms. The pre-test took up about 12-15 minutes.

Interaction phase. After the last trial of the Word or Not Game, the instructions for the Code Breaker Game appeared on the screen. In the Code Breaker Game, the confederate saw a puzzle sequence and four words on her screen, while participants saw four coloured shapes with each having a word below it (see Fig. 2). The confederate described which coloured shape was missing in her puzzle sequence, and the participant's task was to name the word that was written below the respective shape, which in turn the confederate clicked on her screen. Participants were allowed to ask questions during the game, but were specifically instructed not to discuss the spelling and meaning of the word. Each time the confederate clicked on a word, a grey border appeared around the chosen word on the participants' screen, so they could see which word the confederate had picked. The Code Breaker Game took up around 20-25 minutes.

During the trials in which /æ/-/ɛ/ was the target sound contrast, the words below the correct puzzle answers always included the /æ/-vowel. So, participants always pronounced the /æ/ word from the minimal pair (e.g., '*bag*') during /æ/-/ɛ/ trials. The trials that contained word-final /t/-/d/ as target sound contrast always included target words that ended with /d/ (e.g., *'feed'*). Participants were randomly assigned to the /æ/-/ɛ/ or the /t/-/d/ condition of the Code Breaker Game. If participants were assigned to the /æ/-/ɛ/ condition, then their communication would be disrupted on all 12 trials with /æ/-/ɛ/ as the target sound contrast. Whenever participants pronounced target words containing /æ/, the confederate constantly picked the word with the /ɛ/-vowel instead (e.g., *'beg'*), regardless of how accurate the participant had pronounced it (Fig. 2A). Participants received negative feedback on their pronunciation of the /æ/-vowel 12 times out of the 64 trials. The communication was not disrupted when /t/-/d/ and the fillers were considered target sound contrasts, as for these 'control' trials the confederate always picked the correct target word (or whichever word the participant communicated to the confederate; Fig. 2B). If participants were assigned to the /t/-/d/ condition, then the communication was disrupted during all 12 of the 64 trials containing /t/-/d/ as target sound contrast. So, whenever participants pronounced word-final /d/ words, the confederate always clicked on the word ending in /t/ instead (e.g., *'feet'*). The communication was not disrupted during /æ/-/ɛ/ and filler trials; hence the negative feedback was only directed at the /t/-/d/ contrast and specifically the word-final /d/ sound.

Post-test. Finally, instructions to play another round of the Word or Not Game (i.e., the post-test) appeared on the screen after the last Code Breaker Game trial. This phase of the experiment also took up about 12-15 minutes. The participants and the confederate performed the same role they had during the pre-test, depending on in which group the participants had been assigned to (production or perception). The post-test included filler items that were different from the fillers of the pre-test, but the critical items remained the same.

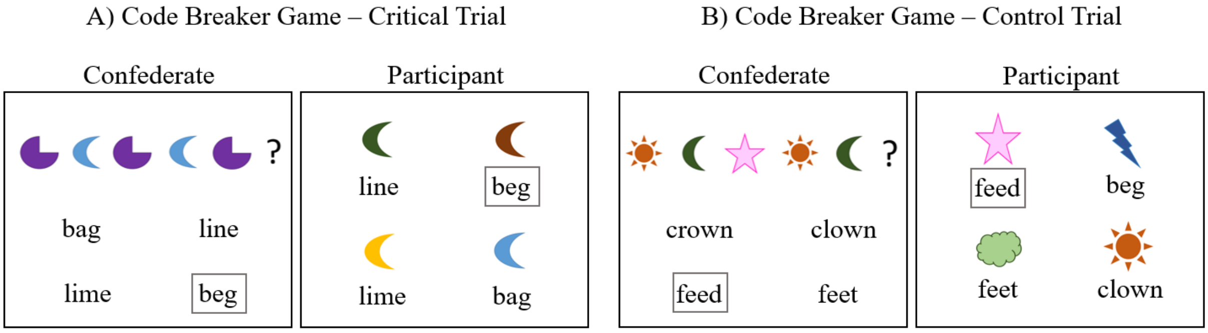

Fig. 2. Sample slides of the two players' screen from the Code Breaker Game illustrating the /æ/-/ɛ/ condition **A.** An example of a critical trial, in which the participant receives negative feedback on the pronunciation of the puzzle answer '*bag'*, as the confederate clicked on the word '*beg'*. **B.** An example of a control trial, in which the participant does not receive negative feedback on the pronunciation of the puzzle answer '*feed'*.

After the interaction. A screen indicating the end of the experiment was presented after the end of the post-test. The confederate then played an audio file indicating that 'her screen says she has to go to the other room', which was the experiment leader's cue to come into the sound-proof studio and pick her up. The screen of the participant included instructions that they should stay seated and wait for the experiment leader. Then the experiment leader guided the confederate to the small room for a 'final task', so participants did not have the chance to speak with the confederate. When the experiment leader returned, participants completed the English version of the LexTALE vocabulary test (Lemhöfer & Broersma, 2012) and filled in a questionnaire on their language background. The questionnaire also included some questions related to the interaction with the confederate. The whole experiment session took up about 1 hour.

All participants were debriefed after the experiment about the fact that the negative feedback they received during the Code Breaker Game was pre-programmed, and that the native English partner was actually a confederate. Only those who expressed some disbelief about the confederate's speech being real were also debriefed at that time about the use of pre-recorded speech in the experiment. After the data collection was completed, all participants were debriefed via email about the goal of the experiment and the use of pre-recorded speech.

RESULTS

Production

For the production group, we examined whether communication disruptions induced by the confederate's negative feedback concerning participants' production of the critical sounds lead to adaptations in the pronunciations of these sounds after the interaction.

The production analysis was only conducted on the critical *real* English words from the LD task, since the pronunciation of non-words is unpredictable and therefore irrelevant. The audio recordings from the experiment of the production group were cut into separate recordings of the pre- and post-test. Then, the speech data from each test was automatically segmented into separate recordings of all items and automatically aligned with its orthographic transcript using Praat (Boersma & Weenink, 2015). We used Montreal Forced Aligner (McAuliffe, Socolof, Mihuc, Wagner, & Sonderegger, 2017) to automatically align the speech recordings with their phonetic transcripts using the CMU Pronouncing Dictionary (Carnegie Mellon University, 2019), which contains the phonetic transcripts from over 134,000 American English words. Separate analyses were conducted for our two sound contrasts of interest. The first two formants (F1 and F2) and vowel duration were taken as the acoustic measures for the analysis of /æ/ and /ɛ/ productions. For the analysis of words ending with /t/-/d/, we took the amount of voicing as the acoustic measure.

/æ/-/ɛ/ contrast. *Formant analysis.* For the formant frequency analysis, we compared each participant's formant space to that of a 'model speaker' to see if participants' improve their pronunciations of /æ/ and /ɛ/ due to the interlocutor's implicit negative feedback. Since the exact same words were included in the perception and production experiment, we were able to use the audio files that were played in the perception version of the experiment as our model speaker data. Her speech data was also automatically segmented and aligned with its phonetic transcripts. For each participant and our model speaker, we automatically extracted F1 and F2 values (measured in Hz) from the /æ/ and /ɛ/ words using Praat (Boersma & Weenink, 2015). All formant values were then normalised in order to account for interspeaker variability, as factors such as gender, age but also individual differences in physiology and anatomy of the vocal tract have been shown to influence the formant frequencies (Ladefoged & Broadbent, 1957). This was done following Lobanov's (1971) method of formant frequency normalisation, described in the formula below:

$$F_n{}^N = \frac{F_n - \mu(F_n)}{\sigma(F_n)}$$

In other words, Lobanov normalises vowel formant frequencies (F_n) for each individual speaker by subtracting the mean frequency from the formant values of all vowels ($\mu(F_n)$) from it, and then dividing it by its standard deviation ($\sigma(F_n)$). The Lobanov transformed formant values were acquired with the *phonR* package in R (McCloy, 2012; R Core Team, 2019). We then calculated the Euclidean distance between each participant's and the model speaker's formant space in the pre- and post-test, using the *joeyr* package (Stanley, 2019). The Euclidean distance is the ordinal line between the participant's and model speaker's formant space, bigger distances indicate greater differences in formant space while smaller distances indicate more similarity. It was expected that the Euclidean distance would become smaller in the post-test for those participants who received implicit negative feedback on the /æ/-/ɛ/ contrast during the Code Breaker Game.

