*INTERSPEECH 2021*
30 August – 3 September, 2021, Brno, Czechia

# SPEECHADJUSTER: A tool for investigating listener preferences and speech intelligibility

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

[1] Language and Speech Laboratory, University of the Basque Country, Vitoria-Gasteiz, Spain
[2] Ikerbasque (Basque Science Foundation), Vitoria-Gasteiz, Spain

olympia.simantiraki@ehu.eus, m.cooke@ikerbasque.org

## Abstract

Most of what we know about speech perception has been gleaned from tests in which listeners respond to stimuli chosen by an experimenter. This paper presents SPEECHADJUSTER, an open source tool that reverses the roles of listener and experimenter by allowing listeners direct control of speech characteristics in real-time. This change of paradigm enables listener preferences – reflecting factors such as cognitive effort, naturalness or distortion – to be measured directly, without recourse to rating scales. Incorporation of a test phase in which listener preferences are frozen also enables intelligibility to be estimated within the same trial. Offline computation and smooth online interpolation within the tool permits the impact of changes in practically any target speech feature (e.g. fundamental frequency or spectral slope) or background characteristic (e.g. noise spectrum), regardless of complexity, to be measured. The paper describes the tool's capabilities, presents a range of visualisations, and notes some potential applications and limitations.

**Index Terms**: 'supra-intelligibility' factors, listener preferences, real-time modifications, software tool

## 1. Introduction

Listening to synthetic or artificially-modified speech is an increasingly common experience, with widespread use of voice output in applications such as public-address systems and voice assistants, often in acoustic contexts where other sound sources are present. Enhancement techniques applied to the signal prior to presentation can lead to substantial improvements in intelligibility for speech presented in challenging conditions [1, 2, 3, 4]. However, a sole focus on intelligibility enhancement risks ignoring the effect of modified or synthetic speech on other characteristics that could affect a listener's experience. One such 'supra-intelligibility' factor is the additional cognitive demand that modified speech may well impose upon listeners [5, 6, 7, 8]. Other such factors include quality, naturalness, distortion and pleasantness.

Intelligibility is readily measurable, but a different approach is required to capture attributes above and beyond word or sentence scores. Subjective preferences have traditionally been measured using rating scales [9, 10, 11], but these paradigms require a participant to map a large and potentially-complex subjectively-interpreted concept such as quality on to a rather artificial and usually discrete set of values such as 'very natural', 'quite natural' and the like. Furthermore, while intelligibility and subjective factors can be measured in the same task, for practical reasons these measurements are sequential and hence delayed relative to the stimulus, raising issues such as whether individual differences in working memory capacity might affect the outcome.

In this paper we present an alternative approach which attempts to avoid issues of interpretation and delayed responses. The technique is to provide listeners with the ability to manipulate some (usually continuous) dimension of interest in real-time, and to select a parameter setting that they judge to be in some sense optimal. Instructions given to participants are deliberately neutral, and the emphasis is simply on the discovery of a preferred setting that allows them to recognise as many words as possible. Participants are able to spend as long as necessary on exploring the available stimulus space governed by the parameter of interest. Having chosen an 'optimal' setting, listeners carry out a short task with the parameter setting frozen. For instance, the task might involve a small number of test phrases in which participants identify words in sentences.

This strategy inverts the traditional paradigm in which an experimenter decides *a priori* the specific points at which intelligibility is to be measured. Instead, listeners themselves modify the parameter being studied – which might be a speech feature such as spectral tilt, or a characteristic of an interfering signal, or some property of the overall auditory scene – with the explicit stated goal of maximising intelligibility, but with an implicit potential for tuning supra-intelligibility features. Such preferences can be expected to vary with listening environment [12, 13] and hearing impairment [14, 15] as well as having an individual component [13]. A better understanding of the basis for listener preferences promises to inform the design of speech modification algorithms that are capable of both increasing intelligibility and reducing listening effort, providing an improved overall listening experience.

