# The Effect of Sentence Accent on Non-native Speech Perception in Noise

*Odette Scharenborg*[1], *Elea Kolkman*[1], *Sofoklis Kakouros*[2], *and Brechtje Post*[3]

[1] Centre for Language Studies, Radboud University Nijmegen, The Netherlands
[2] Department of Signal Processing and Acoustics, Aalto University, Finland
[3] Department of Theoretical and Applied Linguistics, University of Cambridge, UK
o.scharenborg@let.ru.nl

## Abstract

This paper investigates the uptake and use of prosodic information signalling sentence accent during native and non-native speech perception in the presence of background noise. A phoneme monitoring experiment was carried out in which English, Dutch, and Finnish listeners were presented with target phonemes in semantically unpredictable yet meaningful English sentences. Sentences were presented in different levels of speech-shaped noise and, crucially, in two prosodic contexts in which the target-bearing word was either deaccented or accented. Results showed that overall performance was high for both the native and the non-native listeners; however, where native listeners seemed able to partially overcome the problems at the acoustic level in degraded listening conditions by using prosodic information signalling upcoming sentence accent, non-native listeners could not do so to the same extent. These results support the hypothesis that the performance difference between native and non-native listeners in the presence of background noise is, at least partially, caused by a reduced exploitation of contextual information during speech processing by non-native listeners.

**Index Terms**: prominence detection, phoneme detection, sentence accent, native listening, non-native listening, noise

## 1. Introduction

Sentence accent plays an important role in speech comprehension [1][2]. For instance, compare the following two sentences, which consist of the same words but have different sentence accent (denoted by upper case), and consequently have a different meaning:

*a.* *The GIRL was cleaning the table*
*b.* *The girl was cleaning the TABLE*

Where in sentence *a* it is emphasised that it was the *girl*, rather than e.g., a boy, who was cleaning the table, in one reading of sentence *b* it is emphasised that the *table* was cleaned, and not some other object. Sentence accent thus expresses semantic focus. Rapid and effective processing of accent placement in an utterance is thus highly important in efficient comprehension of meaning (for a review: [2]) as it is pivotal in understanding the important parts of a speaker's message.

In optimal listening conditions, native listeners are able to exploit prosodic cues in the speech signal signalling upcoming sentence accent to actively focus their attention to those parts of the sentence where accent will fall [1]. Non-native listeners, at least those with a high proficiency in the non-native language, have been shown to be able to detect sentence prominence [3][4] and to use similar acoustic, prosodic cues as native listeners for prominence detection [1][4]. Nevertheless, non-native listeners display a reduced efficiency in using prosodic information signalling sentence accent for the processing of incoming speech [1]. Moreover, differences in the operationalisation of focus between a native and non-native language lead to increased difficulty of handling non-native accentual focus structures in perception [5].

Although background noise is prevalent in everyday listening conditions, research on prominence detection has so far only been carried out in clean listening conditions (but see [6][7] for native prosody perception in noise). Background noise affects speech perception due to its masking of acoustic cues in the speech signal. Prosodic cues correlating with prominence, such as fundamental frequency (F0), are expected to better survive the degrading effect of background noise as cues may survive in different frequency regions (e.g., [8]). It might therefore be expected that prominence detection by native listeners only starts to suffer when listening conditions are quite bad. The picture for non-native listeners might be different. Speech processing in the presence of background noise is hard, and even harder in a non-native language [8]. This difficulty can only partly be explained by phonetic differences between the native and non-native languages. There is now accumulating evidence that this increased deteriorating effect of noise on non-native speech recognition is caused by the non-native listener's less effective use of higher-level information to compensate for loss of information at lower processing levels during speech recognition [9][10].

This is the first study which investigates whether native and non-native (Dutch and Finnish) listeners of English exploit prosodic information signalling sentence accent to aid speech perception in the presence of background noise, while pulling apart the role of preceding prosodic cues and accent on sentence accent detection. The English-Dutch language pair allows us to investigate the influence of prosodic information on non-native spoken-word recognition without vital mismatches at the phonological level and with a reasonably small mismatch at the sound level, as Dutch and English prosodic structures for sentence accent and prosodic processing are highly similar [1]. Where English and Dutch are often characterised as intonation and stress-timed languages, Finnish is regarded as an accent- and syllable-timed language, although the latter claim is somewhat controversial [11]. In Finnish, the most important acoustic cue to prominence is F0, whereas intensity and duration are less important; instead, word order is an important cue for prominence [12]. By including Finnish non-native listeners of English, we will be able to isolate the difficulties due to cross-linguistic differences in cues to accentuation as opposed to a generalised inability for second language learners to exploit such prosodic cues in noisy listening conditions.

