Interspeech 2018
2-6 September 2018, Hyderabad

Impact of different speech types on listening effort

Olympia Simantiraki[1], *Martin Cooke*[1], *Simon King*[2]

[1]Language and Speech Laboratory, Universidad del País Vasco, Vitoria, Spain
[2]Center for Speech Technology Research, University of Edinburgh, Edinburgh, UK

olympia.simantiraki@ehu.eus, m.cooke@ikerbasque.org, Simon.King@ed.ac.uk

Abstract

Listeners are exposed to different types of speech in everyday life, from natural speech to speech that has undergone modifications or has been generated synthetically. While many studies have focused on measuring the intelligibility of these distinct speech types, their impact on listening effort is not known. The current study combined an objective measure of intelligibility, a physiological measure of listening effort (pupil size) and listeners' subjective judgements, to examine the impact of four speech types: plain (natural) speech, speech produced in noise (Lombard speech), speech enhanced to promote intelligibility, and synthetic speech. For each speech type, listeners responded to sentences presented in one of three levels of speech-shaped noise. Subjective effort ratings and intelligibility scores showed an inverse ranking across speech types, with synthetic speech being the most demanding and enhanced speech the least. Pupil size measures indicated an increase in listening effort with decreasing signal-to-noise ratio for all speech types apart from synthetic speech, which required significantly more effort at the most favourable noise level. Naturally and artificially modified speech were less effortful than plain speech at the more adverse noise levels. These outcomes indicate a clear impact of speech type on the cognitive demands required for comprehension.

Index Terms: listening effort, pupil response, speech perception, synthetic speech

1. Introduction

In our everyday life we are exposed to a variety of speech types, both naturally and artificially produced. Talkers modify their speech when exposed to noise. Live and recorded public address announcements may involve modifications designed to enhance intelligibility. Synthetically-generated speech is commonplace in mobile devices and telephone enquiry systems. Correct message reception is critical in many situations and consequently a great deal of effort has been devoted to understanding the effect of differing speech styles on intelligibility [1]. However, far less emphasis has been placed on investigating the effort required to understand distinct speech types. Exposure to conditions that require a listener to exert substantial effort and the engagement of additional cognitive resources may lead to long term fatigue. Such conditions might be degraded source signals, sound transmission interference or limitations of the receiver (see review in [2]). The current study examines listening effort for distinct speech types under conditions of additive noise.

Listening effort has been estimated using subjective measures such as questionnaires, behavioural metrics (e.g. response time), and via physiological measures such as pupillometry [3]. For instance, [4] obtained psychophysiological recordings (heart rate, skin conductance, skin temperature and electromyographic activity) during speech perception tasks with intelligibility close to ceiling, but with varying task demands involving digit presentation to one or both ears. Increased mean skin conductance and electromyographic activity were observed when task demand increased. Multi-task paradigms are typically employed in studies for measuring behavioural responses. In [5], a dual-task paradigm was used to assess listening effort at a wide range of signal-to-noise ratios (SNRs). Reaction times, in line with subjective effort measures, showed less effort exertion for lower SNRs. In [6] the benefit of a digital noise reduction algorithm was tested using dual-task paradigms either in quiet or in the presence of a 4-talker babble masker at various SNRs. Noise reduction was found to both reduce effort and benefit performance in simultaneous tasks.

Behavioural measures alone cannot systematically measure changes in effort, but such changes can be revealed through variations in pupil size [7]. Several studies have used the pupillary response as an objective indicator of listening effort while measuring speech perception under noisy conditions. Features typically used to estimate effort are the mean pupil dilation, peak pupil dilation (PPD) or delay to reach the peak (latency). These features show an increasing trend with decreasing intelligibility [8]. PPD has been shown to reflect listening effort when tested in speech performance tasks involving sentences presented in conditions of informational or energetic masking [8, 9], with more effort observed for a competing talker masker than stationary or fluctuating maskers [10, 11]. Listening effort typically is maximised for speech-in-noise tasks at intelligibility levels of around 50% [9, 12, 5]. Pupillometric measures of effort have also been obtained as a function of syntactic complexity [13], attention to location [14] and spectral resolution [15].