The key requirement for gathering preferences via direct signal manipulation is the ability to apply the modification continuously, and in real-time. An early tool demonstrating real-time modification of speech properties was Lexicon Varispeech [16], the first commercially-available pitch shifter which in 1972 allowed users to modify speech rate while preserving normal pitch, a technique subsequently applied in speech pathology research. More recently, digital tools for real-time audio feature modifications have been suggested. In [17] and [18], listeners had to adjust certain parameters of a speech synthesizer until the speech sound was matched with a target stimulus. Listeners were presented with a grid of buttons, clicking on any of which led to playback of the corresponding stimulus. Assmann and Nearey [19] asked listeners to adjust fundamental frequency and formant frequencies to levels where the speech signal was perceived as at its most natural. Adjustments could be made by moving the computer mouse to the left and right to lower or raise the values of these properties. Tsiaras et al. [20] demonstrated a software tool for real-time speech enhancement based on reallocating speech signal energy in the frequency and time domains. Audio loudness preferences in realistic environments were investigated by Kean et al. [12] via a USB knob controller.



http://dx.doi.org/10.21437/Interspeech.2021-324

The influence of environmental noise on audio preferences was also tested by [13] where mobile audio listening was simulated. Listeners were able to adjust the background-foreground balance and overall level through a virtual interface. Real [21] and virtual [22] adjustment devices have also been used to allow participants to actively-control speech rate. Torcoli et al. [23] introduced the Adjustment/Satisfaction Test, a user-adjustable system complemented by user satisfaction assessment, applied to the evaluation of dialogue enhancement. Listeners controlled the relative level of speech with a knob and scored their satisfaction level using a rating scale. A virtual knob was used in [24] to allow listeners to modify a local signal-to-noise ratio (SNR) criterion for retaining or removing time-frequency regions of speech.

The above-mentioned approaches utilise a disparate collection of tools to implement real-time speech modification. The current paper describes SPEECHADJUSTER, an open-source, cross-platform software tool that aims to generalise these techniques, allowing manipulation of virtually any aspect of speech, and supporting joint elicitation of listener preferences and intelligibility measures. SPEECHADJUSTER operates by precomputation of modified speech for a fixed set of points on a given modification continuum, and uses smooth, rapid, mid-utterance switching to produce the sensation of continuously-variable speech alteration.

## 2. Applications

SPEECHADJUSTER has been used to investigate the effect of changes in speech rate [25] and spectral energy reallocation, including spectral tilt modifications [26]. Precomputation of stimuli permits many types of speech transformation, of arbitrary complexity, to be investigated. Examples of more complex processes include gradations in degree of foreign accent, emotional valency, or more general voice morphing. SPEECHADJUSTER could also be used to explore user preferences in some dimension of interest in speech synthesis, or to choose between families of synthesis algorithms. Other applications include the determination of optimal parameters in audio engineering in which the level of one audio signal is reduced by the presence of another signal [27] or of the proper balance between intelligibility and supra-intelligibility aspects of speech important for near-end listening enhancement algorithms [4].

In addition to testing listeners' preferences directly, SPEECHADJUSTER can help in the selection of starting parameters for conventional listening experiments with fixed conditions, and has been used in this manner in experiments on distorted speech involving sine-wave and noise-vocoded speech generation. Precomputation also allows for the possibility of experimental screening and modification of stimuli to enable artefact-removal. Figure 1 shows an example for three types of distorted speech.

## 3. SPEECHADJUSTER

For the purpose of illustration in what follows, imagine we wish to examine the possible influence of the mean fundamental frequency (F0) of a target speech signal in the presence of background noise. Participants are provided with the means to modify mean F0, and are instructed to use this control to adjust F0 in such a way as to recognise as many words as possible. The task might be described as being akin to tuning a radio set to produce the best possible signal.