We used R and *lme4* (Bates, Mächler, Bolker, & Walker, 2015) to perform linear mixed-effects modelling (LMEM) on the data with Euclidean distance as our dependent variable. We started our modelling with including predictors that were most relevant to our theoretical hypothesis as our fixed effects, namely Test (pre vs. post), Phoneme (/æ/ vs. /ɛ/), Condition (control: feedback on /t/-/d/ vs. critical: feedback on /æ/-/ɛ/), and their interactions. Subjects and items were included as random effects. Results show that even after removing non-significant effects and interactions from our model one by one, none of the fixed effects nor any interactions affected the Euclidean distances between participants' and the model speaker's formant space significantly (see the full model in Table 2.). The *p*-values were obtained using the Sattherthwaite approximation for degrees of freedom from *lmerTest* (Kuznetsova, Brockhoff, & Christensen, 2017; Luke, 2017).

Table 2.

Summary of linear mixed-effects model results on the relationship between Euclidean distances and Test, Condition, Phoneme and their interactions as fixed effects. Significant effects are printed in **bold**.

Fixed effects:					
	Estimates	**SE**	***t***	**df**	***Pr*(>|t|)**
(Intercept)	1.201	0.118	10.141	51.16	**< .001**
Test Post	-0.059	0.082	-0.720	916.85	.472
Condition Critical	-0.158	0.112	-1.410	53.56	.164
Phoneme /æ/	0.076	0.149	0.510	39.82	.613
Test Post x Condition Critical	0.152	0.116	1.315	916.85	.189
Test Post x Phoneme /æ/	0.006	0.116	0.056	916.85	.955
Condition Critical x Phoneme /æ/	0.132	0.116	1.139	916.85	.255
Test Post x Condition Critical x Phoneme /æ/	-0.098	0.163	-0.599	916.85	.549

We also performed LMEM for the F1 and F2 separately, to examine the effect of interactional feedback on participants' /æ/ and /ɛ/ formants individually. The $F1_{Lobanov}$ was first entered as our dependent variable. The same three theoretical predictors (Test, Condition, Phoneme) and their interactions were included in the model as fixed effects, again with random intercepts for subjects and items. The model with the best fit showed a significant effect of Phoneme on the $F1_{Lobanov}$ values (β = 0.841, SE = 0.240, *p* = .030; see Table 3), as participants pronounce /æ/ vowels with higher F1 than the /ɛ/ vowels (see Fig. 3). The model also shows that Test type is significantly related to the $F1_{Lobanov}$ values (β = 0.177, SE = 0.052, *p* < .001), as both /æ/ and /ɛ/ vowels are generally pronounced with higher F1 in the post-test than in the pre-test. Modelling showed no significant interaction between Phoneme and Test or a main effect of Condition on the $F1_{Lobanov}$ values.

Table 3.

Summary of final linear mixed-effects model with $F1_{Lobanov}$ as dependent variable, and Test and Phoneme as fixed effects. Significant effects are printed in **bold**.

Fixed effects:

	Estimate	**SE**	***t***	**df**	***Pr*(>|t|)**
(Intercept)	-0.510	0.169	-3.019	3.64	**.044**
Test Post	0.177	0.052	3.381	930.98	< **.001**
Phoneme /æ/	0.841	0.240	3.569	3.47	**.030**

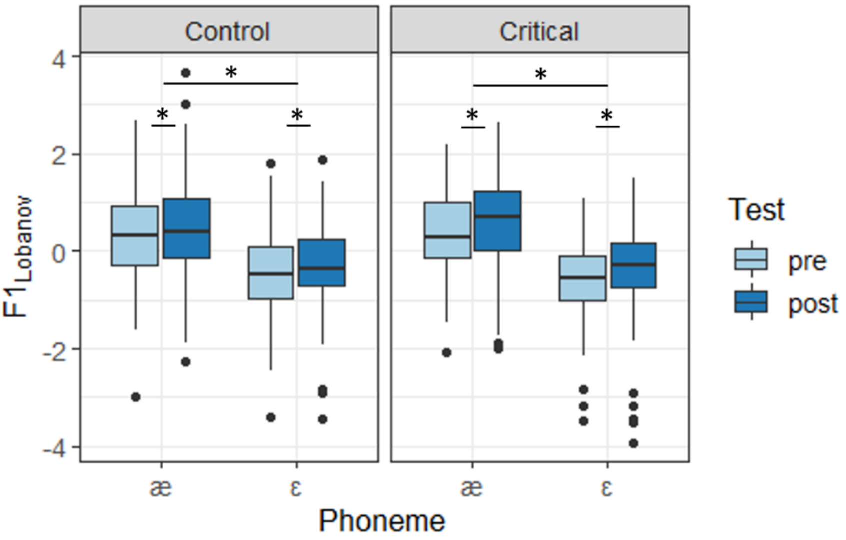

Fig 3. Boxplot displaying the Lobanov transformed F1 frequencies ($F1_{Lobanov}$) in the control (feedback on /t/-/d/) and critical (feedback on /æ/-/ɛ/) condition, plotted by Phoneme (/æ/ vs. /ɛ/) and Test (pre vs. post). The two vowels are pronounced for both conditions with higher F1 in the post-test, but German speakers of English pronounce the /æ/-vowel with even higher F1 than the /ɛ/-vowel (* indicates *p* < .05).

For the separate analysis on the F2, we included $F2_{Lobanov}$ as the dependent variable, and the same fixed and random effects mentioned before in the model. Results show that the F2 is also not affected by Condition. The final model including the $F2_{Lobanov}$ displayed no main effects of Test or Phoneme (both *p* > .05; see Table 4). However, the model demonstrated a significant interaction between Test and Phoneme: participants pronounce /æ/-vowels with lower F2 in the post-test compared to the pre-test while there are no differences between pre- and post-test for /ɛ/-vowels (β = 0.274, SE = 0.090, *p* < .01; see Fig. 4).

Table 4.

Summary of final linear mixed-effects model with $F2_{Lobanov}$ as dependent variable, with Test, Phoneme and their interaction as fixed effects. Significant effects are printed in **bold**.

Fixed effects:

	Estimate	**SE**	***t***	**df**	***Pr*(>|t|)**
(Intercept)	-0.049	0.191	0.256	23.84	.800
Test Post	0.003	0.069	0.047	936.58	.963
Phoneme /æ/	0.036	0.271	0.131	23.84	.897
Test Post x Phoneme /æ/	-0.274	0.098	-2.793	936.58	**< .010**

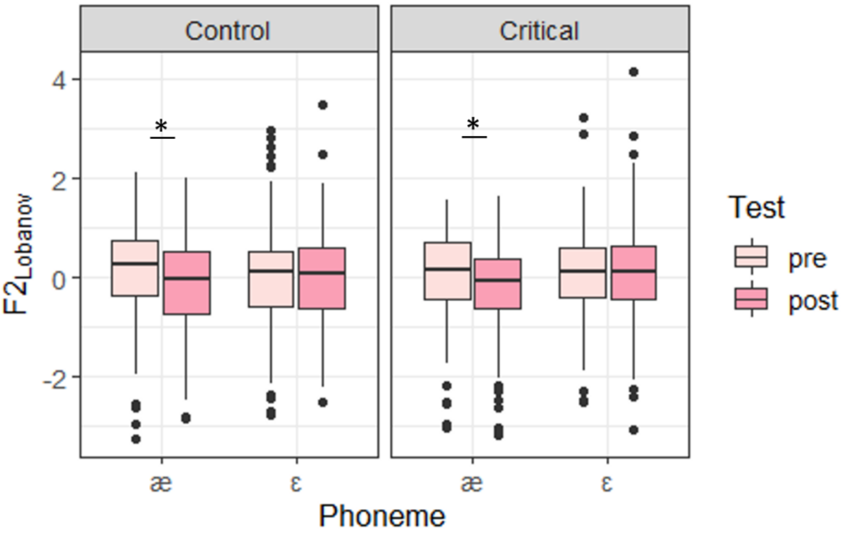

Fig. 4. Boxplot illustrating the Lobanov transformed F2 frequencies ($F2_{Lobanov}$) in the control (feedback on /t/-/d/) and critical (feedback on /æ/-/ɛ/) condition, plotted by Phoneme (/æ/ vs. /ɛ/) and Test (pre vs. post). German speakers of English from both conditions pronounce the /æ/-vowel with lower F2 in the post-test (* indicates *p* < .05).

