Universidad del País Vasco
Euskal Herriko Unibertsitatea

Investigating supra-intelligibility aspects of speech

Olympia Simantiraki

Departamento de Filología Inglesa y Alemana y Traducción e Interpretación

Universidad del País Vasco (UPV/EHU)

Supervised by

Prof. Martin Cooke and Prof. María Luisa García Lecumberri

2022

To my family

Acknowledgements

I would like to express my special thanks to my supervisors, Professor Martin Cooke and Professor María Luisa García Lecumberri. Professor Martin Cooke provided me with guidance and feedback throughout this project. His insight and knowledge into the subject matter steered me through this research.

A big thanks to all the members of LASLAB, and especially to Rubén (who was the first member of the LASLAB that I met), Edurne, Asier, Marta and Monica, for the pleasant atmosphere in the office and their patience to help me with the Spanish bureaucracy when needed.

A big thanks also to all the ENRICH members with whom we shared many experiences, traveling a lot, meeting new places, acquiring a lot of knowledge, sharing concerns on our PhDs, and having a lot of fun.

I could definitely not miss out to thank my friends Maria and Igor, as well as other Spanish friends who I made during my stay in the Basque Country, Olaia, Arantxa, Diego, Maiara, Rubén and Sergio. They really facilitated my stay there.

A special thanks to my beloved Greek friends Eleni, Kallia, Eirini, Giorgos, Andreas for the psychological support and Anastasia who was one of the persons who encouraged me to apply for this PhD.

Last but not least I would like to thank my family for all the support they showed me through this research. They are always standing by my side whatever I choose to do. Specifically, my husband Pavlos and our 8 months old baby Antonis, my parents Nantia and Manolis, my sisters Penny, Natalia, Ismini and their families for providing me with guidance and a sounding board when required. Their frequent visits to me in Vitoria made my stay easier. Finally, my husband Pavlos, thanks for all your support, without which I would have stopped these studies a long time ago.

Abstract

Synthetic and recorded speech form a great part of our everyday listening experience, and much of our exposure to these forms of speech occurs in potentially noisy settings such as on public transport, in the classroom or workplace, while driving, and in our homes. Optimising speech output to ensure that salient information is both correctly and effortlessly received is a main concern for the designers of applications that make use of the speech modality. Most of the focus in adapting speech output to challenging listening conditions has been on intelligibility, and specifically on enhancing intelligibility by modifying speech prior to presentation. However, the quality of the generated speech is not always satisfying for the recipient, which might lead to fatigue, or reluctance in using this communication modality. Consequently, a sole focus on intelligibility enhancement provides an incomplete picture of a listener's experience since the effect of modified or synthetic speech on other characteristics risks being ignored. These concerns motivate the study of 'supra-intelligibility' factors such as the additional cognitive demand that modified speech may well impose upon listeners, as well as quality, naturalness, distortion and pleasantness.

This thesis reports on an investigation into two supra-intelligibility factors: listening effort and listener preferences. Differences in listening effort across four speech types (plain natural, Lombard, algorithmically-enhanced, and synthetic speech) were measured using existing methods, including pupillometry, subjective judgements, and intelligibility scores. To explore the effects of speech features on listener preferences, a new tool, SpeechAdjuster, was developed. SpeechAdjuster allows the manipulation of virtually any aspect of speech and supports the joint elicitation of listener preferences and intelligibility measures. The tool reverses the roles of listener and experimenter by allowing listeners direct control of speech characteristics in real-time. Several experiments to explore the effects of speech properties on listening preferences and intelligibility using SpeechAdjuster were conducted. Participants were permitted to change a speech feature during an open-ended adjustment phase, followed by a test phase in which they identified speech presented with the feature value selected at the end of the adjustment phase. Experiments with native normal-hearing listeners measured the consequences of allowing listeners to change speech rate, fundamental frequency, and other features which led to spectral energy redistribution. Speech stimuli were presented in both quiet and masked conditions.

Results revealed that listeners prefer feature modifications similar to those observed in naturally modified speech in noise (Lombard speech). Further, Lombard speech required the least listening effort compared to either plain natural, algorithmically-enhanced, or synthetic speech. For stationary noise, as noise level increased listeners chose slower speech rates and flatter tilts compared to the original speech. Only the choice of fundamental frequency was not consistent with that observed in Lombard speech. It is possible that features such as

fundamental frequency that talkers naturally modify are by-products of the speech type (e.g. hyperarticulated speech) and might not be advantageous for the listener.

Findings suggest that listener preferences provide information about the processing of speech over and above that measured by intelligibility. One of the listeners' concerns was to maximise intelligibility. In noise, listeners preferred the feature values for which more information survived masking, choosing speech rates that led to a contrast with the modulation rate of the masker, or modifications that led to a shift of spectral energy concentration to higher frequencies compared to those of the masker. For all features being modified by listeners, preferences were evident even when intelligibility was at or close to ceiling levels. Such preferences might result from a desire to reduce the cognitive effort of understanding speech, or from a desire to reproduce the sound of typical speech features experienced in real-world noisy conditions, or to optimise the quality of the modified signal.

Investigation of supra-intelligibility aspects of speech promises to improve the quality of speech enhancement algorithms, bringing with it the potential of reducing the effort of understanding artificially-modified or generated forms of speech.

