Interspeech 2024
1-5 September 2024, Kos, Greece

Listeners' F0 preferences in quiet and stationary noise

Olympia Simantiraki[1], *Martin Cooke*[2,3]

[1]Institute of Applied and Computational Mathematics, FORTH, Greece
[2]Ikerbasque, Bilbao, Spain
[3]University of the Basque Country, Vitoria-Gasteiz, Spain
simantiraki.o@iacm.forth.gr, m.cooke@ikerbasque.org

Abstract

Talkers – and increasingly speech output technology – typically alter speech characteristics when faced with challenging communicative conditions, but the impact of these changes on interlocutors is not fully understood. One such characteristic is fundamental frequency (F0), whose mean and range tend to increase when talking in noise or when communicating with inexperienced listeners. However, speech perception experiments have yet to demonstrate any intelligibility advantage for F0 modifications. The current study asked listeners to alter mean F0 or F0 variation with real-time feedback, in order to maximise comprehensibility in quiet and noise. Listeners chose a lower mean F0 than that of the original, and F0 variation similar to the original. Masker level had no effect on preference, suggesting that while listeners express clear choices, adjustment of F0 has no impact on intelligibility, and may instead reflect considerations such as naturalness or listening effort.

Index Terms: F0, speech modification, intelligibility, preferences

1. Introduction

Listeners are frequently confronted by pre-recorded or synthetic speech in domestic and public settings. These non-live forms of speech offer the potential to modify the audio signal prior to output. Most previous work on speech modification has been aimed at improving intelligibility in adverse conditions [1, 2, 3], and has been highly-effective for both natural and synthetic speech [4, 5]. Other studies have measured the effect of speech modification on characteristics such as naturalness [6], quality [7, 8] and listening effort [9, 10].

A key issue concerns what types of modification are effective, and why. One way to address this question is by examining the forms of adjustment made by talkers, especially when faced with challenging listening conditions or when conversing with interlocutors lacking experience with the language. Among other characteristics that talkers modify in such contexts, changes in fundamental frequency (F0) are often observed, typically taking the form of increases in mean F0 [11, 12] or exaggerated F0 modulation [13]; see [14] for a review. However, the overall effect of F0 modifications on listeners is currently unclear. While F0 plays a clear role in conveying prosody and distinguishing competing talkers [15], there is little evidence to date that changes in mean F0 benefit intelligibility, at least under quiet and stationary noise conditions [16, 17], although the absence of F0 variation can lead to reduced intelligibility [18, 19, 20]. A higher F0 may also facilitate signal detection in the infant brain: infants mature earlier in their perception of higher-pitched sounds [21].

Relying on modifications made by talkers risks conflating changes that are under a talker's control with those that arise as a side-effect. For example, it has long been known that F0 increases occur as a consequence of increased vocal effort [22]. However, recent technological advances in real-time speech modification [23] facilitate an alternative experimental paradigm that can be used to disentangle valuable from inconsequential modifications. The idea is to provide listeners with the ability to adjust a speech signal with immediate auditory feedback, with the goal of finding a setting that meets some criterion such as maximising comprehensibility. This listener-centric approach to finding effective ways to modify speech is both very natural for participants and efficient in permitting a fine-grained search through the space of possible modifications. Following pioneering work in [24], more recent studies have used listener-centric methods to examine preferences involving speech rate [25], formant/F0 relationships [26], spectral energy allocation [27] and local signal-to-noise ratio (SNR) [28].

The current study uses the listener-centric approach to examine listeners' F0 preferences for sentences presented in quiet and in varying levels of stationary noise. In separate trials, listeners modified either mean F0 or F0 variation, with real-time feedback. Having selected a preferred value of either feature, they then identified keywords in sentences whose F0 characteristics were modified at the chosen value. Our first research question (RQ1) concerns whether listeners show distinct preferences for F0 (mean or variation) under conditions where intelligibility is expected to be at ceiling levels. The second question (RQ2) asks whether preferences in more challenging conditions differ from those in quiet. If F0 makes an independent contribution to intelligibility, we would expect to observe differences in F0 preferences in quiet and noisy conditions. Further, on the basis of aforementioned studies into talkers' modifications in challenging conditions we would anticipate these changes to result in increases in mean F0 or exaggerated F0 variation.