In summary, the results from the vowel formants analysis demonstrated that none of the three theoretical predictors (Test, Phoneme and Condition) were related to the Euclidean distances between participants' and the model speaker's formant space. The independent formant analyses revealed that the F1 of /æ/ and /ɛ/ vowels is affected by Test and Phoneme, while only a significant interaction between Test and Phoneme was found for the F2. These results indicate that German speakers' pronunciation of the English /æ/-vowel becomes more native-like over time in terms of F1 and F2 separately; however, this improvement is not related to whether their communication was disrupted or not during the trials involving the /æ/-/ɛ/ contrast in the Code Breaker Game.

Vowel duration analysis. We also analysed the effect of communication disruptions on German's pronunciation of /æ/ and /ɛ/ by means of vowel duration. The durations (s) of /æ/ and /ɛ/ vowels were automatically extracted from each participant's speech data using Praat. Durations that were too far (i.e., more or less than 2.5 SD) from the mean were manually corrected in Praat.

We fitted another LMEM to the data with Duration as our dependent variable. The same three theoretical predictors and all interactions were entered as our fixed effects, while subjects and items were entered as random effects. The three-way interaction between Test, Phoneme and Condition turned out to be non-significant. No main effects of Condition and Phoneme were found, but crucially, both Condition and Phoneme significantly interacted with Test. The interaction between Test and Condition indicates that the durations of /æ/ and /ɛ/ in the post-test are increased in the critical condition compared to the control condition (β = 0.014, SE = 0.004, *p* < .001). In other words, the participants who received negative feedback on the /æ/-/ɛ/ contrast pronounced both vowels longer in the post-test compared to those did not receive negative feedback on the /æ/-/ɛ/ contrast during the Code Breaker Game. In order to know whether there is a difference in duration between the two vowels, we have to look at the interaction between Test and Phoneme. This interaction shows us that the durations in the post-test are different for the two sounds: both vowels are pronounced with longer durations, but the duration for /æ/ is longer compared to /ɛ/ (β = 0.008, SE = 0.004, *p* < .05). In short, the disrupted communication during the /æ/-/ɛ/ trials of the Code Breaker Game seemed to have affected German speakers' pronunciation of the two vowels in terms of duration.

Table 5.

Summary of final linear mixed-effects model with Duration (s) as dependent variable, with Test, Condition, Phoneme and the two-way interactions as fixed effects. Significant effects are printed in **bold**.

Fixed effects:					
	Estimates	**SE**	***t***	**df**	**Pr(>|t|)**
(Intercept)	0.127	0.013	9.971	36.75	**< .001**
Test Post	-0.005	0.003	-1.439	913.00	.150

Condition Critical	0.003	0.014	0.229	19.24	.822
Phoneme /æ/	0.022	0.011	1.910	24.55	.068
Test Post x Condition Critical	0.014	0.004	3.782	913.00	**< .001**
Test Post x Phoneme /æ/	0.008	0.004	2.205	913.00	**.028**
Condition Critical x Phoneme /æ/	-0.010	0.004	-2.585	913.00	**< .010**

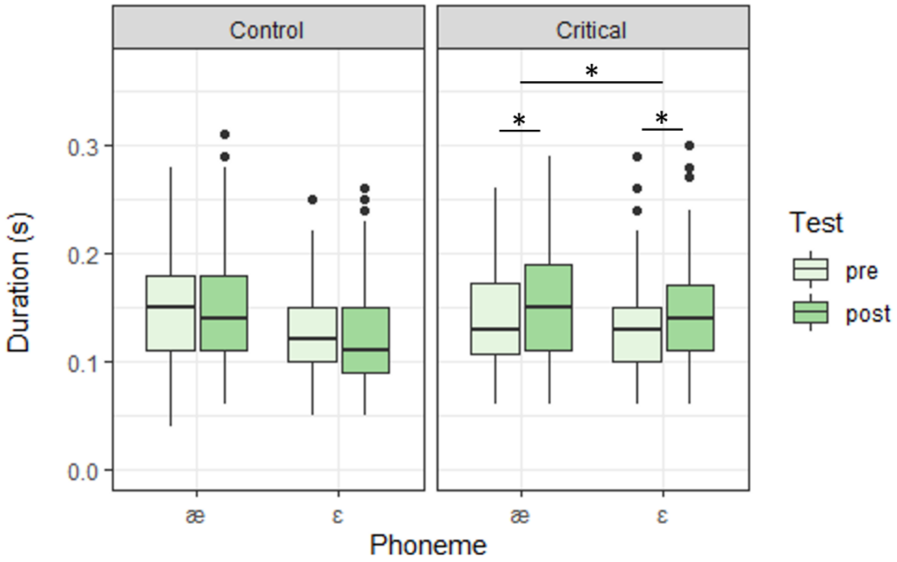

Fig. 5. Boxplot illustrating the vowel durations (s) in the control (feedback on /t/-/d/) and critical (feedback on /æ/-/ɛ/) condition, plotted by Phoneme (/æ/ vs. /ɛ/) and Test (pre vs. post). The figure shows that the duration for both vowels are increased in the post-test of the critical condition, and that the /æ/-vowel is pronounced longer than the /ɛ/-vowel (* indicates *p* < .05)

/t/-/d/ contrast. We decided to take the amount of voicing as our acoustic measure for the analysis of the word-final /t/-/d/ contrast. We automatically extracted the 'fraction of locally unvoiced frames' from word-final /t/ and /d/ segments with Praat's Voice Report function, using the same parameter settings that were mentioned in Eager (2015). These fractions indicate the amount of voiceless parts in a specific segment, with 1 indicating a complete voiceless segment while 0 indicates that there are no voiceless fragments within the segment. We are aware of the fact that it is more common practice to examine the speech processing of stop consonants by analysing the durations of vowels preceding these stop consonants (e.g., Port & O'Dell, 1985; Smith et al., 2009; Warner, Jongman, Sereno, & Kemps, 2004), as vowel duration has been shown to be a reliable cue for distinguishing voiced and voiceless stop consonants (Raphael, Dorman, Freeman, & Tobin, 1975). However, the stop consonants were not immediately preceded by vowels for the majority of our items containing the word-final voicing contrast. Therefore, we were unable to use vowel duration as a reliable measure to investigate the productions of words ending with /t/ or /d/ and analysed the amount of voicing instead.

We initially ran a generalised linear mixed-effects model (GLMM) to analyse our data, since we are dealing with proportional data. However, we decided to switch to a LMEM instead after encountering multiple issues with the GLMM. Unfortunately, the GLMM failed to converge, as it indicated that we were overfitting the data. This could be due to low power (which we will return to in the 'Discussion' section) or to the fact that the variance of the random effect 'Item' was close to 0, indicating that our model structure was too complex for the data. Even though the GLMM would theoretically be the most appropriate statistical test for our binomially distributed data, we cannot rely on the output without accounting for the item-variation in our data as we would then violate the most important assumption of performing mixed-effects models, namely the independence assumption (Winter, 2013). In addition, linear models have shown to robustly handle lack of normality in data (Winter, 2013).

We therefore decided to perform another LMEM on the /t/-/d/ data, with the fraction of unvoiced frames as our dependent variable. Like in our other models, we included the three predictors Test (pre vs. post), Phoneme (/t/ vs. /d/) and Condition (control: feedback on /æ/-/ɛ/ vs. critical: feedback on /t/-/d/) as our fixed effects, with subjects and items as random effects. Note that the levels from 'Condition' now refer to the trials from the Code Breaker Game that included word-final /t/-/d/ as target sound contrast. The final model (Table 6) shows that there is a main effect of Phoneme on the fraction of locally unvoiced frames. Overall, participants pronounce words that end with /t/ more devoiced than words ending with /d/ (β = -0.250, SE = 0.022, *p* < .001), indicating that German speakers of English are able to make the distinction between /t/ and /d/ in terms of voicing. The model also shows a significant interaction between Phoneme and Test: The /d/ has less portions of unvoiced frames (i.e., is pronounced more voiced) in the post-test than in the pre-test, whereas there is no difference in the number of unvoiced frames between pre- and post-test for words ending with /t/ (β = -0.065, SE = 0.021, *p* < .01; see Fig. 6). Importantly, results show that these results are independent of Condition as the model shows that disruption of communication is not significantly related to the voicing of /t/ and /d/ sounds.

