LISTEN-AND-REPEAT TRAINING OF NON-NATIVE VOWEL QUALITY – PRELIMINARY FINDINGS FROM SPEAKERS OF NAMIBIAN LANGUAGES

Katja Haapanen[1], Antti Saloranta[1], Kimmo U. Peltola[1], Henna Tamminen[1], Lannie Uwu-khaeb[3], Paavo Alku[2], Maija S. Peltola[1]

[1]Phonetics and Learning, Age & Bilingualism Laboratory, University of Turku, Finland

[2]Department of Information and Communications Engineering, Aalto University, Finland

[3]FutureTech Lab, University of Turku in Windhoek, Namibia

Abstract

This study investigated how listen-and-repeat training affects the production of non-native (Swedish) vowels /y/ and /ʉ/ by speakers of different Namibian languages. 17 speakers, who did not have /y/ or /ʉ/ categories in their L1, participated in the experiment. Training effects were measured with acoustic analysis and an identification task performed by 40 proficient Swedish speakers to see whether the acoustic quality and the perceptual salience of the speakers' non-native production evolved during training. We expected the speakers' production to change as a function of training and the change to be reflected on the vowel formant values and the identification task. The results showed that the speakers produced /ʉ/ close to the trained target already in the first session, but changed their production away from the target after the first training. The speakers' production of /y/ did not change significantly. The speakers did not reach a perceivable spectral difference between the two non-native vowels. The participants' productions remained inconsistent throughout the experiment. There was great inter-speaker variation, which could not be accounted for by the speakers' language backgrounds.

Keywords: non-native speech learning, vowel production, vowel quality, phonetic training, identification

Introduction

Learning to speak a foreign language poses the learner with the challenge of learning new speech sound perception and production patterns that might not be relevant in their native language (L1). This study investigated whether listen-and-repeat training affects non-native vowel quality production by speakers of Namibian languages by studying both the acoustical characteristics of the produced speech and the listener perception. The study aimed to answer the following questions: First, can adult native speakers of different Namibian languages learn to produce the theoretically difficult non-native sounds /y/ and /ʉ/, which are phonemic in Swedish, through short-term listen-and-repeat training? Second, does listen-and-repeat training affect the identification of the speakers' productions by listeners who are familiar with the trained vowel categories?

Theories of second language sound learning

Theories of second language (L2) sound learning, such as the Speech Learning Model (SLM and the revised SLM-r) (Flege, 1995; Flege & Bohn, 2021) and the Perceptual Assimilation Model (PAM and the modified PAM-L2) (Best, 1994, 1995; Best & Tyler, 2007) predict that challenges in L2 speech sound perception and production arise from the relative differences between L1 and L2 phonological categories. In other words, if an L2 sound is not phonologically relevant in the speaker's L1, the perception and production of the L2 sound can cause differing degrees of difficulties before a new L2 category is formed. The new revised version of the Speech Learning model (SLM-r) more specifically states that the degree of relative difference between the L1 and L2 sounds determine the degree of difficulties in L2 sound perception and production (Flege & Bohn, 2021). In addition, the SLM-r proposes that the precision of L1 sound categories along with the quality and quantity of L2 input influence the formation of new L2 phonetic categories (Flege & Bohn, 2021). The PAM and PAM-L2 predict that when two L2 sounds assimilate equally to one L1 category (single category assimilation), the perception and production of the L2 sounds is the most difficult for learners (Best,

1994, 1995; Best & Tyler, 2007). The PAM-L2 also predicts that learners are likely to have initial difficulties in the perception and production of two L2 sounds that assimilate to one L1 category unevenly (category goodness assimilation), i.e., one of the L2 sounds is perceived as a better exemplar of the L1 sound (Best & Tyler, 2007).

The original SLM proposed that the perception of L2 sound categories limits the accuracy of L2 category production (Flege, 1995). The SLM-r, however, suggests that there is a bidirectional connection between L2 segmental perception and production. In other words, according to the SLM-r, the perception and production of L2 sound categories coevolve without one preceding the other (Flege & Bohn, 2021). Therefore, according to the predictions of the SLM-r, the perceptual component of the listen-and-repeat training paradigm should also in part support non-native sound production learning. This proposition is supported by previous studies showing both passive listening and listen-and-repeat training to be effective methods for L2 sound production (Bradlow et al., 1997, 1999; Immonen et al., 2021; Immonen, Alku, et al., 2022; Immonen, Peltola, et al., 2022; Iverson et al., 2012; K. U. Peltola et al., 2017, 2020; Taimi et al., 2014) and perception learning (Bradlow et al., 1997, 1999; Iverson et al., 2012; Saloranta et al., 2020; Tamminen et al., 2015, 2021; Ylinen et al., 2010) in different L1 and age groups.

Phonetic training studies

Several studies have examined the effectiveness of different phonetic training methods in L2 sound quality perception and production learning. For example, a study by Bradlow et al. (1997) found that perceptual training of the English /r/–/l/ contrast had a significant effect on L1 Japanese speakers' /r/–/l/ perception and production. The subjects participated in 45 sessions of high-variability identification training with feedback over 3–4 weeks. The results showed that the Japanese participants' identification of English /r/ and /l/ improved significantly as a function of training and, most interestingly, that their /r/–/l/ productions received higher identification scores from L1 English

listeners after training. A follow-up study provided evidence that the improvements in /r/–/l/ perception and production were retained even after three months, suggesting long-term learning effects as a function of high-variability phonetic training (Bradlow et al., 1999). In addition, Iverson et al. (2012) found that both experienced and inexperienced L2 learners benefitted from high-variability auditory training of English vowels when learning was measured with perception and production tests. The high-variability auditory training paradigm used by Iverson et al. (2012) included eight 45-minute sessions of English vowel identification tasks with feedback over the course of 1–2 weeks. The results showed modest but significant improvements in vowel identification and production as well as category discrimination in both groups (experienced and inexperienced) as a function of training. Another high-variability training study by Ylinen et al. (2010) found that phonetic training changed the cue weighting of L2 sounds when training effects were monitored with behavioral identification tasks and pre-attentive mismatch negativity (MMN) response recordings. The participants were native Finnish speakers and the trained stimuli were the English vowels /i/ and /ɪ/. The participants completed ten 25-minute sessions of identification training with feedback in three weeks. The behavioral and MMN results showed that L1 Finnish speakers' cue weighting of the trained vowels changed after training as they started to rely more on spectral cues over duration cues. Taken together, these previous findings offer evidence that perceptual phonetic training supports L2 sound perception and production learning (Bradlow et al., 1997; Iverson et al., 2012; Ylinen et al., 2010) and that learning effects can even be retained after several months (Bradlow et al., 1999). Although these studies have mainly focused on high-variability phonetic training methods, which were not used in the present study, the results offer relevant empirical insight into the effects of phonetic training on non-native speech sound quality learning.

