The impact of one's own voice and production skills on word recognition in a second language

Nikola Anna Eger & Eva Reinisch
Institute of Phonetics and Speech Processing, Ludwig Maximilian University Munich
eger@phonetik.uni-muenchen.de, evarei@phonetik.uni-muenchen.de

Running head: A self-benefit for spoken-word recognition in L2

Author note

Nikola Anna Eger and Eva Reinisch, Institute of Phonetics and Speech Processing,
Ludwig Maximilian University Munich, Germany.
This project was funded by a grant from the German Research Foundation (DFG; grant nr. RE 3047/1-1) to the second author. This work will be part of the first author's PhD project. Parts of the work have been presented at the workshop "How Words Emerge and Dissolve" 2016 in Munich, Germany, and at the "Second Workshop on Psycholinguistic Approaches to Recognition in Adverse Conditions" 2016 in Nijmegen, the Netherlands. We would like to thank Miquel Llompart and Will Schuerman for discussion of a previous version of the paper. We would also like to thank Rosa Franzke for help with segmenting the acoustic data.
Correspondence concerning this article should be addressed to: Nikola Anna Eger,
Institute of Phonetics and Speech Processing, Ludwig Maximilian University Munich,
Schellingstraße 3, 80799 Munich, Germany. E-mail: eger@phonetik.uni-muenchen.de

Abstract

Second language (L2) learners often speak with a strong accent, which can make them difficult to understand. However, familiarity with an accent enhances intelligibility. We propose that L2 learners are even more familiar with their own accented speech patterns and may thus understand self-produced L2 words better than others' accented productions, presumably due to adaptation. This hypothesis was tested by asking German learners of English to identify English words from minimal pairs that are distinguished by difficult L2 sound contrasts. Words had been spoken by the learners themselves or other learners who produced the contrasts equally well. Self-produced words were identified significantly better than others' productions. A second experiment revealed that better producers can exploit acoustic cues in perception more than poor producers, especially when the produced acoustic cues to the minimal pairs were clearly differentiated. The self-benefit, however, did not depend on production skills. We conclude that L2 learners adapt not only to their L1 accent in general but also to their own specific speech patterns. Speculating about L2 acquisition more generally, these results may raise the question whether adaptation to own, accented productions may be one reason why learners have difficulties to improve their pronunciation, since they may not notice a need to improve.

Keywords: Spoken-word recognition; Second language learning; Non-native sound contrasts; Own-voice perception; Foreign accent

The impact of one's own voice and production skills on word recognition in a second language

Learning a new language brings along many challenges. Even learners who have been using a second language (L2) for a very long time often retain a perceptible foreign accent in their L2 pronunciation. That is, their production of sounds and prosodic characteristics of the second language differs from how native speakers of that language would typically produce them. Interestingly, L2 learners are often well aware that their fellow learners (e.g., in the classroom) speak with a strong accent. Anecdotal evidence comes from the many jokes about learners' failure to produce certain sounds correctly (e.g., a German sailor saying "[s]ink positive"). However, given this awareness of others' errors, the question arises as to why listeners would not (always) be able to use that kind of information to improve their own accent.

One possible reason for this failure may be because L2 learners are highly familiar with their own accents. Familiarity with foreign-accented speech in general is beneficial to understanding the accent. The other side of the coin to also better understanding one's own accent, however, may be reduced awareness of one's own errors. In the present study, we test the part of this suggestion that L2 learners understand their own speech better than the speech of other L2 learners, presumably as a consequence of greater experience with and exposure to "self" speech compared to "other" speech. This seems plausible since the perception of others has been shown to differ in several aspects from the perception of oneself. Differences in the perception of self vs. others have been shown with regard to face recognition (see e.g., Devue & Brédart, 2011, for an overview), body (e.g., Ionta, Gassert, & Blanke, 2011), odor (e.g., Platek, Burch, & Gallup, 2001), and, importantly, voice and word recognition (e.g., Douglas & Gibbins, 1983; Schuerman, Meyer, & McQueen, 2015; Shuster, 1998; Xu, Homae, Hashimoto, & Hagiwara, 2013).

The sound of our voice that we are used to hearing while speaking differs from the sound we hear on a recording. This is due to the different routes via which the sound is conducted: air vs. air and bone (Shuster & Durrant, 2003). However, people have been shown to be good at recognizing their own

voice on recordings and differentiate it from unfamiliar voices as indicated by a variety of behavioral and physiological measures (Aruffo & Shore, 2012; Douglas & Gibbins, 1983; Graux, Gomot, Roux, Bonnet-Brilhault, Camus, & Bruneau, 2013). Graux and colleagues, for example, used event-related potentials (ERPs) to show that neural processes involved in discriminating one's own voice from others' voices differs from processes involved in the discrimination of two unknown voices. It appeared that brain activity pertaining to attentional processing was reduced when listening to one's own voice relative to others' voices. Xu and colleagues (2013) showed that speakers were significantly better at identifying their own voice than voices of other, familiar speakers. This effect was especially strong in difficult listening conditions, where the recordings were filtered so that only frequencies above 2000Hz were available (Xu et al. 2013). This benefit was explained by richer and more stable self-representations that could result from higher auditory familiarity with one's own voice relative to others' voices, and from strong associations with motor/articulatory representations. In line with this account, Shuster (1998) found that children with a phonological disorder accepted recordings of their own words as correctly produced more frequently than words with similar errors that were uttered by other children. That is, children recognized their own errors less often than others' errors. Shuster (1998) argued that this lower awareness of self-produced errors resulted from the experience the children had with their own, erroneous productions that eventually might have led to imprecise representations and hence imprecise perception and articulation models (see also Strömbergsson, Wengelin, & House, 2014).

"Unusual" pronunciation is common not only in the field of phonological impairments but also in the field of second language learning, where learners often retain a foreign accent. L2 acquisition models such as the Speech Learning Model (SLM, Flege, 1995) propose that foreign accents can be explained through sound similarities between the learners' first and second language (in addition to developmental and social factors)[1]. The general idea is that L2 sounds are perceived through a native-language (L1) filter. That is, L2 sounds that do not occur in a learner's first language are preferentially interpreted as the closest native category. In other words, second language learners

are (at least initially) bad perceivers and this has repercussions on their productions in the second language where similar assimilation processes take place.

The perception and production of an unfamiliar sound contrast in the L2 is especially difficult if both sounds are mapped onto a single native category. The English vowel contrast /ɛ/ – /æ/, for instance, challenges learners whose first language is Dutch or German (e.g., Bohn & Flege, 1992; Broersma, 2012; Eger & Reinisch, in press; Escudero, Hayes-Harb, & Mitterer, 2008; Llompart & Reinisch, 2017; for learners of other L1's see e.g., Flege, Bohn, & Jang, 1997; Ingram & Park, 1997; Tsukada, Birdsong, Bialystok, Mack, Sung, & Flege, 2005). For learners of both languages, one of the sounds, /ɛ/ (as in *bet*), is familiar as it sounds similar to their native unrounded front open-mid vowel. Using their native substitute for this sound usually results in acceptable productions as perceived by native English listeners. It is hence an "easy" sound to acquire. However, the more open /æ/ (as it occurs in *bat*) does not occur in either German or Dutch and is difficult to acquire, therefore it is frequently substituted with the nearest L1 sound /ɛ/. To native listeners the learners' intended /æ/ hence sounds accented or simply "wrong" because to them it does not sound like the intended category. Other problematic contrasts for learners of English are voicing contrasts in word-final obstruents. Although German does have voicing contrasts in word-initial and medial obstruents, in word-final position all obstruents are devoiced and this process of devoicing is transferred to L2 English (Smith, Hayes-Harb, Bruss, & Harker, 2009; for learners of other L1's see e.g., Broersma, 2005, 2010; Cebrian, 2000; Cho & McQueen, 2006; Flege, Munro, & Skelton, 1992; Hayes-Harb, Smith, Bent, & Bradlow, 2008; Weber, Broersma, & Aoyagi, 2011; Xie & Fowler, 2013).

Critically, whatever sound contrast should be produced, for native and non-native speakers, there is usually more than one way to signal the contrast by using a combination of different cues. The word-final voicing contrast in English stops, for instance, is signaled by the duration of the preceding vowel, closure duration, duration of the burst/aspiration as well as voicing during the closure (Barry, 1979; Port & Dalby, 1982; Wright, 2004). One central observation when looking at such difficult but lexically relevant contrasts in second language learning is that even if a difficult contrast is not neutralized by the L2 learner, this does not necessarily mean that it is maintained in a native-like

manner, neither in production nor in perception. Instead, the cues that L2 learners produce to indicate a difficult L2 contrast might be less differentiated and/or different from native cues. Similar differences for the use of cues can also be found in perception. In many cases, learners transfer their habitual use of cues from the native language to the target language, even if these cues are only secondary or irrelevant in the L2 (e.g., Bohn, 1995; Iverson, Kuhl, Akahane-Yamada, Diesch, Tohkura, Kettermann, & Siebert, 2003; Levy & Law II, 2010; McAllister, Flege, & Piske, 2002). Importantly, it has been demonstrated that patterns of transfer may be even learner-specific (Schertz, Cho, Lotto, & Warner, 2015, 2016; Smith & Hayes-Harb, 2011) and/or differ between perception and production (Eger & Bohn, 2015; Kartushina & Frauenfelder, 2014; Kassaian, 2011; Schertz et al. 2015).

Given this influence of the learners' first language sound patterns on their accent, it has been suggested that other non-native speakers of a target language are often as good or even better at perceiving accented speech than native speakers of the target language (Imai, Flege, & Walley, 2003; Munro, Derwing, & Morton, 2006), specifically when listeners and speakers share the same native language (Bent & Bradlow, 2003). This advantage has been termed the matched interlanguage intelligibility benefit and has been argued to result from shared knowledge about the phonetics of the first language (Bent & Bradlow, 2003, see also, Hayes-Harb et al. 2008). It may also result from incorrect perception such that L2 learners who perceive L2 speech through the perceptual filter of their native language don't suffer from mismatches if the accented production (at least partially) matches the L1 filter. This possibility seems likely given that the interlanguage intelligibility benefit has mainly been found for learners at low levels of proficiency, that is, when low-proficiency learners listen to utterances produced by other low-proficiency speakers of the L2 (Hayes-Harb et al. 2008; Pinet, Iverson, & Huckvale, 2011; van Wijngaarden, Steeneken, & Houtgast, 2002; Xie & Fowler, 2013).

In addition to this L1-filtered perception, L2 learners, especially those who learn their L2 in a classroom in their home country, have ample exposure to accented speech, for instance, from their fellow learners and often even from the teacher. A large body of research has shown that native listeners are able to quickly adapt to or tune in to non-native speech, such that with experience

accented speech becomes easier to understand (e.g., Bradlow & Bent, 2008; Clarke & Garrett, 2004; Reinisch & Weber, 2012; Sidaras, Alexander, & Nygaard, 2009; Witteman, Weber, & McQueen, 2013). This may also be the case for second language learners. Weber, Di Betta and McQueen (2014), for instance, showed that Dutch listeners had little difficulty in processing Italian-accented English when they were familiar with an Italian accent in their L1 Dutch. That is, listeners transferred their knowledge on a specific foreign accent from their first language to a second language (see also Reinisch, Weber, & Mitterer, 2013). In addition to accent adaptation for other speakers, second language learners may also adapt to their own personal accent, that is, their very own specific L2 production patterns.

In terms of second language sound representations the suggested adaptation can be illustrated using a belief-updating model (Kleinschmidt & Jaeger, 2015). The model suggests that listeners perceive sounds as probabilistically falling into one or the other category. For our purposes of describing L2 listening we refer to them as the global sound distributions that represent the L2 categories that learners perceive through their L1 filter and that are formed by or abstracted from native and non-native tokens that listeners encountered. In addition, the model suggests that listeners track consistencies within the speech signal for a given situation. If consistencies are found for a given situation, the global distributions are adapted. That is, listeners are likely to have distributions for accented (L2) speech that they frequently encounter (see e.g., Cutler, 2015, for suggestions about "accented" representations). These adapted distributions can be reapplied in perception when the situation or speaker is recognized again, hence facilitating perception. Further adaptation to the situation or speaker would then proceed from this model.

Importantly, when listening to accented speech, L2 learners have experience not only with their fellow learners' accented speech but also with their own accented productions. That is, every time they speak their L2, they hear themselves speak and receive proprioceptive feedback from their (accented) productions (i.e., articulation patterns; Abbs, Gracco, & Cole, 1984; Guenther, 2006; Postma, 2000). Critically, in this way they may be even more familiar with their own speech patterns than with others' accented speech. If through this reinforcement, learners adapted to their own

personal accent, they then may show a benefit when recognizing words in their own voice in the L2 relative to other learners' voices. An advantage when perceiving one's own voice could be in principle the same as the above mentioned matched interlanguage benefit, differing in that the adaptation is even more particular and applies to one's own voice rather than to an entire language-specific accent. Our aim is to experimentally test whether learners are indeed better at understanding words produced by themselves than words produced by other, equally proficient L2 learners.

Experiment 1 tested this by means of a word-reconstruction (or word-identification) task with minimal word pairs containing difficult sound contrasts for German learners of English. The contrasts were the vowel contrast /ɛ/ – /æ/ and the voicing distinction in word-final fricatives and stops. /ɛ/ and the voiceless obstruents have corresponding sounds in German and are therefore supposedly "easy" to produce. /æ/ and word-finally voiced obstruents do not occur in German and are hence difficult for Germans to produce. These sounds tend to be produced as the respective other sounds of the contrasts (Bohn & Flege, 1992; Smith et al. 2009).

German learners of English were recorded producing these words and two months later were invited back for a perception experiment. There they were presented the words one at a time and asked to decide which word of the minimal pair was intended. Critically, they were presented with words they had produced themselves as well as productions of other learners from the sample. Speakers were grouped separately for each sound contrast such that they were matched in the type and magnitude of cues they produced. If a self-benefit was found here, that is, if learners understood/reconstructed the words better when presented with their own voice, this may suggest that familiarity with their own productions (through adaptation or stored representations) affected L2 processing.

As for the magnitude of the benefit of understanding one’s own voice it would be reasonable to assume that the benefit may be larger for poor learners. This second hypothesis is motivated by previous findings that low-proficient speakers were better at understanding strongly accented speech than high-proficient and native listeners (Hayes-Harb et al. 2008). If exposure to one’s own productions played a role, poor listeners might benefit even more from experience with their own

accent, since their productions are overall more difficult to understand. Reversely, a self-benefit might be not so large for better learners, since the acoustic cues that they produce – even though possibly not native-like – might already be sufficient for good perception. With regard to the words within a minimal pair, the self-benefit may be larger for words with the difficult sounds, that is, the ones that do not occur in the learners' L1, as each learner may use different sets of cues to keep these sounds apart from the easy ones (i.e., the ones that are present in the L1). The role of the magnitude of produced acoustic difference between the "easy" and "difficult" sounds of the contrasts across proficiency groups will be further explored in Experiment 2.

