


Listeners' Spectral Reallocation Preferences for Speech in Noise

Olympia Simantiraki ^{1,*}  and Martin Cooke ²¹ Language and Speech Laboratory, University of the Basque Country, 01006 Vitoria-Gasteiz, Spain² Ikerbasque (Basque Science Foundation), 48009 Bilbao, Spain; m.cooke@ikerbasque.org

* Correspondence: olympia.simantiraki@ehu.eus

Abstract: Modifying the spectrum of recorded or synthetic speech is an effective strategy for boosting intelligibility in noise without increasing the speech level. However, the wider impact of changes to the spectral energy distribution of speech is poorly understood. The present study explored the influence of spectral modifications using an experimental paradigm in which listeners were able to adjust speech parameters directly with real-time audio feedback, allowing the joint elicitation of preferences and word recognition scores. In two experiments involving full-bandwidth and bandwidth-limited speech, respectively, listeners adjusted one of eight features that altered the speech spectrum, and then immediately carried out a sentence-in-noise recognition task at the chosen setting. Listeners' preferred adjustments in most conditions involved the transfer of speech energy from the sub-1 kHz region to the 1–4 kHz range. Preferences were not random, even when intelligibility was at the ceiling or constant across a range of adjustment values, suggesting that listener choices encompass more than a desire to maintain comprehensibility.

Keywords: listener preferences; spectral energy reallocation; glimpses profile



Citation: Simantiraki, O.; Cooke, M. Listeners' Spectral Reallocation Preferences for Speech in Noise. *Appl. Sci.* **2023**, *13*, 8734. <https://doi.org/10.3390/app13158734>

Academic Editor: Douglas O'Shaughnessy

Received: 24 June 2023

Revised: 24 July 2023

Accepted: 26 July 2023

Published: 28 July 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Speech output is in widespread use to deliver information in listening environments such as transport hubs, in vehicles, and in the home, where other sound sources are likely to be co-active. While live speech is still deployed in such contexts, much of the time listeners hear non-live forms of speech, either pre-recorded natural speech that has been produced offline, or speech that has been dynamically generated by synthesis from text. The use of non-live speech provides an opportunity to manipulate the audio signal prior to presentation, usually with the goal of maintaining or enhancing intelligibility in noise [1–7]. Such algorithms have been shown in large-scale evaluations (e.g., [8,9]) to be highly effective in boosting intelligibility when applied to natural (e.g., [10]) or synthetic speech (e.g., [11–13]).

Human speech perception is a complex process that involves bringing linguistic knowledge to bear on the available acoustic evidence. A listener's prior experience with the fundamental structural characteristics of the language, including its phonological rules and phonotactic constraints, significantly shapes speech processing, particularly in adverse conditions. Non-native listeners encounter greater challenges in noisy environments compared to native listeners (see [14] for a review), indicating the importance of linguistic familiarity in understanding speech.

The spectral and temporal characteristics of speech also play a crucial role in how listeners perceive and comprehend spoken language. Even in the absence of deliberate speech manipulation, various factors are known to contribute to speech intelligibility, including spectral change [15], syllable-scale amplitude modulations [16,17], and phonologically relevant sampling instants [18]. Spectral properties in particular impact various aspects of speech perception. For instance, the spectral content of speech is closely linked to many important phonemic contrasts e.g., [19–22], allowing listeners to distinguish between different speech sounds and forming the basis of the accurate recognition of words. Additionally,

the temporal evolution of the spectral envelope contributes to the perception of prosody, rhythm and intonation.

In noisy situations, acoustic evidence is partially masked or distorted. It has been shown that speech can be understood after manipulations that reduce its spectral resolution [23] or bandwidth [24], remove its mid-frequency energy [25], or restrict it to sparse time–frequency regions [26,27]. These findings suggest that the processes underlying human speech perception incorporate substantial latitude in the types of acoustic signals that can be considered speech-like, and motivates the search for manipulations that optimise a listener’s experience.

Several models have been proposed to explain how noise affects intelligibility [28–32]. One common feature of many models is the notion of signal-to-noise ratio (SNR) in each frequency region (e.g., [1]), with each frequency receiving an importance weighting [30]. These models highlighted the potential for exploring speech modifications that alter the spectral energy balance of speech in noisy conditions.

Energy reallocation is one form of manipulation that alters the overall spectral energy balance. The term ‘reallocation’ is used because the effectiveness of speech modification approaches is typically assessed after normalizing the root-mean-square (RMS) energy following modification [8,9], which has the consequence that attenuation or boosting in one part of the spectrum leads to changes in other spectral regions. Spectral reallocation techniques are informed by perceptual studies that indicate the relative importance of different spectral regions for speech recognition (e.g., [28,33–35]), and by measurements of acoustic properties in naturally enhanced speech, such as the speech styles that result from an instruction to speak clearly [36,37] or from noise immersion (Lombard speech; [38–40]). Reallocating energy to the mid-frequencies (1–3 kHz) using spectral tilt adjustment [41,42], high-pass filtering [43,44] or formant amplitude equalisation [45] has been shown to lead to substantial intelligibility increases in noise. For example, Tang and Cooke [44] demonstrated that high-pass filtering led to gains of 45 and 55 percentage points for keyword identification in sentences in high and low SNRs, respectively. While other forms of speech modification (e.g., durational or modulation adjustments) are also worth investigating, the potential gains from the above-cited studies motivated our focus on spectral manipulation in the present study.

Spectral energy reallocation is both simple to implement and highly effective in boosting intelligibility. However, while the delivery of a fully comprehensible message is of paramount importance, other factors, such as naturalness, quality and listening effort, are known to influence a listener’s overall experience (e.g., [46–48]). Relatively little is known about the impact of speech modification on these factors. Nevertheless, a few studies have examined the effect of certain types of speech modification on speech quality and listening effort. These investigations have demonstrated that modifications which successfully promote intelligibility can also have the side effect of reducing speech quality [49], especially when presented in quiet or at comfortable SNRs [50]. Cognitive effort has been shown to increase when listening to synthetic voices, relative to natural speech [51–53]. On the other hand, decreases in listening effort have been observed for two forms of algorithmically modified speech. Using effort ratings scales, Rennie et al. [54] demonstrated that the AdaptDRC algorithm [7] resulted in lower effort than unprocessed speech, while both subjective ratings and pupillometry measures in a study by Simantiraki et al. [53] pointed to a reduced listening effort for the SSDRC algorithm [5]. Collectively, these studies suggest that intelligibility-enhancing speech modifications can have additional impacts—both positive and negative—on a listener’s overall experience when processing speech.

Intelligibility is easy to measure but assessing a listener’s overall speech communication experience is more challenging. Techniques such as ratings scales and questionnaires (e.g., [54,55]), pupillometry (e.g., [56]) and dual-task performance (e.g., [57,58]) have been employed to measure aspects beyond intelligibility (for a review, see [59]).

An alternative paradigm is used in the present study to investigate the wider impact of spectral energy reallocation. The approach is based on providing listeners with the ability to modify some facet of the speech signal, with real-time auditory feedback, allowing listeners to express their preferences directly. This listener-centric technique has been used in the past to explore listeners' preferred choice of formant frequency/fundamental frequency relationship [60], speech rate [61–63], speech level [64] and local SNR for speech enhancement [65]. Listener preferences in the context of spectral modifications have previously been explored for individuals with hearing loss by providing them with direct control over broadband, low-, and high-frequency gain [66] or loudness and degree of spectral tilt [67].

A listener-centric approach has a number of potential benefits compared to the more traditional technique, in which behavioural measurements are obtained for a small number of experimentally defined processing conditions. First, it can be used to explore the parameter space at a finer level of quantisation than is economically efficient in traditional designs. Second, compared to the use of rating scales, the approach is likely to be more natural for a listener, being free of the semantic labels that map multi-faceted stimuli on to a finite number of categories. Further, the approach provides an additional objective metric—adjustment time—that may reflect task difficulty. However, the primary motivation for employing a listener-centric approach in the current study is that it has the potential to allow preferences to emerge in parallel with the optimisation of intelligibility, making it possible to elucidate the relationship between the two.

The objective of the present study is to better understand listeners' spectral reallocation preferences. We first assess the extent to which listeners are capable of using spectral energy modifications to improve speech intelligibility in the presence of masking noise. This issue is examined by comparing the distribution of preferences with keyword recognition scores, to find the preference values that optimise intelligibility.

A second issue is whether listeners solely optimise intelligibility, or if they exhibit additional preferences that are independent of message comprehensibility. To address this question, the space of possible adjustments available to listeners was designed in such a way that intelligibility was likely to be constant across at least part of the adjustment range; in addition, a quiet condition was included to promote ceiling performance. This design enables the observation of changes in preferences, even when intelligibility is constant.

A further question is whether preferences differ in the presence or absence of masking noise. We hypothesise that listeners' preferred adjustment values are more likely to be aimed at the maintenance of intelligibility in the presence of high levels of noise. Conversely, in quiet or low noise conditions, preferences may reflect other factors.

To address our research questions, we conducted two experiments aimed at exploring spectral preferences. We chose to examine spectral preferences due to their established effectiveness in enhancing intelligibility. In the first experiment (Expt. 1; Section 3), we examined modifications across the entire speech bandwidth, while in the second experiment (Expt. 2; Section 4), we investigated bandwidth-limited modifications. In Expt. 1, participants had the choice to directly adjust the spectral tilt of the speech or indirectly modify it by adjusting the magnitude of one of three spectral bands. In the latter experiment, participants were presented with a single band of speech and given control over its spectral extent through three forms of bandwidth modification, or by changing the spectral location of the band. In each experiment, listeners had the ability to adjust the speech in real time for one of four features (eight features in all across the two experiments), thereby altering the speech spectrum. Immediately following the adjustment, participants carried out a speech-in-noise recognition task at the chosen setting.

2. Outline of the Experiments

In two experiments, listeners manipulated one of eight different speech properties, illustrated schematically in Figure 1. The first experiment involved full-bandwidth speech and consisted of the four types of adjustment shown in the left column of the figure, viz.

increasing or decreasing the level of speech in one of three octave bands, or altering spectral tilt (a preliminary report on spectral tilt adjustments was provided in Simantiraki et al. [68]). In contrast, all conditions of Expt. 2 involved narrowband speech. In three conditions, listeners made changes to speech bandwidth by lowpass or highpass filtering, or by direct manipulation of the bandwidth; in a fourth condition, they selected the centre frequency of an octave band of speech.