For the purposes of this study, previous studies on the effects of phonetic listen-and-repeat training on L2 sound learning are of special interest. Some studies have offered evidence that listen-and-repeat training can result in improved L2 sound perception. For example, two studies by

Tamminen et al. (2015, 2021) examined how a three-day listen-and-repeat training protocol without feedback affected L2 speech perception in adult (Tamminen et al., 2015) and elderly (Tamminen et al., 2021) learners. L1 Finnish speakers were trained using stimuli containing the English fricative voicing contrast /f/–/v/. Training effects were measured with behavioral measures (discrimination, reaction time, identification and goodness rating) as well as pre-attentive MMN recordings in both experiments. The results showed clear training effects on both behavioral and pre-attentive perception of the trained contrast for the adult participants (Tamminen et al., 2015). In the case of elderly learners, the training resulted in improvements in identification and goodness rating, but not in discrimination or the MMN (Tamminen et al., 2021). In addition, Saloranta et al. (2020) examined the effects of listen-and-repeat training on the perception of vowel duration contrasts. Speakers with L1s that did not include phonological quantity contrasts participated in four sessions of listen-and-repeat training without feedback over two consecutive days. The trained contrast was /tite/–/tiːte/. Generalization effects were investigated with another vowel duration contrast /tote/–/toːte/ and a sinusoidal tone pair. Training effects were measured with discrimination and production tests as well as pre-attentive brain response recordings (MMN and N1) on three days. The results showed clear improvements in the perception of the trained stimulus pair /tite/–/tiːte/ in both behavioral and pre-attentive measures. However, no generalization effects on untrained stimuli were observed.

Previous studies have also examined the effects of phonetic listen-and-repeat training on L2 sound production. For example, a study by Taimi et al. (2014) found that 7–11-year-old L1 Finnish speaking children learned to produce a difficult L2 vowel after three short sessions of production training without feedback. The trained L2 vowel contrast was the Swedish /y/–/ʉ/ contrast and the stimuli the same as used in this study. The children participated in four short training sessions on two consecutive days. A later study by Immonen et al. (2022) found that the same training paradigm resulted in even faster training effects on L2 vowel production in younger, 6–7-year-old, children who changed their production of the target L2 vowel /ʉ/ after just one session of listen-and-repeat

training, on the first day of the two-day experiment. Even misleading orthography did not interfere with Finnish children's L2 production learning, when the same listen-and-repeat training paradigm was tested with visual cues that guided pronunciation away from target-like L2 production towards existing L1 production patterns (Immonen, Peltola, et al., 2022), while the same training resulted in an opposite effect in adults (K. U. Peltola et al., 2015). Furthermore, another study by Immonen et al. (2021) found that even passive listening training with the same auditory stimuli resulted in L2 production learning after three training sessions in 9–11-year-old L1 Finnish speaking children. Similar training paradigms have also been tested on adult learners with varying results. For example, a study by K. U. Peltola et al. (2017) found that L1 Finnish and L1 American English speakers continued to produce the target L2 vowels according to their existing L1 production patterns even after a one-day listen-and-repeat training paradigm without feedback. Interestingly, their formant standard deviations decreased after training, indicating more systematic production patterns, but no other training effects were found. However, a later study by K.U. Peltola et al. (2020) showed that a two-day listen-and-repeat training without feedback improved L1 Finnish adults' identification and production of the target L2 vowel, while active listening training did not result in changes in L2 perception or production. Overall, results from previous studies indicate that listen-and-repeat training can be an effective method for L2 perception and production learning, especially in younger learners (Immonen, Alku, et al., 2022; Immonen, Peltola, et al., 2022; Taimi et al., 2014), but it can also benefit adult L2 learners with sufficient amount of training (K. U. Peltola et al., 2020).

Aims and scope of the current study

Rather than studying learning or production of a specific language, the approach typically used in the study area, the goal of the current study was to investigate the effects of short-term listen-and-repeat training on the production of non-native sounds that are L1 irrelevant. Namibia is a linguistically diverse nation with numerous indigenous and Germanic languages spoken across the country. There

are 28 indigenous Namibian languages, most of which are Bantu and Khoisan languages (Norro, 2022). The classification of these languages is not always straightforward, as they are often referred to using different lower and upper level names in different contexts (see e.g., Lusakalalu, 2007; Norro, 2022). According to the 2011 census, the most common indigenous languages in Namibia are Oshiwambo, Nama/Damara, Kavango and Otjiherero (Norro, 2022). The speakers who participated in the present experiment spoke ten different L1s and therefore represent the diversity and variety of Namibian languages in general.

The models of L2 sound learning (Best, 1994, 1995; Best & Tyler, 2007; Flege, 1995; Flege & Bohn, 2021) were used in designing the training protocol to ensure that the trained sounds /y/ and /ʉ/ were L1 irrelevant for the speakers of Bantu and Khoisan languages. The phonological inventories of the Bantu and Khoisan languages spoken in Namibia typically have five or seven vowel systems which include only the rounded back vowel /u/ from the close rounded vowel continuum /y/–/ʉ/–/u/ (Kilarski & Dziubalska-Kołaczyk, 2012; Maddieson & Sands, 2019; Odden, 2015). Therefore, for the purposes of this study, a stimulus pair containing the target vowels /y/ and /ʉ/ from the close rounded /y/–/ʉ/–/u/ vowel space was used in the experiment. The selected vowel pair /y/–/ʉ/ represents a theoretically difficult non-native sound contrast in general, as one or both sounds can be predicted to assimilate to some degree to existing L1 categories according to the PAM-L2 (Best & Tyler, 2007) and /ʉ/ can be perceptually relatively close to the L1 back vowel /u/ according to the SLM (Flege, 1995) and SLM-r (Flege & Bohn, 2021). In addition, the spectral difference between the vowels /y/ and /ʉ/ in itself can be predicted to be challenging, as the vowels /y/–/ʉ/–/u/ exist in the same phonetic space and the location of their category boundaries depends on the L1 sound system (M. S. Peltola et al., 2012).

The study used a simple listen-and-repeat training protocol to measure the effects of phonetic training on non-native vowel production from adult speakers in Namibia. The listen-and-repeat training protocol was chosen for the purposes of this study because it has traditionally been used to

teach pronunciation in L2 classrooms. The productions were acoustically analyzed to determine the quality of the produced vowels. The first hypothesis was that the listen-and-repeat training would result in changes in the speakers' non-native vowel production. More specifically, we expected the training effects to be reflected on the second formant (F2) values of the target vowels. The hypothesis was based on previous results showing short listen-and-repeat training to be an effective method of L2 production and perception learning when training effects are measured in laboratory conditions (Immonen, Alku, et al., 2022; Immonen et al., 2021; Immonen, Peltola, et al., 2022; K. U. Peltola et al., 2017, 2020; Saloranta et al., 2020; Taimi et al., 2014; Tamminen et al., 2015, 2021).

The utterances recorded from the speakers during the training protocol were then used in an identification (ID) task to see how the speakers' productions developed during training, i.e. whether the speakers' productions were perceived as the target words by experienced listeners. Similar identification tasks and listener ratings by experienced and/or native listeners have been used in previous studies to measure L2 production learning in phonetic training settings (Akahane-Yamada et al., 1998; Bradlow et al., 1997). The listeners were given the option to label the produced utterances as uncategorized if they thought the word could not be identified. The hypothesis was that training effects on the speakers' vowel production could be reflected on the identification scores in two ways: First, if the speakers started to produce the trained non-native vowels more recognizably in their speech, but were not yet consistent in their productions, the number of "uncategorized" responses in the ID task would decrease. Second, if the speakers learned to produce the two target vowels consistently and accurately in their speech, the number of correct categorizations in the ID task would increase. The ID results were then viewed together with the results of the acoustic analysis to see how the perceptual categorization of the words developed across sessions compared to the actual formant values. Both the listen-and-repeat training setting and the ID task are described in more detail below.