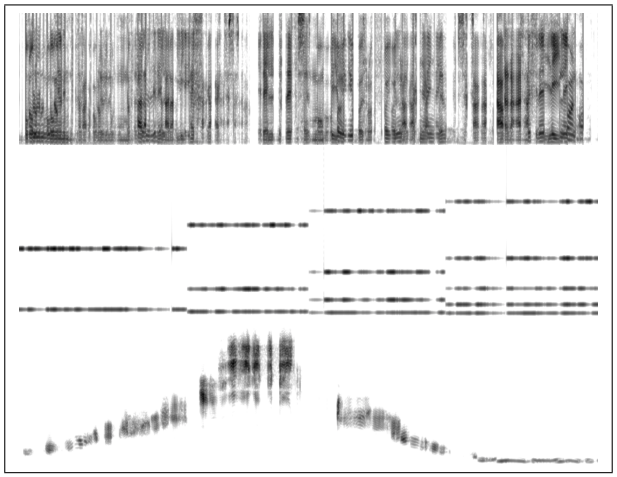

Figure 1: *Examples of the use of* SPEECHADJUSTER *to explore the parameter space of three forms of distorted speech. Top: locally time-reversed speech (varying time window size); middle: tone-vocoded speech (varying number and location of frequency bands); bottom: speech heard through a narrow spectral slit (varying the centre frequency of the slit).*

### 3.1. Adjustment and test phases

A SPEECHADJUSTER experiment consists of a sequence of trials, each of which is made up of an open-ended adjustment phase, optionally followed by a fixed-length test phase. In the adjustment phase, the listener is presented with speech material such as words, phrases or continuous speech, with or without masking noise, and the task is to explore the parameter space for the characteristic under study (e.g., mean F0) in order to find a value which the listener considers optimal in terms of understanding as many words as possible. When the participant decides that the adjustment is complete, the endpoint value of the chosen parameter is used to generate one or more test stimuli that the user responds to, just as in a traditional speech intelligibility task. During the test phase, listeners supply responses to stimuli by typing in an input box. To avoid memory load, listeners are permitted to start typing at any point after the onset of the stimulus.

### 3.2. Virtual control of speech parameters

There are many ways to elicit continuous uni-dimensional preferences. Here, five different mechanisms were explored: (1) a pair of up/down arrow buttons; (2) a mouse wheel; (3) a normal scrollbar; (4) a scrollbar whose value returns to the midpoint when released; and (5) a virtual rotary knob. In pilot experiments, five normal-hearing adult listeners with Spanish as a native language adjusted the volume level of Spanish sentences, first in quiet and then in stationary masking noise, using each of these mechanisms in independent trials. Listeners were consistent in reporting that the rotary knob and up/down arrows were the easiest to understand and apply. Consequently, SPEECHADJUSTER provides the experimenter with the choice of these two forms of input (Fig. 2).

The selection of which mechanism is the most appropriate to use will be task-dependent. The pair of arrow buttons avoids the listener having a visual indication of the current value of the parameter (apart from feedback that the upper or lower extreme has been reached). Since the listener has no indication of

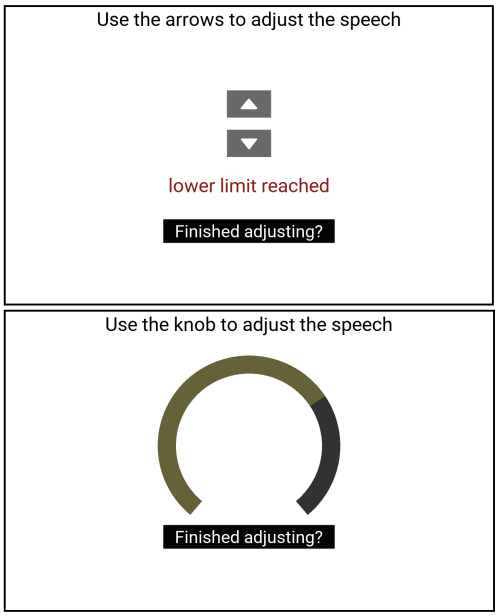


Figure 2: SPEECHADJUSTER *GUI options during the adjustment phase: (a) a pair of arrow buttons (up/down) and (b) a virtual rotary knob. During the test phase, a text input box is added.*

the parameter value at the start of each trial, they are prevented from adopting a strategy based on using the same parameter settings as on the previous trial, purely on a visual basis, since this may or may not be the most appropriate setting. On the other hand, the rotary knob can be used to simulate realistic scenarios where listeners are aware of the current parameter value and the need to adjust it from trial to trial on the basis of clear between-trial acoustic changes. To some extent, the choice will depend on how trials are blocked across conditions. For example, the choice of a mean F0 value in the presence of stationary noise might be expected to be similar from one trial to the next if the experiment is blocked by noise type, motivating the use of up/down buttons. Conversely, if the masker changes from trial to trial, or if the masker is the same but varies in some property that is likely to interact with the target speech (e.g., a competing speech masker), the mean F0 chosen might differ from trial to trial, in which case the visually more intuitive rotary knob would be preferred.