Extracto

El habla sintética y el habla grabada forman gran parte de nuestra experiencia auditiva diaria, y la mayoría de nuestra exposición a estas formas de habla ocurre en entornos potencialmente ruidosos, como el transporte público, las aulas o el lugar de trabajo, mientras conducimos y en nuestros hogares. Optimizar la salida del habla para garantizar que la información destacada se reciba correctamente y sin esfuerzo es una preocupación principal para los diseñadores de aplicaciones que hacen uso de la voz. La mayoría de los trabajos destinados a adaptar el habla a condiciones auditivas desafiantes se han centrado en la inteligibilidad, y específicamente en mejorar la inteligibilidad mediante la modificación del habla antes de su presentación. Sin embargo, la calidad del discurso generado no siempre es satisfactoria para el receptor, lo que puede llevar a la fatiga o reticencia en el uso de esta modalidad de comunicación. En consecuencia, un enfoque exclusivo en la mejora de la inteligibilidad proporciona una imagen incompleta de la experiencia de un oyente, ya que el efecto del habla modificada o sintética en otros aspectos corre el riesgo de ser ignorado. Estas consideraciones motivan el estudio de factores de 'supra-inteligibilidad', como la demanda cognitiva adicional que el habla modificada puede suponer para los oyentes, así como su calidad, naturalidad, distorsión y lo agradable que pueda resultar dicha habla.

Esta tesis se centra en la investigación de dos factores de supra-inteligibilidad: el esfuerzo de escucha y las preferencias del oyente. Se midieron las diferencias en el esfuerzo de escucha ante cuatro tipos de habla (natural simple, habla Lombard, habla mejorada algorítmicamente y habla sintética) utilizando los métodos existentes, incluida la pupilometría, juicios subjetivos y puntuaciones de inteligibilidad. Para explorar los efectos de las funciones de voz en las preferencias del oyente, se desarrolló una nueva herramienta, SpeechAdjuster. SpeechAdjuster permite la manipulación de prácticamente cualquier aspecto del habla así como la elicitación conjunta de las preferencias del oyente y las medidas de inteligibilidad. La herramienta invierte los roles de oyente y experimentador al facilitar a los oyentes el control directo de las características del habla en tiempo real. Se realizaron varios experimentos para explorar los efectos de las propiedades del habla en las preferencias de escucha y la inteligibilidad utilizando SpeechAdjuster. A los participantes se les permitió cambiar una función de voz durante una fase de ajuste abierta, seguida de una fase de prueba en la que identificaron el discurso presentado con el valor de la función seleccionado al final de la fase de ajuste. Los experimentos con oyentes nativos con capacidad auditiva normal midieron las consecuencias de permitir que los oyentes cambien la velocidad del habla, la frecuencia fundamental y otras características destinadas a la redistribución de la energía espectral. Los estímulos de habla se presentaron tanto sin ruido como con ruido de fondo (enmascaramiento).

Los resultados revelaron que los oyentes prefieren modificaciones de características similares a las observadas en el habla modificada naturalmente en presencia de ruido (habla Lombard).

Además, el habla Lombard requirió el menor esfuerzo de escucha en comparación con el habla natural simple, el habla mejorada algorítmicamente o el habla sintética. En condiciones de ruido estacionario, a medida que aumentaba el nivel de ruido, los oyentes eligieron velocidades de habla más lentas e inclinaciones espectrales más planas en comparación con el discurso original. Sólo la elección de la frecuencia fundamental no fue consistente con la observada en el habla Lombard. Es posible que características como la frecuencia fundamental que los hablantes modifican naturalmente sean subproductos del tipo de habla (por ejemplo, habla hiperarticulada) y no sean ventajosas para el oyente.

Los resultados también sugieren que las preferencias de los oyentes proporcionan información sobre el procesamiento del habla más allá de lo que la inteligibilidad indica. Una de las preocupaciones de los oyentes fue maximizar la inteligibilidad. En condiciones de ruido, los oyentes preferían los valores de características en los que más información sobrevivía al enmascaramiento, eligiendo velocidades de habla que condujeran a un contraste con la tasa de modulación del enmascarador, o modificaciones que condujeran a un cambio de concentración de energía espectral a frecuencias más altas en comparación con las del enmascarador. Para todas las características modificadas por los oyentes, las preferencias eran evidentes incluso cuando la inteligibilidad estaba en o cerca de niveles máximos. Tales preferencias pueden obedecer a un deseo de reducir el esfuerzo cognitivo para comprender el habla, o a un deseo de reproducir el sonido típico de las características del habla experimentadas en condiciones ruidosas del mundo real, o de mejorar la calidad de la señal modificada.

La investigación de los aspectos suprainteligibles del habla promete optimizar la calidad de los algoritmos de mejora del habla, y por consiguiente, el potencial de reducir el esfuerzo de comprensión de formas de habla modificadas o generadas artificialmente.

Contents

Chapter 1

Introduction

1.1 Motivation

In our everyday life we are exposed to a variety of speech types, both naturally and artificially produced. Both speakers and speech enhancement developers attempt to help the listener by modifying the speech characteristics. Talkers modify their speech when exposed to noise, producing Lombard speech. Live and recorded public address announcements may involve modifications designed to enhance intelligibility. Synthetically-generated speech is commonplace in mobile devices, voice assistants and telephone enquiry systems. Speech understanding in ideal conditions is automatic and effortless. However, several factors, such as ambient noise and the listener's limitations, may have a negative effect on the perception of speech.