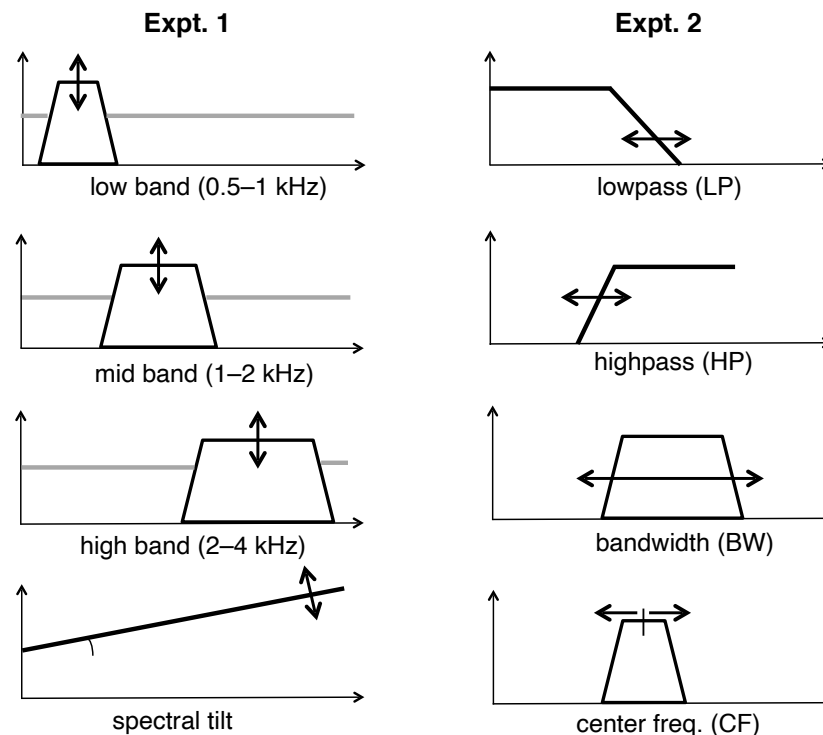


Figure 1. Schematic of the eight listener-controlled modifications tested in the present study. Arrows indicate the direction of modification, while axes represent frequency (x) and magnitude (y). Implementation details for Expt. 1 and Expt. 2 are provided in Sections 3.2 and 4.2, respectively.

Spectral energy modifications in both experiments were pre-computed at a sufficiently fine range of parameter values to produce the perceptual effect of continuous variation (including smooth mid-utterance transitions) as listeners move a virtual knob or click up/down buttons. Stimulus presentation and response collection were implemented using a software tool (SpeechAdjuster; [69]). In the current study, on each trial, listeners initially adjusted a continuous sequence of sentences, with no time limit, using up/down buttons, then underwent a brief test phase, in which they were asked to identify keywords in sentences presented at the parameter value chosen during the adjustment phase.

One issue that arises in any attempt to measure listener preference is the possibility of bias due to the form of the instructions given to participants. The current study did not seek to distinguish between factors that could conceivably contribute to a listener's preferred setting. To avoid any bias towards one or other of these potential factors, listeners were asked to adjust the signal to make it understandable, and to consider the task as being analogous to adjusting the volume when watching a film: too high a volume causes discomfort, while too low a volume makes understanding the words difficult.

3. Expt. 1: Spectral Reallocation Preferences for Full-Bandwidth Speech

3.1. Listeners

Thirty-five native Spanish listeners (30 females) aged between 18 and 34 years (mean 20.1; SD 2.6) took part in Expt. 1. Participants were recruited through an announcement shared with the students of the English and German philology and translation and inter-

pretation department of the University of the Basque Country. The recruitment criteria included being native Spanish listeners or bilingual Spanish and Basque listeners, and having no known hearing impairment. All listeners passed an audiological screening with a hearing level better than 25 dB at frequencies in octave steps in the range 125 Hz to 8 kHz in both ears. Listeners were paid for their participation.

3.2. Stimuli

Speech stimuli were derived from the Sharvard Corpus [70], which consists of Spanish sentences spoken at a normal speaking rate. The linguistic complexity of this corpus is similar to that of the original English Harvard Corpus [71]. Each sentence contains five keywords used for scoring. Target sentences were drawn from the male Sharvard talker. Sentence numbers 1–380 and 381–540 were used in the adjustment and test phases, respectively (these phases are defined in Section 3.3 below).

Stimuli were presented in quiet and in three additive stationary speech-shaped noise (SSN) conditions at SNRs of -6 , -3 and 0 dB. The masker was generated by filtering random uniform noise with the long-term spectrum of 700 concatenated sentences of the female talker of the Sharvard corpus.

Four kinds of variation (for brevity, referred to as ‘features’ below) were tested in Expt. 1, namely, changes in spectral tilt and changes in the energy of octave bands in the low-, mid- and high-frequency parts of the spectrum. For each feature, the entire sentence set (both adjustment and test phase sentences) was processed at each level of adjustment available to the listener, resulting in multiple examples of each sentence that differed only in the degree of modification applied to them. Collectively, the sentence sets formed the precomputed input to the SpeechAdjuster tool. The number of modification steps was chosen in pilot testing as the minimum number required to give the impression of continuous change to the listener. Different numbers of steps (indicated below) were used for different feature types. The amplitude of each sentence was normalised using a fixed root-mean-square criterion following modification to ensure that each sentence had the same presentation level as every other sentence, and to remove any contribution to intelligibility from changes in the overall energy.

3.2.1. Spectral Tilt Adjustment

Spectral tilt adjustment made use of pre-emphasis or de-emphasis to increase or decrease tilt, respectively, implemented in the time domain using digital filtering with the transfer function $1 - \lambda z^{-1}$ for pre-emphasis, and its reciprocal for de-emphasis. The parameter λ controls the degree of emphasis/de-emphasis and was drawn linearly from the interval $[0.2, 1]$ to produce tilts in the range $[-10.85, 0.59]$ dB/octave. Some 23 steps were used, the central one having the original (unmodified) tilt value of -5.25 dB/octave, along with 11 having a progressively flatter tilt than the original, and 11 with a progressively steeper tilt. Spectral tilt adjustments were implemented using the `filter` function in Matlab 2016b.

3.2.2. Spectral Band Filtering

The three spectral band conditions had cut-off frequencies forming octave intervals corresponding to the frequency ranges 0.5–1 kHz, 1–2 kHz and 2–4 kHz. We will refer to these as the low, mid and high bands. In total, 21 energy adjustment steps were available to listeners. These were determined according to the function $y = a - b(c^{x-1})$ for $a = 15$, $b = 100$, $c = 0.8$ and for step x taking on all integer values in the range 1–21, with the level difference y expressed in decibels. Constants a – c were chosen to provide a range that spanned an attenuation of 85 dB to a boost of nearly 15 dB. The lower value provided the complete removal of speech energy in the band, while the upper value was chosen to avoid clipping in the output. Band energy changes were implemented by multiplying 1024-point Fourier amplitude spectra computed in successive 30 ms Hann windows with 50% overlap, followed by inverse Fourier transformation, with the phase component unmodified.

3.3. Procedure

The experiment was divided into four blocks according to SNR (quiet plus three masked conditions). Within each block, listeners were presented with five trials for each of the four features being modified, for a total of 20 trials. The presentation order of the 20 trials was randomised, meaning that listeners would typically adjust different features on consecutive trials (i.e., there was no blocking by feature type).

Each trial consisted of an adjustment phase followed by a test phase. In the adjustment phase, sentences were presented in a random order, with a 0.5 s gap between sentences. In order to prevent listeners from simply selecting the same adjustment value as on the previous trial, the modification value applied to the initial sentence in the adjustment phase was randomly selected from the available steps. Participants could listen to as much speech as desired during the adjustment phase but they were forced to listen to at least 5 s of speech before proceeding to the test phase. Listeners made their adjustments using on-screen buttons labelled with up and down arrows. They were instructed to “use the buttons to alter the speech in order to recognise as many words as possible”. In the test phase of each trial, intelligibility was evaluated by a speech perception task using the value chosen at the end of the adjustment phase. Participants listened to two sentences and typed what they heard into an on-screen text box immediately after presentation of each of the two sentences. Prior to the experiment, participants underwent a task-familiarisation phase consisting of five trials (two in quiet and three in noise), using sentences from the adjustment set i.e., different from those used in any test phase.

Block ordering was counterbalanced across listeners using a balanced Latin square design. Stimuli were presented through Sennheiser (Wedemark, Germany) HD380 headphones at a fixed presentation level, different for each condition (approximately 84, 79, 77 and 75 dB SPL for the -6 , -3 , 0 dB SNR and quiet conditions, respectively) as measured by a Brüel and Kjaer (Nærum, Denmark) type 4153 artificial ear and a Brüel and Kjaer type 2260 sound level meter. Listeners were seated in a sound-attenuating booth in a purpose-built speech perception laboratory at the University of the Basque Country (Vitoria Campus).

3.4. Postprocessing

Participant responses in the test phase were scored by counting the number of keywords identified correctly. Prior to scoring, user responses were normalised by the automatic application of the following processes: conversion to lower case, removal of vowel stress marks, removal of non-alphanumeric characters, substitution of homophones, and conversion of numbers represented as digits to orthographic form. In addition, limited word-level error correction was applied based on a list of 84 common issues (typos, missing tildes, and spelling mistakes) identified manually in a previous study [72] that used the Sharvard Corpus. Finally, participants were encouraged to spend as much time as necessary during the adjustment phase; thus, trials with long response times were also considered valid. Therefore, all collected data were included in the statistical analysis to ensure a representative understanding of participants' preferences.

3.5. Results

3.5.1. Preferences

For each of the four features, Figure 2 depicts listeners' preferred adjustments relative to the unmodified speech baseline (left column), mean keywords correct scores (middle), and the time spent in the adjustment phase (right). For all features, listeners made the smallest adjustment relative to the original in quiet, and adjustments became progressively larger as the noise level increased. While listeners boosted energy in the mid and high bands, the opposite was the case for the low band, where listeners preferred to attenuate its energetic contribution. Since the attenuation of the low band results in the boosting of the other spectral regions, the strategies employed by listeners point to a consistent goal of increasing the SNR in the mid-high frequency region. Listeners also preferred flatter

spectral tilts with increasing noise, again suggesting a preference for reallocation of energy to the mid-high frequencies.

