The Phonetician
Journal of the
International Society of Phonetic Sciences

Editorial board

President:
Ruth Huntley Bahr
e-mail: rbahr@usf.edu

Editor-in-Chief:
Mária Gósy
e-mail: gosy.maria@nytud.mta.hu

Angelika Braun,
University of Trier,
Germany

Nick Campbell,
Trinity College Dublin,
Ireland

Jens Edlund,
KTH, Sweden

Susanne Fuchs,
Centre for General
Linguistics, Germany

Hilmi Hamzah,
Universiti Utara,
Malaysia

Valerie Hazan,
University College
London, England

Katarzina Klessa,
Adam Mickiewicz
University, Poland

Jens-Peter Köster,
University of Trier,
Germany

Marko Liker,
University of Zagreb,
Croatia

Alexandra Markó,
Eötvös Loránd University,
Hungary

Vesna Mildner,
University of Zagreb,
Croatia

Sylvia Moosmüller,
University Wien, Austria

Tilda Neuberger,
Research Institute for
Linguistics, MTA,
Hungary

Daniel Recasens,
Autonomous University
of Barcelona, Spain

Judith Rosenhouse,
Swantech Ltd., Israel

Radek Skarnitzl, University
of Karlova, Czech Republic

Elisabeth Shriberg,
SRI International, USA

Vered Silber-Varod,
The Open University of
Israel, Israel

Masaki Taniguchi,
Kochi University, Japan

Jürgen Trouvain,
University of
Saarbrücken, Germany

Editorial assistant:
Tekla Etelka Gráczi
e-mail: graczi.tekla.etelka@nytud.mta.hu

Submissions should be sent to:
e-mail: j.phonetician@gmail.com
ISPhS
International Society of Phonetic Sciences

President:
Ruth Huntley Bahr

Honorary President:
Harry Hollien

Secretary General:
Mária Gósy

Vice Presidents:
Angelika Braun
Marie Dohalská-Zichová
Mária Gósy
Damir Horga
Heinrich Kelz
Stephen Lambacher
Asher Laufer
Judith Rosenhouse

Past Presidents:
Jens-Peter Köster
Harry Hollien
William A. Sakow †
Martin Kloster-Jensen†
Milan Romportl †
Bertil Malmberg †
Eberhard Zwirner †
Daniel Jones †

Honorary Vice Presidents:
A. Abramson
S. Agrawal
L. Bondarko
E. Emerit
G. Fant †
P. Janota †
W. Jassem †
M. Kohno
E.-M. Krech
A. Marchal †
H. Morioka
R. Nasr
T. Nikolayeva †
R. K. Potapova
M. Rossi
M. Shirt
E. Stock
M. Tatham
F. Weingartner
R. Weiss

President’s Office:
Prof. Dr. Ruth Huntley Bahr
Dept. of Communication Sciences and Disorders,
University of South Florida
4202 E. Fowler Ave., PCD 1017 Tampa,
FL 33620-8200 USA
Tel.: ++1-813-974-3182
Fax: ++1-813-974-0822
e-mail: rbahr@usf.edu

Secretary General’s Office:
Prof. Dr. Mária Gósy
Dept. of Phonetics, Research Institute for Linguistics,
Hungarian Academy of Sciences
Benczúr u. 33 H-1068 Budapest
Hungary
Tel.: ++36 (1) 321-4830 ext. 172
Fax: ++36 (1) 322-9297
e-mail: gosy.maria@nytud.mta.hu
The Phonetician

A Peer-Reviewed Journal of ISPhS/International Society of Phonetic Sciences

ISSN 0741-6164

Number 113 / 2016

CONTENTS

Note from the President ........................................................................................................ 5

Perception of Consonant Length Opposition in Hungarian Stop Consonants
  by Tilda Neuberger ........................................................................................................... 6

Pronunciation Norms and Pronunciation Habits of Orthographic <ä, äh> in
  Standard Austrian German by Bettina Hobel, Sylvia Moosmüller & Christian Kaseß
  ........................................................................................................................................ 24

Formant Analysis of the Mid-front Vowel as Realized in Hesitation Disfluencies in
  Hebrew by Vered Silber-Varod and Noam Amir ................................................................. 49

Temporal Patterns of Error Repairs and Appropriateness Repairs in Hungarian
  by Mária Gósy, Irene Vogel & Viola Váradi ......................................................................... 61

IZÉ ‘stuff’: A Corpus-based Study on a Curious Hungarian Filler
  by Dorottya Gyarmathy ....................................................................................................... 79

Call for Papers .......................................................................................................................... 100

ISPhS Membership Application Form .................................................................................. 104

News on Dues .......................................................................................................................... 105
NOTE FROM THE PRESIDENT

I am excited to announce the initiation of our new peer-reviewed journal, *The Phonetician*. For over 50 years, the International Society of Phonetic Sciences (ISPhS) has published a newsletter that contained summaries of recent conferences, brief notes on colleagues, important updates in the phonetic sciences, and one or two articles per issue. Since ISPhS moved to a totally online format, we have been publishing this newsletter once a year using guest editors who have solicited articles within a particular topic. So many people have shown an interest in submitting research articles to our newsletter that Dr. Mária Gósy suggested that we formalize our newsletter as an official peer-reviewed journal. I thought that this was an excellent idea.

The current issue serves as the inaugural volume of our new journal, *The Phonetician*. This new journal is peer-reviewed and focuses on issues related to speech science. It will continue to be edited by guest editors under the leadership of the ISPhS President, a Chief Editor, Mária Gósy, and an Editorial Board of distinguished Phonetics from around the world. This open access journal will serve as an excellent venue for phoneticians to share their scientific projects and results. Finally, instead of having the cover represent the contents of the issue, we will standardize *The Phonetician’s* cover. In this way, people will begin to recognize this publication as an official journal.

We hope that you enjoy this issue and the new format. We also hope that you will consider organizing an issue that reflects your particular interests in speech science. The Editorial Board is excited to see this online, open access journal flourish.

*Ruth Huntley Bahr*
President of ISPhS
PERCEPTION OF CONSONANT LENGTH OPPOSITION IN HUNGARIAN STOP CONSONANTS

Tilda Neuberger
Dept. of Phonetics, Research Institute for Linguistics,
Hungarian Academy of Sciences
e-mail: neuberger.tilda@nytud.mta.hu

Abstract
A specific speech sound is characterized by strong variability in articulation. Nevertheless, despite great variability, human listeners are able to identify phonemes successfully. The most important acoustic cue that distinguishes geminate from singleton stop consonants in production is the closure duration. The aim of this paper is to examine the role of duration as a perceptual distinction between single and geminate stops using a binary discrimination test of stops with systematically manipulated closure duration. Results confirmed that closure duration is a sufficient and adequate perceptual cue in the distinction of single and geminate stops; however, differences between the response curve of original geminates and that of original singletons indicate that cues other than closure duration may contribute to the length contrast in Hungarian stops.

Keywords: geminate, quantity, perception, stop consonant, closure duration, manipulation

1 Introduction
One of the most interesting questions in speech perception research is how the continuous speech signal can be recognised as sequences of speech sounds. Furthermore, identification of speech sounds as discrete phonemes is also an important issue. It has been shown that a specific speech sound is characterized by strong variability in articulation, including timing of articulation (Lehiste, 1970; Lisker, 1974; Rosen, 1992). In spite of this great variability, however, human listeners are able to identify phonemes successfully, irrespective of speaker or speech condition. During the identification of consonants, listeners make decisions about place of articulation, manner of articulation, voicing contrast and length contrast.

The present paper investigates the perception of consonantal length contrast in Hungarian, which is a language that has distinctive long consonants. The aim of this study is to examine how native listeners discriminate phonological categories of quantity (singleton and geminate) along a continuous durational scale of the phonetic realisations.

There has been much research indicating that duration of singleton and geminate consonants shows considerable overlap in production (e.g., Lisker, 1958; Pickett et
al., 1999; Hirata & Whiton, 2005; for Hungarian consonants see Gósy, 2004; Olaszy, 2006; Beke & Gyarmathy, 2010; Gráci, 2012; Neuberger, 2015), which might make discrimination difficult to listeners in some cases, particularly during first or second language acquisition. In Hungarian children, for instance, the most difficult perceptual process at the phonological-phonetic level has proved to be the consonantal length contrast discrimination (Gósy, 1989; Horváth & Gyarmathy, 2010). Still, adult listeners seem to have the ability to effectively cope with this task in everyday speech situations. The effectiveness of this process may be due to the extraction of acoustic cues and phonetic information that speech sounds contain. More precisely, different phonetic categories correlate with various acoustic cues facilitating human perception of different phonemes. However, there is not a one-to-one mapping between acoustic features and perceptual categorization of speech sounds (Blumstein & Stevens, 1985).

According to the acoustic invariance theory of speech perception, invariants establish a link between acoustics and perception during the identification of phonetic segments (Stevens & Blumstein, 1978, 1981, 1985; Blumstein & Stevens, 1979; Lisker, 1985). It has been suggested that only a limited set of articulatory configurations leads to stable acoustic patterns. These patterns are the ones that define the finite set of speech sounds. Similarly, from the perspective of speech perception, the human auditory system can judge differences between sounds along a physical scale. Two items on the continuum which cannot be discriminated by the perceptual system are identified as realisations of the same phoneme, while the items that are discriminated are deemed to belong to different phonetic classes (Blumstein & Stevens, 1979). It is assumed that the acoustic signal contains invariant information that is present in all instances that correspond to the perceived linguistic unit (Wright et al., 1999).

A considerable number of studies have contributed to the elaboration of this theory. Most of these studies focused on invariant acoustic properties that can be used to classify stop consonants according to place of articulation (Fant, 1960; Stevens & Blumstein, 1978; Sussmann et al., 1991). Besides place-of-articulation categories, the acoustic and perceptual correlates of other consonantal contrasts have been examined, such as the voiced-voiceless distinction (Liberman et al., 1958; Lisker & Abramson, 1964; Stevens & Klatt, 1974; Williams, 1977; Sussmann et al., 1991) or the single-geminate distinction (Pickett & Decker, 1960; Lisker, 1974; Abramson, 1986, 1987; Hankamer & Lahiri, 1988; Hankamer et al., 1989; Schmidt & Flege, 1995). This study will focus on the single-geminate distinction.

Production and perception studies have revealed that the primary acoustic attribute that distinguishes geminates from singletons is duration (e.g., Pickett & Decker, 1960; Klatt, 1976; Hankamer et al., 1989; Ylinen et al., 2005; Ridouane, 2010). Geminates are longer than matched singletons in many languages, such as Arabic, Japanese, Italian, Finnish or Hungarian. In other languages, consonant length is not contrastive, geminates only arise from morpheme concatenation (also known as fake geminates). Hence, geminates may occur across morpheme boundaries, but not morpheme-internally (e.g., English *top pick* vs. *topic*, French *Il l’aime* vs. *Il aime*). Both contrastive and non-contrastive geminates were found to be longer than matched
singletons, on average. In Hungarian, investigation of the quantity contrast may be important for several reasons. In this language, all consonants can occur as both short (singleton) or long (geminate). Considerable overlap in the production of singletons and geminates can make classification of the two categories difficult. In addition, the distribution of geminates is restricted in Hungarian. If an underlying geminate occurs next to another consonant, it obligatorily degeminates and must surface as short. Therefore, the degemination process has an influence on the temporal properties of consonants. In this case, the degree of shortening is a key issue (see Siptár & Gráczi, 2014). For stop consonants, studies have commonly shown that the most important acoustic and perceptual cue for length distinction is the duration of the closure phase (Ham, 2001; Ridouane, 2010). Therefore, perception research has largely focused on the measurement of this parameter.

### 1.1 Previous work on perception of consonant length

As background to the present study, several perception experiments which applied an artificial manipulation technique on closure duration are described here. Listeners’ responses to incrementally manipulated durations allow one to determine the point at which a singleton percept shifts to a geminate percept. Thus, durational correlates of the length distinction can be examined. Some of the following studies focused on the role of absolute closure duration as a measure of length distinction, while other studies emphasized the role of relational timing in various languages.

**Absolute duration** was investigated by the manipulation of closure duration in many languages, such as Marathi, Arabic, Turkish, Bengali and English. Lisker’s (1958) perceptual experiment dealt with the length distinction of Marathi stop consonants. Closure duration of [t] in the word *matə* ‘mind’ was artificially manipulated to increase in 20 ms steps, while the closure duration of the long counterpart in the word *mattə* ‘drunk’ was manipulated to decrease in 20 ms steps. The perceptual boundary value was noted between 140 and 160 ms in the former case and between 140 and 120 ms in the latter case. For Arabic, a very similar result to Lisker’s (1958) results was found by Obrecht (1965). In this case, the perceptual boundary between geminate and non-geminate [bː]-[b] were between 140 and 160 ms.

The region between 120 and 160 ms of closure duration proved to be an important interval in both Turkish and Bengali length distinctions. Hankamer et al. (1989) employed two sets of stimuli with closure duration varying incrementally between that of geminate and non-geminate stops in Turkish and Bengali. Several minimal pairs were used in this experiment and the task of the listeners was to write down the word they thought they heard (e.g., Turkish [ata] ‘horse (dat)’ or [atːa] ‘horse (loc)’ and Bengali [paːta] ‘leaf’ or [paːtːa] ‘whereabouts’). Responses to stimuli created from original geminate differed from those for non-geminates in both languages. The original geminates were identified as geminates more frequently than the original non-geminates at closure durations between 120 and 160 ms. The authors concluded that the responses were “biased by secondary features of the acoustic signal when the closure duration cue is in the ambiguous region between 120-160 ms” (Hankamer et
al., 1989: 295). However, examination failed to reveal the exact secondary cue. Thus, it was assumed that the bias may be due to a combination of cues.

A perceptual distinction experiment was carried out using English test sentences with an intervocalic single consonant *(topic)* and its double counterpart *(top pick)* as stimuli (Pickett & Decker, 1960). The closure duration of *[p]* was altered by inserting or removing magnet tape. The effect of closure duration and rate of utterance was tested. Findings showed that closure durations shorter than 150 ms were judged as a single consonant, while closures longer than 250 ms were judged as geminate. It was also found that as the rate of utterance increased (from 2 to 8 syllables per second), the threshold closure duration decreased (from 320 to 140 ms). This raises the question of whether acoustic invariance exists in an absolute form.

Numerous researchers have assumed that acoustic invariance is **relational invariance**, and relational timing plays a crucial role in the perception of length contrast. In other words, not only, and not primarily, absolute closure duration plays a role in length distinction, but other variables, such as various durational ratios are critical. In addition, it is important to take speaking rate into consideration as well.

Pind (1995) proposed relational timing as an important acoustic property which is able to define durational categories across speaking rates. His findings showed that the ratio of vowel to rhyme duration was a stable acoustic feature in distinguishing VːC versus VCːsyllable categories in Icelandic, and it remained invariant at different speech rates.

The closure duration of labial and dental stops was found to discriminate singletons and geminates in a production and a perception experiment in Italian (Pickett et al., 1999). In production, the duration of singletons and geminates showed overlap across different speaking rates, however, the ratio of consonant duration to preceding vowel duration remained stable across speaking rate. It was hypothesised that manipulation of the C/V ratio affected the perception of the quantity distinction and resulted in perceptual shifts. A significant main effect of C/V ratio was found, indicating that listeners tended to change the category of their responses as a result of the timing manipulation.

In accordance with the findings in Italian by Pickett et al. (1999), Hansen’s (2004) production data for Tehrani Persian demonstrated that increased speaking rate had a greater influence on the closure duration of geminates than it did on the duration of singletons. His pilot study revealed that different threshold values were required to discriminate between singletons and geminates for isolated and for connected speech because of the overlap of isolated singleton durations and connected geminate durations.

Japanese stop quantity distinction was investigated in the theoretical framework of relational acoustic invariance (Amano & Hirata, 2010). The authors analysed the perceptual boundary between single and geminate stops across speaking rates (fast, normal, slow). They found that closure duration at the perception boundaries ranged between a relatively large interval across speaking rates (34–213 ms, mean: 110 ms). It was also found that the durational ratio of stop closure to word (CW ratio) is an invariant parameter in distinguishing the two phonemic categories.
Japanese stop quantity distinction was examined with respect to another durational ratio by Idemaru and Guion-Anderson (2010). They created test stimuli varying in previous mora duration and following vowel duration, while leaving closure duration and VOT unaltered. Listeners’ ‘geminate’ responses increased as C/Mora\textsubscript{1} ratio increased, whereas a less important change was observed due to the C/V\textsubscript{2} variation.

Perception of consonant length was investigated among non-native listeners as well (Hayes, 2002; Wilson et al., 2005; Sonu et al., 2013). Results of a perception experiment of Japanese length contrast differentiation by English listeners suggested that non-native listeners did not use cues that vary by speaking rate but instead used absolute durational criteria (Wilson et al., 2005). Similar findings were observed in Korean learners of Japanese (Sonu et al., 2013). These findings suggest that non-native listeners tend to have difficulty identifying length contrasts affected by speaking rate variations.

1.2 Research questions and hypothesis of the present work

Three research questions were addressed in the present study. First, how does closure duration contribute to native listeners’ discrimination of Hungarian singleton and geminate stops? Second, does place of articulation or voicing affect the differentiation of singleton and geminate stops? Third, do the response curves displayed across varied closure durations show the same pattern for the original singletons as for the original geminates?

To answer the questions above, a two-alternative forced choice test was applied for stops with systematically manipulated closure duration. The stimuli contained Hungarian stops \([p, t, k, b, d, g]\)\textsuperscript{1} and their geminate counterparts.

Based on previous results for other languages, it was hypothesised that (i) closure duration would be a sufficient cue to quantity discrimination for native Hungarian listeners. It was also assumed that (ii) place of articulation and stop voicing might have an effect on the closure durations associated with perceptual boundaries. Finally, it was hypothesised that (iii) the original quantity of the stops would affect the listeners’ responses to some extent due to probable secondary cues (e.g., preceding vowel duration, closure voicing or combination of acoustic characteristics) in the acoustic signal.

2 Method

2.1 Baseline experiment

A baseline experiment was carried out to ensure high reliability in constructing the test stimuli for the main perception experiment. Therefore, a set of nonwords containing single and geminate stops was created for this pre-experiment. VCV and VC:V sequences were recorded by a 27-year-old, Hungarian native female speaker. Being a qualified phonetician, she produced the sequences maintaining fundamental

---

\textsuperscript{1} Stops \([c, j]\) were not analysed because there is no consens on their manner of articulation whether they are stops or affricates.
frequency and sound pressure relatively constant in each token. The recording was made with an AT 4040 side-address condenser microphone using GoldWave software in a sound-proof booth located within the Phonetic Department of the Research Institute for Linguistics of the Hungarian Academy of Sciences. Recordings were digitized with a sampling rate of 44.1 kHz (storage: 16 bits, 86 kbytes/s, mono).

The consonants of the sequences were \([p, t, k, b, d, g]\) and their long counterparts, whereas the preceding and the following vowels were identical. Vowels were \([i\), \([aː]\), or \([u]\), since they represent high, low and back vowels, respectively. The speaker read 36 items \([6 \text{ (consonants) } \times 2 \text{ (quantity) } \times 3 \text{ (adjacent vowels)}\] , such as \([ipi], [uku], [iti], [aːdaː]\) etc. She produced each item three times, resulting in 108 productions. These nonwords were prepared for the baseline experiment: 54 nonsense word realisations contained short stop consonant sequence, and 54 nonsense word realisations contained long stop consonants.

The perceptual robustness of the length distinction was evaluated in the baseline experiment. Eleven Hungarian-speaking adults (9 females, 2 males) participated in this experiment. Their mean age was 27 years (range = 22–34 years). Each listener completed a discrimination test using Praat software (Boersma & Weenink 2015). The task of the listeners was to listen to the samples and make a binary decision about whether the heard consonant was long or short, for example \([iti]\) or \([iti]\). If it was heard as long, they chose ‘long’ response, whereas when the presented consonant seemed to be short, they had to click the ‘short’ answer on the screen. Items were played in random order. Reaction time were measured by means of Praat. Identification accuracy was measured for each token. Identification accuracy was defined as 100% when all participants gave the same ‘short’ or ‘long’ response to a token containing short or long consonant corresponding with the speaker’s intention.

Results of the baseline experiment revealed that listeners identified consonant length at an accuracy rate of 82–100%. In general, the poorest result was in the case of low vowels, while the best result was in the case of the front high vowels (identification accuracy: 98.5% on average). Therefore, listeners’ responses to each token containing \([i]\) were further analysed, while tokens containing \([u]\) and \([aː]\) were excluded from further investigation. Thus, listeners’ responses to 36 tokens containing vowel \([i]\) and consonants \([p, t, k, b, d, g]\) and their long counterparts were examined with respect to identification accuracy and reaction time.

Identification accuracy of each token can be seen in Figure 1. Correct identification accuracy of 100% means unanimous agreement among the listeners. The lower the identification accuracy, the more listeners gave incorrect responses to the token. It was found that all listeners gave correct responses to at least one of the three produced token regarding place of articulation (labial, alveolar, velar) and voicing (voiceless, voiced) of stops. Given the results of the baseline identification experiment, 12 tokens were selected for manipulation in the main perception test. Selection was made based on the combination of high accuracy rate (100% in each case) and relatively short reaction time (< 1.2 ms). Overall, the mean reaction time duration was 1.5 ms.
Figure 1. Correct identification accuracy for each token in the baseline experiment

2.2 Perceptual test stimuli

In order to construct the perceptual test stimuli with different closure durations, as mentioned above, 12 tokens were selected for manipulation: 6 tokens with singleton and 6 tokens with germinate stops. Using these tokens, two sets of stimuli were created. For one of them, tokens with a short consonant were used (e.g., [ipi] or [idi]). For the other one, tokens with long consonants were used (e.g., [ipi] or [idi]). In the first case, stimuli were made by artificially lengthening the closure duration of original singletons [p, t, k, b, d, g] in 10 ms steps up to the closure duration of the matched geminates [pː, tː, kː, bː, dː, gː]. For instance, closure duration of the originally singleton [t] was manipulated from 104 ms to 224 ms in twelve 10 ms steps in order to create a continuum from [ipi] to [ipːi]. For the other set of stimuli, the closure duration of the original geminates was shortened likewise in 10 ms steps, resulting in nonsense words
with artificially shortened geminate. Minimum and maximum values of closure durations and the number of steps of manipulation are listed in Table 1 for each consonant. The reason for having two sets of stimuli, i.e., shortened geminates and lengthened singletons, was to observe any possible differences in listeners’ ‘short’ or ‘long’ responses for original singletons and original geminates with equal closure duration. It is assumed that if response curves for the originally geminate stimuli are located somewhere else than those of the originally singleton stimuli, acoustic cues other than closure duration may play a role in singleton vs. geminate stop distinction.

Incremental manipulation of closure duration was conducted using a Praat script. In order to do this, sequences had to be first segmented into speech sounds. Consonant closure boundaries (start time and end time) were marked based on visual observation of the oscillogram and spectrogram, as well as auditory feedback. Closure duration was defined as the time interval between the termination of the preceding vowel and the stop burst. Start time of the closure duration was measured at the offset of the vertical striations of the preceding vowel’s formants, while the end time of closure duration was measured right before the release burst. The content of closure duration, namely the silent interval in voiceless stops and voicing in voiced stops, was preserved during the manipulation. Temporal and spectral properties of the adjacent vowels, the VOT of voiceless stops, and the burst releases remained unaltered. Altogether, 138 tokens were created.

<table>
<thead>
<tr>
<th>Consonant</th>
<th>Minimum-maximum closure duration (ms)</th>
<th>Number of steps of manipulation</th>
</tr>
</thead>
<tbody>
<tr>
<td>[p]-[pː]</td>
<td>103–204</td>
<td>10</td>
</tr>
<tr>
<td>[t]-[tː]</td>
<td>104–224</td>
<td>12</td>
</tr>
<tr>
<td>[k]-[kː]</td>
<td>104–204</td>
<td>10</td>
</tr>
<tr>
<td>[b]-[bː]</td>
<td>102–222</td>
<td>12</td>
</tr>
<tr>
<td>[d]-[dː]</td>
<td>104–224</td>
<td>12</td>
</tr>
<tr>
<td>[g]-[gː]</td>
<td>105–175</td>
<td>7</td>
</tr>
</tbody>
</table>

2.3 Perceptual test subjects
Forty-four native speakers (33 females, 11 males) of standard Hungarian participated in this experiment. They were recruited from a university in Budapest. All of them were undergraduate students of linguistics, having little experience in phonetic studies. Their age ranged from 18 to 27 years, and the mean age was 21 years. No participant reported being diagnosed with a speech or hearing disorder.

2.4 Perceptual test conditions
The experiment was run in a quiet room at the university in Budapest. Each participant listened to the recorded samples through Sennheiser HD 419 headphones. The nonwords were played one by one in random order. The listeners’ task was to make a binary decision, choosing between ‘short’ or ‘long’ responses, by clicking the
appropriate button in Praat software (i.e., the same procedure as in the baseline experiment described above).

2.5 Analyses

For evaluating the binary data, listeners’ responses to each item were summarized. Percentage of geminate responses at each duration was measured in the case of all stop consonants. Then the two sets, i.e. originally singleton and originally geminate stimuli, were analysed separately. Fitting a logistic function to the sets and plotting response curves was carried out using MATLAB 2015b.

Binary logistic regression was used for statistical analysis in SPSS 20.0 software. A generalized linear mixed model (GLMM) was constructed by ‘responses’ as the target (or dependent) variable, ‘closure duration’ was the fixed effect, and ‘speaker’ was the random effect. The interaction of closure duration and place of articulation, as well as closure duration and voicing, were analysed (as fixed effects). To compare responses to stimuli created from originally singleton and originally geminate stops, paired samples t-test in SPSS were used.