# 2. Method

## 2.1. Participants

Forty-five native Dutch listeners (36 females and 9 males; mean age=22.1, SD=2.7), recruited from the Radboud University subject pool, 46 native English listeners (28 females and 18 males; mean age=20.8, SD=2.7), all students from the University of Cambridge, UK, and 49 native Finnish listeners (24 females and 25 males; mean age=27.5, SD=6.7) from Aalto University, Finland participated in the experiments. None of the participants had a history of language, speech, or hearing problems. The participants were paid for their participation. Listeners' proficiency of English was assessed using LexTale [13] (English: mean= 98.6, SD=2.6; Dutch: mean=69.2, SD=17.4 (upper intermediate proficiency); Finnish: mean=84.6, SD=12.1 (lower advanced proficiency)). The difference between the native and non-native listener groups (Dutch: $t(51.3)=-11.8$, $p < .001$; Finnish: $t(52.7)=-7.9$, $p < .001$) as well as the difference between the Dutch and Finnish ($t(87.4)=-5.1$, $p < .001$) listener groups on the LexTale task was significant.

## 2.1. Materials

### *2.1.1. Target phonemes and sentences*

Target phonemes /p, t, k/ always appeared word-initially and in lexically-stressed syllables. Target-bearing words were nouns consisting of up to three syllables. They were not controlled for lexical frequency as frequency has not been found to influence phoneme monitoring [14][15]. The target-bearing words were embedded in sentences, and could appear early or late in the sentence but always minimally 4 words from the start of the sentence. Examples of an early and late target phoneme position (indicated in bold):

a. The owner of the **p**awn shop checked the customer's items.
b. The actions of the crew led to the **t**est lab's evacuation.

For the phoneme-monitoring task, a set of 48 experimental and 48 filler distractor sentences was created. This set was adapted and extended from the set of 24 experimental and 24 distractor sentences created by [1]. All sentences had similar syntactic structure, were semantically unpredictable and only contained one 'critical' target phoneme per sentence (indicated prior to each sentence). Half of the distractor sentences also contained a target phoneme, while the other half did not. Moreover, all 48 experimental sentences were also recorded with prosody that did not signal (upcoming) sentence accent on the target-bearing word. These 'prosodically neutral' sentences were used as a second type of filler sentences.

All sentences were recorded by a male native speaker of British English, using the front internal microphone on a Samson Zoom H2 recorder. All recordings were made at 44.1 kHz, 16 bit, stereo, in a quiet room.

### *2.1.2. Background noise*

Four levels of noise were used in the experiment: clean (no noise was added), and three levels of stationary speech-shaped noise (SSN). SSN is a pure energetic masker, it has a fixed spectrum and no significant temporal modulations [16]. The three SNRs that were used were +5 dB, 0 dB, and -5 dB. The SSN noise was automatically added to all experimental and filler sentences using a PRAAT script [17]. All sentences had 200 ms of leading and trailing SSN noise. A Hamming window was applied to the noise, with a fade in of 10 ms for the leading noise and a 10 ms fade out for the trailing noise.

## 2.2. Prosodic contexts

Sentence accent was manipulated so that the target-bearing words could occur in two prosodic contexts. All sentences contained prosodic context preceding the target-bearing word signalling sentence accent on the upcoming target-bearing word; however, in the 'deaccented' condition, the target-bearing word was in fact deaccented, i.e., incongruent with the preceding context, while it was accented in the 'accented' condition, i.e., congruent with the preceding context. To create the two prosodic contexts all sentences were recorded with an early and a late focal sentence accent (reflecting narrow focus on the words in upper case), and subsequently manipulated:

a. The remains of the **CAMP** were found by the tiger hunter.
b. The remains of the **camp** were found by the TIGER hunter.
c. The remains of the **CAMP** were found by the tiger hunter.

Following the cross-splicing procedure used in [1], for the deaccented condition, the target-bearing word (in bold) from sentence **b** was spliced into sentence **a**. For the accented condition, the target-bearing word from sentence **c,** which is a different rendition of the same sentence as in **a,** was spliced into sentence **a**. The deaccented and accented conditions thus had identical prosodic information preceding the target-bearing words. Differences between the two conditions can thus only be attributed to absence or presence of sentence accent on the target-bearing word.

