PERCEPTION OF GERMAN TENSE AND LAX VOWEL CONTRAST BY CHINESE LEARNERS

Yingming Gao[1], *Hongwei Ding*[2]*, *Peter Birkholz*[1], *Rainer Jäckel*[1], *Yi Lin*[2]
[1]*Institute for Acoustics and Speech Communication, TU Dresden, Germany*
[2]*School of Foreign Languages, Shanghai Jiao Tong University, China*
yingming.gao@mailbox.tu-dresden.de,{hwding, carol.y.lin}@sjtu.edu.cn
{peter.birkholz, rainer.jaeckel}@tu-dresden.de

Abstract: First language phonological categories strongly influence the late learners' perception of second language (L2) categories. In the present study, we examined whether duration or vowel quality is more important for Mandarin Chinese learners of German to distinguish German tense and lax vowels compared to German native speakers. In order to have good control of the two factors, 70 German words were synthesized with an articulatory speech synthesizer (VocalTractLab). For each of the 70 words, three "incorrectly pronounced" variants were generated by manipulating the target vowel in terms of *duration* (long vs. short), *quality* (tensing vs. laxing), or *both*. Then 10 native German speakers, 16 Chinese advanced learners, and 19 Chinese beginners were asked to listen to the synthesized and manipulated words and identify the correctly pronounced words. It was found that German listeners could distinguish tense-lax vowels with almost 100% accuracy, and the vowel quality seemed to be the primary cue. Chinese listeners achieved an accuracy of 82% for advanced learners and 76% for beginners. Moreover, Chinese learners relied more on duration than quality in their identification. This suggests that durational cues in vowel perception are easier to learn even if the duration is not a distinctive feature for vowels in their native language, while quality dimensions are more difficult for L2 learners to acquire. The findings can shed some light on L2 speech acquisition.

1 Introduction

Generally speaking, German tense-lax vowels differ both in quality (spectral cues) and in quantity (duration) [1], while Mandarin Chinese has no tense-lax vowel pairs and duration is not a distinctive feature for vowels [2]. On the other hand, phonological characteristics of the native language are thought to interfere with the L2 speech acquisition. There have been many production and perception studies examining the effects of Mandarin Chinese L2 English acquisition. Among the few studies on the production of Mandarin learners on the L2 German acquisition, it was demonstrated that Chinese speakers have difficulties in distinguishing German tense-lax vowels in their production [3]. However, it is not clear whether the production difficulty is due to perception problems. It remains to be explored: (1) Whether the Chinese learners can distinguish the minimal pairs of tense-lax vowels in German; (2) Whether they rely on duration or vowel quality to distinguish tense-lax vowels in L2 German compared to German native speakers, and between different levels of learners.

In order to generate the speech stimuli for the experiment, the articulatory speech synthesizer VocalTractLab 2.2 [4, 5] was employed, which provides a suitable tool for synthesizing German vowels and manipulating their durations while controlling other factors.

*Corresponding author.

2 Method

First, 70 German carrier words were selected, then stimuli were synthesized and manipulated based on these words, and finally an identification test was conducted.

2.1 Word Selection

The German monophthongs comprise seven tense/lax vowel pairs that can occur in rhythmically strong syllables and include four front-vowel pairs ([iː]-[ɪ], [yː]-[ʏ], [eː]-[ɛ], [øː]-[œ]), one open-central-vowel pair ([aː]-[a]), and two back-vowel pairs ([oː]-[ɔ], [uː]-[ʊ]). To avoid difficulties of the foreign learners' perception due to unfamiliarity with unknown utterances, we chose words that are simple and frequent. One- or two-syllable words with a 'CV(C)' structure were preferred with the target vowel embedded in non-syllable final positions. For each vowel, five simple carrier words which are frequently used in real life communication situations were chosen from the word list of "the Goethe-Zertifikat A1: Start Deutsch 1" [6]. Hence, 70 words (14 target vowels x 5 words per vowel) were subsequently synthesized with VocalTractLab.

2.2 Stimuli Creation

VocalTractLab is an articulatory speech synthesizer, which simulates the articulation process, specified by a gestural score, and simultaneously produces acoustic signals. A gestural score is organised in eight tiers corresponding to supraglottal places of articulation, glottal settings, and lung pressure. The realization of each phoneme is considered to comprise multiple gestures, which are coordinated and distributed over gestural tiers. Each gesture consists of three parameters: a gestural value, a duration, and a time constant, which define target positions of articulators, their lasting time, and how quickly the participating articulators reach the targets from their previous states, respectively. Hence, the synthesizer allows individual control of vowel quality and duration in the present work.