These results indicate that German speakers of English are able to distinguish word-final /t/ from /d/ in their production in terms of voicing and improve this distinction over time, but that they accomplish this without the aid of negative feedback on their production of the word-final /d/ sound in the Code Breaker Game. This is contrary to the /æ/-/ɛ/ contrast, which showed that negative feedback on the pronunciation of the /æ/-vowel does improve participants' distinction between the two sounds.

Table 6.

Summary of the final linear mixed-effects model with the 'fraction of locally unvoiced frames' as dependent variable, with Test, Phoneme and their interaction as fixed effects. Significant effects are printed in **bold**.

Fixed effects:

	Estimate	SE	*t*	df	*Pr*(>|t|)
(Intercept)	0.897	0.025	36.603	39.94	**< .001**
Test Post	-0.009	0.015	-0.611	915.89	.541
Phoneme D	-0.250	0.022	-11.389	38.57	**< .001**
Test Post x Phoneme D	-0.065	0.021	-3.156	915.91	**.002**

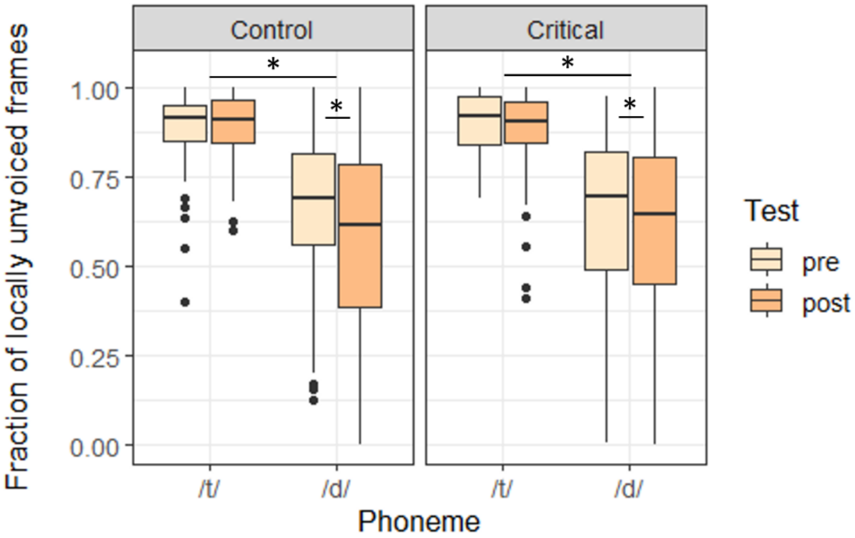

Fig. 6. Boxplot illustrating the amount of voicing ('fraction of locally unvoiced frames) in the control (feedback on /æ/-/ɛ/) and critical (feedback on /t/-/d/) condition, plotted by Phoneme (/t/ vs. /d/) and Test (pre vs. post). German speakers from both conditions pronounce the final /d/ sound more voiced in the post-test, whereas the amount of voicing for the /t/ sound was not altered (* indicates *p* < .05).

Perception

We also wanted to investigate whether disruptions in communication, by means of interactional feedback addressing participants' production abilities, can lead to changes in the perception of the two sound contrasts. For the analysis of the perception experiment, responses to both real words *and* non-words from the auditory LD task containing the four critical sounds were included in the analysis. Participants' responses to these items were rated on accuracy.

Figure 7 displays participants' percentage of "yes" responses to the /æ/-/ɛ/ items of the LD tasks for the control (feedback on /t/-/d/) and critical (feedback on /æ/-/ɛ/) conditions. In the pre-test of the *control* condition, 93.0% of the real words containing /æ/-/ɛ/ were correctly accepted as real English, while the acceptance rate in the post-test was 97.0% for real words. However, the non-words including /æ/-/ɛ/ were often perceived as real words: 73.8% of the non-words were accepted as real words in the pre-test and 72.9% in the post-test. The acceptance rate of the /æ/-/ɛ/ items from those in the *critical* condition showed extremely similar patterns: the real words were accepted in 92.4% of the cases for the pre-test and 95.4% for the post-test, whereas 72.8% of the non-words were accepted as real words in the pre-test compared to 73.3% in the post-test. The similar patterns in "yes" responses between the two conditions suggest that the perception of the two sounds is not affected by Condition (i.e., receiving negative feedback on their pronunciation of the /æ/-vowel or not).

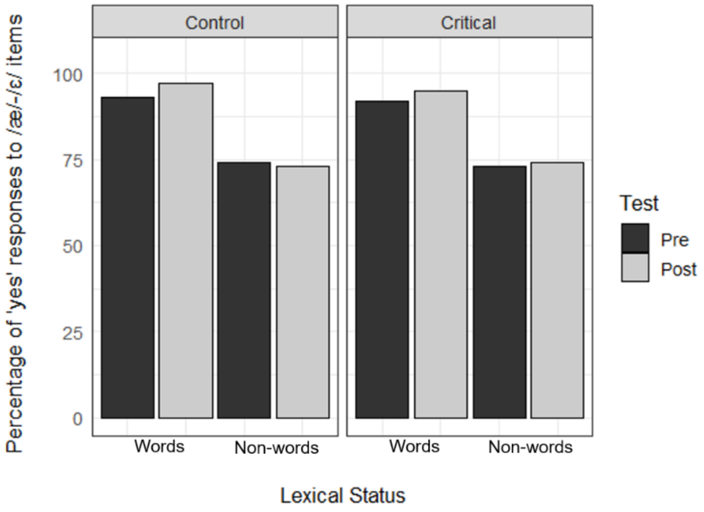

Fig. 7. Percentage of 'yes' responses to the /æ/-/ε/ items of the Lexical Decision tasks for the control (feedback on /t/-/d/) and critical (feedback on /æ/-/ε/) conditions, plotted by Lexical Status (words vs. non-words) and Test (pre vs. post).

Figure 8 displays participants' percentage of "yes" responses to the /t/-/d/ items of the LD task for the control (feedback on /æ/-/ε/) and critical (feedback on /t/-/d/) conditions. Similar to the acceptance rates of /æ/-/ε/ items, participants also showed high acceptance rates for all the /t/-/d/ items, indicating that negative feedback does not influence the perception of words and non-words ending with /t/ or /d/. In the control condition, participants accepted 97.1% of the real words as existing English words in the pre-test and 97.9% in the post-test. The non-words were accepted as real words in 83.3% and 80.0% of the cases for the pre- and post-tests, respectively. In the critical condition, participants accepted 96.3% (pre-test) and 97.9% (post-test) of the real /t/-/d/ words as existing English words, while 82.9% (pre-test) and 82.1% (post-test) of the non-words were perceived as real words.

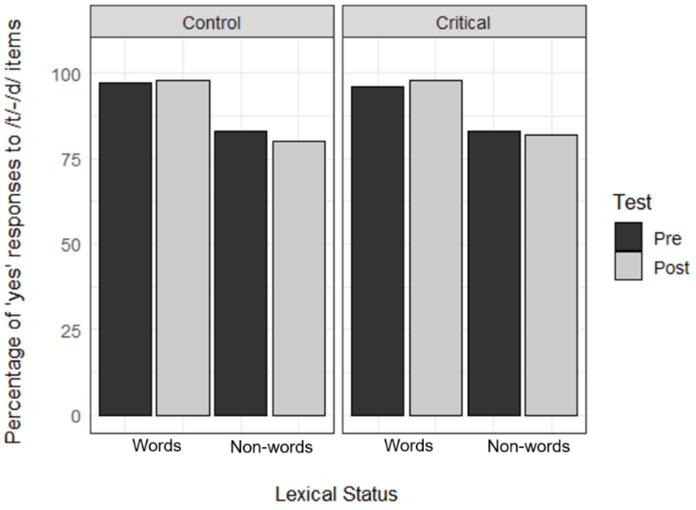

Fig. 8. Percentage of ‘yes’ responses to the /t/-/d/ items of the Lexical Decision tasks for the control (feedback on /æ/-/ɛ/) and critical (feedback on /t/-/d/) conditions, plotted by Lexical Status (words vs. non-words) and Test (pre vs. post).