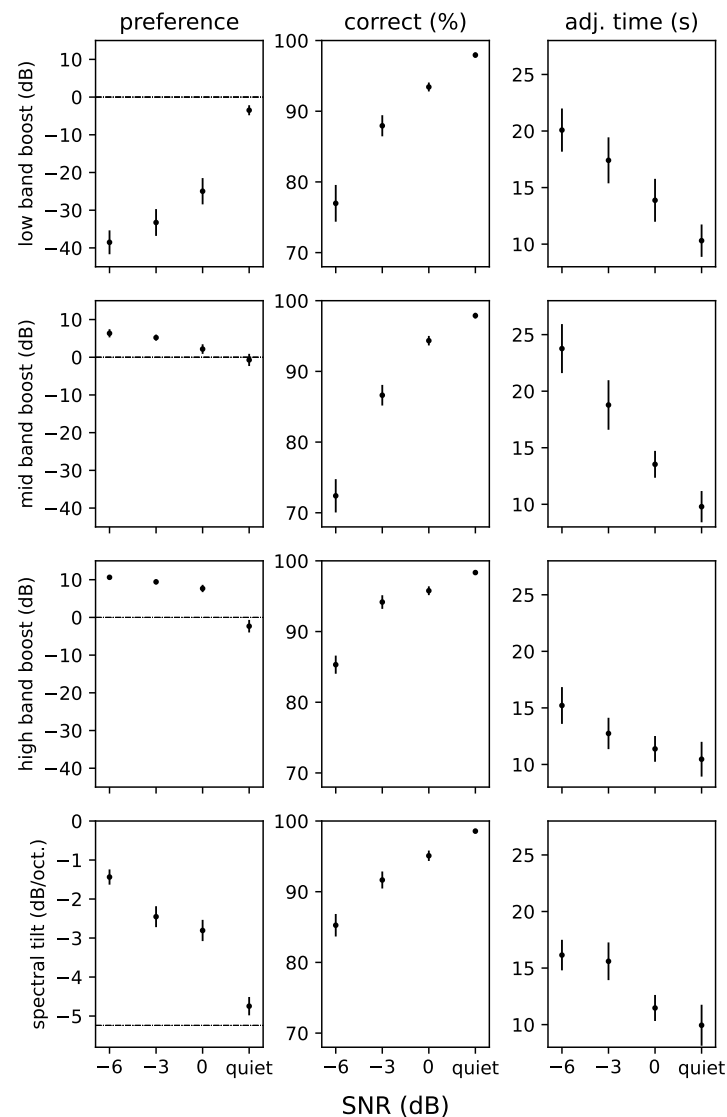


Figure 2. (Left): Listeners' preferred energy adjustments relative to the unmodified speech baseline (dotted line); (middle): mean keywords correct; (right): time spent in the adjustment phase. Error bars here and elsewhere denote ± 1 standard errors.

Since adjustment choices were not always normally distributed (see Figure 3), rank-based Kruskal–Wallis H tests were conducted to examine the effect of SNR on preferences for each feature independently. These tests showed a statistically significant effect of SNR for all features (min $H = 56.3$, all $p < 0.001$). Post hoc pairwise comparisons using Dunn's test, with Holm correction [73] for multiple comparisons, indicated that preferences for all features differed across pairs of SNR conditions (all $p < 0.05$), with the following exceptions: -3 and -6 dB for the low band ($p = 0.07$), and between the pair 0 and -3 dB for both the mid band ($p = 0.08$) and spectral tilt ($p = 0.12$). Statistical comparisons were performed using the `kruskal` and `posthoc_dunn` functions of the `stats.scipy` and `scikit_posthocs` libraries in Python [74].

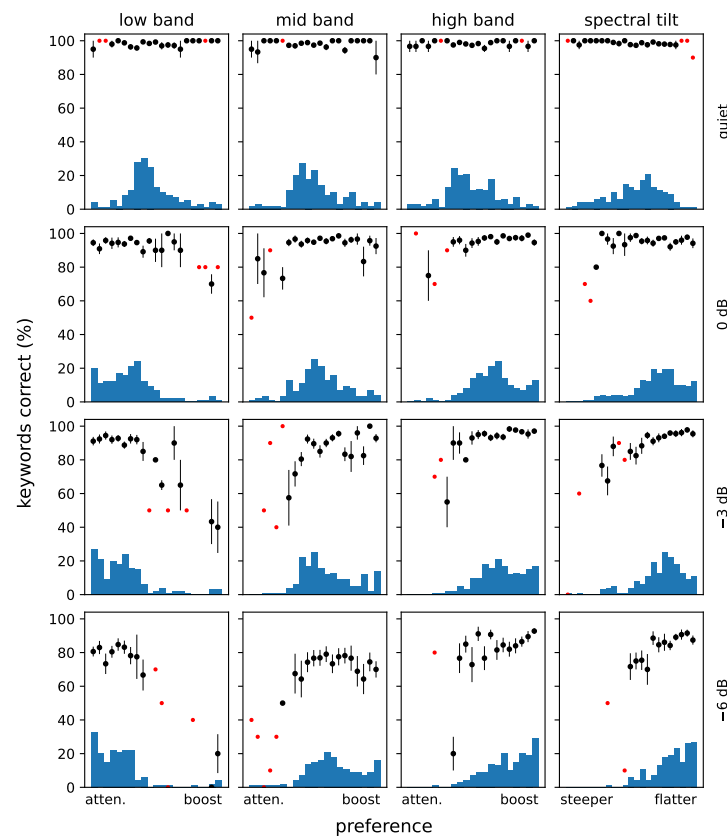


Figure 3. [Colour online] Distribution of preferences (blue bars) and percentage of keywords identified correctly at each preference (points with error bars; points in red correspond to preferences chosen only once). The units on the y-axis represent both percentage correct values (for intelligibility) and preference counts (i.e., height of blue bars). In the latter case, distributions sum to 175 (35 listeners \times 5 trials).

3.5.2. Intelligibility

As expected, listeners achieved the highest intelligibility in quiet. In noise, listeners were able to maintain intelligibility at a reasonably high level, even at the most adverse SNR, when presented with the possibility of modifying energy in the high band or by adjusting spectral tilt; modifications to the low and mid bands led to lower scores in the most adverse conditions. The intelligibility scores for all types of modification tested in this experiment are substantially in excess of those reported in [70] for unmodified speech from the same corpus and talker, and using the same type of masking noise, where psychometric functions indicate that listeners would achieve scores of 83%, 72% and 54% at 0, -3 , and -6 dB SNR, respectively.

A generalised linear mixed-effects model using the `glmer` function from the `lme4` library [75] in R [76] was constructed to predict the proportion of keywords identified correctly in each trial. This model had fixed effects of SNR and feature, by-subject random intercepts and per-feature slopes. The model indicated significant effects of SNR ($\chi^2(3) = 1232, p < 0.001$) and feature ($\chi^2(3) = 78.1, p < 0.001$) together with a significant interaction of the two ($\chi^2(9) = 25.6, p < 0.01$). Post hoc pairwise comparisons using function `emmeans` [77] with Tukey corrections showed that all SNR pairs differed for all features (all $p < 0.001$) except for the $0/-3$ dB SNR pair for the high band ($p = 0.13$). Intelligibility was not statistically different for all features in the quiet (min $p = 0.50$) and SNR = 0 dB (min $p = 0.08$) conditions, indicating that in the least challenging conditions, listeners may have been able to optimise intelligibility to an equivalent level for all types of adjustment. Adjustments to spectral tilt and to the high band were not statistically different at all SNRs, apart from -3 dB, where they tended to differ ($p = 0.054$).

3.5.3. Adjustment Time

Median rather than mean per-participant adjustment times were analysed, i.e., the data presented in the third column of Figure 2 represents the cohort mean of per-participant median adjustment times.

At the level of the listener cohort (i.e., comparing columns 2 and 3 of Figure 2), the time spent during the adjustment phase was strongly inversely correlated with the final score achieved in the test phase (Pearson $r = -0.96$, $N = 16$, $p < 0.001$). Listeners spent nearly 2.5 times longer when adjusting the mid band in the most challenging condition compared to the noise-free condition. Expressed as an average across all conditions, listeners varied substantially in the amount of time taken to reach a preference, ranging from 6 s to 32 s. However, individual listener scores were not significantly correlated with the median adjustment time ($r = 0.30$, $N = 35$, $p = 0.08$), indicating that those listeners who spent more (or less) time exploring the range of possible adjustments did not necessarily achieve higher (or lower) intelligibility rates. A linear mixed-effects model using function `lmer` from the `lme4` library [75], with fixed effects of SNR and feature, and by-subject random intercepts, indicated the main effects on the adjustment time of SNR ($\chi^2(3) = 102$, $p < 0.001$) and feature ($\chi^2(3) = 23.9$, $p < 0.001$), with a modest interaction between the two ($\chi^2(9) = 17.8$, $p = 0.04$). Post hoc tests with Tukey corrections indicated that listeners took longer to adjust in the -6 dB condition than in quiet for all features (all $p < 0.01$ except for the high band, where $p = 0.053$). As was the case for intelligibility, adjustment times did not differ statistically in the quiet and 0 dB conditions for any feature (min $p = 0.18$).

3.5.4. Relationship between Preference and Intelligibility

Figure 3 shows the distribution of listener preferences alongside the percentage of keywords identified correctly as a result of that choice. Listeners exhibited clear preferences when intelligibility was at (or close to) ceiling in quiet and 0 dB SNR conditions in particular. Two-sample Kolmogorov–Smirnov tests using function `ks_2samp` in `scipy.stats` of Python [74] confirmed that listeners' preference distributions were non-uniform for all combinations of feature and SNR (all $p < 0.01$).

In some conditions, listeners did not adopt extreme preferences, even though intelligibility was not compromised at those extreme adjustment values. This is especially evident in the quiet condition but also prevails over most of the adjustment range for the 0 dB condition. However, as conditions became more adverse, preferences and intelligibility outcomes were more closely related, and listeners appeared more willing to choose the extremes of the adjustment range. This is clearly seen for the low- and high-frequency bands in the -6 dB SNR condition, where the modal value of listeners' choices occupies one or the other extreme of the range.

3.5.5. Energetic Masking

The consequences of choosing a specific adjustment level on speech audibility in noise can be assessed by estimating the amount of energetic masking at each frequency resulting from that specific adjustment choice. Figure 4 plots the proportion of glimpses of the target speech at each frequency as a function of the adjustment level. Glimpses were computed using the extended glimpse proportion metric defined in Tang and Cooke [32]. It is evident from this visualisation that a listener's reallocation strategy in the face of a speech-shaped noise masker is to make adjustments that ensure the availability of speech glimpses in the frequency region above 1 kHz. This is clearly seen for the low and high bands and for spectral tilt, where attenuation and boosting, respectively, lead to a high glimpse density in the 2–4 kHz region in particular. The middle 1–2 kHz band represents an interesting compromise; here, many listeners chose to keep the energy level close to that of the original speech (as evidenced by Figure 3), presumably because boosting this band leads to a loss of glimpses at higher frequencies, while attenuation conversely produces a paucity of glimpses in the central frequency region itself. The average preference made by listeners here enables the maintenance of glimpses in the 1–2 kHz range.