2. Methods

Since no appropriate public dataset was available, we collected new data to address our research questions. This study received approval from the Research Ethics Committee of Foundation for Research and Technology – Hellas (FORTH) (Approval Reference: 424/31-6/8.10.2020).

2.1. Participants

Seventeen Greek monolingual listeners (10 female) participated in the experiment. All were young adults (age $19-33$; mean 24.2; standard deviation 3.8). No listener reported hearing problems.

10.21437/Interspeech.2024-14

2.2. Speech stimuli

Speech material was drawn from the GrHavard Corpus [29] which consists of semi-predictable Greek sentences similar in complexity to the English Harvard material [30]. Sentences were uttered by a 31-year-old native Greek male speaker at a normal speaking pace. The mean F0 of the corpus was approximately 130 Hz, with a standard deviation (s.d.) of 20 Hz.

The original F0 contour of each sentence was modified according to eq. 1:

$$f_0{}' = \frac{f_0 - \mu}{\sigma} \cdot \sigma' + \mu' \quad (1)$$

where μ and σ denote the mean and s.d. of F0, and the primed variables indicate modified versions.

Changes in mean F0 were performed as a simple shift in the entire contour i.e. $\sigma' = \sigma$. Conversely, for F0 variation the mean was held constant i.e. $\mu' = \mu$. All F0 modifications were carried out using the PSOLA algorithm [31].

As described in sec. 2.3 below, stimuli were delivered by a tool [23] that uses precomputation of stimuli at a range of modification steps. For each of the two features – mean and variation in F0 – stimuli were constructed at 25 modification steps (denoted $\lambda = 0, 1, ..., 24$). For mean F0, steps were spaced according to eq. 2:

$$\mu' = \mu + k * (1 + r)^{\lambda} + m \quad (2)$$

where the constants $k = 250$, $r = -0.1$ and $m = -65$ were chosen to avoid very low or high values of F0. An exponential spacing was motivated by the finding that listeners prefer F0 values close to those of the original [26]. Changes in F0 variation followed eq. 3:

$$\sigma' = \sigma \cdot (\delta + k * (1 + r)^{\lambda}) \quad (3)$$

where $k = 10$, $r = -0.2$ and $\delta = 1e - 6$, the latter preventing variation going to zero. Any F0 values lower than 75 Hz or higher than 500 Hz were mapped linearly to the range $[75, 80]$ or $[450, 500]$ respectively.

Sentences were presented in quiet, or mixed with speech-shaped noise (SSN) at -3, 0 and $+3$ dB signal-to-noise ratio (SNR). The SSN masker was generated by filtering random uniform noise with the long-term spectrum of the 720 concatenated sentences (without gaps) of the GrHarvard corpus.

2.3. Procedure

Listeners modified speech in real-time via the open source tool SpeechAdjuster [23]. SpeechAdjuster operates by precomputing sentences for each of a range of modification steps, using smooth interpolation to provide the impression of continuous variation. Each trial in SpeechAdjuster consists of an adjustment phase in which participants are allowed to modify as much speech material as necessary to achieve some goal, followed immediately by a test phase, where they are asked to identify one or more sentences presented using the modification values selected at the end of the adjustment process. Different sentences were presented in the adjustment and test phases.

The experiment was blocked by condition (N=4; quiet and 3 SNRs). Each block contained 10 trials, 5 for modifying mean F0 and 5 trials for modifying F0 variation, with a randomised trial order. To ensure that listeners explored the F0 space afresh on each trial, the mean F0 or F0 variation of the initial sentence in any trial was set at random to one of the 25 available steps. Listeners modified the feature by clicking on one of two on-screen buttons labelled with up and down arrows. Participants had to listen to at least 5 s of speech in the adjustment phase before proceeding to the test phase. In this latter phase participants typed what they heard into an on-screen text box. Participants underwent a task familiarisation phase consisting of 3 trials, 1 in quiet and 2 in noise. Sentence IDs 350–575, 576–656, and 714–720 were used for the adjustment, test and practice phases respectively.

