Prosody, voice assimilation, and conversational fillers

Štefan Beňuš[1,2] *Marián Trnka*[2]

[1] Department of English and American Studies, Constantine the Philosopher University, Nitra, Slovakia

[2] Institute of Informatics, Slovak Academy of Sciences, Bratislava, Slovakia

sbenus@ukf.sk, trnka@savba.sk

Abstract

Conversational fillers (CFs), commonly transcribed as *uh*, *um*, or *er*, typically start with a schwa-like vowel, and signal multiple social, interactive, meta-cognitive, and pragmatic functions. They also co-occur with prosodic boundaries, increase saliency of inter-word disjunctures, and participate thus in coding the prosodic structure. Contrary to these functions, CFs are assumed not to participate in the phonological system of a language. This paper uses two types of Slovak conversational speech corpora for investigating the prosodic and phonological behavior of CFs. In Slovak, the vowel inventory does not include a schwa, and word-final obstruents undergo voice assimilation that is triggered by word-initial vowels but interacts with the strength of the prosodic boundary between the two words. Our data show the propensity of CFs to neutralize word-final voicing, and function thus as prosodic breaks, but also non-negligible number of cases of CFs triggering voicing of word-final obstruents, supporting their relevance for cognitive phonology.

Index Terms: conversational fillers, Slovak, non-verbal vocalizations

1. Introduction

1.1. Conversational fillers

Conversational fillers (CFs) are vocalizations such as *uh*, *um*, or *err* that are ubiquitous in everyday speaking, and if considered words, their rates reach 2-5% of all words (e.g. [1], [2]). They convey numerous para-linguistic functions ranging from social approval [3], turn-taking management [4] and meta-cognitive processing [5], to co-creating pragmatic and discourse structures in interactions [6]. CFs are thus necessary for successful, smooth, and socially natural spontaneous conversations ([7], [8]). For illustration, CFs participate in establishing conceptual alignment between interlocutors by signaling given-new distinctions, problems with parsing or understanding preceding speech, or focusing attention on the upcoming material [9]. In this paper we focus on turn-internal CFs that most commonly signal turn-holding hesitation associated with cognitive planning and packaging information in upcoming speech material.

In addition to the frequencies and distributions of CFs, their prosody also provides useful information. For example, in turn-medial position, F0 of CFs was found to systematically correlate with the intonation of the preceding material [10]. CFs are commonly delimited by silent pauses and form thus a separate intonational phrase [2]. However, their role in prosodic chunking when integrally linked to the surrounding material is less clear. On the one hand they are perceived as increasing the salience of the disjuncture between the word that precedes and follows a CF. On the other hand, the pitch re-set, one of the typical signals of prosodic chunking between adjacent units, is very often absent between the units flanking a CF. Hence, it is not clear what role, if any, these prosodically integrated CFs play in prosodic structuring.

1.2. Slovak

Slovak is a West-Slavic language with a vowel inventory containing 5 basic vowel qualities [i, e, a, o, u], phonemic vowel duration, left-most words stress, and minimal reduction of vowel quality in unstressed syllables [11]. Fig. 1 illustrates prosodically conditioned (quantity & stress) Slovak vocalic qualities.

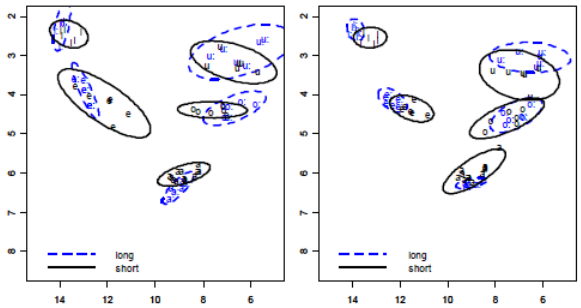

Figure 1: *Slovak vowel quality based on F1 (y-axis) and F2 (x-axis) in Bark for short (solid) and long (dashed) stressed (left) and unstressed (right) monophthongs in CVCa nonsense words from one speaker; adapted from [11].*

Slovak consonantal inventory includes 8 plosives, 4 affricates, and 8 fricatives paired for the underlying phonemic voiced-voiceless contrast. These obstruents are targets for voicing assimilation in the following way. Word-final *voiced* obstruents are *devoiced* if the following word-initial consonant is voiced or if there is a major prosodic break between the words, commonly including a silent pause. Word-final *voiceless* obstruents are *voiced* if the following word starts with a voiced segment, including a vowel, and no silent pause occurs between the two words. While consonant-triggered voicing neutralization is common, vowel-induced voicing of obstruents is rather rare, reported for example for Cracow Polish [12], a variety of another West-Slavic language. Phonetically, voicing neutralization has been found incomplete for languages such as German [13], Polish [14], or Catalan [15], but limited available experimental evidence for Slovak supports complete consonant-induced voicing neutralization across words separated by weak prosodic breaks [16].