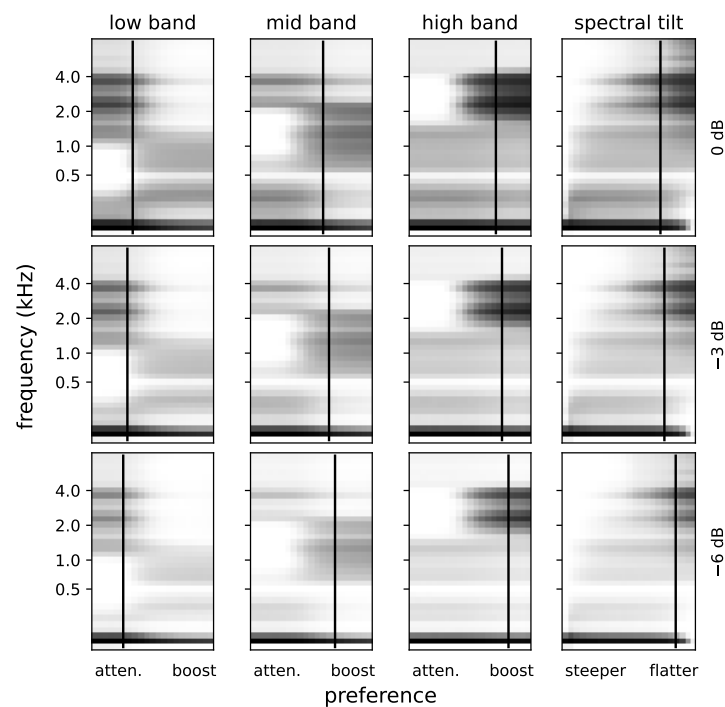


Figure 4. Proportion of glimpses (spectro-temporal regions where the speech is more intense than the masker) in each spectral region as a function of listeners' level adjustment preferences. In each panel, the x-axis represents the adjustment chosen by listeners, while the y-axis represents spectral frequency (34 channels on an ERB-rate scale). Darker areas indicate a higher proportion of glimpses. Vertical bars indicates mean listener preferences.

3.6. Interim Discussion

When provided with the opportunity to adjust spectral tilt or the level of an octave band of speech in the presence of masking noise, listeners made adjustments that tended to maximise intelligibility (Figure 3), enabling the effect of the masker to be counteracted to some extent, and moreover achieved intelligibility scores well in excess of those for unmodified speech [70]. The utility of spectral energy reallocation as a strategy for improving the intelligibility of speech in noise is illustrated by the range of keyword scores across conditions at each SNR (Figure 2): under a constant energy constraint, reallocating energy to the 2–4 kHz region or flattening spectral tilt resulted in approximately half as many keyword errors as modifying speech level in the 1–2 kHz band. A transfer of energy to mid-frequencies is in line with natural modifications made by talkers when speaking in noise (e.g., [78]), or when asked to speak clearly, e.g., [36]. The finding that speech modification can lead to substantial gains in intelligibility echoes earlier studies [1,3,10,13], and confirms the outcome of studies that indicated the value of spectral modification in particular [41,44]. The present study extends previous investigations by providing a listener-centric approach, which has the potential to permit the identification of optimal values for any parameters tested, rather than relying on an a priori set of conditions chosen by the experimenter, as is the case in earlier studies e.g., [42].

Glimpse distributions across frequency were relatively stable over a range of neighbouring adjustment values (Figure 4). This can be explained by the binary nature of glimpses: once the target speech has sufficient energy in a given region, the glimpse distribution does not change by further boosting the target speech energy in that region. This fact provides listeners with some latitude in the choice of adjustment, and it may be that other factors come into play once a glimpse pattern that supports a high level of intelligibility has been established. Indeed, listeners made adjustments at a scale commensurate with the adversity of the listening condition, evidenced by the choice of modifications whose deviation from the unaltered speech increased monotonically as SNR decreased

(Figure 2). Listeners also spent more time selecting the appropriate adjustment in challenging listening conditions. These findings indicate a tension between maintaining a spectral balance close to that of the original speech, and making changes to overcome the masker. The distribution of adjustment values chosen by listeners also hints at bimodality in several conditions (Figure 3). This tendency is best appreciated in the high band and for spectral tilt, where, in addition to a relatively central mode, a secondary and more extreme mode exists. A similar tendency in the opposite direction is evident for the low band.

Reallocation of speech energy to the remainder of the spectrum provides an opportunity for listeners to optimise intelligibility while simultaneously expressing a preference influenced by other factors. In Expt. 1, the entire spectral band of speech was intact, and as such, provides only a partial view into listeners' preferences. To explore choices in a different—and more challenging—scenario, in Expt. 2, listeners were presented with a single band of speech and were able to control its spectral extent via three forms of bandwidth modification, or by changing the spectral location of the band.

4. Expt. 2: Spectral Reallocation Preferences for Band-Limited Speech

In experiment 2, listeners were able to modify the bandwidth of speech by adjusting the low-pass (LP) or high-pass (HP) cut-off frequency of the band, or by modifying the bandwidth (BW) directly for a band with a constant centre frequency. They were also able to modify the centre frequency (CF) of an octave band of speech. Except where specified below, all procedural and post-processing details for Expt. 2 are identical to those for Expt. 1.

4.1. Listeners

Thirty-seven native Spanish listeners (32 females) aged between 18 and 34 (mean 20.1 years; SD 2.6 years) participated in Expt. 2. Most ($N = 32$) of these listeners had also participated in Expt. 1.

4.2. Stimuli

As in Expt. 1, sentences came from the Sharvard Corpus [70]. For the adjustment phase only, the same sentence subset (1–380) as in Expt. 1 was employed. This decision was driven by the limited availability of utterances in the corpus, given that listeners required a potentially unlimited amount of speech material during this phase. While during adjustment listeners may have recognised elements from sentences used in the earlier experiment, which in turn may have led to changes in preferences, we believe that any such changes due to repeated exposure would be modest, and that our main research question involves comparisons of conditions within each experiment rather than across experiments. Note that a different set of utterances (sentences 541–700) from those used in Expt. 1 was used during the test phase.

Features being modified (LP, HP, BW, CF) all involved spectral filtering implemented with frequency-domain FIR filters, designed using a Chebyshev window with 100 dB of attenuation from the pass to stop bands, via the `fir1` and `chebwin` functions in Matlab 2016b. For each of the four conditions, the parameter adjustment range of interest was divided into 25 steps, a number chosen to ensure that the transitions between steps were smooth and to avoid redundant steps. After filtering, every sentence in every condition was normalised to have the same RMS energy.

For the LP and HP conditions, the frequency cut-off varied from 0.45 to 5.4 kHz in equal steps on a logarithmic scale. For the BW condition, all bands were centred on 1470 Hz, and the bandwidth varied from 1.14 to 4.58 octaves in equal steps of 0.143 octaves, corresponding to a bandwidth of 1.2 kHz at the narrowest end of the continuum and 6.9 kHz at the widest. The CF condition consisted of octave bands of speech with centre frequencies equally spaced on a log scale from 0.42 kHz to 5.09 kHz. Parameter ranges for the LP, HP and BW conditions were designed to permit choices that would lead to intelligibility scores close to ceiling in the quiet condition. No such choice was possible in

the CF condition, but the bandwidth was restricted to a single octave for compatibility with earlier studies that examined the intelligibility of octave bands of speech (e.g., [79]).

4.3. Results

4.3.1. Preferences

Figure 5 shows mean listener preferences at the end of the adjustment phase, intelligibility scores in the test phase, and the time spent in the adjustment phase, for the quiet and masked presentation conditions. For each of the four features, listeners' preferences indicate a clear effect of increasing masker level. For the LP feature, listeners chose a cut-off frequency of around 3 kHz in quiet but raised the cut-off frequency to increasingly higher frequencies as the SNR decreased. Listeners also raised the cut-off frequency in the HP condition, from a value of around 700 Hz in quiet to around 1 kHz in noise. As a consequence, listeners chose to transfer energy from the sub-1 kHz to the region above 1 kHz. A similar preference for high frequencies is seen in the CF condition, in which listeners shifted the octave band of speech from a centre frequency of around 1.5 kHz in quiet to a value close to 2.5 kHz in noise. In the condition where listeners were able to choose the bandwidth of speech, a mild increase from 3.25 octaves in quiet to 3.45 octaves in the -6 dB SNR condition was observed. However, this increase is not statistically significant (see below).

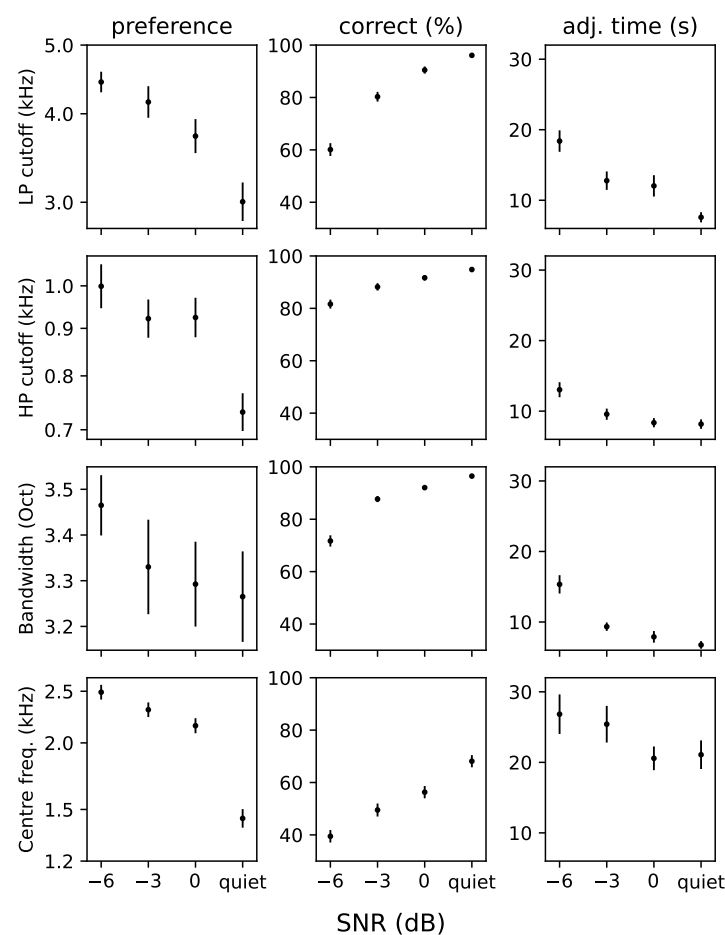


Figure 5. (Left): Listeners' preferred adjustments; (middle): keyword identification rates; (right): adjustment times for the four modifications of Expt. 2. Details as for Figure 2.