We used *lme4* to fit a GLMM (family = binomial) to analyse whether participants’ perception of the four critical sounds changed over time due to negative feedback on their production of either the /æ/-/ɛ/ contrast or the word-final /t/-/d/ contrast. Accuracy was included as our dependent variable, while the predictors Test (pre vs. post), Phoneme (/æ/, /ɛ/, /t/, /d/), Feedback (/æ/-/ɛ/ vs. /t/-/d/) and their interactions were entered as fixed effects. We also included random intercepts for subjects and items. Results show that besides a main effect of Test (β: 0.229, SE: 0.113, *p* = .042), none of the other fixed effects nor any of the interactions significantly affect participants’ accuracy on the auditory LD task (see Table 7). This shows that participants perform slightly better in the post-test of the LD task overall, but that there is no relationship between the perception of the four sounds and whether participants received negative production feedback on the sound contrast or not.

Table 7.

Summary of generalized linear mixed-effects model output of the relationship between Accuracy and Phoneme, Test and Feedback as fixed effects. Significant effects are printed in **bold**.

Fixed effects:

	Estimates	**SE**	***z***	***Pr*(>|z|)**
(Intercept)	1.490	0.892	1.671	.095

Phoneme /ɛ/	0.025	1.169	0.022	.982
Phoneme /t/	-0.199	1.173	-0.170	.865
Phoneme /d/	-0.087	1.174	-0.074	.941
Test Post	0.229	0.113	2.031	**.042**
Feedback /t/-/d/	-0.045	0.475	-0.096	.924

DISCUSSION

The present study investigated the effect of the interlocutor’s negative feedback that concerns the learner’s production of non-native sounds on L2 speech perception and production processes. German speakers of English were tested on their production and perception of a problematic English vowel and consonant contrast before and after they engaged in interaction with a confederate. During this interaction, the communication was systematically disrupted, as participants received implicit negative feedback from the interlocutor on their pronunciation of words containing the /æ/-/ɛ/ contrast or the word-final /t/-/d/ contrast.

More native-like productions of L2 sounds due to interactional feedback

The first aim of the study was to examine whether communication disruptions, induced by the interlocutor’s implicit negative feedback, led to more native-like productions of the two English sound contrasts in the post-test. The results from the production part of the experiment showed that communication disruptions changed German speakers’ production of the vowel contrast in terms of the formants individually and duration, and the word-final consonant contrast in terms of voicing. The analysis on the /æ/-/ɛ/ vowel contrast revealed that, even though the F1 was increased for both vowels after the interaction, the /æ/-vowels were pronounced with lower height than the /ɛ/-vowels. Furthermore, the F2 was decreased for the /æ/-vowel in the post-test, hinting towards a more fronted pronunciation, while the F2 for the /ɛ/-vowel did not change. At last, the durations for both vowels were increased in the post-test, but participants pronounced the /æ/-vowel significantly longer than the /ɛ/-vowel. The analysis on the word-final /t/-/d/ contrast showed that participants pronunciation of words ending with /d/ contained more voiced segments in the word-final consonant than the pronunciation of words ending with /t/. The findings confirm our hypothesis that communication disruptions, by means of the interlocutor’s feedback on German speakers’ L2 production, can alter the pronunciations of non-native sounds, leading to more native-like productions of the L2 sounds.

We also predicted that improvements would mainly occur in the pronunciation of non-native sounds belonging to the sound contrast that participants received negative feedback on during the interaction phase. The interlocutor’s negative feedback on the pronunciations of /æ/-vowels did not influence German speakers’ pronunciations of the /æ/ and /ɛ/ vowels in the formant dimension, as measured by the Euclidean distance and independent formant analyses. On the other hand, the interlocutor’s negative feedback turned out to be beneficial for the distinction between /æ/ and /ɛ/ vowels in terms of duration, as only those Germans who received negative feedback on the /æ/-/ɛ/ contrast pronounced the /æ/-vowel with increased vowel duration compared to the /ɛ/-vowel. However, no difference was found between the pronunciations of word-final /t/ and /d/ from Germans who received negative feedback on the /æ/-/ɛ/ contrast and those who received negative feedback on the word-final /t/-/d/ contrast. Both groups equally pronounced word-final /d/ more voiced than word-final /t/ in the post-test.

To summarize, the production findings suggest that disruptions in communication can indeed lead to adaptions in the production of L2 sounds, however, the type of sound contrast that is addressed in the interlocutor’s feedback is only relevant for the pronunciation of the /æ/ and /ɛ/ vowels. The results from the /æ/-/ɛ/ contrast are in line with our prediction that production improvements are related to the content of the interlocutor’s feedback, however, the findings from the word-final /t/-/d/ contrast do not confirm this. We discuss three possible explanations for the discrepancy between the two sound contrasts.

The first possible explanation for the different effects found for the two sound contrasts could be that the /æ/-/ɛ/ contrast is more difficult for German speakers of English than the word-final /t/-/d/ contrast. Even though it has been shown that Germans struggle with the word-final voicing contrast in stop consonants, they are able to produce them well in word-initial position for instance (Broersma, 2002; Port & O’Dell, 1985; Smith et al., 2009), indicating that they are already capable of distinguishing both sounds in their production. This is probably due to the fact that the two sounds are present in the German phonological inventory, whereas the /æ/-vowel does not exist in German (e.g., Bohn & Flege, 1992). The German speakers are therefore not dealing with an entire new phonological category for the word-final /t/ and /d/ sounds, but rather have to adapt to new phonological rules to already established representations when they want to speak in English. For this reason, it could be that the communication disruptions themselves, regardless of which sound contrast the communication disruption concerned, was already enough for the German speakers to change their pronunciations of words containing the word-final voicing contrast.

Secondly, it could also be that participants generally became more conscious of their L2 productions because of the communication disruptions. Participants either received feedback on the /æ/-/ɛ/ contrast or the word-final /t/-/d/ contrast, but the study did not involve a condition in which participants did not receive negative feedback at all. Therefore, all participants were exposed to the same amount of communication disruptions and negative feedback on their production, which could have increased their overall alertness during the Code Breaker Game. In order to understand the exact role of communication disruptions themselves in L2 sound learning, futures studies could incorporate a ‘true’ control condition, in which the communication between the L2 learner and the interlocutor is *not* disrupted.

Lastly, another explanation for the different effects between the two sound contrasts could be that the orthographic transcripts of the word-final /t/-/d/ minimal pairs in the Code Breaker Game may have raised sufficient awareness for the voicing contrast. Escudero, Hayes-Harb, & Mitterer (2008) demonstrated that L2 learners who were only auditorily exposed to L2 words had more difficulty distinguishing the difficult non-native sound contrast than those who were auditorily and orthographically exposed to novel L2 words containing the sound contrast. The finding suggest that L2 learners use orthographical differences between words as a cue to distinguish phonetically-similar sounds in their pronunciations. Hwang et al. (2015) also found that L2 speakers were better able to distinguish two phonetically-similar sounds in their pronunciations when both members of the minimal pair were orthographically presented on the screen, even though they only investigated pronunciations during the interaction itself. The orthographic transcripts of the word-final /t/-/d/ minimal pairs in the Code Breaker Game could therefore have been an indication for participants to disambiguate words ending with /t/ or /d/ in their pronunciations. However, all participants were also visually exposed to the orthography of /æ/-/ɛ/ minimal pairs during the Code Breaker Game, and for this contrast we clearly see an effect of feedback on the production of the two vowels.

We believe that disruptions in communication (by the interlocutor's negative feedback on production) is only beneficial when L2 learners have not fully established new phonological categories yet in their phonological inventory. As mentioned before, the /t/ and /d/ sounds exist in German (but the /d/ is used differently in word-final position for English) so it may be the case that the learner already had separate phonological categories for these sounds. The /æ/-vowel does not exist in German, and as a consequence, the English /æ/ and /ɛ/ vowels could have been mapped onto the same German phoneme that is most similar to these sounds in terms of acoustic or articulatory features, as predicted by PAM (Best & Tyler, 2007). Even though participants might visually notice the difference between *'bag'* and *'beg'*, their phonological representations of the two sounds are originating from the same native category, restricting them from distinguishing the two sounds in their pronunciation. The disruption in communication could be seen as a (more) clear indication of a phonological difference between /æ/ and /ɛ/, raising more awareness for it and eventually leading to altered pronunciations.

