INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

The effect of masking noise on listeners' spectral tilt preferences

Olympia Simantiraki[1], *Yannis Pantazis*[1], *Martin Cooke*[2]

[1]Institute of Applied and Computational Mathematics, FORTH, Greece
[2]Ikerbasque (Basque Science Foundation), Spain

simantiraki.o@iacm.forth.gr, pantazis@iacm.forth.gr, m.cooke@ikerbasque.org

Abstract

Speech enhancement algorithms often focus on optimising intelligibility while neglecting other aspects of speech such as naturalness, quality and listening effort which may affect a listener's experience. This paper investigates the impact of spectral tilt on listeners' preferences, using a new corpus of Greek utterances. Participants adjusted spectral tilt with real-time feedback to select their preferred tilt in quiet and in the presence of speech-shaped noise at eight signal-to-noise ratios. Listeners displayed distinct preferences, with a tendency to select flatter tilts with increasing noise. Preferences were not random even for constant intelligibility, indicating that their adjustments were influenced by factors beyond the need to maintain comprehensibility. These findings have the potential to inform the design of speech enhancement algorithms that jointly optimise intelligibility and a listener's overall experience.
Index Terms: listener preferences, SPEECHADJUSTER, spectral energy reallocation, glimpses profile

1. Introduction

Listeners routinely encounter pre-recorded or synthetic speech. While speech enhancement techniques have been used to improve intelligibility in potentially challenging listening conditions, these approaches typically do not account for aspects such as naturalness and quality that may also affect a listener's experience. Speech quality can have a significant impact on cognitive effort during listening tasks, even when word recognition is held constant. Synthetic voices can require increased effort compared to natural speech [1, 2, 3]. Increasing spectral resolution in a cochlear implant simulation reduced listening effort during a dual-task paradigm [4], while attending to clear (as opposed to plain) speech in the presence of babble noise resulted in lower cognitive effort [5]. Consequently, it is of interest to study listeners' preferences for features such as spectral tilt which talkers adjust naturally (resulting in a flatter spectrum) when speaking in noisy conditions (i.e. Lombard speech).

Listener preferences provide insight into listeners' overall experience with speech, taking into account factors such as naturalness, pleasantness, and loudness. One approach to studying listener preferences is through real-time auditory feedback, where listeners can modify speech characteristics until they reach their preferred settings. Previous studies have used this method to investigate preferences for formant frequency/fundamental frequency relationships [6], speech rate [7, 8], speech level [9], and local SNR [9]. Spectral modifications have been explored in studies involving individuals with hearing loss, with investigations into preferences for broadband, low-, and high-frequency gain [10] and degree of spectral tilt [11]. Our study builds on recent work into spectral tilt preferences [12], but instead of using a small set of noise levels and tilt adjustments, here we examine preferences across a broad range of SNRs and for a different target language. These aspects of the current study highlight the novelty of our approach which ought to contribute to a better understanding of the effects of spectral modifications in noisy conditions.

This study has two objectives (i) to provide a better understanding of the impact of masking noise on listeners' preferences for spectral tilt; (ii) to predict listener preferences. The latter objective is motivated by the fact that, while subjective evaluation of speech enhancement algorithms is generally considered more reliable than prediction, it can be impractical, time-consuming, and rules out using evaluation outcomes to optimise modification techniques. A large-scale validation study demonstrated that commonly-used measures of listening effort are not consistently or strongly intercorrelated [13]; other studies have shown that subjective measures are correlated with task performance [14, 15].

Automatic prediction of speech characteristics beyond intelligibility can be a valuable tool for developing speech enhancement algorithms. A DNN-based listening effort predictor, quantified via the degradation of phoneme posteriorgrams, and not requiring prior knowledge of the processed speech, was proposed in [16]. Another model to predict preferences [12] used an objective measure of energetic masking to capture intelligibility and a further Gaussian component to model supra-intelligibility factors.