The current study uses pupillometry, subjective judgements and intelligibility scores to investigate the effect of four distinct speech types on listening effort. In addition to plain natural speech, we examine one naturally-modified form (Lombard speech), one algorithmically-modified form designed to enhance intelligibility, as well as synthetic speech, using the same set of sentences in each case. Listeners heard sentences presented in one of three levels of speech-shaped noise noise.

2. Methods

2.1. Participants

Twenty-six normal-hearing, native British English participants (6 males and 20 females between 18 and 24, with a mean age of 20.5, $SD = 1.8$) were recruited. Participants were asked not to wear glasses and eye makeup [8]. Participants underwent pure-tone hearing screening; all had a hearing level less than or equal to 25 dB in both ears. Data from two participants was excluded from the analysis due to technical problems during recording.

10.21437/Interspeech.2018-1358

2.2. Speech and masker materials

Speech and noise material for the different speech types was selected from the Hurricane Challenge corpus [16] which consists of Harvard sentences [17] subsequently mixed with noise. The four speech types used in the current study are detailed below.

Plain: Plain sentences were uttered by a British English male talker in quiet when asked to speak normally.

Lombard: Lombard sentences were uttered by the same talker in the presence of a speech-shaped noise (SSN) masker. Lombard speech is a naturally-modified form which exhibits acoustic differences over plain speech such as a decrease in spectral tilt, increased F0 and changes to segmental durations [18].

Synthetic speech (TTS): Sentences were generated by a high quality speaker-adaptive HMM-based speech synthesis system [19] adapted to the male talker's voice. In the Hurricane Challenge TTS was less intelligible than non-TTS speech types.

SSDRC-modified speech: The Spectral Shaping and Dynamic Range Compression (SSDRC) method [20] was used to generate this algorithmically-modified form of speech. SSDRC was inspired by observations from both clear and Lombard speech styles, and involves several stages of spectral modification followed by dynamic range compression. The technique operates independently of the masker type and level, and can lead to substantial intelligibility gains [1].

Each speech style used the same utterances, and all were mixed with a SSN masker at one of three SNRs (-1, -3 and -5 dB), a range of SNRs at which neither "giving-up" or "effortless" effects are observed for the plain speech under stationary masking conditions as shown in [12]. We chose a range of SNRs in order to explore conditions where intelligibility is at near-ceiling levels and where it falls below ceiling levels. The speech signal was rescaled to achieve the desired SNR for the portion where it overlapped the masker. Speech-plus-noise stimuli were then normalised to the same root mean square level and 20 ms half-Hamming ramps were applied to reduce onset and offset transients.

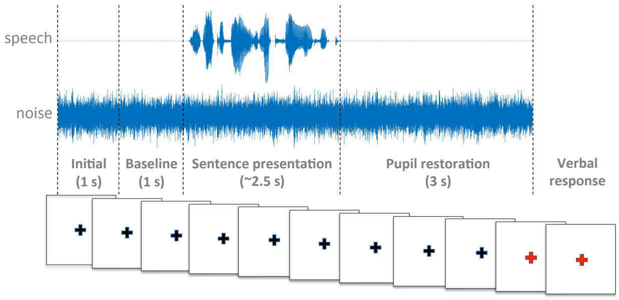

Figure 1: *Illustration of a single trial. The participant's gaze was focused on the cross.*

2.3. Procedure

Masking noise was present throughout each trial. Sentence onset occurred two seconds into the trial, and the masker continued for three seconds following sentence offset. Pupil data from the interval 1-2 seconds (ie immediately preceding sentence onset) was used for calibration (see 2.4 below). Participants were asked to look at the black cross and to respond by repeating the words they had heard when the cross changed colour to red. The procedure is illustrated in Figure 1.