## 2.3. Procedure

Twenty-four separate experimental lists were created. Each list contained all 48 experimental and 48 distractor sentences. In each list, 8 experimental sentences were presented in each of the four background noise conditions. Within each set of 8 experimental sentences, the target phoneme, position of the target-bearing word, and the two prosodic contexts were evenly distributed. The filler sentences were distributed over the experimental lists following the same procedure.

In order to ensure that listeners processed the sentences for comprehension, and not just focussed on detecting the target phoneme, participants were first instructed that they were participating in an experiment on sentence comprehension, and were told they would be tested on the content of the sentences after the experiment. Afterwards, they were asked to listen within a sentence for the presence of a target sound that was specified for each sentence separately. The target phoneme appeared on the screen for 1 s prior to auditory presentation of the stimulus sentence. Listeners were asked to press the space bar as fast as possible upon hearing the target phoneme. Participants were tested individually in a sound-proof booth. They were randomly assigned to one of the 24 experimental lists. Audio stimuli were presented binaurally through headphones. Participants were comfortably seated in front of a computer screen in a sound-proof booth.

After the experiment, participants were presented with 48 sentences from the main task (equally sampled from all noise and prosodic conditions), in which one word was left blank, and had to indicate which of four alternative word choices they thought had appeared in the sentence in the main experiment. The word recognition task confirmed that both the native (47.8% correct, averaged over all noise and prosodic conditions) and the non-native (Dutch: 40.2% correct; Finnish: 42.0% correct) participants had indeed engaged with the experimental materials, although the non-native listeners did, unsurprisingly, worse on the task than the native listeners.

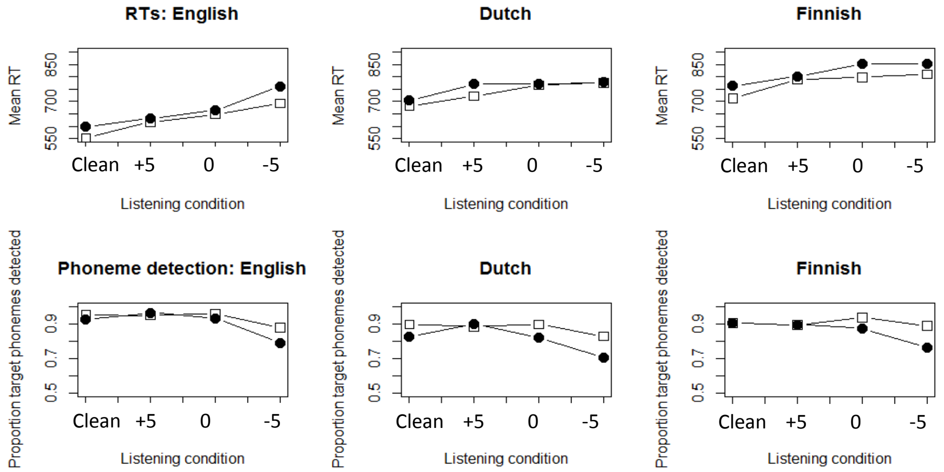

Figure 1. *Mean reaction times (top panels) and the proportion of detected target phonemes (bottom panels) for the three listener groups and the four background noise conditions. The deaccented condition is marked by the bullets, the accented condition by the squares.*

Table 1. *Fixed effect estimates for the best-fitting models of performance for the RT analyses, n=3851.*

| Fixed effect | β | *SE* | *t* |
|---|---|---|---|
| Intercept | 6.394 | .052 | 123.29 |
| Language: Dutch listeners | .218 | .042 | 5.17 |
| Language: Finnish listeners | .281 | .041 | 6.84 |
| Noise | .083 | .014 | 5.76 |
| Prosodic Condition | -.047 | .010 | -4.77 |
| Language: Dutch listeners × Noise | -.037 | .011 | -3.32 |
| Language: Finnish listeners × Noise | .036 | .011 | -3.39 |

Table 2. *Fixed effect estimates for the best-fitting models of performance for the target phoneme detection accuracy analyses, n=4512.*

| Fixed effect | β | *SE* | *p*< |
|---|---|---|---|
| Intercept | 3.194 | .500 | .001 |
| Noise | -1.039 | .237 | .001 |
| Language: Dutch listeners | -1.080 | .289 | .001 |
| Language: Finnish listeners | -.850 | .290 | .01 |
| Prosodic condition | .094 | .179 | n.s. |
| Language: Dutch listeners × Noise | .198 | .115 | .09 |
| Language: Finnish listeners × Noise | .237 | .116 | .05 |
| Prosodic condition × Noise | .259 | .089 | .01 |