Figure 3 depicts the adjustment phase of a typical series of trials where a listener was able to control mean F0. The plot shows all the speech modifications that a listener performed via up/down buttons during the adjustment phase for each of five trials. Initial F0 values were chosen at random. These traces exhibit typical features of user-controlled exploration of parameter space: listeners sample the entire range during some trials, while other trials show more rapid adjustment phases and faster decisions, and overall there is a high level of consistency in the final value chosen. In this instance, listeners were not allowed to proceed to the test phase until five seconds had elapsed. This user-configurable value ensures that participants spend at least the specified time exploring the space of possible adjustments before signalling that they are ready for the test phase.

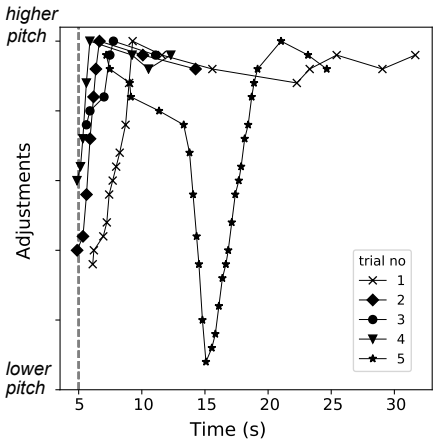


Figure 3: *A listener's F0 adjustments (y-axis) across time (x-axis) for five independent trials. The vertical dotted line indicates the time point (here* 5 *s) when the completion button (denoted Finished adjusting? in Fig. 2) in the adjustment phase was activated.*

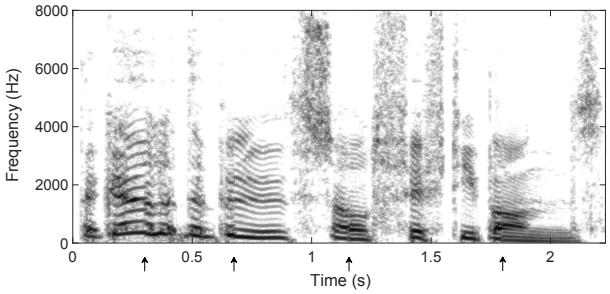


Figure 4: *A spectrogram of the speech sample 'The girl at the booth sold fifty bonds' that results from a listener making changes to mean F0 at the four time instants denoted by the arrows. The initial mean F0 is around* 290 $Hz$*, while the final value is* 130 $Hz$*.*

### 3.3. Stimulus preparation

SPEECHADJUSTER requires each stimulus (e.g. word, phrase or longer speech passage) to be precomputed at each of a range of discrete parameter values. For instance, in the case of F0, each experimental stimulus will be processed offline to produce $N$ different exemplars that differ only in mean F0, with the $N$ points along the F0 continuum chosen by the experimenter to meet some criterion such as equal-spacing on a semitone scale. The number $N$ of such levels is customisable and only impacts on the amount of offline storage required, and does not affect the latency of online processing. In our own experiments [25, 26] we have found that 20-25 discrete values are adequate to produce the impression of continuous change.

In online operation, all $N$ versions of the same stimulus (e.g. the same sentence with different mean F0) can be considered to be activated in parallel, and the user's actions control which one is actually chosen to be output by SPEECHADJUSTER at any given time point. In practice, the signal that the listener hears is merely the concatenation of segments. Switch-over is low latency and to minimise artefacts a short fade-out ramp is applied to the current segment and a similar fade-in ramp applied to the next segment corresponding to the new

stimulus.