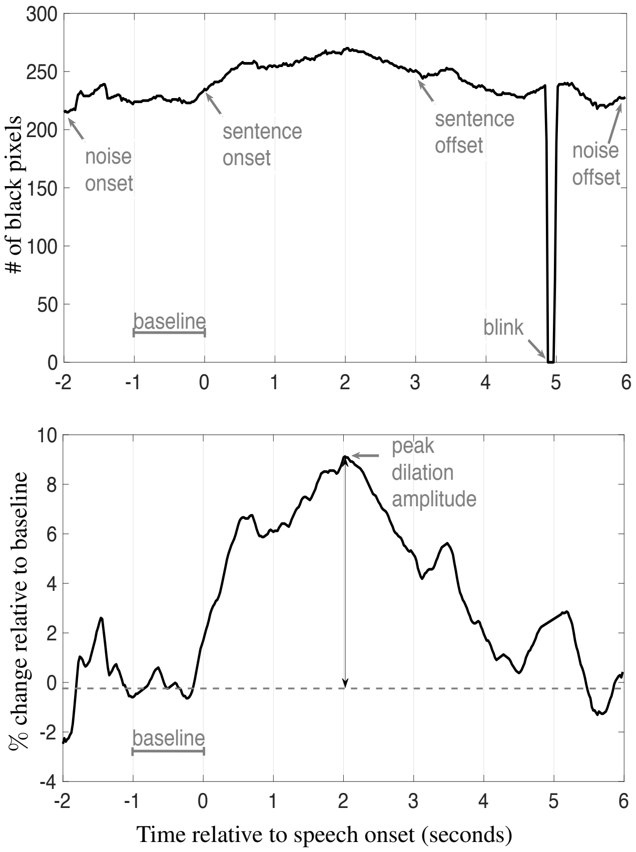

Figure 2: *Pupil size variation during a single trial. Upper: uncalibrated pupil area; lower: calibrated pupil diameter. Times are relative to sentence onset at 0s.*

Stimuli were blocked by speech type and SNR, resulting in 12 blocks, each containing 20 sentences. Order of block presentation was balanced across listeners, and stimuli within each block were presented in a random order. No listener heard the same sentence more than once. The data from the first five sentences in each block were excluded from further analysis as their purpose was to familiarise the participant with the condition.

The experiment took place in a sound proof studio at the University of Edinburgh. Pupil data was collected using the remote EyeLink 1000 eye-tracker while participants listened to sentences through headphones.

At the end of each block each participant was asked "How much effort did it take to listen and understand the sentences in this block?" and provided an answer on a scale from 0 (no effort) to 10 (very effortful). The experiment lasted around 1 hour with a five minute break in the middle.

2.4. Calibration

Pupil data from the left eye was used [8]. Pupil area data was first downsampled to 50 Hz and converted to pupil diameter, followed by noise removal (e.g., detecting blinks) using a similar procedure to that of [9]. The cleaned traces were then calibrated following [21]:

$$ERPD = (observation - baseline)/baseline * 100$$

where ERPD stands for Event-Related Pupil Dilation, $observation$ is the uncalibrated pupil diameter and $baseline$ is the mean pupil diameter during the one second interval preceding the onset of the speech. Figure 2 provided an example of the uncalibrated pupil area and calibrated pupil diameter (bottom). In common with [11, 12], the peak dilation amplitude was used for assessing listening effort. Peak pupil dilation (PPD) is defined as the maximum size that the pupil reaches after the cognitively demanding event relative to the baseline (see lower panel of Figure 2). A visual inspection on the data for artifacts was also performed and blocks with fewer than 66.67% correct trials were excluded, leading to the removal of around 11% of the pupil data.

2.5. Statistical analyses

Intelligibility scores were converted to rationalised arcsine units [22]. Linear mixed-effect models were used for statistical analysis. Effect sizes were estimated using d-prime (d') for fixed effects and Chi-square (χ^2) for random effects, computed using the *SensMixedUI* tool [23].

