Sentence Accent Perception in Noise by French Non-Native Listeners of English

Odette Scharenborg[1], *Fanny Meunier*[2], *Sofoklis Kakouros*[3], *and Brechtje Post*[4]

[1] Centre for Language Studies, Radboud University Nijmegen, The Netherlands
[2] University Côte d'Azur, CNRS, BCL, Nice, France
[3] Department of Signal Processing and Acoustics, Aalto University, Finland
[4] Department of Theoretical and Applied Linguistics, University of Cambridge, UK

o.scharenborg@let.ru.nl, fanny.meunier@unice.fr, sofoklis.kakouros@aalto.fi, bmbp2@cam.ac.uk

Abstract

This paper investigates the use of prosodic information signalling sentence accent and the role of different acoustic features on sentence accent perception during native and non-native speech perception in the presence of background noise. A phoneme detection experiment was carried out in which English native listeners and French highly proficient non-native listeners of English were presented with target phonemes in English sentences. Sentences were presented in different levels of speech-shaped noise and in two prosodic contexts in which the target-bearing word was either deaccented or accented. Acoustic analyses of the two prosodic conditions showed that the target-bearing words in the accented condition carried more energy, had a higher F0, and more spectral tilt than those in the deaccented condition. Results of the behavioural data showed that the native listeners outperformed the French listeners in the clean condition but not in the noise conditions and that the effect of noise was smaller for the non-native compared to the native listeners. Possibly, the non-native listeners use more and different acoustic cues than the native listeners who primarily relied on more local cues for sentence accent detection.

Index Terms: sentence accent, phoneme detection, native listening, non-native listening, noise, acoustic features

1. Introduction

In optimal listening conditions, high-proficiency non-native listeners have been shown to be able to detect sentence accent [1][2][3][4] and exploit prosodic cues in the speech signal signalling upcoming sentence accent [1][5][6]. Moreover, non-native listeners have been shown to use similar acoustic, prosodic cues as native listeners for prominence detection [4][6]. Nevertheless, they show reduced performance compared to native listeners [1] and a reduced efficiency in using prosodic information signalling sentence accent for the processing of incoming speech [6]. [1], for instance, observed that in a sentence accent perception task reaction times were slower and accuracies were lower for non-native listeners (Dutch and Finnish) compared to native listeners (English). Nevertheless, all listener groups used relevant prosodic information in the preceding context to exploit upcoming sentence accent.

Background noise is known to have a different effect on speech perception by native and non-native listeners (e.g., [7][8][9][10]). The degrading effect of background noise is due to its masking of local, acoustic cues [11]. Consequently prominence-related prosodic cues which are (more) local, such as energy and duration, can be obscured. Several prosodic cues which correlate with prominence, however, are more widely distributed (referred to as 'non-local' cues). These non-local prosodic cues, 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., [10]). [1] indeed found that the presence of background noise reduced native and non-native listeners' (Dutch and Finnish listeners of English) ability to exploit sentence accent for speech processing, however, all listener groups were (still) able to use prosodic information signalling upcoming sentence accent, although the non-native listeners to a lesser extent than the native listeners. It is possible that the relative similarity between the prosodic cues to prominence in the three languages facilitated phoneme detection in the accented words for the non-native listeners, while the degrading effect of the presence of background noise could be due to the use of more local prosodic cues in these languages. The current study investigates these possibilities by examining a third group of non-native listeners whose native language is much further removed from English in its cueing and use of prosodic prominence.