1.3. Research questions

The majority of Slovak CFs begin with a schwa-like vowel [17]. This, together with the nature of Slovak voicing assimilation targeting word-final obstruents, offers a unique

10.21437/SpeechProsody.2014-3

testing ground for the role of CFs in the sound system of a language. If CFs trigger voicing of word-final voiceless obstruents, or block de-voicing of word-final voiced ones, CFs might be considered as regular words participating in the phonology of Slovak. If CFs trigger de-voicing of voiced obstruents, or block voicing of the voiceless ones, this would support their analysis as boundary signals, similarly to pauses, participating in the establishment of prosodic structuring. In this sense, CFs might be an optional modal voice boundary marker similar to pre-boundary lengthening or pitch-reset. Alternatively, CFs might be invisible and filtered out for voicing assimilations or prosodic boundary marking. Finally, we want to test if the phonetic realization of fillers, such as their initial glottalization or their duration, affects their behavior in voicing assimilation. A positive answer would be consistent with their analysis as a prosodic boundary marker.

The above questions concern the nature of CFs and their role in the Slovak sound system. A possibility of CF-triggered voicing assimilation might also shed light on the nature of the assimilation process itself. Some analyses treat word-final voicing of obstruents as a phonological phenomenon accounted for in a discrete fashion (i.e. alternations with +/- voice irrespective of the formal framework applied, e.g. [18], [19]). Other approaches employ dynamic modeling to account for both discrete and continuous nature of word-final voicing, especially in light of incomplete neutralization discussed above, e.g. [20]. Finally, some assume that voicing of word-final pre-sonorant obstruents is a purely phonetic phenomenon, e.g. [12]. Hence, if CFs trigger (or do not block) voicing of word-final obstruents preceding them, then CFs must be specified for [+voice], which then spreads to, or licences the voicing of, the preceding obstruent. Alternatively, CFs might be treated as non-linguistic entities incapable of receiving a [+voice] specification and their participation in voicing alternations would be taken as support for the phonetic nature of pre-sonorant word-final obstruent voicing.

2. Methodology

2.1. Corpora

Two corpora of conversational Slovak are analyzed in this paper. The first is a corpus of task-oriented dyadic collaborative conversations accompanying the interactive game designed to elicit dialogues and adapted from the OBJECT Games of Columbia Games Corpus [21], [17]. A pair of subjects saw images on their computer screens and without seeing each other, had to agree on the location of the target object with respect to other objects. One of the subjects (Placer) then dragged this image with the mouse on the location matching as closely as possible the location of this image on the other subject's (Describer) screen. Each session consisted of 14 such placement tasks in which the roles of the Placer and Describer were equally divided between the two players. We will refer to this corpus as *SK-Games*.

In this paper we analyze a subset of *SK-Games* including 6 sessions and totaling almost 4 hours of speech (3h, 54m) from 7 subjects (3 females 4 males; 5 subjects played the game twice with a different partner and 2 male subjects played only one game). There are 21773 words in total, out of which 763 are conversational fillers (labeled as *uh* or *mm*) [17].

The second corpus contains recordings of a courthouse TV show in which semi-professional actors play attorneys and plaintiffs and plead their cases in front of a judge. To match approximately the number of CFs of interest in first corpus (156, see 3.1), we randomly selected 188 conversational fillers following words ending in obstruents. We disregarded the cases in which the disjuncture between the word and the filler corresponded to a #4 ToBI break, that is a silent pause in most cases, focusing thus on the core cases meeting the environment of voicing assimilation. Finally, the extensive size of this corpus allowed for rough balancing of underlying voicing of obstruents. This was not done in SK-Games corpus and resulted in a strong bias toward underlyingly voiceless obstruents (110 vs. 46; see Table 2), which is partially lessened in this corpus (107 vs. 81; see Table 3). We will refer to this second corpus as *Court-TV*.