Unlike in Expt. 1, none of the adjustment options open to listeners in Expt. 2 corresponded to the original speech, since in all cases the bandwidth of speech was restricted. Listeners might have been expected to choose the adjustment value that led to the widest speech band in the quiet condition but this was not the case (see Section 4.4 below). For the

LP condition, the mean preferred cut-off of 3 kHz is substantially lower than the 5.4 kHz cut-off that would have led to the widest band of speech; likewise, for high-pass filtering, the chosen value of just above 700 Hz is well above the 450 Hz value that produces the widest band. A similar outcome is present in the BW condition, where listeners were able to choose a 4.58 octave band but instead preferred a bandwidth closer to 3.25 octaves on average. The notion of the widest band is less straightforward in the CF case since all bands were 1 octave wide; in terms of linear Hz bandwidth, the highest cut-off frequency (5.09 kHz) produced the widest band but presumably listeners were also sensitive to optimising the location in the frequency of this band, resulting in a choice of around 1.5 kHz in the quiet condition.

Kruskal–Wallis H tests were conducted to compare the effect of SNR on preferences for each of the four features, indicating significant effects for LP, HP and CF conditions (min $H = 43.3, p < 0.001$) but not for BW ($H = 6.2, p = 0.10$). Post hoc tests suggest that all pairs of SNRs differed at the $p = 0.05$ level, except for the following: preferences in the -6 and -3 dB conditions were identical ($p = 0.23$) in the LP condition; for the HP condition, the masked conditions did not differ statistically (min $p = 0.25$) but all differed from the quiet condition.

4.3.2. Intelligibility

Apart from the CF feature, intelligibility was at or close to ceiling in the quiet condition and, as expected, reduced progressively with noise level. Scores were generally lower in noise than in Expt. 1 due to the restricted speech bandwidth available. To analyse the impact of bandwidth manipulations in Expt. 2, a general linear mixed-effects model with the same effects structure as its counterpart in Expt. 1 was constructed. The CF feature was omitted from this analysis since it is rather different from the other three features in having a fixed bandwidth on an octave scale, and intelligibility is clearly lower for the CF feature at all SNRs. The mixed-effects model indicated significant effects of SNR ($\chi^2(3) = 1303, p < 0.001$) and feature ($\chi^2(2) = 51.3, p < 0.001$) together with a significant interaction ($\chi^2(6) = 94.0, p < 0.001$). Post hoc pairwise comparisons showed that intelligibility was different at all SNRs for all features (max $p = 0.003$). Intelligibility was identical in the quiet (min $p = 0.11$) and 0 dB (min $p = 0.42$) conditions for all 3 features but differed in the -3 and -6 dB conditions (all $p < 0.001$) apart from in the -3 dB condition, where high-pass filtering and bandwidth modification led to the same intelligibility.

4.3.3. Adjustment Time

Listeners took longer to make their choices as the noise level increased, in all conditions, and for those conditions with lower overall scores, the time taken was significantly longer. Across all conditions and noise levels, a very clear inverse relationship between adjustment time and intelligibility is evident ($r = -0.98, N = 16, p < 0.001$). As in Expt. 1, there is no significant correlation between an individual listener's adjustment time and intelligibility ($r = 0.22, N = 37, p = 0.20$). A linear mixed-effects model analysis (again omitting the CF condition) indicated significant effects of SNR ($\chi^2(3) = 161, p < 0.001$) and feature ($\chi^2(2) = 33.5, p < 0.001$) together with a significant interaction ($\chi^2(6) = 16.2, p < 0.05$). As in Expt. 1, listeners spent longer on adjustments in the -6 dB condition than in quiet for the three features (all $p < 0.001$).

4.3.4. Relationship between Preference and Intelligibility

Figure 6 shows the distribution of listener preferences and intelligibilities at each preference value. Two-sample Kolmogorov–Smirnov tests confirmed that listeners' preference distributions were non-uniform (all $p < 0.01$). Unlike the mean preferences shown in Figure 5, modal choices tended to be at values most similar to the original speech, i.e., at adjustment values that produce the widest speech bandwidth. Indeed, for almost all of the LP, HP and BW adjustments (the exception being the HP/ -6 dB condition), the most

frequently chosen adjustment led to the widest possible speech band. These preferences are apparent even when other choices lead to a similar intelligibility as seen by the generally uniform range of scores obtained for a range of adjustment values for many of the conditions. Intelligibility scores in the case where listeners were able to adjust bandwidth appear to be largely independent of the chosen bandwidth. In contrast, the relationship between intelligibility and preference is quite close for most of the other conditions. This correspondence is clearest in the CF condition.

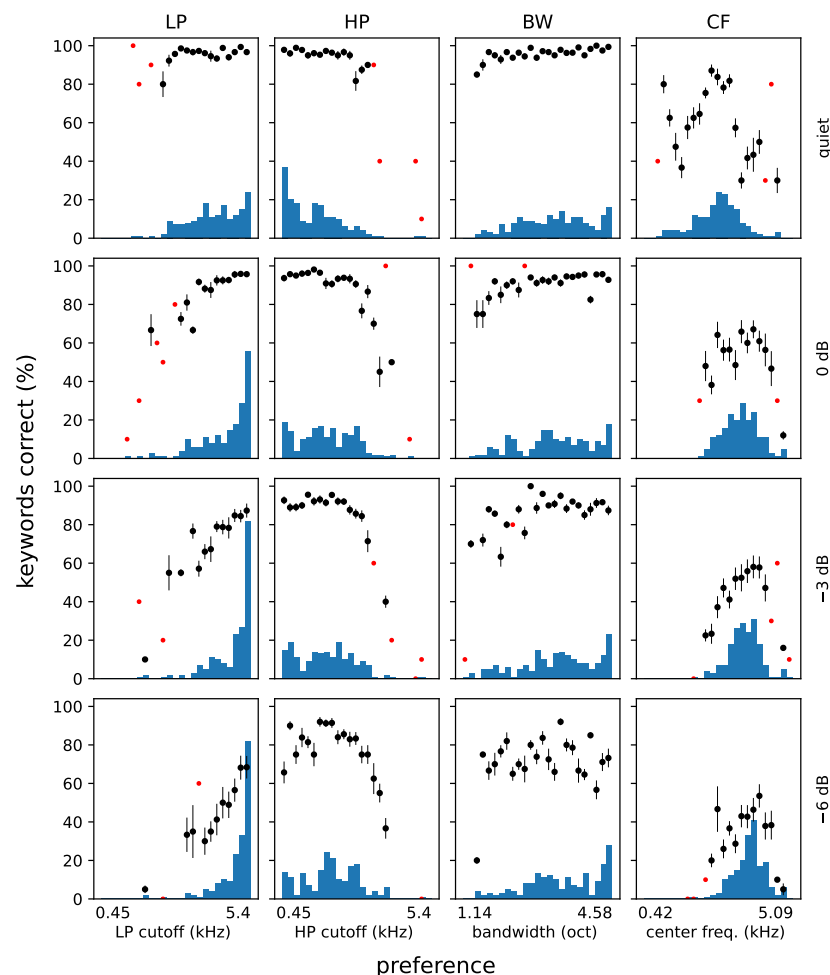


Figure 6. [Colour online] Distribution of preferences (blue bars) and percentage of keywords identified correctly at each preference (points with error bars; points in red correspond to preferences chosen only once). Details as for Figure 3, except distributions, sum to 185 (37 listeners \times 5 trials).

4.3.5. Energetic Masking

Figure 7 depicts spectral regions that survive energetic masking for the conditions of Expt. 2. Unlike the case in Expt. 1, many spectral regions contain no speech information for most of the range of adjustment available to listeners. As in Expt. 1, preferences suggest a desire to ensure that speech information is audible in regions above 1 kHz, and especially in the 2–4 kHz range.

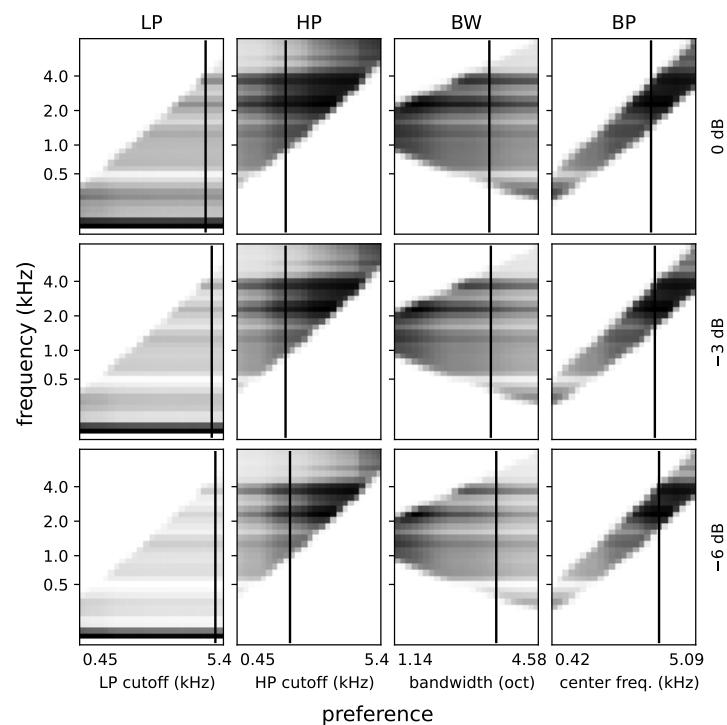


Figure 7. Proportion of glimpses in each spectral region as a function of listeners' adjustment preferences. Vertical bars indicates mean listener preferences. Details in Figure 4.