Listeners from different language backgrounds use different prosodic cues to detect sentence accent, depending on the way prominence is expressed in their native language [12]. The most important prosodic cues for prominence expression and detection in English are duration, energy, and F0 [13][14][15]. Prominence is used contrastively, marking variable word stress and sentence accents that signal information structure. Unlike English, French does not use prominence to mark word accent (or stress) contrastively [16][17]. In fact, French listeners have been claimed to be 'deaf' to stress both in their native language and in non-native languages [16][17]. However, relative phrasal prominence can be present in French, but to a lesser degree than in English (see [18] for a discussion). French listeners have also been found to rely more heavily on F0 to perceive prominence than English listeners, and less on duration [18] and amplitude [19] (arguably more 'local' cues). As mentioned above, the three languages at play in [1] all use acoustic cues for prominence marking in a similar way. By investigating French non-native listeners of English, we are able to investigate whether they are able to exploit prominence in processing non-native speech; whether they use 'English' acoustic cues to do so (as other non-native listener groups have been claimed to do [3][4]); and whether English listeners suffer more from the effect of background noise on sentence accent perception than French listeners.

2. Methods

Following [1][6], a phoneme detection experiment was carried out. English and French listeners were presented with target

phonemes in sentences in different levels of speech-shaped noise and, crucially, in two sentence accent contexts. The rationale was that words with sentence accent are processed faster and more deeply than words without sentence accent. Thus if listeners are able to exploit prosodic information, this should result in higher phoneme detection rates [6].

2.1. Participants

Thirty-two native French listeners (14 males; 18 females; mean age=29.8, SD=8.6), recruited from the University of Nice, France, and 47 native English listeners (29 females and 18 males; mean age=20.8, SD=2.7), all students from the University of Cambridge, UK, participated in the experiment. The English listeners are a superset of those reported in [1]. 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 [19] (English: mean= 98.6, SD=2.6; French: mean=80.3, SD=8.9 (lower advanced proficiency); $t(34.7)=-11.3$, $p < .001$).

2.2. Materials

Sentence accent perception was investigated by means of a phoneme detection task (following [1][6]). The set-up of the experiment, the stimuli, and procedure were identical to those reported in [1], and will be summarised here.

2.2.1. Target phonemes and sentences

Three target phonemes were used: /p, t, k/. The target phonemes always appeared word-initially in word-initial stressed nouns consisting of up to three syllables. 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 (a) and late (b) target phoneme position (indicated in bold) are given here:

a. The woman with the **p**arrot went into the teacher's office
b. The actions of the crew led to the **t**est lab's evacuation.

A set of 48 experimental and 48 filler distractor sentences was created (using the 24 experimental and 24 distractor sentences from [6] as a starting point). All sentences had a 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, the 48 experimental sentences were also recorded with 'neutral' prosody which did not signal (upcoming) sentence accent. These 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.

2.2.2. Background noise

The background noise applied to the sentences consisted of three levels of stationary speech-shaped noise (SSN) [21]: +5 dB, 0 dB, and -5 dB. The SSN noise was automatically added to all experimental and filler sentences using a PRAAT script [22]. All sentences had 200 ms of leading and trailing SSN noise. A Hamming window was applied to the noise, with a fade in/out of 10 ms for the leading/trailing noise. All sentences were also presented without added background noise (clean).

2.3. Prosodic contexts

Sentence accent was manipulated so that the target-bearing words occurred in one of 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 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:

a1. The remains of the **CAMP** were found by the tiger hunter.
a2. The remains of the **CAMP** were found by the tiger hunter.
b. The remains of the **camp** were found by the TIGER hunter.

To ensure that both the deaccented and accented conditions had identical prosodic information preceding the target-bearing words, for the accented condition, the target-bearing word (in bold) from sentence **a2,** which is a different rendition of the otherwise identical sentence in **a1,** was spliced into sentence **a1**, while for the deaccented condition, the target-bearing word from sentence **b** was spliced into **a1**. Differences between the two conditions can thus only be attributed to absence or presence of sentence accent on the target-bearing word.

2.4. Procedure

Participants were 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. This was done to ensure that listeners processed the sentences for comprehension, and not just focussed on detecting the target phoneme. After this initial instruction, they were asked to also listen within a sentence for the presence of a target sound (p, t, or k) that was specified on a computer screen for each sentence separately. 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. Audio stimuli were presented binaurally through headphones.