EXPERIMENT 1

METHOD

Participants

Twenty-four female[2] students at the University of Munich participated for pay. They were native speakers of German and reported no history of speech, language, or hearing problems. The mean age was 22.4 ranging from 19 to 28. All speakers had learned English at school starting at an average age of 10.0 years and following classes for an average of 8.2 years. None of them had lived in an English-speaking country for longer than 6 months. Since on German television films and series are dubbed into German, exposure to native speakers of English was likely limited. All participants took part in two sessions: one for the recordings, and one for the perception experiment a few weeks later. In addition, they filled in a language background questionnaire with special focus on their history of learning English. The production data is shown in Figures B.1 to B.3 in the Appendix.

Materials

Thirty-one English minimal word pairs that differed in sound contrasts that have been shown to cause problems for German learners were selected (Bohn & Flege, 1992; Smith et al. 2009). Eleven minimal pairs contained the vowel contrast /ɛ/ – /æ/, seven pairs a word-final voicing contrast in fricatives and thirteen pairs a word-final voicing contrast in stops. Within each pair, one word was considered to contain an easy sound for the learners (i.e., /ɛ/, and the voiceless sounds in word final position). The other word contained a difficult sound (/æ/ and the word final voiced sounds). The

labels easy and difficult were based on whether or not the critical sounds occur in the German sound inventory (German does not have the vowel /æ/) and in the given word position (German word-final phonologically voiced obstruents are canonically produced as devoiced). An additional 22 words were selected to serve as fillers in the recording session. Some of them contained other difficult sounds that do not occur in German (e.g., /θ/ or /w/) to distract participants from the main purpose of the study. Words are listed in Appendix A.1.

Recordings

For the recordings all words were randomly assigned to one of ten semantically neutral carrier sentences such as *The next word is...* (see Appendix A.2). Target words always occurred in sentence final position. The assignment of sentences to words and the order of words within the recording session were randomized separately for each participant with the restriction that the words of a minimal pair could not follow one another. Each word was repeated twice for a total of 160 sentences[3].

Participants received all instructions in English and were asked to read out the entire sentence at a comfortable pace. The sentences including the target words were presented one by one on a screen, and a small light signaled when to start speaking. The recordings were made in a soundproof recording room using a diaphragm microphone (Neumann Microphone, type TLM 103) and the software Speechrecorder (Draxler & Jänsch, 2004), which stored each sentence as a separate wav file on a computer. After the session, participants had to review a list of all target words (in randomized order) and mark those words that seemed unknown or unfamiliar to them. The whole session lasted approximately 50 minutes.

Acoustic analyses

Several acoustic measures were taken using Praat (Version 5.4.08, Boersma & Weenink, 2015) for each sound contrast in order to group participants by production pattern for the perception experiment. The grouping of participants was done separately for each type of sound contrast, based on how well speakers differentiated the two words of the minimal pairs in production. We will refer to this grouping according to production patterns by the term “proficiency”, note however, that this

measure is not based on how native listeners judged the individual productions. Rather, with the term proficiency we refer to the acoustically measured magnitude of the produced difference between the critical sounds in the words of the minimal pairs (for native speakers' goodness ratings of a subset of the present productions see Eger & Reinisch, in press). We use the term "proficiency" rather than "production accuracy" or other terms highlighting production measures since in the perception experiments "proficiency" will refer to the *listeners'* proficiency (i.e., listeners' own production accuracy) rather than the quality of the sound material that they were listening to (for details see below).

The magnitude of the speakers' produced contrasts was assessed relative to the other speakers in the experiment such that differences between speakers within a group were minimized. Figures B.1 to B.3 in the Appendix show the main acoustic measures for each type of contrast for each of the eventually-formed proficiency groups. For the vowels, these cues were the between-category differences in the first two formants and duration. For the word-final fricatives, these were the duration of the preceding vowel and the duration of the fricative (combined as vowel duration divided by fricative duration), and the voiced portion of the fricative. For the word-final stops, the duration of the aspiration, the duration of the preceding vowel and the voiced portion of the closure were taken into account. These are the cues that are reported in the literature to be the most important ones for native speakers of English (see e.g., Deterding, 1997; Hillenbrand, Getty, Clark, & Wheeler, 1995, for the vowels; e.g., Broersma, 2010; Wright, 2004, for the fricatives; e.g., Barry, 1979; Smith et al. 2009, for the stops). The cues to the production of each contrast were weighted in the order named above. Specifically, we looked how *differently* the acoustic measures of the two categories were produced. A good contrast was defined as a large difference between the means of these measures within minimal pairs. "Good" also indicated that categories were discrete as reflected in smaller standard deviations for each measure and thus less overlap between the words of the minimal pairs. Looking at these measures of how large a difference the twenty-four participants had produced for each of the three sound contrasts it was decided that a split into three "proficiency" groups (A= best, B = middle, C = worst) of eight participants would best capture the

main differences. The assignment to groups was done separately for each contrast and followed this procedure: First, for each type of contrast separately, the eight best speakers with the clearest contrasts were assigned to the most proficient group A. Next, the eight speakers with the smallest/worst contrasts were assigned to group C. The remaining eight speakers were then assigned to group B.

In order to reduce the number of unfamiliar voices per participant presented in the perception experiment, within each of the three proficiency groups two subgroups were formed such that each group contained only four instead of eight voices (i.e., each listener was presented with only three unfamiliar voices per contrast). In this way, it was ensured that a sufficient amount of data could be collected for the analysis for tokens in one's own voice (see also Design below). The subgroups were formed such that speakers in one sub-group of four were not only similar in the overall amount of the produced difference for each contrast, but also according to which cues they had produced the largest difference in. This was especially important for speakers who had produced better contrasts using multiple cues since the goal was to match the productions of the participants' own voices with the presented others' voices as closely as possible. For instance, one speaker in group A may use duration of the preceding vowel to indicate a phonologically voiced fricative in *rise* in contrast to *rice*, with a longer vowel in the former, whereas another speaker from group A may produce the word-final voiced fricative with a long period of voicing in the fricative rather than a longer preceding vowel. However, for the statistical comparisons between the proficiency groups, only the three overall groups were considered, since differences between the subgroups were rather small and a comparison between all six groups appeared no more informative than a comparison between the three overall groups. Notably proficiency had to be used as a grouped variable rather than a continuous measure because the manipulation of Voice (listening to one's own voice vs. others' voices) had to be compared within sets of participants that listened to each other.

Design

For the perception experiment, the words of the minimal pairs were spliced out of the carrier sentences to be presented in isolation. Of the 31 recorded pairs, three were excluded due to

incorrect production of sounds other than the target sounds. Additionally, the word pair *latter-letter* was excluded because several participants indicated in the questionnaire that they did not know the meaning of the word “latter”. The final word set consisted of the remaining 50 words (27^4 pairs, see Appendix A.1).

The stimulus set of the experiment was prepared separately for each participant for each of the three sound contrasts. For each contrast, participants were presented with their own productions and those of three other speakers (henceforth other voices) that had been selected to match this participant in terms of use of cues for this contrast ("proficiency"; see above). That is, overall the stimulus set for each participant consisted of a 25% of own productions. The other voices were assigned separately for each type of contrast to maximally match the production patterns such as to isolate as far as possible the effect of voice over the types and magnitude of acoustic cues used to produce the contrasts. The total number of other unfamiliar voices in the experiment varied between 5 and 9 (though mostly 7 or 8) per participant. This was because a specific unfamiliar voice could occur in one, two, or even all three sets of contrasts. Within each contrast each participant always heard three other voices. That is, although one’s own voice was heard more frequently than any other single, unfamiliar voice in the experiment, overall it was heard less often than unfamiliar voices (i.e., 25%). Using groups of four learners per sound contrast appeared the best way to divide up our set of participants to tightly control production patterns between the own and other voices for each sound contrast. At the same time, it allowed us to collect enough data points for own-voice trials which would have been substantially lower had we compared every participant to everyone else or lowered the number of own-voice trials to the number of trials for each individual voice in the group.

For each voice, the two recorded repetitions of each word were presented twice each (i.e., for a total of 4 repetitions per voice per word). All words were presented once before they were repeated. The experiment consisted of a total of 864 trials (27 pairs x 2 words x 4 speakers x 2 spoken repetitions x 2 blocks) of which 216 tokens were in the participants’ own voice and 648 in an unfamiliar voice.

Procedure

Participants returned for the perception experiment approximately six weeks after they were recorded. They were informed about the procedure by means of written instructions. It was noted (although not emphasized) that among several unfamiliar voices they would hear themselves. Instructions were written in English as to set participants into an English language mode without influencing their perception by talking to them with a specific accent. Participants were seated in a sound-proof cabin in front of a laptop computer. On each trial, participants saw the two words of the minimal pair written on the left and right side of the computer screen. After 400 ms they were presented one of the words over headphones at a comfortable listening level. Their task was to indicate by button press which of the words was intended by the speaker. They pressed the 1-key on the computer keyboard if they thought the speaker intended the word on the left, and the 0-key for the word on the right. The material was presented in randomized order and the position (left or right) of the two response alternatives was counterbalanced according to correct answer so that participants were not biased towards left or right position. The experiment was implemented in Psychopy2 (Version 1.83.01; Peirce, 2007), and took approximately 50 minutes to complete. Every 70 trials participants were allowed to take a self-paced break. After the experiment participants were asked whether they had recognized their own productions throughout the experiment which most of them confirmed.

Results

Listeners' responses were categorized into correct and incorrect answers depending on whether they chose the intended word (i.e., the word the speaker intended to produce) or the other member of the minimal pair. Correct vs. incorrect (coded as 1 and 0 respectively) was used as the dichotomous dependent variable in a series of linear mixed-effects models with a logistic linking function (Jaeger, 2008). The models were implemented in R (Version 3.3.0, R Core Team, 2017) using the lme4 package (Bates, Mächler, Bolker, & Walker, 2015). The random-effects structure included random intercepts for participant and word with random slopes for fixed factors that were manipulated within participants and items respectively. Note that the models with a full random effects structure

(Barr, Levy, Scheepers, & Tily, 2013) did not always converge. In this case, it was determined via model comparisons using log-likelihood ratio tests whether simpler models would fit the data just as well. The best fitting model with the largest random effects structure that converged will be reported.

Since our main hypothesis was that listeners would be better at recognizing the intended word when hearing their own voice than other speakers' voices, the main variable of interest was Voice (i.e., whether they heard themselves or not). This variable was contrast coded to 0.5 = self, -0.5 = other. Note that "other" was collapsed across the different other (not-self) voices since within each proficiency group (i.e., groups of listeners that had produced similar cues to the sound contrasts) each voice contributed to "self" as well as "other" trials.

An overall model on all data with only Voice as fixed factor revealed a significant effect of Voice ($b_{(Voice)}$=0.24, SE=0.07, z=3.41, p<.001; $b_{(Intercept)}$=1.18, SE=0.17, z=6.87, p<.001) showing that overall in the experiment more correct answers were given if the stimulus was in the participant's own voice than someone else's voice ($M_{(self)}$ = 75.1 % correct, SD = 0.43; $M_{(other)}$ = 71.9 %, SD = 0.45) of the same proficiency level (i.e., because voices were matched on production patterns).

To test whether the effect of hearing one's own voice only emerged over the course of the experiment because listeners heard their own voice more often than any single other voice (since proficiency was matched within sound contrast), Trial number and its interaction with Voice were added as fixed factors to the model. Trial Number was centered and scaled from -1 to 1. As in the overall model, there was a significant effect of Voice ($b_{(Voice)}$=0.24, SE=0.07, z=3.42, p<.001; $b_{(Intercept)}$=1.18, SE=0.17, z=6.86, p<.001). However, neither Trial Number ($b_{(TrialsNumber)}$=0.05, SE=0.04, z=1.49, p=.14) nor the interaction between Trial Number and Voice were significant ($b_{(Voice:TrialNumber)}$=-0.03, SE=0.07, z=-0.42, p=.68). This indicates that the effect of Voice did not change over the course of the experiment. Trial Number was therefore not included in the subsequent analyses. The effect of Voice was then tested in relation to a number of other independent variables: sound Contrast (vowels, fricatives, stops), Proficiency (grouped into A, B, and C) and Sound Type (easy, difficult).

Sound Contrast

To test whether the effect of Voice held for all three types of sound contrasts (i.e., minimal pairs differing in the vowel contrast /ɛ/–/æ/, the final voicing contrast in fricatives or stops) a model was fitted with Voice, sound Contrast, and their interaction as fixed factors. Sound Contrast was coded as a factor with three levels with the vowel contrast mapped onto the intercept. Results showed that participants performed better overall for words from the stop contrast ($M_{(stop)}$ correct = 81.1%, SD = 0.39) than for words from the vowel contrast ($M_{(vowel)}$ correct = 67.2%, SD = 0.47; $b_{(Intercept\_vowel)}$=0.47, SE=0.17, z=2.68, p<.01; $b_{(Contrast\_stop)}$=1.76, SE=0.20, z=8.85, p<.001) but overall performance for the fricative contrast did not differ from the vowel contrast ($M_{(fricative)}$ correct = 66.5 %, SD = 0.47; $b_{(Contrast\_fricative)}$=0.46, SE=0.26, z=1.77, p=.08). Critically, for the vowel contrast that had been mapped onto the intercept there was a significant effect of Voice ($b_{(voice)}$=0.20, SE=0.08, z=2.46, p<.05) with better performance if participants heard their own voice. There was no interaction between Voice and sound Contrast for either other level of this factor suggesting that the magnitude of the effect of Voice that was found for the vowels was not different for the stop or fricative contrasts (fricatives: b=0.08, SE=0.12, z=0.65, p=.52; stops: b=0.22, SE=0.21, z=1.05, p=.29). Given that the effect of Voice was not modulated by sound Contrast, this factor was not included in any further analyses and data were collapsed across contrasts. Note that this did not affect the grouping of participants, which had been conducted for each sound contrast separately, and was coded in the variable Proficiency.

Proficiency

To test the hypothesis that participants with a lower proficiency in English would benefit more from hearing their own voice, a model was fitted with Voice, Proficiency and their interaction as fixed factors. Proficiency was taken to refer to the groups A, B, and C that listeners were assigned to according to their productions, that is, the magnitude with which each of the contrasts had been produced. Consequently, when we henceforth refer to the "high-proficiency group", or "group A", we refer to those participants who had produced a well-differentiated contrast between the words of the minimal pairs. Participants from the "low-proficiency group", group C, had produced the smallest contrast according to the acoustic measures. Participants from group B performed at an

intermediate level. As described in the Methods section in more detail, this grouping was done separately for each sound contrast. Note that since in this experiment participants were listeners, we will refer to this factor as also "listener proficiency". While in Experiment 1, participants only heard stimuli from their own proficiency group, and listener and speaker proficiency (i.e., how well the stimuli had been produced) are confounded, this distinction will be relevant for Experiment 2. The factor Proficiency was coded as numeric with A = 0.5, B = 0, and C = -0.5. With this coding, the grand mean is mapped onto the intercept and effects can be interpreted as main effects.