3 Results

Since closure duration is indeed a major cue to the length distinction in production, it is expected that listeners also use this parameter in perception. Figures 2 and 3 show listeners’ responses by means of response curves. Response curves represent the percentage of ‘long’ responses (y-axis) at different closure durations (x-axis) for each stop. A typical sigmoid function was plotted in accordance with the closure duration of the consonants. As Figures 2 and 3 show for voiceless and voiced stops (respectively), listeners judged consonants with relatively long closure durations as ‘long’ and they hardly judged consonants with relatively short closure durations as ‘long’. It is also worth noting that approximately 100 ms closure duration induced a total agreement of ‘short’ response among the 44 participants (in this case, proportion of ‘long’ responses was close to zero, see Figures 2 and 3). In contrast, closure durations approaching 200 ms triggered unanimous ‘long’ decisions (in this case, the proportion of ‘long’ responses was close to the maximum, see Figures 2 and 3). Binary logistic regression showed that closure duration had a main effect on listeners’ 0-1 (short-long) responses: $F(1, 6067) = 1317.391; p < 0.001$.

A comparison of the response curves for the voiceless and voiced stops revealed similar patterns for [p, t, k, b, d, g]. As can be seen in Figures 2 and 3, when moving more posterior in place of articulation, response curves were arranged at shorter closure durations along the time axis. Statistical analysis revealed a significant interaction between closure duration and place of articulation for the stop: $F(2, 6067) = 29.427; p < 0.001$ and between closure duration and voicing of the stop: $F(1, 6067) = 50.082; p < 0.001$. This means that not only closure duration but other variables have effects on listeners’ responses as well.
These differences in the response curves for different stops are related to the fact that stops of different places of articulation exhibit closure duration differences. Previous production studies provided objective data on closure duration differences among labial, dental and velar stops in Hungarian. For instance, Gósy and Ringen (2009) measured that the mean closure duration was 69 (±14) ms for labial, 59 (±13) ms for dental, and 53 (±14) ms for velar voiced stops in the intervocalic position in isolated words. In Gráczi’s (2013) data using nonsense words, mean closure duration of intervocalic voiceless stops seemed to be shorter by moving more posterior in place of articulation: 90 (±13) ms in labial, 72 (±15) ms in dental, 70 (±13) ms in velar stops on average. The mean closure duration of intervocalic voiced stops was shorter than that of voiceless stops: 70 (±10) ms for labial, 52 (±10) ms for dental, and 60 (±12)
for velar stops (Gráczi, 2013). For spontaneous speech material, the mean closure duration was 79 (±11) ms for labial, 71 (±18) ms for dental, 63 (±18) ms for velar stops (Neuberger, 2015). The former values represent singleton stop consonants. Spontaneous speech data for voiceless geminate stops showed an average closure duration of 115 (±20) ms for labial, 122 (±31) for dental, 106 (±27) ms for velar stops (Neuberger, 2015). Considering these mean values, it was expected that the perceptual shift in listeners’ responses from singleton to geminate would follow the same tendency described by production data.

In order to observe the perceptual shift from singleton to geminate, in the next step of analysis, closure duration at the 50% perceptual boundary point was measured. This represents the 50% rate between the identification of a singleton versus a geminate. The values of 50% boundary closure duration are shown in Table 2 for each stop. The more posterior the place of articulation, the shorter the 50% boundary closure duration. This result shows consistency with the above mentioned production data. More specifically, since stops proved to be produced with shorter closure durations by moving more posterior in place of articulation, it could be expected that 50% boundary closure duration would show this tendency as well. The 50% boundary durations of stops with respect to place of articulation showed this tendency, however, statistical analysis did not confirm that these differences were significant ($p > 0.05$). In terms of place of articulation, the shortest duration boundary involved velar stops and the longest duration was attributed to labial stops. Listeners judged voiced consonants as ‘long’ at a shorter closure duration than voiceless ones. The shortest closure duration perceptually belonged to the voiced velar stop, while the longest one belonged to the voiceless labial stop. Nevertheless, voicing did not show a significant main effect on 50% boundary values ($p > 0.05$). The average boundary value in the entire data set was 153 ms. Listeners tended to judge consonants as ‘short’ below this duration, and as ‘long’ above this duration.

Table 2. 50% boundary closure duration (ms) of stop one by one and in total, as well as across place of articulation and voicing

<table>
<thead>
<tr>
<th></th>
<th>[p]</th>
<th>[t]</th>
<th>[k]</th>
<th>Voiceless</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>163</td>
<td>160</td>
<td>153</td>
<td>159</td>
</tr>
<tr>
<td>[b]</td>
<td>159</td>
<td>147</td>
<td>137</td>
<td>147</td>
</tr>
<tr>
<td>[d]</td>
<td></td>
<td></td>
<td></td>
<td>Voiced</td>
</tr>
<tr>
<td>[g]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Labial</td>
<td>161</td>
<td>153</td>
<td>145</td>
<td>153</td>
</tr>
<tr>
<td>Alveolar</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Velar</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

A comparison of the response curves of originally short and originally long consonants was made in order to reveal the possible presence of any secondary cues that may help listeners in discriminating the two phonological categories. It was assumed that in case of differences between response curves of the two conditions (stimuli made by originally singletons vs. originally geminates), closure duration may not be the only cue to length discrimination. The results showed that response curves
of stimuli created from different original quantity were not identical (Figures 4 and 5 for voiceless and voiced stops respectively). In each graph, blue curves show listeners’ responses to stimuli created from originally singleton stops, while red curves shown listeners’ responses to stimuli created from originally geminate stops (x-axis: manipulated closure duration; y-axis: percentage of listeners’ ‘long’ responses, horizontal yellow line: 50% boundary). A paired samples t-test showed the difference was significant in [k]: $t(10) = -3.798, p = 0.003$; in [b] $t(13) = 3.879; p = 0.002$; and in [g] $t(7) = -3.815; p = 0.007$.

**Figure 4.** Response curves for voiceless stops with respect to the original length of stimuli ([p] – left; [t] – middle, [k] – right)

**Figure 5.** Response curves for voiced stops with respect to the original length of stimuli ([b] – left; [d] – middle, [g] – right)

The difference between the 50% boundary closure duration of the originally singleton and originally geminate stops indicated that listeners’ reactions differed according to the original quantity of some of the stop consonants (see Table 3). The largest shift in the boundary location occurred in [b] with 27 ms, while the smallest shift occurred in case of [d] with 2 ms. In the latter case, listeners’ responses to original geminates and original singletons were basically the same. The difference between the two stimuli conditions
proved to be minor in case of [p] and [t] as well (minor differences in response curves of [p], [t] and [d] can also be observed in Figures 4 and 5).

For the alveolar and velar stops, the perceptual boundary of originally geminate consonants appeared earlier than in the case of responses to stimuli created from original singletons. The displacement of 50% boundary closure durations was the largest between [kː] and [k] (15 ms). At the 50% perception boundary closure, duration was 147 ms in original geminates and 162 ms in original singletons. Interestingly, labial stops showed a different behaviour. In the case of [p] and even more in the [b], response curves of original singletons were positioned at shorter closure durations. The shift in the boundary location was larger in the case of voiced labial stops than in the case of voiceless labial stops. The possible reasons for this displacement are discussed later in this article.

Table 3. 50% boundary closure duration (ms) with respect to the original length of stimuli

<table>
<thead>
<tr>
<th></th>
<th>[p]</th>
<th>[t]</th>
<th>[k]</th>
<th>[b]</th>
<th>[d]</th>
<th>[g]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Originally singleton</td>
<td>162</td>
<td>165</td>
<td>162</td>
<td>146</td>
<td>150</td>
<td>141</td>
</tr>
<tr>
<td>Originally geminate</td>
<td>165</td>
<td>160</td>
<td>147</td>
<td>173</td>
<td>148</td>
<td>137</td>
</tr>
</tbody>
</table>

4 Discussion and conclusion

Production of speech sounds has been shown to be highly variable, which may pose difficulties for listeners, especially concerning the perception of those temporal cues which are linguistically significant, i.e. define relevant phonemic contrasts, like distinctive consonant length. There has been notable research on how listeners deal with this problem in many languages but there is only a limited number of studies about the perception of consonant length contrast in Hungarian. The present study reported data on the role of closure duration in the perceptual distinction between Hungarian single and geminate stops in non-word items. Examination of native listeners’ quantity discrimination was conducted with the intention of obtaining detailed information about the relationship between acoustic and perceptual cues of consonant length in Hungarian.

The first hypothesis – closure duration would be a sufficient cue to quantity discrimination for native Hungarian listeners – was established by the present data. Results of the present study confirmed the statement, which has been formulated for many languages that closure duration is the main perceptual cue to the stop consonant quantity distinction in the Hungarian language as well. This was substantiated by statistical analysis. The stimuli scale of closure duration between 100 ms and 200 ms provided a good medium to observe how listeners’ responses shifted from ‘singleton’ to ‘geminate’. Response curves created from the present data formed S shapes on the dimension of closure duration, which implies that listeners’ perception operates more or less categorically during the process.
As mentioned above in the introduction, studies describing production data on Hungarian singleton and geminate consonants reported largely overlapping durations. This overlap is due to varying speaking rates, individual articulatory properties, stress and phonetic positions of the given consonant etc., and mostly occurs in the interval of 80–120 ms of duration (e.g., production data of Gráczi, 2012; Neuberger, 2015). The question arises whether this overlap in duration causes difficulties for listeners in identifying consonant length. Based on the responses in the 80–120 ms interval in the present perception data, it would be expected for listeners to judge these consonants as singleton, even if they were geminates. But this was not the case in everyday speech situations. This is the reason why it is supposed that, in continuous speech, the perception of quantity extends to larger units than speech sounds (syllable or word-sized units) and listeners identify consonant length considering relational properties and not absolute duration of consonants.

The second hypothesis was that place of articulation and voicing of the stops might have an effect on the closure durations related to perceptual boundaries. The present data supported this hypothesis. Values of the 50% boundary closure duration showed correspondence with the position of the tongue in the mouth during articulation of stop: The more posterior the place of articulation, the shorter the 50% boundary closure duration. We can conclude that this parameter is associated with the different durations of intervocalic stops across places of articulation in production (Gósy & Ringen, 2009; Gráczi, 2012). In the case of shorter closure duration (which is more common in posterior stops), the perceptual boundary between singletons and geminates is situated at lower values.

The third hypothesis was that the original quantity of the manipulated stops would affect the listeners’ responses. Results of the present investigation, along with results from earlier studies, suggested that acoustic cues other than closure duration may play a role in the identification of quantity contrast. The plausible presence of other acoustic cues was revealed by differences in response curves comparing originally geminate to originally singleton stimuli. The displacement of the two response curves along the time axis varied in size across consonant quality. The greatest displacement was manifested in velar stops, on average. This means that possible secondary cues are more salient in these consonants than in labial or alveolar ones. In most of the analysed stops consonants, the 50% boundary closure durations were lower in the case of originally geminate stimuli than in the case of originally singletons. This finding indicates that listeners tended to identify originally geminate stimuli as ’long’ even at shorter closure durations than they did in case of originally singleton stimuli. This suggests that the [+long] feature might be coded in the speech signal by other acoustic properties besides closure duration (e.g., preceding vowel duration, closure voicing), and that the listeners’ auditory system might be sensitive to these articulatory/acoustic characteristics. At this stage in the research process, many questions remained unanswered. What acoustic cues cause the differences in listeners’ responses between the two stimuli conditions? Further acoustic and perceptual analysis is required to determine the relevant secondary cues. The issue of cues other than closure duration
is more significant in languages in which word-initial voiceless stop geminates exist (for instance Pattani Malay, see Abramson, 1987; or Cypriot Greek, see Muller, 2003; or Swiss German, see Kraehenmann & Lahiri, 2008). Despite the fact that the silent interval in the closure of singleton and geminate voiceless stops in utterance-initial position is not audible, listeners are capable of differentiating these two categories. Kraehenmann and Lahiri (2008) revealed by using electropalatography that even if audible closure duration is missing, speakers articulate initial geminate stops with longer oral closures than singletons (based on duration of contact of the tongue and hard palate). Still, there must be other perceptual cues besides closure duration in determining the length distinction of initial stops, such as intensity of stop burst, rate of formant transitions, fundamental frequency perturbations (Abramson, 1987) or voice onset time (Muller, 2003) etc. These acoustic features may function as secondary cues in quantity distinction of non-initial stop geminates. The task of investigating the role of other cues in Hungarian length distinction will be addressed in a future study.

There was an unexpected result which ought to be interpreted. That is, the 50% boundary closure duration of originally singleton [b] and geminate [bː] also showed significant difference, but the response curves of the originally singleton stimuli were positioned at shorter closure durations than that of the geminate. This result may lead to the conclusion that the originally geminate [bː] stimuli did not contain secondary cues which seemed to be evident in the velar and alveolar stops. Or, which is more likely, there was some aspect of the stimuli that made identification complicated. After reanalysis, it can be seen that the percentage of voicing in closure phase (which was not 100% in the original geminate [bː] nor in the original singleton [b]) remained stable along with manipulation of closure duration. Using aerodynamic modelling, Westbury and Keating (1986) stated that closure of a relatively long intervocalic stop is likely to be initially voiced and then voiceless. Moreover, Ohala (1983) concluded that geminates not voiced through release are generally categorized as voiceless. By shortening the closure duration of [bː], the voiced part also decreased, and by lengthening closure duration of [b], the voiceless part also increased, which might have caused uncertainty in the identification of stop quality in [p] or [b]. (It must be noted that all [d] and [g] tokens in the material were fully voiced, therefore, the above mentioned problem was not present in cases of these consonants.) Previous research has confirmed that voiceless stops can be distinguished from voiced counterparts based on voicing in closure phase and are (significantly) longer that voiced counterparts, thus, consonant duration may function as secondary cue to voicing contrast (Maddieson, 1997; Olaszy, 2006; Gósy & Ringen, 2009; Gráczi, 2011). It is supposed in this study that an inverse phenomenon also exists, that is, voicing properties (partially-voiced closure) can be a secondary cue to length distinction, by influencing the percentage of voicing in closure that distinguishes singleton and geminate stop consonants. Nevertheless, further examination is needed to provide evidence for what effects partial voicing have on the percept of geminate stops.
Another important issue to investigate is the Hungarian length contrast in the face of variation in speaking rate and on the basis of larger units than speech sounds. Findings from an investigation of acoustic and perceptual correlates of various phonetic distinctions may be useful in understanding many issues, such as the development of speech perception, second language learning or processing models of speech perception.

Acknowledgments
The paper was supported by the János Bolyai Research Scholarship and by the Postdoctoral Fellowship Program of the Hungarian Academy of Sciences. This paper has benefited considerably from contributions of Mária Gósy, András Beke and Anita Auszmann. The author thanks reviewers for their suggestions and constructive feedback.

References


PRONUNCIATION NORMS AND PRONUNCIATION HABITS OF ORTHOGRAPHIC <Ä, ÄH> IN STANDARD AUSTRIAN GERMAN

Bettina Hobel¹, Sylvia Moosmüller², Christian Kaseß³
¹,²,³Acoustics Research Institute, Austrian Academy of Sciences, Vienna, Austria
e-mail: ¹bettina.hobel@oeaw.ac.at, ²sylvia.moosmueller@oeaw.ac.at, ³christian.kasess@oeaw.ac.at

Abstract
The aim of this study is to examine the pronunciation of orthographic <ä, äh> in Standard Austrian German. Various factors, such as the historical development of e-vowels, various prescriptions for pronunciation, regional differences in pronunciation standards, the questionable phonological status of /ɛː/, as well as the influences of the situation’s formality, have an impact on the phonetic realisation. In general, the expected [eː]-pronunciation for <ä, äh> has been confirmed. Only the group of older males exhibited a higher F₁ for <ä, äh> than for <e, ee, eh>. An auditory analysis revealed that in read speech, sporadically [ɛː] is used for <ä, äh>, whereas in spontaneous speech, only [eː] occurs. These occasional exceptions can be explained by old prescriptive norms, which still exist in some speakers’ minds and can be activated in reading tasks due to the influence of orthography and the formality of the situation.

Keywords: sociophonetics, Standard Austrian German, pronunciation norms, long front mid vowels

1 Introduction
The question of [ɛː] can be seen from a variety of perspectives. Each perspective sheds its own light on the topic adding some explanations for this controversial sound and its questionable phonemic status. In the following section, a closer look will be taken on each of the four perspectives, in order to better understand the ambiguities about [ɛː].

For a start, the historical perspective further splits into two different points of view. First, the actual development of different sounds from Proto-Indo-European to today’s sounds are presented¹. Second, in the course of time, different attempts for

¹ We are well aware of the fact that our presentation of the historical development of the e-vowels does by far not mirror the discussion of historical linguistics or the debate about the reconstructed sounds themselves. Yet, a full account of the historical development would go beyond the scope of our endeavor to discuss the pronunciation of orthographic <ä, äh> in Standard Austrian German. However, by touching on this topic, we wanted to point out that the development of orthographic <ä, äh> is by no means straightforward.
standardisation of pronunciation have been made, which interfered with the respective pronunciation.

The second perspective is the regional one: different conventions about the standard pronunciation exist in the different German-speaking countries. Therefore, both the norms of the respective standard languages and the diverse dialectal pronunciations have to be considered.

The phonological perspective raises the question of a possible phonemic status for \([e:]\) and provides arguments for and against \([e:]\) being a phoneme.

Finally, a short summary on the situational influence, the formal context, and the influence of orthography is made.

### 1.1 Historical perspective

#### 1.1.1 Development of e-sounds from Proto-Indo-European to New High German

The development of the different e-sounds from Proto-Indo-European to New High German is rather complex, lacking a straightforward development over time. In this section, we aim to describe which e-sounds existed at certain stages of the German language rather than tracing the exact development of these sounds. The written letters for e-vowels are presented shortly; for details on conventions and regulative control of orthography, see section 1.1.2. In our description, we concentrate on the most important developmental stages, which are also presented in Table 1.

In the reconstruction of Proto-Indo-European (PIE), only two e-sounds are assumed: short \(*e*\) and long \(*\ddot{e}\). Additionally, \(*a\), connected to ablaut, was assumed. Later on, this vowel changed to *a* in Proto-Germanic (Szulc, 2002: 41, 48-51; von Kienle, 1960: 18).

During the development of PIE to Proto-Germanic (PG), stress changed from a free expiratory accent to an accent on the first syllable of the stem (Karstien, 1939: 46). Long PIE \(*\ddot{e}\) was changed to \(*\dddot{e}\) in Proto-Germanic, pronounced \([\dddot{a}:]\) at those times and then undergoing different regional developments, e.g. to \([\dddot{a}:]\) in Old High German. In PG, a long \(*\dddot{e}\) is assumed as well, but its origin is not absolutely clear, probably from PIE \(*\dddot{e}\) or \(*e + \) dark vowel; later on, it developed into a diphthong. Therefore, it is not relevant for the current investigation. As concerns the short e-sounds, PIE *e* remained *e* in PG (in some cases, it changed to *i*) (Szulc, 2002: 57f., 60f., 64; von Kienle, 1960: 18f.).

The most important change in the vowel system of Old High German (OHG, 750 to 1050 A.D.) was due to the *i*-umlaut. Proto-Germanic short \(*a\) changed to a closed e-vowel in the typical *i*-umlaut condition, with the presence of \(*i, \dddot{i}, j\) in the subsequent syllable (primary umlaut, “Primärumlaut”); this umlaut was already represented in

---

2 In the introductory part, the transcription follows the convention of using italics for sounds of earlier stages of German, although Szulc (2002) uses slashes. For reconstructed Proto-Indo-European and Proto-Germanic sounds, the asterisk \(*\) is used.
orthography at that time. In a second step, a further change termed ‘secondary umlaut’ ("Sekundärumlaut") took place. PG short \*a changed to an open e-vowel (in Central German, a very open e-vowel in Alemannic and Bavarian) when \*i ï, j were preceded by one of the following consonant clusters: \*ht, \*hs, or consonant + \*w or when \*i ï, j was found in the syllable after the next syllable. The long and back vowels were also affected by the secondary umlaut. This change was not orthographically represented until early Middle High German (von Kienle, 1960: 26-28; Russ, 1978: 56f.). In addition, OHG featured a short open ɛ (written <ë>, von Kienle, 1960: 22) and a long ë from Proto-Germanic \*ei which evolved from [ɛː] to [eː] during the development of OHG (von Kienle, 1960: 24).

Only in Classical Middle High German (CMHG, 1170-1250 A.D.)\(^3\), the allophones caused by the i-umlaut received phonemic status after the triggering i ï, j had been dropped in unstressed final syllables (Donegan, 1993: 120; Russ, 1978: 64). Therefore, those e-sounds which stem from Proto-Germanic \*e were written <ë> (pronounced [ɛ]) and those which were a consequence of the i-umlaut, were represented by the following graphemes: <e>, <ë> or <ë>, pronounced [e]. The vowel quality [æ], the result of the secondary umlaut a to å, was now written as <å> or <ä>. Consequently, in CMHG, there were three short e-sounds, e <ë>, e <ë>, e and å <å, å>. Two pronunciations for long vowels are attested: æ <æ> going back to the umlaut of OHG å, being pronounced as [æ] and ë <ë> going back to OHG ë, pronounced [e:] (Russ, 1978: 73; Szulc, 2002: 93, 129, 133; von Kienle, 1960: 42, 45). In Early New High German (ENHG, from 1250-1520 (Szulc, 2002: 143) or 1350-1650 (Russ, 1978: 8)), a complicated development concerning the e-vowels took place when the CMHG syllable structure changed and led to a shortening and lengthening of vowels (Szulc, 2002: 150-153). From now on, a clear distinction between spoken dialects and written language had to be made, since the one-to-one relationship between pronunciation and orthography was lost. First, regarding the spoken dialects, the three short e-sounds from CMHG collapsed into two sounds; in most cases, å and e merged into æ; and e was maintained. The long æ from CMHG, tracing back to the umlaut of OHG å, varied developmentally in different regions, ranging from [aː] to [æː] and [æː] or even diphthongs, such as [eɪ] and [aɪ] or [æa]. The other long CMHG e-vowel, ë <ë>, mostly remained long [eː]. However, in hardly any dialect did the two long e-sounds collapse (Schirmunski, 1962: 213, 236, 241; Szulc, 2002: 153, 155f.). In written language, less variation is observed, as most e-phonemes are written as <e>. Occasionally, in Upper German\(^4\), <å> was used for CMHG å. For long CMHG

---

\(^3\) The literary language of MHG is usually referred to as Classical MHG. Other sources date its period from 1050 to 1350, thus including what Szulc named Late Old High German. As language evolves continuously, such boundaries are never clear-cut and open for discussion (Russ, 1978: 6f.).

\(^4\) The English terminology for German *Hochdeutsch* in the dialectological / geographical sense is not unanimous. Upper German is used e.g. by Russ (1978).
$\ddot{e} <\ddot{e}>$, $e$ was used. Although CMHG $\dddot{a} <\dddot{a}>$ was typically represented as $e$, it was sometimes also represented by $\dddot{a}$. In the late 15th century, the rule of writing $\dddot{a}=\dddot{a}$ was established, if the vowel traced back to $\dddot{a}$; for details, see section 1.1.2. (Russ, 1978: 73; Szulc, 2002: 156).

**Table 1. Overview of the development of $e$-vowels from Proto-Indo-European to New High German**

<table>
<thead>
<tr>
<th>Time span</th>
<th>Short $e$-vowels</th>
<th>Long $e$-vowels</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Phoneme</td>
<td>Pronunciation</td>
</tr>
<tr>
<td>Proto-Indo-European (PIE)</td>
<td>~3000 B.C.</td>
<td>$e$</td>
</tr>
<tr>
<td>Proto-Germanic (PG)</td>
<td>~1000 B.C. – ~50 B.C.</td>
<td>$e$</td>
</tr>
<tr>
<td>Old High German (OHG)</td>
<td>1170-1250</td>
<td>$e$</td>
</tr>
<tr>
<td>Classical Middle High German (CMHG)</td>
<td>1170-1250</td>
<td>$e$</td>
</tr>
<tr>
<td>Early New High German (ENHG)</td>
<td>1520</td>
<td>$e$</td>
</tr>
<tr>
<td>New High German (NHG)</td>
<td>since 1520</td>
<td>$e$</td>
</tr>
</tbody>
</table>

The development from CMHG $e$-vowels to today’s New High German (starting around 1520) vowels of the standard pronunciation is complex, as not only quality, but also quantity changed. In principle, a development of almost all CMHG $e$-vowels to the three NHG vowels $e$, $\ddot{e}$, and $\ddot{a}$ is conceivable. However, the three different CMHG short $e$-vowels were reduced to just one short $e$-vowel $e$ in standard pronunciation. Yet, there are still two long $e$-vowels, $[e]$ and $[\ddot{a}]$ – at least in the prescribed standard pronunciation –, but the latter is only pronounced for graphemic $<\ddot{a}>$, cf. Siebs (1957). In the dialects, the situation is even more complex, since in some regional varieties, more variants of $e$-vowels exist which are used differently from the standard pronunciation (Russ, 1978: 73-75; Szulc, 2002: 233–237).

5 The digraph $<\ddot{a}>$ developed later to the monograph $<\ddot{a}>$ (Sanders, 1972: 42).
6 Here, Late Old High German is included into OHG, not into CMHG (compare Footnote 3), because the phoneme system is almost the same, whereas it changed in CMHG.
1.1.2 Prescriptive regulations on pronunciation and spelling

Regulations about orthography started not long after the increasing use of writing in the times of Old High German. In monasteries, so-called “written dialects” (“Schriftdialekte”) evolved which resulted from converging several dialects in contact. That means, scriptures in the local dialects were abandoned in favour of a more unified writing system. Thus, the one-to-one representation of local dialect pronunciation and orthography was dismissed (Szulc, 2002: 89). In the course of the following centuries, an increase of written interchange took place, leading to a convergence of the different orthographies for a better mutual understanding.