3. Results

3.1. Pupil data

Figure 3 depicts the mean across-participant ERPD for each speech style and SNR. For the most favourable SNR, TTS shows the greatest change in pupil dilation over the baseline, followed by plain speech. A similar trend is seen at the intermediate SNR. For the adverse SNR plain speech exhibits the largest relative increase in pupil size. Lombard speech generally results in the lowest ERPD at each noise level.

A linear mixed model fitted to the peak pupil dilation data revealed a significant effect speech type ($F[3, 200] = 13.05, p < .001, d' = 0.64$) and SNR ($F[2, 200] = 4.10, p < .05, d' = .31$). PPD also differed significantly across participants ($\chi^2(1, N = 23) = 71.75, p < .001$). Post-hoc analyses showed that at SNR = -1dB, TTS led to significantly higher PPDs than Lombard, SSDRC and plain speech ($max\ p < .01$). At the intermediate SNR, Lombard speech produced significantly lower PPD than both plain and TTS ($max\ p < .01$). At this SNR TTS had a larger PPD than SSDRC ($p < .05$), while the difference between plain speech and SSDRC was marginal ($p = .052$). Finally, at the adverse SNR the PPD for SSDRC was significantly different from both TTS ($p < .05$) and plain speech ($p < .01$); Lombard also differed from plain speech ($p < .05$).

3.2. Subjective ratings

Mean subjective ratings for the different speech types across the three SNRs are depicted in Figure 4, revealing an unambiguous ranking of speech types. Synthetic speech was considered the most effortful style and SSDRC the least. Effort decreased with increasing SNR. A linear mixed model confirms the significant effect of SNR ($F[2, 200] = 68.64, p < .001, d' = 1.31$) and speech type ($F[3, 200] = 242.82, p < .001, d' = 2.77$), as well as an interaction ($F[6, 200] = 2.48, p < .05, d' = .36$).

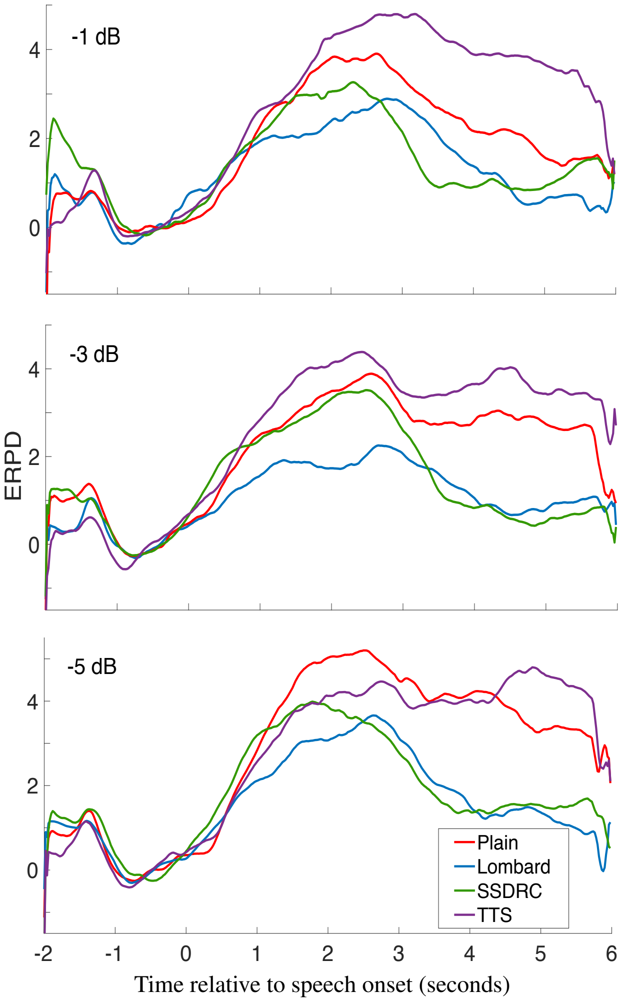

Figure 3: *Mean pupil size increase over baseline as a function of speech type and SNR.*

3.3. Intelligibility

The mean percentage of correct words repeated by participants for the different speech types and SNRs is shown in Figure 5. A ceiling effect can be observed for the SSDRC and the Lombard styles for the most favourable SNR, and for SSDRC at the moderate SNR. As for the subjective measure of effort, intelligibility scores shows a clear ranking of effort across speech types that is the inverse of the subjective effort rating, as well as the expected decrease in scores with increasing noise. A linear mixed model indicated a significant effect of SNR ($F[2, 40] = 73.82, p < .001, d' = 1.81$) and speech type ($F[3, 161] = 591.01, p < .001, d' = 4.33$), as well as a significant effect of their interaction ($F[6, 161] = 10.63, p < .001, d' = .74$).