Correct message reception is critical in many situations. Consequently, a great deal of effort has been devoted to evaluating the effect on intelligibility of different speech styles [Cooke et al., 2013a] and changes in distinct speech properties [Nejime and Moore, 1998; Adams and Moore, 2009; Lu and Cooke, 2009a]. Near-end listening enhancement algorithms can achieve significant improvements in speech understanding compared to unprocessed speech under adverse conditions [Taal and Jensen, 2013; Schepker et al., 2015]. However, the perceived speech after the near-end listening enhancement might not be completely satisfying for the listener, since commonly used speech enhancement algorithms mainly focus on intelligibility improvements. Other subjective aspects of perceived speech, such as listening effort, quality, naturalness, pleasantness and overall listener preferences, also need to be considered. To refer to speech attributes above and beyond word recognition, the term 'supra-intelligibility' is used. The objective of this thesis is twofold: to study supra-intelligibility aspects of speech in terms of both listening effort and listener preferences, and to develop a tool for investigating supra-intelligibility aspects of speech.

Complementary to the speech clarity dimension is the overall listener's experience, which has been far less investigated. Listening can become hard, even when intelligibility is at ceiling. Listening effort reflects the cognitive resources necessary for speech understanding. A great exertion of effort is sometimes necessary in situations with background noise, low speech intensity, poor mobile connection, accented speech, or listener's high motivation (e.g. larger peak pupil dilation for higher rewards; for a recent review see Carolan et al. [2022]). Variations in listening effort can also be found among different populations. For instance, non-native listeners exert greater effort compared to native listeners, even when performing a task to the same level

[Borghini and Hazan, 2018].

A high allocation of cognitive resources imposes a great handicap on the listener, leading to reduced performance in multi-tasks [Sarampalis et al., 2009], a stronger feeling of listening and/or mental fatigue, or rejection of social life. In a more extreme case [World Health Organization. Regional Office for Europe., 2011], working in an unpleasant environment with frequent and loud announcements may lead to ill health. Listening effort has been estimated using subjective measures such as questionnaires, behavioural metrics (e.g. response time), and physiological measures such as pupillometry (e.g. see review by McGarrigle et al. [2014]).

Listening effort can be considered as one of the several individual aspects of listener preferences. Listener preferences arise from the generalised judgement of speech perception that includes factors such as intelligibility, naturalness and pleasantness. Listener preferences can be collected by allowing listeners to modify speech properties using adjustment tools. Listeners are familiar with the concept of smooth audio modifications, such as that used for volume adjustment on television and radio. Listeners' responses derived from speech-adjustment tools can be precise, since the speech can be fine-tuned, in contrast to traditional tests in which the listener is provided only with few options. Previous studies have suggested different real-time audio feature modifications [Assmann and Nearey, 2007; Kean et al., 2015; Zhang and Shen, 2019; Novak III and Kenyon, 2018]. Such preferences can be expected to vary according to the listening environment [Kean et al., 2015; Walton et al., 2016], and any hearing impairment [Buyens et al., 2014; Shirley et al., 2017], as well as having an individual component [Walton et al., 2016]. A better understanding of the basis for listener preferences promises to inform the design of speech modification algorithms that are capable of both increasing intelligibility and reducing listening effort, providing a better overall listening experience.

The motivation for this thesis is to elucidate the 'hard to listen' effect (i.e. the condition in which speech requires more effort from the listener) by evaluating the contribution of distinct speech factors to this effect for different listening conditions. The overarching objective of this thesis is to determine whether listeners exhibit supra-intelligibility preferences when they are given the power to manipulate distinct speech properties that speakers naturally modify to produce Lombard speech. The main hypothesis is when intelligibility is at ceiling levels, listeners will attempt to reduce listening effort and maintain speech quality. It is assumed that, for the conditions where intelligibility is maximised, the relationship between listener preferences and a wide range of speech feature values (e.g. spectral slope) will be a bell-shaped distribution, as shown in Fig. 1.1. Besides the user-centric aspect of listener preferences, when listeners of the same group (e.g. younger vs older adults, normal-hearing vs hearing impaired, native vs non-native) are listening under the same conditions (e.g. stationary noise, competing speech), it is expected that similar speech feature values will excite similar cognitive or hearing patterns (e.g. frequencies to which the listener is particularly sensitive, speech intensity, prosodic features of a language that are familiar to the listener).

1.2 Outline

Chapter 2 surveys previous work on speech perception in terms of intelligibility and supra-intelligibility aspects of speech. Studies of speech quality, listening effort and listener preferences are included, and different ways to measure supra-intelligibility aspects that have previously been used are described. Figure 1.2 illustrates the main focus of each of chapters 3 to 7.

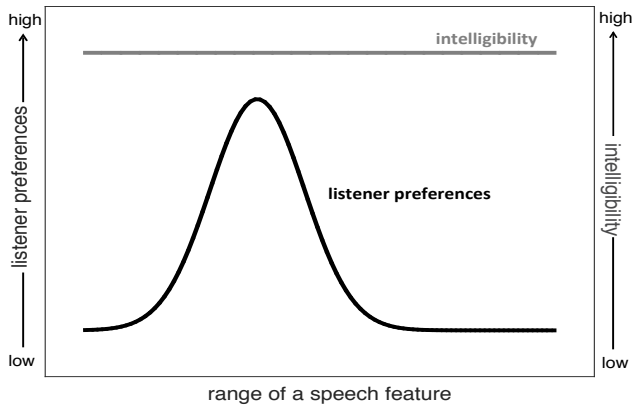


Figure 1.1: *The plot illustrates how the listener preferences (left axis) are expected to vary when intelligibility (right axis) is at ceiling as a function of a speech feature (x-axis).*

In chapter 3, the impact of different speech types and of nativeness on listening effort were studied. Three measures of listening effort were investigated: (i) an objective measure of intelligibility, (ii) a physiological measure of listening effort (pupil size), and (iii) listeners' subjective judgements. The examined speech types were plain (natural) speech, speech produced in noise (Lombard speech), speech enhanced to promote intelligibility, and synthetic speech.