4.4. Interim Discussion

When given the ability to control the cut-off frequency of low-pass and high-pass filtered speech, listeners' most frequent choice was for values that led to the widest possible speech bandwidth (Figure 6), although this was by no means a universal preference (Figure 5). A similar finding was observed when listeners controlled the bandwidth of speech directly. For the HP and BW conditions, listeners were able to achieve similar intelligibility rates over a wide range of bandwidths. This can be explained on the basis of the existence of audible speech glimpses in the region above 1 kHz across a range of parameter values (Figure 7) for the HP and BW conditions; conversely, the LP and CF conditions indicate an absence of adjustment latitude if the goal is to maintain mid-high frequency speech. Our findings for the HP condition also align with the results reported by Tang and Cooke [44], who employed machine learning techniques and an intelligibility model to determine appropriate spectral weightings. Their study found that enhancing frequencies above 1 kHz was optimal for improving speech intelligibility in the SSN condition at -6 dB SNR, consistent with the findings in our study. Additionally, both studies observed a similar trend concerning the optimal frequency range for the different SNRs. Specifically, for lower SNR, there was a preference for a greater cut-off frequency.

The finding that listeners preferred to increase the high-pass cut-off frequency by around 50% in noise is unexpected but explicable on the basis of trading off energy in the sub- and supra-1 kHz regions. The relative unimportance for message decoding of the sub-1 kHz frequency region is consistent with the observation that talkers, when confronted by high-pass filtered noise, also increase energy in the mid-frequency region rather than attempting to exploit the (unmasked) lower frequency zone [80].

In the more adverse masking conditions, listeners appeared to trade off bandwidth increases against the availability of audible speech in important frequency regions. This search for a compromise is most evident in the HP and BW manipulation conditions. Listeners sacrificed frequency components below 1 kHz (LP/BW modifications) and above 4 kHz (HP/BW modifications), leading to the boosting of the mid-frequencies.

In noise-free conditions, listeners chose an octave band of speech centred on about 1.4 kHz. Our work extends Warren et al. [79] (Expt. 1), who measured intelligibility for low-predictability SPIN sentences [81] in six different 1-octave bands, finding the highest score of approximately 55% for a band centred on 2 kHz, and the next highest score of 40% for the 1 kHz band. Our findings suggest that intelligibility is rather sensitive to the choice of centre frequency: selecting bands centred on 1 or 2 kHz leads to a substantial drop in scores (top right panel of Figure 6).

5. General Discussion

5.1. Do Listeners' Preferences Optimise Intelligibility?

When provided with a simple instruction to adjust some characteristic of a speech signal in order to recognise as many words as possible, much of the time listeners responded by choosing a setting that maximised intelligibility. This finding suggests that a listener-centric approach can be used to design algorithmically modified speech in a way to optimise intelligibility. A further benefit of the experimental paradigm is in delivering findings at a finer level of granularity than is normal in a traditional experiment that tests a limited number of conditions. This is exemplified by the ability to fine tune centre frequency (Expt. 2), leading to a more refined estimate of the optimal choice than would be possible using the smaller number of octave bands typically studied (e.g., [79]). In addition, preference distributions indicate how sensitive the choice of adjustment is to small perturbations.

However, while much of the time listeners' choices did maximise intelligibility, this outcome has to be tempered by the finding that in a substantial number of trials, listeners' preferences were sub-optimal as evidenced by the reduced intelligibility at values chosen by some listeners on some trials (see Figures 3 and 6). The difficulty in choosing the optimal value is most evident when adjusting the centre frequency of an octave band of speech in Expt. 2, especially in the quiet condition. In that condition, listeners spent nearly twice as long attempting to optimise the band centre than for any of the other seven features tested in the two experiments. The extreme sensitivity to the precise location of an octave band of speech is an unexpected finding and merits further study using a more fine-grained adjustment scale.

5.2. Do Listeners Solely Optimise Intelligibility?

The second motivation for the present study was to determine whether listeners optimise intelligibility alone. The outcomes of both experiments strongly suggest that the answer is no. This is most clear in the quiet conditions where, apart from the centre frequency adjustment, intelligibility was high and close to ceiling levels throughout all or most of the range of adjustment, yet preference distributions were non-uniform. In quiet, listeners tended to prefer small adjustments relative to the unmodified speech.

What is less clear is what drives preferences when intelligibility is already at ceiling. One possibility is that leaving speech close to its original form in terms of spectral balance favours naturalness. Moore and Tan [82] assessed how the perceived naturalness of speech is affected by several spectral distortions in quiet, finding that spectral tilt modifications degrade naturalness, especially when they are applied over the whole frequency range. The situation is somewhat different in the presence of masking noise, where any reduction in naturalness due to spectral distortion may be less perceptually salient.

Speech quality may also play a role in listener preferences. In the present study, listeners shifted the spectral centre of gravity to higher frequencies in noise. High-frequency regions are known to include cues that contribute to the perception of sound quality. Gabriellson et al. [83] evaluated the effect of different frequency weightings on speech quality. The preferred weighting with quiet and with noise at 10 dB SNR was characterised by a flat response at frequencies below 1 kHz and a 6 dB/octave increase in the 1–4 kHz range. This high-frequency boosting led to improvements in characteristics such as brightness, clarity, and spaciousness. One study found that hearing aid users who were allowed to

select their own parameters reported better sound quality than those whose parameters were selected by a clinician [67].

Another possibility is that listeners made their choices in a way that decreased processing demands, reducing listening effort. It has been proposed that lexical retrieval is less demanding when more acoustic information is available, even when intelligibility is at ceiling, due to a reduction in the mismatch between the target speech and its mental representation [84]. A recent study by Borghini and Hazan [85] supports this speculation. They conducted a pupillometry experiment to test the impact of clear versus plain speech in noise on listening effort. Even though intelligibility was equalised for the two types of speech, pupil data suggest that clear speech requires less listening effort. One of the characteristics that differentiates clear from plain speech is greater energy in the mid frequencies, allowing more energy to escape the babble noise masker, perhaps contributing to reduced effort.

5.3. How Do Preferences Change When Listening in Noise?

The other issue addressed concerns whether listeners' preferences change in noise. Apart from bandwidth modification in Expt. 2, all other features showed clear monotonic changes with noise level (Figures 2 and 5). In all cases, the modification direction was the one that promoted audibility in the region above 1 kHz. There is also some evidence that preferences become more consistent as noise level increases. Combined across the 32 preference distributions of the two experiments, distributional entropy is positively correlated ($r = 0.58, p < 0.001$) with intelligibility. The entropy of preference distributions for 6 of the 8 features showed a reduction from quiet to the most adverse SNR, suggesting that noise has a constraining effect on listener choice. Taken together, these lines of evidence suggest that a large component of listener preference in masked conditions is aimed at the maintenance of comprehension rather than expressing choices influenced by supra-intelligibility factors. One explanation might be that listeners need a degree of informational richness as the basis for adjusting speech in ways that address factors other than comprehensibility, and some of the necessary information may not survive energetic masking, or may be sacrificed in the interest of maintaining intelligibility. For example, pitch cues and related intonational information are likely to be less salient in noise following spectral modifications that transfer energy from the sub-1 kHz region.

5.4. Limitations and Further Studies

One major limitation of the listener-centric preferences paradigm adopted in the present study is its inability to distinguish between those signal characteristics that listeners are seeking to optimise when intelligibility is maximised, some of which are explored in Section 5.2 above. Studies like the present one, in which the existence of supra-intelligibility factors and the range of parameter variation over which intelligibility is at a plateau can be assessed, should be regarded as precursors to more detailed investigations, using, for example, pupillometry or dual-task paradigms, which target specific potential characteristics, such as cognitive effort.

It is also not clear whether findings with a speech-shaped masker can be extrapolated to stationary maskers with different spectral profiles, or to non-stationary maskers, such as competing speech. In the latter case, it is possible that listeners will favour modifications that enhance cues that help in sound source segregation, such as those based on the fundamental frequency of the target and masker voices.

Further, the available forms of manipulation were unidimensional, and therefore unrepresentative of both human speech production modifications and those implemented algorithmically. For example, Assmann et al. [86] demonstrated that shifts in fundamental frequency that are not accompanied by appropriate formant frequency changes have a negative impact on naturalness.

While it is normal practice in speech perception studies, such as the present one, to use sentences of moderate linguistic predictability, embedded in maskers at relatively adverse

noise levels, further studies will be needed to determine whether listeners' preferences show similar patterns for the high-context speech material and more benign SNRs, which are more typical of everyday speech communication.

Finally, although allowing listeners to modify speech features to satisfy preferences may be beneficial due to the potential for customising applications to individual needs (see Section 5.5), it may be necessary to use larger sample sizes in future studies in order to provide a more fine-grained view of listeners' preferences without issues of data sparsity that could result from providing listeners with a wide spectrum of response possibilities.

5.5. Potential Impact

In terms of practical applications, the listener-centred paradigm opens up the possibility for future speech technology adapting to individual needs. For example, readily available feedback from listeners in the form of comfortable volume settings might be used to assess the effectiveness of audio adjustments in a process of the continual online adjustment of speech parameters by speech output devices. This paradigm might be applied in the development of algorithms that are specifically designed to enhance speech intelligibility for listeners with hearing loss. For example, we would expect that individual variations in frequency selectivity exhibited by hearing-impaired listeners will result in distinct spectral modification preferences.

Additionally, our study provides significant insights that can be utilised in near-end listening enhancement (NELE) algorithms, such as the algorithm proposed by Chermaz and King [10], which is grounded in audio engineering knowledge. Specifically, for that algorithm, our findings can contribute to the identification of the optimal spectral power distribution (Figure 1 in their paper) for enhancing speech. NELE algorithms in general can benefit from the decision making, regarding the selection of different speech feature values for speech enhancement.

Finally, this experimental paradigm offers the potential to go beyond intelligibility and explore additional applications. One such application is the prevention of fatigue and annoyance in environments where audio announcements are frequently used. By testing different speech modifications, it may be possible to optimise the speech of audio announcements to be more pleasant and less effortful to understand, enhancing the overall listener experience.

6. Conclusions

When provided with real-time auditory feedback, listeners are capable of modifying properties such as the spectral balance or bandwidth of speech, in ways that enhance intelligibility in stationary noise. Moreover, listeners exhibit distinct preferences that appear to be governed by factors other than intelligibility, most notably in quiet conditions but also in moderate levels of noise. Further studies are required to uncover the factors that drive listeners' choices.

Author Contributions: Software, O.S.; investigation, O.S.; data collection, O.S.; methodology, M.C.; supervision, M.C.; conceptualisation, M.C.; formal analysis, O.S. and M.C.; writing, O.S. and M.C.; reviewing and editing, O.S. and M.C. All authors have read and agreed to the published version of the manuscript.