Listeners were asked to modify the speech until they could "recognise as many words as possible". All information was provided to listeners both orally and in written form in their native language. Stimuli were presented through Sennheiser HD380 Pro headphones. The presentation level was not controllable by listeners, but was preset at a level that pilot experiments indicated produced a comfortable listening level. Across participants, block order was counterbalanced using a Latin square design. Experiments on average lasted around one hour and participants were able to take a short break at the end of each block. The experiment took place in a sound-proof room at the Speech Signal Processing Laboratory, University of Crete.

2.4. Post-processing

Intelligibility scores were computed based on the number of keywords correctly recalled on each trial (2 test phrases $\times$ 5 keywords per phrase). Prior to scoring, all accents over vowels were removed and letter/diphthongs with the same pronunciation were replaced with a unique letter.

Outlier analysis, based on identifying participants with scores or adjustment times (combined across SNRs) more than 1.5 times the inter-quartile range either below the first quartile or above the third quartile boundaries, led to the removal of data from one participant who was a low-scoring outlier for both the mean F0 and F0 variation features. Analyses are based on the remaining 16 participants.

2.5. Statistical procedures

Separate mixed-effects models were constructed in R [32] to predict (i) listeners' preferences (ii) intelligibility and (iii) time spent in the adjustment phase. Random effects included slopes and intercepts for participants and sentences; the maximal random-effects structure for each model was assessed using model comparison and is specified in secs. 3.1–3.3 below.

3. Results

3.1. Preferences

Across-listener mean F0 preferences relative to the F0 of the original (unmodified) speech, computed via Eqn. 2 using the per-listener median step chosen, are depicted in the upper panel of Fig. 1. These indicate that in quiet and in all masked conditions, on average listeners preferred to adjust mean F0 to a lower value than that of the original speech. A similar computation for F0 variation based on Eqn. 3 and expressed as the ratio σ'/σ suggests that listeners modified the speech signal to have a variation close to that of the original (lower panel of Fig. 1).

Separate linear mixed-effects models with SNR as a fixed effect and per-participant random intercepts predicted the median step chosen by listeners for mean F0 and F0 variation respectively, via the `lmer` function from the `lmerTest` package [33]. These analyses indicated that (i) listeners chose to decrease mean F0 in all conditions [$t(60.4) = -2.83, p <$

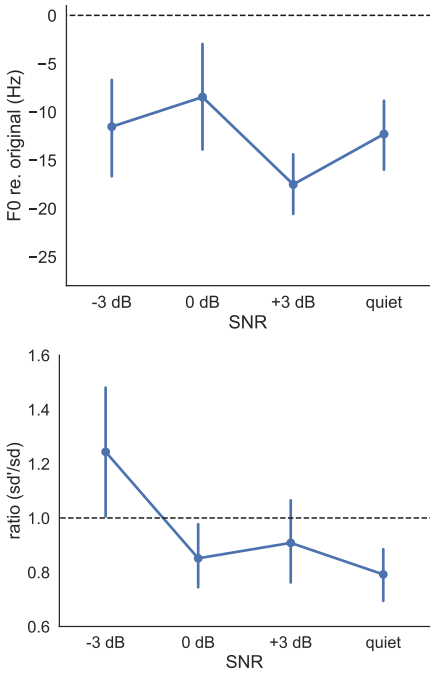

Figure 1: *Upper: Across-listener average of mean F0 preferences relative to the F0 of the unmodified speech, in quiet and at 3 SNRs. Lower: Listeners' F0 variation preferences, expressed as a ratio relative to that of the unmodified speech. Dotted lines show values for the unmodified speech. Error bars, here and elsewhere, denote ± 1 standard error.*

.01]; (ii) SNR has no effect on choice of mean F0 [$\chi^2(3) = 2.54, p = .47$]; and (iii) F0 variation preferences did not differ statistically from the original speech [$t(48.1) = -1.30, p = .20$], although there was a tendency towards an effect of SNR [$\chi^2(3) = 6.73, p = .08$] due to the preference for greater F0 variation in the -3 dB condition.