No transfer effects from L2 speech production to perception during interaction

The second aim of our study was to gain more insight into the L2 speech perception and production link, making us one of the first studies to investigate the perception-production relationship in an interactive context. We examined whether communication disruptions, induced by the interlocutor's negative feedback on the L2 learners' production of the two sound contrasts, led to improvements in the perception of the critical sounds. The results from the perception version of the experiment showed that German speakers slightly improved their perceptual distinction of the four critical sounds in the post-test compared to the pre-test. However, no differences were observed between the perception of the four sounds themselves, and between participants that received feedback on the /æ/-/ɛ/ contrast and participants with feedback on the word-final /t/-/d/ contrast. The overall high false-positive rates suggest that German speakers of English find it difficult to distinguish both /æ/-/ɛ/ and word-final /t/-/d/ sounds in the perceptual domain, as they accepted more than three-quarters of all the non-word items containing the critical sounds as real English words.

The fact that German speakers showed improved perception of the critical sounds in the post-test could suggest that communication disruptions can lead to adaptations in L2 perception. However, the effect was only marginally significant and no differences were found for the other two predictors. Therefore, we cannot claim with certainty that communication disruptions themselves can result in altered L2 perception abilities, as the main effect of Test could also indicate that participants simply became better at the LD task over time.

Nonetheless, the findings clearly indicate that the interlocutor's feedback addressing the learner's production of the critical sounds does not lead to adaptations in the perception of these sounds, suggesting that learning effects in the production domain do not cross-over to the perceptual domain. Our findings are in line with previous studies that also failed to find a direct relationship between L2 speech production and perception representations (Baese-Berk & Samuel, 2016; Thorin et al., 2018). The results do not confirm PAM's prediction that phonological representations are shared between speech production and perception processes (Best & Tyler, 2007), because otherwise we would have seen adaptations in German speakers' L2 perception. We suggest four possible reasons for the lack of transfer effects found in our study.

Firstly, we may have not found any evidence for the link between L2 speech perception and production, because we investigated the effect of production feedback on speech perception, whereas most studies that demonstrated a positive transfer effect examined the reverse direction, namely the role of perceptual training on L2 production (e.g., Bradlow et al., 1999, 1997; Wang et al., 2003). This could indicate that it generally may be harder to study the crossover effect from production to perception than vice versa.

Secondly, the lack of transfer effects could also be attributed to the use of an implicit method in our study to provide production feedback. The previous studies that did find positive transfer effects from production to perception all included explicit production training methods (Herd et al., 2013; Kartushina et al., 2015; Linebaugh & Roche, 2013, 2015; Wong, 2014), while studies that investigated the effect of implicit production training on L2 perception did not find evidence for a direct link between speech production and perception processes (Baese-Berk & Samuel, 2016; Thorin et al., 2018), including the current study. We decided to study the effect of production feedback on L2 perception in an implicit way, because we believe that implicit methods are more relatable to L2 learning in real-life than explicit methods, as learners do not have access to explicit information about the acoustic properties of non-native sounds outside the experimental setting.

Along the same lines, another reason for the lack of transfer effects could be that the interactive context of the study, which makes our research particularly relatable to L2 learning in real-life, may have hindered the perceptual learning process. As participants completed the experiment while they were interacting with the confederate, our study could have required participants to use more cognitive resources compared to studies that employ simpler computer tasks that can be done in isolation. Our results demonstrate the importance of studying L2 learning in more realistic settings such as conversational interactions, because it shows that more factors play a role in L2 learning than usually accounted for in laboratory settings.

Lastly, we believe that the lack of perceptual improvements could possibly be due to the drawback that the perception version of the pre- and post-test contained less communicative relevant reasons for the participants to perform well than the production version. While participants from the production group believed that pronunciations of the LD items served as the auditory input for the confederate, the performance of the perception group (i.e., the perceptual judgments) had no consequences for the confederate at all. The difference in communicative factors between the two types of pre- and post-tests support the view that communicatively relevant reasons may play an important role in L2 sound learning (Hwang et al., 2015).

While our results do not confirm PAM's hypothesis that L2 speech production and perception representations are shared, the results also do not provide direct evidence for SLM's prediction that perceptual phonological representations are at the basis of production abilities (Flege, 1995). Even though the results suggest no bidirectional relationship between L2 speech production and perception, we cannot conclude whether perception really precedes production because the interlocutor's feedback was specifically meant to induce changes in the *production* domain rather than the perceptual domain. Therefore, we cannot conclude whether the lack of transfer effects to the perceptual domain conform to SLM's hypothesis, or are due to other reasons like the demands of the experiment.

The role of error detection and awareness in L2 sound learning

The results from the production part of the experiment are partially in line with earlier studies that found instant adaptations of pronunciations in native speakers who were being misunderstood by the interlocutor (Buz et al., 2016; Schertz, 2013), suggesting that error detection may have raised their awareness for the phonological difference between two phonetically-similar sounds. The present study extended previous literature on the role of awareness in SLA by studying L2 speakers and letting them participate in a more naturalistic type of interaction. Our findings demonstrate, first of all, that *L2 speakers* can also adapt their pronunciations of phonetically-similar sounds due to the interlocutor's negative feedback, and secondly, that this can also happen during *human-human* interaction. The fact that German speakers overall improved their pronunciations of the four critical sounds after they received negative production feedback from the interlocutor could indicate that error detection may indeed be an important mechanism for L2 sound learning. However, the discrepancy between the results from the two sound contrasts of the production group and the lack of perceptual learning in the perception group raise several questions for our proposal that error detection leads to increased awareness of phonological details in the L2.

As discussed earlier, the interlocutor's feedback related to a target sound contrast only affected the distinction between the /æ/ and /ɛ/ vowels, whereas for the word-final /t/ and /d/ sounds it did not matter on which sound contrast the feedback was directed at as both feedback groups equally distinguished word-final /d/ from /t/. In that respect, our results provide evidence for the usefulness of the interlocutor's feedback targeted at phonological contrasts that have not been fully established yet in the L2 sound inventory, but do not clearly indicate how error detection leads to improved pronunciations of 'easier' L2 sounds. Error detection could still play a role in the adaptations of existing phonological categories, as it could lead to overall alertness for the phonetic details of the L2 for example, but future studies could examine the L2 productions from participants whose communication was not disrupted during the interaction (in a 'true' control condition) to investigate the exact role of error detection in learning 'easier' phonological contrasts.

The lack of perceptual learning for both sound contrasts during the perception part of the experiment also questions the exact role of error detection in L2 sound learning. Currently, the findings suggest that error detection may increase the awareness of phonological details in the L2 and lead to altered representations, but only in the modality in which the errors occurred. However, the results do not inform us whether error detection of L2 productions is really unhelpful for the perception of L2 sounds, nor do the findings totally exclude the possibility that error detection in one domain can lead to learning effects in the other domain.

For example, it may be the case that participants did became aware of the contrast between two non-native sounds during the interaction phase, but failed to improve their perceptual performance on the post-test due to the high attentional demands of the pre- and post-tests. The fact that we investigated participants perception and production of the L2 sounds via a pre-post-test design allowed us to base our conclusion on generalisation effects, since the critical items were different between the interaction phase and the pre- and post-tests, but generally could have affected participants' concentration levels as the tests took quite some time.

Thorin et al. (2018) also relate the presence of crossover effects in their study, but its absence in the study of Baese-Berk & Samuel (2016) to a difference in cognitive load of the experimental tasks. Whereas in the first study participants produced words *after* the perceptual judgment, participants from the other study were required to produce words *before* they were making the perceptual judgment. Therefore, the task design of a study could constrain the learning mechanism.

Nevertheless, our results provide new insights into of the role of awareness in SLA. We also contribute to the existing literature by studying the relationship in the phonological domain, as previous theories have mainly been focused on L2 grammar acquisition (Schmidt, 1990; Tomlin & Villa, 1994).

Study limitations and implications

Our results increase the understanding of the link between L2 speech production and perception and pose important implications for future studies of SLA. However, we are also left with unanswered questions partially due to the study's limitations.

Three of our limitations are caused by the time constraints of the research project. The first limitation concerns the amount of negative feedback participants received on their productions of the target sounds in the Code Breaker Game. Participants only received negative feedback on their production from the interlocutor during 12 of the 64 trials in the Code Breaker Game. While we had to make a trade-off between the naturalness of the interaction and providing sufficient feedback to induce learning, we believe that the relatively low number of 'learning opportunities' during the interaction may have been enough to evoke adaptations in the production but insufficient for the perception domain.