4. Discussion

The current study demonstrates that listening effort varies with type of speech, as judged both by participants' own ratings and

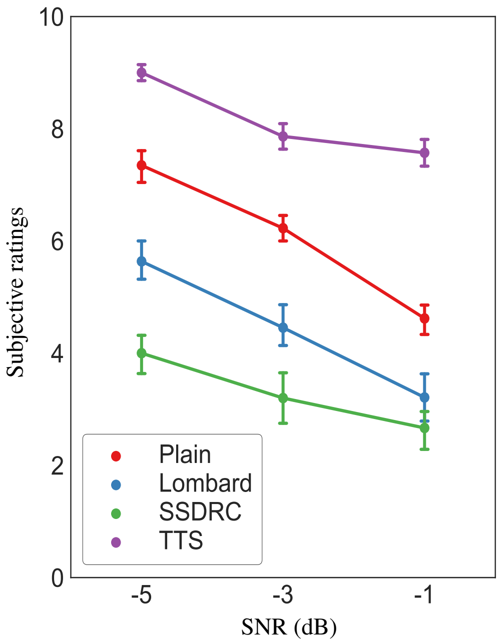

Figure 4: *Mean subjective listening effort ratings. Error bars denote* ±1 *standard error.*

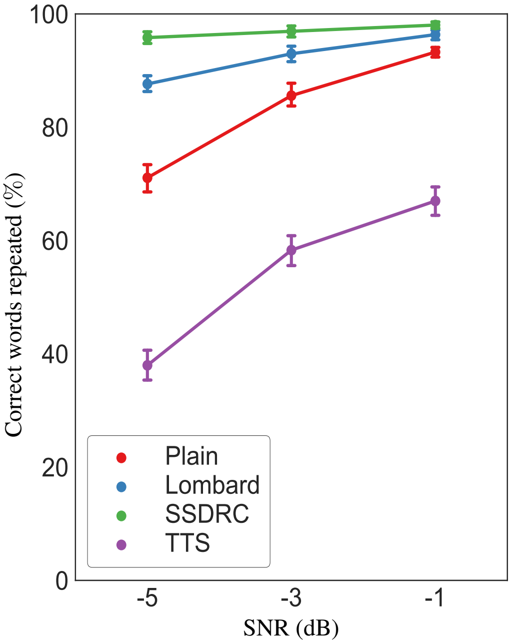

Figure 5: *Mean intelligibility scores. Error bars denote* ±1 *standard error.*

by the physiological measure of pupil size change. Synthetic speech was the least intelligible type, producing the greatest subjective ratings of effort, and the largest increase in pupil diameter over baseline in the less adverse SNRs. Speech modified algorithmically to enhance its intelligibility was most intelligible and least effortful according to participants' judgements, although peak pupil size changes were somewhat intermediate, being statistically-equivalent to Lombard speech. Lombard speech is a naturally-modified type of speech that talkers produce in response to noise. It is interesting to note that this not only results in more intelligible speech than the plain natural form, but also leads to smaller peak pupil changes at the intermediate and adverse noise levels.

Comparing our results for SSDRC and plain speech, listening effort might be related also to listeners' preference. In [24], at low SNRs, listeners showed a preference for SSDRC-modified speech over plain speech. In our results, at the adverse SNR, the difference in these two speech types is significant and this difference decreases at higher SNRs while at the most favourable SNR there is no difference. It is of interest to strengthen this finding by investigating the behaviour of pupillary responses of these speech types at noise-free conditions in which modified speech can be perceived as less natural so the effort (if it is related to preference) would be greater. On the other hand, at adverse SNRs SSDRC is more intelligible [1] and less effortful as it is shown in our study, this might be due to the fact that features related to the naturalness are less noticeable.