We adopted the time structure model of the syllable to organise all gestures involved in a word [7]. The temporal alignment of all the phones within a syllable follows some simple principles: the initial consonant and the vowel share the same onset of the syllable, and the other phones are sequentially aligned after the vowel of the syllable. Accordingly, a gestural score for a word is organized as follows (here, the German word "Buch" as illustration in Figure 1): the initial consonantal gesture and the vocalic gesture start at the syllable onset; the gestures of others phones are sequentially arranged from the offset of the vowel gesture. The glottal gestures, controlling the phonation type for each phone, are aligned with the respective supraglottal gestures. The consonantal glottal gesture dominates the vocalic glottal gesture in the overlapped time slot. During the initialization stage, the phone durations were set to German phoneme inherent durations measured by Kohler [8]. The time constants were set to 10 milliseconds for lip gestures, 15 milliseconds for tongue-tip gestures (with an exception of 5 milliseconds for the lateral gesture), and 20 milliseconds for tongue-body gestures [5]. The time constant for a vowel follows that of its preceeding consonatal gesture. Because VocalTractLab employs a target approximation model, all gestural targets are gradually approached, thus making each real articulatory process occur always a little later than its corresponding gesture (see the trajectory of lung pressure gestures in Figure 1). Therefore, the lung pressure tier starts with an empty lung pressure gesture of an empirical duration (100%, one-half or one-third of the initial consonant for stops, fricatives or sonorants, respectively). An extra lung pressure gesture of 100 milliseconds is appended after the last phone to provide pressure during its realization. After the initialization of the gestural scores, manual adjustments of duration and time constants were required to make the synthetic speech sound natural and make all segments have exactly their

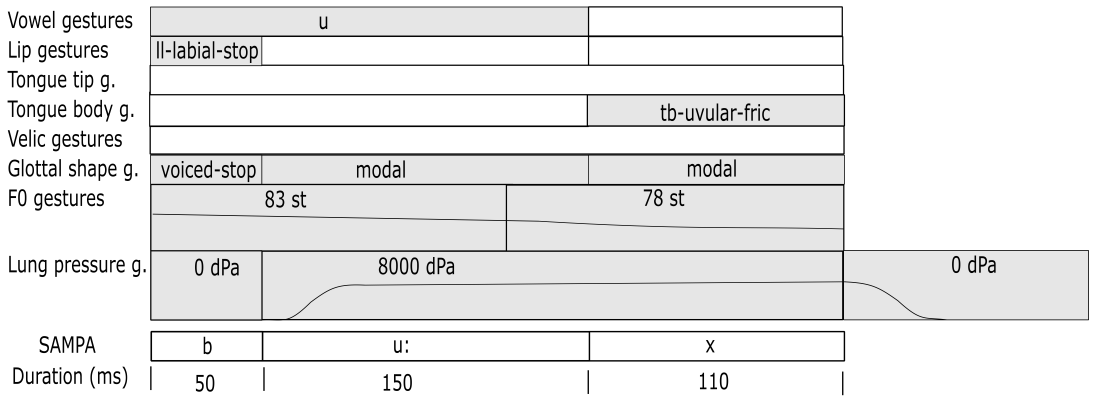


Figure 1 – The time-structure-model based initialization of the gestural score of German word "Buch" ("book" in English)

inherent duration as given in [8]. The duration of the last phone in each word was extended by 30% to simulate the final lengthening. In the f_0 tier, we defined two f_0 gestures with equal duration but a declination pitch contour.

For each word, the gestural score was subsequently manipulated with regard to vowel quality and/or duration of the target vowel, to yield three variants with "incorrect" pronunciations, which resembled the manipulation and perception test by Sendlmeier [9]. For the first manipulated variant, the gesture of target vowel in the carrier word was substituted with its tense/lax counterpart (tense vowels were replaced with their lax counterparts, and vice versa) but the duration was kept the same; for the second manipulated variant, the gestural duration of the target vowel in the original stimulus was manipulated to have the inherent duration as its tense/lax counterpart, while the vowel quality was kept the same; for the third manipulated variant, the duration of the replaced vowel gesture in the first manipulated variant was further manipulated to match its inherent duration (duration and quality were both altered). Hence, we created a basic correct version and three manipulated variants with incorrect pronunciations for each word, thus obtaining 280 stimuli in total (70 words x 4 versions per word).