An innovation – and a first regulation – was performed in the late 15th century, when grammarians established the “etymological principle”. The letter <aͤ> should be used in paradigms for ä, æ, e and ē when CMHG <a> alternates with e-sounds due to primary or secondary umlaut. However, due to a lack of etymological knowledge, not all words were correctly transformed according to the new rule, leading to inconsistencies. Examples are eltern < alt (“parents” < “old”) or behende < hand (“nimble” < “hand”), which were not written with the letter <aͤ>. In other lexemes, <ä> was written without having a cognate with <a>, but stems from CMHG e, e.g. Bär “bear”, gebären “give birth” (Siebs, 1912: 39f.; Szulc, 2002: 156f., 197). During the time of Early New High German, spelling pronunciation became popular. Thus, the pronunciation followed the graphemic representation of the two sounds, resulting in what was dubbed an “artificial phoneme” /ɛː/ by von Polenz (2000: 151).

From the early 16th century onwards, efforts were made to unify German orthography. In this context, Martin Luther’s Bible translations and the East Central German of the Saxon Chancery made a significant contribution. With respect to our research question, all CMHG e-sounds coincided in written <e> in Luther’s scriptures; <ä> was only rarely used for very open e-sounds or for a reference to the i-umlaut of <a>. In East Central German, e was written <ä> or <ä> when going back to the primary umlaut, but only, as mentioned earlier, when the connection to <a> was recognised (Szulc 2002: 185–197). Regarding pronunciation: around 1700, a standard pronunciation (“Oralisierungs-norm”) evolved by means of a small well-educated group of Central German speakers who tried to talk the newly established written variety. They strongly adhered to a spelling pronunciation based on their own dialectal competence. Later on, these conventions were spread by churches and then by schools, where reading lessons were used for practising pronunciation (Schmidt, 2005: 284f.).

However, it was only in 1902, at the 2nd orthographic conference in Berlin, that a common orthography, mandated for the German Reich, Austria, and Switzerland, was brought into use (Reuter, 2005). A few years preceding the orthographic conference, a meeting on the standardisation of pronunciation for the stages was held in Berlin, convened by Theodor Siebs. The results of the conference were published in 1898 and, in the subsequent editions, became the basis for a unified pronunciation for
German speaking countries (Moosmüller, 2015). According to “Siebs” (1912: 41f., 44) long lax [ɛː] should be pronounced when one of the following criteria is met:

1. <ä>, e.g. wählen “to choose”
2. <ä> in open syllables, e.g. Träne “tear”
3. <ä> in closed syllables with a single consonant, e.g. spät “late”
4. in foreign words, as, e.g. Ära “era”, Sphäre “sphere”, Rabelais, Portière “portiere”.

Luick, the Austrian representative at the 1898 conference in Berlin, was dissatisfied with the results of the conference, since Austrian pronunciation characteristics were disregarded. In 1904, he published his “Deutsche Lautlehre”; however, with regards to the pronunciation of long <ä, äh>, he gave no clear recommendation. More than 100 years ago, he already summed up the dilemma that is still valid today: “Keinesfalls besteht ein fester Brauch, nicht einmal die Übereinstimmung in der Vorstellung von dem, was als besser anzusehen ist” (1904: 75f.). He cautiously predicted a tendency for an [eː]-pronunciation in all situational contexts except in the pronunciation on stage where [ɛː] dominates. Similarly, Braune (1904: 29) reached no conclusion either, but only stated ironically that everyone should pronounce the e-vowels as they wished.

However, Luick’s orthoepy, which achieved three editions, was neither included in the educational system in Austria nor was it considered in the training of news anchors. Rather, the many editions of Siebs served as a model for pronunciation in Austria. Consequently, in 1957, the editors of the 16th edition (Siebs, 1957) decided to produce a four-page document, authored by Austrian scholars, such as E. Kranzmayer, W. Steinhauser, F. Trojan and others, the so-called “Austrian Supplement to Siebs ‘Deutsche Hochsprache – Bühnenaussprache’ which, regarding our research question, contained the prescription that long <ä, äh> should be pronounced as a long, open vowel. Wächter-Kollpacher (1996) notes that trainers instructing news anchors draw on various materials; concerning the pronunciation of long <ä, äh>, she states that the vowel should not be pronounced too openly (1996: 280) – obviously a compromise solution, which results in a speaker-specific realisation of orthographic <ä, äh> in the pronunciation of news anchors (Wiesinger 2009: 238). More recently, news anchors draw on a pronunciation database and on the various editions of the Duden Aussprachewörterbuch (Soukup & Moosmüller, 2011). Yet, a general inconsistency in the speech of news anchors is observed (Hollmach, 2007: 106). With respect to the pronunciation of long <ä, äh>, news anchors pursue their own pronunciation preferences (Patocka, 1988; Wiesinger, 2009). For this

---

7 Unfortunately, the first edition is not available in any Austrian library. Therefore, we cite the 10th edition, which is publicly accessible.
8 “There is certainly no fixed convention, not even an agreement on the idea of what is considered the better [pronunciation]” (translation by authors).
reason, it is not advisable to draw on the pronunciation of news anchors as a
determiner of Standard Austrian German (Moosmüller, 2015), as was done by Muhr
(2007) in the “Österreichische Aussprachewörterbuch”\textsuperscript{9}. Apart from news anchors’
speech, recommendations for pronunciation standards in Austria, in a recently
published pronunciation dictionary (Krech, Stock, Hirschfeld & Anders, 2009), were
elaborated on by Wiesinger (2009), who, on the basis of empirical investigations,
proposes to pronounce [eː] for <ä, äh>, except <ä, äh> preceding <r> (Wiesinger,

1.2 Regional perspective
As elaborated in 1.1.1, the historical development of the respective e-vowels
rendered different regional outputs. In the Austrian context, it is important to mention
that a distinction of /eː/ and /ɛː/ was preserved in the Alemannic dialects, thus in the
province of Vorarlberg, the distinction is realised. For long <ä, äh>, the pronunciation
[eː] prevails in the other parts of Austria, which belong to the Bavarian dialect group
(Kleiner, 2011ff.). Thus, in the Austrian dialects, the situation concerning the
pronunciation of <ä, äh> is less complicated than in Germany, where a merger in
favour of [eː] is mainly found in the northern parts of the country\textsuperscript{10}. While in the
western and southern parts, [eː] is pronounced for <ä, äh>, in South-West Germany, a
merger is found in favour of [ɛː] (König, 1989a: 45; König, 1989b: 112f.).

Since the Austrian Bavarian dialects belong to the non-distinguishing area, and
since Standard Austrian German is largely a compromise between Middle Bavarian
varieties and German pronunciation norms as put down in Siebs (Moosmüller, 2015),
speakers of Standard Austrian German might, at times, remember the prescriptions
learnt in school and pronounce [eː] for orthographic <ä, äh>. This happens in very
formal situations, such as reading word lists, as shown in an investigation performed
by Iivonen (1987: 297, 325) for educated speakers from Vienna. Nevertheless, even for
this very formal speaking task, female subjects tended to merge the two vowels in favour
of [ɛː], while men more readily made the distinction. However, in the analysis presented
in Moosmüller (2007), no distinction was made between [eː] and [ɛː]. Without
exception, all of her subjects, both male and female, realised [eː] for orthographic <ä,
äh>. A similar result was obtained by Ehrlich (2009: 82): in her data, she could
observe hardly any realisations of [ɛː].

Recently, Sloos (2013) proposed an unmerger of [eː] and [ɛː] in Standard Austrian
German. However, the study has too many drawbacks to reliably support this
hypothesis. First, speakers of all dialect regions, including Vorarlberg, were lumped

\textsuperscript{9} Also for other reasons, Muhr’s Aussprachewörterbuch has been subjected to serious
critique (Hirschfeld, 2008; Kleiner, 2008).

\textsuperscript{10} The prescriptions put down in Siebs (1898 and subsequent editions) did not follow the
Low German phonetic qualities (“Lautwerte”) in the case of <ä, äh>, but claimed to
distinguish /eː/ and /ɛː/.
together. Second, no differentiation between speakers who were raised in the dialect and speakers who were raised in a standard variety was made. Third, the phonetic context was disregarded: tense vowels preceding /r/ become lax in regions which feature r-vocalisation (Moosmüller, Schmid & Brandstätter, 2015: 342; Wiesinger, 2009: 237). Paradoxically, her example of the pair Beeren, “berries”, and Bären, “bears” refutes her own hypothesis. According to this hypothesis, Bären should be pronounced with [ɛː] (merger) and is – due to an unmerger – now pronounced with [eː]. However, given the above rule that any vowel preceding /r/ has to be pronounced lax in Standard Austrian German, neither Beeren nor Bären have ever been pronounced with [eː]; the pronunciation of both Beeren and Bären is [ˈb̥ɛɐ̯rɛn] or, alternatively, [ˈb̥ɛɐ̯n̩]. As a consequence, Beeren and Bären are homophonous. Mainly for these reasons, her study cannot serve as a basis for a change or even an unmerger in Austrian varieties. Additionally, an unmerger presupposes a state in which the vowels were distinguished and subsequently have been merged (Russ, 1978: 25). This state never existed in the Bavarian varieties, since the secondary umlaut evolved to a resp. ā (Schirmunski, 1962: 243; von Kienle, 1960: 45).

1.3 Phonological perspective

Numerous scholars have discussed the doubtful phonemic status of /ɛː/. Scholars in the structuralist tradition complained the loss of symmetry (Mangold, 1966: 39). Moulton (1962: 68) concluded that /ɛː/ “is not well integrated into the German vowel system”; or as Becker (2012: 37) put it, it is “eine Störung des Lautsystems”11. As a solution, Becker (2012: 38) suggested: “Man kann somit getrost [ɛː] als tiefen Vokal auffassen; obendrein verhält er sich wie ein tiefer Vokal: Er ist ungespannt wie a, hat keinen runden Partner wie a und unterscheidet sich von dem kurzen Partner in erster Linie durch Dauer (…) Die Verwirrung um [ɛː] lässt sich vermeiden, indem man dafür besser [æː] schreibt.”12 This argument is, however, far-fetched, since [ɛː] cannot simply be replaced by [æː] in order to meet a certain phonological theory. Wiese (1996: 11, 17) assumed /ɛː/ but put it into brackets. No comment is given on that, except that /ɛː/ and /ɛː/ preceding /r/ are neutralised. However, neutralisation is no explanation for setting brackets. Kohler (1995: 172f.) mentions a merger in Northern Germany, but, generally, assumes phonemic /ɛː/.

Mangold (1966: 37f.) argued that a strong influence of orthography on pronunciation is normal, hence, spelling pronunciation in which the two long e-vowels are distinguished is nothing artificial. Yet, most scholars agree that /ɛː/, if assumed, is rather artificial, i.e., an “artificial phoneme based on a grapheme” (von Polenz, 2000: 151), “an

11 “a disturbance of the phonemic system” (translation by authors).

12 “[ɛː] can safely be understood as an open vowel; additionally, it acts in the same way as an open vowel: it is lax as a, it has no rounded counterpart in the same way as a, and it differs mainly in duration from the short counterpart (…) The confusion about [ɛː] can be avoided if [æː] is assumed” (translation by authors).
artificial phoneme of recent invention, without a phonological history, based only on the spelling system” (Moulton, 1962: 69), or a “ghost phoneme” (Sanders, 1972). Szulc (2002: 211) proposes to view /ɛː/ as a “graphogeneous” phoneme.

One argument often brought forward in favour of a phonemic status of /ɛː/ concerns the existence of minimal pairs, such as, e.g., ich gebe “I give” – ich gäbe “I would give” or Seele “soul” – Säle “hall, pl.”. However, this argument is based on the graphemic representation, thus, a non-valid argument for a phonological representation.

Generally, for Standard Austrian German, phonemic /ɛː/ is not assumed (Moosmüller, Schmid & Brandstätter, 2015). Muhr (2007: 42) mentions a possible phonemic distinction due to minimal pairs in the written language, but in spontaneous speech, it has merged with [ɛ]. Wiesinger (2009: 237f.) based his recommendation for Standard Austrian German on an investigation of trained speakers and educated untrained speakers. In his data, he found both [ɛː] and [ɛː] for the graphemes <ä, äh>. According to his results, a speaker-specific use of both [ɛː] and [ɛː] is employed by trained speakers in official language use. Preceding /ʀ/, [ɛː] is used by all speakers. Educated, but untrained speakers from Vienna, Burgenland, and Upper and Lower Austria occasionally make use of [ɛː]. The use of [ɛː] is, as a matter of fact, a trait of the Alemannic region. Consequently, [ɛː] is always employed by speakers from Vorarlberg. However, he did not discuss the possible phonemic status of /ɛː/.

To our knowledge, Iivonen (1987: 297, 332) is the only one to assume /ɛː/ in Standard Austrian German. As stated above, female subjects in his study hardly ever distinguished [ɛː] and [ɛː], whereas male subjects showed a slight tendency for a distinction. Consequently, Iivonen (1987: 332) also questioned the phonemic status of /ɛː/. Since Iivonen’s study was based on a highly formal speaking task (reading isolated words from cards), the speaking task and the speaking situation have to be taken into account in the discussion on a possible assumption of the phonemic status of /ɛː/.

1.4 Situational and individual influence

As mentioned above, there is no doubt that the situation and its formality have a strong impact on the pronunciation of long <ä, äh>. Two related aspects deserve special attention: the degree of formality of the speech situation and the various gradations in reading tasks (minimal pairs vs. continuous text). As stated above, Mangold (1966: 38, 41) emphasised the strong influence of written language on spoken language, meaning that pronunciation is not only influenced in reading, but also in spontaneous speech, especially in formal speech situations. Thus, /ɛː/ occurs more often in formal than in informal situations, as well as in “humourous mock-formal situations” (Moulton, 1962: 69). Conversely, in informal situations, [ɛː] prevails (Wängler, 1981: 44).

In reading, the influence of orthography might lead to so-called spelling pronunciations, which might even produce hypercorrect forms. The more unnatural

---

13 The trained speakers followed mainly the norms set up in Siebs.
the reading task, e.g. reading word lists or even minimal pairs, the higher the awareness of certain peculiarities, such as the grapheme <ä, äh> and the more likely the occurrence of spelling pronunciations (Kleiner 2011ff.). König (1989a: 45) contrasted the realisation of <ä, äh> in a reading task of word lists and of minimal pairs. The results showed that the more formal the task, the more likely the distinction between <ä, äh> and <e, ee, eh> was preserved in most regions of Germany (except for the very northern parts). Conversely, a higher likelihood of <ä, äh> being pronounced as [eː] in less formal situations was observed. These results correspond to the findings in Stearns and Voge (1979: 141, cited in König, 1989a: 45).

In addition, a speaker-specific use of /ɛː/ was observed. Some “careful” speakers try to maintain the distinction of /eː/ – /ɛː/, because they consider upholding the distinction as the most refined pronunciation (Pilch, 1966: 257). Moulton (1962: 68f.) assumed that the use of /ɛː/ could also depend on a speaker’s education. He states that “probably all educated speakers use long /eː/ = [ɛː] as the name of the letter ä” as well as for the distinction of minimal pairs, such as gäben “would give” – geben “to give”. Regarding Austria, the “Österreichische Wörterbuch” (1999: 24) allows a distinction in very formal styles or in minimal pairs.

1.5 Aim of this study

Considering all the above-mentioned factors, especially the historically different developments, the influence of orthography, the formality of the situation, and the variation in regional standard varieties, it is our intention to shed some light on this ongoing issue. In addition, the questionable results presented in Sloos (2013) call for a review of the pronunciation habits in Standard Austrian German. Thus, the following research questions developed from the issues discussed in our introduction:

A. How is long <ä, äh> pronounced in Standard Austrian German?
B. Is there a difference in the pronunciation of <ä, äh> in read speech and in spontaneous speech, due to the formality of the speaking task and the influence of orthography?
C. Are there differences in the pronunciation of <ä, äh> due to gender or age?
D. Can we observe an unmerger of [eː] and [ɛː] in Standard Austrian German, as proposed by Sloos (2013)?

2 Method

2.1 Subjects

A total of 24 speakers of Standard Austrian German (SAG), as defined in Moosmüller (1991), were analysed for this study. SAG is spoken by people with a higher educational background: the speakers themselves, as well as at least one parent, held either an academic degree or a school leaving examination (“Matura”). All speakers

---

14 In the phonological representation, Moulton did not mark length, but noted that phonetically, the long vowel [ɛː], equivalent to IPA [ɛː], is meant.
were born and raised in Vienna with at least one parent having been raised in Vienna as well. We chose subjects from Vienna, because the standard variety as spoken in Vienna holds the highest prestige (Goldgruber, 2011; Moosmüller, 1991; Soukup, 2009). In order to capture possible changes in pronunciation habits, subjects were split into two age groups (20-25 years and 45-63 years). Thus, the corpus comprised four groups of six subjects each, balanced for age and gender. None of the speakers completed a professional speaker training.

2.2 Recordings and material
Since the degree of formality of the speaking task and the orthographic representation seem to have the greatest influence on the pronunciation of <ä, äh> (see section 1.4.), data of semi-spontaneous interviews, dialogue tasks, and reading material were used for further analysis.

Each of the 24 subjects participated in two recording sessions. The first meeting was a face-to-face recording session with the interviewer. The recordings were obtained using a semi-structured interview (approx. 25 minutes) and several reading tasks. The task of reading sentences was performed twice. The second meeting was a few weeks later, where two subjects were recorded in separate booths, connected via headphones and microphones, while carrying out several dialogue tasks.

In this study, we were interested in disyllabic words with stress on the first syllable containing the vowels graphemically represented as <ä, äh> and <e, ee, eh>. We focused on long vowels, since in short vowels, no distinction in the vowel qualities of <e> and <ä> is produced, both being pronounced as [ɛ] as exemplified in the homophonous words Fälle “case, pl.” – Felle “fur, pl.” (Iivonen, 1987: 297; Mangold, 1966: 39; Siebs, 1957: 41f). No target vowels were followed by /ʀ/, since vowels preceding /ʀ/ become lax (Moosmüller, Schmid & Brandstätter, 2015: 342; Wiesinger, 2009: 237). Because of the coarticulatory influences of nasalisation, vowels preceding nasals were excluded as well. The words of the reading task which contained orthographic <ä, äh> were subjected to a word frequency analysis.

2.3 Data extraction and analysis
Each word following the above mentioned requirements and containing orthographic <ä, äh> or <e, ee, eh> was considered for analysis. Segmentation, annotation, and acoustic measurements were carried out with STx15. The first three formant frequency trajectories (F₁, F₂, and F₃) have been extracted via Linear Prediction Coding (window length: 46ms long Kaiser-Bessel window with an overlap of 95%). Segmentation was performed from the onset to the offset of the vowel. Duration measurements were performed simultaneously. The formants calculated were subjected to a plausibility check, meaning that a combination of phonetic knowledge and auditory impression formed the basis for deciding whether a specific formant needed correction or not. Extracted formants of every single vowel went through this plausibility check. STx

15 http://www.kfs.oeaw.ac.at (Noll, White, Balazs & Deutsch, 2007)
provides a number of facilities for parameter correction. For read speech, a total of \( n = 432 \) <ä, äh> and \( n = 382 \) <e, ee, eh> productions were subjected to analysis; for spontaneous speech, \( n = 157 \) for <ä, äh> and \( n = 285 \) for <e, ee, eh>.

Acoustically, [ɛː] und [eː] are differentiated by a higher \( F_1 \) for [ɛː] and a lower \( F_2 \) and \( F_3 \) for [eː] (Iivonen, 1987). Therefore, a quantitative approach was employed to find out whether significant differences in formant values exist. For the statistical analysis, linear mixed-effects models were fitted to the data of each formant separately with “subject” and “word” as random factors, and “age”, “gender”, “task”, and “vowel quality” as fixed factors. Linear mixed-effects modelling was carried out in R (version 3.1.1; R Core Team 2014) using the *lmer*-function of the lme4-toolbox (Bates, Maechler, Bolker & Walker, 2015). Significance of the factors was determined using the Type-III ANOVA of the lmerTest-toolbox (Kuznetsova, Brockhoff & Bojesen Christensen, 2015). The denominator degrees of freedom were derived using the Satterthwaite approximation as implemented in lmerTest. In case of significant effects, pairwise Tukey post-hoc tests were performed on all possible combinations with regard to the significant effect using the lsmeans-toolbox (Russel, 2015). The fourfold interaction of age*gender*task*vowel was not modelled for any formant. In addition, a qualitative analysis was performed for some exemplary cases where an unexpected [eː] occurred.

3 Hypotheses

Based on the literature, the following hypotheses were formulated:

Hypothesis 1: No differences between long <ä, äh> and <e, ee, eh> exist in pronunciation; both graphemic representations are realised as [eː].

Hypothesis 1a: In certain instances, speakers adhere to orthography in read speech, due to the formality of the task, thus realising [ɛː].

Hypothesis 1b: In spontaneous speech, no differences exist in the realisation of graphemic <ä, äh> and <e, ee, eh> in spontaneous speech.

Hypothesis 1c: No age-specific differences exist.

Hypothesis 1d: No gender-specific differences exist in the pronunciation of <ä, äh> and <e, ee, eh>.

4 Results

4.1 Quantitative analysis

For better clarity, the quantitative analysis will be presented separately for each formant. For an overview, Table 2 presents the mean formant frequencies of \( F_1 \), \( F_2 \), and \( F_3 \) of both vowels, according to the factors speaking task (spontaneous / read), gender (male / female), and age (< 25 / > 45).
Table 2. Mean and Standard deviation of F$_1$-F$_3$.

<table>
<thead>
<tr>
<th>Vowel</th>
<th>Task</th>
<th>Age</th>
<th>Gender</th>
<th>F$_1$ [Hz] mean</th>
<th>F$_1$ [Hz] sd</th>
<th>F$_2$ [Hz] mean</th>
<th>F$_2$ [Hz] sd</th>
<th>F$_3$ [Hz] mean</th>
<th>F$_3$ [Hz] sd</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;ä, äh&gt;</td>
<td>spont</td>
<td>&lt;25</td>
<td>m</td>
<td>334</td>
<td>36</td>
<td>1954</td>
<td>209</td>
<td>2622</td>
<td>171</td>
</tr>
<tr>
<td>&lt;e, ee, eh&gt;</td>
<td>spont</td>
<td>&lt;25</td>
<td>m</td>
<td>321</td>
<td>59</td>
<td>1859</td>
<td>196</td>
<td>2527</td>
<td>228</td>
</tr>
<tr>
<td>&lt;ä, äh&gt;</td>
<td>spont</td>
<td>&gt;45</td>
<td>m</td>
<td>324</td>
<td>57</td>
<td>2097</td>
<td>165</td>
<td>2593</td>
<td>169</td>
</tr>
<tr>
<td>&lt;e, ee, eh&gt;</td>
<td>spont</td>
<td>&gt;45</td>
<td>m</td>
<td>298</td>
<td>66</td>
<td>2055</td>
<td>193</td>
<td>2634</td>
<td>243</td>
</tr>
<tr>
<td>&lt;ä, äh&gt;</td>
<td>read</td>
<td>&lt;25</td>
<td>m</td>
<td>320</td>
<td>67</td>
<td>1983</td>
<td>196</td>
<td>2719</td>
<td>217</td>
</tr>
<tr>
<td>&lt;e, ee, eh&gt;</td>
<td>read</td>
<td>&lt;25</td>
<td>m</td>
<td>296</td>
<td>42</td>
<td>2011</td>
<td>159</td>
<td>2739</td>
<td>208</td>
</tr>
<tr>
<td>&lt;ä, äh&gt;</td>
<td>read</td>
<td>&gt;45</td>
<td>m</td>
<td>366</td>
<td>85</td>
<td>2083</td>
<td>214</td>
<td>2653</td>
<td>205</td>
</tr>
<tr>
<td>&lt;e, ee, eh&gt;</td>
<td>read</td>
<td>&gt;45</td>
<td>m</td>
<td>304</td>
<td>59</td>
<td>2154</td>
<td>223</td>
<td>2747</td>
<td>241</td>
</tr>
<tr>
<td>&lt;ä, äh&gt;</td>
<td>spont</td>
<td>&lt;25</td>
<td>f</td>
<td>426</td>
<td>84</td>
<td>2466</td>
<td>207</td>
<td>3163</td>
<td>336</td>
</tr>
<tr>
<td>&lt;e, ee, eh&gt;</td>
<td>spont</td>
<td>&lt;25</td>
<td>f</td>
<td>453</td>
<td>68</td>
<td>2353</td>
<td>187</td>
<td>3039</td>
<td>220</td>
</tr>
<tr>
<td>&lt;ä, äh&gt;</td>
<td>spont</td>
<td>&gt;45</td>
<td>f</td>
<td>343</td>
<td>77</td>
<td>2344</td>
<td>182</td>
<td>2928</td>
<td>221</td>
</tr>
<tr>
<td>&lt;e, ee, eh&gt;</td>
<td>spont</td>
<td>&gt;45</td>
<td>f</td>
<td>388</td>
<td>58</td>
<td>2318</td>
<td>161</td>
<td>2904</td>
<td>197</td>
</tr>
<tr>
<td>&lt;ä, äh&gt;</td>
<td>read</td>
<td>&lt;25</td>
<td>f</td>
<td>438</td>
<td>82</td>
<td>2480</td>
<td>204</td>
<td>3186</td>
<td>226</td>
</tr>
<tr>
<td>&lt;e, ee, eh&gt;</td>
<td>read</td>
<td>&lt;25</td>
<td>f</td>
<td>429</td>
<td>60</td>
<td>2483</td>
<td>176</td>
<td>3202</td>
<td>208</td>
</tr>
<tr>
<td>&lt;ä, äh&gt;</td>
<td>read</td>
<td>&gt;45</td>
<td>f</td>
<td>400</td>
<td>57</td>
<td>2486</td>
<td>230</td>
<td>3081</td>
<td>235</td>
</tr>
<tr>
<td>&lt;e, ee, eh&gt;</td>
<td>read</td>
<td>&gt;45</td>
<td>f</td>
<td>399</td>
<td>49</td>
<td>2455</td>
<td>236</td>
<td>3048</td>
<td>220</td>
</tr>
</tbody>
</table>

4.1.1 Results on F1

For the first formant, the threefold interaction of age*gender*vowel was statistically significant ($p = 0.004$). A post-hoc Tukey test revealed that older men distinguish <ä, äh> and <e, ee, eh> ($p = 0.008$), independently of the task, showing a higher formant frequency for [ɛː] than for [eː]. All other statistically significant contrasts included a difference between both genders which is a consequence of the significant gender effect ($p < 0.0001$). Table 3 shows all results for the ANOVA with Satterthwaite approximation for degrees of freedom on F$_1$.