Figure 4 shows an example speech spectrogram that results as a consequence of a listener adjusting mean F0 at several points during the utterance.

### 3.4. Configuration

The experimenter can adjust many parameters of SPEECHADJUSTER, allowing it to be adapted to the requirements of each listening task and to the linguistic background of the participants. Options include those that control the tool's appearance, the textual content and language of all interface components, participant instructions, inter-stimulus delays and numbers of trials. Other options control the size of audio chunks used during streaming, chosen to ensure that user-controlled changes are applied rapidly, but without audible artefacts. A complete list of options can be found in the user guide that is provided with the application.

### 3.5. Outputs

SPEECHADJUSTER collects detailed information during both the adjustment and test phases. In the former phase, the tool makes available both raw data in the form of time-stamps for all adjustments, and summary data on the initial and final parameter values and the total time taken to move to the test phase. Textual responses are collected during the test phase. SPEECHADJUSTER can also produce a range of figures that depict experimental outcomes. Specifically, the tool can (1) visualise the adjustments that a listener performed in a trial (as shown in Fig. 3); (2) produce a histogram of listeners' preferences; (3) generate box plots of listeners' choices and the time needed for the adjustments across the different experimental conditions; and (4) display a two-dimensional heatmap showing each listener's preferences for each of the tested phrases.

An example of the use of data produced by SPEECHADJUSTER on listeners' mean F0 preferences coupled with intelligibility scores is shown in Fig. 5. This figure illustrates that preferences tap into information over and above intelligibility: in this case, the proportion of words identified correctly is at or near ceiling across the entire range of mean F0 values, but listeners express a clear preference for values at or above the mean F0 of the original (unmodified) speech material.

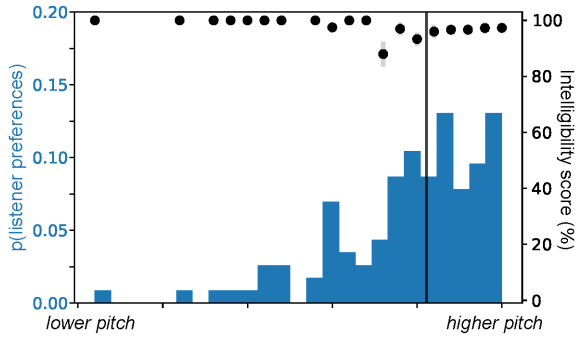


Figure 5: *Probability of each mean F0 value (histogram, left-axis), along with the percentage of words recalled correctly (black dots, right-axis). Error bars represent ± 1 standard error. The vertical line corresponds to the mean F0 of the original speech.*

### 3.6. Implementation, platforms and availability

SPEECHADJUSTER is open-source software with a GNU General Public License v3.0. SPEECHADJUSTER is written using the Python programming language and makes use of cross-platform libraries, specifically Kivy [28] for the graphical user interface and PyAudio [29] for audio streaming. Consequently, SPEECHADJUSTER can be used on Windows, OSX and Linux variants.

SPEECHADJUSTER can be installed with the command 'pip install speechadjuster'. The source code is available at `https://github.com/osimantir/speechadjuster`.

## 4. Limitations

While SPEECHADJUSTER supports the elicitation of listeners' preferences and generates information that is clearly complementary to intelligibility scores (cf. Fig. 5), it does so in a holistic manner, and consequently is unable to say anything about the weighting of individual factors that influence listeners' preferences, which may be due to listening effort, naturalness, pleasantness, attractiveness, familiarity, distortion or other quality-related considerations, and which can be expected to show substantial inter-participant variability. Further studies are needed to relate outcomes from tools such as SPEECHADJUSTER to other measures. For instance, it would be of interest to compare SPEECHADJUSTER-elicited preferences with listening effort metrics based on pupillometry [30, 31], EEG [32, 33] or self-reports [34, 35].