2.3 Subjects

The Mandarin Chinese learners of German were recruited for two proficiency levels. The advanced learners were German major students who had learned German for more than three years with about eight hours every day, while the beginners were students who had learned German as a second foreign language with two hours per week resulting in a total of German classes between 50-200 hours. All the Chinese subjects were students from Shanghai Jiao Tong University, and the German subjects were from TU Dresden. More information is listed in Table 1.

Table 1 – Information of Subjects

Subject Group	**Number**
German native listeners	10 (males)
Chinese advanced learners	16 (13 females and 3 males)
Chinese beginners	19 (14 females and 5 males)

2.4 Identification Test

In the perception experiment, a group of listeners were asked to identify the "correctly pronounced" stimuli from pairs of stimuli, with one cononical and one manipulated form of a word in each pair. The experiment consisted of three sessions. In the first session, the minimal pairs (the canonical realization and its first manipulated variant) had different vowel qualities but the same inherent duration of the canonical vowel. In the second session, the minimal pairs (the canonical realization and its second manipulated variant) had different duration values but the same vowel quality. In the third session, the minimal pairs (the canonical realization and its third manipulated variant) had different vowel qualities and different duration values (with the tense vowel longer than its lax counterpart). Hence, different acoustic cues were contrasted in different sessions. Vowel pairs in session 1 differed only in *quality* (spectral characteristics), in session 2 only in *duration* (vowel length), and in session 3 in both *quality* and *duration*.

Three groups of subjects conducted the perception experiment individually in quiet rooms. The order of the 70 words was randomized in each session. Each time one word was prompted on a computer screen and the two stimuli of a pair were randomly played to the listeners, and they were asked to judge which stimulus matched the word presented on the screen. They could replay the stimuli as many times as they wanted. Before the identification test started, a training session with six examples had to be carried out to help listeners get familiar with the procedure.

3 Results

Results are first presented in overall identification rates for the three listener groups in three conditions, followed by comparisons between tense-lax vowels and different vowel pairs.

3.1 Overall Comparison

Figure 2 shows the average identification accuracy of different groups in three conditions. If the vowel pairs differed only in *quality*, the German listeners could identify the correct pronunciations with an accuracy of 91.1%, while the Chinese listeners could only identify 68.4% and 63% for advanced learners and beginners respectively. If the vowel pairs differed only in *duration*, the accuracy was increased for all groups with a higher increase for the Chinese groups. However, when the modified stimulus differed in both *quality & duration*, the accuracy rate kept rising for German listeners as expected, but it started to drop for advanced learners and remained almost the same for beginners.

A series of mixed-effects models were calculated with the subject as a random effect and perception accuracy as the dependent variable. The results showed that significant differences were found between German native listeners and Chinese listeners with $p < 0.001$ for both advanced and beginners. However, no significant difference was found between Chinese groups with $p = 0.054$. Moreover, significant differences were found between any two conditions for German listeners with $p < 0.05$. But for Chinese listeners, significant differences were only found between *quality* and *duration*, *quality* and *quality & duration*, but not between *duration* and *quality & duration* with $p = 0.45$ for Chinese advanced learners and $p = 0.68$ for beginners.

3.2 Tense/Lax Comparison

Furthermore, we compared the average identification rate for tense and lax vowels separately, which can be seen in Figure 3. For tense vowel identification, the German listeners achieved similar accuracy in either *quality* or *duration*, but a highly increased accuracy in both *quality & duration*; the Chinese listeners showed an increased identification rate in *duration* and a slight