Chapter 4 describes SpeechAdjuster, an open source tool that reverses the roles of listener and experimenter by allowing listeners direct control of speech characteristics in real-time. This change of paradigm enables listeners' preferences to be measured directly, without recourse to rating scales. Incorporation of a test phase in which the preferences are frozen also enables intelligibility to be estimated within the same trial. Offline computation and smooth online interpolation within the tool permit measurement of the impact of changes in practically any target speech feature (e.g. fundamental frequency or spectral slope) or background characteristic (e.g. noise spectrum), regardless of complexity.

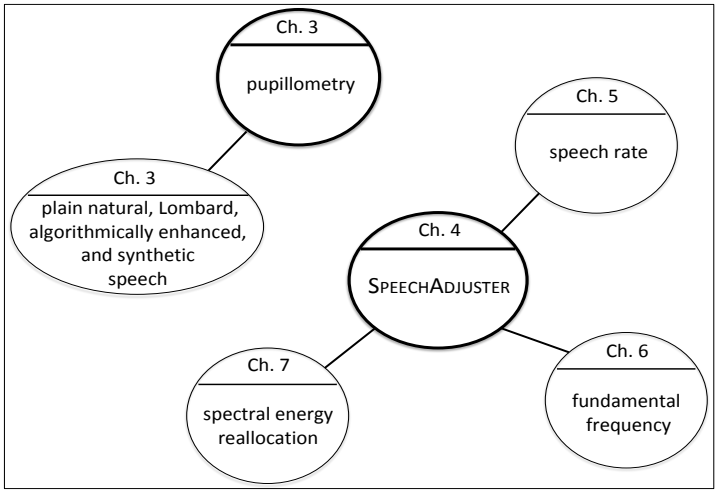


Figure 1.2: *An overview of the different studies conducted in this thesis. The ellipses with the thick black lines correspond to the methods used for investigating the supra-intelligibility aspects of speech and the attached nodes show the speech types or the distinct features that were studied.*

Several experiments were conducted using SpeechAdjuster to explore the effects on listening preferences and intelligibility of speech properties that are typically modified in naturally

enhanced speech (chapters 5-7). In chapter 5, the relationship between speech rate and masker properties was investigated. Listeners adjusted speech rate while listening to word sequences in quiet, in stationary noise at different noise levels, and in modulated noise for 5 envelope-modulation rates. After selecting a preferred rate, participants went on to identify words presented at that rate. Preferences regarding the fundamental frequency for speech presented in energetic and in competing speech maskers were tested in chapter 6, while in chapter 7, spectral properties (spectral tilt, spectral band energy modifications, frequency filter characteristics) of speech presented under conditions of energetic masking were investigated.

This thesis studies the effects of speech features *other than signal intensity* on listener preferences. Throughout, stimuli are normalised to have the same RMS energy before and after modification. This approach is common, for example, in evaluating the performance of speech enhancement algorithms [Cooke et al., 2013a; Rennies et al., 2020], and leads to a focus on speech modifications that benefit listeners independent from the simple expedient of increasing audibility by raising signal level. One consequence of normalisation is that speech modifications always represent the joint outcome resulting from both the direct effect of the modified parameter itself (e.g. flatter spectral tilt), and the effect on any change in local SNR across time and frequency due to the subsequent normalisation (e.g. more energy in mid frequencies). In order to assess changes in local SNR that result from speech parameter modification, chapter 6 introduces a new metric that measures the distribution of speech glimpses across frequency.

The pupillometry study was used as a reference study to understand the effort involved in speaking styles that can be found in real-life conditions. Each of the tested speech types involves changes of one or more of the features studied in chapters 5, 6 and 7. Thus, insights can be derived as to whether the preferred values of the distinct speech features have contributed to reducing the listening effort required.

Finally, chapter 8 describes the main findings, novel contributions and conclusions. Indirect comparisons between listener preferences and listening effort are provided and future paths are also discussed.

1.3 Research questions

In this thesis I have tried to answer the following research questions:

1. How is listening effort affected by the presence of noise?
2. What is the difference between naturally produced, artificially enhanced, and synthetically produced speech in terms of cognitive processing load?
3. Do listener preferences change under challenging conditions?
4. Do listeners always choose a preference that maximises intelligibility?
5. What is the nature of preferences when intelligibility is constant?

Chapter 2

Supra-intelligibility aspects of speech

Intelligibility is the main factor that listeners take into account when judging speech quality. When intelligibility is constant, other dimensions emerge, such as pleasantness, naturalness and listening effort [Preminger and Tasell, 1995]. In noisy conditions, improving speech intelligibility is more important for the listener, while for more favourable SNRs speech quality has an impact [Tang et al., 2018]. Previous studies have shown that, for equivalent speech intelligibility, other factors can increase the cognitive effort: e.g. when attending to plain rather than clear speech [Borghini and Hazan, 2020], or to synthetic speech rather than natural [Pisoni et al., 1987]. While intelligibility has been widely used for evaluating speech, supra-intelligibility aspects (i.e. speech aspects beyond intelligibility) have been far less investigated. Section 2.1 describes one approach to the speech understanding mechanism and intelligibility. Section 2.2 presents the methods traditionally used to evaluate supra-intelligibility. The remaining sections survey the different subjective and objective measures for specifically evaluating speech quality (sec. 2.3), listening effort (sec. 2.4), and listener preferences (sec. 2.5).