3.2. Intelligibility

Fig. 2 indicates that listeners were able to maintain scores close to ceiling in the quiet and 3 dB SNR conditions, with the expected progressive fall-off in scores as the level of the masker increased. There was no advantage in manipulating mean F0 over F0 variation, or vice versa: both features led to similar scores.

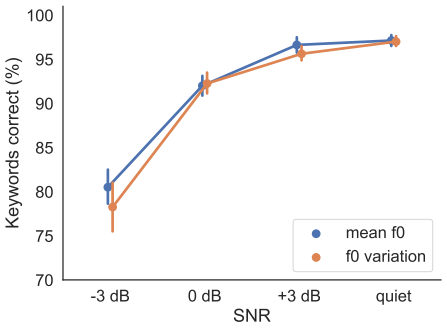

Figure 2: *Intelligibility scores for quiet and masked conditions as a result of modifying mean F0 or F0 variation.*

A generalised linear mixed-effects model via the `glmer` function from the `lme4` library [34] predicted the proportion of keywords identified correctly in each trial. The minimal model had SNR and feature as fixed effects, and both by-participant and by-sentence random intercepts and per-SNR slopes. The model indicated the expected effect of SNR [$\chi^2(3) = 107, p < .001$] but no difference between the features [$p = .33$], nor any interaction with SNR [$p = .81$]. Post-hoc tests using the `emmeans` library [35] suggested that intelligibility in the quiet and least noisy conditions were equivalent [mean F0: $p = .99$; F0 variation: $p = .37$].

3.3. Adjustment time

Listeners spent increasing amounts of time during the adjustment phase as noise level increased, and took more time to explore F0 variation than mean F0 (fig. 3).

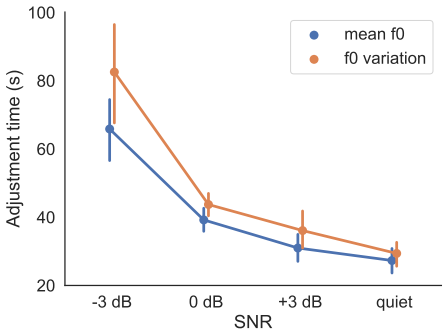

Figure 3: *Time spent during the adjustment phase for quiet and masked conditions as a result of modifying mean F0 or F0 variation.*

A linear mixed-effects model was constructed using `lmer` to predict adjustment time on each trial. The minimal model had SNR and feature as fixed effects, and by-participant random intercepts and per-SNR slopes, and indicated a moderate interaction between SNR and feature [$\chi^2(3) = 16.1, p < .01$] and main effects of SNR [$\chi^2(3) = 18.2, p < .001$] and feature [$\chi^2(1) = 8.6, p < .01$]. Post-hoc tests indicated that adjustment time was significantly longer only in the -3 dB condition [$t(60.5) = 4.9, p < .001$], and that adjustment times in noise differed from those in quiet except in the least noisy condition for both features [mean F0: $p = .69$; F0 variation: $p = .20$].

3.4. Preferences vs intelligibility

The mean preferences shown in fig. 1 provide only a partial picture of listeners' choices, and disguise a substantial dispersion in the final adjustment steps chosen. Fig. 4 depicts listeners' preferences as a distribution across steps, along with intelligibility scores based on responses at each step. These distributions show that listeners had clear preferences even when intelligibility was near to ceiling levels. Indeed, 2-sample Kolmogorov-Smirnov tests using function `ks_2samp` in `scipy.stats` [36] demonstrated that preference distributions were non-uniform for both types of modification at all SNRs [max $p = .007$]. In general, these distributions and corresponding intelligibilities highlight a lack of dependence of intelligibility on either mean F0 or F0 variation across a wide range of modifications, apart from a hint that much lower mean F0 might be detrimental in the most adverse masking condition.