2.2. Data processing and labeling

Speech was manually transcribed including the transcription of conversational fillers. Transcripts with the audio signal were used for automatic forced alignment using the SPHINX toolkit adjusted for Slovak [22]. Three dimensions, described in Table 1, represented the core labeling effort: characterizing the articulatory activity of the vocal cords initiating the filler, voicing of the obstruent ending the word preceding the filler, and the degree of disjuncture between the filler and the preceding word within the ToBI framework [23].

Table 1. *Labeling scheme for voice assimilation types, "UR" stands for underlying representation*

Label	Dimension	Function description
M	Vocal cords	Modal voice; smooth amplitude increase
G		Glottalization
B		Burst; glottal stop with a observable burst
D	Voice assimilation	Devoicing: UR [+v] obstruent →[-v]
N		No application: UR [+v] obstruent remains [+v]
V		Voicing: UR [-v] obstruent →[+v]
K		Blocking of voicing: UR [-v] obstruent remains [-v]
1	Prosodic disjuncture	No perceivable break
2		Perceived lengthening of word-final rhyme
3		Minor break, commonly associated with tonal marking and/or small pause
4		Major break with significant silent pause between the word and the filler

Additionally, to assess the hypothesis that CFs might be invisible for the purposes of voice assimilation, for all cases with possible voicing (disjunctures 1-3) in *Court-TV*, we noted if the voicing of word-final obstruents respects the generalization described in Section 1.2 should the intervening filler be disregarded and treated as transparent. In other words, we checked the voicing agreement between the final obstruent before the CF and the initial sound of the word following the CF and labeled as respecting (1) if agreement was present or if word-final obstruent was voiceless and a silent pause followed the CF, and as not respecting (0) in the complementary cases.

Finally, the only continuous feature analyzed in this paper is CF duration that is assumed to correlate positively with the boundary strength: longer CF signal stronger prosodic boundaries. Given the uncontrolled nature of the speech in the corpora and questionable reliability of word alignment to the

signal, we employ raw duration and leave normalization effort for subsequent research.

3. Results

3.1. SK-Games corpus

The distribution of voice assimilation types in the corpus of 763 fillers is shown in Table 2. The first four rows correspond to the assimilation types based on the second dimension of Table 1, and the last row is a 'catch all' in which CFs were either preceded by sonorants or a major silent pause inhibiting the application of voicing assimilation. The second and third columns represent the underlying representation (UR) and surface form (SF) of the voicing feature respectively, and the following four columns represent ToBI's boundary indices.

Table 2. *Distribution of voice assimilation types in SK-Game corpus*

Type	UR	SF	1	2	3	4	Total
D	V+	V-	21	1	1	0	23
N	V+	V+	22	1	0	0	23
V	V-	V+	9	0	0	0	9
K	V-	V-	81	15	5	0	101
N-rest			59	40	28	480	607
Total			192	57	34	480	763

If we take the 4 most relevant categories, there are 156 cases in which a word-final obstruent is followed by a filler with a disjuncture lower than ToBI's #4 break. These cases represent the core data for evaluating the questions set in Section 1.3. There are several observations that can be made from the table. First, all possible situations can be observed (with frequency counts of more than 5), and thus any process responsible for voicing alternations potentially triggered by conversation fillers is optional and/or variable. Second, the distribution of underlyingly voiced and voiceless obstruents in this corpus (46 vs. 110) significantly deviates from the expected equal split given the same number of voiced and voiceless obstruents in Slovak ($p < 0.001$ with exact binomial calculation). The same applies to surface forms that are more likely to be voiceless than voiced (124 vs. 32). Third, the 2x2 contingency table for voiced/voiceless obstruents underlyingly and on the surface gives a significant deviation from the expected values, X^2 [1] = 32.9, $p < 0.001$. The analysis of residuals shows that D and V types are significantly under-represented while N is significantly over-represented. Seen in this way, the absence of CF-triggered voicing (K) is to be expected and retention of UR voicing occurs significantly more often than chance. However, changes in underlying specification of voicing (D & V) occur significantly less often.

The preference of voiceless obstruents before the fillers, might be triggered by the voiceless glottal stop initiating the filler. The labeling of the glottal activity shows the presence of burst (B) as extremely rare (9 cases, i.e. 1.2%) and modal voice (M) and glottalization (G) as the most frequent ones. The frequencies of M and G types varied significantly with the boundary strength; X^2 [3] = 55.6, $p < 0.001$. The analysis of residuals showed that their frequencies in the #4 break did not differ significantly but they did for the other 3 break strengths: G-tokens were more frequent for #2 and #3 breaks and less frequent for #1 breaks than M-tokens. This supports previous research in that glottalizations signal prosodic boundary.