As can be observed in Figure 1, the benefit of hearing one's own voice was present for all three of our proficiency groups. This was confirmed by statistical analyses. There was a significant effect of Voice ($b_{(Voice)}$=0.24, SE=0.08, z=3.04, p<.01; $b_{(Intercept)}$=1.21, SE=0.15, z=8.00, p<.001). As expected, Proficiency had a significant effect such that the higher the proficiency the more correct responses were given ($b_{(Proficiency)}$=0.55, SE=0.13, z=4.22, p<.001). The interaction between Voice and Proficiency was not significant suggesting that the effect of Voice was not different between proficiency groups ($b_{(Voice:Proficiency)}$=-0.17, SE=0.14, z=-1.19, p=.23).

(Insert Figure 1 about here)

Sound Type

Sound Type refers to whether the intended word of the minimal pair contained the sound that listeners know from their L1 (i.e., /ɛ/ and the voiceless obstruents as "easy" sounds; contrast coded as 0.5) or not (i.e., /æ/ and the voiced obstruents as "difficult" sounds; coded as -0.5). An interaction with Voice would indicate that the effect found for Voice (with more correct responses for one's own voice) differed according to whether the word contained an easy or a difficult sound. Sound Type was added to the model including Proficiency since even though Proficiency did not interact with Voice in the analysis reported above, the identification of difficult sounds may especially challenge participants with lower proficiency (Hayes-Harb et al. 2008; Pinet et al. 2011; van Wijngaarden et al.

2002; Xie & Fowler, 2013). Note that since all dependent variables were contrast coded, again the grand mean is mapped onto the intercept and effects can be interpreted as main effects.

Results are shown in Figure 2 and Table 1. Again, there was a main effect of Voice (more correct responses for one's own voice), and Proficiency (more correct responses the higher the proficiency; that is, the better participants had produced the contrasts themselves). The effect of Sound Type was not significant. However, these effects were modulated by a two-way interaction between Sound Type and Proficiency as well as the three-way interaction between all factors. These interactions are illustrated in Figures 2 and 3. While there were more correct responses for words containing the difficult sound category for participants of proficiency group A, in proficiency group C more words containing the easy sound were recognized correctly. The three-way interaction suggests that this modulation of Sound Type by Proficiency had repercussions on the Voice effect. That is, the effect of Voice differed between proficiency groups when considering easy and difficult categories separately.

Follow-up analyses on the three-way interaction testing each Sound Type separately revealed that in words containing the easy sound, there was a significant effect of Voice with more correct answers when hearing one's own voice ($b_{(Voice)}$=0.28, SE=0.13, z=2.19, p<.05; $b_{(Intercept)}$=1.32, SE=0.18, z=7.33, p<.001). Proficiency and the interaction between Proficiency and Voice just failed to reach significance ($b_{(Proficiency)}$=0.28, SE=0.15, z=1.89, p=.06; $b_{(Voice:Proficiency)}$=0.38, SE=0.21, z=1.78, p=.08). The benefit when hearing one's own voice for easy words hence did not differ between proficiency groups. If anything there was a slight tendency for the self-benefit to become larger for higher proficiency groups (i.e., from proficiency group C to A, Figure 3 left panel).

For words with the difficult sounds, an effect of Proficiency emerged with more correct responses the higher the proficiency group ($b_{(Proficiency)}$=0.96, SE=0.18, z=5.35, p<.001; $b_{(Intercept)}$=1.19, SE=0.19, z=6.17, p<.001) and an interaction between Proficiency and Voice ($b_{(Voice:Proficiency)}$=-0.61, SE=0.22, z=-2.77, p<.01). The effect of Voice was not significant ($b_{(Voice)}$=0.15, SE=0.13, z=1.12, p=.26). That is, when looking at the difficult words, the self-benefit appears larger the lower the proficiency (i.e., in proficiency group C; see Figure 3 right panel).

(Insert Table 1 about here)

(Insert Figure 2 about here)

(Insert Figure 3 about here)

Discussion

Experiment 1 demonstrated that second language learners are better at recognizing L2 words containing sounds from a difficult L2 sound contrast if they had produced the words themselves than when listening to other learners of the same L1 that used similar acoustic cues to produce the L2 words. Although the overall ease of identifying the intended word differed between sound contrasts, the effect of Voice did not differ. This suggests that the effect of Voice was not specific to one of the contrasts but may constitute a more general effect.

In an overall analysis including Voice and Proficiency as factors, the effect of Voice did not differ between proficiency groups (i.e., the variable that refers to listeners' own production skills). That is, as could be expected, there was a main effect of Proficiency such that the higher the proficiency of the learners, and hence the clearer the sound contrasts had been produced, the more correct responses were given. However, the hypothesis that the effect of hearing one's own voice may be larger for poor learners (i.e., poor producers), could not be confirmed in an all or nothing fashion (i.e., lack of an interaction in this analysis).

Interestingly, the effect of Voice was modulated in a three-way interaction with Proficiency and Sound Type. The latter defines whether the specific word of the minimal pair contains the sound that is similar to the learner's first language (i.e., "easy") or the sound that is exclusive of the L2 (i.e., "difficult"). Looking first at the results for words with the easy sound, that is, the L2 sound that is similar to the corresponding sound in the learners' L1, there was a main effect of voice (i.e., better recognition of words produced in one's own voice) and this self-benefit was not different between the three proficiency groups (though see Figure 3, left panel, for the tendency that for the easy sounds the self-benefit was somewhat larger the higher the proficiency). Considering words

containing the difficult sounds, in contrast, the self-benefit was larger the lower the proficiency of the learners (i.e., interaction between Voice and Proficiency in the follow-up analyses; see Figure 3, right panel). This confirms our suggestion that the self-benefit may be largest for poor learners, at least for difficult sounds.

The reason for the modulation of the effect of Voice by Sound Type together with Proficiency is likely to be found in the acoustics that the learners produced or failed to produce for the difficult sounds. As can be seen from the summary of the production data (Figures B.1 to B.3 in the Appendix), the learners that were assigned to Group A produced large contrasts between the words of the minimal pairs, specifically by better cuing the difficult sounds. The low-proficiency group, in contrast, produced rather small differences between the words in the minimal pairs (henceforth “poor” productions). Therefore, for the difficult sounds poor learners could benefit more from knowing their own patterns, since they had to rely on overall much smaller and/or less reliable cues to identify the difficult sounds. Note that despite these modulations of the effect of Voice, listeners from all proficiency groups benefitted from listening to their own productions in most conditions.

Why would listeners benefit from listening to their own productions? As discussed in the introduction, one reason may be the frequent exposure to one’s own accent. That is, whenever one speaks, one also hears one’s own productions, hence, unless an L2 learner is only passively exposed to the L2, their own voice is likely one of those heard most often also in the second language. Therefore, listeners are likely highly familiar with the relation between their own productions (or production strategies) and the acoustic consequences of these strategies for perceiving these difficult sound contrasts. Note that this does not necessarily mean that our listeners would remember how they read the specific word list during recordings six weeks prior to the perception experiment. Rather they may rely on adapted representations of accented sounds as their L2 targets.

In order to compare the perception of listeners’ own vs. other voices, participants in Experiment 1 were matched according to how clearly and with what types of cues they had produced the contrasts. Consequently, participants who had produced better contrasts were also presented with “rich” tokens, and low-proficiency speakers were presented with "poor" material. The next question

then is what would happen if poor learners had larger/more cues available and the good learners smaller/fewer cues. In order to investigate the influence of the availability of cues and the learner's proficiency, a second experiment was designed. It aimed at showing whether in Experiment 1 low-proficiency participants performed worse than high-proficiency participants only because they were presented with "poor" tokens or because they are also less capable of picking up acoustic cues. Moreover, if being a better producer helped L2 perception in general then the high-proficiency learners should outperform low-proficiency learners regardless of the magnitude of available cues.

EXPERIMENT 2

The aim of this experiment was to test the interaction between the availability of acoustic cues in the L2 speech signal and how well participants had produced the cues themselves (i.e., what we labeled "proficiency"). Specifically, we asked whether poor learners would benefit in perception from receiving more differentiated acoustic cues to the difficult sound contrasts. To address this question, those participants of Experiment 1 who differentiated the contrasts most and those who differentiated them least in production were invited back for a second perception experiment (i.e., groups A and C). They performed the same word identification task as before, but this time they were presented with stimuli of the opposite proficiency group. That is, participants who had been assigned to the high-proficiency group based on their productions were now presented with the "poor" tokens (i.e., those produced with only few and small acoustic cues) and participants who were assigned to the low-proficiency group were presented the "rich" tokens (i.e., those produced with various and clearer cues). Since the availability of cues may play a crucial role for word recognition, we expected that, when presented with "rich" productions, low-proficiency learners may perform better than when presented with poor productions (as in Experiment 1). However, the crucial question was whether despite this expected improvement, they would reach the level of the high-proficiency listeners found in Experiment 1. If not, then the availability of cues may play a role in L2 perception but the ability to use these cues may be related to the learners' own production abilities[5]. As concerns the high-proficiency learners, the question is whether they may be better at recognizing the words than the low-proficiency learners regardless of the quality of the stimuli (i.e., when

presented the rich and poor tokens). This would suggest that the ability to pick up available cues in perception is strongly related to the learner's ability to produce these cues. An additional comparison of group C listeners' perception of poor vs. rich material with the perception of the stimuli that they had produced themselves (i.e., in Experiment 1) will show how the effects of listener proficiency, hearing one's own voice, and availability of acoustic cues relate to each other.

METHOD

Participants

A subset of participants from Experiment 1 was invited to return for a third session. To test for differences in production and perception skills, the subset of those participants was selected who differentiated the contrasts best and those who differentiated them least in production. Specifically, we selected those participants who had been assigned to group A or C in Experiment 1 for at least two of the three sound contrasts and who were not in the opposite proficiency group for the third contrast (for example, a high-proficiency participant was either only in A-groups for all contrasts, or she was assigned to A for the stops and the vowels, and B for the fricatives). Seven participants from the high-proficiency group returned as did eight participants from the low-proficiency group. The experiment was run 4 to 11 weeks (mean=6.7 weeks) after the first perception experiment.

Materials

Since in Experiment 1 the sets of stimuli varied for each participant (due to the individual combination of the participant's own and several other voices), two sets of four participants from Experiment 1 were chosen from good speakers and two from poor speakers. The two sets from good speakers consisted exclusively of tokens produced by speakers from the high-proficiency group A. The two sets from poor speakers consisted of only tokens that had been produced by speakers from group C. Two sets per condition were chosen in order to use a representative sample of voices similar to Experiment 1. Each set contained between 8 and 10 voices across the three sound contrasts. Each participant was presented with one of these sets. Note that since in this experiment participants listened to stimuli of speakers from the "opposite" proficiency group, all voices were unfamiliar to them.

Design and Procedure

The design and procedure were the same as in Experiment 1. The participants received the same instructions but were told that this time they would not hear their own voice, but only “new“, unfamiliar voices. They heard the words in isolation and had to decide which word of the minimal pair had been produced.

Analyses

Again, listeners' responses were categorized into correct and incorrect responses depending on whether they chose the intended word or the other word of the minimal pair. Responses were coded as 1 = correct and 0 = incorrect, respectively, and used as the dichotomous dependent variable in a set of linear mixed-effects models with a logistic linking function (Jaeger, 2008). As for Experiment 1, the random-effects structure included random intercepts for participant and word with random slopes for fixed factors that were manipulated within participants and items. The best fitting model with the largest random effects structure that converged will be reported.

To compare high- and low-proficiency learners when presented with both, the “rich“ and “poor“ tokens, subsets of data from Experiment 1 were included in the present analyses. Note that our variable proficiency was based on production. However, since this grouping was done separately for each sound contrast, some of the participants invited back for Experiment 2 had been assigned to group B for one of the contrasts. In order to restrict our analyses to data from group A or C, all trials from contrasts for which a given participant had been assigned to group B in Experiment 1 were excluded from the analyses. For example, a given participant in Experiment 2 was assigned to group C for the fricatives and the stops in Experiment 1, because she produced the contrasts with very few and small acoustic cues, but she was assigned to group B for the vowel contrast. After participating in Experiment 2 as a listener of the overall proficiency group C, her responses to words from the vowel contrast were discarded. In this way, we controlled for both the quality of the tokens and listener proficiency to be restricted to groups A and C.

For the first model, only data for “other” voices were used, since in Experiment 2 voices were necessarily others’ voices. The main variables of interest were then the availability of acoustic cues

(i.e., Material with the levels rich and poor, coded as 0.5 and -0.5 respectively) and Proficiency of the listener (coded as 0.5 for the participants from the high-proficiency group A, and -0.5 for participants from the low-proficiency group C; note that again proficiency refers to the production as discussed in Experiment 1). In addition, an interaction between these variables was specified. As for Experiment 1 additional analyses included the factors Contrast, to test whether results held for all three sound contrasts, and Sound Type, to test whether effects would differ for easy and difficult sounds.

Results

The analysis of the model with Material, Proficiency, and their interaction as fixed factors revealed a significant effect of Material ($b_{(Material)}$=1.14, SE=0.14, z=8.40, p<.001; $b_{(Intercept)}$=1.11, SE=0.12, z=9.42, p<.001) with more correct responses for the rich than the poor tokens (see Figure 4, in white and dark grey boxes). There was a significant effect of listener Proficiency ($b_{(Proficiency)}$=0.54, SE=0.15, z=3.64, p<.001) with more correct responses for listeners from the highly-proficient group. Moreover, Proficiency was involved in an interaction with Material ($b_{(Material:Proficiency)}$=0.60, SE= 0.20, z=2.95, p<.01), indicating that the effect of Material was different in the two proficiency groups. Results are shown in Figure 4 in white and dark grey boxes.

To follow up on the interaction, two additional analyses were run to test the effect of Proficiency within each Material set. Proficiency was contrast-coded to A=0.5 and C=-0.5, as before. The results revealed significant effects of Proficiency for both the rich and the poor material set, with more correct responses when the listener was from the high-proficiency group A (rich material: $b_{(Proficiency)}$=0.84, SE=0.23, z=3.59, p<.001; $b_{(Intercept)}$=1.69, SE=0.16, z=10.32, p<.001; poor material: $b_{(Proficiency)}$=0.25, SE=0.11, z=2.36, p<.05; $b_{(Intercept)}$=0.54, SE=0.10, z=5.27, p<.001). This together with the difference in regression weights (i.e., higher $b_{(Proficiency)}$ for the rich than poor material set) suggests that the interaction between Proficiency and Material is driven by the magnitude of the effects. That is, both high- and low-proficiency learners can benefit in word recognition from hearing rich over poor cues, but learners from the high-proficiency group benefit to a larger extent.

Since poor learners appear to benefit from rich material, the question arises as to how this effect compares to the effect of Voice (i.e., the self-benefit) found in Experiment 1. Therefore, in an

additional analysis, the responses to "self"-trials from Experiment 1 were added to the dataset described above - again only for those participants who participated in both experiments. To compare poor learners' performance on rich material to all other conditions, one combined variable with six levels was included in the model instead of the previously used variables Proficiency and Material. Two of those levels defined the "self"-trials for each proficiency group (i.e., *A self*, *C self*). The other four consisted of trials in other voices, once with material from the same and once with material from the opposite proficiency group: listeners *A* hearing *poor* material, listeners *A* hearing *rich* material, listeners *C* hearing *poor* material, and listeners *C* hearing *rich* material.