Additionally, the threefold interaction of age*task*vowel was statistically significant ($p = 0.005$). The post-hoc Tukey test showed that four contrasts are significant. Spontaneously spoken <e, ee, eh> is different from read <ä, äh> in the older age group ($p = 0.017$, read > spontaneous). Furthermore, in the older age group, <ä, äh> in reading differs from <ä, äh> in spontaneous speech ($p = 0.0002$, read > spontaneous) and pronunciation of <ä, äh> in spontaneous speech differs between younger and older speakers ($p = 0.011$, older > younger). Besides these three significant effects, the pronunciation of <ä, äh> is different in the reading task of younger speakers from the spontaneous speech of older speakers ($p = 0.043$, younger > older). However, this contrast does not lend itself to a meaningful interpretation, as in both cases, [eː]-pronunciations are prevailing.

The twofold interactions of age*task ($p < 0.0001$), task*vowel ($p = 0.036$), age*gender ($p = 0.030$), gender*vowel ($p < 0.0001$) and task*gender ($p = 0.027$) reached significance. However, except for task*gender, all interactions are included in the two statistically significant threefold interactions and thus, they are not further analysed. The interaction of task*gender manifests itself as a minor difference of about 10 Hz between males and females in the (non-significant) task effect.
Table 3: ANOVA table for $F_1$ using the Satterthwaite approximation for degrees of freedom (DF).

<table>
<thead>
<tr>
<th>Sum of Squares</th>
<th>Numer. DF</th>
<th>Denom. DF</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>age</td>
<td>11271</td>
<td>1</td>
<td>20.6</td>
<td>3.665</td>
</tr>
<tr>
<td>task</td>
<td>7548</td>
<td>1</td>
<td>46.9</td>
<td>2.454</td>
</tr>
<tr>
<td>voc</td>
<td>4898</td>
<td>1</td>
<td>42.1</td>
<td>1.593</td>
</tr>
<tr>
<td>gender</td>
<td>154846</td>
<td>1</td>
<td>20.4</td>
<td>50.355</td>
</tr>
<tr>
<td>age:task</td>
<td>103409</td>
<td>1</td>
<td>1239.8</td>
<td>33.628</td>
</tr>
<tr>
<td>age:voc</td>
<td>1557</td>
<td>1</td>
<td>1233.5</td>
<td>0.506</td>
</tr>
<tr>
<td>task:voc</td>
<td>14396</td>
<td>1</td>
<td>46.4</td>
<td>4.682</td>
</tr>
<tr>
<td>age:gender</td>
<td>16665</td>
<td>1</td>
<td>20.3</td>
<td>5.419</td>
</tr>
<tr>
<td>voc:gender</td>
<td>94976</td>
<td>1</td>
<td>1240.1</td>
<td>30.885</td>
</tr>
<tr>
<td>task:gender</td>
<td>15081</td>
<td>1</td>
<td>1246.1</td>
<td>4.904</td>
</tr>
<tr>
<td>age:task:voc</td>
<td>24556</td>
<td>1</td>
<td>1233.5</td>
<td>7.985</td>
</tr>
<tr>
<td>age:voc:gender</td>
<td>25215</td>
<td>1</td>
<td>1225.4</td>
<td>8.200</td>
</tr>
<tr>
<td>task:voc:gender</td>
<td>0</td>
<td>1</td>
<td>1240.1</td>
<td>0.000</td>
</tr>
<tr>
<td>age:task:gender</td>
<td>1370</td>
<td>1</td>
<td>1245.3</td>
<td>0.445</td>
</tr>
</tbody>
</table>

In Figure 1 and 2 (see Appendix) the results on $F_1$ are presented, showing no difference in the pronunciation of <ä, äh> and <e, ee, eh>, except for the male subjects of the older generation in Figure 1. In Figure 2, the results on $F_1$ for female subjects are shown.

4.1.2 Results on $F_2$

As concerns the second formant, again a statistically significant threefold interaction of age*gender*vowel was found ($p = 0.008$). The post-hoc Tukey test revealed only significant differences involving different genders, which was again a significant main effect ($p < 0.0001$). Table 4 shows all results for the ANOVA with Satterthwaite approximation for degrees of freedom on $F_2$.

Table 4. ANOVA table for $F_2$ using the Satterthwaite approximation for degrees of freedom (DF).

<table>
<thead>
<tr>
<th>Sum of Squares</th>
<th>Numerator DF</th>
<th>Denom. DF</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>age</td>
<td>6141</td>
<td>1</td>
<td>20.1</td>
<td>0.374</td>
</tr>
<tr>
<td>task</td>
<td>644204</td>
<td>1</td>
<td>44.3</td>
<td>39.236</td>
</tr>
<tr>
<td>voc</td>
<td>3466</td>
<td>1</td>
<td>39.7</td>
<td>0.211</td>
</tr>
<tr>
<td>gender</td>
<td>610847</td>
<td>1</td>
<td>20.1</td>
<td>37.205</td>
</tr>
<tr>
<td>age:task</td>
<td>49081</td>
<td>1</td>
<td>1236.8</td>
<td>2.989</td>
</tr>
<tr>
<td>age:voc</td>
<td>10378</td>
<td>1</td>
<td>1234.3</td>
<td>0.632</td>
</tr>
<tr>
<td>task:voc</td>
<td>49505</td>
<td>1</td>
<td>43.9</td>
<td>3.015</td>
</tr>
<tr>
<td>age:gender</td>
<td>24228</td>
<td>1</td>
<td>20.0</td>
<td>1.476</td>
</tr>
<tr>
<td>voc:gender</td>
<td>140218</td>
<td>1</td>
<td>1241.7</td>
<td>8.540</td>
</tr>
<tr>
<td>task:gender</td>
<td>9688</td>
<td>1</td>
<td>1243.3</td>
<td>0.590</td>
</tr>
<tr>
<td>age:task:voc</td>
<td>4254</td>
<td>1</td>
<td>1234.3</td>
<td>0.259</td>
</tr>
<tr>
<td>age:voc:gender</td>
<td>115580</td>
<td>1</td>
<td>1227.3</td>
<td>7.040</td>
</tr>
<tr>
<td>task:voc:gender</td>
<td>19582</td>
<td>1</td>
<td>1241.7</td>
<td>1.193</td>
</tr>
</tbody>
</table>
Also, a significant main effect of the task, reading versus spontaneous speech, was found \((p < 0.0001)\); with lower formant frequencies in spontaneous speech than in the reading task. The threefold interaction of age\(\times\)task\(\times\)gender almost reached significance \((p = 0.0503)\). The twofold interaction of gender\(\times\)vowel is significant \((p = 0.003)\). It was, however, not further analysed as it was part of a significant threefold interaction. In Figures 3 and 4 (see Appendix) the results on F\(\_2\) are presented, for males and females, respectively.

### 4.1.3 Results on F\(\_3\)

Turning to the third formant, the interaction of age\(\times\)gender\(\times\)vowel was again statistically significant \((p = 0.004)\), but the post-hoc Tukey test did not show any results regarding differences between \(<\text{ä}, \text{äh}>\) and \(<\text{e}, \text{ee}, \text{eh}>\); gender differences not considered \((p < 0.0001\) for the main gender effect). The task effect was significant \((p < 0.0001, \text{read} > \text{spontaneous})\), and the twofold interactions of gender\(\times\)vowel \((p = 0.002)\) and age\(\times\)vowel \((p = 0.049)\) were significant, but they were included in the threefold interaction. Table 5 shows all results for the ANOVA with Satterthwaite approximation for degrees of freedom on F\(\_3\). Figures 5 and 6 (see Appendix) show the results on F\(\_2\).

**Table 5.** ANOVA table for F\(\_3\) using the Satterthwaite approximation for degrees of freedom (DF).

<table>
<thead>
<tr>
<th>Sum of Squares</th>
<th>Numerator DF</th>
<th>Denom. DF</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>age</td>
<td>56359</td>
<td>1</td>
<td>20.2</td>
<td>1.915</td>
</tr>
<tr>
<td>task</td>
<td>1023294</td>
<td>1</td>
<td>43.8</td>
<td>34.779</td>
</tr>
<tr>
<td>voc</td>
<td>21</td>
<td>1</td>
<td>40.0</td>
<td>0.001</td>
</tr>
<tr>
<td>gender</td>
<td>1480210</td>
<td>1</td>
<td>20.2</td>
<td>50.308</td>
</tr>
<tr>
<td>age:task</td>
<td>7001</td>
<td>1</td>
<td>1232.0</td>
<td>0.238</td>
</tr>
</tbody>
</table>
| age:voc        | 117573       | 1         | 1227.3| 3.996 | 0.05  |*
| task:voc       | 53285        | 1         | 43.9  | 1.811 | 0.19  |
| age:gender     | 36742        | 1         | 20.1  | 1.249 | 0.28  |
| voc:gender     | 280851       | 1         | 1241.5| 9.545 | 0.002 |**|
| task:gender    | 72193        | 1         | 1244.9| 2.454 | 0.12  |
| age:task:voc   | 59839        | 1         | 1227.3| 2.034 | 0.15  |
| age:voc:gender | 246579       | 1         | 1227.4| 8.381 | 0.004 |**|
| task:voc:gender| 1            | 1         | 1241.5| 0.000 | 1     |
| age:task:gender| 84037        | 1         | 1242.7| 2.856 | 0.09  |*

### 4.2 Qualitative analysis

In addition to the quantitative analysis, some special cases were analysed qualitatively. Interesting instances of the \([\varepsilon:\text{ä}]\)-pronunciation occasionally appeared in read speech, giving hints to underlying mechanisms.

First, the reading task was repeated once, so that each subject read the list of sentences twice. Occasionally, some speakers pronounced graphemic \(<\text{ä}, \text{äh}>\) as \([\varepsilon:]\) to follow
some imaginary norms they had in their minds. Interestingly, they did not always use \( [\varepsilon:] \) for both occurrences of a word, but in one instance only. Based on this observation, we conclude that the prescriptive norms in those speakers’ minds are rather unstable; they are insecure concerning the proper realisation of graphemic \(<\ddot{a}, \ddot{a}h>\). It can be concluded that \([\varepsilon:]\)-pronunciation for \(<\ddot{a}, \ddot{a}h>\) is disappearing. This conclusion is corroborated by the result that only \([\varepsilon:]\)-pronunciations occur for \(<\ddot{a}, \ddot{a}h>\) in spontaneous speech. In addition, in the reading task, the most frequent word\(^{16}\) \( \text{täglich}, \text{“daily”} \), is consistently pronounced with \( [\varepsilon:] \). Only once, \(<\ddot{a}>\) in \( \text{täglich} \) is pronounced with \( [\varepsilon:] \) by one subject, however, in the second round, he pronounced \( [\varepsilon:] \).

Next, some selected results from the interview are presented, where the subjects explicitly talked about their use of \(<\ddot{a}, \ddot{a}h>\). First, one speaker (p285) talked about self-correction; he is one of the three speakers occasionally producing \([\varepsilon:]\) in the reading task. When first reading \( \text{Säle}, \text{“hall, pl.”} \), he pronounced \([\varepsilon:]\) without hesitating or commenting on it. In the repetition of the reading task, first, he pronounced \([\varepsilon:]\) again, but after a short hesitation pause, he, murmuringly, tried to pronounce the word with \( [\varepsilon:] \) and \( [\varepsilon:] \), concluding that \( [\varepsilon:] \) is more adequate. He then repeated the whole sentence with the \([\varepsilon:]\)-pronunciation of the target word. In the interview later on, he referred to these insecurities:

“\( \text{Saal, Säle, wird eher wie ein } [\varepsilon:] \text{ ausgesprochen. Und/ und..., das hat mir dann nicht gefallen, Säle } [\varepsilon:] \text{sagt man nicht. Aber ich war irgendwie so falsch drinnen und daher habe ich es noch einmal wiederholt.} \)”\(^{17}\) (sentences 41-43, spontaneous speech of subject p285)

This example shows the speaker’s awareness of pronunciation, as he corrected himself when he had the feeling that the pronunciation was wrong. Additionally, it shows that the \([\varepsilon:]\)-pronunciation is somehow related to specific lexemes, probably again linked to a word’s frequency (see above and section 5.3.). The recording of this speaker generally gives the impression that he provides a high priority to pronunciation, but nevertheless, even in the reading task, he uses \( [\varepsilon:] \) in the most frequent target words.

Another speaker (p013) expressed the belief that \( [\varepsilon:] \) is not used in everyday life, but that the laboratory setting of the recording can trigger an \([\varepsilon:]\)-pronunciation. He cannot give any reasons, except intuition, why for some words, such as \( \text{säen}, \text{“to saw”} \), \( [\varepsilon:] \) is an acceptable pronunciation in this unnatural setting, whereas for other words, e.g. \( \text{Gräte}, \text{“fishbone”} \), he refuses an \([\varepsilon:]\)-pronunciation.

\(^{16}\) According to the online word counter of the University of Leipzig (Quasthoff & Richter, 2005).

\(^{17}\) “Hall, halls, is rather pronounced as \([\varepsilon:]\). And/and..., I did not like it, you don’t say halls (with \([\varepsilon:]\)). But I got it the wrong way and therefore, I repeated it” (translation by authors).
5 Discussion

The present study was designed to determine the pronunciation of <ä, äh> as compared to <e, ee, eh>. Spontaneous speech and reading tasks were recorded with 24 speakers, balanced for age and gender. With this study design, influences of the speaking task as well as influence of age and gender were tested.

Regarding our research question on the pronunciation of long <ä, äh> in Standard Austrian German, the statistical analysis revealed no differences in the pronunciation of <ä, äh> and <e, ee, eh> (Hypothesis 1). Focusing on our four groups of subjects, the interaction of age*gender*vowel reached significance in all three formants, but for F2 and F3, all significant contrasts can be easily explained as a consequence of the gender main effect, i.e. only those contrasts were significant in which different genders were involved. These contrasts were expected to be significant, as gender differences in formant frequencies always occur; thus, these results are not relevant for our research question.

5.1 Speaking task and speaker specific characteristics

A main effect of the speaking task was found for F2 and F3, indicating higher formant frequencies for the reading task than for spontaneous speech. For the second formant, the size of the effect was different depending on age and gender. These results are consistent with those of previous studies (Moosmüller, 2007). For F1, an interaction of task*gender was found, indicating a small difference of about 10 Hz between genders which can be ignored as it is not expected to affect perception.

Concerning the results on F1, the group of older male speakers differs from the other groups. The statistical analysis revealed that the F1 of these speakers was raised for <ä, äh> as compared to <e, ee, eh>, indicating that for <ä, äh>, [ɛː] is pronounced. A subsequent auditory analysis indicated that the speaking task effects were most significant. In the reading task of some speakers, several [ɛː]-realisations occurred, (research question B and hypothesis 1a). Within the group of older male speakers, mainly two speakers distinguish consciously between <ä, äh> and <e, ee, eh> in most words during the reading task, resulting in the above-mentioned statistically significant differences. This distinction was also true for one young male speaker, but did not affect any statistical results. This was also true for single realisations of other speakers. Conversely, in spontaneous speech, no instances of [ɛː] occurred, (Hypothesis 1b). The fact that only a few individual speakers differ from the others indicate that the pronunciation of <ä, äh> might be a speaker–specific characteristic. However, this issue needs to be investigated on a larger corpus.

During the interview, some speakers also stated that, when reading the sentences, they were unsure about the correct pronunciation of <ä, äh> in some words, while they clearly preferred [ɛː] for <ä, äh> in other words. Other speakers reported that they definitely prefer [ɛː] in all words with <ä, äh>. The analysed formant frequencies, as well as the comments on the pronunciation of <ä, äh>, show that only a few speakers try to adhere to an old prescriptive norm as determined, e.g., in Siebs (1912) and succeed in doing so, when orthography demands their attention.
5.2 Differences in age and gender

The results of our research question on age and gender corroborated our hypotheses (Hypotheses 1c and 1d). No clear differences were found between the two age groups. Only one statistically significant difference regarding age was obtained: the realisation of <ä, äh> in spontaneous speech differs between the younger and the older age group, with older speakers exhibiting a lower F_1 than younger ones. However, most importantly, no difference occurred between <ä, äh> and <e, ee, eh>, in either group.

Regarding gender, a gender main effect was expected. It was found in all three formants, showing the well-known fact that women exhibit higher formant frequencies than men. Statistical analysis rendered no unexpected differences (see Hypothesis 1d).

5.3 Word frequency

As a by-product of our analysis, we observed an influence of word frequency. In the reading task, word frequency might also affect the implementation of norms. Word frequency also plays a decisive role in sound changes, with high frequency words being affected first (Bybee, 2012). Therefore, we assume a tendency that higher word frequency is associated with higher likelihood of [eː]-pronunciations of <ä, äh>. Due to the small data set, this could not be analysed statistically, but only qualitatively.

An online search of the word frequency of words with <ä, äh> occurring in the reading task was performed with the online tool of the University of Leipzig (Quasthoff & Richter, 2005). In our reading task, the word täglich, “daily”, was by far the most frequent word (of all <ä, äh> words). The expected prevailing [eː]-pronunciation was confirmed in the auditory analysis, as only in one single occasion a speaker pronounced the word with [eː]. It should be noted that, in the repetition of the reading task, this speaker pronounced [eː]. By and large, the frequency counting of the University of Leipzig and the likelihood that [eː] is pronounced for <ä, äh> seem to be related; the less frequent a word is, the more likely it is pronounced with [eː] in the reading task. Only one word did not fit into this pattern; gäbe, “would give”. This word occurred rather frequently, but is often pronounced with [eː]. A possible explanation for this can be the subjunctive mood which is differentiated by the pronunciation of [eː] vs. [eː] from the indicative. However, it has to be noted that nowadays, these subjunctive forms are hardly ever used in spontaneous speech, but only in (formal) writing. Mangold (1966: 40) referred to the “preserving influence” of the subjunctive for an [eː]-pronunciation.

5.4 Merger or unmerger?

Our hypotheses stated at the beginning of the study can be confirmed. In general, no differences are made in the pronunciation of long <ä, äh> and <e, ee, eh> (Hypothesis 1), although some speakers adhere to obsolete norms of [eː]-pronunciation when reading (Hypothesis 1a), but not so in spontaneous speech (Hypothesis 1b). Thus, the use of [eː] is limited to the formal context of a reading task, and even then only some individuals consider it important to distinguish graphemic <ä, äh> and <e, ee, eh>. The present study confirms the previous findings of Iivonen (1987) and Moosmüller (2007). Consequently, no reversal of a merger of <ä, äh> and
<e, ee, eh>, as proposed in Sloos (2013), can be found (research question D), and her hypothesis can safely be refuted.

Considering the historical development of Standard Austrian German, it can be concluded that the two vowels were hardly ever discerned. It was only when the prescriptive norms recorded in Siebs were expected to be realised in Austria, that importance was placed on this kind of distinction in education and speaker training. Yet, in Standard Austrian German, the distinction is rather artificially produced, and, as long as the graphemic distinction is carried out, some residual influences on pronunciation will persist, especially in reading.

Further investigations could focus exclusively on the effect of orthography. This could be achieved by setting up different kinds of reading tasks, such as reading minimal pairs, word lists, sentences and continuous texts.

Acknowledgements

The subjects for the current investigation were recruited within the project “Soziolekte in Wien – die mittelbairischen Varietäten” (“Sociolects in Vienna – the Middle Bavarian variants”) funded by the cultural department, Vienna (Kultur Wien). We are grateful for the helpful comments of two anonymous reviewers on an earlier version of the paper. Special thanks go to Jonnie White for proof-reading.
Males

Figure 1. $F_1$ (in Hz) of $\langle e, ee, eh \rangle$ and $\langle ä, äh \rangle$ in spontaneous and read speech of younger and older male subjects

Females

Figure 2. $F_1$ (in Hz) of $\langle e, ee, eh \rangle$ and $\langle ä, äh \rangle$ in spontaneous and read speech of younger and older female subjects
Figure 3. $F_2$ (in Hz) of <e, ee, eh> and <ä, äh> in spontaneous and read speech of younger and older male subjects

Figure 4. $F_2$ (in Hz) of <e, ee, eh> and <ä, äh> in spontaneous and read speech of younger and older female subjects
Figure 5. F$_3$ (in Hz) of <e, ee, eh> and <ä, äh> in spontaneous and read speech of younger and older male subjects

Figure 6. F$_3$ (in Hz) of <e, ee, eh> and <ä, äh> in spontaneous and read speech of younger and older female subjects
References


FORMANT ANALYSIS OF THE MID-FRONT VOWEL AS REALIZED IN HESITATION DISFLUENCIES IN HEBREW

Vered Silber-Varod¹, Noam Amir²
¹The Research Center for Innovation in Learning Technologies, The Open University of Israel, Israel
²Department of Communication Disorders, Faculty of Medicine, Tel Aviv University, Ramat Aviv, Israel

E-mail: ¹vereds@openu.ac.il, ²noama@post.tau.ac.il

Abstract
This study attempts to characterize the timbre of the default type of hesitation disfluency (HD) in Israeli Hebrew: the mid-front vowel /e/. We used a triple categorization of hesitation disfluencies: 1. Elongated word-final syllables, 2. Final [e] appended to the lexical word, and 3. Filled pauses. The goal was to find consistent differences/similarities in hesitation types based on their spectral features. To this end, we measured the formant frequencies (F₁, F₂, and F₃) of 575 HD pronunciations taken from the Corpus of Spoken Israeli Hebrew (CoSIH), containing all three types of HDs. We also compared the results to two former studies that were carried out on the vowel /e/ in Hebrew fluent speech, in order to determine whether [e] in hesitation is more centralized when compared with the usual (fluent) [e]. The findings show that, in general, HDs do not tend to be more centralized. Elongated word-final syllables and final [e]s are pronounced with the same amount of openness as the fluent [e], while filled pauses tend to be more open (lower F₁), and more frontal (higher F₂). In light of these results, we suggest using different IPA symbols for the different types of HDs, rather than just the phonemic mid-front /e/.

Keywords: Hesitation disfluency, filled pauses, e vowel, LPC analysis, formants, spontaneous speech, Israeli Hebrew.

1 Introduction
Hesitation disfluencies (HDs) are vocalic fillers speakers utter spontaneously. The most prominent observed effect of HDs is the lengthening of rhymes or syllables. This suggests that speakers are hesitating using such vocal fillers in the same way as they might use a filled pause (Shriberg, 2001, p. 161). However, the phenomenon of filled pauses is considered language dependent with respect to its vowel quality (Shriberg 1994): “Filled pauses have variants in many (perhaps all) languages, but their vowel quality varies predictably with the vowel inventory of the language” (Shriberg, 1994, p. 24-25). Shriberg (2001) demonstrates how disfluency is also associated with alternations in vowel quality – an elongated the is pronounced as the variant [ði] (Shriberg, 2001, p. 163). Fox (2000) suggests that the ‘hesitation vowel’ used by speakers is different even in speakers of different varieties of English. When
discussing filled pauses, he says that the ‘neutral or rest position’ is not always the same for different languages (Fox, 2000, p. 30). Cruttenden (1997) provides additional examples of the differences in the sounds of filled pauses between dialects and languages: “In R.P. and in many other dialects of English the latter [the filled pause] typically involves the use of a central vowel [ə] ... in Scottish English a sound in the region of the vowel in gate and play is typical while in Russian an alveolar nasal is more common than a bilabial nasal” (Cruttenden, 1997, p. 30). In French, it is the closed-mid front vowel [ɔ], commonly written as euh or [œ] but also as an [e], while in American English, it is the open-mid back vowel [ʌ] (Shriberg 2001, p. 164). For European Portuguese, the two most common phonetic forms of filled pauses are the near-open central vowel [ɐ] and the mid-central vowel [ə] (Candeias, Celorico, Proença, Veiga & Perdigão, 2013). In Estonian, it was found that hesitation formants are likely to be centralized or posterior and opened in comparison with related phonemes (Nemoto, 2013). In European Portuguese, a small tendency for higher F1 and lower F2 in vocalic fillers (termed FP in the present research) was found (Proença, Celorico, Veiga, Candeias & Perdigão, 2013).