Each participant was presented with one of 24 experimental lists. 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. Target phonemes, position of the target-bearing word, and the two prosodic contexts were evenly distributed over all noise conditions. The filler sentences were distributed over the lists following the same procedure. The target phoneme appeared on the screen for 1 s prior to auditory presentation of the sentence.

2.5. Acoustic features

2.5.1. Feature extraction

The speech signals were initially downsampled from 44.1kHz to 8kHz and four main acoustic features were extracted that are known to correlate with the occurrence of prominence in speech: (i) energy, (ii) F0, (iii) spectral tilt, and (iv) duration (see, e.g., [14][23][24]). For the computation, windows of 25 ms were used with a frame shift of 10 ms. Specifically, F0 was computed using YAAPT [25], spectral tilt by computing mel frequency cepstral coefficients (MFCCs) and by taking the first (C1) MFCC [26][27], and word duration was obtained from manual segmentations. Following the computation of the raw feature values: (i) energy was logarithmically normalised, (ii)

F0 was semitone normalised relative to the minimum F0 in each utterance, and (iii) tilt was exponentially normalised – in this case, the exponential function provides a near linear scaling of the tilt estimates to positive real numbers for ease of interpretation. For all features, except word duration, two word-level aggregate measures were computed for the target words and the immediately preceding (3 words) and following (1) word context: the mean and max. Only one word was used for the following context as in many utterances the target-bearing word occurred at the penultimate position in the sentence. Additionally, the mean and max energy were computed over the target-bearing phonemes only.

2.5.2. Acoustic analyses

Differences in the acoustic features of the target-bearing words (first analysis) and of the three target phonemes (/p, t, k/; second analysis) between the two prosodic conditions (accented and deaccented) were statistically analysed using the Wilcoxon rank-sum test statistic and Cohen's *d* for effect size. For the first analysis, data were pooled over all target-bearing words within a prosodic condition, and over all target-bearing words for each of the individual target phonemes separately within a prosodic condition in the second analysis.

The analysis showed that target-bearing words in the accented condition carry more energy (max_energy: $Z = -5.21$, $p < 0.001$, $d = 1.42$; mean_energy: $Z = -2.77$, $p < 0.01$, $d = 0.62$), have a higher F0 (max_F0: $Z = -5.54$, $p < 0.001$, $d = 1.33$; mean_F0: $Z = -4.89$, $p < 0.001$, $d = 1.05$), and increased spectral tilt (smaller slope - max_tilt: $Z = -4.18$, $p < 0.001$, $d = 0.94$; mean_tilt: $Z = -3.84$, $p < 0.001$, $d = 0.80$) than target-bearing words in the deaccented condition. Duration did not differ between the accented and deaccented condition ($Z = -0.61$, $p = 0.55$, $d = 0.17$). The words directly preceding and following the target-bearing word showed small but non-significant differences in energy, F0, spectral tilt, and duration between the prosodic contexts. The acoustic measures for the target phonemes did not differ between the two prosodic conditions. These results are, however, not unexpected, as all target phonemes are unvoiced plosives, while effects related to accentuation are typically observed at sonorant parts of the words which are typically found at syllabic nuclei.

2.6. Statistical analyses

Statistical analyses on the reaction times (RT; on the correctly detected phonemes) and the number of target phoneme detections (following [6]) on the experimental sentences were carried out using (generalised, in the accuracy analyses) linear mixed-effect models (e.g., [28]), containing fixed and random effects. To obtain the final, best-fitting model containing only statistically significant effects, we used a backward stepwise selection procedure, in which interactions and predictors that proved not significant at the 5% level were removed one-by-one from the model (see e.g., [29]). Fixed factors were Prosodic Condition (accented and deaccented, latter on the intercept), Noise (clean on intercept, SNR +5, 0, -5 dB), and Language (English on intercept). 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. Moreover, the acoustic features calculated at the target-bearing word level (centered and scaled) were added as fixed factors and in interaction with Noise and Language: Energy_max, Energy_mean, Tilt_max, Tilt_mean, F0_max, F0_mean, Duration.