### 2.4. Statistical analysis

Statistical analyses on the reaction times (RT; on the correctly detected phonemes) and the number of target phoneme detections on the experimental sentences were carried out using (generalised, in the accuracy analyses) linear mixed-effect models (e.g., [18]), containing fixed and random effects. To obtain the final, best-fitting model containing only statistically significant effects, we used the procedure as, e.g., described in [19]. Fixed factors were Prosodic Condition (accented and deaccented, latter on the intercept), Noise (4 levels: clean (on the intercept), SNR +5, 0, -5), and Language (English (on the intercept), Dutch, and Finnish). Target-bearing Word, Target Phoneme, and Subject were entered as random factors. Random by-stimulus slopes and by-subject slopes for Noise were added and tested through model comparisons in all analyses.

## 3. Results

Analysis of the RTs in the clean condition showed that the RTs of the non-native listeners were significantly slower than those of the English native listeners (Dutch: β=.200, *SE*=.052, *t*=3.83; Finnish: β=.268, *SE*=.051, *t*=5.28), with no significant difference in RTs between the Dutch and Finnish listeners. Moreover, a simple effect of Prosodic Condition was observed (β=-.059, *SE*=.012, *t*=-2.98): RTs in the deaccented condition were significantly slower than those in the accented condition. Native and non-native listeners were thus faster to detect the target phoneme when not only the preceding context indicated upcoming sentence accent but when the target-bearing word also carried sentence accent.

Table 1 shows the estimates of the fixed effects and their interactions in the best-fitting model for the RT analysis. As can also be observed in the top panels of Figure 1, the non-native listeners are significantly slower in detecting the target phonemes than the native listeners. RTs became significantly slower with deteriorating listening conditions, but less so for the non-native listeners compared to the native listeners (see the General Discussion). Interestingly, there was a simple effect for Prosodic Condition: RTs for the accented condition (line with squares in Figure 1) were significantly lower than those for the deaccented condition (bullets in Figure 1). The lack of an interaction between Prosodic Condition and Language, however, illustrates there is no differential use of prosodic information signalling sentence accent between the three listener groups.

The accuracy analysis in the clean listening condition showed that the accuracy for the Dutch non-native listeners was significantly lower than that of the English native listeners (β=-.958, *SE*=.338, *p*<.001) while the accuracy of the Finnish listeners was similar in number to that of the English (β=-.540, *SE*=.344, *p*>.1). In contrast to the analysis of the RTs in the clean condition, no effect of Prosodic Condition was observed.

Table 2 shows the the best-fitting model for the accuracy analysis. Overall, both non-native listener groups detected significantly fewer target phonemes than the English listener group (see also bottom panels of Figure 1). Moreover, significantly fewer target phonemes were detected with increasingly more difficult listening conditions, although this deterioration was significantly smaller for the Finnish listeners (and marginally so for the Dutch) than the English listeners. Interestingly, there was a differential effect of Prosodic Condition on Noise: significantly fewer target phonemes were detected with increasingly more difficult listening conditions, but less so for the accented condition compared to the deaccented condition (compare the line with squares to that of the line with bullets in Figure 1). Finally, the maximal random slope structure of the model included a target word random slope for Noise, indicating that target phoneme detection decreases faster for some target words than others when listening conditions deteriorate.

As with the RT data, the accuracy data did not show an interaction between Language and Prosodic Condition, indicating that the native and non-native listeners did not differ in their uptake of prosodic information signalling sentence accent. Independent analyses of the listener groups indeed confirmed that all listener groups more often detected a target phoneme when the target-bearing word carried sentence stress compared to the condition when it did not. For both the English (β=.443, *SE*=.193, *p*<.05) and Dutch (β=.599, *SE*=.159, *p*<.001) listeners, a simple effect for Prosodic Condition was found, while for the Finnish listeners, as in the main analysis, a differential effect of Prosodic Condition on Noise was found (β=.396, *SE*=.146, *p*<.01).