The notion that lengthened vowels might also be distinctive in their spectral characteristics is not new. In languages with short and long vowels, it is often observed that short and long counterparts have somewhat different vowel qualities. For example, the distinction between long and short vowels in two dialects of Arabic was found not only in duration, as the vowel space for the short vowels was found to be more centralized than that of the long vowels (Amir, Tzenker, Amir & Rosenhouse, 2012). In Thai, Abramson and Ren (1990) found that there are sufficient differences in formant patterns to yield slight, but audible differences, in vowel quality associated with the length feature.

In the present study, we focus on hesitation disfluencies in spoken Israeli Hebrew. HDs were found in Silber-Varod (2013a; 2013b) as largely produced by the mid-front vowel /e/. This vowel, together with four others /a, i, o, and u/, comprise the five vowels in the Israeli Hebrew (IH) vowel system (Bolozky, 1997). This vowel system does not include a phonological distinction of vowel length. Moreover, lax/tense differences are not phonemic in Israeli Hebrew, nor does it includes the neutral Schwa [ə] as a phoneme. According to the International Phonetic Alphabet (2005), diacritics indicating lowering [ɛ] or rising [ɛ], may be used to represent the quality of mid-front unrounded vowel (nevertheless, in the following, the symbol [e] will be used to represent this vowel). /e/ is the phonetically unmarked vowel in IH and is the vowel most likely to break up undesirable consonant clusters and to be affected by casual vowel deletion (Bolozky, 1999).

In this paper we will use triple categorization of HDs in IH, i.e. three distinct manifestations with regard to word-level phonology (cf. Silber-Varod, 2010; Silber-Varod, 2013a; Silber-Varod, 2013b):

A. Elongated word-final syllables (EWFS). Elongated syllables are mostly monosyllabic function words, such as /ve/ ‘and’; /be/ ‘in/at’; /je/ ‘that’, and more. A duration threshold of 230 ms was applied for inclusion as an EWFS,
as justified in Silber-Varod 2013a p. 70-73). A spectrogram of this type is presented in Figure 1.

**Figure 1.** An example of an elongated word-final syllable (EWFS) as realized on the word [be] ‘in’, within the translated utterance “I have to take an exam at the end of the year in: (0.969 s pause) women’s history”.

B. Final [e] vowels that are appended after a word, with no silent pause in between (i.e., within the same intonation unit). This type will be further referred as Final [e] (henceforth: F[e]). A spectrogram of this type is presented in Figure 2.

**Figure 2.** An example of a Final [e] vowel appended after a word (F[e]) as realized attached to the word [ʃel] ‘of’, within the translated utterance “It took a time span of: (0.483 s pause) forty-two hours”.

C. HDs between silent pauses, known also as filled pauses (FP). This type is realized only as elongated [e], and sometimes with the nasal consonants in coda position, as [em]. A spectrogram of this type is presented in Figure 3.
Figure 3. An example of a Filled Pause [e] as realized between two silent pauses of 205 ms and 118 ms, respectively, within the translated utterance “It’s the capital of (0.205 s pause) [e:] (0.118 s pause) the the South Gobi”.

A similar distinction between the first and the third realisations (types A and C above) was described in (Shriberg, 1994). It is important to note that although types B and C are only realized by the /e/ vowel in IH, elongated syllables occur in word final syllables, or monosyllable lexemes, and thus may be realized by each of the five vowels in the IH vowel system. For example, the most common elongated lexeme in spontaneous Hebrew is the definite article [ha] ‘the’ (Silber-Varod, 2013b), which of course is realized by an elongated /a/.

This study focuses on two questions:
Q1: Do the three types of hesitation disfluencies have similar formant values?
Q2: Are formants of [e] when realized by hesitations more likely to be centralized in comparison with fluent [e]?

Regarding the first question, we expected to find similar formants for the three types of HDS (EWFS, F[e], and FP). As to the second question, previous studies on the formants of hesitation disfluencies showed differences from their fluent counterparts (Shriberg, 2001; Nemoto, 2013; Proença, Celorico, Veiga, Candeias & Perdigão, 2013), therefore we expected to find differences also in IH.

To summarize, in the present study, formant analysis of three types of HDs produced with the vowel /e/ was carried out and the results were compared to two other studies on the acoustics of vowels in IH: The first is on the acoustic properties of Hebrew vowels as manifested in isolated utterances (Most, Amir & Tobin, 2000); The second is on the phenomenon of vowel reduction in spontaneous speech (Amir & Amir, 2007), a study that was carried out in part on the same corpus as in the present research.

Filled pause detection relying on vocal tract characteristics (and hence the formants) has been shown to outperform other methods (Audhkhasi, Kandhway, Deshmukh & Verma, 2009). Therefore, this study may contribute to various applications that require automatic detection of filled pauses. It may also be used to evaluate the spoken fluency skills of the speaker and to improve the performance of automatic speech recognition systems (Audhkhasi, Kandhway, Deshmukh & Verma, 2009).
2 Method

2.1 Recordings

Speech samples were taken from the Corpus of Spoken Israeli Hebrew (CoSIH). The recordings, which were carried out during 2001-2002, are of authentic IH everyday conversations. Participants wore two unobtrusive microphones on their lapels, during an entire waking day. Each dialogue consists of conversations between the one core speaker and various interlocutors with whom the speaker interacted during this day.

The present research is based on recordings of 19 subjects who carried the recording device on their body. In addition, the speech of 28 interlocutors was also transcribed, resulting in a total of 47 different speakers (19 men and 28 women). The compilation method is described in Silber-Varod (2013a, pp. 37-38), and is generally based on the volume of extractable data. Each recording selected was over 9 minutes long and contained at least 1,000 words. All participants were native speakers of Hebrew, residing in Israel since birth. Mean age was 30.5 years (SD = 14.6) for the men, and 26.3 years (SD = 5.5) for the women. The research corpus consists of 31,760 word-tokens (over 5 hours of speech) of which 4,289 are word-types. 56% of the conversations are face-to-face dialogues, while 44% are one-sided telephone conversations, in which only the participant with the recording device can be heard. More about the recording process can be found on the cosih.com website.

2.2 The database of HDs

Amongst 764 HDs that were annotated in the corpus, the most frequent elongated word-types were open monosyllabic words, of which the most frequent words were: 82 occurrences of [ha] ‘the’, 39 occurrences of [ve] ‘and’, 30 occurrences of [ze] ‘this’, 28 occurrences of [ba] ‘in the/at the’, 25 occurrences of [je] ‘that’, etc. It is worth noting that these are function words. Excessive elongation also occurred frequently in disyllabic words such as [ha’ze] ‘that’ or in monosyllabic CCV structures such as in [kʃe] ‘when’ and [kmo] ‘as’. However, not only words with final [e] underwent elongation. As summarized in Silber-Varod (2010), the most frequent elongated vowel in spontaneous spoken Hebrew is [a] (36% of elongated EWFS), mainly when speakers lengthen the definite article [ha] ‘the’. The vowel [e] was found to be elongated in 34% of the EWFS (more than its 27% total frequency in the corpus).

In the current research, we analysed 575 HDs of the vowel [e], and these were found to be of all three types (A, B, and C) defined above. The remaining 189 HDs consisted of other elongated vowels (mostly [a] in the Hebrew definite article /ha/ ‘the’). This demonstrates that the [e] vowel is the default HD in Hebrew (75.2% HDs with [e] vs. 24.7% HDs with the other four vowels). Figure 4 demonstrates the distribution of the three types of HDs between women and men: 70 elongated word-final syllables were collected from men’s speech, and 162 from women’s speech; 105 F[ε]s were collected

1 Available at: http://cosih.com/english/index.html
from men, and 202 from women; lastly, only 3 filled pauses, were collected from men and 33 from women.

![Figure 4. Occurrences of the three realisations of hesitation disfluencies spoken by women and men](image)

Similar to the unbalanced distribution of HDs among men and women, HDs are not distributed equally among speakers. Some speakers uttered a single HD during the conversation, while others uttered many more. For example, in one phone conversation, the subject uttered over one hundred HDs.

### 2.3 Formant extraction method

Formant extraction was performed in the same manner used by Praat, through downsampling and applying Linear Predictive Coding (LPC). This was applied to segments of 30-50ms, from the centre of each HD, judged visually to be stationary. The first three formant frequencies were identified as the frequencies of the first three conjugate pole pairs obtained from this analysis. Generally, the results of this type of analysis may vary considerably when applied automatically, depending on sampling rate and order of the LPC algorithm. Therefore, we performed a manually supervised analysis of each HD separately. A certain amount of background noise in this type of corpus is inevitable. However we have accumulated considerable experience with this database and formant extraction in general (e.g., Amir, Tzenker, Amir & Rosenhouse, 2012). In our experience, formant values are not extremely sensitive to background noise, since it introduces variability which is much smaller than the normal variability in production. In any case, recordings with strong background noise were discarded.
3 Results

The results are divided into two parts: The first part describes intrinsic findings on the measurements of the first three formants in the three examined HDs – EWFS, F[e], and FP. In the second part we present extrinsic findings which compare the results of the current study to the results in the two previous studies mentioned above (Amir & Amir, 2007; Most, Amir & Tobin, 2000), which will be termed the “reduced /e/” study and “read /e/” study, respectively. A summary of the measured formant values (means and standard deviation values) is presented in Table 1.

3.1 Intrinsic comparison

The intrinsic comparison is a comparison between the three types of HDs – elongated word-final syllables (EWFS), Final [e] (F[e]), and filled pauses (FP) described above. Descriptive results are presented for all 575 items in the database in Table 1. Several trends are evident:

- For men and women, F1 in FPs is lower than in EWFSs and F[e]s.
- For men and women, F2 is lowest in EWFSs, higher in F[e]s, and highest in FPs.
- For men, F3 is lowest in EWFSs, higher in F[e]s, and highest in FPs.
- For women, F3 is lowest in F[e]s, slightly higher in EWFSs, and much higher in FPs.

Table 1. Descriptive statistics of F1, F2, and F3 in the three types of HDs (EWFS, F[e], and FP), for men (M) and women (W)

<table>
<thead>
<tr>
<th>Formants</th>
<th>Gender</th>
<th>EWFS Mean</th>
<th>SD</th>
<th>F[e] Mean</th>
<th>SD</th>
<th>FP Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>F1 (Hz)</td>
<td>M</td>
<td>494</td>
<td>85</td>
<td>499</td>
<td>88</td>
<td>448</td>
<td>51</td>
</tr>
<tr>
<td></td>
<td>W</td>
<td>585</td>
<td>88</td>
<td>580</td>
<td>88</td>
<td>535</td>
<td>85</td>
</tr>
<tr>
<td>F2 (Hz)</td>
<td>M</td>
<td>1683</td>
<td>246</td>
<td>1716</td>
<td>245</td>
<td>1933</td>
<td>284</td>
</tr>
<tr>
<td></td>
<td>W</td>
<td>1977</td>
<td>233</td>
<td>2024</td>
<td>283</td>
<td>2308</td>
<td>427</td>
</tr>
<tr>
<td>F3 (Hz)</td>
<td>M</td>
<td>2818</td>
<td>400</td>
<td>2853</td>
<td>383</td>
<td>2928</td>
<td>488</td>
</tr>
<tr>
<td></td>
<td>W</td>
<td>3025</td>
<td>286</td>
<td>3013</td>
<td>321</td>
<td>3307</td>
<td>543</td>
</tr>
</tbody>
</table>

As F1 values reflect the open-closed vowel characteristic, results suggest that FP is pronounced in a more closed manner than EWFS and F[e]. As F2 values reflect the front-back distinction, results suggest that FP is pronounced in a more frontal manner than EWFS and F[e].

The model used in the statistical analysis was selected to account for the imbalance of HDs amongst the different speakers and HD types. Mixed model tests were performed for each formant, using the SAS PROC MIXED model, for men (M) and women (W) separately, for each of the first three formants, with HD type as a within-subject factor.

No significant results were found, for men, for any of the three formants (F1: F(2, 13) = 0.56, p = 0.5861, F2: F(2, 13) = 1.67, p = 0.227, F3: F(2, 13) = 1.67, p = 0.25).
For women, a significant main effect was found for all three formants ($F_1$: $F(2, 27) = 4.47$, $p = 0.0211$, $F_2$: $F(2, 27) = 19.37$, $p < 0.001$, $F_3$: $F(2, 27) = 11.5$, $p < 0.001$)

Contrast analysis was therefore carried out for women only, revealing:

1. $F_1$:
   A. EWFS is significantly different from FP ($p = 0.019$)
   B. P is significantly different from F[e] ($p = 0.031$)
   C. EWFS is not significantly different from F[e] ($p = 0.951$).

2. $F_2$:
   A. EWFS is significantly different from FP ($p < 0.001$)
   B. FP is significantly different from F[e] ($p < 0.001$)
   C. EWFS is not significantly different from F[e] ($p = 0.321$).

3. $F_3$:
   A. EWFS is significantly different from FP ($p < 0.001$)
   B. FP is significantly different from F[e] ($p < 0.001$)
   C. EWFS is not significantly different from F[e] ($p = 0.980$).

In general, the patterns for women are clear and consistent. For men no significance was achieved, possibly due to the smaller number of occurrences.

3.2 Extrinsic comparison

The comparison to previous studies was carried out qualitatively. Mean $F_1$ and $F_2$ values are presented in Table 2 and graphically in Figure 5.

The main findings are the following:

1. $F_1$: In comparison to the mean $F_1$ of “read /e/” (M: 455, W: 548), the mean $F_1$ values are higher in EWFS (M: 494, W: 585) and F[e] (M: 499, W: 580). This reflects a more open manner of disfluent [e] in these two HDs. On the other hand, mean $F_1$ is lower in FP (M: 448, W: 535), which reflects a more closed [e] in FPs, as compared to read /e/.

2. $F_2$: In comparison to men’s fluent [e] (in the read speech and the reduced /e/ studies), $F_2$ of the three HDs is higher. In women’s speech, the mean $F_2$ of F[e] and EWFS is lower than in the “read /e/” study (Most, Amir & Tobin, 2000), but all three disfluent $F_2$s are higher than the $F_2$ of reduced /e/.

Figure 5 illustrates the data in Table 2, for men and women separately (the solid ellipse surrounds men’s values, while the dotted ellipse surrounds the women’s values). In general, results showed that EWFSs are closer in pronunciation to F[e]s and farthest from FPs. It is evident that for both men and women, $F_1$ of filled pauses (FP) is closer to the read /e/ values, while $F_1$ of EWFS and F[e] is closer to the reduced /e/. This reflects a similar open manner of production of the [e] vowel in fluent and disfluent speech.

With regard to $F_2$, it seems that disfluent $F_2$ of all three HDs is higher than the reduced $F_2$, thus reflecting a more frontal manner. More specifically, men pronounce $F_2$ in F[e] and EWFS in a similar frontal manner (higher $F_2$), in comparison to the
“read /e/”. $F_2$ of FPs, for both men and women, is higher than $F_2$ of /e/ in the two previous studies. This should probably be verified on a more extensive corpus.

Table 2. Extrinsic comparison of $F_1$ and $F_2$ in the three types of HDs and in two previous work on [e] formants in fluent speech

<table>
<thead>
<tr>
<th>Formants</th>
<th>Gender</th>
<th>$F_1$ (Hz)</th>
<th>SD</th>
<th>$F_2$ (Hz)</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Read /e/</td>
<td>M</td>
<td>455</td>
<td>40</td>
<td>1662</td>
<td>171</td>
</tr>
<tr>
<td></td>
<td>W</td>
<td>548</td>
<td>60</td>
<td>2200</td>
<td>172</td>
</tr>
<tr>
<td>Reduced /e/</td>
<td>M</td>
<td>491</td>
<td>91</td>
<td>1514</td>
<td>175</td>
</tr>
<tr>
<td></td>
<td>W</td>
<td>597</td>
<td>126</td>
<td>1729</td>
<td>257</td>
</tr>
<tr>
<td>EWFS</td>
<td>M</td>
<td>494</td>
<td>85</td>
<td>1683</td>
<td>246</td>
</tr>
<tr>
<td></td>
<td>W</td>
<td>585</td>
<td>88</td>
<td>1977</td>
<td>233</td>
</tr>
<tr>
<td>F[e]</td>
<td>M</td>
<td>499</td>
<td>88</td>
<td>1716</td>
<td>245</td>
</tr>
<tr>
<td></td>
<td>W</td>
<td>580</td>
<td>88</td>
<td>2024</td>
<td>283</td>
</tr>
<tr>
<td>FP</td>
<td>M</td>
<td>448</td>
<td>51</td>
<td>1933</td>
<td>284</td>
</tr>
<tr>
<td></td>
<td>W</td>
<td>535</td>
<td>85</td>
<td>2308</td>
<td>427</td>
</tr>
</tbody>
</table>

Figure 5. Extrinsic comparison of $F_1$ and $F_2$ in the three types of HDs (FP, F[e], and EWFS) and in two previous works on fluent /e/ in lexical words (“reduced /e/” and “read /e/”), for men (M) and women (W)

The ratios of $F_2/F_1$ in the three HD types are presented in Table 3, with comparison to the ratios found for read speech [e], as reported in (Most, Amir & Tobin, 2000).
Table 3. $F_2/F_1$ ratios of the three HD types and of read speech [e].

<table>
<thead>
<tr>
<th>Gender</th>
<th>Read /e/ (Most, Amir &amp; Tobin, 2000)</th>
<th>EWFS $F_1$</th>
<th>EWFS $F_2$</th>
<th>FP $F_1$</th>
<th>FP $F_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>M</td>
<td>3.675</td>
<td>3.481</td>
<td>3.530</td>
<td>4.388</td>
<td></td>
</tr>
<tr>
<td>W</td>
<td>4.046</td>
<td>3.44</td>
<td>3.569</td>
<td>4.391</td>
<td></td>
</tr>
</tbody>
</table>

4 Discussion and Future Research

In this study, we attempted to characterize the timbre of the default type of hesitation disfluency (HD) in Israeli Hebrew: the mid-front vowel /e/. The first three formants of the hesitation disfluencies of 49 speakers were measured on excerpts taken from the Corpus of Spoken Israeli Hebrew (CoSIH). Our goal was to find consistent differences or similarities regarding the hesitation types based on their spectral features. We analyzed three types of hesitation disfluency: elongated word-final syllables (EWFS), a final [e] appended to the lexical word (F[e]), and filled pauses (FP).

The findings show that, in general, HDs do not tend to be more centralized. Of the three HD types, elongated word-final syllables and Final [e]s formant measurements are similar to each other while both are different from filled pauses. This is true for all three formants. Filled pauses in Hebrew are less central with comparison to the two other types of HD (Final [e] and elongated word-final syllables), i.e., they are pronounced with a more closed manner (lower $F_1$) and with a more frontal manner (higher $F_2$).

In comparison to the fluent /e/ discussed in the literature, $F_1$ of elongated word-final syllables and Final [e] is closer to reduced [e] in fluent spontaneous speech, while $F_1$ of filled pauses is closer to [e] in read monosyllabic non-words. We suggest that this reflects discrimination between the lexically attached types of HD (EWFSs and F[e]s), which have a reduced (centralized) $F_1$ nature, versus the isolated type of HDs (FPs), which have a more closed nature, as in read speech.

$F_2$ of all three types of HDs is higher than the fluent [e] in men’s pronunciation. In women’s speech, $F_2$ of elongated word-final syllables and Final [e] is higher than the reduced [e] but lower than the read [e]. In general, the $F_2$ of filled pauses is higher than both fluent /e/s.

Following the above analysis, we can thus suggest two options for the phonetic representation of the studied types of HDs. The first suggestion includes two different IPA symbols for HDs: Filled pauses may be indicated by [e] – the IPA symbol for closed-mid vowel; while elongated syllables and final [e] may be indicated by [ɛ] – the open-mid IPA symbol. This suggestion is farfetched since it assumes two different mid vowels in Israeli Hebrew.

The second option is to use IPA diacritics to indicate lowered vowel: [e̞] (for filled pauses) and raised vowel: [ɛ̝] (for elongated word-final syllables and Final [e]s).

The two options are presented in the following IPA vowel chart (Figure 6). The two vowels [e] and [ɛ̝] are located as Closed-mid and Open-mid IPA vowels, and the two diacritics locations are added artificially to the vowel chart.
These variations are yet to be investigated perceptually. However, in the present study, we observed a potential for a diachronic shift from length to quality in at least part of the vowel system. The present results thus contribute to the understanding of the historic rise and fall of distinctive vowels, as suggested by (Abramson & Ren, 1990, p. 268).

**Figure 6.** The vowel chart with the four possible symbols for the representation of hesitation disfluencies in Israeli Hebrew: [e], [ɛ]; [ɛ̝], [ɛ̞] (The International Phonetic Association. Retrieved from: http://www.internationalphoneticalphabet.org/ipa-charts/vowels/)

### 5 Conclusions
In this study, we analysed the first three formants of hesitation disfluencies in spoken Israeli Hebrew. Results showed that our triple categorization of hesitation disfluencies: Elongated word-final syllables, final appended [e]s, and filled pauses, has a phonetic justification, at least with regard to their spectral features. As to the question whether [e] realized by hesitation is centralized as compared to fluent [e], we found that filled pauses in Hebrew are less central in comparison to both reduced or fluent speech and to the two other types of HD (Final [e] and Elongated word final syllable). Hence, filled pauses are pronounced in a more closed manner (lower F1) and with a more frontal manner (higher F2), thus becoming closer in pronunciation to the Hebrew [i] vowel.

### Acknowledgment
We would like to thank Ms. Adva Weiss for her assistance. The International Phonetic Alphabet chart is made possible thanks to the International Phonetic Association (Department of Theoretical and Applied Linguistics, School of English, Aristotle University of Thessaloniki, Thessaloniki 54124, Greece).

### References
International Congress of Phonetic Sciences, 849–852.
CoSIH: The Corpus of Spoken Israeli Hebrew. Available at: http://humanities.tau.ac.il/~cosih.
TEMPORAL PATTERNS OF ERROR REPAIRS AND APPROPRIATENESS REPAIRS IN HUNGARIAN

Mária Gósy¹, Irene Vogel², Viola Váradi³
¹Dept. of Phonetics, Research Institute for Linguistics, Hungarian Academy of Sciences, Budapest, Hungary
²University of Delaware, Newark, Delaware, USA
³Eötvös Loránd University, Budapest, Hungary
e-mail: ¹gosy.maria@mta.nytud.hu, ²ivogel@udel.edu, ³varadi.viola@btk.elte.hu

Abstract
Natural conversational speech often exhibits interruptions and modifications of the speech stream when the speaker “repairs” what has been said. We examine two types of repairs, those involving real errors and those involving appropriateness considerations, in an eleven-hour corpus produced by 26 Hungarian speakers. Since both the reasons and the supposed processes of the speech planning are different in the two types of repairs, we hypothesize that these differences will be reflected by the temporal patterns of the editing phases, reparanda and repairs. Based on the analysis of the occurrences and temporal properties of the repairs in our corpus, we demonstrate that there are, in fact, distinct patterns in the two types of repairs involving the articulation rate of the speech preceding the editing phase and in that following the editing phases (i.e., in both the reparanda and the repair strings).

Keywords: error repair, appropriateness repair, editing phase, temporal properties.

1 Introduction
Spontaneous utterances reflect both a speaker’s thoughts and grammatical competence, and thus a great deal of planning and action must take place in order to produce a message as a grammatical spoken structure (Levelt, 1989). Among other things, the speaker must make many linguistic decisions regarding the morphological, syntactic, phonological and semantic structure of the message, and these, in turn, will lead to the execution of a series of behaviors that ultimately result in the articulation of the intended output. Given the complexity of speech planning and articulation, it is not surprising that we observe a variety of error phenomena, often resulting in disfluencies in spontaneous speech, as well as repairs of about half of these errors (Postma, 2000). The fact that speakers often immediately provide corrections of their errors or inappropriate elements, referred to as self-initiated self-repairs in the conversation-analytic and psycholinguistic literature (e.g., Schegloff et al., 1977; Levelt, 1983; 1989; Swerts, 1998; Nooteboom, 2005; Hartsuiker, 2006; Fox et al., 2010; Pouplier et al., 2014), can be taken as evidence of active monitoring of speech production for grammaticality at all linguistic levels, as well as for utterance
appropriateness (e.g., Levelt, 1983, 1989; Wheeldon & Levelt, 1995; Postma, 2000; Hartsuiker & Kolk, 2001). Although there is discussion in previous research about different sources for the two types of error phenomena (e.g., Levelt, 1983; Postma, 2000; Hartsuiker & Kolk, 2001), relatively little attention has been paid to their temporal patterns. Plug (2011), however, provides a study of the phonetic relationship between errors and their repairs in a corpus of 83 instances from Dutch conversations. The present study investigates both the distribution and temporal patterns of error repairs and appropriateness repairs in a large corpus of natural Hungarian speech in order to determine similarities and differences between these two types of behaviors. Specifically, we examine recordings of over 11 hours of spontaneous speech produced by 26 young Hungarian speakers and demonstrate that the different types of repairs exhibit different temporal patterns in both the errors themselves and their corrections.

2 Error and Appropriateness Repairs

Before presenting the Hungarian data, we first briefly introduce the assumptions and terminology we adopt in our analysis. We also provide illustrations drawn from our corpus since some of the phenomena observed in Hungarian are not found in more familiar languages, such as English.

2.1 The nature of repair types

We use the terms “reparandum” and “error” here to refer to any portion of an utterance that is unintended by the speaker, as evidenced by a disfluency or interruption in the speech stream. More specifically, we distinguish between “speech repairs” in the case of either ungrammatical or in the actual context semantically incorrect productions and “appropriateness repairs” in the case of productions that are felt by the speaker to be undesirable in some other way. The error may be of any size, type of speech, location, reason for their occurrence, or even whether they are followed by a repair or not (e.g., Dell, 1986; Levelt, 1989; Frisch & Wright, 2002; Postma, 2000). When the speaker repairs what s/he said, we refer to it with the type of repair that occurred (i.e., error repair or appropriateness repair).