2.1 Speech understanding mechanism

Speech perception is a process involving three sequential steps; a speech sound is heard, interpreted, and understood [Moore et al., 2008]. Specifically, the auditory information is received; subsequently it is transformed to a neural signal; and finally, the phonetic information is processed. Speech processing is automatic and effortless when it happens under ideal conditions. In quiet, the speech information in frequency and time is in excess of that required for perceiving speech accurately by normal-hearing listeners [Moore, 2008]. In noise, however, the speech perception task becomes more difficult and additional work is required from the automatic processes. One hypothesis is that two automatic processes are involved in the perception of speech: i.e. bottom-up and top-down processes. During the bottom-up process, the incoming speech signal is analysed, while the top-down process is based on the listener's prior knowledge. The brain is capable of isolating certain sound sources and filtering out others ('selective gain' mechanism [Kerlin et al., 2010]). In the literature, several techniques have been suggested as being involved in the automatic mechanism. Some of these are the clustering and stitching of speech pieces into a single signal (i.e. auditory grouping [Bregman, 1990]), extraction of

time-frequency regions where the target speech is less masked (i.e. glimpsing [Cooke, 2003]), or spatial source separation between the target and the masker when located in different regions [Hawley et al., 2004]. Finally, visual cues are also a useful mechanism for distinguishing phonemes in noise [Macleod and Summerfield, 1987].

Although there are a number of factors that can interfere with optimal speech comprehension, normal-hearing listeners are able to understand speech under severe conditions [Diehl, 2008]. In order to achieve successful communication, talkers naturally modify their speaking style to take account of the environmental conditions and their interlocutor [for a review see Cooke et al., 2014a]. Such environmental conditions can be additive ambient noise, in which the talker produces the so-called 'Lombard' speech (e.g. at a cocktail party) [Summers et al., 1988; Hazan and Baker, 2011], reverberation [Brunskog et al., 2009], or wide separation between talker and interlocutor [Pelegrín-García et al., 2011]. On the other hand, speech types related to interlocutor's characteristics include speech directed to infants [Burnham et al., 2002], children with learning disabilities [Bradlow et al., 2003], hearing-impaired listeners [Lam and Kitamura, 2012], non-natives [Sankowska et al., 2011], machines [Mayo et al., 2012], or pets [Burnham et al., 2002].

The talker's intention is to facilitate the listener's comprehension by increasing the speech clarity and reducing the required cognitive effort. They achieve this by making acoustic and linguistic adaptations, separately or in combination. For the acoustic modifications in particular, one mechanism is to improve audibility by increasing vocal intensity [Picheny et al., 1986; Castellanos et al., 1996; Pelegrín-García et al., 2011], raising the fundamental frequency to shift the spectrum to frequencies to which the ear is more sensitive [Bond and Moore, 1990; Pittman and Wiley, 2001], enhancing voiced sounds in intensity and duration [Boril and Pollak, 2005], and reallocating spectro-temporal energy [Lu and Cooke, 2008]. Another mechanism is to increase the speech coherence in the presence of competing sounds by increasing speech modulation [Krause and Braida, 2004; Boril and Pollak, 2005], with changes in the first two formants [Picheny et al., 1986; Bradlow et al., 2003], or by inserting pauses between words [Picheny et al., 1986]. Finally, linguistic level modifications can also be applied, such as using a simpler vocabulary.

To study the effect of different speech types on speech perception, researchers usually evaluate the intelligibility. Intelligibility can be defined as the percentage of words accurately recognised (word recognition rate). Factors that can reduce intelligibility include imperfect listening conditions, with or without aspects such as ambient noise or reverberation; the interlocutor's limitations, such as being a non-native, or hearing-impaired; and the talker's limitations, such as accented speech. Intelligibility decreases as a function of SNR: i.e. a lower SNR leads to lower intelligibility. Additionally, intelligibility reduction is affected differently for the different types of speech, SNRs, distortions, maskers and reverberation [Picheny et al., 1985, 1986; Summers et al., 1988; Robinson et al., 2002]. For example, elongated speech has been shown to increase intelligibility in babble noise [Adams and Moore, 2009], while in stationary noise no significant gains were observed [Nejime and Moore, 1998]. Moreover, spectral tilt flattening led to gains in intelligibility in the presence of noise, but increasing the fundamental frequency did not have any impact [Lu and Cooke, 2009a].

2.2 Traditional measures of supra-intelligibility aspects of speech

Supra-intelligibility aspects of speech are highly subjective in nature and thus they have traditionally been measured using subjective judgements. One of the widely used methods is the collection of subjective ratings. Subjective ratings are fast, easily distributed (no special equipment is needed) and easily developed. These paradigms require a participant to map a large and potentially-complex subjectively-interpreted concept, such as quality, on to a rather artificial and usually discrete set of values, such as 'very natural', 'quite natural' and the like. Furthermore, while intelligibility and subjective factors can be measured in the same task, for practical reasons these measurements are sequential and hence delayed relative to the stimulus, raising issues such as whether individual differences in working memory capacity might affect the outcome.