In order to specifically test poor listeners' responses when presented with rich material relative to their performance when hearing their own voice (with poor cues) the level *C rich* was mapped on the Intercept. Results are presented in Table 2 and Figure 4 (*C rich* is indicated by thicker lines) and revealed significant differences between *C rich* and all other levels of the variable. That is, listeners from the low-proficiency group that were presented with rich material (*C rich*) performed significantly better than listeners from the same proficiency group when presented with their own (*C self*) or with others' productions from the poor material set (*C poor*). Moreover, those listeners were also better than listeners from the high-proficiency group when presented with poor material (*A poor*). Finally, low-proficiency listeners presented with rich material performed significantly worse than high-proficiency listeners hearing rich material, in both cases where high-proficiency listeners heard their own voice (*A self*) or other voices (*A rich*). This suggests that for low-proficiency listeners who themselves differentiate difficult contrasts only by using few and poor cues, the advantage when being presented with rich cues goes beyond and above the self-benefit. For learners who have already reached a high level in production in the L2 and produce more differentiated cues to the contrasts, the self-benefit appears to be on top of the material effect (see the effect of Voice in Experiment 1 for all proficiency groups).

(Insert Table 2 about here)

(Insert Figure 4 about here)

Sound Contrast

Having established that the effect of Material for the poor listeners goes above and beyond the self-benefit, for ease of interpretation, the remaining analyses will focus on the effect of Material without the factor Voice. To test whether the effects of Proficiency and Material reported above differed between sound contrasts, additional analyses were run involving the factor sound Contrast with the level "vowels" mapped onto the intercept (i.e., as in Experiment 1). Effects for the other contrasts are then interpreted relative to this reference level. Results are given in Table 3. They suggest that all effects found in the overall analyses held for the vowels (i.e., effect of Material, Proficiency, and their interaction). Furthermore, the same effects in the other contrasts were not significantly different from the effects in the vowels (i.e., Proficiency:Contrast was not significant for either fricatives or stops). Only two significant differences were found between the vowel contrast and either fricatives or stops: First, words from the stop contrast were overall identified better than words of the vowel contrasts (effect of Contrast for Stops, $M_{(stops)}$ = 80.2 % correct, SD = 0.40; $M_{(vowels)}$ = 65.6 % correct, SD = 0.47; $M_{(fricatives)}$ = 65.8 % correct, SD = 0.47). Secondly, the interactions between Contrast and Material for the fricatives and stops indicate that the effect of Material, that is, the difference in overall correct responses between rich and poor tokens, was larger for the fricatives (21.6 % difference of correct responses; $p=.06$) and stops (20.3 %; $p<.05$) than the vowels (10.0 %; see also Table 3).

(Insert Table 3 about here)

Sound Type

The effect of Sound Type was entered to the model with Material and listener Proficiency as fixed factors. This was to test whether – as in Experiment 1 – the effects of Material and Proficiency differed between words with an easy or a difficult sound (again coded as 0.5 and -0.5, respectively). Results are given in Table 4 and Figure 5. As in the overall analysis, there was an effect of Material and an effect of listener Proficiency in the same directions as before, and the interaction between

these two. Again, the interaction indicates that high-proficiency listeners benefit more from rich cues than listeners from the low-proficiency group do. Furthermore, there was no main effect of Sound Type, but one significant interaction between Sound Type and Material: As can be observed in Figure 5, the effect found for Material, that is, that words produced with rich cues (white boxes) are understood better than words with fewer and poorer cues (grey boxes), was larger for words containing the difficult sound category (two-way interaction between Material and Sound Type). The marginally significant three-way interaction between Sound Type, Material and Proficiency suggests that this effect was somewhat larger for the high-proficiency listeners (see also Figure 5).

(Insert Table 4 about here)

(Insert Figure 5 about here)

Discussion

Experiment 2 tested to what extent the complex pattern of results found in Experiment 1 with regard to proficiency and contrast could be due to the availability of larger acoustic differences between the words of the minimal pairs for the high-proficiency group. Specifically, we asked, firstly, whether poor learners would benefit in perception if they were also presented with large acoustic differences between the words of the minimal pairs and specifically more differentiated cues on the difficult sounds (i.e., the sounds that they don't know from their L1). Secondly, we asked whether good learners would outperform poor learners regardless of the magnitude of available cues.

Results showed that indeed, poor learners benefit from “rich” speech material, that is they were better at recognizing words produced by speakers from group A than their "own" group C productions in Experiment 1. However, despite this benefit for the "rich" materials, low-proficiency learners did not reach the level of performance of the proficient listeners (see Figure 4). Critically, high-proficiency learners outperformed poor proficiency learners for both the rich and poor material, however, the effect was larger for the rich materials. This suggests that differentiating L2 contrasts better in production (cf. our definition of proficiency) allows learners to perceive even small cues to

an L2 contrast better than poor learners, however this benefit is seen especially then, when sounds are cued in a fashion that approaches native production.

A comparison of the effect of Material with the effect of Voice (i.e., the self-benefit shown in Experiment 1) revealed that even though poor listeners benefit from hearing their own poor productions over others' poor productions, the availability of rich cues in the signal leads to even better word recognition. That is, the effect of rich material was above and beyond the self-benefit for low-proficiency learners. When the learners' productions were already clearly differentiated (i.e., in proficiency group A), the benefit when hearing one's own, good productions was on top of the effect of rich material produced by other speakers.

The benefit of being presented "rich" over "poor" productions was present for all three sound contrasts, but differed in its magnitude. It was larger for the fricative and stop contrasts than for the vowel contrast. As can be seen from the production data in Appendix B.1-B.3 the produced acoustic differences between the words of the voicing contrast in word-final stops or fricatives was larger than for the vowel contrast. Together with the results of Experiment 1 where listeners also showed overall better performance for the word-final voicing contrasts, this underlines the relevance of available acoustic cues in L2 word identification: the more differentiated the cues to a difficult sound contrast, the better L2 perception is.

The relevance of acoustic cues is further confirmed when looking at the easy and difficult sounds within each contrast (Sound Type). Words with the difficult sounds, that is, sounds that do not occur in the learners' L1, were overall harder to recognize when they were produced by poor speakers (i.e., the two-way interaction between Material and Sound Type, the grey boxes in the right panel of Figure 5). When the words with difficult sounds were produced by good speakers (i.e., were part of our "rich" productions, white boxes) then both listener groups showed a benefit, but with the above-mentioned difference between high- and low-proficient listeners (tentatively confirmed by the marginally significant three-way interaction between all factors).

Taken together, the results underline that, in order to recognize L2 words that differ in a difficult contrast, learners benefit most from two sources: (i) acoustic cues to be used for perception that

result from a differentiated production of the minimal word pairs and (ii) having already acquired a reasonable level of proficiency, here established as good production skills. This can be especially observed in words containing a difficult sound. Even though low-proficiency learners had a larger benefit when hearing rich cues than when hearing their own, poor productions, the self-benefit helped identify the intended word, and was on top of the material effect when both sets contained rich cues (i.e., in the high-proficiency learners).

GENERAL DISCUSSION

The present study showed that learners of a second language understand L2 words better when they were spoken in their own voice than others' voices, even when the speech material was matched according to the speakers' proficiency, that is, production patterns. The question of whether a self-benefit could be found in L2 learners was motivated by the hypothesis that due to frequent exposure to their own accented speech, learners are highly familiar with their own L2 sound patterns - above and beyond the sound patterns that are typical of their L1's accent - and that this facilitates the identification of the words they had produced themselves.

The self-perception benefit that we found in Experiment 1 did not differ between the three sound contrasts and proficiency groups. The hypothesis that the self-perception benefit would be larger for poor than proficient learners could only be confirmed in the three-way interaction of Voice, Proficiency, and Sound Type. Whereas the self-benefit did not differ between groups when presented with words containing the "easy" sound of the contrast (i.e., the one that was similar to the participants' L1), for words with the difficult sound the self-benefit was mainly apparent in the low-proficiency speakers. This difference for words with the difficult sound could be explained by differences in the produced acoustic cues between proficiency groups. Tokens produced by the high-proficiency group contained more or better cues to identify the words within a sound contrast, specifically by producing clearer cues to the difficult sounds. This likely helped not only in self-perception but also in the identification of words produced by the other speakers of the group who were matched in their production patterns. In other words, the proficient groups could be

reasonably sure to identify the difficult sounds as the intended ones by identifying the acoustic cues to the sounds even when they were produced by other learners. The self-benefit for the easy sounds in the high-proficiency group could stem from knowledge that despite the other good tokens in the experiment, German learners of English in general tend to use this sound as a substitution for the difficult sound. Hence the good learners were less confident to really hear the easy sound if it was produced by others than if it was produced by themselves. The low-proficiency participants, in contrast, had overall more trouble to identify the intended sounds since they themselves as well as the others they heard during the experiment had produced very small differences between the sounds of the contrasts, which made all words hard to identify (see also the main effect of proficiency). This suggests that the self-benefit is found relative to proficiency-matched other learners especially under difficult listening conditions.

Experiment 2 further investigated the role of the quality of the input in the perception of others' L2 productions. Being presented with rich material, that is, words that had been produced with more and clearer acoustic cues to the contrasts, enhanced word recognition in both proficiency groups compared to when hearing poor material. However, there was also an overall effect of proficiency: High-proficiency listeners outperformed learners from the low-proficiency group regardless of the availability of acoustic cues, that is, when presented with poor and when presented with rich productions. Given that “proficiency” was based on the learners’ productions, this confirms L2 models suggesting that perception and production abilities are somehow linked. Moreover, the finding that high-proficiency learners outperformed low-proficiency learners, but that this difference was larger when presented with rich material, suggests that the pattern that better producers make better perceivers is more complex, and depends also on the type of input. A comparison with productions in the learners’ own voices further revealed that the effect of material was stronger than the self-benefit. Crucially, however, this was true only for learners from the low-proficiency group, who themselves had produced only small cues to the contrasts. Those learners who have already acquired more advanced production skills also benefitted from rich compared to poor material, but even more when they heard words that have been produced by themselves.

While the idea that L2 perception and production abilities are somehow linked is commonly-accepted, the exact relation and direction of causation remains yet unclear. Studies comparing perception and production abilities in L2 learners show mixed results (e.g., Flege et al. 1997; Hattori & Iverson, 2010; Kartushina & Frauenfelder, 2014; Kartushina, Hervais-Adelman, Frauenfelder, & Golestani, 2015; Peperkamp & Bouchon, 2011; Schertz et al. 2015). Tsukada and colleagues (2005), for example, showed that Korean bilinguals who started to speak English early in life produced English vowels in a native-like fashion, but their ability to perceptually discriminate them failed to reach a native-like level (Tsukada et al. 2005). Results like these (see also, e.g., Kassaian, 2011; Kluge, Rauber, Reis & Bion, 2007; Sheldon & Strange, 1982) challenge the probably most common proposal that L2 perception leads production (Flege, 1995). Training studies also give insights into the relation between production and perception, with the specific focus on development, but again, results about cross-modal transfer of training are mixed (e.g. Bradlow, Pisoni, Akahane-Yamada, & Tohkura, 1997; Herd, Jongman, & Sereno, 2013; Kartushina et al. 2015). However, one drawback of these previous studies is that they often used different tasks and types of material, for instance, acoustic measurements or intelligibility ratings by native listeners to determine production skills on one hand, identification and discrimination tests using native productions or synthetic stimuli to determine perception skills on the other (e.g., Flege et al. 1997; Hattori & Iverson, 2010).

Although the results of the present study cannot decide on the direction of causality between improvements in perception and production abilities either, its contribution was to compare L2 perception by groups of listeners who had produced the stimuli themselves and were grouped based on their production abilities. In this sense, the present results contribute to the understanding of L2 production and perception in that they show how learners of different proficiency levels exploit the cues in naturally produced stimuli. The Speech Learning Model (Flege, 1995) proposes that perception may lead production, in the sense that "the production of an L2 phonetic segment will typically be no more native-like than its perceptual representation" (Flege, 2003: 322). The present study investigated in more detail how difficult L2 sound contrasts are produced and perceived by focusing on the use of relevant cues. They showed that both high and low-proficiency speakers were

able to make use of available acoustic cues in perception, but the more proficient learners to a larger extent. Critically, across all proficiency groups the present study showed that listeners are better able at reconstructing words they had produced themselves than other learners, when they were matched in proficiency (i.e., magnitude and type of produced cues). Although this effect was smaller than the effect of Material or overall Proficiency (as shown in Experiment 2, Figure 4) it provides further insights into the learners' representations of L2 speech.

That is, in terms of modelling the perception process (i.e., access to these representations), the present results indicate that learners make use of different (sub)sets of cues including speaker-independent cues as well as fine-tuned cues typical of their own productions. In addition to adapting to other speakers whose productions are characterized by non-native ways of differentiating difficult sounds or contrasts (e.g., Bradlow & Bent, 2008; Clarke & Garrett, 2004; Reinisch & Weber, 2012; Reinisch et al., 2013; Witteman et al. 2013), learners might also adapt to their own productions in a second language. These language, accent, and speaker-specific adapted sound targets could be stored and accessed depending on context (Kleinschmidt & Jaeger, 2015; see Reinisch, 2016a,b, for similar suggestions concerning other speaker-specific cues). In other words, slightly different auditory targets may be accessed depending on whether one's own voice is perceived or another, unfamiliar voice. The idea that listeners adapt to their own productions receives support from the link between production and perception such that experience with one's own productions differs from experience with others' primarily by the combined sensorimotor and auditory feedback the learner receives immediately during speaking (Guenther, 2006; Perkell, Guenther, Lane, Matthies, Stockmann, Tiede, & Zandipour, 2004). This coupling may contribute to stronger adaptation to one's own voice than to others and may be part of the reason why the self-benefit goes even beyond the mere interlanguage intelligibility benefit that should apply to all participants in the current study, when listening to other learners of a similar proficiency.

If familiarity with one's own specific production patterns was a critical factor in our self-perception benefit, the question arises as to how strong it would be relative to benefits of other familiar voices. As simple as it seems to test this issue, a well-controlled empirical study is hard to design as it would

be unclear how familiar, for example, pairs of friends would be to listening to each other in their second language - provided that their overall L2 proficiency was similar as well. However, the present comparison to others' voices that were closely matched in production proficiency and types of cues produced provides first insights into the contribution of one's own voice in processing L2 speech. One's own voice is special since every time a person speaks not only acoustic but also proprioceptive feedback is experienced (Guenther, 2006; Perkell et al. 2004). Critically, changes in a speaker's own production patterns can cause changes in perception of the relevant sound contrasts (Lametti, Rochet-Capellan, Neufeld, Shiller, & Ostry, 2014; Shiller, Sato, Gracco, & Baum, 2009). Given this special role of one's own voice, the experimental design was matched as much as possible a natural situation, in which one person is confronted with multiple other, unfamiliar voices, yet hearing one's own voice more than any single other. That is, one's own voice may always be the most familiar.