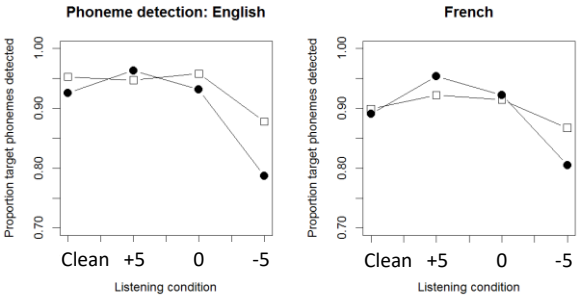

Figure 1. *Proportion of detected target phonemes for the native English listeners (left panel) and the French non-native listeners of English (right panel) for the four background noise conditions. The deaccented condition is marked by the bullets, the accented condition by the squares.*

Table 1. *Fixed effect estimates for the best-fitting models of performance for the phoneme detection analysis, noise only n=1896.*

Fixed effect	β	*SE*	*p* <
Intercept	7.130	1.566	.001
Prosodic Condition	-1.070	.586	.068
Noise	-2.380	.626	.001
Prosodic Condition × Noise	.603	.242	.013

3. Results

First the English and French listeners' phoneme detection rates in the clean and the noise conditions were compared. Subsequently, the use of acoustic cues was investigated for the clean and noise condition for the two listener groups separately.

Figure 1 shows the results on the phoneme detection task for the English listeners (left panel) and the French listeners (right panel) for the four listening conditions, and for the deaccented (bullets) and accented (open squares) conditions separately. Detection rates of the target phonemes /k, p, t/ were 89.2%, 94.0%, and 92.0%, respectively, for the English listeners, and 87.2%, 90.6%, and 90.9% respectively, for the French listeners. The first striking result is that the French listeners, despite being called 'deaf to prominence' [16][17] were very good at the task. Nevertheless, the accuracy analysis in the clean listening condition showed that the accuracy for the French non-native listeners was significantly lower than that of the English native listeners (β=-.6144, *SE*=.306, *p*=.045).

Table 1 shows the estimates of the fixed effects and their interactions in the best-fitting model for the phoneme detection analysis for only the noise conditions. The statistical analysis showed that significantly fewer phonemes were detected with increasing noise levels, and this is especially the case for the deaccented condition (see also the bulleted lines in Figure 1). Importantly, no significant differences in overall performance nor of the effect of noise on phoneme detection were observed between the native and non-native listener group.

Table 2 shows the results for of the acoustic feature analyses in the clean condition. For the English listeners, a lower phoneme detection accuracy was associated with increasing mean F0 for target-bearing words and this was especially so for the accented condition. For the French non-native listeners of English, significantly more target phonemes were detected in the accented condition compared to the deaccented condition, while decreasing max energy and increasing max spectral tilt were associated with higher phoneme detection rates.

Table 2. *Fixed effect estimates for the best-fitting models of performance for the phoneme detection analysis with acoustic parameters – clean condition only.*

Fixed effect	β	*SE*	*p* <
English listener group, n=368			
Intercept	1.705	1.175	.147
Prosodic Condition	.392	.484	.418
F0_mean	-2.800	1.429	.050
Prosodic Condition × F0_mean	1.347	.624	.031
French listener group, n=256			
Intercept	-4.174	3.158	.186
Prosodic Condition	2.703	1.355	.046
Energy_max	-2.377	1.178	.044
Tilt_max	2.338	1.131	.039

Table 3. *Fixed effect estimates for the best-fitting models of performance for the phoneme detection analysis with acoustic parameters – clean (on the intercept) vs. noise.*