In the current study participants were given the ability to adjust the spectral tilt of speech in quiet and in 8 levels of speech-shaped noise. The main questions addressed are: do listener preferences show a pattern different from intelligibility, and can the spectral profile of the masked speech signal be used to predict listener preferences?

2. Experimental design

2.1. Listeners

Twenty-three Greek listeners (6 females; 19-28 years, mean 23.7 years) were recruited for the experiment. Listeners reported no known hearing problems. An incentive of 10 euros was given for each participation.

2.2. Sentence material

A Greek corpus [17] provided sentence material for the experiments. The corpus consists of 720 semi-predictable 5-9 word sentences in modern Greek with a similar level of difficulty to the original English Harvard sentences [18]. Each sentence contains 5 keywords for scoring. Meaningful words resembling everyday language were used. An example is 'Θα χόψω το

10.21437/Interspeech.2023-1604

φρούτο σε τρία ίσα μέρη.' ('I will cut the fruit into three equal pieces.'); the keywords are underlined.

2.3. Speech material

A 31-year-old native Greek male talker was recruited to read the complete corpus. The talker was asked to read each sentence at a normal speaking rate and was able to repeat any utterance if necessary. Recording took place in a sound studio at the Speech Signal Processing Laboratory at the University of Crete (Heraklion, Greece) using Pro Tools 12 software with an RME Fireface 400 recorder. A Neumann KMS104 handheld vocal condenser microphone (cardioid directional polar pattern) was placed on a desktop microphone stand on a table at a fixed distance of 15 cm from the talker's mouth. Recordings were made at a sampling rate of 44.1 kHz. Sentences were segmented using a custom amplitude-based pause detector based on the normalised envelope of the signal. The algorithm's effectiveness and the quality of the recordings were screened manually: signals were checked for clipping, correct utterance segmentation, and common speaking style. Where necessary, recordings were repeated. Sentences had a mean duration of 2.8 s ($S.D.$ 0.3 s). For the experiment, phrases were downsampled to 16 kHz, a 20 ms half-Hamming ramp was applied at the beginning and end of each recording, and each stimulus was normalised to a common root-mean-square level.

2.4. Stimuli

Stimulus design was informed by findings in [12], whose listeners did not select extremely steep spectral tilts in any condition and in the most challenging condition may have preferred even flatter spectral tilts had they been available. Changes in spectral tilt were achieved by filtering the speech signal with a digital filter twice (*filter* function in Matlab 2016b) to produce a $|H(\omega)|^2$ system. The rational transfer function for pre-emphasis was $H(z) = 1-\lambda z^{-1}$ and for de-emphasis $H(z) = 1/(1-\lambda z^{-1})$, with λ drawn linearly from the range $[-0.16, 0.80]$, where positive and negative values correspond to pre-emphasis and de-emphasis respectively. In total 25 modification levels were constructed (4 with spectral tilt steeper than the original, 1 with the original spectral tilt, and 20 with flatter than the original) corresponding to tilts in the range $[-4.40, 2.66]$ dB/octave.

Stimuli were presented in quiet and in speech-shaped noise (SSN) at 8 SNRs: -7.5, -6, -4.5, -3, -1.5, 0, $+3$, and $+6$ dB. The masker was generated as in [12] by filtering random uniform noise with the long-term spectrum of the 700 concatenated sentences of the female talker in the Sharvard corpus [19], without gaps. The desired SNRs were obtained by rescaling the noise. The amplitude of each sentence was normalized using a fixed root-mean-square criterion.