Materials and methods

Speech data

Speakers

17 speakers participated in the study (aged 20–27 years, mean age 22.6, 11 females). All participants gave written informed consent and answered a language background questionnaire prior to testing. The speakers' reported L1s were Khoe-khoegowab (5), Otjiherero (3), Oshiwambo (2), Oshikwanyama (1), Subia (1), Setswana (1), Mbalangwe (1), Rukwangali (1) and English (2). All the listed L1s except English are Bantu and Khoisan languages. All speakers also spoke English daily and most of them spoke at least one other local language from those listed above in addition to their L1. Other reported languages included Afrikaans (12), German (2), Portuguese (1), Spanish (1), and Silozi (1) but none of the speakers reported speaking these languages as an L1.

Stimuli

Two semisynthetic, bisyllabic pseudowords /ty:ti/ and /tʉ:ti/ were used as stimuli in the listen-and-repeat training protocol to test the effects of the training on speakers' vowel quality production. The stimulus words represent the /y/–/ʉ/ vowel contrast. The same stimuli and training protocol have been used in previous studies with different language and age groups (Immonen, Alku, et al., 2022; Jähi et al., 2015; K. U. Peltola et al., 2017, 2020; Taimi et al., 2014).

The stimuli were created with the Semisynthetic Speech Generation method (Alku et al., 1999). The vowel contrast /y/–/ʉ/ present in the stimuli is phonological in Swedish. Therefore, the stimuli were based on the natural productions of a 24-year-old Finnish-Swedish male speaker. The glottal pulse waveform (i.e. the air flow excitation signal of voiced speech generated by the vocal folds) was first extracted from a voiced speech signal. The first, second and third formant (F1, F2 and F3) structures of the vowels were then synthesized over the natural glottal pulse waveform to create

the pseudo word pair /tyːti/–/tʉːti/. The front vowel /y/ embedded in the stimulus word /tyːti/ had the F1 value of 269 Hz, the F2 value of 1866 Hz and the F3 value of 2518 Hz. The F1, F2 and F3 values of the central vowel /ʉ/ in the word /tʉːti/ were 338 Hz, 1258 Hz and 2177 Hz respectively. A more detailed description of the stimuli can be found in Taimi et al. (2014).

Training procedure

The listen-and-repeat training protocol consisted of a short familiarization phase followed by three recordings and two training sessions. The familiarization phase included three repetitions of the stimuli to allow the participants to set the volume to a comfortable level and get accustomed to the task. A baseline recording was followed by the first training session, then a second recording, the second training and the final recording. In both the training and recording sessions, the stimuli were presented in /tʉːti/–/tyːti/ order with an interstimulus interval (ISI) of three seconds. Each stimulus was presented 10 times in the recordings and 30 times in the training sessions. The input received during recordings formed a part of the overall training, as the listen-and-repeat task was the same in both the recording and training sessions. The speakers did not receive any feedback during training. The participants were instructed to carefully listen to and repeat each word as they heard it. Participants were told that the stimuli were nonsense words. The total duration of the protocol was approximately 20 minutes.

The stimuli were presented diotically with the Beyerdynamic MMX300 headphones from a Dell Latitude 5320 laptop running the Sanako Study Recorder software. Speech data was recorded using the onboard microphone of the laptop. The recordings were performed in a quiet office space and the testing was conducted in English.

Acoustic analysis

The primary acoustic difference between the close rounded vowels /y/ and /ʉ/ lies in the F2 value, which is higher in the front vowel /y/ than in the central vowel /ʉ/. Both /y/ and /ʉ/ have higher F2

values than the back vowel /u/. Therefore, during acoustic analysis, special attention was paid to the first vowel F2 values of each utterance and their development across sessions.

The speakers' productions were subjected to acoustic analysis using Praat (version 6.2.20; Boersma & Weenink, 2022). The F1 and F2 values were extracted using the Linear Predictive Coding (LPC) Burg algorithm. The maximum frequency was set at 5500 Hz for most speakers and lowered to 5000 Hz for speakers with low fundamental frequency (F0 < 100 Hz). The F1 and F2 values were extracted from the center of the first syllable vowels of each utterance (20 utterances per recording). Individual average F1 and F2 values for /y/ and /ʉ/ in each session were then calculated. The development of the average formant values was observed across three time points: pre-training (first recording session), mid-training (second recording session) and post-training (third recording session). The average formant values were compared between sessions by subjecting them to paired samples t-tests using the SPSS Statistics (version 27.0.1.0) software to see whether the listen-and-repeat training resulted in significant changes in target vowel qualities.

Identification of the utterances

Listeners

40 listeners (aged 18–39 years, mean age 23.4, 38 females) volunteered to complete an identification (ID) task where they heard utterances produced by the speakers. All the listeners were students at a Swedish speaking university and used Finland Swedish daily. This ensured the listeners' familiarity with the vowels /y/ and /ʉ/, as the sound categories are phonemic in Swedish.

Identification task and stimuli

The first and last /tyːti/ and /tʉːti/ produced within each recording session were used in the ID task. In other words, the first two utterances and the last two utterances of each session were extracted from the original recordings from all 17 speakers (12 utterances per speaker). A total of 204 individual

utterances were included in the ID task and each of them was repeated 3 times. Due to the high number of utterances, the ID task was divided between listeners so that one listener identified utterances from 4 to 5 speakers. All utterances were categorized 3 times by 10–16 listeners. This means that the number of identifications per utterance ranged between 30 and 48. All utterances were normalized to 65 dB SPL in Praat (version 6.2.20) (Boersma & Weenink, 2022).

The ID protocol was an online forced choice task, designed using jsPsych (de Leeuw, 2015). The listeners heard the utterances in a randomized order, one speaker at a time, and were asked to identify them by clicking a button labelled "/ty:ti/", "/tʉ:ti/" or "neither" (uncategorized). The listeners were instructed to pay special attention to the first syllable vowels. They were told that the vowel in the word /ty:ti/ is the same as in the Swedish word *byta* and the vowel in /tʉ:ti/ is the same as in the Swedish word *hus*. The listeners were instructed to use the option "neither" only in cases where the utterance could not in any way be identified as /ty:ti/ or /tʉ:ti/. The listeners were not given any information about the speakers and they were instructed to use headphones to complete the task. The total duration of the ID task was about 15 minutes including self-paced breaks between speakers.

Analysis

The ID score for each utterance was calculated as % of responses, as the number of categorizations per utterance varied between 30 and 48 depending on the number of listeners. The maximum ID score for each utterance was therefore 100%, with the three ID categories /ty:ti/, /tʉ:ti/ and uncategorized. The average ID scores were statistically analyzed by subjecting them to paired samples t-tests using the SPSS Statistics (version 27.0.1.0) software to discover possible changes in identification compared to the baseline.

Results

Acoustic analysis

The average formant values extracted from the first syllable vowels /y/ and /ʉ/ during acoustic analysis were first observed visually to see whether there were any visible changes in the F1 and F2 values across sessions. The formant values were then subjected to appropriate statistical tests to investigate whether the observed changes were significant. Since the primary acoustic difference between the vowels /y/ and /ʉ/ lies in the F2 value, the analyses focused on the development of the target vowel F2 values across sessions.

The speakers' average F1 and F2 values (shown in Table 1 and Figure 1) indicate some changes in the F2 values of both vowels across recording sessions, while the F1 values seem to remain unchanged. In the first recording session (pre-training), the average F2 values in both vowels (/ʉ/ F2=1269 Hz, /y/ F2=1366 Hz) were produced very close to the stimulus F2 values in /tʉːti/ (F2=1258 Hz). In the second recording session, the speakers' average F2 values (/ʉ/ F2=1427 Hz, /y/ F2=1456 Hz) rose towards the stimulus F2 values in /tyːti/ (F2=1866 Hz), but still remained considerably lower than in the stimulus. Between the second and third recording session, the speakers' average F2 values seem to lower slightly (/ʉ/ F2=1362 Hz, /y/ F2=1408 Hz), but still remain higher than in the first recording session. In order to gain a more in-depth understanding of the formant data, the group's average standard deviations (SD) of the formant values were also calculated (Table 1). The F1 values were produced rather consistently throughout the experiment, whereas there appeared to be rather large variation in the F2 values in all sessions.