Although the type of glottal activity varied significantly for the four assimilation types; X^2 [3] = 13.47, $p < 0.01$, only one cell – glottalization in the blocking of voicing (K) – was significantly different from the expected frequency. Surprisingly, this frequency was significantly lower than expected. This suggests that the phonetic realization of the filler (initiated with a glottal closure or without it) does not predict whether the filler triggers voicing assimilation. Additionally, modal voice was more frequent for all assimilation types, but glottalization appeared for each type as well. These observations suggest that the phonetic realization of fillers and triggering voice assimilation are not linked.

Finally, we examine CF duration as the only continuous feature of this analysis. Fig. 2 shows the data. We observe a tendency for the CFs following surface voiceless obstruents (D, K) to be longer than those following the voiced ones. A mixed models test [24] showed a weak overall effect of voicing type on CF duration (F = 3.02) and Monte-Carlo simulations revealed that the only significant pair-wise difference was between the K and N voicing types ($p = 0.003$).

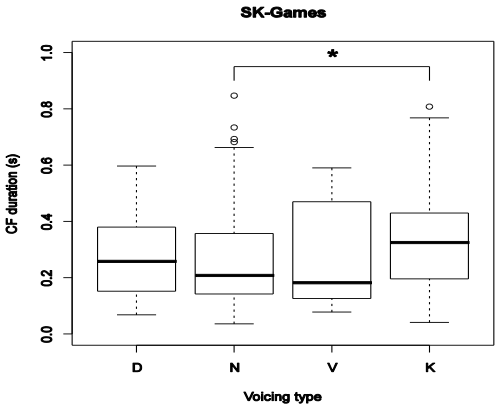

Figure 2: *Duration of conversational fillers for the four voicing types from Table 1.*

3.2. Court-TV corpus

Table 3 summarizes the behavior of fillers with respect to word-final voicing; see text above Table2 for the explanation of labels. The effect of the prosodic strength of the disjuncture corroborates the tendency of fillers to act as prosodic breaks (and either induce or not block word-final devoicing).

Table 3. *Voice assimilation types in Court-TV corpus*

Type	UR	SF	1	2	3	4	Total
D	V+	V-	59	7	4	NA	70
N	V+	V+	11	0	0	NA	11
V	V-	V+	14	0	0	NA	14
K	V-	V-	77	13	3	NA	93
Total			161	20	7	0	188

Contrary to *SK-Games*, the 2x2 contingency table for voiced/voiceless obstruents underlyingly and on the surface shows no significant deviation from the expected values; X^2 [1] = 0.1, $p = 0.92$. This corroborates a strong bias for SF [-voice] obstruents preceding a CF irrespective of their UR.

Regarding the possible interaction between word-final voicing alternations and the vocal cord activity initiating the

fillers, data in this corpus show no token with the burst, very few cases of glottalizations (N=24, i.e. 12.8%), and thus predominantly the initiation of fillers with modal voice. Moreover, all cases of glottalization occurred in tokens with surface voiceless word-final obstruents (11 for D and 13 for K types). Hence, in this corpus, the phonetic realization of the CFs reflected their phonological behavior: glottalization, assumed to reflect CF-initial voiceless glottal stop, co-occurred with devoicing of word-final obstruents resulting thus in the voicing agreement in these 24 tokens. Nevertheless, surface voiceless word-final obstruents were significantly more likely to be followed by CF initiated with regular modal voice than with glottalization (139 vs. 24). This provides additional support for CFs functioning as prosodic breaks rather than participating in phonetic voicing assimilations. Finally, the Fisher's exact test shows that the observed frequency of glottalizations for #2 and #3 breaks (combined to prevent low cell counts) is significantly greater than the expected one. This supports the observation from *SK-Games* and literature that initial glottalizations for vowel-initial words serves as one of the signals for prosodic boundaries.