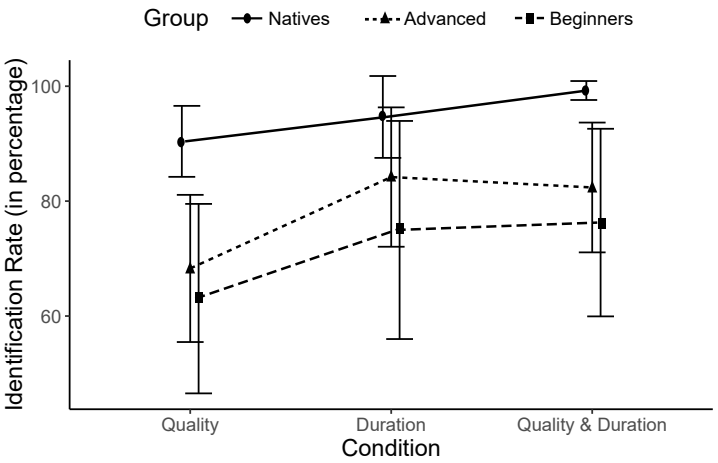
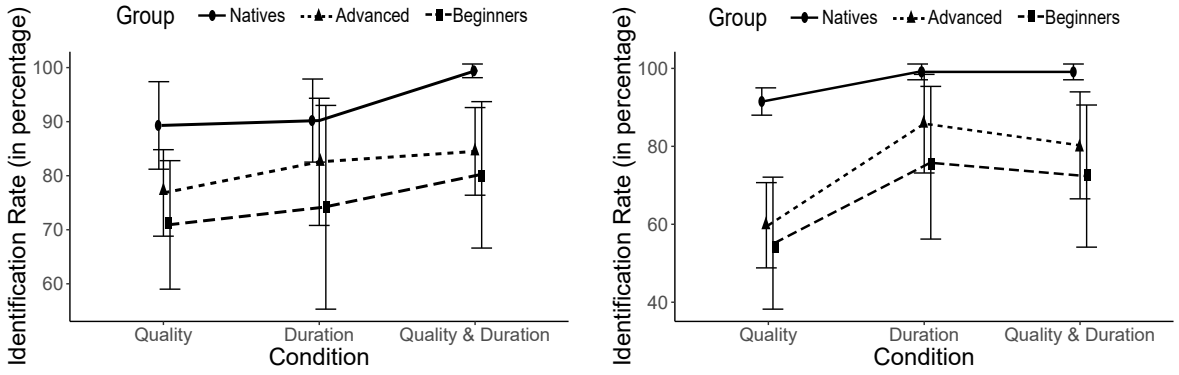


Figure 2 – Mean identification rate (±sd) of three groups in three conditions for all vowels (with solid, dotted and dashed lines for German natives, Chinese advanced learners and beginners, respectively).

increase in both *quality & duration*. For lax vowel identification, the German listeners demonstrated an increase in *duration* but similar accuracy when the quality cue was also presented in *quality & duration*; the Chinese listeners displayed a sharp increase from the *quality* condition to the *duration* condition, but a clear drop in the *quality & duration* condition.



(a) Mean identification rate for tense vowels.

(b) Mean identification rate for lax vowels.

Figure 3 – Mean identification rate (±sd) of three groups in three conditions for tense and lax vowels (with solid, dotted and dashed lines for German natives, Chinese advanced learners and beginners, respectively).

3.3 Vowel Pair Comparison

Different groups showed different identification accuracy patterns for different vowel pairs. In Figure 4, condition *C1*, *C2*, and *C3* represent *Quality*, *Duration*, and *Quality & Duration* respectively, and the vowels are represented in German SAMPA transcription for the sake of clear display in the figure.

It is clear that for German native listeners, the perception accuracy was reduced if only the quantity cue was contrasted in following pairs: [e:]-[ɛ], [i:]-[ɪ], [u:]-[ʊ], but the perception accuracy drastically increased for [a:]-[a] pair. Though they demonstrated different patterns from condition *quality* over *duration* to *quality & duration*, the German listeners could finally reach almost 100% accuracy in condition *C3* for all vowel pairs. For Chinese learners, from condition *quality* to *duration* both groups demonstrated an increase in the identification rate for all pairs except for [e:]-[ɛ], and the accuracy increased differently for different vowel pairs with