Another commonly used evaluation method over different systems is the paired comparison test. To test the performance of a system over N others all the possible pairs with the reference system have to be presented and evaluated separately. This method can be time-consuming, while another disadvantage is that the results can be either binary or limited to a 4-point scale (i.e. comparison category rating test) [Loizou, 2011].

Differences in the internal standards of listener groups can result in great variability in the evaluation. For example, in Larsby et al. [2005], elderly adults did not report greater listening effort than young adults, despite their worse performance in the task. Another limitation of subjective judgements is that individuals might interpret the notions under investigation, e.g. listening effort, differently. More specifically, previous studies showed that subjective ratings of listening effort were correlated with task performance [Gosselin and Gagné, 2011; Johnson et al., 2015; Seeman and Sims, 2015]. In addition, listener's judgements might be influenced by their experiences. In Tang et al. [2018], modified speech affected perceptual quality for listeners who preferred plain over modified speech under quiet conditions.

Despite the wide use of subjective methods for evaluating supra-intelligibility aspects of speech, they are not always consistent with objective methods. For example, a large-scale validation study was conducted for evaluating the convergent validity and sensitivity of commonly used measures of listening effort, concluding that listening effort measures are not consistently or strongly intercorrelated [Strand et al., 2018]. In line with this study, other studies have shown that subjective measures of listening effort are correlated with listeners' task performance and not with objective measures of listening effort [Gosselin and Gagné, 2011; Johnson et al., 2015; Seeman and Sims, 2015].

The following sections present the different methods that have been used in the literature for evaluating speech quality, listening effort, and listener preferences. Apart from the usually used forced-choice paradigms, listener-driven tools to explore supra-intelligibility aspects of speech, by allowing the listener to modify speech properties in real-time using adjustment tools, are also presented.

2.3 Speech quality

Speech quality can be influenced by several perceptual attributes, such as intelligibility, listening effort, pleasantness, naturalness, loudness, and overall experience. Sometimes the speech quality

term is not defined in the experiment, but instead, individuals are free to judge according to their listening experience [e.g. Tang et al., 2018]. Previous studies have shown that perceived speech quality is poorer when the speech signal is attenuated in the range where the pitch and harmonic information occur (below 1000 Hz) [Gabrielsson et al., 1988]. Speech quality also varies as a function of changes to the frequency response [Gabrielsson et al., 1988, 1990]. Temporal modifications have been shown to have a more negative impact on speech quality, compared to spectral modifications and methods of enhancing specific time–frequency regions [Tang and Cooke, 2010]. However, listeners might not notice the deteriorated speech quality under conditions in which the masker covers artefacts or distortions of the speech signal [Taal et al., 2014].

2.3.1 Subjective measures

For the subjective quality of speech in noise, an evaluation methodology was suggested in the ITU-T P.835 [ITU-T, 2003]. First, three separate quality ratings have to be assessed. Listeners have to attend to the speech signal alone, the background noise alone and the speech plus noise, and rate them separately on a five-point scale. Finally, the mean opinion score (MOS) and the absolute category rating (ACR) are derived. This methodology has been applied for evaluating speech enhancement algorithms subjectively [Rohdenburg et al., 2005; Hu and Loizou, 2006, 2007, 2008].

A different category rating scale was used in Preminger and Tasell [1995] for investigating the relation between speech quality and intelligibility. Listeners had to assign a number to quantify the speech they heard, along with the dimension of interest, by pointing to a numbered location on a line using the mouse. Each quality dimension (i.e. intelligibility, pleasantness of tone, listening effort, loudness, and total impression) was rated on a scale from 0 to 100. In a first experiment, in which the intelligibility varied, results showed that the ratings of loudness, effort, and total impression could be predicted by the judged intelligibility. In a second experiment, in which the intelligibility was kept constant, results revealed that the listeners interpreted the speech quality dimensions differently, while none of them was highly correlated with the total impression. In Taal et al. [2014], a different approach was used. Among other factors, listeners subjectively evaluated the speech quality of methods that enhance intelligibility, using the AB-preference test. They heard two versions of the same sentence and were asked to choose the sentence that they preferred in terms of speech quality. Results showed that the proposed algorithm, which optimally redistributes the speech energy over time and frequency based on a perceptual distortion measure, apart from improving intelligibility was also able to preserve the speech quality.

2.3.2 Objective measures

The most widely used metric for objective assessment of speech quality is the perceptual evaluation of speech quality [PESQ; Rix et al., 2002] standardised by the ITU-T recommendation (P.862; ITU-T [2001], P.862.2; ITU-T [2005]). An overview of the PESQ measure is the following. First, both the reference and the degraded signals are aligned to a standard listening level. Then, they are filtered to model a standard telephone handset, and signals are aligned in time and processed via an auditory transform. Finally, two distortion parameters are extracted, aggregated in frequency and time, and mapped to a prediction of subjective mean opinion score.

This measure can handle different degradations, such as background noise, filtering [Beerends et al., 2002] and applying time–frequency-varying gain functions [Hu and Loizou, 2008] and source separation algorithms [Mowlaee et al., 2012]. To compute the predictions, two signals are compared, the reference/original and the degraded/modified speech signal, and results show high correlation under a variety of conditions. For noise-dependent speech algorithms, PESQ may affect the speech quality differently if the masker varies [Tang and Cooke, 2010].