As discussed in Section 2.2, the annotation of this corpus included binary coding that disregarded CFs and marked the voicing of word-final obstruents as either consistent or inconsistent with the general assimilation processes as described in Section 1.2. This served to test the hypothesis that fillers are transparent and function as neither prosodic breaks that induce devoicing nor as vowel-initial words triggering voicing. Data from this labeling suggest the rejection of this hypothesis. Disregarding CFs shows a slightly greater frequency of cases respecting the assimilation processes than those not respecting them (104 vs. 72), and exact binomial calculation shows that this split is significantly different from 88-88 split ($p = 0.019$). However, by the same token, 72 cases (41%) in which the surface word-final voicing with omitted CF did not respect typical Slovak voicing alternation is sufficient for rejecting the hypothesis. Moreover, the distribution of these two categories did not vary significantly in the four major assimilation types; $X^2 [3] = 0.3$, $p = 0.96$.

Finally, Fig. 3 completes the analysis with CF durations in the similar way to Fig. 2 for *SK-Games*. Although the tendencies in the two figures are similar, a mixed-models test showed neither any overall significant effect nor any significant pair-wise comparison.

4. Discussion

The data from both corpora revealed an overwhelming tendency for surface devoicing of word-final obstruents. While *SK-Games* showed this only for underlyingly voiceless obstruents, *Court-TV* showed this trend irrespective of the underlying voicing of the word-final obstruents preceding the filler. This result supports the hypothesis that CFs function as prosodic boundary markers even if the disjuncture between a word and the following CF is minimal (ToBI's #1 break).

Regarding a possible phonetically-based explanation for the observed voicing assimilations preceding CFs, the results from both corpora reject the hypothesis that the phonetic realization of the filler, i.e. whether it is initiated with a glottal closure or without it, predicts whether the filler supports surface voicing of preceding word-final obstruents. Moreover, the data in both corpora agree with the expected function of glottalization in vowel-initial of increasing saliency of disjunctures: In both corpora, CFs initiated with glottal voice

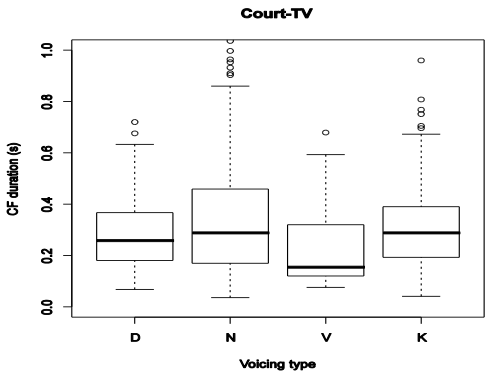

Figure 3: *Duration of conversational fillers for the four voicing types from Table 1.*

were more frequent in #2 and #3 breaks than those in #1 breaks. Interestingly, in *SK-Games*, there was no difference in the major #4 breaks, which suggests that glottalization may function as an additional boundary signal in Slovak for disjunctures of lower boundary strengths.

In spite of the overwhelming tendency for word-final devoicing before CFs, both corpora include small, but non-negligible, numbers of tokens in which CFs trigger (V), or at least do not block (N), voicing of word-final obstruents. These tokens suggest that CFs' participation in voicing assimilation cannot be completely refuted. Furthermore, following the discussion in Section 1.3, the account for these tokens require either treating CFs as 'regular' words or treating voicing assimilation as a purely phonetic process. Interestingly, further analysis of these tokens in *SK-Games* reveals that 21 out of 23 N-type tokens include prepositions like *od* and *z*, both meaning 'from'. Short Slovak prepositions (consonantal or mono-syllabic ones) tend to restructure with the following word and form a single prosodic word. The intervening surface filler and its tendency to function as a prosodic disjuncture might thus clash with the intended single prosodic word unit.

Regarding the duration of fillers, our data showed tendencies for the longer CFs to be preceded by voiceless obstruents on the surface (more so for the K type in which underlyingly voiceless consonant remains voiceless in the environment of the following CF). This observation is consistent with the analysis that CFs function as prosodic boundary markers since the longer the CF, the stronger is presumably the prosodic break signaled by this CF, and thus greater is the tendency for devoicing of word-final obstruents. However, this should be treated as a preliminary finding since CF durations were not normalized and possible silent pauses between the CF and the following words were not considered.

In sum, despite a strong bias for word-final devoicing preceding CFs, and thus their functioning as prosodic boundary markers on par with silent pauses, the hypothesis that CFs participate in voicing assimilation cannot be rejected.

5. Acknowledgements

We thank E. Cyran for inspiring discussion on Krakow Polish voicing and fillers. This research was supported by the ERDF's Research & Development Operational Programme, grant ITMS 26240220064 (RPKOM).