Previous studies [9, 12] have shown that PPD varies non-monotonically with intelligibility. Among the speech types tested, only TTS resulted in lower effort (as gauged by pupil responses) at the most adverse SNR, most likely due to its difficulty. This outcome agrees with [12], in which for plain speech the greatest effort revealed at intelligibility level around 50%. Although we only have measurements at three SNRs, our data hint at a non-monotonic relationship between SNR (and hence intelligibility) and listening effort for the TTS speech style.

Subjective ratings of listening effort are not always consistent with physiological measures. In our study, pupillary responses to synthetic speech across the different SNRs showed no clear relationship to the subjectively reported effort. Participants might have reported their opinion of their performance on the task in response to the subjective question they were asked. This finding is also supported by other studies [8, 10] concluding that subjective ratings and pupil dilation may well represent different aspects of effort [13].

5. Conclusions

In this study we examined the impact of different speech types on listening effort. Four speech types in a range of signal-to-noise ratios mixed with a speech-shaped noise masker were tested. Listening effort was measured using the peak pupil dilation and participants' ratings while comparisons with the intelligibility level were also performed. Our findings demonstrate that speech type has an impact on the effort exerted when someone is attempting to perceive speech. Lombard and modified (SSDRC) speech were shown to require less effort than synthetic (TTS) and plain speech. Further research is required in order to better understand how to synthesise speech that will be perceived with less effort. Future work will also investigate the effect of different speech types on effort for non-native listener groups.

Acknowledgements: The authors would like to thank Dr. Anita Wagner from University of Groningen. Olympia Simantiraki was funded by the European Commission under the Marie Curie European Training Network ENRICH (675324).

6. References

[1] M. Cooke, C. Mayo, C. Valentini-Botinhao, Y. Stylianou, B. Sauert, and Y. Tang, "Evaluating the intelligibility benefit of speech modifications in known noise conditions," *Speech Communication*, vol. 55, no. 4, pp. 572–585, 2013.

[2] S. L. Mattys, M. H. Davis, A. R. Bradlow, and S. K. Scott, "Speech recognition in adverse conditions: A review," *Language and Cognitive Processes*, vol. 27, no. 7-8, pp. 953–978, 2012.

[3] R. McGarrigle, K. J. Munro, P. Dawes, A. J. Stewart, D. R. Moore, J. G. Barry, and S. Amitay, "Listening effort and fatigue: What exactly are we measuring? A British society of audiology cognition in hearing special interest group 'white paper'," *International Journal of Audiology*, 2014.

[4] C. L. Mackersie and H. Cones, "Subjective and psychophysiological indexes of listening effort in a competing-talker task," *Journal of the American Academy of Audiology*, vol. 22, no. 2, pp. 113–122, 2011.

[5] Y. H. Wu, E. Stangl, X. Zhang, J. Perkins, and E. Eilers, "Psychometric functions of dual-task paradigms for measuring listening effort," *Ear and Hearing*, vol. 37, no. 6, pp. 660–670, 2016.

[6] A. Sarampalis, S. Kalluri, B. Edwards, and E. Hafter, "Objective measures of listening effort: Effects of background noise and noise reduction," *Journal of Speech, Language, and Hearing Research*, vol. 52, no. 5, pp. 1230–1240, 2009.

[7] J. E. Peelle, "Listening effort: How the cognitive consequences of acoustic challenge are reflected in brain and behavior," *Ear and Hearing*, vol. 39, no. 2, pp. 204–214, 2017.

[8] A. A. Zekveld, S. E. Kramer, and J. M. Festen, "Pupil response as an indication of effortful listening: The influence of sentence intelligibility," *Ear and Hearing*, vol. 31, no. 4, pp. 480–490, 2010.