Secondly, the number of critical items in the pre- and post-tests was also relatively low compared to the total number of items in each test, as only 25% of the items were considered critical for the perception group, while for the production group only 12.5% of the items were critical (since we only analysed productions of the real English words). We decided not to include more critical items in the pre- and post-tests, because we wanted to conceal the presence of the two critical sound contrasts as much as possible. However, by doing so, we also decreased the power of our study. As mentioned in the analysis of word-final /t/ and /d/ productions (see the '/t/-/d/ contrast' subsection of the Results), our initial analysis with the GLMM indicated that we were perhaps overfitting our data, causing us to analyse the production /t/-/d/ results with the LMEM instead.

Finally, the relatively low number of participants in our research due to time constraints also form a limitation of the study and additionally resulted in the low power of the study. Participants were divided between the perception and the production group, and were also divided between the two sound contrasts, resulting in only 10 participants per experimental group. In the future, we could perhaps increase the power of the study by incorporating a within-subjects design in order to deal with less experimental groups, in which L2 production could be studied by analysing the pronunciations of the items during the Code Breaker Game instead of analysing it via pre- and post-tests performance. As mentioned before, we decided to analyse the pronunciations of items in the pre- and post-tests and not during the interaction to make sure learning effects are also generalisable to new words. However, this would still be possible since participants only get one chance in the Code Breaker Game to pronounce the target words correctly.

A similarity between the three limitations mentioned above besides timing issues is that they are also partially caused by the fact that we investigated two sound contrasts. For instance, if we focused the study on one of the two sound contrasts instead, we would have been able to lower the total number of items and increase the number of critical items in the pre- and post-tests, and have two experimental groups with a higher number of participants. However, we purposely decided to investigate two sound contrasts rather than one, and consider this as one of the strengths of the present study. By studying and comparing two sound contrasts, we have the possibility to evaluate whether communication disruptions and interactional feedback affect the representations of different phonological distinctions to the same extent, which has not been the case for our study.

The study also contained limitations due to other factors than time constraints. For example, the interlocutor's feedback was always corrective, regardless of how accurate participants pronunciations were. Participants could have been confused by the interlocutor's negative feedback, as the confederate pretended to misunderstand the participant during all 12 critical trials. This could have led to the lack of perceptual improvements in the study, as participants were constantly being told that they were wrong in their production without having the chance to correct themselves. However, the fact that we were able to find learning in the production domain could be an argument against the notion that the constant corrective feedback affected the perceptual learning process negatively.

Yet, we believe it would be interesting for future studies to investigate whether stronger beneficial effects of communication disruptions would occur when participants get the opportunity to correct themselves during the interaction. Previous studies have shown that native speakers are able to instantly adapt their pronunciations when they clarify misunderstood speech during human-computer interactions (Burnham et al., 2010; Buz et al., 2016; Schertz, 2013). However, this has not yet been demonstrated for L2 speakers and more importantly, during human-human interactions.

The present study has several implications for future studies concerning SLA and the L2 perception-production interface. We highly recommend to study any L2 sound learning process in more naturalistic settings of language learning in order to take aspects such as social or communicative factors into account that are also present in real-life, but often absent in the laboratory. Furthermore, as previous studies showed that speakers elicit different speech towards computer partners then human interlocutors (Burnham et al., 2010; Oviatt et al., 1998), we suggest to study L2 speech processing during human-human interaction so the results are more generalisable to realistic contexts of L2 learning. We have demonstrated with the use of the ventriloquist paradigm (Felker et al., 2018) that it is possible to have full control over the phonetic input while maintaining the ecological validity of the interaction.

Conclusion

In conclusion, the current study did not demonstrate a direct link between L2 speech perception and production processes in an interactive context. We were able to demonstrate that communication disruptions by means of interactional negative feedback on L2 learners' production enhances the pronunciation of L2 sounds, as German speakers' pronunciations of the vowel and consonant contrasts became more native-like after the interaction. The results provide evidence for the role of error detection in L2 sound learning, as error detection could be a mechanism to raise the awareness for phonological details in the non-native language that otherwise may not have come to the learner's attention. However, the nature of the learning effect is dependent on the degree of difficulty of the respective phonological distinction, suggesting that the interactional feedback is only beneficial when L2 learners have yet to establish phonological representations of the non-native sounds in their L2 sound inventory.

ACKNOWLEDGEMENTS

I would like to express eternal gratitude towards Aurora Troncoso-Ruiz and Emily Felker for their daily supervision during the master’s research project. Without their expertise, encouragement but foremost their kindness, I would not have been able to complete such a complicated yet fun and original study. Furthermore, I want to express many thanks to the research assistants who played the role of the experiment leader during the experiment. Without their convincing performance, we would not have been able to achieve our high success rate for the current version of the ventriloquist paradigm. I would also like to thank Femke de Groot from Radboud in’to Languages who was willing to help us recruit native German speakers for our study. At last, I would like to thank my main supervisor Dr. Mirjam Broersma, who was willing to give me the opportunity to conduct my master’s research project in her research group, as well as giving me a lot of freedom and authority for the research project.

REFERENCES

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

Bates, D., Mächler, M., Bolker, B. M., & Walker, S. C. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1).

Best, C. T., McRoberts, G. W., & Sithole, N. M. (1988). Examination of perceptual reorganization for nonnative speech contrasts: Zulu click discrimination by English-speaking adults and infants. *Journal of Experimental Psychology: Human Perception and Performance*, *14*(3), 345–360.

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

Boersma, P., & Weenink, D. (2015). Praat: doing phonetics by computer. *Retrieved from Http://Www.Praat.Org/*.

Bohn, O. S., & Flege, J. E. (1990). Interlingual identification and the role of foreign language experience in L2 vowel perception. *Applied Psycholinguistics*, *11*(3), 303–328.

Bohn, O. S., & Flege, J. E. (1992). The Production Of New And Similar Vowels By Adult German Learners Of English. *Studies in Second Language Acquisition*, *14*(2), 131–158.

Bradlow, A. R., Akahane-Yamada, R., Pisoni, D. B., & Tohkura, Y. (1999). Training Japanese listeners to identify english /r/and /1/: Long-term retention of learning in perception and production. *Perception and Psychophysics*, *61*(5), 977–985.

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

Broersma, M. (2002). Comprehension of non-native speech: Inaccurate phoneme processing and activation of lexical competitors. *7th International Conference on Spoken Language Processing, ICSLP 2002*, 261–264.

Broersma, M., & Cutler, A. (2008). Phantom word activation in L2. *System*, *36*(1), 22–34.

Broersma, M., & Cutler, A. (2011). Competition dynamics of second-language listening. *Quarterly Journal of Experimental Psychology*, *64*(1), 74–95.

Burnham, D., Joeffry, S., & Rice, L. (2010). Computer- and human-directed speech before and after correction. *Thirteenth Australasian International Conference on Speech Science and Technology (SST)*, 13–17.

Buz, E., Tanenhaus, M. K., & Jaeger, T. F. (2016). Dynamically adapted context-specific hyper-articulation: Feedback from interlocutors affects speakers' subsequent pronunciations. *Journal of Memory and Language*, *89*, 68–86.

Carnegie Mellon University. (2019). The CMU Pronouncing Dictionary. Retrieved December 6, 2019, from http://www.speech.cs.cmu.edu/cgi-bin/cmudict

Costa, A., Pickering, M. J., & Sorace, A. (2008). Alignment in second language dialogue. *Language and Cognitive Processes*, *23*(4), 528–556.

Eager, C. D. (2015). Automated voicing analysis in Praat : statistically equivalent to manual segmentation. *Proceedings of the 18th International Congress of Phonetic Sciences*.

Eger, N. A., & Reinisch, E. (2019). The role of acoustic cues and listener proficiency in the perception of accent in nonnative sounds. *Studies in Second Language Acquisition*, *41*(1), 179–200.

Ellis, R. (2009). Corrective Feedback and Teacher Development. *L2 Journal*, *1*(1), 2–18.

Escudero, P., Hayes-Harb, R., & Mitterer, H. (2008). Novel second-language words and asymmetric lexical access. *Journal of Phonetics*, *36*(2), 345–360.

Felker, E., Troncoso-Ruiz, A., Ernestus, M., & Broersma, M. (2018). The ventriloquist paradigm: Studying speech processing in conversation with experimental control over phonetic input. *The Journal of the Acoustical Society of America*, *144*(4), EL304–EL309.

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* (pp. 233–275). Timonium, MD: York Press.

Goto, H. (1971). Auditory perception by normal Japanese adults of the sounds "L" and "R." *Neuropsychologia*, *9*(3), 317–323.