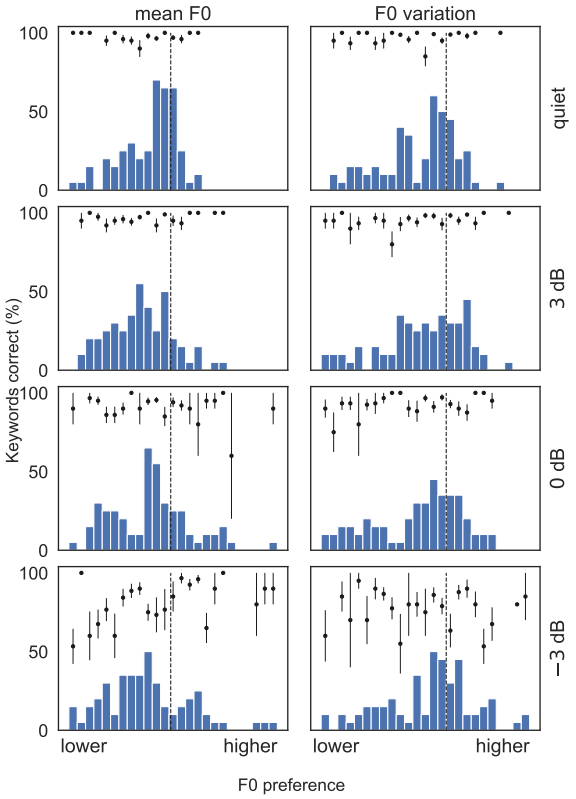

Figure 4: *Distribution of steps chosen by listeners. Dots indicate the mean percentage of keywords identified correctly at that step value, across all listeners. Dotted lines denote the step that corresponds to mean F0 or F0 variation of the original speech.*

4. Discussion

4.1. Listeners exhibit clear F0 preferences

When asked to modify a continuous stream of sentences in order to identify as many words as possible, listeners exhibited clear preferences in their choice of mean F0 and F0 variation. These preferences were present even in quiet and low noise conditions where intelligibility was at ceiling levels (RQ1). Two pieces of evidence suggest that listeners completed the task assiduously: (i) although they could listen to as little as 5 s of speech during adjustment, on average listeners spent at least 30 s per trial in the most favourable conditions; and (ii) the F0 applied to the first sentence was uniformly-distributed over the available modification steps; if listeners didn't care about F0, we would observe a more uniform distribution of final preferences.

It is notable from the preference distributions of Fig. 4 that participants generally avoided very high mean F0. Several studies [37, 38] have found that very high mean F0 values lead to poor vowel identification compared to lower F0 values, possibly due to a reduction in the number of resolved harmonics.

4.2. F0 preferences do not change in challenging conditions

In moderate and high noise levels, listeners' preferences were broadly similar to those in quiet and at a low noise level (RQ2). Listeners chose to reduce mean F0 by more than 12 Hz relative to an original mean F0 of 130 Hz, but did not significantly modify F0 variation. Listeners spent around twice as long selecting an appropriate F0 value in more challenging conditions, and spent longer when manipulating F0 variation than mean F0. The finding that listeners chose similar reductions in quiet and noise supports earlier findings [16, 17] that mean F0 plays little if any role in intelligibility under noisy conditions. The fact that listeners preferred to reduce rather than increase mean F0, as would have been expected from studies of speech production in noise (Lombard speech; [17]), suggests that F0 changes in Lombard speech occur as a consequence of other changes, notably increases in vocal effort. These outcomes suggest that even though listeners were asked to optimise comprehensibility, their final preferences were influenced by other considerations such as naturalness. Speech is perceived as more natural [39] and is preferred by listeners [26] when F0 and formants are in an appropriate relationship; it might have been the case that a lower F0 provided a more natural setting for the talker's voice.

4.3. Implications

In contrast to speech modification algorithms influenced by (or trained on) Lombard speech e.g. [2], which will typically increase mean F0, our findings suggests that to optimise the overall listening experience (i.e. incorporating naturalness, cognitive effort or speech quality as well as intelligibility), the impact of reductions in mean F0 should be investigated. Further, providing the technological means to adjust speech characteristics such as F0 may lead to a better listening experience.