[9] A. A. Zekveld and S. E. Kramer, "Cognitive processing load across a wide range of listening conditions: Insights from pupillometry," *Psychophysiology*, vol. 51, no. 3, pp. 277–284, 2014.

[10] T. Koelewijn, A. A. Zekveld, J. M. Festen, and S. E. Kramer, "Pupil dilation uncovers extra listening effort in the presence of a single-talker masker," *Ear and Hearing*, vol. 33, no. 2, pp. 291–300, 2012.

[11] ——, "The influence of informational masking on speech perception and pupil response in adults with hearing impairment," *Journal of the Acoustical Society of America*, vol. 135, no. 3, pp. 1596–1606, 2014.

[12] B. Ohlenforst, A. A. Zekveld, T. Lunner, D. Wendt, G. Naylor, Y. Wang, N. J. Versfeld, and S. E. Kramer, "Impact of stimulus-related factors and hearing impairment on listening effort as indicated by pupil dilation," *Hearing Research*, vol. 351, pp. 68–79, 2017.

[13] D. Wendt, T. Dau, and J. Hjortkjær, "Impact of background noise and sentence complexity on processing demands during sentence comprehension," *Frontiers in Psychology*, vol. 7, 2016.

[14] T. Koelewijn, H. de Kluiver, B. G. Shinn-Cunningham, A. A.Zekveld, and S. E. Kramer, "The pupil response reveals increased listening effort when it is difficult to focus attention," *Hearing Research*, vol. 323, pp. 81–90, 2015.

[15] M. B. Winn, J. R. Edwards, and R. Y. Litovsky, "The impact of auditory spectral resolution on listening effort revealed by pupil dilation," *Ear and Hearing*, vol. 36, no. 4, pp. 153–165, 2015.

[16] M. Cooke, C. Mayo, and C. Valentini-Botinhao, "Intelligibility-enhancing speech modifications: the Hurricane Challenge," *Interspeech*, 2013.

[17] E. H. Rothauser, W. D. Chapman, N. Guttman, H. R. Silbiger, M. H. L. Hecker, G. E. Urbanek, K. S. Nordby, and M. Weinstock, "IEEE recommended practice for speech quality measurements," *IEEE Transactions on Audio and Electroacoustics*, vol. 17, pp. 225–246, 1969.

[18] W. V. Summers, D. B. Pisoni, R. H. Bernacki, R. I. Pedlow, and M. A. Stokes, "Effects of noise on speech production: Acoustic and perceptual analyses," *Journal of the Acoustical Society of America*, vol. 84, no. 3, pp. 917–928, 1988.

[19] J. Yamagishi, T. Nose, H. Zen, Z. H. Ling, T. Toda, K. Tokuda, S. King, and S. Renals, "Robust speaker-adaptive HMM-based text-to-speech synthesis," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 6, pp. 1208–1230, 2009.

[20] T. C. Zorila, V. Kandia, and Y. Stylianou, "Speech-in-noise intelligibility improvement based on spectral shaping and dynamic range compression," *13th Annual Conference of the International Speech Communication Association*, 2012.

[21] A. Wagner, P. Toffanin, and D. Başkent, "How hard can it be to ignore the pan in panda? Effort of lexical competition as measured in pupil dilation," *18th International Congress of Phonetic Sciences (ICPhS)*, 2015.

[22] G. A. Studebaker, "A rationalized arcsine transform," *Journal of Speech, Language, and Hearing Research*, vol. 28, no. 3, pp. 455–462, 1985.

[23] P. B. Brockhoff, I. de Sousa Amorim, A. Kuznetsova, S. Bech, and R. de Lima, "Delta-tilde interpretation of standard linear mixed model results," *Food Quality and Preference*, vol. 49, pp. 129–139, 2016.

[24] Y. Tang, C. Arnold, and T. J. Cox, "A study on the relationship between the intelligibility and quality of algorithmically-modified speech for normal hearing listeners," *Journal of Otorhinolaryngology, Hearing and Balance Medicine*, vol. 1, no. 5, 2017.