One prediction that falls out of our account on own-voice sound representations is that listeners must recognize their own voice for this effect to occur. Schuerman et al. (2015; see also Schuerman, 2017) provide tentative evidence for this. When listeners were asked to transcribe noise-vocoded words that originally had been spoken by themselves vs. a speaker whose voice represented the voice of an "average speaker" (i.e., voice characteristics were most similar to all voices used in the experiment) participants did not show a self-benefit. However, in this study most listeners did not recognize their own voice. In another study where listeners transcribed short sentences produced by themselves or others in speech-shaped background noise, a self-benefit did appear (Schuerman, 2017). In that study about half of the participants reported that they did recognize their own voice. In the present study, exclusively natural stimuli were used to investigate differences between the perception of words produced by participants themselves or others. Participants were informed that they may hear their own voice and importantly, they also reported to recognize themselves during the experiment. The present study is thus in line with the suggestion that in normal listening conditions where listeners are likely to recognize their own productions, they show enhanced perception abilities for their own productions.

A final question to our account is how the self-benefit found in the present study may be related to the observation that learners have difficulties improving their foreign accent. We speculate that better recognition of one's own productions and lower awareness of potential errors may be two sides of the same coin though findings may depend on the task. A case in point is Shuster (1998) who tested the identification of speech errors by children with a phonological disorder. There the children were worse at identifying their own erroneous productions as incorrect than judging other children's incorrectly produced sounds. In the present study, the task was not to report the correctness of the productions but to understand the words. We hypothesized that if a self-benefit can be found, this suggests that learners have adapted to their own accented (i.e., non-native) production patterns. The familiarity with one's own "errors" (that is, own productions that are not well differentiated) may appear as an advantage in reconstructing the intended word. We speculate that as the other side of the coin this benefit may be a drawback, since learners have fewer difficulties to understand their own than fellow learners' productions, which may be one reason why the need for improvement may not be obvious.

Summing up the present study, we found that learners of all proficiency groups are able to perceptually discern difficult second-language sound contrasts, given the availability of sufficient acoustic cues marking the sounds. Listeners who are better producers themselves show an advantage in exploiting cues to second language contrasts over poor producers especially if the cues are strong. Even though clear acoustic cues contribute most to understanding difficult L2 contrasts, listeners are better at recognizing second language words produced by themselves compared to when produced by an unfamiliar speaker using similar production patterns. We hypothesized that due to frequent exposure to their own accented speech, learners are highly familiar with their own L2 sound patterns - above and beyond the sound patterns that are typical of their L1's accent - and that this facilitates the identification of the words they had produced themselves. Future research will have to show how adaptation to one's own speech patterns and already acquired production skills in a L2 relate to the ease or difficulty to improve one's accent in a foreign language.

REFERENCES

Abbs, J. H., Gracco, V. L., & Cole, K. J. (1984). Control of multimovement coordination: Sensorimotor mechanisms in Speech motor programming. *Journal of Motor Behavior, 16*, 195-232. DOI: 10.1080/00222895.1984.10735318

Aruffo, C., Shore, D. I. (2012). Can you McGurk yourself? Self-face and self-voice in audiovisual speech. *Psychonomic Bulletin & Review, 19,* 66–72. DOI: 10.3758/s13423-011-0176-8

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*, 255-278. DOI: 10.1016/j.jml.2012.11.001

Barry, W. J. (1979). Complex encoding in word-final voiced and voiceless stops. *Phonetica*, *36*, 361-372. DOI: 10.1159/000259973

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*. DOI: 10.18637/jss.v067.i01

Bent, T., & Bradlow, A. R. (2003). The interlanguage speech intelligibility benefit. *The Journal of the Acoustical Society of America*, *114*, 1600-1610. DOI: 10.1121/1.1603234

Best, C. T., & Tyler, M. D. (2007). Nonnative and second-language speech perception: Commonalities and complementarities. In O.-S. Bohn, & M. J. Munro (Eds.), *Language experience in second language speech learning: In honor of James Emil Flege* (pp. 13-34). Amsterdam, NL: John Benjamins. DOI: 10.1075/lllt.17.07bes

Boersma, P., & Weenink, D. (2015). Praat: doing phonetics by computer [Computer program]. Version 5.4.08, retrieved 24 March 2015 from http://www.praat.org/

Bohn, O.-S. (1995). Cross-language speech perception in adults: First language transfer doesn't tell it all. In W. Strange (Ed.), *Speech Perception and Linguistic Experience: Issues in Cross-language Research* (pp. 279-304). Timonium MD: York Press.

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*, 131-158. DOI: 10.1017/s0272263100010792

Bradlow, A. R., & Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition*, *106*, 707-729. DOI: 10.1016/j.cognition.2007.04.005

Bradlow, A. R., Pisoni, D. B., Akahane-Yamada, R., & Tohkura, Y. I. (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*, 2299-2310.

Broersma, M. (2005). Perception of familiar contrasts in unfamiliar positions. *The Journal of the Acoustical Society of America, 117*, 3890-3901. DOI: 10.1121/1.1906060

Broersma, M. (2010). Perception of final fricative voicing: Native and nonnative listeners' use of vowel duration. *The Journal of the Acoustical Society of America, 127*, 1636-1644. DOI: 10.1121/1.3292996

Broersma, M. (2012). Increased lexical activation and reduced competition in second-language listening. *Language and Cognitive Processes*, *27*, 1205-1224. DOI: 10.1080/01690965.2012.660170

Cebrian, J. (2000). Transferability and productivity of L1 rules in Catalan-English interlanguage. *Studies in Second Language Acquisition*, *22*, 1-26. DOI: 10.1017/s0272263100001017

Cho, T., & McQueen, J. M. (2006). Phonological versus phonetic cues in native and non-native listening: Korean and Dutch listeners' perception of Dutch and English consonants. *The Journal of the Acoustical Society of America, 119*, 3085-3096. DOI: 10.1121/1.2188917

Clarke, C. M., & Garrett, M. F. (2004). Rapid adaptation to foreign-accented English. *The Journal of the Acoustical Society of America, 116*, 3647-3658. DOI: 10.1121/1.1815131

Cutler, A. (2015). Representation of second language phonology. *Applied Psycholinguistics*, *36*, 115-128. DOI: 10.1017/s0142716414000459

Deterding, D. (1997). The formants of monophthong vowels in Standard Southern British English pronunciation. *Journal of the International Phonetic Association*, *27*, 47-55. DOI: 10.1017/S0025100300005417

Devue, C., & Brédart, S. (2011). The neural correlates of visual self-recognition. *Consciousness and Cognition, 20*, 40-51. DOI: 10.1016/j.concog.2010.09.007

Douglas, W., & Gibbins, K. (1983). Inadequacy of voice recognition as a demonstration of self-deception. *Journal of Personality and Social Psychology*, *44*, 589-592. DOI: 10.1037/0022-3514.44.3.589

Draxler, C., & Jänsch, K. (2004). SpeechRecorder: A Universal Platform Independent Multi-Channel Audio Recording Software. In M. T. Lino, M. F. Xavier, F. Ferreira, R. Costa, & R. Silva (Eds.), *Proceedings of Language Resources and Evaluation* (pp. 559-562). Lisbon, Portugal: Universidade Nova de Lisboa.

Eger, N. A., & Bohn, O. S. (2015). Picking up the cues to a new consonant contrast: Danish learners' production and perception of English word-final /s/-/z/. In The Scottish Consortium for ICPhS 2015 (Ed.), *Proceedings of the 18th International Congress of Phonetic Sciences*. Glasgow, UK: the University of Glasgow. http://www.icphs2015.info/pdfs/Papers/ICPHS0648.pdf

Eger, N. A., & Reinisch, E. (in press). The role of acoustic cues and listener proficiency in the perception of accent in nonnative sounds. *Studies in Second Language Acquisition*. DOI: 10.1017/S0272263117000377

Escudero, P., Hayes-Harb, R., & Mitterer, H. (2008). Novel second-language words and asymmetric lexical access. *Journal of Phonetics*, *36*, 345-360. DOI: 10.1016/j.wocn.2007.11.002

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. 233-277). Timonium MD: York Press.

Flege, J. E. (2003). Assessing constraints on second-language segmental production and perception. In N. Schiller, & A. Meyer (Eds.), *Phonetics and Phonology in Language Comprehension and*

Production: Differences and Similarities (pp. 319–358). Berlin, Germany: Mouton de Gruyter. DOI: 10.1515/9783110895094.319

Flege, J. E., Bohn, O. S., & Jang, S. (1997). Effects of experience on non-native speakers' production and perception of English vowels. *Journal of Phonetics, 25*, 437-470. DOI: 10.1006/jpho.1997.0052

Flege, J. E., Munro, M. J., & Skelton, L. (1992). Production of the word-final English/t/–/d/contrast by native speakers of English, Mandarin, and Spanish. *The Journal of the Acoustical Society of America, 92*, 128-143. DOI: 10.1121/1.404278

Graux, J., Gomot, M., Roux, S., Bonnet-Brilhault, F., Camus, V., & Bruneau, N. (2013). My voice or yours? An electrophysiological study. *Brain Topography, 26*, 72-82. DOI: 10.1007/s10548-012-0233-2

Guenther, F. H. (2006). Cortical interactions underlying the production of speech sounds. *Journal of Communication Disorders*, *39*, 350-365. DOI: 10.1016/j.jcomdis.2006.06.013

Hattori, K., & Iverson, P. (2010). Examination of the relationship between L2 perception and production: an investigation of English/r/-/l/perception and production by adult Japanese speakers. In *Interspeech Workshop on Second Language Studies: Acquisition, Learning, Education and Technology*. Tokyo: Waseda University.

Hayes-Harb, R., Smith, B. L., Bent, T., & Bradlow, A. R. (2008). The interlanguage speech intelligibility benefit for native speakers of Mandarin: Production and perception of English word-final voicing contrasts. *Journal of Phonetics, 36*, 664-679. DOI: 10.1016/j.wocn.2008.04.002

Herd, W., Jongman, A., & Sereno, J. A. (2013). Perceptual and production training of intervocalic/d, ɾ, r/in American English learners of Spanish. *The Journal of the Acoustical Society of America, 133*, 4247-4255. DOI: 10.1121/1.4802902

Hillenbrand, J., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *The Journal of the Acoustical society of America*, *97*, 3099-3111. DOI: 10.1121/1.411872

Imai, S., Flege, J., & Walley, A. (2003). The recognition of accented and unaccented English words by native speakers of Spanish and English. *The Journal of the Acoustical Society of America, 113*, 2255. DOI: 10.1121/1.4780439

Ingram, J. C., & Park, S. G. (1997). Cross-language vowel perception and production by Japanese and Korean learners of English. *Journal of Phonetics*, *25*, 343-370. DOI: 10.1006/jpho.1997.0048

Ionta, S., Gassert, R., & Blanke, O. (2011). Multi-Sensory and Sensorimotor Foundation of Bodily Self-Consciousness – An Interdisciplinary Approach. *Frontiers in Psychology, 2*, 113-120. DOI: 10.3389/fpsyg.2011.00383

Iverson, P., Kuhl, P. K., Akahane-Yamada, R., Diesch, E., Tohkura, Y. I., Kettermann, A., & Siebert, C. (2003). A perceptual interference account of acquisition difficulties for non-native phonemes. *Cognition*, *87*, B47-B57. DOI: 10.1016/S0010-0277(02)00198-1

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language, 59*, 434-446. DOI: 10.1016/j.jml.2007.11.007

Kartushina, N., & Frauenfelder, U. H. (2014). On the effects of L2 perception and of individual differences in L1 production on L2 pronunciation. *Frontiers in Psychology*, *5*, 1246. DOI: 10.3389/fpsyg.2014.01246

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 a. *The Journal of the Acoustical Society of America*, *138*, 817-832.

Kassaian, Z. (2011). Age and gender effect in phonetic perception and production. *Journal of Language Teaching Research*, *2*, 370–376. DOI: 10.4304/jltr.2.2.370-376

Kleinschmidt, D. F., & Jaeger, T. F. (2015). Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological Review, 122*, 148. DOI: 10.1037/a0038695

Kluge, D. C., Rauber, A. S., Reis, M. S., & Bion, R. A. H. (2007). The Relationship between the Perception and Production of English Nasal Codas by Brazilian Learners of English. In *Proceedings of Interspeech* (pp. 2297-2300). Antwerp, Belgium.

Kuhl, P. K., Conboy, B. T., Coffey-Corina, S., Padden, D., Rivera-Gaxiola, M., & Nelson, T. (2008). Phonetic learning as a pathway to language: new data and native language magnet theory expanded (NLM-e). *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *363*, 979-1000. DOI: 10.1098/rstb.2007.2154

Lametti, D. R., Rochet-Capellan, A., Neufeld, E., Shiller, D. M., & Ostry, D. J. (2014). Plasticity in the human speech motor system drives changes in speech perception. *Journal of Neuroscience*, *34*, 10339-10346. DOI: 10.1523/JNEUROSCI.0108-14.2014

Levy, E. S., & Law II, F. F. (2010). Production of French vowels by American-English learners of French: Language experience, consonantal context, and the perception-production relationship. *The Journal of the Acoustical Society of America, 128*, 1290-1305. DOI: 10.1121/1.3466879

Llompart & Reinisch, (2017). Articulatory information helps encode lexical contrasts in a second language. *Journal of Experimental Psychology: Human Perception and Performance, 43*, 1040-1056. doi: http://dx.doi.org/10.1037/xhp0000383

McAllister, R., Flege, J. E., & Piske, T. (2002). The influence of L1 on the acquisition of Swedish quantity by native speakers of Spanish, English and Estonian. *Journal of Phonetics*, *30*, 229-258. DOI: 10.1006/jpho.2002.0174

Munro, M. J., Derwing, T. M., Morton, S. L. (2006). The mutual intelligibility of L2 speech. *Studies in Second Language Acquisition, 28*, 111-131. DOI: 10.1017/S0272263106060049

Peirce, J. W. (2007). PsychoPy—psychophysics software in Python. *Journal of Neuroscience Methods*, *162*, 8-13. DOI: 10.1016/j.jneumeth.2006.11.017

Peperkamp, S., & Bouchon, C. (2011). The relation between perception and production in L2 phonological processing. In *Proceedings of Interspeech* (pp. 161-164). Florence, Italy.