2.5. Procedure

The experiment consisted of 5 trials in each of 9 conditions (quiet + 8 SNRs), split into 3 blocks of 15 trials in a random order. Each trial consisted of an adjustment phase followed by a test phase. In the adjustment phase, sentences were presented randomly, starting at a random feature value, with a 0.5 s gap between sentences. Participants were required to listen to at least 5 s of speech before moving on to the test phase, but they could listen to as much speech as desired during the adjustment phase. In this phase, a total of 250 unique sentences were presented. Once all 250 sentences were heard, they were shuffled and were available for presentation. This process was repeated until the experiment was completed. The test phase evaluated intelligibility via a speech perception task using the feature's value chosen at the end of the adjustment phase. Participants were presented with a sequence of two sentences and asked to type what they heard into an on-screen text box after each sentence presentation. All the sentences in the test phase were unique. Participants underwent a task familiarization phase consisting of three trials (in quiet and at -7.5 dB and +6 dB SNR).

Participants modified spectral tilt in real-time using SPEECHADJUSTER [20], receiving the instruction to adjust the speech in order to recognise as many words as possible. Adjustments were made using up/down keys on a computer keyboard. Stimuli were presented at a fixed presentation level over Sennheiser HD380 headphones. Listeners were seated in a sound-attenuating booth located in the Speech Signal Processing Laboratory at the University of Crete.

Intelligibility scores were based on the number of keywords correctly recalled in each trial (2 test phrases x 5 keywords per phrase). Written responses were post processed to remove accents over vowels and replace letter/diphthongs with the same pronunciation with a unique letter.

3. Results

Listeners preferred progressively flatter spectral tilts (all flatter than the original) as SNRs decreased (Fig. 1, top). A linear mixed effects model (*lmer* function from the *lme4* package in R) with SNR as ordered factor (the quiet condition excluded from the analysis) and participant as a random effect, indicated that SNR had a significant effect on preferences $[F(7, 1142) = 25.34, p < .001]$. The ability of linear, quadratic, cubic, and reciprocal models in predicting tilt were evaluated using leave-one-out cross-validation. The average variance of the mean spectral tilts across the 8 iterations of the cross-validation method was 0.173. Reciprocal and quadratic models predict the mean listener preferences with similar accuracy (Fig. 2).

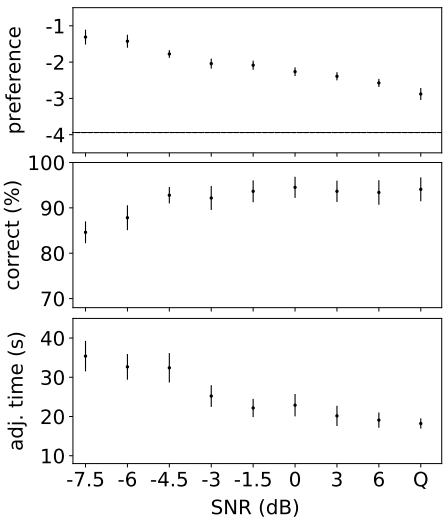

Figure 1: *Top: listeners' preferred spectral tilt relative to unmodified speech (horizontal line); middle: mean number of correctly identified keywords; lower: time spent in the adjustment phase. Q denotes the quiet condition. Error bars represent* $\pm$ *one standard error.*

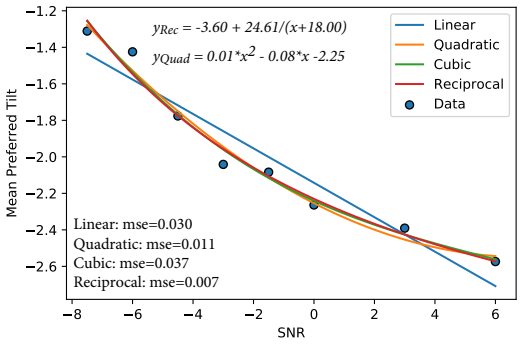

Figure 2: *Tilt model fits.*

Listeners' tilt preferences permitted them to maintain high intelligibility (Fig. 1, middle) even in challenging levels of noise with scores of 93% at -4.5 dB SNR, dropping to 85% in the most adverse condition. A linear mixed-effects model with SNR as a fixed effect and participant as a random intercept indicated a significant main effect of SNR on intelligibility [$F(8, 1234) = 16.72, p < .001$]. Post-hoc tests with Tukey corrections demonstrated that there was no significant intelligibility differences between SNRs of -4.5 dB and higher.