Fixed effect	β	*SE*	*p* <
English listener group, n=1472			
Intercept	3.435	.249	.001
Noise	-.462	.0923	.001
Energy_max	-.116	.212	.584
Noise × Energy_max	.204	.090	.024
French listener group, n=960			
Intercept	2.019	.788	.010
Noise	-.230	.108	.033
Energy_max	-1.134	.337	.001
Energy_mean	.419	.287	.145
Prosodic Condition	.416	.319	.192
Tilt_mean	3.529	1.068	.001
Noise × Energy_max	.342	.122	.005
Noise × Energy_mean	-.320	.156	.041
Prosodic Condition × Tilt_mean	-1.177	.367	.002

Table 3 shows the results for the acoustic feature analyses in clean versus noise for the two listener groups. For the English listeners, we find a significant effect of noise: significantly fewer phonemes are detected in deteriorating listening conditions compared to the clean, and this is especially the case for target phonemes in target-bearing words with higher max energy. Seemingly, max energy is less reliably used for sentence accent detection when noise conditions are adverse. Potentially, at low SNRs the differences in max energy are not as reliable as in the clean condition as the intensity level of speech is masked by the intensity level of the added noise (thus, not acting as a reliable cue for prominence). For the French listeners, we observe more acoustic features that play a role in sentence accent detection. First, as for the English listeners, we find a significant effect of noise. Moreover, a higher max energy is associated with significantly fewer detected target phonemes, specifically when listening conditions get harder. A higher mean tilt is associated with significantly more detected target phonemes, and this is even more the case for the accented condition. Finally, significantly fewer phonemes are detected in deteriorating listening conditions, but this is less so for target phonemes in target-bearing words with higher mean energy.

4. Discussion

This paper investigates the effect of background noise on the use of acoustic cues for prominence in native and non-native listening: English, with more 'local' acoustic cues for prominence marking which are arguably less easy to pick up in background noise, was the native language, and French the non-native language, whose listeners are claimed to be 'deaf' to prominence [16][17]. Moreover, we investigated whether high-proficiency French listeners use similar acoustic cues for prominence detection as native English listeners.

French non-native listeners were found to be surprisingly good at exploiting sentence accent for phoneme detection. Although they detected significantly fewer target phonemes than the English listeners in the clean condition, there was no significant difference between the two listener groups in noise. The French listeners thus suffered less from the presence of background noise than the English listeners. Moreover, the French listeners were found to be able to use preceding prosodic information signalling upcoming sentence accent to the same extent as the English listeners, in both the clean and noisy listening conditions. These results extend the results of [1], who found that prosodic context is an equally robust cue for native and Dutch and Finnish non-native listeners of English, to non-native listeners with a native language (French) that has no contrastive sentence accent marking. The current results are in line with the hypothesis that high proficiency helps non-native listeners to overcome differences at the prosodic level between the native and non-native language [1].

The acoustic analyses of the two prosodic conditions showed that the target-bearing words in the accented condition carried more energy, had a higher F0, and more spectral tilt than those in the deaccented condition. In line with these results, we found that significantly more target phonemes were detected in the accented compared to the deaccented condition. The subsequent analyses of the use of acoustic features for prominence exploitation, though, showed that both the native and non-native listeners did not use the acoustic cues as expected. Only a few acoustic cues were found to predict proportion of detected target phonemes and often in the opposite direction of what one would expect: e.g., a higher mean F0 was associated with fewer detected target phonemes in the clean condition by the natives; and increasing max energy was associated with fewer detected target phonemes for both groups. However, for the French listeners, more spectral tilt was associated with an increase in the proportion of detected target phonemes and especially for the accented condition. Moreover, they were able to use mean energy to compensate for worse listening conditions: although increasing background noise levels led to fewer detected target phonemes, this effect was reduced for target-bearing words with higher mean energy.

To conclude, French non-native listeners of English were found to be able to exploit sentence accent for improved target phoneme detection, similar to non-native listeners with native languages that cue and use prominence quite differently [1]. This confirms that non-native listeners can overcome certain differences at the prosodic level between the native and non-native language, at least at high proficiency levels. Moreover, the effect of background noise was smaller for the non-native than the native listeners. Arguably, the highly proficient non-native listeners exploit more and different acoustic cues when intelligibility is compromised, while the native English listeners continue to primarily rely on the usual more 'local' cues.