Perkell, J. S., Guenther, F. H., Lane, H., Matthies, M. L., Stockmann, E., Tiede, M., & Zandipour, M. (2004). The distinctness of speakers' productions of vowel contrasts is related to their

discrimination of the contrasts. *The Journal of the Acoustical Society of America, 116*, 2338-2344. DOI: 10.1121/1.1787424

Pinet, M., Inverson, P., & Huckvale, M. (2011). Second/language experience and speech-in-noise-recognition: Effects of talker-listener accent similarity*. The Journal of the Acoustical Society of America, 130,* 1653-1662. DOI: 10.1121/1.3613698

Platek, S. M., Burch, R. L., & Gallup, G. G. (2001). Sex differences in olfactory self-recognition. *Physiology & Behavior, 73*, 635-640. DOI: 10.1016/s0031-9384(01)00539-x

Port, R. F., & Dalby, J. (1982). Consonant/vowel ratio as a cue for voicing in English. *Perception & Psychophysics*, *32*, 141-152. DOI: 10.3758/bf03204273

Postma, A. (2000). Detection of errors during speech production: A review of speech monitoring models. *Cognition, 77*, 97-132. DOI: 10.1016/s0010-0277(00)00090-1

R Core Team (2017). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/.

Reinisch, E. (2016a). Natural fast speech is perceived as faster than linearly time-compressed speech. *Attention, Perception, & Psychophysics, 78*, 1203-1217. DOI: 10.3758/s13414-016-1067-x

Reinisch, E. (2016b). Speaker-specific processing and local context information: the case of speaking rate. *Applied Psycholinguistics*, 37, 1397-1415. DOI: http://dx.doi.org/10.1017/S0142716415000612

Reinisch, E., Weber, A. (2012). Adapting to suprasegmental lexical stress errors in foreign-accented speech. *The Journal of the Acoustical Society of America, 132*, 1165-1176. DOI: 10.1121/1.4730884

Reinisch, E., Weber, A., & Mitterer, H. (2013). Listeners retune phoneme categories across languages. *Journal of Experimental Psychology: Human Perception and Performance, 39*, 75-86. DOI: 10.1037/a0027979

Schertz, J., Cho, T., Lotto, A., & Warner, N. (2015). Individual differences in phonetic cue use in production and perception of a non-native sound contrast. *Journal of Phonetics*, *52*, 183-204. DOI: 10.1016/j.wocn.2015.07.003

Schertz, J., Cho, T., Lotto, A., & Warner, N. (2016). Individual differences in perceptual adaptability of foreign sound categories. *Attention, Perception, & Psychophysics*, *78*, 355-367. DOI: 10.3758/s13414-015-0987-1

Schuerman, W. L. (2017). *Sensorimotor Experience in Speech Perception.* Doctoral dissertation. Radboud University Nijmegen.

Schuerman, W. L., Meyer, A., & McQueen, J. M. (2015). Do we perceive others better than ourselves? A perceptual benefit for noise-vocoded speech produced by an average speaker. *PLoS ONE*, 10: e0129731. DOI: 10.1371/journal.pone.0129731

Sheldon, A., and Strange, W. (1982). The acquisition of /r/ and /l/ by Japanese learners of English: evidence that speech production can precede speech perception. *Applied Psycholinguist, 3*, 243–261. doi: 10.1017/S0142716400001417

Shiller, D. M., Sato, M., Gracco, V. L., & Baum, S. R. (2009). Perceptual recalibration of speech sounds following speech motor learning. *The Journal of the Acoustical Society of America*, *125*, 1103-1113. DOI: 10.1121/1.3058638

Shuster, L. I. (1998). The perception of correctly and incorrectly produced /r/. *Journal of Speech, Language, and Hearing Research, 41*, 941-950. DOI: 10.1044/jslhr.4104.941

Shuster, L. I., Durrant, J. D. (2003). Toward a better understanding of the perception of self-produced speech. *Journal of Communication Disorders*, *36*, 1-11. DOI: 10.1016/s0021-9924(02)00132-6

Sidaras, S. K., Alexander, J. E., & Nygaard, L. C. (2009). Perceptual learning of systematic variation in Spanish-accented speech. *The Journal of the Acoustical Society of America, 125*, 3306-3316. DOI: 10.1121/1.3101452

Smith, B. L., & Hayes-Harb, R. (2011). Individual differences in the perception of final consonant voicing among native and non-native speakers of English. *Journal of Phonetics*, *39*, 115-120. DOI: 10.1016/j.wocn.2010.11.005

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*, 257-275. DOI: 10.1016/j.wocn.2009.03.001

Strömbergsson, S., Wengelin, Å., & House, D. (2014). Children's perception of their synthetically corrected speech production. *Clinical Linguistics & Phonetics*, *28*, 373-395. DOI: 10.3109/02699206.2013.868928

Tsukada, K., Birdsong, D., Bialystok, E., Mack, M., Sung, H., & Flege, J. (2005). A developmental study of English vowel production and perception by native Korean adults and children. *Journal of Phonetics*, *33*, 263-290. DOI: 10.1016/j.wocn.2004.10.002

van Wijngaarden, S. J., Steeneken, H. J., & Houtgast, T. (2002). Quantifying the intelligibility of speech in noise for non-native listeners. *The Journal of the Acoustical Society of America, 111,* 1906-1916. DOI: 10.1121/1.1456928

Weber, A., Broersma, M., & Aoyagi, M. (2011). Spoken-word recognition in foreign-accented speech by L2 listeners. *Journal of Phonetics*, *39*, 479-491. DOI: 10.1016/j.wocn.2010.12.004

Weber, A., Di Betta, A. M., & McQueen, J. M. (2014). Treack or trit: Adaptation to genuine and arbitrary foreign accents by monolingual and bilingual listeners*. Journal of Phonetics*, *46*, 34-51. DOI: 10.1016/j.wocn.2014.05.002

Witteman, M. J., Weber, A., & McQueen, J. M. (2013). Foreign accent strength and listener familiarity with an accent codetermine speed of perceptual adaptation. *Attention, Perception, & Psychophysics*, *75*, 537-556. DOI: 10.3758/s13414-012-0404-y

Wright, R. (2004). A review of perceptual cues and cue robustness. In B. Hayes, R. Kirchner, & D. Steriade (Eds.), *Phonetically Based Phonology* (pp. 34-57). Cambridge: University Press.

Xie, X., & Fowler, C. A. (2013). Listening with a foreign-accent: The interlanguage speech intelligibility benefit in Mandarin speakers of English. *Journal of Phonetics, 41*, 369-378. DOI: 10.1016/j.wocn.2013.06.003

Xu, M., Homae, F., Hashimoto, R. I., & Hagiwara, H. (2013). Acoustic cues for the recognition of self-voice and other-voice. *Frontiers in Psychology, 4*. DOI: 10.3389/fpsyg.2013.00735

NOTES

[1] Similar predictions based on phonetic (dis-)similarities between first and second language sound contrasts are made by the Perceptual Learning Model (PAM-L2, Best & Tyler, 2007) and the Native Language Magnet Model (NLM-e, Kuhl, Conboy, Coffey-Corina, Padden, Rivera-Gaxiola, & Nelson, 2008).

[2] Since participants had to listen to their own voice and other unfamiliar voices, we decided to restrict participants to one gender in order to avoid large acoustic differences between their own and other's voices.

[3] The words *bet*, *bat*, *bed* and *bad* were used for the stop voicing contrast as well as for the vowel contrast, but each word was recorded only twice.

[4] The words *bet*, *bat*, *bed* and *bad* were used for the stop voicing contrast as well as for the vowel contrast, therefore the word set consisted of 27 pairs (25 plus additional 2) but only 50 single words.

[5] Note that we refer to production abilities here, since we grouped participants by production measures. However, we cannot determine the direction of causality. Our high-proficient participants could be good producers because they are good perceivers rather than the other way around. This will be further considered in the General Discussion.

Tables

Table 1: Results of the mixed-effects model fitted with all factors (Voice, Proficiency, and Sound Type) in Experiment 1.

Fixed effect	*b*	SE	*z*	*p*
Intercept	1.23	0.16	7.89	<.001
Voice	0.22	0.08	2.89	<.01
Sound Type	-0.26	0.15	-1.76	=.08
Proficiency	0.56	0.12	4.80	<.001
Voice:Sound Type	0.13	0.21	0.61	=.54
Voice:Proficiency	-0.15	0.15	-0.99	=.32
Sound Type:Proficiency	-0.77	0.21	-3.61	<.001
Voice:Sound Type:Proficiency	1.00	0.30	3.29	<.001

Table 2: Results of the mixed-effects model to compare the effects of Proficiency, Material, and Voice with reference to the poor listeners and rich material (i.e., *C rich* mapped on the intercept; see text for details) in Experiment 2.

Fixed effect	*b*	SE	*z*	*p*
Intercept (C rich)	1.14	0.12	9.13	<.001
C self	-0.66	0.07	-9.29	<.001
C poor	-0.72	0.05	-14.42	<.001
A rich	0.86	0.13	6.53	<.001
A self	1.39	0.16	8.72	<.001
A poor	-0.50	0.13	-3.92	<.001

Table 3: Results of the mixed-effects model fitted with Material, Proficiency, Contrast, and their interactions in Experiment 2.

	Fixed effect	*b*	SE	*z*	*p*
Vowels	Intercept	0.42	0.16	2.62	<.01
	Material	0.73	0.19	3.94	<.001
	Proficiency	0.55	0.23	2.38	<.05
	Proficiency:Material	0.96	0.29	3.32	<.001
Fricatives	Contrast	0.45	0.27	1.65	=.10
	Material:Contrast	0.59	0.31	1.89	=.06
	Proficiency:Contrast	0.04	0.36	0.12	=.91
	Proficiency:Material:Contrast	0.11	0.48	0.23	=.82
Stops	Contrast	1.51	0.13	11.51	<.001
	Material:Contrast	0.65	0.28	2.32	<.05
	Proficiency:Contrast	-0.21	0.20	-1.08	=.28
	Proficiency:Material:Contrast	-0.66	0.47	-1.40	=.16

Table 4: Results of the mixed-effects model fitted with Material, Proficiency and Sound Type in Experiment 2.

Fixed effect	*b*	SE	*z*	*p*
Intercept	1.14	0.12	9.56	<.001
Material	1.16	0.14	8.39	<.001
Proficiency	0.58	0.15	3.89	<.001
Sound Type	0.12	0.20	0.66	=.51
Proficiency:Material	0.67	0.20	3.26	<.01
Proficiency:Sound Type	-0.11	0.33	-0.32	=.75
Material:Sound Type	-0.43	0.20	-2.16	<.01
Proficiency:Material:Sound Type	-0.52	0.27	-1.89	=.06

Figure captions

Figure 1: Proportion of correct responses in Experiment 1 for three proficiency groups, shown for participants' own voice (self) and others' voices, averaged over contrast. Data points are shown aggregated over repetitions and words. Chance performance is at 0.5.

Figure 2: Proportion of correct responses in Experiment 1 for the three proficiency groups, shown for easy and difficult sounds, averaged over contrast. Data points are shown aggregated over repetitions and words. Chance performance is at 0.5.

Figure 3: Illustration of the three-way interaction between Sound Type (easy, difficult), Proficiency (A, B, C), and Voice (self, other) in Experiment 1. Proportion of correct responses are averaged over contrast. Data points are shown aggregated over repetitions and words. Chance performance is at 0.5.

Figure 4: Proportion of correct responses in Experiment 2 for the two material conditions (rich and poor) and the tokens in the participants' own voices, shown for two subgroups of participants from the proficiency categories A and C, averaged over contrast. A subset of the data was added from Experiment 1 (the responses to "self"-trials and to tokens produced by others from the same proficiency group). Data points are shown aggregated over repetitions and words. Chance performance is at 0.5. The box with thick lines refers to the condition that was mapped onto the intercept in the statistical analyses (see text for details).

Figure 5: Proportion of correct responses in Experiment 2 for the two proficiency groups of participants (A and C), for poor and rich speech Material split by Sound Type (i.e., easy vs. difficult sounds). Data points are shown aggregated over repetitions and words. Chance performance is at 0.5.