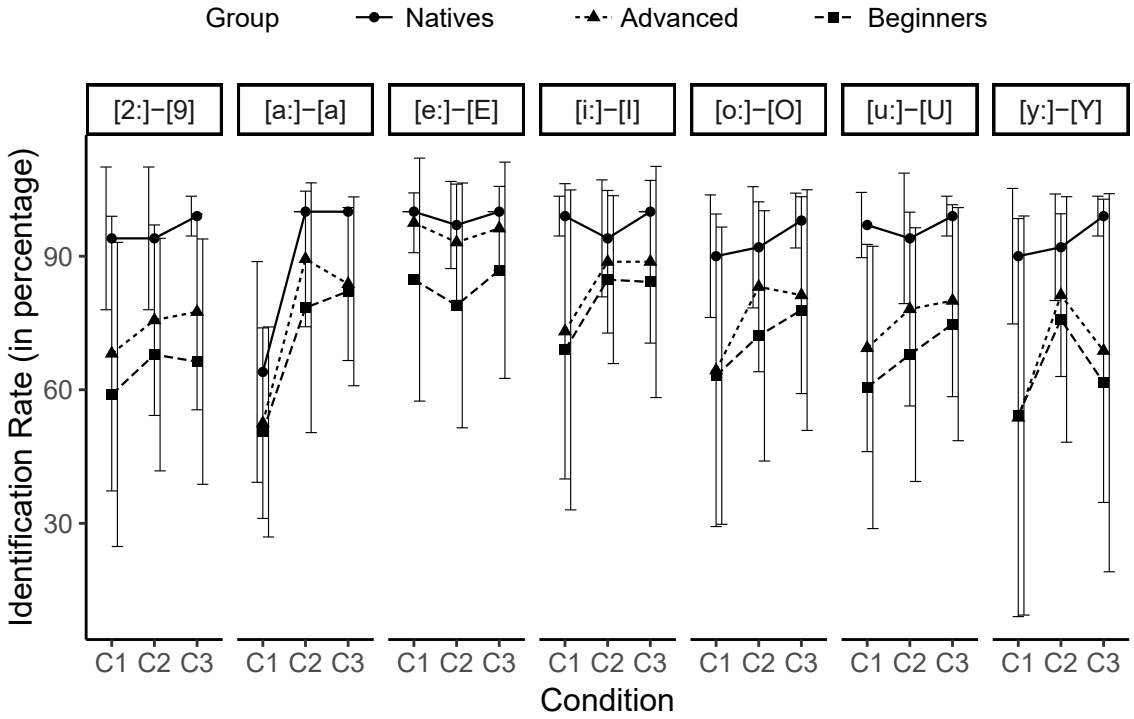


Figure 4 – Mean identification rate (±sd) of three groups in three conditions for vowel pairs (with solid, dotted and dashed lines for German natives, Chinese advanced learners and beginners, respectively).

the largest increase for [a:]-[a] and [y:]-[ʏ]; from condition *duration* to *quality & duration*, both learner groups showed an increase except for the beginner group for some vowel pairs (e.g. [ø:]-[œ], [i:]-[ɪ], and [y:]-[ʏ]) and advanced learner group in other pairs (e.g. [a:]-[a], [i:]-[ɪ], [o:]-[ɔ], and [y:]-[ʏ]).

4 Discussion

As it was pointed out in [2], the problem of quality versus length in the perception of German vowels is a complex one. However, on the basis of the results, we can propose answers to the questions posed at the beginning: (1) German native speakers can distinguish tense-lax vowel contrast successfully when both quality and duration cues are presented, while Chinese learners cannot; (2) Compared to German native listeners, Mandarin Chinese learners rely more on duration to distinguish tense-lax vowel contrast, and advanced learners demonstrated a better performance in every condition for each vowel than beginners, which means that the identification accuracy can be improved if the German proficiency level has been raised. For German native speakers, our findings are in agreement with the previous literature:

- For the high vowels (such as [i:]-[ɪ] and [u:]-[ʊ]), quality seems to be more important than duration, whereas for the low vowel pair [a:]-[a], duration plays a more important role [2].
- For tense vowels, the length cue is not so important. That means, tense vowels do not change their category if the duration is shortened; however, duration is an important cue for lax vowels.
- There is strong evidence that the relationship of quality and duration is an inverse relationship: if the quality is distinctive enough, more variation is allowed for duration; if

the quality is not distinctive enough for the discrimination, the duration cue is more critical. Finally, with both quality and duration, tense-lax vowel pairs can be successfully distinguished [2].

- A detailed look at the individual vowels proves that the tense vowels in these pairs were responsible for the reduced identification accuracy for these pairs. The main reason could be that the duration of these tense vowels in the canonical realization was too long to be natural for the German native listeners, while a shortened tense vowel sounded more natural for them. Moreover, because the duration of tense vowels varies in different syllable structures and contexts, a constant inherent value is not the optimal one in each of the stimulus words [9].

For Chinese learners, we have obtained some new findings:

- They rely more on quantity than quality to identify vowel categories. Provided with the durational cue, Chinese learners can obtain comparable identification accuracy (or even better for Chinese advanced learners) for some vowels without the vowel quality cue.