5. Acknowledgements

This work was sponsored by a Vidi-grant from the Netherlands Organisation for Scientific Research (NWO; 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 English experiment.

6. References

[1] O. Scharenborg, E. Kolkman, S. Kakouros, B. Post, "The effect of sentence accent on non-native speech perception in noise," *Interspeech*, San Francisco, CA, 863-867, 2016.

[2] A. Eriksson, E. Grabe, and H. Traunmueller, "Perception of syllable prominence by listeners with and without competence in the tested language", *Proceedings Speech Prosody*, Aix-en-Provence, France, pp. 275-278, 2002.

[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] 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.

[6] 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.

[7] 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.

[8] O. Scharenborg, J. Coumans, R. van Hout, "The effect of background noise on the word activation process in non-native spoken-word recognition", Journal of Experimental Psychology: Learning, Memory, and Cognition. doi:10.1037/xlm0000441, 2017.

[9] 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.

[10] 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.

[11] M. Cooke, "Discovering consistent word confusions in noise", *Interspeech*, Brighton, 2009.

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

[13] C. Gussenhoven, *On the grammar and semantics of sentence accents*. Dordrecht: Foris, 1983.

[14] D. B. Fry, "Duration and intensity as physical correlates of linguistic stress", *Journal of the Acoustical Society of America*, vol. 27(4), 765-768, 1955.

[15] D. B. Fry, "Experiments in the perception of stress", *Language and speech*, Vol. 1(2), 126-152, 1958.

[16] Dupoux, E., Sebastián-Gallés, N., Navarrete, E., Peperkamp, S., "Persistent stress 'deafness': The case of French learners of Spanish," *Cognition*, vol. 106(2), pp. 682-706, 2008. https://doi.org/10.1016/j.cognition.2007.04.001

[17] S. Peperkamp, E. Dupoux, and N. Sebastián-Gallés,"Perception of stress by French, Spanish, and bilingual subjects", *Proceedings of Eurospeech*, pp. 2683-2686, 1999.

[18] D. Frost, "Stress and cues to relative prominence in English and French: A perceptual study", *Journal of the International Phonetic Association*, Vol. 41(1), 67-84, 2011.

[19] A. Séguinot, "L'accent d'insistance en français standard," *Studia Phonetica **12**: L'accent d'insistance: Emphatic stress*, edited by F. Carton, D. Hirst, A. Marschal & A. Séguinot, pp. 1-58. Paris: Didier, 1977.

[20] 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.

[21] 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.

[22] P. Boersma, and D. Weenink, "Praat. Doing phonetics by computer (Version 5.1)", 2005.

[23] S. Shattuck-Hufnagel and A.E. Turk, "A prosody tutorial for investigators of auditory sentence processing", *Journal of psycholinguistic research*, *25*(2), pp. 193-247, 1996.

[24] A.M. Sluijter and V.J. van Heuven, "Spectral balance as an acoustic correlate of linguistic stress", *The Journal of the Acoustical society of America*, *100*(4), pp. 2471-2485, 1996.

[25] S.A. Zahorian and H. Hu, "A spectral/temporal method for robust fundamental frequency tracking", *The Journal of the Acoustical Society of America*, *123*(6), pp. 4559-4571, 2008.

[26] P. Tsiakoulis, A. Potamianos, & D. Dimitriadis, "Spectral moment features augmented by low order cepstral coefficients for robust ASR", *IEEE Signal Processing Letters*, *17*(6), pp. 551-554, 2010.

[27] S. Kakouros, O. Räsänen, A. Paavo, "Evaluation of Spectral Tilt Measures for Sentence Prominence Under Different Noise Conditions," *Interspeech*, Stockholm, Sweden, 3211-3215, 2017.

[28] 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.

[29] 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.