6. References

[1] Shriberg, E., “To "Errrr" is human: Ecology and acoustics of speech disfluencies,“ Journal of the International Phonetic Association 31(1): 153-169, 2001.

[2] Beňuš, Š, Enos, F., Hirschberg, J., Shriberg, E., “Pauses and deceptive speech,” 3rd International Conference on Speech Prosody, Dresden, 2006.

[3] Christenfeld, N., “Does it hurt to say um?” Journal of Nonverbal Behavior, 19: 171 – 186, 1999.

[4] Stenström, A., “Pauses in monologue and dialogue,“ In J. Svartvik (ed.) *London-Lund Corpus of Spoken English: Description and Research*, Lund: Lund University Press, 1990.

[5] Brennan, S., Williams, M., “The feeling of another’s knowing: prosody and conversational fillers as cues to listeners about the metacognitive states of speakers,” Journal of Memory and Language, 34: 383–398, 1995.

[6] Swerts, M., “Conversational fillers as markers of discourse structure, “ Journal of Pragmatics 30: 485–496, 1998.

[7] Bortfeld, H., Leon, S., Bloom, J., Schober, M., Brennan, S., “Disfluency Rates in Conversation: Effects of Age, Relationship, Topic, Role, and Gender,“ Language and Speech 44(2): 123-147, 2001.

[8] Taboada, M., “Spontaneous and non-spontaneous turn-taking,” Journal of Pragmatics 16(2-3): 329-360, 2006.

[9] Stewart, O., Corley, M., “Hesitation disfluencies in spontaneous speech: the meaning of um,“ Language and Linguistics Compass 4: 589–602, 2008.

[10] Shriberg, E., Lickley, R., “Intonation of clause-internal filled pauses,“ Phonetica 50: 172-179, 1993.

[11] Beňuš, Š., Mády, K., “Effects of lexical stress and speech rate on the quantity and quality of Slovak vowels,” Proceedings of 5th Speech Prosody Conference, 2010.

[12] Cyran, E., “Polish voicing,” Lublin: KUL, 2013.

[13] Port, R., Crawford, P., “Incomplete neutralization and pragmatics in German,” Journal of Phonetics, 17: 257-282, 1989.

[14] Slowiaczek, L., Dinnsen, D. “On the neutralizing status of Polish word-final devoicing,” Journal of Phonetics, 13: 325-341, 1985.

[15] Dinnsen, D., Charles-Luce, J., “Phonological neutralization, phonetic implementation and individual differences,” Journal of Phonetics, 12: 49-60, 1984.

[16] Bárkányi, Z., Kiss, Z., “Phonological categoricality vs. phonetic gradience: The laryngeal properties of Slovak three-consonant clusters”, paper presented at the 11th Old-World Conference in Phonology, Leiden, the Nederlands, 2014.

[17] Beňuš, Š., “Cognitive aspects of communicating information with conversational fillers in Slovak”, Proceedings of 4the IEEE Conference of Cognitive Infocommunication, 2013.

[18] Rubach, J., “Nonsyllabic analysis of voice assimilation in Polish.“ Linguistic Inquiry, 27:69–110, 1996.

[19] Lombardi, L., “Positional faithfulness and voicing assimilation in Optimality Theory,“ Natural Language and Linguistic Theory, 17:267–302, 1999.

[20] Gafos, A., Beňuš, Š., “Dynamics of phonological cognition,” Cognitive Science, 30: 905-943, 2006.

[21] Gravano, A., Hirschberg, J. and Beňuš, Š., ”Affirmative cue words in task-oriented dialogue,“ Computational Linguistics, 38(1): 1-39, 2012.

[22] Darjaa, S., Cerňak, M., Trnka, M., Rusko, M., Sabo, R., “Effective triphone mapping for acoustic modeling in speech recognition,” INTERSPEECH, pp. 1717–1720, 2011.

[23] Beckman, M. E., Hirschberg, J., and Shattuck-Hufnagel, S., "The original ToBI system and the evolution of the ToBI framework," S.-A. Jun, ed., Prosodic Typology: The Phonology of Intonation and Phrasing. Oxford University Press, pp. 9–54, 2004.

[24] Baayen, R. H., “Analyzing Linguistic Data. A Practical Introduction to Statistics Using R,“ Cambridge: Cambridge University Press, 241–302, 2008.