The increase in noise level led to an increase in the time required by listeners to finalise their tilt preferences (Fig. 1, lower). A linear mixed-effects model with SNR as a fixed effect and participant as the random intercept indicated a significant effect of SNR on adjustment time [$F(8, 1234) = 31.04, p < .001$]. Post-hoc tests with Tukey correction indicated no significant differences in adjustment time between SNRs of -1.5 dB and higher, nor between SNRs of -4.5 dB and lower.

Listener preference distributions plotted alongside intelligibility (Fig. 3, left column) show distinct preferences even when intelligibility is effectively constant, with the distributional mass shifting from steeper to flatter spectral tilts with increasing noise level. Two-sample Kolmogorov-Smirnov tests (*ks_2samp* in *scipy.stats* of Python) confirmed the non-uniformity of preference distributions at all SNRs [all $p < .001$].

The impact of spectral tilt modifications on audibility was evaluated using an energetic masking model, the Extended Glimpse Proportion metric [21]. The right column of Fig. 3 shows the proportion of glimpses of the target speech at each frequency as a function of tilt. Listeners appear to adjust spectral tilt to ensure that glimpses are available across a wide range of frequencies. This observation motivated the construction of a model to predict listener preferences as detailed below.

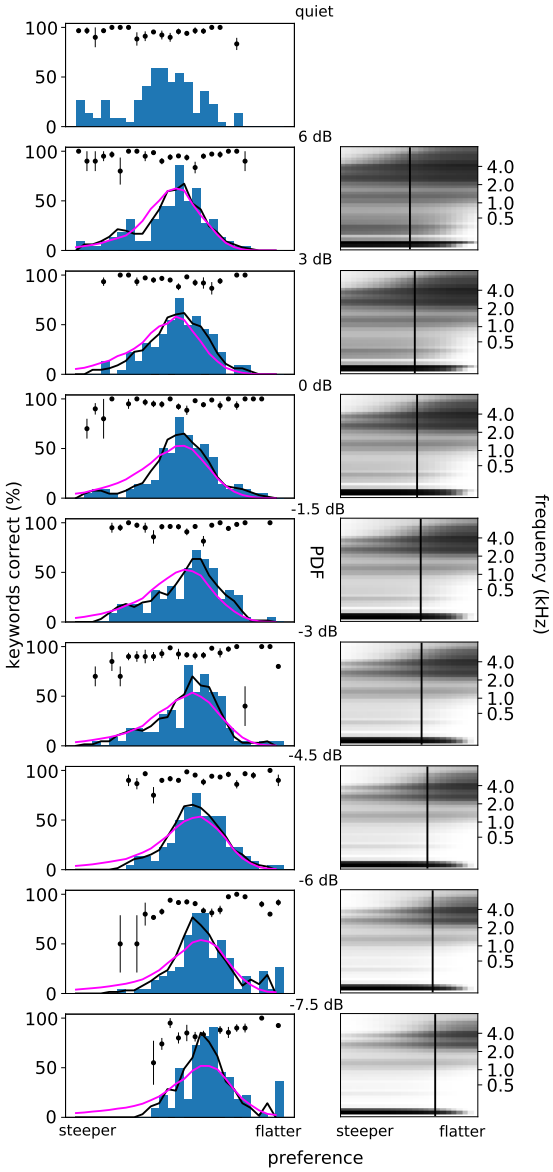

Figure 3: *Left: distribution of actual preferences (blue bars and black line), predicted preferences (magenta line), and percentage of keywords identified correctly at each step (points with error bars). Right: proportion of glimpses in each spectral region as a function of spectral tilt levels (higher proportions indicated by darker colours). Vertical bars represent the mean listener preferences. Panel depict spectral tilt on the x-axis and spectral frequency on an ERB-rate scale on the y-axis.*