More speech quality measures which require both signals (the reference/original and the degraded/modified speech signal) have been suggested. However, a limitation of such metrics is that for some conditions the reference signal is not available. Some of them are the computational model, which computes the auditory spectrum distance [Karjalainen, 1985], the perceptual speech quality measure (PSQM) which was used to predict the quality of speech codecs [Beerends and Stemerdink, 1994], and the measuring normalising block technique (MNB), which uses a perceptual transformation and a distance measure [Voran, 1999]. Another speech quality measure is the perceptual analysis/measurement system (PAMS), which was designed for evaluating the quality of telephone networks taking into account issues of previous models caused by linear filtering and variable delay packet-based transmission [Rix and Hollier, 2000]. Finally, the perceptual objective listening quality assessment [POLQA; Beerends et al., 2013], standardised by the ITU-T as Recommendation P.863 [ITU-T, 2011], has been developed to assess speech quality. This algorithm includes two parts: a temporal alignment part, with which a wide variety of complex distortions can be aligned—e.g. different delay variations in utterances or temporal stretching/compression of the degraded signal—and a perceptual model part, which calculates the internal representation of the reference and degraded signals. Finally, there is the hearing aid speech quality indices [HASQI; Falk et al., 2015] algorithm, which uses a comparison of the time-frequency envelope between the two signals and a cross-correlation measurement.

Additionally, objective measures have been suggested for predicting the quality of noisy speech enhanced by noise suppression algorithms [see Hu and Loizou, 2008]. For making predictions, the time-domain, frame-based segmental SNR [SegSNR; Hansen and Pellom, 1998] considers only the frames with segmental SNR in the range of -10 to 35 dB, the frequency-weighted segmental SNR [fwsegSNR; Tribolet et al., 1978], while the weighted spectral slope metric [WSS; Klatt, 1982] is based on an auditory model of 36 overlapping filters and finds a weighted difference between the spectral slopes in each band. Finally, linear predictive coding [Quackenbush et al., 1988], such as the log-likelihood ratio (LLR), Itakura-Saito distance measure (IS), and cepstrum distance measures (CEP). In Hu and Loizou [2008], the investigators reported that the segSNR measure was poorly correlated with the subjective quality ratings. They thus concluded that it is unsuitable for evaluating the performance of enhancement algorithms. In addition, the same authors reported that, amongst the tested objective measures, the predictions using PESQ had the highest correlation with the overall quality.

2.4 Listening effort

The attention or cognitive resources and processes required for comprehending speech are referred to collectively in the literature as listening effort. One definition for the listening effort is ‘the mental exertion required to attend to, and understand, an auditory message’ [McGarrigle et al., 2014]. A model for describing the processes involved while listening to speech has been suggested in Rönnberg [2003] and Rönnberg et al. [2013] (ease-of-language understanding).

Talkers may try to decrease the cognitive effort of the interlocutor. Some of the techniques that they use is to slow down their speaking rate [Picheny et al., 1986; Uther et al., 2007; Bradlow et al., 2003], use simpler vocabulary [Zampini et al., 2012], vary the fundamental frequency for giving emphasis to the significant information [Fernald and Mazzie, 1991], and enhance articulatory movements [Fitzpatrick et al., 2015]. However, exposure to conditions that require a listener to exert substantial effort and the engagement of additional cognitive resources may lead to long-term fatigue, social life withdrawal, or may have a negative impact on dual-task performance. Such conditions can be limitations of the source signal (e.g. degraded speech, accented speech), sound transmission interference (e.g. noise), or limitations of the receiver (e.g. hearing impaired or non-native listener).

2.4.1 Subjective measures

Subjective listening effort is usually assessed using questionnaires or rating scales. In the questionnaires, listeners have to respond to questions that refer to their everyday listening experience. One questionnaire is the Speech, Spatial and Qualities of Hearing Scale [SSQ; Gatehouse and Noble, 2004], which is designed to measure a range of hearing disabilities across several domains. An example question on listening effort is 'Do you have to put in a lot of effort to hear what is being said in conversation with others?' and the listener has to give an answer in the range from 0 (a lot of effort) to 10 (no effort). In Dawes et al. [2014] the SSQ was used to examine the changes in listening effort subsequent to acclimatisation to hearing aids.

On the other hand, rating scales are used to judge the effort of each testing condition. Usually, rating scale techniques accompany physiological techniques. In Luts et al. [2010], listeners rated the listening effort on a 13-point scale. There were 7 subcategories, ranging from 'no effort' (0) to 'extreme effort' (6) with 1 empty button in between. A similar test was used by Brons et al. [2013], with the differences that the scale was a 9-point rating scale and they used 5 labelled buttons instead of 7. Different rating scales have been used in studies where listeners were asked to answer a similar question to 'How much effort did it take to perceive the speech during the block?'. In Larsby et al. [2005] and Koelewijn et al. [2012], the subjective evaluation of listening effort for different noise backgrounds ranged from 0 (none at all/no effort) to 10 (extremely great/very effortful). Similar ratings have been used in Zekveld et al. [2010] for evaluating the listening effort for speech of different intelligibility levels. Other studies have used larger rating scales. In van Esch et al. [2013], in which an auditory profile test battery was evaluated, after each trial listeners had to estimate the exerted effort on a 100-point rating scale from 0 (no effort) to 100 (maximum effort). In Rudner et al. [2012], the relation was investigated between subjective effort ratings and other measures during aided speech recognition in noise. Listeners rated the listening effort using a visual analogue scale of 11.7 cm. The left-hand end represented 'no effort' (at 0 cm) and the right-hand end 'maximum possible effort' (at 11.7 cm), and the rating was computed as the distance from 0 cm to the point marked by the participant. Finally, the National Aeronautics and Space Administration Task Load Index (NASA-TLX) questionnaire [Hart and Staveland, 1988] was designed to elicit a participant's workload performance for a variety of dimensions using a visual-analogue rating scale, but it has also been used for assessing listening effort. In the NASA-TLX questionnaire, a question related to listening effort is 'How hard did you have to work to accomplish your level of performance?' Mackersie and Cones [2011] used this questionnaire to examine the listening effort under conditions of near-ceiling-level performance, while Peng and Wang [2019] used it

to test a wide range of realistic classroom acoustic conditions while varying talker accent and listener English proficiency.