4.4. Limitations and further work

As noted in sec. 4.2, the observed non-uniform distribution of preferred F0 indicates that listeners were most likely making their decision based on other factors such as cognitive effort or naturalness which the current experimental paradigm cannot distinguish. More focused follow-up studies will be required, for example, using techniques such as pupillometry [40] to provide a physiological measure of effort as a function of F0. A further limitation is that the study involved stationary masking noise. It is possible that F0 preferences will change for other maskers, particularly those that contain F0 information (e.g. competing talkers). Finally, the study involved just a single male talker; further work is required to determine whether F0 preferences are talker- or gender-specific.

5. Conclusions

When given the option to modify the F0 of a talker's voice in real-time, on average listeners chose to lower mean F0 and to retain the original F0 variation, in both quiet and noisy conditions. These findings suggest that the increased mean F0 and exaggerated F0 range observed in talkers' speech when communicating under challenging conditions is a side-effect of increased vocal effort, or related to non-intelligibility goals such as promoting attentional salience or increasing speech affect.

6. Acknowledgements

This work has been co-financed by the European Commission under the Marie Curie European Training Network ENRICH (675324) and 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] B. Sauert and P. Vary, "Near end listening enhancement: Speech intelligibility improvement in noisy environments," in *Int. Conf. Acoustics, Speech and Sig. Proc.*, 2006, pp. 493–496.

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

[3] H. Schepker, J. Rennies, and S. Doclo, "Speech-in-noise enhancement using amplification and dynamic range compression controlled by the speech intelligibility index," *J. Acoust. Soc. Am.*, vol. 138, no. 5, pp. 2692–2706, 2015.

[4] 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.

[5] C. Chermaz and S. King, "A sound engineering approach to near end listening enhancement," in *Proc. Interspeech*, 2020, pp. 1356–1360.

[6] S. Moller, *Assessment and prediction of speech quality in telecommunications*. Springer, Berlin, 2000.

[7] T.-C. Zorilă and Y. Stylianou, "On the quality and intelligibility of noisy speech processed for near-end listening enhancement," in *Proc. Interspeech*, 2017, pp. 2023–2027.

[8] Y. Tang, C. Arnold, and T. Cox, "A study on the relationship between the intelligibility and quality of algorithmically-modified speech for normal hearing listeners," *J. Otorhinolaryngol. Hear. Balance Med.*, vol. 1, no. 1, 2018.

[9] J. Rennies, A. Pusch, H. Schepker, and S. Doclo, "Evaluation of a near-end listening enhancement algorithm by combined speech intelligibility and listening effort measurements," *J. Acoust. Soc. Am.*, vol. 144, no. 4, pp. EL315–EL321, 2018.

[10] A. Govender and S. King, "Measuring the cognitive load of synthetic speech using a dual task paradigm," in *Proc. Interspeech*, 2018, pp. 2843–2847.

[11] A. R. Bradlow, N. Kraus, and E. Hayes, "Speaking Clearly for Children With Learning Disabilities," *J. Speech Lang. Hear. Res.*, vol. 46, no. 1, pp. 80–97, 2003.

[12] 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," *J. Acoust. Soc. Am.*, vol. 84, no. 3, pp. 917–928, 1988.

[13] M. Uther, M. Knoll, and D. Burnham, "Do you speak E-NG-LI-SH? A comparison of foreigner- and infant-directed speech," *Speech Comm.*, vol. 49, no. 1, pp. 2–7, 2007.

[14] M. Cooke, S. King, M. Garnier, and V. Aubanel, "The listening talker: A review of human and algorithmic context-induced modifications of speech," *Comp. Speech & Lang.*, vol. 28, pp. 543–571, 2014.

[15] J. Bird and C. J. Darwin, "Effects of a difference in fundamental frequency in separating two sentences," *Psychophysical and Physiological Advances in Hearing, edited by Palmer A. R., Rees A., Summerfield A. Q., and Meddis R.*, pp. 263–269, 1998.

[16] P. F. Assmann, T. M. Nearey, and J. M. Scott, "Modeling the perception of frequency-shifted vowels," *Int. Conf. Spoken Lang. Proc.*, pp. 425–428, 2002.