Table 1. The average formant values (Hz) and their average standard deviations (reported in brackets) extracted from the first syllable vowels produced by the 17 speakers. The average F1 and F2 values were calculated from the 10 repetitions per session of the words /ty:ti/ and /tʉ:ti/.

		Stimulus	1st Recording	2nd Recording	3rd Recording
F2	/ʉ/	1258	1269 (139)	1427 (111)	1362 (130)
	/y/	1866	1366 (117)	1456 (126)	1408 (125)
F1	/ʉ/	338	370 (23)	363 (24)	380 (29)
	/y/	269	366 (27)	362 (25)	375 (28)

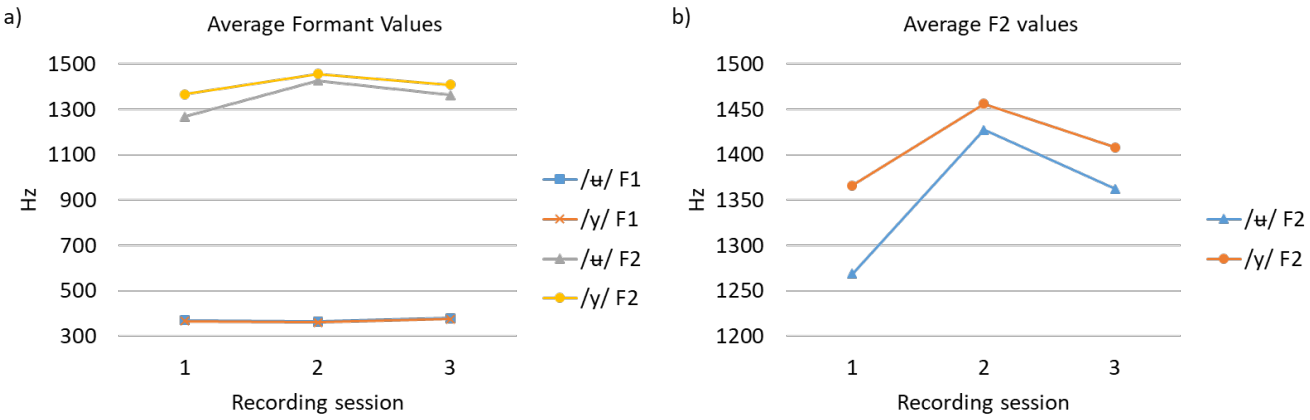

Figure 1. The average formant values extracted from the first syllable vowels /y/ and /ʉ/, as shown in Table 1. The F2 values shown in (b) are the same as in (a).

The average F2 values were statistically analyzed. First, the distribution of the F2 values was inspected with Shapiro-Wilks test of normality, which showed that the data were normally distributed. Next, development of the vowels' F2 values was examined by comparing the F2 values of both vowels between sessions with paired samples t-tests. The analysis of the /ʉ/ F2 values revealed a significant change between the first and second recording session ($t(16) = -3.116$, $p = 0.007$; Cohen's d = 0.756), as well as between the second and third session ($t(16) = 2.248$, $p = 0.039$; Cohen's d = 0.545). The analysis revealed no significant difference in the /ʉ/ F2 values between the first and third

recording sessions. This can be explained by looking at the overall development of the formant values shown in Table 1 and Figure 1. The /ʉ/ F2 rises from 1269 Hz to 1427 Hz between the first and second recording session, and then drops back to 1362 Hz in the third session. Therefore, the acoustic difference between the /ʉ/ F2 values in the first and third recording remains relatively small, even with the significant change between the second and third sessions. No significant changes were observed in the /y/ F2 values across sessions.

The F2 values shown in Figure 1 suggest that the acoustic difference between the target vowels decreased after the first recording session. In other words, it appears that the speakers produced a smaller acoustic contrast between /y/ and /ʉ/ in the second and third recording session compared to the first session. The /y/ and /ʉ/ F2 values were compared within each recording session with paired samples t-tests to determine whether the acoustic difference between the two vowels was statistically significant. The analysis revealed a significant difference between the /y/ and /ʉ/ F2 values in the first ($t(16) = -2.493$, $p = 0.024$; Cohen's d = 0.605) and third ($t(16) = -2.174$, $p = 0.045$; Cohen's d = 0.527) recording sessions but not in the second session ($t(16) = -1.159$, $p = 0.263$).

There was great inter-speaker variation in the participants' F2 values and their development across sessions, as can be seen in Figures 2–4. Firstly, it seems that the speakers were divided in two groups: those who mostly produced relatively low F2 values (< 1400 Hz) for both /y/ and /ʉ/ (e.g., speakers 4, 5, 8, 10, 12, 13, 15, 16), and those who produced higher F2 values (> 1400 Hz) for both vowels. In addition, the individual F2 averages revealed at least three different patterns in F2 development. Some speakers (e.g., speakers 9, 11, 13, 14 and 17) produced clearly higher F2 values for both vowels in the second session, and then reverted close to the lower baseline values in session 3. Others (e.g., speakers 3, 6, 7, 15, 16) showed a rising trend, as their /y/ and /ʉ/ F2 values were higher in sessions 2 and 3 in comparison with the first baseline session, while a couple of speakers (4 and 12) did the exact opposite and lowered their F2 values after the first session.

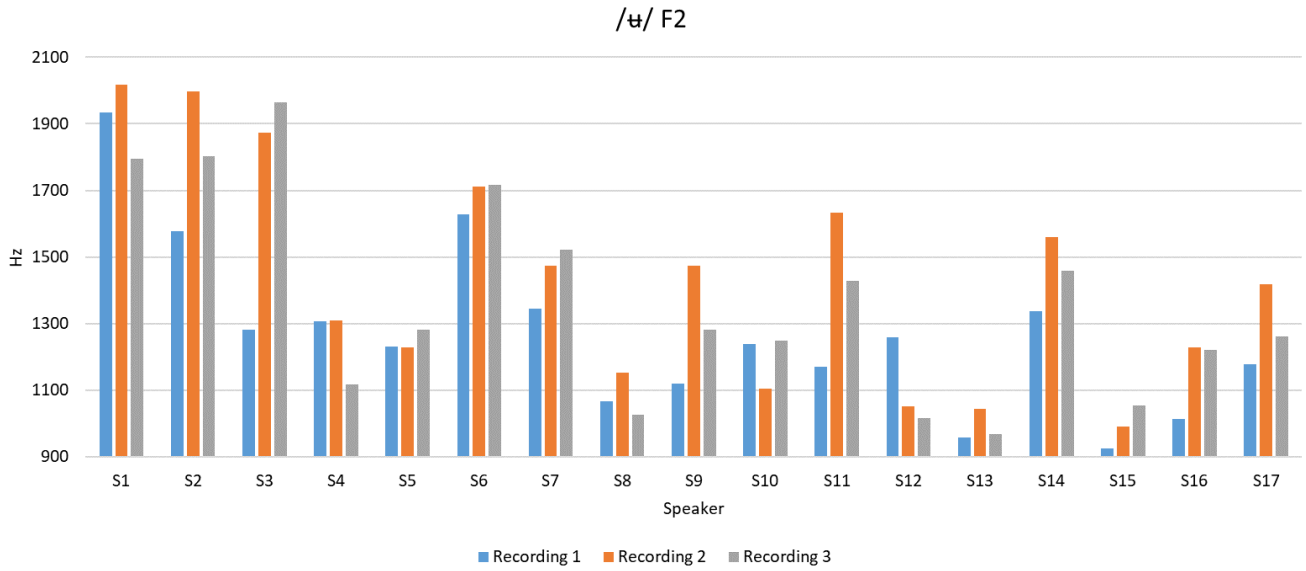

Figure 2. Each speakers' (S1–S17) average /ʉ/ F2 values across the three recording sessions.