Although seemingly the non-native listeners do not deviate from the native listeners in their use of prosodic information signalling sentence accent, more targeted analyses on the accuracies of the listener groups separately do seem to suggest a difference in 'breaking point' for the native listeners on the one hand and the non-native listeners on the other: at SNR 0 (Dutch: β=.813, *SE*=.362, *p*<.05; Finnish: β=.786, *SE*=.377, *p*<.05) and -5 (Dutch: β=.886, *SE*=.296, *p*<.01; Finnish: β=1.050, *SE*=.30, *p*<.001) the non-native listeners detected significantly fewer target phonemes in the deaccented compared to the accented condition (see bottom panels in Figure 1) while this difference was only observed at SNR -5 for the English listeners (β=.727, *SE*=.303, *p*<.05).

## 4. Discussion

This paper investigated the uptake and use of prosodic information signalling sentence accent during native and non-native speech perception in the presence of background noise. In line with previous results obtained in clean listening conditions (e.g., [1][3][4]), we found that native and non-native listeners of English are able to exploit prosodic information signalling sentence accent. Overall, the native and non-native listeners were faster and more accurate to detect a target phoneme when not only the preceding context indicated upcoming sentence accent but when the target-bearing word also carried sentence accent. When listening conditions deteriorated, RTs became slower and detection accuracies lower. Importantly, no differential effect of the use of prosodic context was found between the listener groups.

In the clean condition, the overall RTs of the non-native listeners were significantly slower than those of the native listeners. For all listener groups, target phonemes were faster detected in the fully accented condition compared to the deaccented condition. Native and non-native listeners were thus faster to detect the target phoneme when not only the preceding context indicated upcoming sentence accent but when the target-bearing word also carried sentence accent. The Dutch listeners detected fewer target phonemes than the native English, while the Finnish and English listeners did not differ significantly. There was no difference in the number of detected target phonemes between the two prosodic contexts.

When listening conditions deteriorated, the RTs became significantly slower and the number of target phonemes that was detected decreased significantly for all listener groups. This decrease in RTs and number of detected target phonemes was however (marginally) significantly smaller for the non-native listeners than the native English listeners: the native listeners showed a sharp decline in performance at the highest noise level from close to ceiling performance; the non-native listeners showed less of a decline, but they were already performing less well. We speculate that this discrepancy between the listener groups should be explained in terms of the relative robustness of native listener perception, rather than non-native listening behaviour.

Importantly, as was the case in the clean condition, in deteriorating listening conditions, listeners were significantly slower to detect a target phoneme when the target-bearing word did not receive sentence stress compared to when it did. Moreover, significantly fewer target phonemes were detected with increasingly more difficult listening conditions, but less so for the accented condition compared to the deaccented condition. So, while in acoustically more challenging listening conditions both native and non-native listeners are still relatively well able to detect and use sentence accent when the target-bearing word is accented, this use deteriorates when only prosodic information signalling upcoming sentence accent is available. However, no differences in the use of the prosodic cues for sentence accent detection was observed between the three listener groups, which suggests that prosodic context in a non-native language is an equally robust cue for native and non-native listeners.

When listening conditions deteriorated, native listeners used the preceding prosodic information to partially overcome the problems at the acoustic level, which resulted in less of a drop in phoneme detection accuracy compared to the Dutch non-native listeners for the deaccented condition, and less of an increase in reaction time compared to the Dutch and Finnish non-native listeners. Non-natives' speed of use of contextual information thus falls short of that of native listeners even if they are able to use the relevant prosodic cues.

Surprisingly, we did not observe a difference between the Dutch and Finnish listener groups in spite of the prosodic differences between the two languages. A possible explanation is that the two listener groups were not matched closely enough for proficiency in their L2. Possibly, (high) proficiency helps the non-native listener to overcome differences at the prosodic level between the native and non-native language. This needs further exploration.

To conclude, both native and non-native listeners use acoustic and prosodic information for phoneme detection; however, where native listeners seem able to partially overcome the noise-induced problems at the acoustic level by using prosodic information signalling upcoming sentence accent, non-native listeners cannot do so to the same extent, even when the key cues are very similar in their own native language as is the case for Dutch or when the non-native proficiency is high as is the case for the Finnish. Non-native listeners need more prosodic information than the native English listeners to reach a performance level similar to that of the native listeners. These experiments provide support for the hypothesis that the performance difference between native and non-native listeners in the presence of background noise is, at least partially, caused by a reduced exploitation of higher-level, such as prosodic, information during speech processing.