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

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by the Comité de Ética para las Investigaciones con Seres Humanos (CEISH-UPV/EHU), affiliated with the Universidad del País Vasco/Euskal Herriko Unibertsitatea (UPV/EHU) (protocol code M10_2016_112 and 8-11-2016).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Data available on request.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

RMS	Root-Mean-Square
SNR	Signal-to-Noise Ratio
LP	Low-pass
HP	High-pass
BW	Bandwidth
CF	Centre Frequency
SSN	Speech-Shaped Noise
NELE	Near-End Listening Enhancement
OIM	Objective Intelligibility Measures

References

1. Sauert, B.; Vary, P. Near end listening enhancement: Speech intelligibility improvement in noisy environments. In Proceedings of the International Conference on Acoustics, Speech and Signal Processing, Toulouse, France, 14–19 May 2006, pp. 493–496.
2. Skowronski, M.D.; Harris, J.G. Applied principles of clear and Lombard speech for automated intelligibility enhancement in noisy environments. *Speech Commun.* **2006**, *48*, 549–558. [[CrossRef](#)]
3. Yoo, S.D.; Boston, J.R.; El-Jaroudi, A.; Li, C.C.; Durrant, J.D.; Kovacyk, K.; Shaiman, S. Speech signal modification to increase intelligibility in noisy environments. *J. Acoust. Soc. Am.* **2007**, *122*, 1138–1149. [[CrossRef](#)] [[PubMed](#)]
4. Brouckxon, H.; Verhelst, W.; Schuymer, B.D. Time and frequency dependent amplification for speech intelligibility enhancement in noisy environments. In Proceedings of the Ninth Annual Conference of the International Speech Communication Association, Brisbane, Australia, 22–26 September 2008; pp. 557–560. [[CrossRef](#)]
5. Zorila, T.C.; Kandia, V.; Stylianou, Y. Speech-in-noise intelligibility improvement based on spectral shaping and dynamic range compression. In Proceedings of the Thirteenth Annual Conference of the International Speech Communication Association, Portland, OR, USA, 9–13 September 2012; pp. 635–638. [[CrossRef](#)]
6. Taal, C.H.; Hendriks, R.C.; Heusdens, R. Speech energy redistribution for intelligibility improvement in noise based on a perceptual distortion measure. *Comp. Speech Lang.* **2014**, *28*, 858–872. [[CrossRef](#)]
7. Schepker, H.; Rennie, J.; Doclo, S. Speech-in-noise enhancement using amplification and dynamic range compression controlled by the speech intelligibility index. *J. Acoust. Soc. Am.* **2015**, *138*, 2692–2706. [[CrossRef](#)] [[PubMed](#)]
8. Cooke, M.; Mayo, C.; Valentini-Botinhao, C.; Stylianou, Y.; Sauert, B.; Tang, Y. Evaluating the intelligibility benefit of speech modifications in known noise conditions. *Speech Commun.* **2013**, *55*, 572–585. [[CrossRef](#)]
9. Rennie, J.; Schepker, H.; Valentini-Botinhao, C.; Cooke, M. Intelligibility-enhancing speech modifications—The Hurricane Challenge 2.0. In Proceedings of the Interspeech, Shanghai, China, 25–29 October 2020; pp. 1341–1345. [[CrossRef](#)]
10. Chermaz, C.; King, S. A sound engineering approach to near end listening enhancement. In Proceedings of the Interspeech, Shanghai, China, 25–29 October 2020; pp. 1356–1360. [[CrossRef](#)]
11. Valentini-Botinhao, C.; Yamagishi, J.; King, S.; Stylianou, Y. Combining perceptually-motivated spectral shaping with loudness and duration modification for intelligibility enhancement of HMM-based synthetic speech in noise. In Proceedings of the Interspeech, Lyon, France, 25–29 August 2013; pp. 3567–3571. [[CrossRef](#)]
12. Erro, D.; Zorila, T.C.; Stylianou, Y. Enhancing the intelligibility of statistically generated synthetic speech by means of noise-independent modifications. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2014**, *22*, 2101–2111. [[CrossRef](#)]
13. Paul, D.; Shifas, M.P.; Pantazis, Y.; Stylianou, Y. Enhancing speech intelligibility in text-to-speech synthesis using speaking style conversion. In Proceedings of the Interspeech, Shanghai, China, 25–29 October 2020; pp. 1361–1365. [[CrossRef](#)]
14. García Lecumberri, M.L.; Cooke, M.; Cutler, A. Non-native speech perception in adverse conditions: A review. *Speech Commun.* **2010**, *52*, 864–886. [[CrossRef](#)]
15. Stilp, C.E.; Kluender, K.R. Cochlea-scaled entropy, not consonants, vowels, or time, best predicts speech intelligibility. *Proc. Natl. Acad. Sci. USA* **2010**, *107*, 12387–12392. [[CrossRef](#)]
16. Drullman, R.; Festen, J.M.; Plomp, R. Effect of reducing slow temporal modulations on speech reception. *J. Acoust. Soc. Am.* **1994**, *95*, 2670–2680. [[CrossRef](#)]
17. Ghitza, O. On the Role of Theta-Driven Syllabic Parsing in Decoding Speech: Intelligibility of Speech with a Manipulated Modulation Spectrum. *Front. Psychol.* **2012**, *3*, 238. [[CrossRef](#)]
18. Fogerty, D.; Kewley-Port, D. Perceptual contributions of the consonant-vowel boundary to sentence intelligibility. *J. Acoust. Soc. Am.* **2009**, *126*, 847–857. [[CrossRef](#)]
19. Stevens, K. *Acoustic Phonetics*; Current Studies in Linguistics; MIT Press: Cambridge, MA, USA, 2000.
20. Kent, R.; Read, C. *The Acoustic Analysis of Speech*; Singular/Thomson Learning: London, UK, 2002.
21. Johnson, K. *Acoustic and Auditory Phonetics*; Wiley-Blackwell: Hoboken, NJ, USA, 2011.
22. Ladefoged, P.; Johnson, K. *A Course in Phonetics*; Cengage Learning: Boston, MA, USA, 2014.

23. Shannon, R.V.; Zeng, F.G.; Kamath, V.; Wygonski, J.; Ekelid, M. Speech recognition with primarily temporal cues. *Science* **1995**, *270*, 303–304. [[CrossRef](#)]
24. Warren, R.M.; Riener, K.R.; Bashford, J.A.; Brubaker, B.S. Spectral redundancy: Intelligibility of sentences heard through narrow spectral slits. *Percept. Psychophys.* **1995**, *57*, 175–182. [[CrossRef](#)]
25. Lippmann, R.P. Accurate consonant perception without mid-frequency speech energy. *IEEE Trans. Speech Audio* **1996**, *4*, 66–69. [[CrossRef](#)]
26. Cooke, M. A glimpsing model of speech perception in noise. *J. Acoust. Soc. Am.* **2006**, *119*, 1562–1573. [[CrossRef](#)]
27. Kjems, U.; Boldt, J.B.; Pedersen, M.S.; Lunner, T.; Wang, D. Role of mask pattern in intelligibility of ideal binary-masked noisy speech. *J. Acoust. Soc. Am.* **2009**, *126*, 1415–1426. [[CrossRef](#)]
28. French, N.R.; Steinberg, J.C. Factors governing the intelligibility of speech sounds. *J. Acoust. Soc. Am.* **1947**, *19*, 90–119. [[CrossRef](#)]
29. Dau, T.; Püschel, D.; Kohlrausch, A. A quantitative model of the “effective” signal processing in the auditory system. I. Model structure. *J. Acoust. Soc. Am.* **1996**, *99*, 3615–3622. [[CrossRef](#)]
30. ANSI S3.5-1997; American National Standard: Methods for Calculation of Speech Intelligibility Index. American National Standards Institute, Inc.; New York, NY, USA, 1997.
31. Christiansen, C.; Pedersen, M.S.; Dau, T. Prediction of speech intelligibility based on an auditory preprocessing model. *Speech Commun.* **2010**, *52*, 678–692. [[CrossRef](#)]
32. Tang, Y.; Cooke, M. Glimpse-based metrics for predicting speech intelligibility in additive noise conditions. In Proceedings of the Interspeech, San Francisco, CA, USA, 8–12 September 2016; pp. 2488–2492. [[CrossRef](#)]
33. Kryter, K.D. Methods for the calculation and use of the articulation index. *J. Acoust. Soc. Am.* **1962**, *34*, 1689–1697. [[CrossRef](#)]
34. Ma, J.; Hu, Y.; Loizou, P. Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions. *J. Acoust. Soc. Am.* **2009**, *125*, 3387–3405. [[CrossRef](#)] [[PubMed](#)]
35. Healy, E.W.; Yoho, S.E.; Apoux, F. Band importance for sentences and words reexamined. *J. Acoust. Soc. Am.* **2013**, *133*, 463–473. [[CrossRef](#)] [[PubMed](#)]
36. Krause, J.C.; Braida, L.D. Acoustic properties of naturally produced clear speech at normal speaking rates. *J. Acoust. Soc. Am.* **2004**, *115*, 362–378. [[CrossRef](#)]
37. Uchanski, R.M. Clear speech. In *The Handbook of Speech Perception*; John Wiley & Sons, Ltd.: Hoboken, NJ, USA, 2005; pp. 207–235.
38. Summers, W.V.; Pisoni, D.B.; Bernacki, R.H.; Pedlow, R.I.; Stokes, M.A. Effects of noise on speech production: Acoustic and perceptual analyses. *J. Acoust. Soc. Am.* **1988**, *84*, 917–928. [[CrossRef](#)] [[PubMed](#)]
39. Junqua, J.C. The Lombard reflex and its role on human listeners and automatic speech recognizers. *J. Acoust. Soc. Am.* **1993**, *93*, 510–524. [[CrossRef](#)]
40. Garnier, M.; Henrich, N. Speaking in noise: How does the Lombard effect improve acoustic contrasts between speech and ambient noise? *Comp. Speech Lang.* **2014**, *28*, 580–597. [[CrossRef](#)]
41. Takou, R.; Seiyama, N.; Imai, A. Improvement of speech intelligibility by reallocation of spectral energy. In Proceedings of the Interspeech, Lyon, France, 25–29 August 2013; pp. 3605–3607. Available online: https://www.isca-speech.org/archive/interspeech_2013/takou13_interspeech.html (accessed on 20 July 2023).
42. Cooke, M.; Mayo, C.; Villegas, J. The contribution of durational and spectral changes to the Lombard speech intelligibility benefit. *J. Acoust. Soc. Am.* **2014**, *135*, 874–883. [[CrossRef](#)] [[PubMed](#)]
43. Jokinen, E.; Takanen, M.; Vainio, M.; Alku, P. An adaptive post-filtering method producing an artificial Lombard-like effect for intelligibility enhancement of narrowband telephone speech. *Comp. Speech Lang.* **2014**, *28*, 619–628. [[CrossRef](#)]
44. Tang, Y.; Cooke, M. Learning static spectral weightings for speech intelligibility enhancement in noise. *Comp. Speech Lang.* **2018**, *49*, 1–16. [[CrossRef](#)]
45. Hall, J.L.; Flanagan, J.L. Intelligibility and listener preference of telephone speech in the presence of babble noise. *J. Acoust. Soc. Am.* **2010**, *127*, 280–285. [[CrossRef](#)]
46. Moller, S. *Assessment and Prediction of Speech Quality in Telecommunications*; Springer: Berlin, Germany, 2000.
47. Zekveld, A.A.; Kramer, S.E.; Festen, J.M. Pupil response as an indication of effortful listening: The influence of sentence intelligibility. *Ear Heart* **2010**, *31*, 480–490. [[CrossRef](#)] [[PubMed](#)]
48. Dall, R.; Yamagishi, J.; King, S. Rating naturalness in speech synthesis: The effect of style and expectation. In Proceedings of the Speech Prosody 2014, Dublin, Ireland, 20–23 May 2014; pp. 1012–1016. [[CrossRef](#)]
49. Zorilă, T.C.; Stylianou, Y. On the quality and intelligibility of noisy speech processed for near-end listening enhancement. In Proceedings of the Interspeech, Stockholm, Sweden, 20–24 August 2017; pp. 2023–2027. [[CrossRef](#)]
50. Tang, Y.; Arnold, C.; Cox, T. A study on the relationship between the intelligibility and quality of algorithmically-modified speech for normal hearing listeners. *J. Otorhinolaryngol. Hear. Balance Med.* **2018**, *1*, 5. [[CrossRef](#)]
51. Delogu, C.; Conte, S.; Sementina, C. Cognitive factors in the evaluation of synthetic speech. *Speech Commun.* **1998**, *24*, 153–168. [[CrossRef](#)]
52. Govender, A.; King, S. Using pupillometry to measure the cognitive load of synthetic speech. In Proceedings of the Interspeech, Hyderabad, India, 2–6 September 2018; pp. 2838–2842. [[CrossRef](#)]
53. Simantiraki, O.; Cooke, M.; King, S. Impact of different speech types on listening effort. In Proceedings of the Interspeech, Hyderabad, India, 2–6 September 2018; pp. 2267–2271. [[CrossRef](#)]