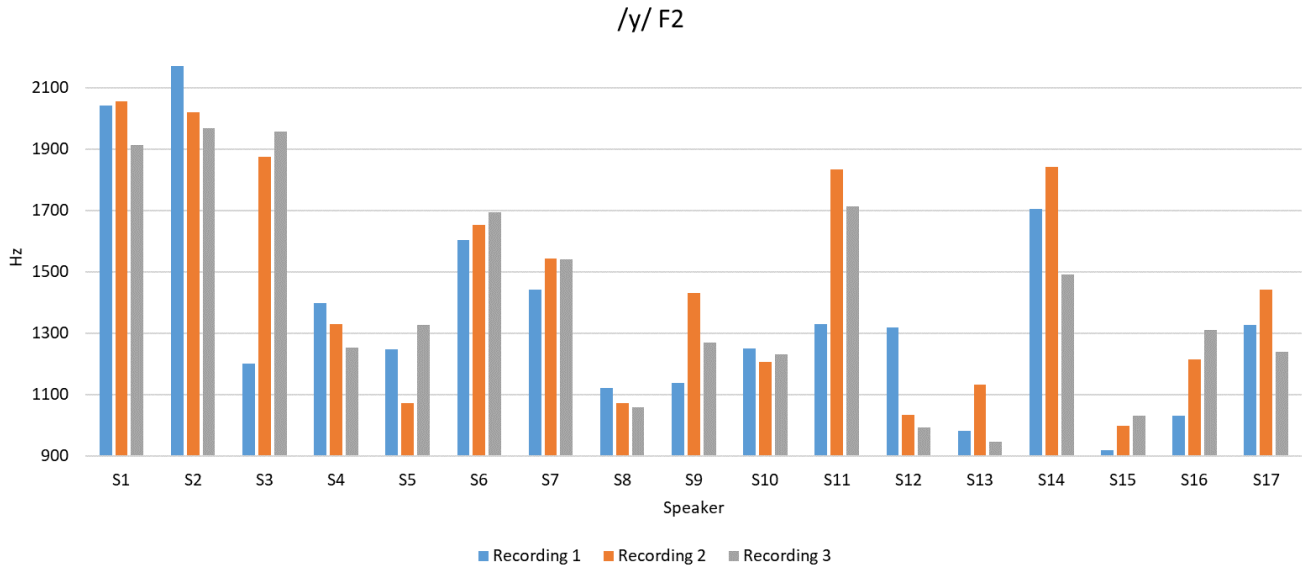

Figure 3. Each speakers' (S1–S17) average /y/ F2 values across the three recording sessions.

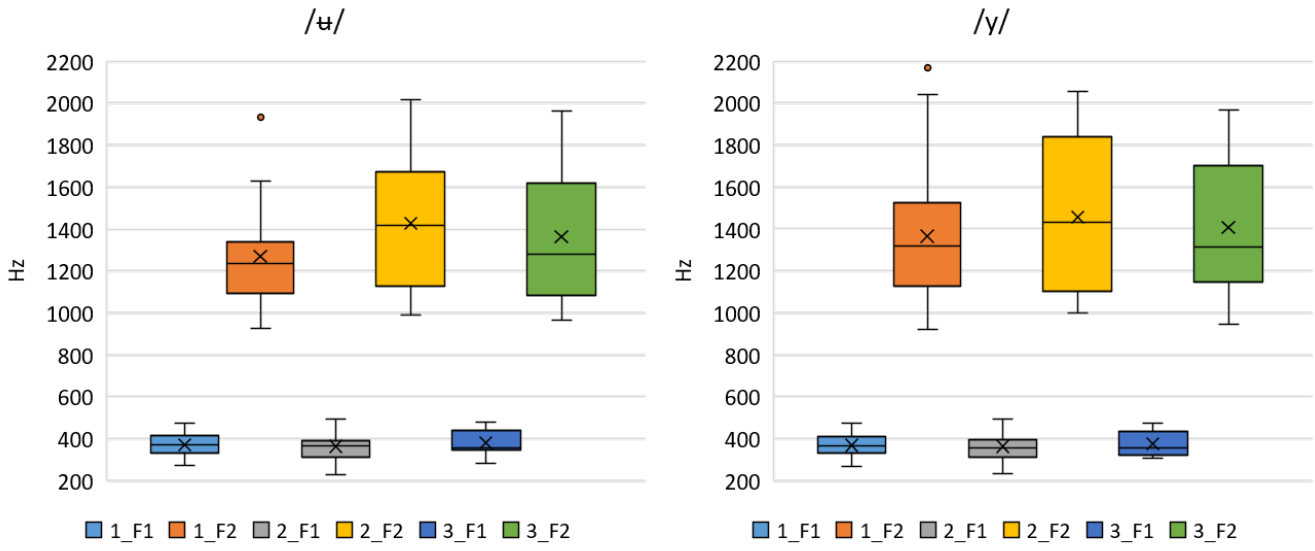

Figure 4. The distribution of the F1 and F2 values for /ʉ/ and /y/ across the three recording sessions.

Identification of the utterances

The average identification scores (% of responses) for /tyːti/ and /tʉːti/ are displayed in Figures 5 and 6. The average ID scores showed that, overall, the speakers' productions were identified more as /tʉːti/ than as /tyːti/, regardless of the target vowel. In other words, the speakers' productions were mostly perceived as /tʉːti/ across sessions, even if the target word was /tyːti/. However, there seemed to be some changes in identification across sessions, especially in the /tʉːti/ utterances. For example, the percentage of correct /tʉːti/ identifications decreased from 71% to 55% between the beginning of the first and second sessions (Figure 5), which seems to suggest that the speakers' production became less target-like in the second session.

The average ID scores were subjected to paired samples t-tests to discover possible changes in identification compared to the baseline. The ID responses for the first utterance of the first recording session (1_first, baseline) were compared to the first and last utterance of each three recording sessions. The statistical analysis revealed some significant changes in the identification responses for the /tʉːti/ utterances in the second and third sessions (Figure 5). The first /tʉːti/ utterance of the second session was identified significantly less as /tʉːti/ ($t(16) = 2.603$, $p = 0.019$; Cohen's d =

0.631), and significantly more as /ty:ti/ ($t(16) = -3.479$, $p = 0.003$; Cohen's d = 0.844) compared to the baseline. The last /tʉ:ti/ utterance of the second session received significantly less "neither" responses ($t(16) = 2.234$, $p = 0.040$; Cohen's d = 0.542), meaning that the utterances were overall easier to categorize for the listeners. The last /tʉ:ti/ utterance of the second session was also identified significantly more as /ty:ti/ ($t(16) = -3.799$, $p = 0.002$; Cohen's d = 0.921) compared to the baseline. In addition, the last /tʉ:ti/ utterance of the third session was identified significantly more as /ty:ti/ ($t(16) = -2.463$, $p = 0.026$; Cohen's d = 0.597) compared to the baseline. The analysis revealed no significant changes in the identification of /ty:ti/ utterances across sessions (Figure 6).