- They showed a better performance for some pairs [a:]-[a], [e:]-[ɛ], and [i:]-[ɪ] compared to other pairs [ø:]-[œ], [o:]-[ɔ], [u:]-[ʊ], and [y:]-[ʏ]. The reasons can be various. Possible explanations may be that the former three vowel pairs are located in the peripheral area of the vowel space, which can be easily distinguished in perception. Moreover, the Chinese learners have acquired a similar vowel contrast in their first foreign language of English. On the contrary, the latter four vowel pairs are located in a relatively centralized area in the vowel space, which predicts more difficult for Chinese listeners who do not have central vowels in their native phonological system, and furthermore, some vowels like [ø:]-[œ] [y:]-[ʏ] are new phonemes which have not been learned in English. An accurate identification is thus also difficult for them.

- Last but not the least, a potential negligence in the experiment that might bring about additional difficulties for some Chinese learners was the relationship between spelling and pronunciation. For some learners it is not quite clear which vowel should be pronounced as tense and which should be pronounced as lax in the orthographic word form. Thus orthographic influences should also be taken into consideration [10].

Compared to German native speakers, Mandarin Chinese learners rely more on duration than quality to distinguish tense-lax vowels. Some efforts can be made to improve the identification experiment in the future. For the experimental design, the effect of orthography and word familiarity influence can also be taken into consideration; for the generation of stimuli, the duration of the vowels should be adjusted according to the syllable structure and phoneme combination because the influence has been demonstrated in many experiments [11]; for the experiment procedure, fillers can be added, each stimulus can be repeated for several times, and reaction time can be calculated with other psycholinguistic software programmes.

5 Conclusion

This study employed an articulatory speech synthesizer and successfully manipulated vowel quality and duration of speech stimuli separately. On the basis of the neatly controlled cues, German natives and Mandarin Chinese learners have been shown to employ different perceptual cues to distinguish German tense and lax vowels.

6 Acknowledgements

This research work is partially sponsored by the Major Program of National Social Science Foundation of China (18ZDA293), Shanghai Social Science project (2018BYY003), the Interdisciplinary Program of Shanghai Jiao Tong University (14JCZ03) and China Scholarship Council.

References

[1] WEISS, R.: *Relationship of vowel length and quality in the perception of German*. *Linguistics*, 12(123), pp. 59–70, 1974.

[2] WIESE, R.: *Underspecification and the description of Chinese vowels*. In J. WANG and N. SMITH (eds.), *Studies in Chinese Phonology*, pp. 219–249. The Hague: Mouton de Gruyter, 1997.

[3] DING, H., O. JOKISCH, and R. HOFFMANN: *Perception and analysis of Chinese accented German vowels*. *Archives of Acoustics*, 32(1), pp. 89–100, 2007.

[4] BIRKHOLZ, P.: *Modeling consonant-vowel coarticulation for articulatory speech synthesis*. *PLoS ONE*, 8(4): e60603, 2013.

[5] BIRKHOLZ, P., L. MARTIN, Y. XU, S. SCHERBAUM, and C. NEUSCHAEFER-RUBE: *Manipulation of the prosodic features of vocal tract length, nasality and articulatory precision using articulatory synthesis*. *Computer Speech & Language*, 41, pp. 116–127, 2017.

[6] *Goethe-Zertifikat A1: Start Deutsch 1 – Wortliste*. 2004. URL `https://www.goethe.de/pro/relaunch/prf/de/A1_SD1_Wortliste_02.pdf`.

[7] XU, Y. and F. LIU: *Tonal alignment, syllable structure and coarticulation: Toward an integrated model*. *Italian Journal of Linguistics*, 18(1), pp. 125–159, 2006.

[8] KOHLER, K. J.: *Zeitstrukturierung in der Sprachsynthese*. In A. LACROIX (ed.), *ITG-Tagung Digitale Sprachverarbeitung*, pp. 165–170. vde-Verlag, Berlin, Bad Nauheim, 1988.

[9] SENDLMEIER, W. F.: *Der Einfluß von Qualität und Quantität auf die Perzeption betonter Vokale des Deutschen*. *Phonetica*, 38, pp. 291–308., 1981.

[10] NIMZ, K. and G. KHATTAB: *L2 sound perception: Does orthography matter?* In *ICPhS XVIII*. 2015.

[11] HARRINGTON, J., F. KLEBER, U. REUBOLD, and J. SIDDINS: *The relationship between prosodic weakening and sound change: evidence from the German tense/lax vowel contrast*. *Laboratoy Phonology*, 6(1), pp. 87–117, 2015.