54. Rennies, J.; Pusch, A.; Schepker, H.; Doclo, S. Evaluation of a near-end listening enhancement algorithm by combined speech intelligibility and listening effort measurements. *J. Acoust. Soc. Am.* **2018**, *144*, EL315–EL321. [CrossRef]
55. Brons, I.; Houben, R.; Dreschler, W.A. Perceptual effects of noise reduction with respect to personal preference, speech intelligibility, and listening effort. *Ear Hear* **2013**, *34*, 29–41. [CrossRef] [PubMed]
56. Zekveld, A.A.; Kramer, S.E. Cognitive processing load across a wide range of listening conditions: Insights from pupillometry. *Psychophysiology* **2014**, *51*, 277–284. [CrossRef] [PubMed]
57. Sarampalis, A.; Kalluri, S.; Edwards, B.; Hafter, E. Objective measures of listening effort: Effects of background noise and noise reduction. *J. Speech Lang. Hear. Res.* **2009**, *52*, 1230–1240. [CrossRef]
58. Govender, A.; King, S. Measuring the cognitive load of synthetic speech using a dual task paradigm. In Proceedings of the Interspeech, Hyderabad, India, 2–6 September 2018; pp. 2843–2847. [CrossRef]
59. McGarrigle, R.; Munro, K.J.; Dawes, P.; Stewart, A.J.; Moore, D.R.; Barry, J.G.; Amitay, S. Listening effort and fatigue: What exactly are we measuring? A British Society of Audiology Cognition in Hearing Special Interest Group white paper. *Int. J. Audiol.* **2014**, *53*, 433–440. [CrossRef]
60. Assmann, P.F.; Nearey, T.M. Relationship between fundamental and formant frequencies in voice preference. *J. Acoust. Soc. Am.* **2007**, *122*, EL35–EL43. [CrossRef]
61. Wingfield, A.; Ducharme, J.L. Effects of age and passage difficulty on listening-rate preferences for time-altered speech. *J. Gerontol. Ser. B* **1999**, *54B*, P199–P202. [CrossRef]
62. Novak, J.S.; Kenyon, R.V. Effects of user controlled speech rate on intelligibility in noisy environments. In Proceedings of the Interspeech, Hyderabad, India, 2–6 September 2018; pp. 1853–1857.
63. Simantiraki, O.; Cooke, M. Exploring listeners’ speech rate preferences. In Proceedings of the Interspeech, Shanghai, China, 25–29 October 2020; pp. 1346–1350. [CrossRef]
64. Torcoli, M.; Freke-Morin, A.; Paulus, J.; Simon, C.; Shirley, B. Preferred levels for background ducking to produce esthetically pleasing audio for tv with clear speech. *J. Audio Eng. Soc.* **2019**, *67*, 1003–1011. [CrossRef]
65. Zhang, Z.; Shen, Y. Listener preference on the local criterion for ideal binary-masked speech. In Proceedings of the Interspeech, Graz, Austria, 15–19 September 2019; pp. 1383–1387. [CrossRef]
66. Boothroyd, A.; Mackersie, C. A “Goldilocks” Approach to Hearing-Aid Self-Fitting: User Interactions. *Am. J. Audiol.* **2017**, *26*, 430–435. [CrossRef] [PubMed]
67. Sabin, A.T.; Tasell, D.J.V.; Rabinowitz, B.; Dhar, S. Validation of a Self-Fitting Method for Over-the-Counter Hearing Aids. *Trends Heart* **2020**, *24*, 2331216519900589. [CrossRef]
68. Simantiraki, O.; Cooke, M.; Pantazis, Y. Effects of spectral tilt on listeners’ preferences and intelligibility. In Proceedings of the International Conference on Acoustics, Speech and Signal Processing, Barcelona, Spain, 4–8 May 2020; pp. 6254–6258.
69. Simantiraki, O.; Cooke, M. SpeechAdjuster: A tool for investigating listener preferences and speech intelligibility. In Proceedings of the Interspeech, Brno, Czechia, 30 August–3 September 2021; pp. 1718–1722. [CrossRef]
70. Aubanel, V.; García Lecumberri, M.L.; Cooke, M. The Sharvard Corpus: A phonemically-balanced Spanish sentence resource for audiology. *Int. J. Audiol.* **2014**, *53*, 633–638. [CrossRef] [PubMed]
71. Rothaus, E.H.; Chapman, W.D.; Guttman, N.; Silbiger, H.R.; Hecker, M.H.L.; Urbanek, G.E.; Nordby, K.S.; Weinstock, M. IEEE recommended practice for speech quality measurements. *IEEE Trans. Audio Electroacoust.* **1969**, *17*, 225–246.
72. Cooke, M.; García Lecumberri, M.L. How reliable are online speech intelligibility studies with known listener cohorts? *J. Acoust. Soc. Am.* **2021**, *150*, 1390–1401. [CrossRef]
73. Holm, S. A simple sequentially rejective multiple test procedure. *Scand. J. Stat.* **1979**, *6*, 65–70.
74. Virtanen, P.; Gommers, R.; Oliphant, T.E.; Haberland, M.; Reddy, T.; Cournapeau, D.; Burovski, E.; Peterson, P.; Weckesser, W.; Bright, J.; et al. SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nat. Methods* **2020**, *17*, 261–272. [CrossRef]
75. Bates, D.; Mächler, M.; Bolker, B.; Walker, S. Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* **2015**, *67*, 1–48. [CrossRef]
76. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2021.
77. Lenth, R.V. *Emmeans: Estimated Marginal Means, aka Least-Squares Means*, R Package Version 1.5.5-1; 23-06-2023; 2021. Available online: <https://cran.r-project.org/web/packages/emmeans/index.html> (accessed on 20 July 2023).
78. Lu, Y.; Cooke, M. Speech production modifications produced by competing talkers, babble, and stationary noise. *J. Acoust. Soc. Am.* **2008**, *124*, 3261–3275. [CrossRef]
79. Warren, R.M.; Bashford, J.A., Jr.; Lenz, P.W. Intelligibilities of 1-octave rectangular bands spanning the speech spectrum when heard separately and paired. *J. Acoust. Soc. Am.* **2005**, *118*, 3261–3266. [CrossRef] [PubMed]
80. Lu, Y.; Cooke, M. Speech production modifications produced in the presence of low-pass and high-pass filtered noise. *J. Acoust. Soc. Am.* **2009**, *126*, 1495–1499. [CrossRef]
81. Bilger, R.C.; Nuetzel, J.M.; Rabinowitz, W.M.; Rzeczkowski, C. Standardization of a test of speech perception in noise. *J. Speech Hear Res.* **1984**, *27*, 32–48. [CrossRef] [PubMed]
82. Moore, B.C.J.; Tan, C.T. Perceived naturalness of spectrally distorted speech and music. *J. Acoust. Soc. Am.* **2003**, *114*, 408–419. [CrossRef] [PubMed]

83. Gabrielsson, A.; Schenkman, B.; Hagerman, B. The effects of different frequency responses on sound quality judgments and speech intelligibility. *J. Speech Lang. Hear Res.* **1988**, *31*, 166–177. [[CrossRef](#)]
84. Rönnerberg, J.; Lunner, T.; Zekveld, A.; Sörqvist, P.; Danielsson, H.; Lyxell, B.; Dahlström, O.; Signoret, C.; Stenfelt, S.; Pichora-Fuller, M.; et al. The Ease of Language Understanding (ELU) model: Theoretical, empirical, and clinical advances. *Front. Syst. Neurosci.* **2013**, *7*, 31. [[CrossRef](#)] [[PubMed](#)]
85. Borghini, G.; Hazan, V. Effects of acoustic and semantic cues on listening effort during native and non-native speech perception. *J. Acoust. Soc. Am.* **2020**, *147*, 3783–3794. [[CrossRef](#)] [[PubMed](#)]
86. Assmann, P.F.; Dembling, S.; Nearey, T.M. Effects of frequency shifts on perceived naturalness and gender information in speech. In Proceedings of the Ninth International Conference on Spoken Language Processing, Pittsburgh, PA, USA, 17–21 September 2006; pp. 889–892.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.