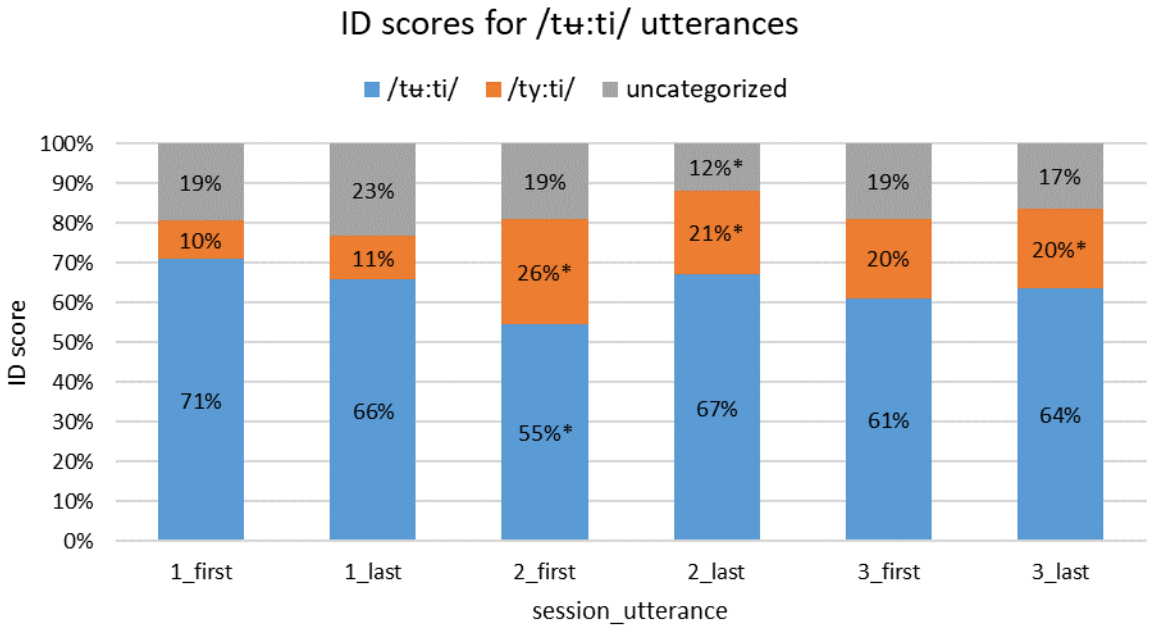

Figure 5. The average identification scores (%) for /tʉ:ti/ utterances (all speakers). Significant changes in the average identification % compared to the baseline utterance (1_first) are marked with an asterisk (paired samples t-tests).

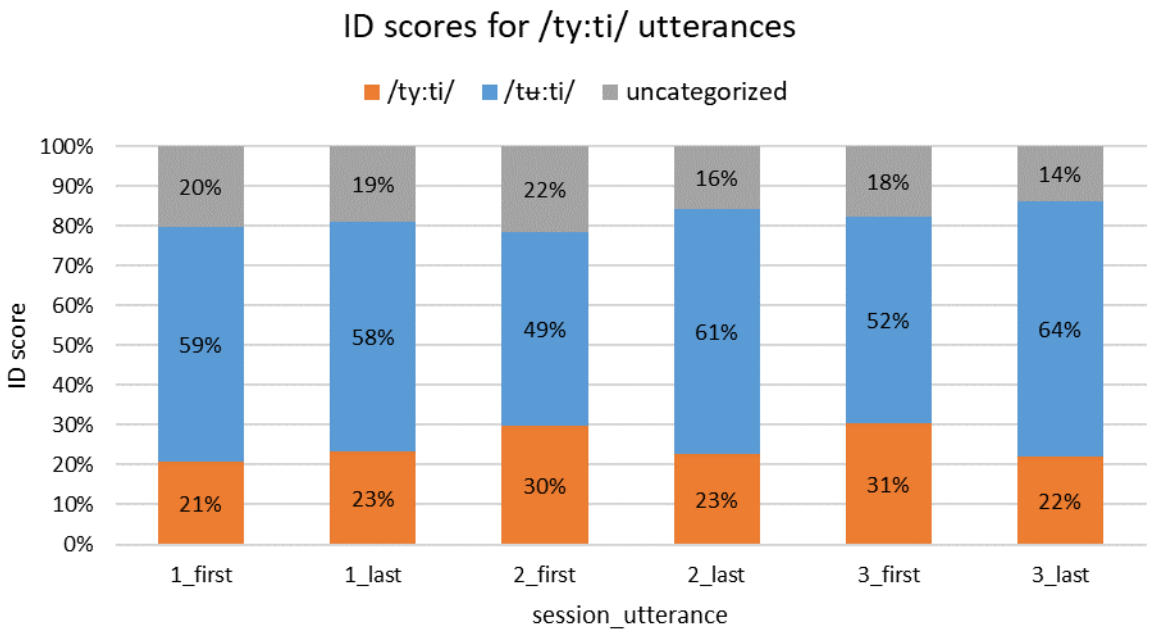

Figure 6. The average identification scores (%) for /ty:ti/ utterances (all speakers). No significant changes in the average identification % compared to the baseline utterance (1_first) were revealed in the paired samples t-tests.

There was also great inter-speaker variation in the identification scores received by the speakers (Figure 7). The individual ID scores revealed four main trends in the perceived quality of the speakers' productions: 1. all productions predominantly identified as /ty:ti/ (e.g., speaker 1), 2. all productions predominantly identified as /tʉ:ti/ (e.g., speaker 10), 3. all productions predominantly identified as neither (e.g., speaker 9), and 4. no clear pattern in identifications (e.g., speaker 6).

Figure 7. Individual ID scores received by four speakers (S1, S6, S9 and S10) that demonstrate the four main patterns observed in the listeners' perception of the speech data.

Discussion

This study investigated how phonetic listen-and-repeat training affects non-native vowel quality production by speakers of Namibian languages. Training effects were measured both acoustically and with an ID task completed by experienced listeners. The study aimed to answer the following two research questions: First, can adult native speakers of different Namibian languages learn to produce the theoretically difficult non-native sounds /y/ and /ʉ/ through short-term listen-and-repeat training? Second, does listen-and-repeat training affect the identification of the speakers' productions by listeners who are familiar with the trained vowel categories? The first hypothesis was that the training would result in changes in the speakers' production of /y/ and /ʉ/ F2 values. This hypothesis was only partly confirmed by the results of the acoustic analysis. The second hypothesis was that training related changes in the speakers' vowel production could be reflected on the ID scores in two ways:

First, if the speakers' productions became perceptually more salient as a function of training, the number of correct identifications would increase, and second, the number of "uncategorized" responses would decrease. This hypothesis was not confirmed by the results of the ID task.

Development across sessions

The results showed some training effects in the second session, which were reflected in the F2 values and the ID scores, but the changes did not persist. The results of the acoustic analysis showed that the speakers changed their production of the target vowel /ʉ/ across sessions and the change was reflected in the F2 value of the vowel, partly confirming the original hypothesis. However, the overall development of the /ʉ/ F2 values and identification scores indicated that the observed changes were away from the acoustic target and not consistent across sessions. The formant analysis showed that the speakers' /ʉ/ F2 values were very close to the target F2 values already in the first session and after the rise in the second session, they reverted close to the target values in the final session. The results of the ID task revealed that the speakers' productions were, on average, identified more as /tʉːti/ than as /tyːti/ and that the number of uncategorized utterances remained rather low throughout the experiment. This finding was in keeping with the results of the acoustic analysis, which showed that most speakers' F2 values remained rather low, closer to the /ʉ/ stimulus F2 values than the /y/ stimulus F2 values. The identification scores for /tʉːti/ utterances showed that the speakers' production of /tʉːti/ changed perceptually in the second session, but the change was in the opposite direction than was expected. The changes in identification seen in the second session indicate that the speakers' productions of /tʉːti/ were actually perceived as less target-like (more /tyːti/ responses). The number of "uncategorized" responses actually did decrease at the end of the second session, as hypothesized, but the effect did not last in the third session and the hypothesis was not confirmed. In other words, the results of the ID task supported the findings of the acoustic analysis, as the speakers' productions of /tʉːti/ were more accurately identified in the first and third session compared to the second session.