Actual speech errors may occur at any linguistic level. For example, at the lexical level, such an error would involve the use of an incorrect lexical item. At the morpho-

---


2 Given the substantial body of research on different aspects of speech errors and repairs, the goal here is not to provide a review of the literature, but to refer to key works that pertain directly to our investigation. The reader may consult many of the items cited throughout this paper for more detailed review of previous research.
syntactic level, we might observe an inflection error, and at the phonological level, a mispronounced segment or syllable. Appropriateness repairs typically arise when an utterance is grammatical, but the speaker feels that there is something s/he would prefer to express differently for some reason. For example, a particular word or expression may be ambiguous, or it may not reflect the appropriate level of terminology (e.g., either too technical or not technical enough), or its use may not be consistent with previous usage in the conversation (Levelt, 1983; Kormos, 1999). Appropriateness might also involve a preference for using different stylistic choices in a particular context (e.g., more colloquial, formal, technical).

Different types of speech errors, those involving ungrammatical productions, along with their repairs, are illustrated with Hungarian examples from our corpus in (1) – (3). Here and below, boldface indicates the words involved in the errors and repairs; an asterisk indicates an ungrammatical or wrong (misretrieved) lexical item in the original utterance.

(1) Lexical errors and repairs
a. *jegyet vagy mi bérletet akarunk venni
   ticket.acc or what monthly pass.acc want.1pl buy.inf
   ‘We want to buy a ticket / monthly pass.’

b. *marketing előadó /őö/ hát marketing gyakornok leszek
   marketing lecturer /et/ well marketing assistant be.future.1sg
   ‘I will be a marketing lecturer / assistant.’

c. földi gyakorlatok voltak *limitálva /őö/ szimulálva
   ground exercises were *limited /őö/ simulated
   ‘ground exercises were limited /őö/ simulated’

In these cases, the problem involves lexical retrieval in that the speaker initially retrieved the wrong word from a set of semantically related lexical items. In example (1a), the word jegy refers to a ticket that can be used for only one trip, and it is a distinct lexical item from the word used for a pass that may be used freely during a month, bérlet.

In (2a), the speaker first uses the wrong case suffix, the inessive –ban (which is a grammatical error in Hungarian), and then replaces it with the illative –ba. Note that the verb tense in the repair is changed as well. In (2b), the speaker also initially uses an incorrect case and then replaces it with the correct one; however, here it is the base form that changes, as opposed to the suffix. The person-number suffix following the base, which expresses the case, remains the same.
(2) Morpho-syntactic errors and repairs

a. *Kínában is eljut /silent pause/ Kínába is eljutott
   China.iness also get.3sg /silent pause/ China.illat also get.past.3sg
   ‘He got as far as China.iness / China.illat’

b. *nekem vagy engem sem fertőz
   dat.1sg or acc.1sg nor infect.3sg
   ‘It does not infect to me / me, either.’

(3) Phonological errors and repairs

a. *érdeklődösz /silent pause/ érdeklődsz
   in[ø]quire.2sg /silent pause/ inquire.2sg
   ‘You in[ø]quire / inquire.’

b. *ez egy probo- /silent pause/ probléma az egész osztálynak
   this a probo- /silent pause/ problem the whole class.dat
   ‘This is a probo- / problem for the whole class.’

Both examples contain phonological errors, however, in the first one the full word is uttered since the error appears in the last syllable: a vowel inserted between the last two consonants, [d] and the digraph “sz” which represents [s]. In the second case, the speaker recognizes the erroneous introduction of [ø] before the word is finished, and interrupts the production immediately; there is no Hungarian word that begins probo).

The examples in (4) – (5) illustrate appropriateness repairs. Note that the symbol ‘+’ in these cases does not indicate an ungrammatical element; it just signals the use of a less appropriate word. These items are classified into general categories deduced from the nature of the repair, since the original structures do not contain actual errors.

(4) Ambiguity or clarification repairs

a. *tanárnőm /silent pause/ földrajztanárnőm mondta
   teacher.gen.1sg /silent pause/ geography.teacher.gen.1sg say.past.3sg
egyszer
   once
   ‘My teacher / geography teacher told me once.’

b. mikor bekerültem +középiskolába +gimmázumba
   when got.1 singular high school.illat secondary grammar school.illat
   ‘when I got into high school / secondary grammar school’

In (4a), the speaker first uses the generic word for ‘teacher’, but seems to find it too vague or ambiguous, and thus stops and introduces a more precise compound word, földrajztanárnő ‘geography teacher,’ in its place. In (4b), the speaker uses the word

3 The abbreviations “iness” and “illat” stand for the inessive and illative cases, respectively. Other case abbreviations used in this paper are the more usual “dat” (= dative), “acc” (= accusative), “gen” (= genitive), “abl” (= ablative).
‘high school’ which she thinks is too general, and thus clarifies her meaning with the specification *gimnázium*.

In (5a), the appropriateness repair indicates a preference for a more scientific term in this particular situation as the speaker replaces the general word, “cousinhood”, with the more technical expression, “genetic similarity.” In (5b), the word *ázik* ‘soak’ accurately portrays the speaker’s meaning; however, it is replaced, following the interrupting expression *vagy* ‘or’ with the more professional term *vizesedik*, as the speaker determines this to be better suited to the context. The example in (5c) shows the speaker replacing the somewhat slang expression that uses the word ‘crease’ to refer to a difficult situation with a more formal word for ‘attack’ since this seems more appropriate for the discussion of a historical event.

(5) Terminology / better word choice repairs
   a. nagyobb rügyeket növesztettek akik *rokon kapcsolatot*
      larger sprouts.acc grow. Past.3pl those who *cousinhood*
      sőval *genetikai hasonlóságot* mutattak
      that is *genetic similarity* showed
      ‘larger sprouts grew (when they) showed “cousinhood”’ that is “genetic similarity”
   b. *ázik* vagy *vizesedik* a fal
      soak.3sg or get wet.3sg the wall
      ‘The wall is getting soaked / water-drenched.’
   c. bírták a *gyűrűdést* bírták a *támadást*
      can.take.past.3pl the crease.acc can.take.past.3pl the attack.acc
      a várvédők
      the fortress defender.pl
      ‘The defenders of the fortress were able to withstand the crease / withstand the attack.’

In addition to the appropriateness repairs just seen, where the issue seems to be a matter of terminology, there are also repairs that appear to be in response to a desire for more discourse cohesion or a certain discourse style, as seen in (6).

(6) Discourse repairs
   a. ők voltak a *legjobb* /silent pause/ *legvitézebb*
      they be.past.3pl good.superlative /silent pause/ valiant.superlative
      katonák
      soldier.pl
      ‘They were the best /silent pause/ most valiant soldiers.’
   b. a faluban azt *beszélik / mesélík* hogy
      the village.iness that speak.pl.3. / tell about that
      ‘In the village they *speak / tell about* that.’

In (6a), the speaker first uses a general word “best,” but then selects a different word to emphasize the bravery of the soldiers, more in keeping with the content of the
narrative (i.e., description of a battle). In the sentence in (6b), the speaker replaces the word ‘speak’ with the word ‘tell about’ since the latter is more consistent with the fact that the story she is about to tell is a local tale.

As can be seen from the different types of examples, while it is generally fairly easy to distinguish among the three types of actual errors shown in (1) – (3), it is not always possible to assess the different types of appropriateness errors and repairs shown in (4) – (6) (e.g., Levelt 1983). Thus, our classification is based on the most likely interpretation of the difference between the repair and the reparandum, but it is recognized that in some cases there may be overlap (e.g., a word used for more precision may also be a more technical term preferred in a given context). Given this potential uncertainty, the various types of appropriateness repairs are combined for statistical analysis below.

2.2 Repair process

Levelt (1983) found that appropriateness repairs were often delayed until the end of a word; however, there is generally less information available about the interruption and temporal patterns of appropriateness repairs than repairs of actual speech errors (e.g., Blackmer & Mitton, 1991; Hartsuiker & Kolk, 2001; Plug, 2011). Thus, more systematic data are needed to fully confirm any differences.

Regardless of the type of error, when a repair process arises in speech, we can identify three distinct segments in relation to the interruption point. The reparandum is the string of speech prior to the interruption, specifically the part that contains the error identified by the speaker in the process of self-monitoring. The repair is the continuation of the utterance that contains the correction or modification the speaker wishes to introduce (see among others, Levelt, 1983; 1989; Lickley & Bard, 1996; Roelofs, 2004; Benkenstein & Simpson, 2003; Slevc & Ferreira, 2006; Nooteboom & Quené, 2008). Between the reparandum and the repair, from the interruption point to the onset of the repair, is the cutoff-to-repair interval, or editing term or editing phase (e.g., Levelt, 1989; Hartsuiker & Kolk, 2001). In some cases, the editing phase may also be absent, with the repair immediately following the reparandum (Levelt, 1983). The portions of an error-repair structure are illustrated in (7) with an example from the Hungarian corpus.

(7) Error-Repair Structure

```
<table>
<thead>
<tr>
<th>reparandum</th>
<th>editing phase</th>
<th>repair</th>
</tr>
</thead>
<tbody>
<tr>
<td>*mi a *rev-öö</td>
<td></td>
<td>rövidítése</td>
</tr>
<tr>
<td>what is its ibb-um</td>
<td></td>
<td>abbreviation</td>
</tr>
<tr>
<td>‘What is its ibb-/um/ abbreviation?’</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
```

As can be seen, the interruption takes place after the first syllable of the word for ‘abbreviation’, indicated in the translation by the truncated form ibb- (i.e., instead of abb-). In this case, the speaker detected a phonological error, and interrupted the utterance immediately following the mispronunciation of the vowel in the first
syllable: [ɛ] instead of the intended [ø]. The interruption is followed by a typical filling sound, a relatively long neutral vowel equivalent to English “um”, and then the utterance resumes with the correct pronunciation of rövidítés ‘abbreviation’.

Given the different types of errors and repairs seen above, and the fact that there are three distinct components of an error and repair sequence, questions arise as to whether there are relationships between the error types and the properties of their repair structures. These questions are addressed in the following sections on the basis of the Hungarian corpus.

3 Hungarian Error Repair Investigation

Interruptions may arise either within a word, resulting in truncation, or after a word has been fully uttered. It has been observed that the former is more prevalent with actual speech errors and the latter with appropriateness errors (Levelt, 1989). What is less clear, however, is whether there are also temporal differences in any of the portions of repair structures that distinguish the two error types. We thus first examine the distribution of different error patterns, and then test four specific hypotheses regarding potential differences between speech errors and repairs vs. appropriateness errors and repairs.

Before examining the details of our error and repair data, we first assess the previous observation about the earlier interruption point in actual speech errors as opposed to appropriateness errors in the Hungarian corpus. We then examine the speech rate of these two portions of the utterances, anticipating that the repair will exhibit a faster speech rate than the reparandum, on the assumption that the material in the repair is already somewhat primed by the reparandum. Finally, on the assumption that more consideration is involved in selecting a different, more desirable, word than repairing a grammatical error in a word being used in a sentence, we examine the temporal patterns of the two error types. Specifically, we anticipate that appropriateness repairs will be more substantial than error repairs, both in terms of the timing of the editing phase and its content.

The following hypotheses are thus formulated:

i. Speech errors will be interrupted before a word is fully uttered more frequently than appropriateness errors.
ii. Repairs will be produced by a faster tempo than the reparanda irrespective of repair type.
iii. The duration of editing phases associated with appropriateness errors will be longer than the editing phases associated with speech errors.
iv. The editing phases associated with appropriateness errors will contain more material (e.g., filler sounds, words) than the editing associated with speech errors.

These hypotheses are tested with the speech corpus described below.
4 Methodology

4.1 Corpus

The recordings analyzed for this investigation are part of a large database of spontaneous Hungarian speech, the BEA corpus (Gósy, 2012). Specifically, we examine the speech of a randomly selected group of 26 speakers (13 F, 13 M) between the ages of 22 and 32 years. All of the speakers are from Budapest and speak standard Hungarian, typical of moderate to high education levels. The recordings were made using a variety of topics in a sound-proof room at the Hungarian Academy of Sciences in Budapest. The interviewer was the same in all cases. In total, the corpus comprises 11.5 hours (F: 5.9 hours; M: 5.6), approximately 26 minutes per speaker.

The first and third authors, both native speakers of Hungarian, identified all of the repairs separately from each other; unrepaired errors were not included in this study. Each repair was then classified as an error repair or appropriateness repair, and in the case of errors, the category of the error was specified (i.e., morpho-syntactic, phonological or lexical). Items that were classified differently by the two native speakers were excluded from consideration (N = 3). Subsequently, the selected items were assessed by 12 PhD students specializing in phonetics or in psycholinguistics. The students listened to the context of each repair and the repair itself, and also saw a written version. For each, they indicated whether it was a speech error or an appropriateness error, and again, in the case of speech errors, they specified the type of error. The students worked at their own pace, and could listen to the items multiple times. Only 2% of the items showed discrepancies with respect to the first coding, and they were discarded, leaving a total of 343 items for analysis.

4.2 Annotation and acoustic analysis

The speech samples used for analysis were annotated manually in Praat (Boersma & Weenink, 2015) with indications for the reparanda or errors, the editing phases if present, and the repairs. All segmentations and annotations were carried out individually by the first and third authors, and found to exhibit inter-rater agreement greater than 98%.

The word boundaries were identified in the waveform signal and spectrographic display in conjunction with auditory assessment. Markers for word boundaries and editing phrases were inserted between acoustically distinct regions in the signal, specifically, at the closure and release of stops, at the onset and offset of other obstruents, and at the onset and offset of voicing, as well as between the first and last glottal pulse in the case of vowels, nasals and approximants following standard acoustic-phonetic criteria (see Turk & Shattuck-Hufnagel, 2000).

For each speech error, the type of error was labeled (morpho-syntactic, phonological or lexical). In addition, extensions of 2-3 words (8-9 syllables) preceding and following the error / repair strings were identified for use in assessing speech rate. For the editing phases, the time interval between the last speech sound of the reparandum and the first speech sound of the repair, further specifications were provided as to whether they contained verbal material, and if so, what it was.
The durations of the three components of the repair structures (i.e., the error, editing phase and repair) were calculated using a Praat script. In addition, the articulation rates of the syllables preceding the reparandum and following the repair were determined as the ratio of speech sounds per second (Laver, 1994). A sample annotation is shown in Figure 1.

Figure 1. Annotated speech sample: ez a /silent pause/ ezek a (= ‘this’ /silent pause/ ‘these’).

The string shown in Figure 1 contains a morpho-syntactic error. The reparandum ez a ‘this’ is a singular form; however, the speaker intended to produce the plural form seen in the repair, ezek a ‘these’. Once the error was recognized, the speaker introduced a silent pause, indicated in the annotation with the small box “□”, and then resumed the utterance with the correct form.

All temporal data were normalized (using z-scores) in order to control for differences in the speakers’ articulation rates, and analyzed statistically using a Generalized Linear Mixed Model (with repeated measures analysis within the model), unless otherwise specified. In the case of the editing phase analysis, the gamma log was added to the GLMM in order to model the (probability) distribution. In all cases, significance was at the 95% confidence level, based on calculations using SPSS 19.0.

5 Results
We present the results pertaining to both the occurrences and temporal patterns of the error and appropriateness repairs. With regard to the temporal patterns, we examine the timing of the editing phases in relation to the nature of the reparanda, as well as the speech rates in both the reparanda and repairs. We consider, moreover, whether the patterns are the same in the two types of repairs, or whether they might exhibit noteworthy interactions and / or trade-offs.
5.1 Distribution of repair types

Of the 343 disfluencies in our corpus, 54.5% (N = 187) were classified as involving speech errors and repairs, and 45.5% (N = 156) were classified appropriateness errors; however, a chi-square test determined that the difference was not significant. As to their frequency, error repairs occurred every 3.6 minutes, while appropriateness repairs occurred every 4.3 minutes.

Examination of the type of errors involved in error repairs revealed 49 phonological errors (26.2%), 63 morpho-syntactic (33.7%) and 75 lexical errors (40.1%), a distribution very similar to that found by Levelt (1983). With regard to the location of the interruption within a word or after its completion, the following distribution was observed. While only 33% of the interruptions with appropriateness repairs occurred before the word had been fully uttered, 59% of the error repairs took place before the word was finished. This pattern suggests that speakers typically react more quickly to actual speech errors in their own speech than they do to cases that only involve a preference among well-formed alternatives.

5.2. Duration of editing phases

Overall, there was a significant difference in the mean duration of the editing phases associated with error repairs (720 ms) and appropriateness repairs (850 ms), ($F(1, 342) = 11.301, p = 0.001$), see Figure 2.

![Figure 2. Duration of editing phases of error repairs and appropriateness repairs (medians and ranges)](image)

We also found temporal differences when we considered the location of the interruption. As expected, longer editing phases occurred following full words (874 ms in error repairs and 915 ms in appropriateness repairs) as opposed to interrupted words (617 ms in error repair and 736 ms in appropriateness repairs) in both types of repairs. These differences were significant (error repairs: $F(1, 186) = 16.599, p = 0.001$;
appropriateness repairs: \( F(1, 155) = 9.594, p = 0.004 \), and a check for gender differences revealed no significant effect.

Not all cases of appropriateness repairs involved long editing phases, however. In fact, in some cases, there were particularly short editing phases, or even a lack of an editing phase, resulting in an immediate repair, as illustrated in (9).

\[
(9) \text{vannak} \quad \text{pozitív} \text{ és} \quad \text{negatív} \quad \text{[0 ms]} \quad \text{jó} \quad \text{és} \quad \text{rossz}
\]

In this case, the speaker first uttered the borrowed (or foreign) words pozitív and negatív, but immediately decided that it was preferable or more appropriate to use the Hungarian words jó ‘good’ and rossz ‘bad,’ these were thus inserted without any editing phase.

The mean duration of editing phases arising with appropriateness repairs was quite long, 850 ms. However, examination of the distribution of these durations revealed an interesting pattern, as seen in Figure 3, where the durations of the editing phases are grouped in 200 ms categories.

\[\begin{array}{c|c|c|c|c|c}
\text{Durations of editing phases (ms)} & 0-200 & 201-400 & 401-600 & 601-800 & \text{more than 801} \\
\hline
\text{error repairs} & 26.2 & 33.1 & 32.7 & 19.2 & 11.5 \\
\text{appropriateness repairs} & 20.8 & 16.6 & 10.2 & 7.7 & 6.9 \\
\end{array}\]

\textbf{Figure 3.} Editing phase durations: Comparison of original durations of error repairs and appropriateness repairs

As can be seen, there are more cases of immediate repairs (0 ms) with actual errors than with appropriateness errors. Of the measurable editing phases (i.e., more than 0 ms), the shortest ones, up to 200 ms, are the most common, and they are equally present with both repair types. There are also minimal differences between the repair types in the ranges up to 600 ms, with the longer editing phases becoming less common. Where a difference arises is with the particularly long editing phases, above
600 ms. In this range, although there are relatively fewer instances, it can be seen that almost all occurred with appropriateness repairs. Thus at the extremes, we see most 0 ms editing phases with error repairs, and the longest editing phases with appropriateness repairs.

When we consider the durational differences in the editing phases associated with the three categories of actual errors, several patterns can be observed, as shown in Table 1.

**Table 1. Durations and standard deviations of editing phases with different error type repairs**

<table>
<thead>
<tr>
<th>Error repair type</th>
<th>Duration of editing phases (ms)</th>
<th>mean value</th>
<th>std. dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>phonological error repairs</td>
<td></td>
<td>548</td>
<td>278</td>
</tr>
<tr>
<td>morpho-syntactic error repairs</td>
<td></td>
<td>743</td>
<td>453</td>
</tr>
<tr>
<td>lexical error repairs</td>
<td></td>
<td>814</td>
<td>441</td>
</tr>
</tbody>
</table>

First, it can be seen that the shortest mean durations of the editing phases arise with phonological errors (548 ms). The durations associated with lexical and morpho-syntactic errors are very similar (743 ms and 814 ms); there is no significant difference between them, but both are significantly longer than the mean editing phase duration associated with phonological errors (548 ms) \(F(2) = 6.940, p = 0.001\). When compared with appropriateness repairs, it can be seen, furthermore, that even the longest editing phases occurring with lexical error repairs are on average shorter than those occurring with appropriateness repairs (814 ms vs. 850 ms), although the difference is not significant. What these findings suggest is that the higher an error occurs in the speech process, the longer it takes to repair it. This confirms earlier findings by Blackmer and Mitton (1991) and van Hest (1996) that conceptual errors are repaired significantly more slowly than lexical or phonological errors.

As far as the content of editing phases is concerned, we observe a good deal of variation. We have thus divided the results into five categories: (i) phase absent (i.e., immediate repairs discussed above: 0 ms duration), (ii) silent pause, (iii) filled pause (e.g., the equivalent of English “er”, “um”), (iv) cue phrase present (e.g., equivalent to English “well”, “I mean”, “that is”, etc.), and (v) combination (e.g., filled pause + cue phrase). As can be seen in Figure 4, the various options for editing phases are not distributed in the same way for error and appropriateness repairs; this difference is significant \(\chi^2 (4) = 25.494, p = 0.001\).

While all options are observed in both error and appropriateness repairs, when actual errors occur, the tendency is to insert minimal content in the editing phase (mostly 0 ms or silence), and when appropriateness repairs occur, the tendency is to insert more

---

4 For discussion of different types and content of editing phases, see among others Jefferson (1974), Fraser (1999), Schourup (1999), Fox Tree and Schrock (2002), Schegloff (2007).
substance. It turns out that there are relatively few filled pauses with no other content (i.e., (iii)) for both types of repairs, the only case of similarity between them.

Figure 4. Percentage of different types of content in editing phases for the two types of self-repair

5.3 Rate of speech of reparanda and repairs

The rate of speech provides another view of the temporal patterns of the two types of repairs under investigation. Figure 5 presents the speech rates for error and appropriateness repairs calculated in terms of the number of segments (i.e., vowels and consonants) per second.

As can be seen, the speech rates were the same for the reparanda, regardless of error type: 12.1 and 12.3 segments per second for error and appropriateness repairs, respectively. The articulation rates of the repair portions, however, were statistically different \( F(1, 342) = 12.172, p = 0.001 \), with error repairs being slower than appropriateness repairs (16.8 and 18.7 segments per second for error and appropriateness repairs, respectively). In addition, there was a significant difference in speech rate between the reparanda and repairs for both repair types (i.e., appropriateness repairs: \( F(1, 185) = 9.400, p = 0.002 \); error repairs: \( F(1, 155) = 8.973, p = 0.001 \)). Thus, although we have not determined the speech rate for all of the utterances, what is clear in our results is that speakers consistently had a faster speech rate following any type of error. This would be consistent with the observation that the disfluencies are, in fact, anticipated by the speaker, resulting in “a lengthening of rhymes or syllables preceding the interruption point” (Shriberg, 2001; Plug, 2011).
Figure 5. Speech rates (segments per second) of reparanda and repairs in error and appropriateness repairs (median and range)

6 Discussion and conclusion

We have presented different types of information from Hungarian regarding a common phenomenon observed in the course of daily conversations – speakers interrupting their own utterances. This takes place when speakers feel the need to modify their spoken utterances either because they have made an actual error or because they would simply prefer to express themselves somewhat differently (e.g., Levelt, 1983, 1989; Shattuck-Hufnagel & Cutler, 1999; Postma, 2000; Benkenstein & Simpson, 2003; Plug, 2011). Specifically, we have investigated to what extent the properties of the interruptions for error and for appropriateness repairs are similar or different. The findings about appropriateness repairs are particularly important since this type of interruption has been less studied in the literature than repairs following actual errors. This is most likely due to the fact that appropriateness repairs are somewhat more difficult to characterize since no actual speech mistake has been made, and they are thus more open to interpretation.

First, it was noted that error repairs and appropriateness repairs were not evenly distributed in the analyzed material the latter occurred less frequently, as also reported elsewhere in the literature (Levelt, 1983; Plug, 2011). Several possible explanations seem to be available for these differences. First, there were some speakers who introduced overt error repairs, but who did not seem to be particularly concerned about repairing words or expressions for clarity or coherence or for stylistic purposes (“error repairers”, in our case 13 speakers). By contrast, there were other speakers who paid more attention to the appropriateness of their messages, and somewhat less attention to real errors (“appropriateness repairers”, in our case 10 speakers). In this case, it seemed that the speakers assumed that the listener could correct the errors s/he heard based on their mutual language knowledge, however, they were concerned that their thoughts were properly expressed. No preference was found with 3 speakers in our material.
Similarly to Levelt (1983; 1989), we found that speakers halted production of a word before finishing it more frequently with real errors (in our case: 59%) as opposed to appropriateness errors (in our case: 33%). The assumption of an inner monitoring mechanism suggests that errors may sometimes be detected and intercepted before they are articulated (Hartsuiker & Kolk, 2001). It seemed that the case was slightly different with the appropriateness repairs in this respect, since more appropriateness repairs occurred after finishing the production of the whole word.

A goal of this study was to determine whether the basic difference between error and appropriateness repairs was confirmed by measured data and statistical analysis. As expected, the durations of editing phases in appropriateness repairs turned out to be longer in our material than those in error repairs. This is consistent with the view that the sources of the problems in the two cases are different (Levelt, 1983; 1989; Postma, 2000). That is, while actual errors originate at the levels of the speech planning mechanism associated with grammatical or phonological encoding or word retrieval from the mental lexicon, appropriateness repairs appear to involve higher levels associated with the formulation of concepts and the selection of the necessary lexemes from the mental lexicon. In the former case, the speaker accidentally retrieves an erroneous word or segment, while in the latter case, the selected word is correct in the given context but the speaker, upon consideration, deems it inadequate or inappropriate for some reason. Thus, in the latter case, the speaker must make conceptual comparisons among words or phrases in appropriateness repairs, as opposed to just replacing an erroneous sound, word or expression with the correct one. Indeed, our empirical data revealed that appropriateness repairs needed longer editing phases in more instances than error repairs. That is, when the speaker had several (competing) ways to formulate a given thought (semantically and/or syntactically), longer hesitations were produced when repairing the undesirable utterance.