One limitation of the current version of the tool is that speech transformations cannot involve nonlinear modifications to the length of speech constituents, as would be the case most obviously in speech rate variation, but could also occur in modifications involving mapping between speech styles such as plain, clear or Lombard speech that typically involve changes in segment durations. However, linear elongation has been tested in [25] in which speech rate modifications were applied on single words and the changes while tuning speech were applied from the next word onwards. In principle, while linear speech rate variations are straightforward to implement, nonlinear changes will require some form of segment annotation to ensure that durational changes are applied at the correct time-points when switching from one parameter value to the next.

## 5. Conclusion

This paper describes SPEECHADJUSTER, a tool for simultaneous acquisition of listener preferences and intelligibility measures. SPEECHADJUSTER allows listeners to adjust a target speech or auditory scene property in real-time in order to optimise their listening experience, and supports studies which seek to go beyond intelligibility measures to investigate factors such as cognitive effort, naturalness or listening comfort.

## 6. Acknowledgements

Olympia Simantiraki was funded by the European Commission under the Marie Curie European Training Network ENRICH (675324).

## 7. References

[1] M. Cooke, C. Mayo, and C. Valentini-Botinhao, "Intelligibility-enhancing speech modifications: the hurricane challenge," in *IN-*

*TERSPEECH*, 2013, pp. 3552–3556.

[2] J. Rennies, H. Schepker, C. Valentini-Botinhao, and M. Cooke, "Intelligibility-Enhancing Speech Modifications - The Hurricane Challenge 2.0," in *Proc. Interspeech*, 2020, pp. 1341–1345.

[3] T. Zorila, V. Kandia, and Y. Stylianou, "Speech-in-noise intelligibility improvement based on spectral shaping and dynamic range compression," in *INTERSPEECH*, 2012, pp. 635–638.

[4] C. Chermaz and S. King, "A Sound Engineering Approach to Near End Listening Enhancement," in *Proc. Interspeech*, 2020, pp. 1356–1360.

[5] J. Rönnberg, M. Rudner, T. Lunner, and A. Zekveld, "When cognition kicks in: Working memory and speech understanding in noise," *Noise Health*, vol. 12, no. 49, pp. 263–269, 2010.

[6] C. Pals, A. Sarampalis, and D. Başkent, "Listening Effort With Cochlear Implant Simulations," *J Speech Lang Hear Res*, vol. 56, no. 4, pp. 1075–1084, 2013.

[7] B. Hornsby, "The Effects of Hearing Aid Use on Listening Effort and Mental Fatigue Associated With Sustained Speech Processing Demands," *Ear Hear*, vol. 34, pp. 523–534, 2013.

[8] A. Govender and S. King, "Using Pupillometry to Measure the Cognitive Load of Synthetic Speech," in *Proc. Interspeech*, 2018, pp. 2838–2842.

[9] R. Moore, E. Adams, P. Dagenais, and C. Caffee, "Effects of reverberation and filtering on speech rate judgment," *Int J Audiol*, vol. 46, no. 3, pp. 154–160, 2007.

[10] E. M. Adams and R. E. Moore, "Effects of speech rate, background noise, and simulated hearing loss on speech rate judgment and speech intelligibility in young listeners," *J Am Acad Audiol*, vol. 20, pp. 28–39, 2009.

[11] I. Brons, R. Houben, and W. Dreschler, "Perceptual Effects of Noise Reduction With Respect to Personal Preference, Speech Intelligibility, and Listening Effort," *Ear Hear*, vol. 34, pp. 29–41, 2013.

[12] J. Kean, E. Johnson, and E. Sheffield, "Study of audio loudness range for consumers in various listening modes and ambient noise levels," 2015. [Online]. Available: {http://www.aes.org/technical/documentDownloads.cfm?docID=523}

[13] T. Walton, M. Evans, D. Kirk, and F. Melchior, "Does Environmental Noise Influence Preference of Background-Foreground Audio Balance?" in *AES Conv. 141*, 2016.

[14] W. Buyens, B. van Dijk, M. Moonen, and J. Wouters, "Music mixing preferences of cochlear implant recipients: A pilot study," *Int J Audiol*, vol. 53, no. 5, pp. 294–301, 2014.