4. Predicting listener preferences

We propose a model that estimates the probability of the spectral tilt of a speech signal being the most preferred. Mathematically, this probability is described by $p(y = y^*|x)$, where y denotes the spectral tilt variable while y^* is the most preferred spectral tilt, conditioned on the glimpse profile (denoted by x) which serves as a predictive indicator of listener preferences. By Bayes' theorem, the posterior probability $p(y = y^*|x) = p(x|y = y^*)p(y = y^*)/p(x)$ where $p(x)$ corresponds to the evidence which is independent of the spectral tilt y, while the prior distribution $p(y = y^*)$ is assumed to be uniform (i.e., uninformative) hence also independent of y. Therefore, Bayes' theorem states that $p(y = y^*|x) \propto p(x|y = y^*)$ with the latter being the likelihood.

We modelled the proportion of glimpses over 33 frequency bands, i.e., the likelihood, using a multivariate Gaussian distribution. The first band was excluded from the likelihood calculation due to the negligible number of glimpses observed in that band (Fig. 3 right column) which can lead to numerical instability. The use of the multivariate Gaussian process allowed us to capture inter-dependencies between the proportion of glimpses in different frequency bands, resulting in a more realistic model

SNR	sKLD (Gr)	sKLD (Sp)	sKLD (Sp) [12]
+6	0.09	-	-
+3	0.17	-	-
0	0.12	0.57	0.08
-1.5	0.20	-	-
-3	0.06	0.27	0.13
-4.5	0.24	-	-
-6	0.25	0.43	0.08
-7.5	0.57	-	-

Table 1: *Symmetric KLD between the proposed model and actual preferences for Greek (2nd col.) and Spanish (3rd col.). The 4th column corresponds to the results from the model proposed in [12] (which tested only 3 SNRs). Lower sKLD values indicate more similar distributions.*

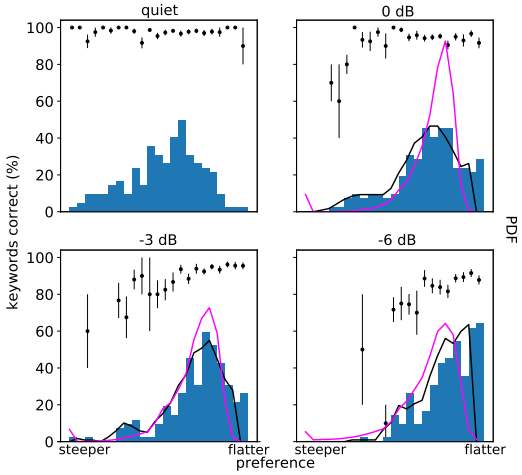

Figure 4: *As for Fig. 3 (left) but for the data of [12].*

of the speech spectral profile. The mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ were estimated using the proportion of glimpses of the most preferred tilt value for all 8 SNR. Given the most preferred tilt for 80% of the Greek speech dataset (i.e., using a training set of 576 utterances), we estimated the model parameters using maximum likelihood estimation. The overall training dataset consists of 4608 samples (8 SNRs x 576 utterances), obtained by computing the corresponding glimpse profiles solely for the most preferred tilt of each SNR condition. For testing, 20% of the Greek speech dataset (i.e. 142 utterances) was used. During testing, the proportion of glimpses of each test utterance for 33 out of 34 frequency bands, and the log-likelihood of the spectral profile were computed. Log-likelihood values were used to determine the similarity between the glimpse profile of the utterance and the model's learned distribution.

To compare the listener preferences distribution with the distribution obtained from the total log likelihoods for each spectral tilt value y we performed some preprocessing. We applied the softmax function to the total log likelihood, given by $L'(y) = \exp(L(y)/T)/\sum_{y'}(\exp(L(y')/T))$ with $T = 6$. To prevent the KLD from becoming infinite, we smoothed the preferences distribution using a moving average filter with a window size of 3. Results (magenta line in Fig. 3 left column) indicated that for most of the conditions the model fits well the listener preferences. The symmetric Kullback–Leibler divergence (sKLD) was computed (using *entropy* from *scipy.stats* in Python) and the results are reported in Table 1.