Therefore, though the listen-and-repeat training did not result in significant changes between the first and third recording session, the formant and ID results showed that the speakers did actually produce the non-native vowel /ʉ/ close to the acoustic model prior to training.

Differentiation of the trained vowels

Furthermore, the two target vowels' formant values developed similarly across sessions and the difference between the /y/ and /ʉ/ F2 values did not reach significance in the second recording session. Overall, the difference between the target vowels' F2 values remained modest throughout the experiment compared to the stimulus values. This finding was supported by the results of the ID task, which showed that the speakers did not produce a perceivable spectral difference between the two trained vowels, as both words were identified more as /tʉːti/ by the listeners. The changes in F2 values observed in the second recording session as well as the large standard deviations indicated confusion between the two non-native sounds, as the speakers started to produce both vowels with higher F2 values and acoustically closer together. As the speakers produced /ʉ/ close to the trained target already in the first session (pre-training), the first training probably drew their attention to /y/, but they were unable to consistently perceive and/or produce the vowels as two separate sounds causing both vowels' F2 values to rise and /ʉ/ to become less identifiable in the second recording session. Based on the predictions made by the PAM-L2 (Best & Tyler, 2007), this could be an indication of category goodness or single category assimilation, meaning that the speakers assimilated the target vowels to one or more of their L1 vowel categories, such as /i/ or /u/ which are situated close to the trained vowels in the articulatory-acoustic vowel space. However, no definite conclusions on the possible assimilation patterns between the L1 and L2 vowels cannot be drawn without additional data on the speakers' L1 vowel production and L2 vowel perception.

Taken together, these results suggest that the speakers did not learn to systematically differentiate between the two target vowels in their production, which could indicate two things:

either the speakers did not reliably perceive the acoustic difference between the stimuli and therefore did not produce it, or they perceived the difference but were unable to produce it accurately in their own speech. Viewed against the predictions of the SLM-r (Flege & Bohn, 2021), suggesting a bidirectional connection between L2 segmental perception and production, both explanations could be true. The spectral difference between the target non-native vowels was probably so difficult that consistent perception and production patterns did not have time to develop, and therefore the speakers did not learn to produce the contrast in their own speech. However, a direct link between the speakers' perception and production of the trained non-native vowels cannot be drawn from these data, because the speakers' perception of the stimuli was not measured in this experiment.

Interestingly, the results of the acoustic analysis and the ID task did not show any changes in the production of /y/. The fact that the speakers (as a group) did not learn to produce /y/ could well be explained by the universal articulatory links between tongue position and lip rounding. Front vowels are naturally unrounded while back vowels are naturally rounded (Knight, 2012; Ladefoged, 1971; Lindau, 1978), meaning that the primary articulation for front vowels includes no lip rounding. Therefore, the rounded front vowel /y/ can be expected to be more challenging to articulate than the rounded central vowel /ʉ/, which could partly explain the findings of this study. Though there were individual speakers who did produce /y/ close to the acoustic model according to the F2 values and ID scores, they, similar to the speakers who did not produce /y/, did not produce a clear perceivable contrast between /y/ and /ʉ/.

These findings indicate that, in contrast with previous findings from similar training studies (Immonen, Alku, et al., 2022; Immonen et al., 2021; Immonen, Peltola, et al., 2022; K. U. Peltola et al., 2017, 2020; Saloranta et al., 2020; Taimi et al., 2014; Tamminen et al., 2015, 2021), the listen-and-repeat training did not result in training effects and that the speakers did not learn to produce /y/ and /ʉ/ as two spectrally distinct sounds, even though their vowel production changed during the experiment. It could be that the amount of training used in this experiment was not enough for the

speakers to learn to perceive and produce two non-native vowels. Previous studies have used training paradigms with longer training periods and/or more input (e.g., Bradlow et al., 1997, 1999; Ylinen et al., 2010), or only one non-native sound contrasted with a familiar L1 sound (e.g., Immonen, Alku, et al., 2022; K. U. Peltola et al., 2020; Tamminen et al., 2015). The small changes observed in this study suggest that the speakers did respond to the training, even though training effects were not clear or persistent. Therefore, it could be that more training over a longer period of time would have allowed the speakers to distinguish the two trained vowels and to adapt their own articulatory patterns more consistently to fit the acoustic model.

Inter-speaker variation

The large inter-speaker differences found in the formant data and the ID scores also need to be addressed. The inter-speaker differences in both data sets were investigated by comparing the speakers' language backgrounds. However, no satisfactory explanations were found, as speakers with the same reported L1s or the same additional languages showed varying patterns in the formant and ID results, meaning that the acoustic and perceptual properties of their productions could not be predicted by their L1 or any other language they reported in the background questionnaire. Surprisingly, the direction of the change in the individual average F2 values also did not seem to be connected to the actual F2 values themselves. In other words, the expectation was that if the direction of the change in production was towards the stimulus values (/tyːti/ F2 = 1866 Hz, /tʉːti/ F2 = 1258 Hz), the speakers who produced very high F2 values would lower them towards /ʉ/, and those who produced very low F2 values would raise them towards /y/. Instead, most speakers seem to have followed a single production pattern (rising, falling and rising-falling) across sessions, regardless of the stimulus target and their own baseline formant values. These observations seemed to be at least partly in line with the individual ID scores, as speakers with the highest average F2 values received the highest number of /tyːti/ categorizations for their productions. However, the formant data and the

ID data are not directly comparable, because the F2 values are averages of the ten repetitions within each recording session and the ID task included only the first and last repetition of each session.

Future directions

In order to examine the effects of listen-and-repeat training further, more data with larger and more homogenous language groups and a longer training period need to be gathered. The inter-speaker differences observed in this experiment could not be explained with language background factors due to small number of speakers. Therefore, more data from L1 speakers of different Namibian languages is needed in order to make more definite conclusions on the effects of listen-and-repeat training on the production of non-native vowel qualities.

Conclusions

The results showed that the speakers of different Namibian languages changed their production of the non-native vowel /ʉ/ after one training session but the training effects were not retained after the second training session. Most importantly, acoustic analysis and listener identifications showed that the speakers produced /ʉ/ close to the trained target already in the first session, but changed their production away from the target after the first training, while their production of /y/ did not change. The amount of training was probably not enough for the speakers to learn to produce a sufficient acoustic and perceptual difference between the non-native vowels /y/ and /ʉ/. Formant analysis and listener identifications showed that the speakers' vowel productions were overall closer to the non-native vowel /ʉ/ than the non-native vowel /y/ throughout the experiment. This could partly be explained by the fact that front vowels are naturally unrounded, making the articulation of the rounded front vowel /y/ universally difficult.