[15] B. Shirley, M. Meadows, F. Malak, J. Woodcock, and A. Tidball, "Personalized Object-Based Audio for Hearing Impaired TV Viewers," *J Audio Eng Soc*, vol. 65, no. 4, pp. 293–303, 2017.

[16] F. F. Lee, "Time Compression and Expansion of Speech by the Sampling Method," *J Audio Eng Soc*, vol. 20, no. 9, pp. 738–742, 1972.

[17] K. Johnson, E. Flemming, and R. Wright, "The Hyperspace Effect: Phonetic Targets Are Hyperarticulated," *Language*, vol. 69, pp. 505–528, 1993.

[18] K. Johnson and P. Ladefoged, "A preliminary perceptual study of the vowels of Montana Salish: The method of adjustment as a fieldwork technique," *UCLA Working Papers in Phonetics*, vol. 87, pp. 105–111, 1994.

[19] P. Assmann and T. Nearey, "Relationship between fundamental and formant frequencies in voice preference," *J Acoust Soc Am*, vol. 122, no. 2, pp. EL35–EL43, 2007.

[20] V. Tsiaras, T. Zorila, Y. Stylianou, and M. Akamine, "Real time speech-in-noise intelligibility enhancement based on spectral shaping and dynamic range compression," in *Proc. ICASSP (Show and Tell)*, 2014.

[21] A. Wingfield and J. Ducharme, "Effects of Age and Passage Difficulty on Listening-Rate Preferences for Time-Altered Speech," *J Gerontol: Series B*, vol. 54B, no. 3, pp. P199–P202, 1999.

[22] J. Novak and III and R. Kenyon, "Effects of User Controlled Speech Rate on Intelligibility in Noisy Environments," in *Proc. Interspeech*, 2018, pp. 1853–1857.

[23] M. Torcoli, J. Herre, J. Paulus, C. Uhle, H. Fuchs, and O. Hellmuth, "The Adjustment/Satisfaction Test (A/ST) for the Subjective Evaluation of Dialogue Enhancement," in *AES Conv. 143*, 2017.

[24] Z. Zhang and Y. Shen, "Listener Preference on the Local Criterion for Ideal Binary-Masked Speech," in *Proc. Interspeech*, 2019, pp. 1383–1387.

[25] O. Simantiraki, M. Cooke, and Y. Pantazis, "Effects of Spectral Tilt on Listeners' Preferences And Intelligibility," in *ICASSP - IEEE*, 2020, pp. 6254–6258.

[26] O. Simantiraki and M. Cooke, "Exploring Listeners' Speech Rate Preferences," in *Proc. Interspeech*, 2020, pp. 1346–1350.

[27] 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.

[28] Kivy Team and other contributors, "Kivy." [Online]. Available: https://kivy.org

[29] H. Pham, "PyAudio: Python Bindings for PortAudio." [Online]. Available: https://pypi.org/project/PyAudio/

[30] D. Kahneman, "Attention and Effort," *New Jersey: Prentice-Hall Inc*, 1973.

[31] A. Zekveld, S. Kramer, and J. Festen, "Cognitive Load During Speech Perception in Noise: The Influence of Age, Hearing Loss, and Cognition on the Pupil Response," *Ear Hear*, vol. 32, pp. 498–510, 2011.

[32] P. Sauseng and W. Klimesch, "What does phase information of oscillatory brain activity tell us about cognitive processes?" *Neurosci Biobehav Rev*, vol. 32, no. 5, pp. 1001–1013, 2008.

[33] J. Obleser, M. Wöstmann, N. Hellbernd, A. Wilsch, and B. Maess, "Adverse listening conditions and memory load drive a common alpha oscillatory network," *J Neurosci*, vol. 32, pp. 12 376–12 383, 2012.

[34] S. Gatehouse and W. Noble, "The Speech, Spatial and Qualities of Hearing Scale (SSQ)," *Int J Audiol*, vol. 43, pp. 85–99, 2004.

[35] M. Rudner, T. Lunner, T. Behrens, T. Sundewall, and J. Rönnberg, "Working memory capacity may influence perceived effort during aided speech recognition in noise," *J Am Acad Audiol*, vol. 23, pp. 577–589, 2012.