Herd, W., Jongman, A., & 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*(6), 4247–4255.

Hirata, Y. (2004). Computer Assisted Pronunciation Training for Native English Speakers Learning Japanese Pitch and Durational Contrasts for Native English Speakers Learning Japanese. *Computer Assisted Language Learning*, *17*(3–4), 357–376.

Hwang, J., Brennan, S. E., & Huffman, M. K. (2015). Phonetic adaptation in non-native spoken dialogue: Effects of priming and audience design. *Journal of Memory and Language*, *81*, 72–90.

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

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

Kim, M., Horton, W. S., & Bradlow, A. R. (2011). Phonetic convergence in spontaneous conversations as a function of interlocutor language distance. *Laboratory Phonology*, *2*(1), 1–30.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, *82*(13).

Ladefoged, P. (1982). *A Course In Phonetics* (2nd ed.). New York: Harcourt Brace Jovanovich.

Ladefoged, P., & Broadbent, D. E. (1957). Information Conveyed by Vowels. *The Journal of the Acoustical Society of America*, *29*(1), 98–104.

Lemhöfer, K., & Broersma, M. (2012). Introducing LexTALE: A quick and valid Lexical Test for Advanced Learners of English. *Behavior Research Methods*, *44*(2), 325–343.

Lewandowski, E. M., & Nygaard, L. C. (2018). Vocal alignment to native and non-native speakers of English. *The Journal of the Acoustical Society of America*, *144*(2), 620–633.

Linebaugh, G., & Roche, T. (2013). Learning to hear by learning to speak: The effect of articulatory training on Arab learners' English phonemic discrimination. *Australian Review of Applied Linguistics*, *36*(2), 146–159.

Linebaugh, G., & Roche, T. (2015). Evidence that L2 production training can en-hance perception. *Journal of Academic Language & Learning*, *9*(1), 1–17.

Lively, S. E., Logan, J. S., & 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 An investigation of current models of second language speech perception: The case of Japanese ad. *Citation: The Journal of the Acoustical Society of America*, *94*, 2711.

Llompart, M., & 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*, 1–29.

Lobanov, B. M. (1971). Classification of Russian Vowels Spoken by Different Speakers. *The Journal of the Acoustical Society of America*, *49*(2B), 606–608.

Logan, J. S., Lively, S. E., & Pisoni, D. B. (1991). Training Japanese Listeners To Identify English /R/ And /1/: A First Report. *Journal of the Acoustical Society of America*, *89*(2), 874–886.

Lotto, A. J., Sato, M., & Diehl, R. L. (2004). Mapping the task for the second language learner: The case of Japanese acquisition of /r/ and /l/. *From Sound to Sense: 50+ Years of Discoveries in Speech Communication*, 181–186.

Luke, S. G. (2017). Evaluating significance in linear mixed-effects models in R. *Behavior Research Methods*, *49*(4), 1494–1502.

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017). Montreal forced aligner: Trainable text-speech alignment using kaldi. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, *2017*, 498–502.

McCloy, D. R. (2012). Vowel normalization and plotting with the phonR package. *Technical Reports of the UW Linguistic Phonetics Laboratory*, *3*, 1–8.

Meyer, A. S., Huettig, F., & Levelt, W. J. M. (2016). Same, different, or closely related: What is the relationship between language production and comprehension? *Journal of Memory and Language*, *89*, 1–7.

Oviatt, S., Levow, G.-A., Moreton, E., & MacEachern, M. (1998). Modeling global and focal hyperarticulation during human–computer error resolution. *The Journal of the Acoustical Society of America*, *104*(5), 3080–3098.

Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, *27*(02).

Pickering, M. J., & Garrod, S. (2006). Alignment as the basis for successful communication. *Research on Language and Computation*, *4*(2–3), 203–228.

Port, R. F., & O’Dell, M. L. (1985). Neutralization of syllable-final voicing in German. *Journal of Phonetics*, *13*(4), 455–471.

R Core Team. (2019). R: A language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria*.

Raphael, L. J., Dorman, M. F., Freeman, F., & Tobin, C. (1975). Vowel and nasal duration as cues to voicing in word final stop consonants: spectrographic and perceptual studies. *Journal of Speech and Hearing Research*, *18*(3), 389–400.

Robinson, P. (1995). Attention , Memory , and the “ Noticing ” Hypothesis. *Language Learning*, *45*(2), 283–331.

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

Saito, K., & Lyster, R. (2012). Effects of Form-Focused Instruction and Corrective Feedback on L2 Pronunciation Development of /r/ by Japanese Learners of English. *Language Learning*, *62(2)*(June), 595–633.

Schertz, J. (2013). Exaggeration of featural contrasts in clarifications of misheard speech in English. *Journal of Phonetics*, *41*(3–4), 249–263.

Schmidt, R. W. (1990). The Role of Consciousness in Second Language Learning. *Applied Linguistics*, *11*, 129–158.

Smith, B. L., Hayes-Harb, R., Bruss, M., & Harker, A. (2009). Production and perception of voicing and devoicing in similar German and English word pairs by native speakers of German. *Journal of Phonetics*, *37*(3), 257–275.

Stanley, J. (2019). joeyr: Functions for Vowel Data. R package version 0.3.0. Retrieved from https://rdrr.io/github/JoeyStanley/joeyr/

Thorin, J., Sadakata, M., Desain, P., & McQueen, J. M. (2018). Perception and production in interaction during non-native speech category learning. *The Journal of the Acoustical Society of America*, *144*(1), 92–103.

Tomlin, R. S., & Villa, V. (1994). Attention in cognitive science and second language acquisition. *Studies in Second Language Acquisition*, *16*(2), 183–203.

Truscott, J. (1998). Noticing in second language acquisition: a critical review. *Second Language Research*, *14*, 103–135.

van Heuven, W. J. B., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). SUBTLEX-UK: A new and improved word frequency database for British English. *Quarterly Journal of Experimental Psychology*, *67*(6), 1176–1190.

Wang, Y., Jongman, A., & 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*(2), 1033–1043.

Warner, N., Jongman, A., Sereno, J., & Kemps, R. (2004). Incomplete neutralization and other sub-phonemic durational differences in production and perception: Evidence from Dutch. *Journal of*

Phonetics, *32*(2), 251–276.

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

Winter, B. (2013). *Linear models and linear mixed effects models in R with linguistic applications*.

Wong, J. W. S. (2014). Comparing the effects of perception and production training on the learning of English vowel contrast /e/ and /æ/ by Cantonese ESL learners. *Linguistics and Language Conference 2014 (LILA '14)*, (July).

Appendix A

Table S1.

Experimental items from the Word or Not Games (i.e., pre- and post-tests).

Words				**Non-words**			
/æ/	**/ɛ/**	**Final /t/**	**Final /d/**	**/æ/**	**/ɛ/**	**Final /t/**	**Final /d/**
cash	breath	burst	beard	bleck	banch	ard	blate
class	chess	dirt	blind	cresh	blass	crade	frient
gap	death	fit	board	gless	chack	flird	golt
hatch	deck	hunt	child	leb	dask	knid	jate
lap	dress	kite	field	lemp	franch	poind	nert
map	fresh	moist	glide	mesk	pladge	quode	reat
rank	hedge	most	guard	plenk	prap	sald	shate
scratch	help	part	guide	smesh	prass	skird	sount
slam	sense	skate	proud	spen	skatch	smard	weirt
spank	stretch	spit	shield	splesh	strass	streed	woot
tag	web	trust	speed	trep	tanse	twisd	wount
thank	yell	vote	stand	tresh	wadge	waisd	yart

Appendix B

Table S2.

Critical words for the Code Breaker Game: Minimal pairs containing the /æ/-/ɛ/ vowel contrast and minimal pairs containing the word-final /t/-/d/ contrast.

/æ/	**/ɛ/**	**Final /t/**	**Final /d/**
Bad	Bed	Bright	Bride
Bag	Beg	Court	Cord
Cattle	Kettle	Fate	Fade
Flash	Flesh	Feet	Feed
Had	Head	Greet	Greed
Madly	Medley	Height	Hide
Man	Men	Hurt	Heard
Mansion	Mention	Root	Rude
Mass	Mess	Seat	Seed
Pan	Pen	Sight	Side
Rack	Wreck	Slight	Slide
Ranch	Wrench	Squat	Squad