The model was further tested on a separate dataset published in [12] to validate its accuracy. In that experiment native Spanish participants adjusted Spanish sentences in quiet and at 3 SNRs (-6, -3, and 0 dB). As noted earlier, our experimental setup was similar to that study, but apart from language differences the range of spectral tilts available to listeners was different. Results are presented in Fig. 4 and Table 1.

5. Discussion

Using a real-time adjustment method, this study measured listeners' spectral tilt preferences for speech when masked at various SNRs. Clear tilt preferences are visible at all SNRs with listeners consistently preferring flatter spectral tilts as noise levels increase. The mean preferred tilt was always flatter than that of the original voice. These outcomes are consistent with findings in [12] and from Lombard speech, where talkers tend to naturally modify their speech by transferring energy to mid-frequencies when speaking in noise (e.g., [22]). Additionally, a glimpsing analysis suggested that listeners employed a consistent goal in tilt adjustment, aiming to ensure the availability of speech glimpses across a wide frequency range. Based on these findings, we proposed a model to predict listener's tilt preferences.

Listeners' adjustments were effective in maintaining intelligibility down to an SNR of nearly -5 dB. Longer adjustment times were needed for the -4.5 dB and -3 dB SNRs compared to less noisy conditions. The longer adjustment time required may indicate an increase in the cognitive effort required to process speech in noise, or the difficulty in finding an effective compromise between choosing a signal that sounds natural (i.e. with tilt close to the original) or one that preserves audibility across as much of the spectrum as possible. Data in noise is lacking, but Moore and Tan [23] found that in quiet, spectral tilt modifications have a negative impact on naturalness, particularly when applied over the entire frequency range.

Based on our observations of how listeners tend to choose spectral tilt in noise, we proposed a model to estimate the likelihood of a speech signal being preferred. An alternative approach for predicting listener preferences was suggested in [12], which involved computing glimpses for each preference level using the extended glimpse proportion metric [21] and fitting the derived distribution to the listener preferences distribution. One distribution was fit in each different noise condition. However, this approach has a limitation in predicting preferences for unseen SNRs or spectral tilts. In contrast, our model is not tailored to specific listener preference distributions, making it more capable of generalisation, as demonstrated for the Spanish data of [12]. While a comparison of columns 3 and 4 of Tab. 1 indicates that the current model performed slightly worse than that reported in [12], a fairer comparison would involve training the model used in the Spanish experiment with Greek data and then comparing the predicted listener preferences with the actual Spanish listener preferences. Further research is necessary to assess the generalisability of the current model.

6. Acknowledgements

The work of Olympia Simantiraki and of Yannis Pantazis was supported by the Hellenic Foundation for Research and Innovation (HFRI) through the "Second Call for HFRI Research Projects to support Faculty Members and Researchers" under Project 4753.

7. References

[1] C. Delogu, S. Conte, and C. Sementina, "Cognitive factors in the evaluation of synthetic speech," *Speech Communication*, vol. 24, no. 2, pp. 153–168, 1998.

[2] A. Govender and S. King, "Using pupillometry to measure the cognitive load of synthetic speech," in *Proc. Interspeech*, 2018, pp. 2838–2842.

[3] O. Simantiraki, M. Cooke, and S. King, "Impact of different speech types on listening effort," in *Proc. Interspeech*, 2018, pp. 2267–2271.

[4] C. Pals, A. Sarampalis, and D. Baskent, "Listening effort with cochlear implant simulations," *J. Speech Hearing Language Res.*, vol. 56, pp. 1075–1084, 2013.

[5] G. Borghini and V. Hazan, "Effects of acoustic and semantic cues on listening effort during native and non-native speech perception," *The Journal of the Acoustical Society of America*, vol. 147, no. 6, pp. 3783–3794, 2020.