Figures

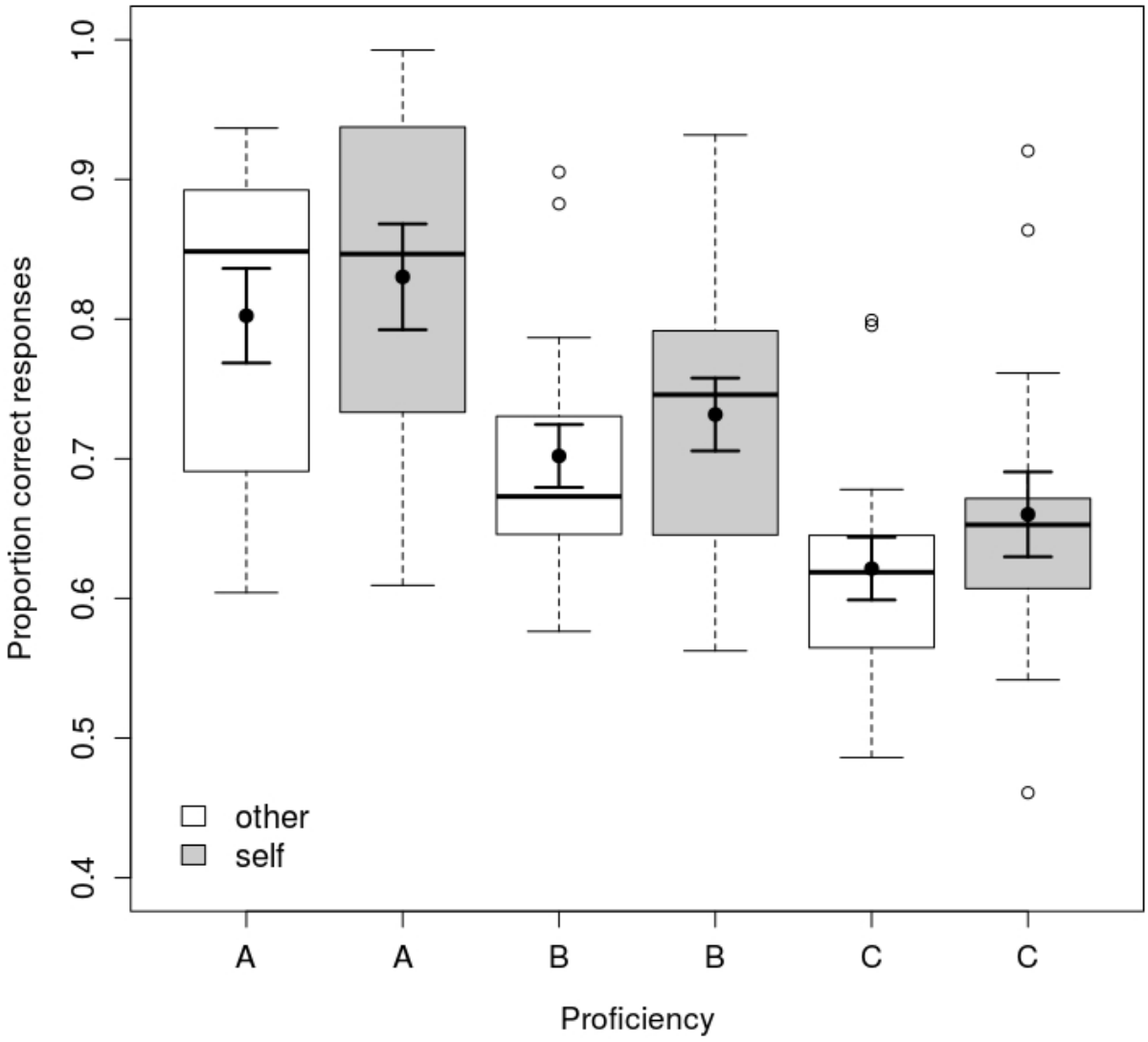

Figure 1

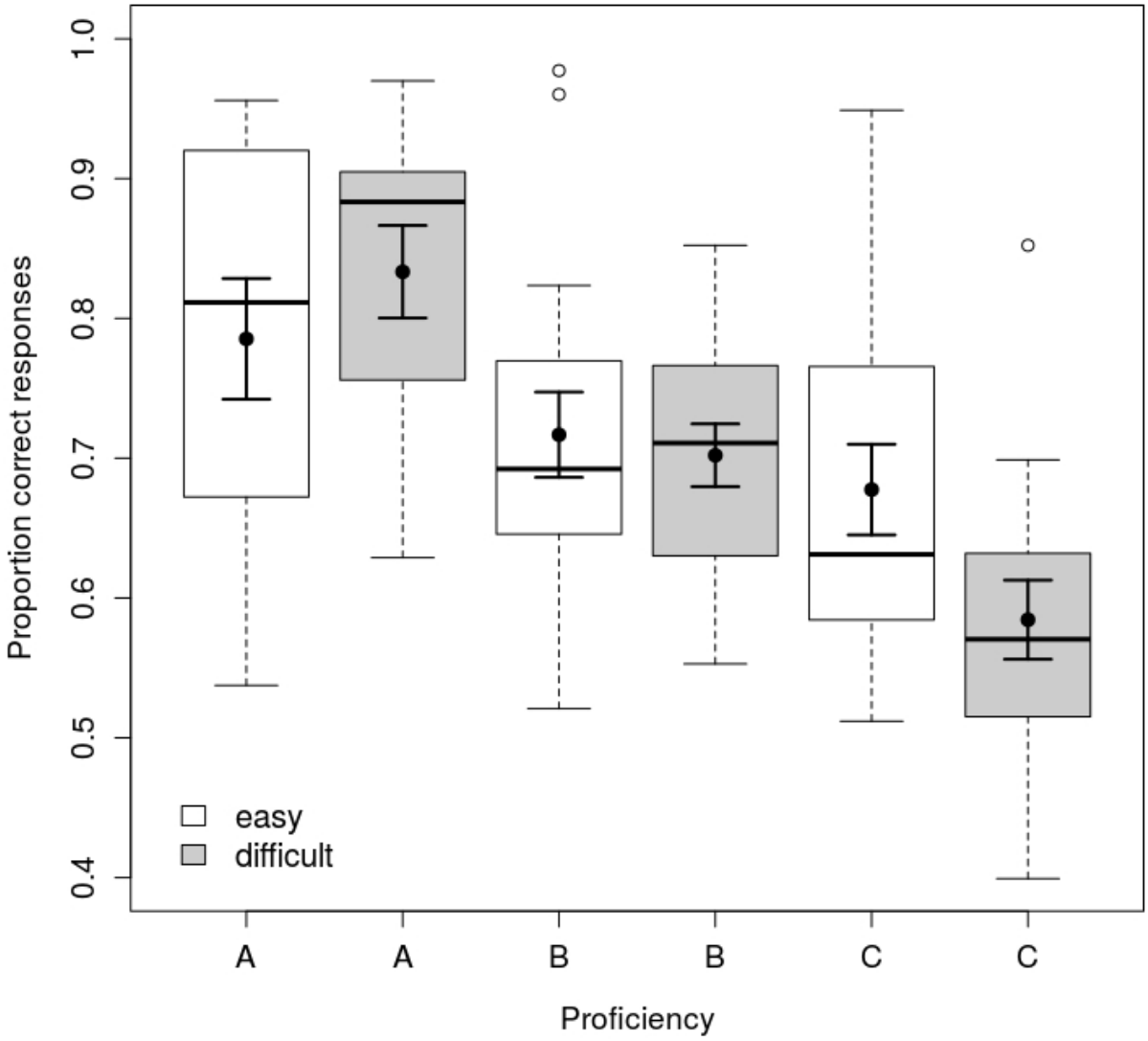
Proportion correct responses
0.4
0.5
0.6
0.7
0.8
0.9
1.0
easy
difficult
A
A
B
B
C
C
Proficiency

Figure 2

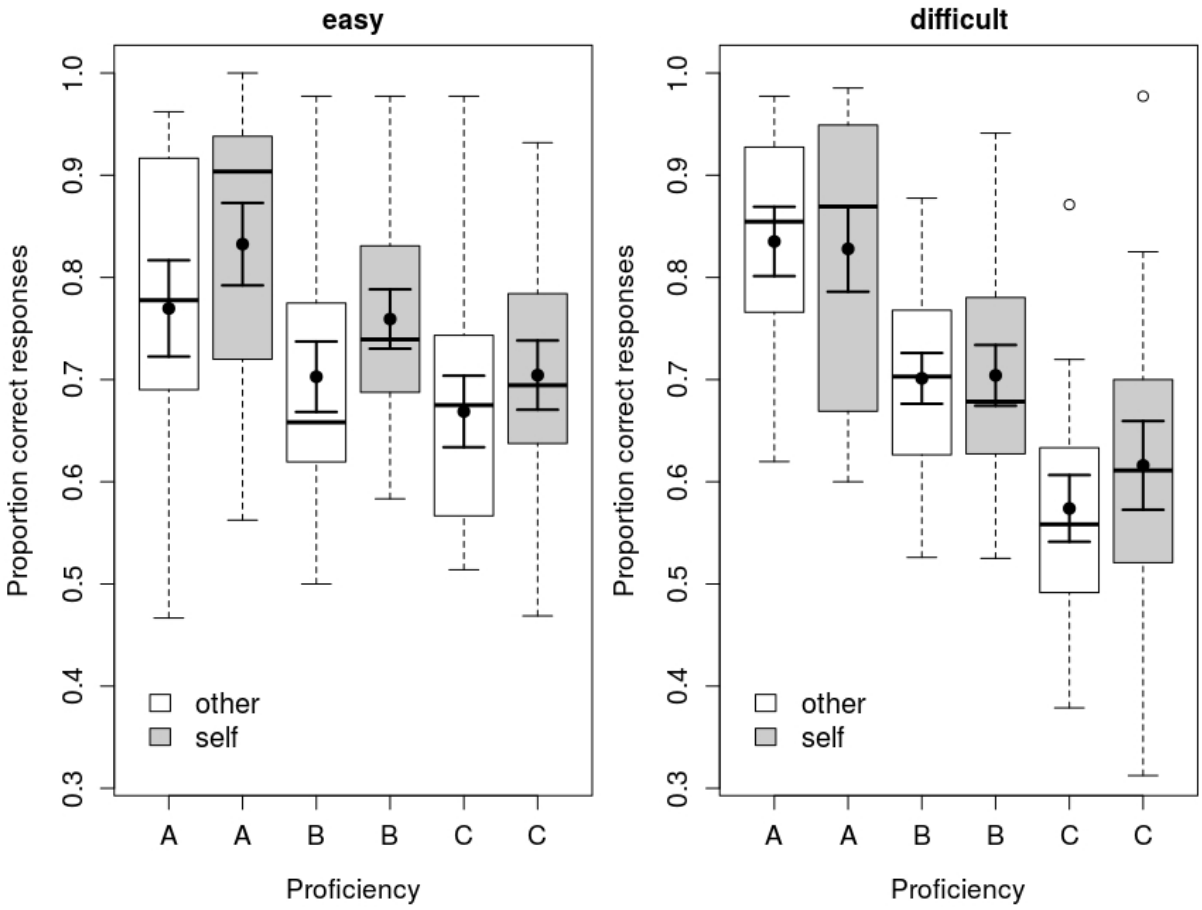
easy
difficult
Proportion correct responses
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
other
self
A
A
B
B
C
C
Proficiency

Figure 3

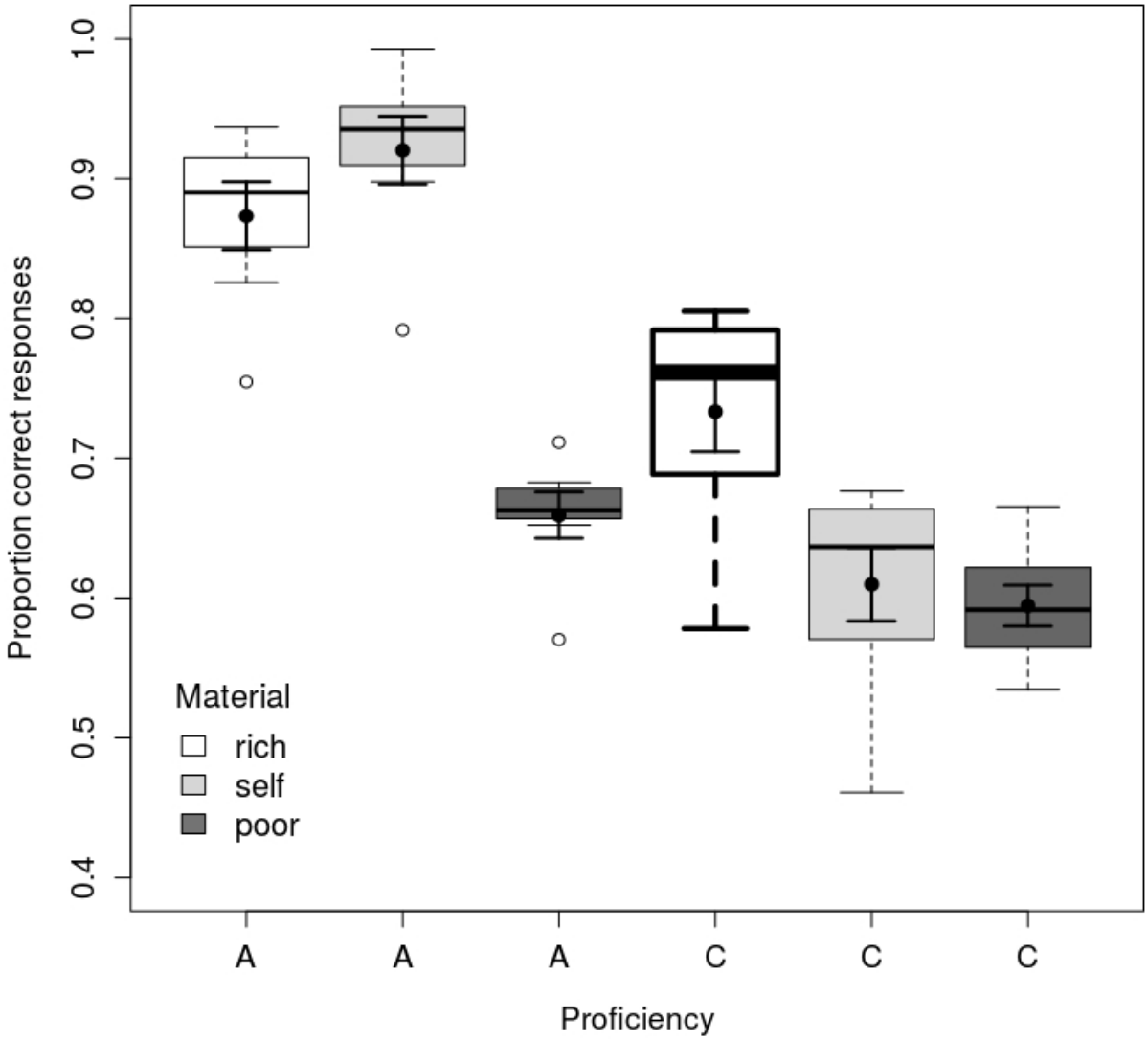
Proportion correct responses
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Material
rich
self
poor
A
A
A
C
C
C
Proficiency

Figure 4

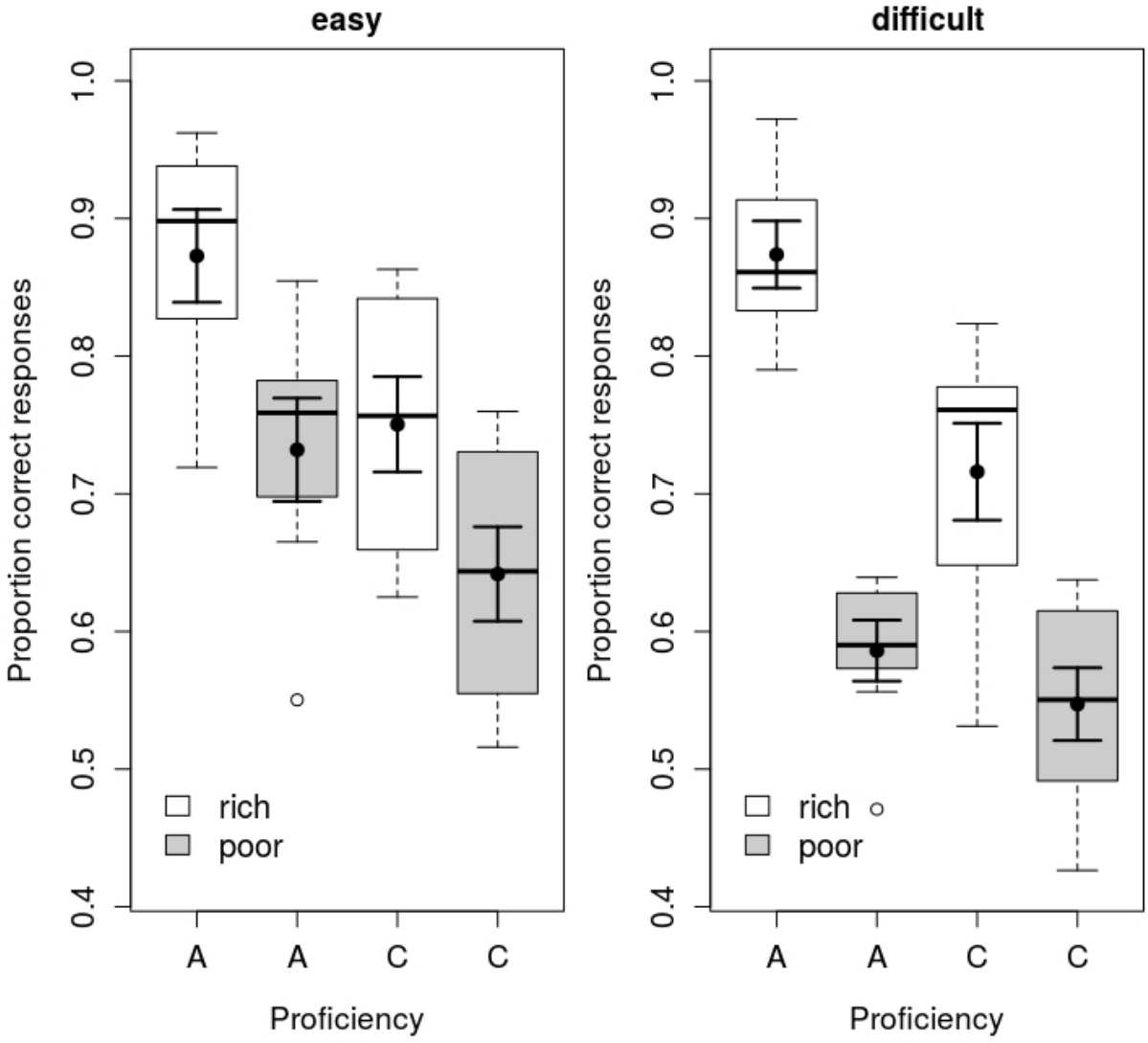
easy
difficult
Proportion correct responses
1.0
0.9
0.8
0.7
0.6
0.5
0.4
rich
poor
A
A
C
C
Proficiency

Figure 5

APPENDIX

APPENDIX A: Materials

Table A.1: Words and word pairs that were recorded in the production session. For the minimal pairs, the word after the dash is the one containing the critical difficult sound. The starred words and their counterparts were recorded and analyzed but excluded from the materials for the perception experiment because several speakers pronounced the vowels in “broad”, “height” and “prove” differently than in the other word of the pair. The word “latter” and its counterpart were excluded because several participants reported that they did not know the meaning of the word.