References

Akahane-Yamada, R., McDermott, E., Adachi, T., Kawahara, H., & Pruitt, J. (1998). Computer-based second language production training by using spectrographic representation and HMM-based speech recognition scores. *5th International Conference on Spoken Language Processing*. Fifth International Conference on Spoken Language Processing, Sydney, Australia. ISCA. https://doi.org/10.21437/ICSLP.1998-722

Alku, P., Tiitinen, H., & Näätänen, R. (1999). A method for generating natural-sounding speech stimuli for cognitive brain research. *Clinical Neurophysiology : Official Journal of the International Federation of Clinical Neurophysiology*, *110*(8), 1329–1333.

Best, C. T. (1994). The emergence of native-language phonological influences in infants: A perceptual assimilation model. *The Development of Speech Perception: The Transition from Speech Sounds to Spoken Words*, *167*, 224.

Best, C. T. (1995). A direct-realist view of cross-language speech perception. In W. Strange (Ed.), *Speech Perception and Linguistic Experience: Issues in Cross-Language Speech Research* (pp. 171–206). York Press, Baltimore.

Boersma, P., & Weenink, D. (2022). *Praat* (6.2.20) [Praat: doing phonetics by computer].

Bradlow, A. R., Akahane-Yamada, R., Pisoni, D. B., & Tohkura, Y. (1999). Training Japanese listeners to identify English /r/and /l/: Long-term retention of learning in perception and production. *Perception & Psychophysics*, *61*(5), 977–985. https://doi.org/10.3758/BF03206911

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

de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior Research Methods*, *47*(1), 1–12. https://doi.org/10.3758/s13428-014-0458-y

Flege, J. E. (1995). Second language speech learning: Theory, findings, and problems. *Speech Perception and Linguistic Experience: Issues in Cross-Language Research*, 233–277.

Flege, J. E., & Bohn, O.-S. (2021). The revised speech learning model (SLM-r). In R. Wayland (Ed.), *Second language speech learning: Theoretical and empirical progress* (pp. 3–83). Cambridge University Press.

Immonen, K., Alku, P., & Peltola, M. S. (2022). Phonetic listen-and-repeat training alters 6–7-year-old children's non-native vowel contrast production after one training session. *Journal of Second Language Pronunciation*, *8*(1). https://doi.org/10.1075/jslp.21005.imm

Immonen, K., Kilpeläinen, J., Alku, P., & Peltola, M. S. (2021). Does Studying in a Music-oriented Education Program Affect Non-native Sound Learning? —Effects of Passive Auditory Training on Children's Vowel Production. *Journal of Language Teaching and Research*, *12*(5), 678–687. https://doi.org/10.17507/jltr.1205.06

Immonen, K., Peltola, K. U., Tamminen, H., Alku, P., & Peltola, M. S. (2023). Orthography does not hinder non-native production learning in children. *Second Language Research*, *39*(2). https://doi.org/10.1177/02676583221076645

Iverson, P., Pinet, M., & Evans, B. G. (2012). Auditory training for experienced and inexperienced second-language learners: Native French speakers learning English vowels. *Applied Psycholinguistics*, *33*(01), 145–160. https://doi.org/10.1017/S0142716411000300

Jähi, K., Alku, P., & Peltola, M. S. (2015). Does interest in language learning affect the non-native phoneme production in elderly learners. *Proc. 18th ICPhS*.

Kilarski, M., & Dziubalska-Kołaczyk, K. (2012). On Extremes in Linguistic Complexity: Phonetic Accounts of Iroquoian, Polynesian and Khoesan. *Historiographia Linguistica: International*

Journal for the History of the Language Sciences/Revue Internationale Pour l'Histoire Des Sciences Du Langage/Internationale Zeitschrift Für Die Geschichte Der Sprachwissenschaften, *39*(2–3), 279–303. https://doi.org/10.1075/hl.39.2-3.05kil

Knight, R.-A. (2012). *Phonetics: A coursebook* (1st edition). Cambridge University Press.

Ladefoged, P. (1971). *Preliminaries to linguistic phonetics*. Univ. of Chicago P.

Lindau, M. (1978). Vowel Features. *Language*, *54*(3), 541–563. https://doi.org/10.2307/412786

Lusakalalu, P. (2007). Media, education and the count of Namibian languages. *Journal of Namibian Studies : History Politics Culture*, *2*, 85–101.

Maddieson, I., & Sands, B. (2019). The sounds of the Bantu languages. In M. Van de Velde, K. Bostoen, D. Nurse, & G. Philippson (Eds.), *The Bantu Languages* (2nd ed., pp. 79–127). Routledge. https://doi.org/10.4324/9781315755946-3

Munro (Eds.), *Language experience in second language speech learning: In honour of James Emil Flege* (pp. 13–34). John Benjamins Publishing Company.

Norro, S. (2022). Factors affecting language policy choices in the multilingual context of Namibia: English as the official language and medium of instruction. *Apples - Journal of Applied Language Studies*, *16*(1), Article 1. https://doi.org/10.47862/apples.107212

Odden, D. (2015). Bantu Phonology. In Oxford Handbooks Editorial Board (Ed.), *Oxford Handbook Topics in Linguistics*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199935345.013.59

Peltola, K. U., Rautaoja, T., Alku, P., & Peltola, M. S. (2017). Adult learners and a one-day production training–Small changes but the native language sound system prevails. *Journal of Language Teaching and Research*, *8*(1), 1–7.

Peltola, K. U., Tamminen, H., Alku, P., Kujala, T., & Peltola, M. S. (2020). Motoric Training Alters Speech Sound Perception and Production—Active Listening Training Does Not Lead into Learning Outcomes. *Journal of Language Teaching and Research*, *11*(1), 10–16.

Peltola, K. U., Tamminen, H., Alku, P., & Peltola, M. S. (2015). Non-native production training with an acoustic model and orthographic or transcription cues. *Proc. 18th ICPhS*.

Peltola, M. S., Tamminen, H., Toivonen, H., Kujala, T., & Näätänen, R. (2012). Different kinds of bilinguals – Different kinds of brains: The neural organisation of two languages in one brain. *Brain and Language*, *121*(3), 261–266. https://doi.org///dx.doi.org/10.1016/j.bandl.2012.03.007

Saloranta, A., Alku, P., & Peltola, M. S. (2020). Listen-and-repeat training improves perception of second language vowel duration: Evidence from mismatch negativity (MMN) and N1 responses and behavioral discrimination. *International Journal of Psychophysiology*, *147*, 72–82. https://doi.org/10.1016/j.ijpsycho.2019.11.005

Taimi, L., Jähi, K., Alku, P., & Peltola, M. S. (2014). Children Learning a Non-native Vowel–The Effect of a Two-day Production Training. *Journal of Language Teaching and Research*, *5*(6), 1229–1235.

Tamminen, H., Kujala, T., Näätänen, R., & Peltola, M. S. (2021). Aging and non-native speech perception: A phonetic training study. *Neuroscience Letters*, *740*, 135430. https://doi.org/10.1016/J.NEULET.2020.135430

Tamminen, H., Peltola, M. S., Kujala, T., & Näätänen, R. (2015). Phonetic training and non-native speech perception—New memory traces evolve in just three days as indexed by the mismatch negativity (MMN) and behavioural measures. *International Journal of Psychophysiology*, *97*(1), 23–29. https://doi.org/10.1016/j.ijpsycho.2015.04.020

Ylinen, S., Uther, M., Latvala, A., Vepsäläinen, S., Iverson, P., Akahane-Yamada, R., & Näätänen, R. (2010). Training the brain to weight speech cues differently: A study of Finnish second-language users of English. *Journal of Cognitive Neuroscience*, *22*(6), 1319–1332.