[6] P. F. Assmann and T. M. Nearey, "Relationship between fundamental and formant frequencies in voice preference," *The Journal of the Acoustical Society of America*, vol. 122, no. 2, pp. EL35–EL43, 2007.

[7] J. S. Novak and R. V. Kenyon, "Effects of user controlled speech rate on intelligibility in noisy environments," in *Proc. Interspeech*, 2018, pp. 1853–1857.

[8] O. Simantiraki and M. Cooke, "Exploring listeners' speech rate preferences," in *Proc. Interspeech*, 2020, pp. 1346–1350.

[9] M. Torcoli, A. Freke-Morin, J. Paulus, C. Simon, and B. Shirley, "Preferred levels for background ducking to produce esthetically pleasing audio for tv with clear speech," *J Audio Eng Soc.*, vol. 67, no. 12, pp. 1003–1011, 2019.

[10] A. Boothroyd and C. Mackersie, "A "Goldilocks" Approach to Hearing-Aid Self-Fitting: User Interactions," *American Journal of Audiology*, vol. 26, no. 3S, pp. 430–435, 2017.

[11] A. T. Sabin, D. J. V. Tasell, B. Rabinowitz, and S. Dhar, "Validation of a Self-Fitting Method for Over-the-Counter Hearing Aids," *Trends in Hearing*, vol. 24, p. 2331216519900589, 2020.

[12] O. Simantiraki, M. Cooke, and Y. Pantazis, "Effects of Spectral Tilt on Listeners' Preferences And Intelligibility," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2020, pp. 6254–6258.

[13] J. F. Strand, V. A. Brown, M. B. Merchant, H. E. Brown, and J. Smith, "Measuring Listening Effort: Convergent Validity, Sensitivity, and Links With Cognitive and Personality Measures," *Journal of Speech, Language, and Hearing Research*, vol. 61, no. 6, pp. 1463–1486, 2018.

[14] J. Johnson, J. Xu, R. Cox, and P. Pendergraft, "A Comparison of Two Methods for Measuring Listening Effort As Part of an Audiologic Test Battery," *American Journal of Audiology*, vol. 24, no. 3, pp. 419–431, 2015.

[15] S. Seeman and R. Sims, "Comparison of Psychophysiological and Dual-Task Measures of Listening Effort," *Journal of Speech, Language, and Hearing Research*, vol. 58, no. 6, pp. 1781–1792, 2015.

[16] R. Huber, M. Krüger, and B. T. Meyer, "Single-ended prediction of listening effort using deep neural networks," *Hearing Research*, vol. 359, pp. 40–49, 2018.

[17] A. Sfakianaki, "Designing a Modern Greek sentence corpus for audiological and speech technology research," in *Proc. of the 14th International Conference on Greek Linguistics*, 2019.

[18] 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.

[19] V. Aubanel, M. García Lecumberri, and M. Cooke, "The Sharvard Corpus: A phonemically-balanced Spanish sentence resource for audiology," *International Journal of Audiology*, vol. 53, no. 9, pp. 633–638, 2014.

[20] O. Simantiraki and M. Cooke, "SpeechAdjuster: A Tool for Investigating Listener Preferences and Speech Intelligibility," in *Proc. Interspeech 2021*, 2021, pp. 1718–1722.

[21] Y. Tang and M. Cooke, "Glimpse-Based Metrics for Predicting Speech Intelligibility in Additive Noise Conditions," in *Interspeech 2016*, 2016, pp. 2488–2492.

[22] Y. Lu and M. Cooke, "Speech production modifications produced by competing talkers, babble, and stationary noise," *The Journal of the Acoustical Society of America*, vol. 124, no. 5, pp. 3261–3275, 2008.

[23] B. C. J. Moore and C.-T. Tan, "Perceived naturalness of spectrally distorted speech and music," *The Journal of the Acoustical Society of America*, vol. 114, no. 1, pp. 408–419, 2003.