## 5. Acknowledgements

This work was carried out by E. K. as part of an internship under the supervision of O. S., and was sponsored by a Vidi-grant from the Netherlands Organisation for Scientific Research (grant number: 276-89-003) to O. S. The authors would like to thank Joop Kerkhoff for creating the Praat script, and Yvonne Flory for running the native experiment.

# 6. References

[1] E. Akker, and A. Cutler, “Prosodic cues to semantic structure in native and non-native listening,” *Bilingualism: Lang. & Cogn.*, vol. 6, pp. 81-96, 2003.

[2] A. Cutler, D. Dahan, and W. van Donselaar, “Prosody in the comprehension of spoken language: A literature review,” *Language & Speech*, vol. 40, pp. 141-201, 1997.

[3] A. Rosenberg, J. Hirschberg, and K. Manis, “Perception of English prominence by native Mandarin Chinese speakers,” *Fifth International Conference on Speech Prosody*, 2010.

[4] P. Wagner, “Great expectations – Introspective vs. perceptual prominence ratings and their acoustic correlates, ” *Proceedings of Interspeech*, Lisbon, Portuagal, pp. 2381-2384, 2005.

[5] M. L. Garcia Lecumberri, “Perception of accentual focus by Basque L2 learners of English,” *ASJU*, pp. 581-598, 1995.

[6] M. Van Zyl and J. J. Hanekom, “Speech perception in noise: a comparison between sentence and prosody recognition,” *Journal of Hearing Science*, vol. 1, no. 2, pp. EA54-56, 2011.

[7] R. Carroll, R. and E. Ruigendijk, “ERP responses to processing prosodic phrasing of sentences in amplitude modulated noise,” *Neuropsychologia*, vol. 82, pp. 91-103, 2016.

[8] M. L. Garcia-Lecumberri, M. Cooke, and A. Cutler, “Non-native speech perception in adverse conditions: A review,” *Speech Communication*, vol. 52, no. 11-12, pp. 864-886, 2010.

[9] A. R. Bradlow, and J. A. Alexander, “Semantic and phonetic enhancements for speech-in-noise recognition by native and non-native listeners,” *Journal of the Acoustical Society of America*, vol. 121, no. 4, pp. 2339-2349, 2007.

[10] A. Cutler, M. L. Garcia Lecumberri, and M. Cooke, “Consonant identification in noise by native and non-native listeners: Effects of local context,” *Journal of the Acoustical Society of America*, vol. 124, pp. 1264-1268, 2008.

[11] A. Iivonen, “Intonation in Finnish”, In: D. Hurst, and A. Di Cristo (eds.), *Intonation Systems. A survey of Twenty Languages*, pp. 314-330, Cambridge University Press, 1998.

[12] M. Vainio, and J. Järvikivi, “Tonal features, intensity, and word order in the perception of prominence,” *Journal of Phonetics*, vol. 34, pp. 319-342, 2006.

[13] K. Lemhöfer and M. Broersma, “Introducing LexTALE: A quick and valid lexical test for advanced learners of English,” *Behavior Research Methods*, vol. 44, pp. 325-343, 2012.

[14] P. D. Eimas, and L. C. Nygaard, “Contextual coherence and attention in phoneme monitoring,” *Journal of Memory and Language*, vol. 31, pp. 375–395, 1992.

[15] D. J. Foss, D. Harwood, and M. A. Blank, “Deciphering decoding decisions, data and devices”, In R. A. Cole (ed.), *Perception and production of fluent speech*, pp. 165–199. Hillsdale, NJ: Erlbaum, 1980.

[16] M. Cooke, M. L. Garcia Lecumberri, O. Scharenborg, W.A. van Dommelen, “Language-independent processing in speech perception: identification of English intervocalic consonants by speakers of eight European languages”, *Speech Communication*, vol. 52, pp. 954-967, 2010.

[17] P. Boersma, and D. Weenink, “Praat. Doing phonetics by computer (Version 5.1)”, 2005.

[18] R. H. Baayen, D. J. Davidson, and D. M. Bates, D.M. “Mixed-effects modeling with crossed random effects for subjects and items”, *Journal of Memory and Language*, vol. 59, pp. 390-412, 2008.

[19] O. Scharenborg, A. Weber, and E. Janse, “The role of attentional abilities in lexically-guided perceptual learning by older listeners,” *Attention, Perception, and Psychophysics,* vol. 77, no. 2, pp. 493-507, 2015.