Vowels /ɛ/ – /æ/	Fricatives	Stops	Filler words	
bed – bad	ice – eyes	back – bag	car	piece
bet – bat	face – phase	bat – bad	eat	shine
dead – Dad	leaf – leave	bet – bed	fine	ship
flesh – flash	proof – prove*	bright – bride	force	sing
head – had	race – raise	brought – broad*	fourth	sit
letter – latter**	rice – rise	feet – feed	forest	skin
men – man	safe – save	heart – hard	get	state
merry – marry		height* – hide	honest	strong
pen – pan		pick – pig	king	time
send – sand		root – rude	kiss	worse
set – sat		rope – robe	nine	worth
		sight – side		
		white – wide		

Table A.2: Carrier sentences used in the production task. Sentences and words were randomly paired for each participant.

Number	Sentence
1	Here is the word
2	She forgot the word
3	He knows the word
4	You say the word
5	You read the word
6	The next word is
7	The correct term is
8	The next term is
9	The next expression is
10	The right expression is

APPENDIX B: Acoustic measures

The Figures below show a selection of acoustic measures that had been taken to determine the produced difference between the words of the minimal pairs for each of the three sound contrasts. These measures were used to assign participants to proficiency groups. For the vowels, the first two formants and duration were measured. For the word-final fricatives, the duration of the preceding vowel and the fricative were combined to the ratio between vowel duration and fricative duration. In addition the voiced portion of the fricative was measured. For the word-final stops, the duration of the aspiration, the duration of the preceding vowel and the voiced portion of the closure were taken into account. Cues to each contrast were weighted in the order named above.

The speakers were assigned one by one to the proficiency groups A, B, and C. Since the whole group of participants had to be distributed, they were split into three groups of 8 speakers each. First, the eight speakers with the clearest contrasts, according to the cues listed above, were assigned to group A. Then, the eight speakers with the smallest contrasts were selected for group C. The remaining eight participants were assigned to group B (see also Method section). In the Figures below, the acoustic measures are averaged over the two repetitions and words. The variability is hence due to inter-speaker differences (8 speakers per box).

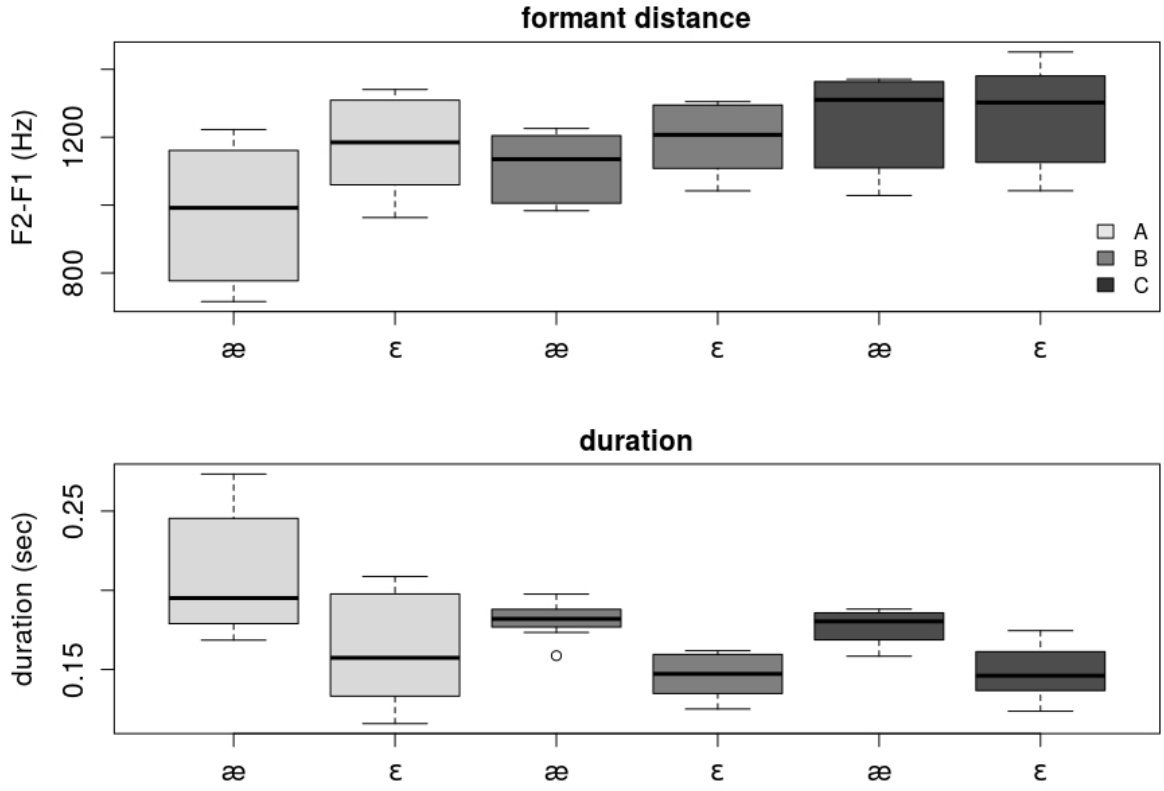

Figure B.1 Upper Panel: Formant values measured as the difference between F2 and F1 in Hz during a stable segment in the vowel for words with either /æ/ or /ɛ/ for the German learners grouped into three groups of 8 (light grey = group A, mid grey = group B, dark grey = group C); Lower Panel: Duration values of the entire vowel for words with either /æ/ or /ɛ/ and the different groups.

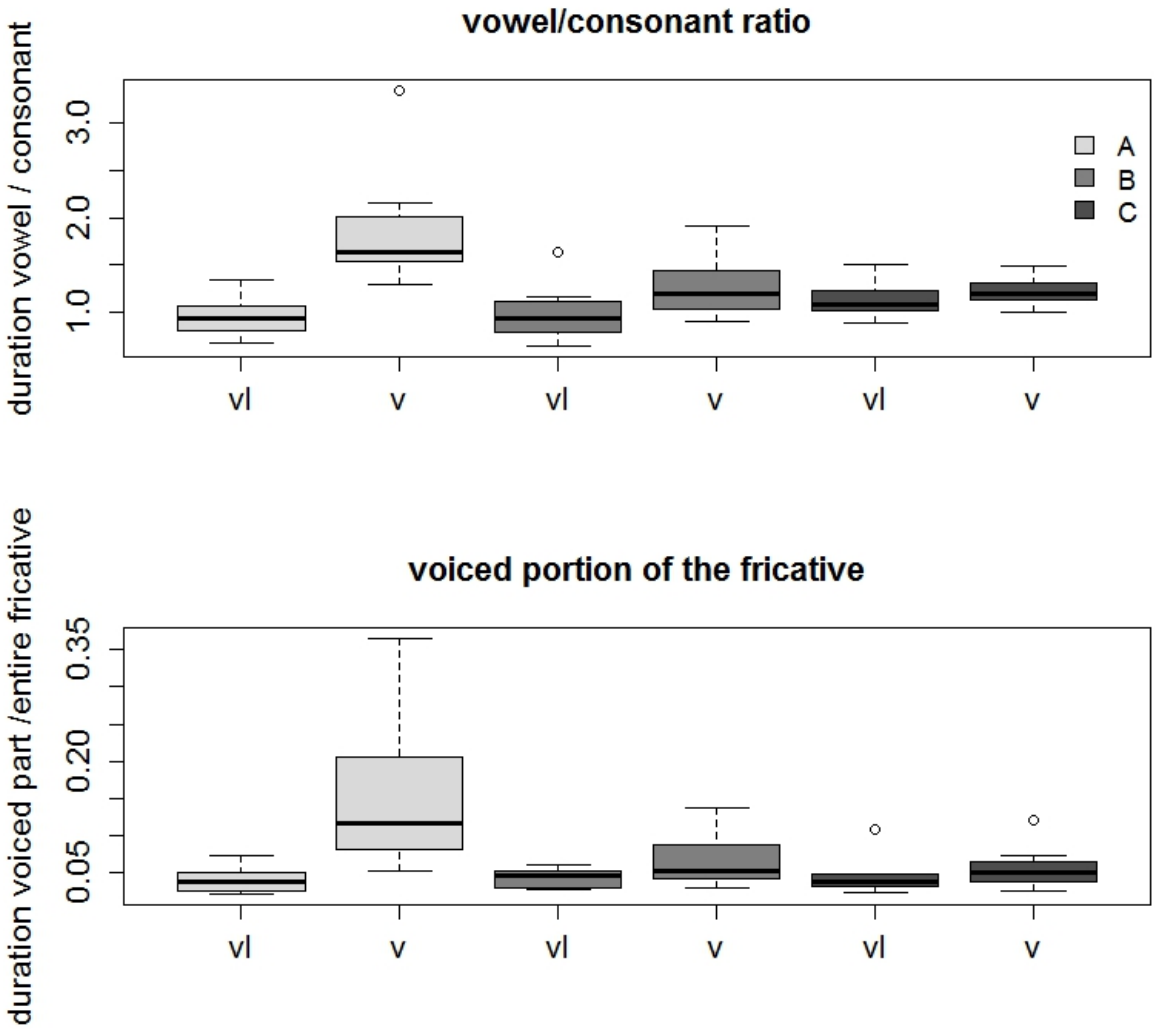

Figure B.2 Upper panel: Vowel/consonant ratios measured as the duration of the vowel divided by the duration of the consonant in words ending in voiced (v) or voiceless (vl) fricatives for the German learners grouped into three groups of 8 (light grey = group A, mid grey = group B, dark grey = group C); Lower Panel: Voiced portion of the fricative measured as the duration of the voiced part of the fricative divided by the total duration.

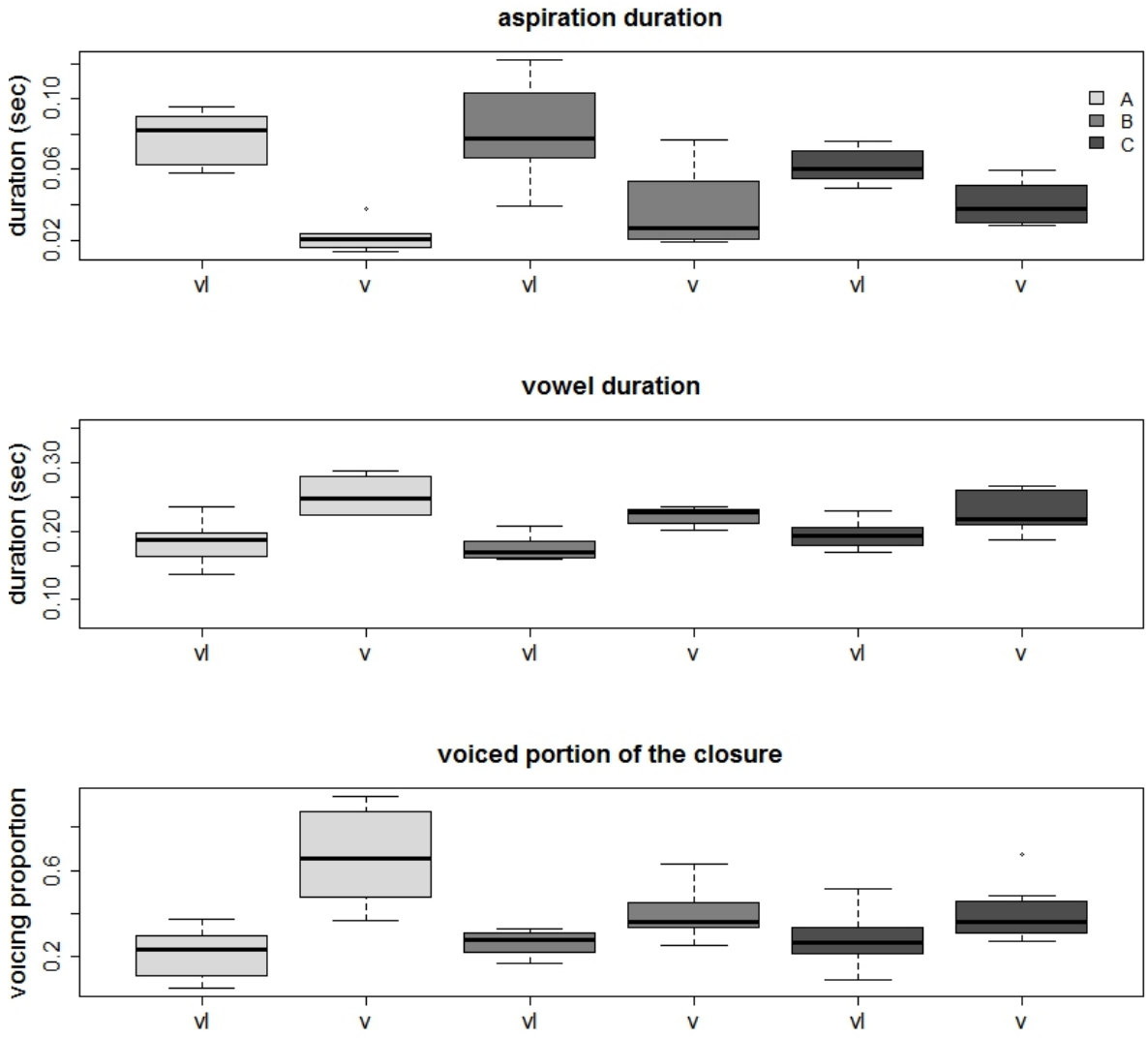

Figure B.3 Top panel: Aspiration duration for words ending in either voiced (v) or voiceless (vl) stops for the German learners grouped into three groups of 8 (light grey = group A, mid grey = group B, dark grey = group C); Mid panel: Duration of the preceding vowel; Bottom panel: Voiced portion of the closure measured as the duration of the voicing during closure divided by the total closure duration. As all other words, words containing a word-final stop were embedded in the end of carrier sentences. All word-final stops were produced as released stops.