It is noteworthy that there was a relatively large number of absent (0 ms) or extremely short editing phases, although Blackmer and Mitton (1991) reported that in almost 50% of the errors they examined, the duration of the cutoff to-repair interval was less than 100 ms. There seemed to be several possible reasons for this behavior. As noted, with error repairs, it is often only part of a word that needs to be adjusted. In the case of appropriateness repairs, however, the short editing phase (although less common) could indicate that both of the competing concepts and structures may have already been activated, making the replacement immediately available for use in place of the item that had just been uttered. In some cases, the short editing phases might also be due to a relatively small number of options available in the language. When this is the situation, if there are two words or expressions that are equally appropriate for a given concept, both lemmas might be selected and undergo phonological activation, allowing immediate replacement if the one selected first is deemed less desirable (Jescheniak & Schriefers, 1999).

With regard to the actual content of the editing phases, it was found, perhaps unsurprisingly, that appropriateness repairs tended to include more “substance” or content words, specifically cue phrases, than actual error repairs. It is possible that
this is evidence that the speaker wishes to continue communicating with the listener even while searching for a more appropriate word or expression; however, it is also possible that simply reflects the fact that the speaker needs to take more time to do the search, and is just filling the gap. The use of items such as “that is” or “you know”, may even serve to draw the listener’s attention to the fact that speaker is searching for a “better” word. In this case, the content during the editing phase may serve a pragmatic role in the broader discourse structure – that of relating the earlier and later parts of the utterance, providing a link between the problematic string and its repair (e.g., Fraser, 1999; Schourup, 1999; Schegloff, 2007). By contrast, when a real error is made, since this is equally obvious for the speaker and listener, the speaker may not feel as inclined to send the “additional message” that a replacement is coming.

It was also noted that the repairs were produced at a significantly faster speech rate than the reparanda irrespective of repair type. Interestingly, however, the speech rate in error repairs was significantly slower than in appropriateness repairs. One possible reason for this is that in the case of error repairs, the speaker may unconsciously be attempting to make the correction more intelligible for the listener by using a slower articulation rate. Since appropriateness repairs are not providing new information, only refining information already present, the speaker may feel that the clarity of articulation is secondary to the clarity of the concept. Taken together, our findings revealed a possible pattern of equivalence or compensation, whereby a longer editing phase was combined with a subsequent faster speech rate in appropriateness repairs, and the opposite pattern was found with error repairs. Although the details were somewhat different, this compensatory pattern may be fundamentally similar to that observed by Levelt and Cutler (1983) in Dutch self-repairs. Specifically, while the majority of error repairs exhibited a pitch accent on the repair stretch, the majority of appropriateness repairs lacked a pitch accent. Since the error repairs in Dutch had a faster speech rate, it is possible that they did not need additional time since the pitch accent was clear enough. The unaccented appropriateness repairs, however, were compensated by their slow speech rate. Ultimately, similar types of comparisons of speech rate and other properties of error and appropriateness repairs must be conducted across different types of languages in order to gain further insight into the different aspects of speech repairs.

Acknowledgement
This work was supported by the OTKA project No. 108762.

References


IZÉ ‘STUFF’:
A CORPUS-BASED STUDY
ON A CURIOUS HUNGARIAN FILLER

Dorottya Gyarmathy
Dept. of Phonetics, Research Institute for Linguistics,
Hungarian Academy of Sciences, Budapest, Hungary
e-mail: gyarmathy.dorottya@nytud.mta.hu

Abstract
Fillers occurring in spontaneous speech are known as specific words or structures that have no propositional meaning. Speakers frequently use them as a strategy to get over their difficulties during speech production. Since fillers do not interrupt the flow of speech, they are hardly noticeable by the listeners. Fillers universally come from function words, e.g. English well, German also, Hungarian tehát, így.

IZÉ is a common filler word in Hungarian spontaneous speech. In the case of difficulties in lexical access this sound sequence can substitute any word of various word classes due to the fact that it can be inflected. Fillers are not repairable among disfluency phenomena; however, izé can be repaired in certain cases. It can be used in two functions: (i) simple filler and (ii) substitutive word.

The aim of the present research is to analyze the filler word izé in a large corpus of Hungarian spontaneous speech. The corpus consists of narratives of 94 adults (aged between 20 and 76 years) from the BEA Hungarian Spontaneous Speech Database. Acoustic analysis was carried out by means of Praat 5.3.19 while the statistical analysis was carried out using SPSS 13.0 software.

The following research question was addressed: (i) Can the functions of izé be distinguished based on their acoustic-phonetic characteristics? Our hypotheses was the following: filler and substitutive izé will show different acoustic parameters. The results confirmed the function-dependent realisations of izé in spontaneous speech.

Keywords: disfluencies, editing phase, repair, filler words.

1 Introduction
Fillers are disfluency markers that usually suggest the speaker’s uncertainty about what to say next. However, the related terminology is not used consistently in the literature. Some researchers use the category ‘filler’ to refer to cases in which the speaker only articulates incongruous speech sounds due to some uncertainty in speech planning (cf. filled pause), as well as to cases in which whole words, word sequences, or expressions are uttered for the same reason (Levelt, 1989; Glücksmannová, 2008; Corley & Stewart, 2008; etc.). This is due to the identical function of the two phenomena: the speaker plays for time in both cases in order for speech planning
processes to be able to take place appropriately. However, as the literature reveals, another group of researchers (cf. Brennan & Williams, 1995; Bortfeld et al., 2001; Gósy, 2002; Horváth, 2004; Gyarmathy, 2010) clearly separate the two phenomena, for formal as well as functional reasons (the terminological diversity will not be further discussed here).

In the case of filled pauses, the speaker usually utters a single sound or a sound sequence. For instance, in English, filled pauses are usually realised as uh or um (e.g., Clark, 1994; Clark & Fox Tree, 2002; Roberts et al., 2009), in Portuguese, they can be uum, [v], or [ə] (Veiga et al., 2011), and in Swedish they are represented as eh (Horne et al. 2003). In Hungarian, they are typically represented as ò [ø], [ɔ], m [m], and less frequently as òh [øh], ôm [om], ôhm [øhm] (Horváth, 2010; Gósy et al., 2013). On the other hand, we can define fillers as (formally) lexemes of the given language (e.g. German also, English well, you know, Hungarian szóval, tehát) that do not organically fit into the utterance they occur in, and do not contribute anything to its contents; their functions are similar to those of filled pauses. Their occurrence in the given context is not required either syntactically or semantically (Glücksmannová, 2008). In most languages, it is usually function words that occur as fillers, although content words can also occasionally turn into fillers like you know in English or úgymond ‘so to say’, azsongya ‘it says’ in Hungarian. In the process of being converted into a filler in spontaneous speech, a lexeme loses its usual content, its original function changes, but the old function may also carry on being used along with the new one. Hence, the process can be seen as a kind of ongoing language change or functional split that takes place as an increasing number of speakers begin to use the given lexeme in its new function (Bybee, 2001; Pierrehumbert, 2001).

Speakers often use fillers as a strategy for problem solving (e.g. Brennan & Williams, 1995; Gósy, 2005), given that their use does not interrupt the flow of communication, whereby their inorganic occurrence is less conspicuous for the listener. It has to be emphasised, however, that fillers may not only occur in spontaneous speech in the time-gaining function referred to above but also as discourse markers. The latter are expressions that link one discourse segment to another and primarily come from the syntactic classes of conjunctions, adverbials, or prepositional phrases. Their main function is to indicate the relationship between segment S₂ (that they introduce) and the previous segment, S₁ (Fraser, 1999). The most conspicuous property of the class of discourse markers is their heterogeneity; the members of the class come from a diversity of part-of-speech classes. The simpler ones can be traced back to verbs, adverbs, conjunctions, interjections, or prepositions; the more complex ones may include whole constructions, or even clauses. Their single shared property is that they do not have uniform formal properties; hence, they can only be defined on a functional basis. Semantically, it is typical of them that they are not relevant for the truth conditions of the utterance and do not influence its propositional content, but they have emotional and expressive functions (Jucker, 1993). Discourse markers do not merely connect discourse segments and help maintain the coherence of the text (Lenk, 1998) but they also have a role in boundary
marking (cf. Jucker, 1993; Fraser, 1999; Fox Tree & Schrock, 2002; Dér, 2005, 2008; Markó & Dér, 2008). In conversations, they indicate the speaker’s intention to go on speaking (Schachter et al., 1991; Bortfeld et al., 2001), or help the interlocutors to coordinate turn taking and organise the conversation in general (Clark, 1994).

Thus, fillers can be seen as discourse markers in the sense that they indicate a temporary disharmony of the speech production process for the listener. Nevertheless, in our view, despite the similarity between discourse markers and fillers, the latter constitute an independent, well-defined category, and cannot be taken to be discourse markers functionally, either (see also Jucker, 1993; Horváth, 2009). In a pragmatic perspective, the functions of fillers, like indication of intention to speak or coordination of turn taking, coincide with discourse markers’ roles.

Formally, then, filled pauses and fillers can be clearly distinguished from one another, and both categories are characterised by language specific properties. There is, however, also a third category whose function is identical to that of filled pauses and fillers, yet formally, it differs from both. This formal difference, in turn, makes it possible for language users to employ items of this third category for resolving disharmony in a special sense. In Hungarian, the sequence *izé* [ize:] occurs as a filler. Its etymology is multiply uncertain; its origin, primary syntactic function, and meaning are unknown. It may be a compound word or a derived word. Its earliest confirmation came in 1604. The root *iz*- may be of Proto-Ugric or even Proto-Finn-Ugric origin, albeit its alleged association either with the noun *iz* ‘thrush, cancer’ or with the pronoun *ez* ‘this’ is unfounded (Benkő, 1967–1984). Handbooks of good Hungarian usage usually take the use of *izé* to be a sign of linguistic negligence and in general discuss it under the rubric of “speech stuffing” or “slipshod style” (Grétsy & Kovalovszky, 1980). The earliest scholarly comment with respect to *izé* comes in 1645. István Geleji Katona writes: “The word *izé* is quasi transcendent in our language: anything whose name does not quickly come to mind is called *izé* by Hungarians” (cited by Prohászka, 1957:245). Similarly, papers dating from the late 19th century (or later) define this word as a substitute for the name of something or somebody that is used by the speaker when the intended expression does not immediately come to mind or she wants to avoid using it deliberately (cf. Simonyi, 1876; Szinnyei, 1912; Prohászka, 1957; Juhász, 1959). Pusztai (2003) provides a nominal and a “sentence-word” interpretation of *izé* as follows: “I. noun <a word used to replace a noun that we cannot recall or do not want to utter>; II. sw <an idle word used to surmount a difficulty in memory or speech production>”.

The earliest collection of examples of the uses of the word *izé* (Gyarmathi, 1794, cited by Prohászka, 1957) shows that *izé* was generally used already in the 18th century; this is also demonstrated by the pronouns, adjectives, and verbs derived from it: *Össze izélt, tele izélt a gatyáját* ‘He made a mess in/all over his pants’; *Ugyan megizélték ökemlét. Mizélte? Mit csinált néki?* ‘He pestered the guy. Pestered? What did he do to him?’; *Láttam az izéjét. Mizéjét? Szégyenlem megmondani.‘I saw his thing. His what? I am ashamed to tell’; *Szörnyű izés ember* ‘He is a terribly finicky man’ (the examples are from Sámuel Gyarmathy’s collection, cited by Prohászka,
Literary examples can also be found: “– A király afféle izé. – Mízé, uram? – Semmízé.” [– The King is a thing. – A thing, my lord? – Of nothing.] (Shakespeare: Hamlet IV/2 – Hungarian translation by János Arany); “Férjém kérdezed? Izé… tudod már… – Kizé? Miféle?” [You’re asking about my husband? Er… you know… – Who’s he? What kind?] (Aristophanes: Thesmophoriazusae – Hungarian translation by János Arany); “Az özvegy menyecske volt nagy kínban, nem győzte paskolni a kezeiket. »Ugyan hagyjanak békét. Ne izéljenek hát stb.«” [The young widow was embarrassed, she went on slapping their hands. “Come on, leave me alone. Don’t do that to me“ etc.] (Mikszáth: Új Zrínyiász).

Of the Hungarian fillers, izé used to be one of the most conspicuous “stuffing items” of spontaneous utterances for quite some time. This was because it was able to replace practically any word in cases where the speaker had some temporary difficulty in organising her speech (Fabulya, 2007). Due to its multifarious suffixability, izé can stand for a verb, a noun, an adjective, a pronoun, an adverb, etc. This behaviour is due to the agglutinative character of Hungarian; this can also account for the fact that it is only in Basque, another agglutinating language, that a filler of similar behaviour has been found to exist. Similarly to izé, Basque zera can take practically any suffix, and thus can replace words of any part-of-speech affiliation. The only difference is that zera, unlike izé, has a meaning and a syntactic function of its own; it is a demonstrative pronoun meaning ‘that’ (Urizar & Samuel, 2014).

The filler izé has a peculiar place among disfluency phenomena, too. In most cases, fillers are among the non-reparable disfluency markers, but izé is an exception in that respect, too. This word has a double function: it can occur in speech either as a filler proper or as a replacement for certain words or phrases. In the latter case, it becomes reparable. Using Fabulya’s (2007) terminology, the former will be called “filler izé”, and the latter will be called “proxy izé” here. Proxy izé invariably occurs in a previously constructed syntactic structure as a temporary or permanent replacement for a word or an expression; one of its most important properties is that it often occurs suffixed, even case-marked, completely fitting into the given syntactic structure. Filler izé, on the other hand, is invariably unsuffixed, and does not fit into the syntactic structure of the given utterance at all. Furthermore, it is also possible to assume a discourse marking function for izé, in a conversation for instance, but its identification is not easy and thus in the present analysis the possibility will merely be mentioned but not discussed in detail.

In sum, the two functions of izé are clearly distinct; the question arises whether they also differ in their articulation and acoustic structure. In order to be able to answer that question, we have analysed izé tokens in large numbers. We have extracted those tokens from a spontaneous speech corpus and have analysed them in terms of fundamental frequency, intensity structure, and temporal relationships. Our hypothesis is that the functional bifurcation of izé can be verified in acoustic parameters, too.
2 Material, methods, and participants

For the purposes of the present paper, we used 94 recordings (narratives, comments, interpreted speech, conversations) of the BEA Hungarian spontaneous speech database (Gósy, 2008) and analysed all occurrences of the lexeme izé, with its possible function dependent pronunciation variants. The spontaneous speech material we analysed was 75 hours and 5 minutes long and came from monolingual native speakers of Standard Budapest Hungarian; there were 55 female and 39 male speakers, their average age was 38.06 years. It was in 62.76% of the material. However, in the speech of 59 participants, no occurrence of izé was found at all. The remaining 35 speakers (20 female and 15 male ones) produced 5 hours, 43 minutes, and 47 seconds of speech. We found a total of 141 instances of izé (4 instances per speaker, on average), that is, one instance every 2.5 minutes. It can be seen that this is a rarely occurring phenomenon, which can be related to its stigmatised nature. Each occurrence of izé was annotated by Praat, version 5.1.34 (Boersma & Weenink, 2009). We determined their duration, average $f_0$, and – where applicable – the duration of their editing phases, that is, the time spent repairing them. The latter duration was calculated as stretching from the end of the last sound segment of the lexeme izé to the beginning of the first sound segment of the intended word (just like in cases of false word access, cf. Horváth & Gyarmathy, 2010); see Figure 1.

![Figure 1. An example for the editing phase in the case of izé](image)

In all instances of izé, we measured fundamental frequency such that we had the Praat program record average $f_0$ for every 10 ms, along with continuous auditive and visual inspection. Of the values obtained, we filtered out extreme outliers assuming that they were measurement errors, and then we calculated average fundamental frequency for the whole duration of izé. We determined the duration and structure of pauses preceding instances of izé, and we analysed their interrelations with the other parameters. For statistical analysis, we used one-way ANOVA, nonparametric Kruskal-Wallis tests and Tukey’s post-hoc tests at the 95% level of significance. For correlation analysis, we calculated Pearson’s correlation coefficients using SPSS 13.0.
3 Results

We analysed the 141 instances of ízé in terms of filler ízé vs. proxy ízé. In the case of the latter, we determined which part-of-speech category the replaced item belonged to, and whether this correlated in any way with the duration and structure of the editing phase. The occurrence of ízé in spontaneous speech primarily indicates a problem in lexical access. The speaker’s internal monitor signals the problem already during speech planning, before the word ízé is uttered. This is often shown by a pause preceding the word ízé.

In the course of categorisation, we performed context analysis, following the criteria defined by Fabulya (2007). A larger portion of all occurrences (61.7%) were proxy ízé; for instance: kétézöööben csináltuk az ízét a Csárdáskirálynől ‘in 2005, we did stuff, The Gipsy Princess’; a személyes élmény hát az, hogy te nem az ízével a fapadosca utazód ‘the personal experience is that you did not fly by stuff, a budget carrier’. A smaller portion (38.3%) belonged to the category of filler ízé; for instance: nekünk az is újdonság volt tudod, hogy úgy nyílik valami, tehát így ízé ‘for us, it was a new thing, you know, that something opens like, so kinda stuff’; hogyha kell, akkor megbuktatok negyven embert meg ízé, de ne szívassuk egymást ‘if I have to, I’ll reject forty people and stuff, but let’s not shaft each other’.

After we established the two groups, we first analysed the temporal properties of each occurrences of ízé. Comparing the average durations of instances in the two categories, we found a difference (Figure 2). Proxy ízé tokens were in general realised as 260.26 ms long, whereas filler ízé tokens were longer, 296.07 ms on average. In terms of statistical analysis, the difference is significant: one-way ANOVA: \( F(1, 140) = 6.060; p < 0.015 \). The different duration of the two types of ízé may be explained by the fact that in the case of proxy ízé, the speaker has to resolve a temporary difficulty of word access, and the retrieval of the intended word from the mental lexicon already happens while the word ízé is being pronounced. If lexical access remains inhibited for some reason, the speaker may still assume that her partner successfully understood what she was going to say, thanks to the context on the one hand, and to the fact that not only speakers but also listeners run certain intentional or unintentional correction mechanisms, on the other (Levelt, 1989). Thus, ízé is produced with a shorter duration in such cases. In the case of filler ízé, the longer duration may be explained by the fact that the speaker does not simply wish to call the listener’s attention to a current problem of speech planning; rather, she wishes to indicate her speech intention: the fact that she does not want to yield the floor just yet. This latter function is characterised by a certain amount of intentionality.

We analysed average fundamental frequency for the full duration of ízé occurrences, for both genders. Female speakers’ proxy ízé tokens were produced at 151 Hz on average (standard deviation: 30.36 Hz; minimum: 96.18 Hz; maximum: 231.75 Hz); their filler ízé tokens were some 20 Hz higher (Figure 3): 172.19 Hz (SD = 36.62 Hz; min.: 96.44 Hz; max.: 253.24 Hz). Statistical analysis confirmed the difference: one-way ANOVA: \( F(1, 51) = 5.195; p < 0.027 \). With male speakers, the difference was not significant: their proxy ízé tokens were produced at 121.41 Hz (SD = 33.96 Hz), and their filler ízé tokens
at 130.98 Hz (SD = 31.58 Hz) (Figure 4). Despite the approximately 10 Hz difference between the average values, the difference is not significant. In the case of proxy izé tokens, lower pitch may be due to the speaker concentrating on lexical selection and therefore paying less attention to the deliberate control over her articulatory processes. Another possibility is that she tries to let her speech partner know, by nonverbal means, that she is having difficulty in word finding; thus the partner is given the chance to stick in a word helpfully, without actually taking the floor. The relatively higher pitch of filler izé tokens may have a kind of discourse organising function as well. In particular, the speaker signals her intention to carry on speaking; also, by uttering izé, she gains some time for giving linguistic form to further utterance units and for the listener to process the foregoing units.

Given that men and women have vocal cords of different length (as part of their gender-based anatomical dissimilarity), and consequently their average fundamental frequency values also differ, we had to study the results separately for men and women. Furthermore, individual differences across speakers persisted. To eliminate these discrepancies, we have normalised $f_0$ values in izé tokens to the mean and to the standard deviation. Normalised $f_0$ means that we provide $f_0$ values measured in the individual izé tokens relative to the average fundamental frequency of the given speaker, with a negative or positive sign indicating whether they were produced at a lower or higher pitch than the given speaker’s own mean $f_0$ value. The results show that proxy izé tokens in general exhibit pitch values that are 0.4976 Hz lower (SD = 1.25 Hz; min.: –4.64; max.: 4.24), while filler izé tokens are 0.1195 Hz higher (SD = 1.44 Hz; min: –4.47; max: 5.54) than the speakers’ average $f_0$ (Figure 5). This difference was statistically confirmed: a one-way ANOVA showed that the difference

![Figure 2. Durations of instances of filler izé and proxy izé (medians and standard deviations)](image)
is significant: $F(1, 131) = 6.789; \ p < 0.010$. That is, speakers are presumably aware of the functional difference between the two types of ızé, as their $f_0$ values suggest.

**Figure 3.** Fundamental frequency data of female speakers in the two types of ızé (medians and standard deviations)

**Figure 4.** Fundamental frequency data of male speakers in the two types of ızé (medians and standard deviations)

Speech planning problems are often only indicated at the surface by silent or filled pauses. However, if repair processes are not successful within that time span, the error itself occurs in the utterance, too. From the length and structure of pauses that precede ızé tokens, we can infer the severity of the problem that the speaker currently faces. In the 141 ızé occurrences that we have analysed, speakers paused for 142.01 ms on
average, but the data were scattered between 0 ms and 5737 ms. We were able to set up five different types of cases (Figure 6).

(i) We have taken the first group to include cases where no pause preceded ْزة, or the speaker uttered a non-lexemic sequence. The latter case occurred in a single instance: *amit lehet még azt is gyorsan el j* (154 ms) ْزةنْهك. ‘what can be, we quickly j...stuff it away’. In this case, the slightly prolonged [j] sound (cf. Gósy, 2004) may result from an anticipatory error, or else it may be a case of false start. The context fails to make it possible to find this out; but the occurrence of ْزة makes it certain that the speaker was unable to resolve her current difficulty of lexical access in the time span available. Most instances, 76.59%, belonged in this group. The fact that no pause precedes the occurrence of ْزة in most cases suggests that whenever speakers face a planning problem they insert ْزة automatically, as a cliché, in order to gain some time for repairing the error.

(ii) The second group was made up of cases where a silent pause preceded ْزة (11.35%): *nem az hogy inkább kiment volna a szembe a* [silent pause] (349 ms) ْزةنْهك *hanem szembejött a motoros azt páff* ‘it’s not that he would go out to the opposite the ... stuff but the motorcyclist came from the other direction and bang!’’. These pauses were 298.63 ms long on average (41–1082 ms; standard deviation: 261.95 ms).

(iii) In the third group (3.55%), ْزة was preceded by a filled pause (marked by ْدْدْد in the paper) that was somewhat longer, 460 ms on average (71–1111 ms; standard deviation: 499.94 ms), e.g.: *hát olyan helyet kell találni ahol sok fa nincs hogy ne tudjanak ْدْدْد* (102 ms) ْزة *építeni hajót* ‘well we got to find a place where there aren’t many trees so that they can’t build er stuff a ship’.

(iv) The fourth category contains combinations of silent pauses with filled pauses (7.80%); these were 1018.55 ms long on average (87–5737 ms; standard deviation: 1596.27 ms), e.g.: *aki a* [silent pause] (380 ms) ْدْدْد (116 ms) (141 ms) [silent pause] ْزةنْهك *ném a Szé-asszem a Szegedi Operánál volt* ‘who was ... er ... in stuff, no, at Sze- I guess at Szeged Opera’.

(v) The last category (combination of silent pause and/or filled pause and linguistic material) contained a single instance: *kiderült hogy engem nem az irodalom ami leköt* [silent pause] (155 ms) *hát* (752 ms) ْدْدْد (113 ms) [silent pause] (567 ms) ْزة *önfejű hülye ember vagyok* ‘it turned out that I am not interested in literature, I’m ... well er ... stuff a stubborn stupid person’. In this case, the speaker spent a total of 1587 ms looking for the appropriate word. Not having found it, he substituted ْزة for it, and then finally he was successful in finding the appropriate lexeme(s). Statistical analysis revealed that there are significant differences between the groups. The results of nonparametric Kruskal-Wallis test ($\chi^2 = 133.095; \ p < 0.001$) confirmed that the structure of preceding pauses determines their duration.
We also analysed pauses preceding izé separately for proxy izé and filler izé items. We found a difference in the mean length of pauses preceding the two types of izé: proxy izé tokens were preceded by 147.89 ms pauses on average (SD = 643.159 ms), while filler izé tokens were preceded by 132.23 ms ones (SD = 313.461 ms). The difference, however, is not statistically significant.