[17] Y. Lu and M. Cooke, "The contribution of changes in F0 and spectral tilt to increased intelligibility of speech produced in noise," *Speech Comm.*, vol. 51, no. 12, pp. 1253–1262, 2009.

[18] A. Wingfield, L. Lombardi, and S. Sokol, "Prosodic Features and the Intelligibility of Accelerated Speech," *J. Speech Lang. Hear. Res.*, vol. 27, no. 1, pp. 128–134, 1984.

[19] J. S. Laures and G. Weismer, "The Effects of a Flattened Fundamental Frequency on Intelligibility at the Sentence Level," *J. Speech Lang. Hear. Res.*, vol. 42, no. 5, pp. 1148–1156, 1999.

[20] P. J. Watson and R. S. Schlauch, "The effect of fundamental frequency on the intelligibility of speech with flattened intonation contours," *American Journal of Speech-Language Pathology*, vol. 17, no. 4, pp. 348–355, 2008.

[21] B. Schneider and S. Trehub, "Sources of developmental change in auditory sensitivity." *L. A. Werner & E. W. Rubel (Eds.), Developmental psychoacoustics*, pp. 3–46, 1992.

[22] I. R. Titze, "On the relation between subglottal pressure and fundamental frequency in phonation," *J. Acoust. Soc. Am.*, vol. 85, pp. 901–906, 1989.

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

[24] A. Wingfield and J. L. Ducharme, "Effects of age and passage difficulty on listening-rate preferences for time-altered speech," *The Journals of Gerontology: Series B*, vol. 54B, no. 3, pp. P199–P202, 1999.

[25] 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.

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

[27] O. Simantiraki and M. Cooke, "Listeners' spectral reallocation preferences for speech in noise," *Applied Sciences*, vol. 13(15), p. 8734, 2023.

[28] Z. Zhang and Y. Shen, "Listener preference on the local criterion for ideal binary-masked speech," in *Proc. Interspeech*, 2019, pp. 1383–1387.

[29] A. Sfakianaki, "Designing a Modern Greek sentence corpus for audiological and speech technology research," in *Proc. of the 14th International Conference on Greek Linguistics (ICGL14)*, 2019.

[30] 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.

[31] F. Charpentier and M. Stella, "Diphone synthesis using an overlap-add technique for speech waveforms concatenation," in *Int. Conf. Acoustics, Speech and Sig. Proc.*, vol. 11, 1986, pp. 2015–2018.

[32] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, 2021.

[33] A. Kuznetsova, P. B. Brockhoff, and R. H. B. Christensen, "lmerTest package: Tests in linear mixed effects models," *J. Statistical Software*, vol. 82, no. 13, pp. 1–26, 2017.

[34] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting Linear Mixed-Effects Models Using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015.

[35] R. V. Lenth, *emmeans: Estimated Marginal Means, aka Least-Squares Means*, 2021, R package version 1.5.4.

[36] P. Virtanen and SciPy 1.0 Contributors, "SciPy 1.0: Fundamental algorithms for scientific computing in Python," *Nature Methods*, vol. 17, pp. 261–272, 2020.

[37] J. H. Ryalls and P. Lieberman, "Fundamental frequency and vowel perception," *J. Acoust. Soc. Am.*, vol. 72, no. 5, pp. 1631–1634, 1982.

[38] P. F. Assmann and T. M. Nearey, "Identification of frequency-shifted vowels," *J. Acoust. Soc. Am.*, vol. 124, no. 5, pp. 3203–3212, 2008.

[39] P. F. Assmann, S. Dembling, and T. M. Nearey, "Effects of frequency shifts on perceived naturalness and gender information in speech," *Proc. Interspeech*, pp. 889–892, 2006.

[40] M. B. Winn, D. Wendt, T. Koelewijn, and S. E. Kuchinsky, "Best Practices and Advice for Using Pupillometry to Measure Listening Effort: An Introduction for Those Who Want to Get Started," *Trends in Hearing*, vol. 22, 2018.