Before 71.69% of the filler izé tokens, the speakers did not produce any pause between the preceding word and izé. In 15.09% of the cases, izé was preceded by a silent pause that was 259.50 ms long on average (41–481 ms; standard deviation: 177.97 ms). In
9.43%, a combination of silent and filled pause was found lasting for 643.80 ms (159–1164 ms; standard deviation: 408.90 ms). We found a single instance (1.88%) where the preceding pause was a filled one (el is dobtad meg mmm (126 ms) izé akkor vissza megint újból tudatosítottad magadba ‘you threw it away and mmm stuff then back again once more you brought it home to yourself’), and another single instance where it was a combination of silent and filled pause and linguistic material (1587 ms, see above for the example). Given that four of the five categories contained a very small number of items, no statistical analysis was performed for this group of cases.

In the case of proxy izé tokens, only four categories of preceding pauses occurred (Figure 7). Here again, the largest category was the one where no surface pause preceded izé (0 ms or linguistic material); 79.54% of the occurrences belong here (0–154 ms; mean length: 2.2 ms; standard deviation: 18.41 ms). In 9.09% of the cases, preceding pauses were silent ones, 337.75 ms on average (58–1082 ms, standard deviation: 334.46 ms); while filled pauses (4.54%) lasted some 200 ms longer on average: 543.50 ms (71–1111 ms, standard deviation: 535.50 ms). The longest average duration was found for preceding pauses realised as combinations of silent and filled pauses (6.82%): 1330.83 ms (87–5737 ms, standard deviation: 2169.09 ms). Analysing the data with Kruskal-Wallis test, a significant difference was found between all four groups: $\chi^2 = 80.626; p < 0.001$. We can conclude again that the structure of preceding pauses determines their duration.

In general, it is true of both types of izé that speakers do not pause before them in most cases: izé tends to fit into the flow of speech unawares, irrespective of its current function.

Figure 7. The duration and structure of pauses preceding proxy izé
Filler words in general belong to the non-reparable kind of disfluency phenomena: they arise due to the speaker’s uncertainty and indicate the current difficulty in the planning process without the error itself appearing in the surface utterance. However, izé in Hungarian is an exception. In its proxy function, it becomes reparable, as the following examples show: hogy lefejeltem múltkor nálatok az izét a csillárt ‘that the other day I bumped your stuff, your chandelier’; miért nem csinálnak izét, ilyen kisoperát kórusból? ‘why don’t they make stuff, a kinda short opera with a choir?’; apjának az izéjével, a batár nagy Volvójával volt ‘he was with his father’s stuff, giant big Volvo’.

With respect to proxy izé tokens, we first analysed what part-of-speech categories they replaced. We used the categories as defined in Keszler (2000). In the recordings that we analysed, we found a total of four word classes (Figure 8); all of them were content words. In most cases, the speakers take recourse to izé when a certain noun fails to come to mind. Approximately 90% of the instances are like this: nem kellett volna kiabálni a szülőszob- vagy az izébe csecsemőothomba; ‘they shouldn’t have shouted in the labour wa... or stuff, maternity home’; űgy fogja ööö dirigálni az izét a az Anyegint ‘he will er conduct the stuff, the the Onegin’. The problem of lexical access, then, concerns nouns for the most part. This result is in harmony with earlier studies on false word access (cf. Gósy 2001; Horváth 2006; Horváth & Gyarmathy 2010). One possible explanation is that the part of speech that occurs the most often in spontaneous speech is the noun (cf. Szende 1973; Keszler 1983). Therefore, chances that errors of word activation concern this part of speech are the highest. The percentage of izé substituting for adjectives was 5.68% (ööö izés volt a felesége, rákos volt ‘his wife had stuff, cancer’), for finite verbs it was 3.41% (az baj mer akkor izélt úúú gáz ha túl drága ‘that’s tough for then it stuff, er, it’s bad if it is too dear’), and for non-finite verb forms we found a single example (1.14%) (elkezdett ott ott ott izélni koldulni ‘he began to stuff, cadge there’).

Figure 8. Part-of-speech distribution of instances of proxy izé
The way the repair of *izé* takes place could only be analysed, obviously, in the group of proxy *izé* tokens. Speakers repaired roughly half (50.57%) of the 88 occurrences (e.g. *voltám annál a izénél a pszichológusnál ‘I went to see that stuff, psychologist’*), while 49.43% remained unrepaired (e.g.: *nyomogatom a telefonomat vagy a laptopomat ott az izé alatt ‘I keep pushing my phone or my laptop there, below the stuff’*). This repair proportion is in harmony with what was found in earlier research (cf. Gyarmathy, 2010, 2012; Neuberger, 2010, 2011). The correction of errors occurring in speech depends both on the way the self-monitoring mechanism works and on the speaker’s deliberate decision. As a combination of those two parameters, four different outcomes are possible: (a) the speaker does not detect any error either via internal or via external monitoring, hence the error remains unrepaired; (b) internal self-monitoring detects the error and repair takes place unawares (in such cases, the speaker remembers neither the error nor the fact that it was repaired); (c) the speaker becomes aware of the mistake and intentionally corrects it; and (d) she is aware of the error but cannot repair it or does not think it is necessary (she thinks the error does not disturb the listener’s comprehension or does not wish to derail her own train of thoughts by inserting a repair sequence). The occurrence of the word *izé* in an utterance is supposed to signal that the speaker is aware of the error; hence, in repairing proxy *izé*, only the last two possibilities are to be discussed. (On the basis of the articulation characteristics, it is practically impossible to claim whether the speaker has been aware of the usage of ‘izé’.)

We analysed the way the duration of proxy *izé* tokens varies with repair vs. no repair. As can be seen in Figure 9, repaired *izé* tokens were shorter on average than unrepaired ones. The mean duration of the former was 256.93 ms (161–460 ms, standard deviation: 74.99 ms), while that of the latter was 263.67 ms (150–437 ms, standard deviation: 71.06 ms). Although the difference cannot be confirmed statistically, there is a clear tendency that *izé* is shorter when it is followed by successful repair. In fact, in such cases, correction takes place during the production of *izé*, while in other cases the speaker needs more time to resolve the planning disharmony.

If we take repair vs. no repair into consideration, the part-of-speech distribution of the words replaced by *izé* turns out to be slightly different (Figure 10). Unrepaired proxy *izé* tokens stand for nouns in 93.18% of the cases, for verbs in 4.54%, and for adjectives in 2.27%. In the group of repaired instances, it was necessarily nouns again that dominated (86.36%), but the share of adjectives is four times as large as in the other group (9.09%). Verbs and non-finite verb forms occur infrequently here, too: 2.27%. The amount of data in the smaller subgroups did not make it possible for us to carry out statistical analysis; but we can conclude that *izé* tokens replacing nouns are often not repaired by the speakers, either because they cannot, or because they do not want to. Apparently, a larger context makes it obvious for the listener what noun the word was used for. Another type of error, also due to difficulties of access to the mental lexicon, is “false word finding”. There has been no investigation for Hungarian concerning the proportions of repaired and unrepaired instances of that error type.
Hence, a comparison with our results concerning proxy izé is a task for the future. Such a comparison could show how necessary (or how unnecessary) the presence of izé in this language is; to what extent it helps resolve major word finding difficulties.

![Figure 9. Average duration of repaired and unrepaired instances of proxy izé](image)

Our speakers repaired roughly half of their proxy izé tokens. For that, they used an average of 526.341 ms; but the data are scattered between 0 ms and 3200 ms. The time devoted to correction, or “editing phase” as it is called, can be calculated between the end of the last sound of izé and the beginning of the first sound of the intended word, just like in the case of wrong word access (cf. Horváth & Gyarmathy, 2010). This is the time span within which correction takes place. We have classified editing phases – just like pauses preceding izé – in terms of their realisation in surface structure (Figure 11). In 52.72% of the cases, no pause was detected after izé; the
editing phase was actually missing. In these cases, repair took place either immediately after *izé* was uttered (ott van nyolcvanhárom darab *izé* csecsemő egy *teremben* ‘there are eighty-three pieces of stuff, babies in a room’, meaning that the editing phase was zero (0 ms); or the speaker used some lexical item (e.g. *ilyen* ‘kinda’, *vagyis* ‘that is’) to fill in the time spent on correction (a túlsó végénél *ilyen jó magas ööö* *meredek* *izé* *ilyen* [148 ms] *domb volt* ‘at the other end, there was a sort of high er steep stuff kinda hill’). Editing phases of this kind, that is, ones that did not interrupt the flow of speech were 83.26 ms long on average (0–530 ms; standard deviation: 146.89 ms). Editing phases realised as silent pauses, on the other hand, were over seven times as long (638.25 ms, standard deviation: 410.96 ms) – but they reflected a mere 9.09% of all instances (*persze én is* *ilyen* *izé* [silent pause] (967 ms) *úri fiú vagyok nem szeretem hogyha büdös emberek ülnek le mellém* ‘of course, I am also kinda stuff ... a spoiled boy, I don’t like sm elly people sitting next to me’. Similarly, there were just a few cases in which the speaker filled the time devoted to correction by a combination of silent and filled pause: *utána végzett izét ööö* (260 ms) [silent pause] (470 ms) *ööö* (143 ms) *f- főiskolát* ‘then she attended stuff er ... er c-college’. These were 701.60 ms long on average (standard deviation: 226.18 ms). In one-fourth of the repaired proxy *izé* tokens, correction took an average of 1310.36 ms. Here, speakers typically combined several time-gaining strategies (silent and/or filled pause and linguistic material). The length and structure of the editing phases clearly show that in this category, the participants had major difficulties in accessing the intended lexeme: *akkor a fene se fog ööö izébe* [silent pause] (600 ms) *ööö* (404 ms) *na* (192 ms) *vízbe* ‘then no one will er into stuff ... er ... into water’. We only found a single instance of the editing phase containing a filled pause: *pörkölet* *izébe* ööö (769 ms) *aszpikba* ‘stew into stuff er jelly’. Statistical analysis confirmed that the length of editing phases depends on their structure: nonparametric Kruskal-Wallis test: \( \chi^2 = 29.880; p < 0.001 \). When the speaker is able to surmount the difficulty of lexical activation in a relatively short time, she either does not break the continuity of the utterance or inserts a silent pause. If, however, she needs more time, she will combine several strategies, thus signalling for her partner that she intends to carry on speaking, and also draws the listener’s attention to her current difficulty in word access, hoping that the other person might help her resolve it.

The length of editing phases also depended on the part-of-speech affiliation of the intended words. Speakers repaired *izé* tokens standing for nouns in 496 ms on average (SD = 694.16 ms). When correcting those standing for adjectives, they needed a lot more time, 788 ms (SD = 978.84 ms). The editing phase of the single instance where a verb was intended was 734 ms, and for the case where *izé* stood for an infinitive, it was 425 ms. Due to the small number of items in the individual groups, statistical analysis was not worth pursuing, but the fact that nouns were repaired faster than adjectives confirms the claim that the former are easier to access. This can be seen as correlated with the properties of the class of nouns. Being the most frequently occurring, and largest, group of lexical items, the semantic fields of its members are also larger than those of the items of other part-of-speech categories.
Speakers employ proxy izé as a strategy for smoothing away difficulties of lexical activation, and they are faced with their current speech-planning problem before pronouncing the sequence izé. Therefore, we thought that an analysis of the potential correlation between preceding pauses and editing phases was also in order. Repaired proxy izé tokens were preceded by pauses of 112.43 ms on average, whereas the mean duration of their editing phase was 526.34 ms. In cases where repair was unsuccessful, preceding pauses were far longer, 183.36 ms on average. However, no mathematically confirmable difference was found between the two groups. With 77.27% of the repaired cases, izé was preceded by no pause, confirming the claim that speakers employed this word as a kind of “crutch”. When both the preceding pause and the editing phase were zero, it is probable that repair took place via covert self-monitoring. This was the case in 20.45% of our data. Potential correlations between the two kinds of duration values were submitted to Pearson’s correlation test that showed no interdependence between the durations of editing phases and those of pauses preceding izé.

5 Conclusions

Filler words occurring in speech are due to the speaker’s uncertainty, a disharmony in the speech planning process. As they do not interrupt the continuity of speaking, they are often employed as a problem solving strategy in order to gain time.

During speech production, the speaker may face a number of planning difficulties. These may appear in utterances partly as various kinds of errors (grammatical, articulatory, sublexical, etc.) but also quite often, the speaker has problems in selecting an appropriate item from the lexicon. In such cases, the speaker may resort
to some kind of time-gaining strategy (e.g., silent pause, repetition, filled pause) or, whenever the given language offers some possibility, she may avert the momentary difficulty of lexical access by some replacement lexeme. In view of the various filler words of familiar languages, the Hungarian sequence izé can be seen as unique in that it is a kind of linguistic jolly joker. In cases of any activation difficulty of the mental lexicon, it can replace any word whatsoever. Due to its suffixability, grammatical coherence is maintained, and thus it is capable of helping the speaker gain time at the surface in the case of any disharmony phenomenon.

Our present results have shown that filler izé and proxy izé are clearly separable from one another. The average length of proxy izé tokens turned out to be some 30 ms shorter than that of filler izé tokens. In the first case, correction of the current lexical error takes less time than in the case where the speaker does not yet know how to go on with her utterance and pronounces filler izé as a kind of temporisation. A similar difference shows up in the normalised pitch value of the two categories; it can be concluded that the speakers are aware of the functional difference between the two kinds of izé, and use both of them consciously in one or the other role.

Disharmony in speech planning is often merely shown by silent or filled pauses occurring in surface structure (Levelt, 1989). The occurrence of izé indicates current planning difficulty of the speaker in all cases, but since it is already part of the problem solving strategy, studying the immediately preceding pauses provides us with important insights. The duration and structure of these pauses suggest conclusions with respect to the character of the current problem. Statistical analyses have confirmed that the structure of preceding pauses determines their duration. In the cases of the two kinds of izé, we have analysed preceding pauses separately. They were shorter on average for filler izé tokens, presumably due to the fact that in the case of proxy izé tokens the speaker uses her internal monitor to counteract the current difficulty of lexical access first, as part of the speech planning process, and it is only as a final “crutch” that she relies on saying izé.

Certain parameters like the part-of-speech category of the intended word, correction, and the length of the editing phase could only be analysed in the group of proxy izé tokens. Speakers mainly use izé where a noun escapes them; this may be related to the frequency of occurrence of that word class and with the fact that temporary difficulty of word access is most characteristic of this category. Although the listener can often guess the word replaced by izé with high probability on the basis of the context, the speaker still tries to repair the error as in any other type of trouble with lexical access. In more than half of the cases, correction is successful, just like in cases of false starts (Horváth & Gyarmathy, 2012) and wrong word finding (Horváth & Gyarmathy 2010). The fact that repair takes place also affects the length of izé, albeit just as a tendency. Repaired proxy izé tokens are realised with a shorter duration, as correction already takes place while they are pronounced.

The stretches of time spent on repair, the editing phases, were analysed in terms of their temporal and structural properties. The structure of the editing phase depends on how much time the process of error correction takes. The more difficult a given repair
procedure is, the longer it takes, and the higher the probability of several time gaining strategies being combined in it. The duration of repair is also affected by the word class of the intended lexeme: speakers supply nouns as repair items a lot faster; but it is unrelated to the length of the pause preceding izé.

The present paper reported on a detailed acoustic investigation of izé, a unique filler word of Hungarian, and confirmed that it has different function-dependent realisations. In spontaneous speech, this word has a double role and the two functions entail two distinct versions of its pronunciation. Our results prompt the conclusion that the speaker may employ the word izé quasi-consciously to help covert speech planning processes; yet the occurrence of this word in spontaneous speech samples is occasional at best. This is mainly due to the negative attitude that handbooks of good usage and many native speakers entertain with respect to this word. The legitimacy of such disapproval of the word izé was already questioned in 1863 by Sámuel Brassai: “Akárhogy fintorgassa reá az errá a tudomány, s akármily comoly képpel tiltakozzék is ellene, de a tény csak tény! Az t. i. hogy van a magyar élő beszédben egy szó, mely minden névnek képviselő synonymája: izé és ennek egy származéka: izél, mely hasonlókép minden kitelhető ige helyett szolgál.” [No matter how scholarship grimaces and protests against it with a wry face, a fact is a fact. Spoken Hungarian has a word that can replace any noun as a synonym: izé, and its derivative izél can likewise stand for any possible verb.] (Akadémiai Értesítő, a Nyelv- és Széptudományi Osztály Közlönye 3:211, footnote, cited by Simonyi, 1876:481–482). The raison d’être and usefulness of izé in Hungarian is also confirmed by the fact that, despite its stigmatization, it has been present in this language for at least four centuries and enjoys the best of health in present-day Hungarian, too.

Acknowledgement
This work was supported by the OTKA project No. 108762.

References


Simonyi, Zs. (1876). Egy deminutiv „-j” képző a magyarban [On a Hungarian diminutive suffix]. *Magyar Nyelvőr, 5*, 481–484.


CALL FOR PAPERS

1 General description

The *Phonetician* will publish peer-reviewed papers and short articles in all areas of speech science including articulatory and acoustic phonetics, speech production and perception, speech synthesis, speech technology, applied phonetics, psycholinguistics, sociophonetics, history of phonetics, etc. Contributions should primarily focus on experimental work, but theoretical and methodological papers also will be considered. Papers should be original works that have not been published and are not being considered for publication elsewhere.

2 Paper submission, procedure, mailing

Your manuscript should be submitted to the chief editor, Mária Gósy (Research Institute for Phonetics, MTA): j.phonetician@gmail.com. The formatting of the first and the final versions are detailed below.

The editors and the assistant editor will contact the corresponding author of the paper during the submission process.

Please, send the first version of the manuscript in .pdf format including the figures and tables at the place where you prefer to place them. Please, make sure not to involve the authors in this version, but list them in the e-mail.

Submissions will be double-blind peer-reviewed by two experts of the field. Their opinion and the editorial comments will be sent to the corresponding author of the manuscript.

When submitting the revised version of your manuscript, please include a letter describing how the author(s) addressed the peer-reviewers’ comments. This manuscript should be submitted in both .pdf and Word (.doc) format. The figures should be in .eps, editable Excel or editable SPSS-output files.

Once accepted, the proof of your paper will be sent to the corresponding author in .pdf format. Please, send your correction requests either written on the pdf or in a separate Word document (not in the e-mail text). If you choose to list the corrections in a Word .doc, please add the page number, the paragraph number, the line number, and a short (at least ±3 words) context.

Once the article is published on the ISPhS website, the corresponding author will be notified via e-mail.

3 Formatting

3.1 First submission

The paper should begin with an abstract. The length of the abstract should be between 150-200 words. It should also include 3-6 keywords. The length of the main text should not exceed the length of 10,000 words.
The first version of the manuscript will be blinded and should not include the names of the authors. The authors’ data (names, affiliation, e-mail addresses) will be listed in the submission e-mail. Please identify the corresponding author.

The manuscript should be sent as a .pdf document containing the figures and tables should be placed within the text, where the author(s) want them.

No specific style setting should be used, but the following formattings are required:

- The paper format is A4 with 2.5 cm / ~ 1 inch margins on all four sides.
- The characters of the text should be Times New Roman 11 pt, and the phonetic symbols should be in Doulos Sil.
- The line height should be set to 1.5, except for figures, that should be single spaced.
- Titles and subtitles should follow these same formatting guidelines. They should be preceded by adding their rank in []). Please do not use more than 3 levels of titles.

Examples:
[tit1] Methods
[tit2] Subjects

The figures and tables should not exceed 6 inch / 15.2 cm. Figures and tables should be Times New Roman or Doulos Sil. In preparing your figures, please remember that the page size of the journal is smaller than the standard page size. This means that the figures may be printed somewhat smaller than in the originally submitted paper. Hence, it is essential that the font size of the figure captions be similar to the font size used in the rest of the text.

The Phonetician is an online journal, therefore we accept colored figures.

Please avoid footnotes and endnotes. If you use footnote or endnote, please, use the same formatting as the main text.

There should be no automatic numbering of sections, examples, tables and figures, and no automatic cross-referencing to such objects; please use automatic numbering only for footnotes and endnotes.

Use \textit{fi} for ‘fundamental frequency’; \textit{f1}, \textit{f2}, \textit{f3} for ‘formant’.

Quotes should be written between quotation marks (““). If the quotation exceeds 40 words, please add it as a block quote: Times New Roman 11pt, single space. In both cases, include the page number(s) in the reference in the text.

3.2 Second submission

The revised version of your paper should also be sent to j.phonetician.gmail.com.

The formatting should be the same as for the first submission (in .pdf or Word .doc format with figures in .eps, editable Excel or editable SPSS format. Please place the figures and tables at the appropriate places within the text.

3.3 References

3.3.1 References in the text
- All references in the text should be placed in parentheses: In case of one author, add the surname followed by a comma (,), then the year. In case of
two authors, write the surnames of the authors separated by a comma (,) with
an ampersand (&) before the last one. If there are more than 2 authors, only
list the first author and ‘et al.,’ followed by the year.

• In case of quotes, add the page number(s) after the year separated by a colon
(·).

• In case of multiple references within the parentheses separate them by a
semicolon (;).

• If you cite more than one paper from the same author(s) from the same year,
please use a, b, c… to refer to the papers in alphabetical order by paper title.
Examples:
(Duez, 1982; Misono, Kiratini, 1990; Schachter et al., 1991)
(Siptár, 2002a: 88)

If you add the name of the cited author within the text, provide the year in
parentheses. In case of two authors, please, name both of them. In case of 3 or more
authors, please, name the first one and write „and her/his colleagues”.

3.3.2 The References section

• The surname of the author(s) be first separated by a comma (,) from the
author's initials. In case of more than one authors, please, separate them by
commas (,), and use an ampersand (&) before the last one.

• The year should follow the authors and be followed by a period. In case of
several cited papers/books by the same author(s) in the same year, please,
add a, b, c… after the year in the parentheses.

• The title of the paper or book follows the year. In case of books the title is in
italic; other titles are in regular font. If the language of the cited item is not
English, please, add the title translated into English between brackets ([]).

• If the cited item is a journal article, the title of the journal is italicized
followed by a comma, the volume number, and if there is any, the issue
number in parentheses.

• If the cited item is part of a book, please add “In”, the editor(s) (formatted
like the author(s), then “(ed.)” or “(eds.)”. The book title in italic follows
separated by a colon, followed by a period. The title of the book is followed
by the place of publication and the publisher. These should be separated by
a colon (:), and followed by a period.

• Papers in proceedings should include the title of the proceedings, the
information for the conference (place, date).

• The page number is last and is followed by a period.

• If the cited item was downloaded, please, include the exact url, and add the
date when you last downloaded it. (“Last retrieved:…”)

Examples for references

• Journal papers with 1, 2, and 3+ authors
11–25.

- **Book chapters/proceedings papers with one or more editors:**

- **Downloaded paper:**

- **Book:**
ISPhS MEMBERSHIP APPLICATION FORM

Please mail the completed form to:

Treasurer:
Prof. Dr. Ruth Huntley Bahr, Ph.D.
Treasurer’s Office:
Dept. of Communication Sciences and Disorders
4202 E. Fowler Ave. PCD 1017
University of South Florida
Tampa, FL 33620 USA

I wish to become a member of the International Society of Phonetic Sciences

Title: _____ Last Name: ___________________ First Name: ___________________
Company/Institution: _______________________________________________________
Full mailing address:
________________________________________________________________________
________________________________________________________________________
Phone: __________________ Fax: __________________
E-mail: ________________________________________________
Education degrees: __________________________________________________________
Area(s) of interest: __________________________________________________________

The Membership Fee Schedule (check one):
1. Members (Officers, Fellows, Regular) $ 30.00 per year
2. Student Members $ 10.000 per year
3. Emeritus Members NO CHARGE
4. Affiliate (Corporate) Members $ 60.000 per year
5. Libraries (plus overseas airmail postage) $ 32.000 per year
6. Sustaining Members $ 75.000 per year
7. Sponsors $ 150.000 per year
8. Patrons $ 300.000 per year
9. Institutional/Instructional Members $ 750.000 per year

Go online at www.isphs.org and pay your dues via PayPal using your credit card.
☐ I have enclosed a cheque (in US $ only), made payable to ISPhS.

Date __________________ Full Signature ________________________________________

Students should provide a copy of their student card
NEWS ON DUES

Your dues should be paid as soon as it convenient for you to do so. Please send them directly to the Treasurer:

Prof. Ruth Huntley Bahr, Ph.D.
Dept. of Communication Sciences & Disorders
4202 E. Fowler Ave., PCD 1017
University of South Florida
Tampa, FL 33620-8200 USA
Tel.: +1.813.974.3182, Fax: +1.813.974.0822
e-mail: rbahr@usf.edu

VISA and MASTERCARD: You now have the option to pay your ISPhS membership dues by VISA or MASTERCARD using PayPal. Please visit our website, www.isphs.org, and click on the Membership tab and look under Dues for “paid online via PayPal.” Click on this phrase and you will be directed to PayPal.

The Fee Schedule:

1. Members (Officers, Fellows, Regular) $ 30.00 per year
2. Student Members $ 10.00 per year
3. Emeritus Members NO CHARGE
4. Affiliate (Corporate) Members $ 60.00 per year
5. Libraries (plus overseas airmail postage) $ 32.00 per year
6. Sustaining Members $ 75.00 per year
7. Sponsors $ 150.00 per year
8. Patrons $ 300.00 per year
9. Institutional/Instructional Members $ 750.00 per year

Special members (categories 6–9) will receive certificates; Patrons and Institutional members will receive plaques, and Affiliate members will be permitted to appoint/elect members to the Council of Representatives (two each national groups; one each for other organizations).

Libraries: Please encourage your library to subscribe to The Phonetician. Library subscriptions are quite modest – and they aid us in funding our mailings to phoneticians in Third World Countries.

Life members: Based on the request of several members, the Board of Directors has approved the following rates for Life Membership in ISPhS:

Age 60 or older: $ 150.00
Age 50–60: $ 250.00
Younger than 50 years: $ 450.00