2.4.2 Behavioural measures

The behavioural measures used for estimating listening effort are the single-task, dual-task, and recall paradigms [for reviews see McGarrigle et al., 2014; Gagné et al., 2017; Strand et al., 2018]. The total processing resources that a person has available to perform tasks are supposed to be limited in capacity and speed [Broadbent, 1958; Kahneman, 1973]. If the overall cognitive resources needed for a task are less than the available resources, then the task is performed optimally. However, if the available resources are less than those needed for the task, then the performance of one of the tasks will decrease. In the experiments they typically ask listeners to give priority to the primary task; thus, the performance of the secondary task is expected to deteriorate.

In single- and dual-task paradigms, listeners are instructed to optimise performance on the primary task (e.g. speech recognition). The single-task paradigm is when a primary task is performed alone. Listeners respond to stimuli, either verbally identifying the heard word/sentence [Gatehouse and Gordon, 1990], or by pressing a response button [Houben et al., 2013]. Response time has been interpreted as reflecting listening effort. In Houben et al. [2013], listeners had to identify three digits in varying levels of stationary noise; slower response times were recorded for the more challenging conditions. A speech recognition task is often used as the primary task [Hicks and Tharpe, 2002; Howard et al., 2010; Gosselin and Gagné, 2011; Govender and King, 2018b].

When a secondary task (e.g. arithmetic operation) is performed simultaneously with the primary task, the paradigm is called dual-task. Listening effort is computed as the loss of performance of the secondary task when it is performed under the dual-task condition, as compared to the same task performed alone (this is illustrated in Fig. 1 in Gagné et al. [2017]). The additional secondary task may include, for example, a tactile pattern-recognition task [Gosselin and Gagné, 2011], a visual recognition task (e.g. random presentations of a probe Hicks and Tharpe [2002], or a visual motor task Govender and King [2018b]). In Wu et al. [2016], a dual-task paradigm was used to assess listening effort at a wide range of SNRs. Reaction times, in line with subjective effort measures, showed less effort exertion for lower SNRs. In Sarampalis et al. [2009], the benefit of a digital noise reduction algorithm was tested using dual-task paradigms, either in quiet or in the presence of a 4-talker babble masker at various SNRs. Noise reduction was found to both reduce effort and benefit performance in simultaneous tasks.

Finally, the recall paradigm (memory task) is also based on the assumption that the more cognitive resources occupied, the poorer the recall performance (i.e. fewer cognitive resources are free for processing a new message). Some of the recall paradigms that have been used are the word recall task, the serial recall task, and the paired-associates recall task. For the word recall task, listeners were asked to respond to word lists or sentences while holding words in memory [Johnson et al., 2015; Sarampalis et al., 2009]. For the serial recall task, participants listened to lists of words. Word reproduction was stopped randomly and they had to recall the last three words presented [McCoy et al., 2005; Sommers and Phelps, 2016]. For the paired-associates recall task, listeners were asked to memorise lists of 5 word pairs. After some seconds they were presented with the first word from one of the 5 pairs and were asked to recall the second word of the pair [Murphy et al., 2000; Picou et al., 2011].

2.4.3 Physiological measures

Behavioural measures alone cannot systematically measure changes in effort. Several measures related to the activity of the central and autonomic nervous systems have been used to assess listening effort (McGarrigle et al. [2014]; Gagné et al. [2017]; Guijo and Cardoso [2018] review the existing physiological measures). The most widely used measure is pupillometry. A greater pupillary response (i.e. greater pupil diameter) is observed when the task requires more effort. Features typically used to estimate effort are the mean pupil dilation, peak pupil dilation or delay to reach the peak (latency). These features show an increasing trend with decreasing intelligibility [Zekveld et al., 2010]. The peak pupil dilation has been shown to reflect listening effort when tested in speech performance tasks involving sentences presented under conditions of informational or energetic masking [Zekveld et al., 2010; Zekveld and Kramer, 2014], with more effort observed for a competing talker masker than stationary or fluctuating maskers [Koelewijn et al., 2012, 2014]. Furthermore, it has been shown that pupil dilation is sensitive to speech quality. Differences between natural and synthetic speech have been observed [Govender and King, 2018a] and it is argued that, in quiet, pupil dilation reflects attention and engagement [Govender et al., 2019]. Pupillometric measures of effort have also been obtained as a function of syntactic complexity [Wendt et al., 2016], attention to location [Koelewijn et al., 2015] and spectral resolution [Winn et al., 2015]. It has been demonstrated that pupil size varies with regard to the noise type, i.e. a larger peak when listening to an informational masker compared to an energetic masker [Koelewijn et al., 2012]; SNR, i.e. forming an inverted U-shaped curve, with the peak pupil dilation at intermediate intelligibility levels and lower values when comprehension is easier or harder (at very adverse SNR levels listeners tend to give up the task) [Zekveld et al., 2010; Zekveld and Kramer, 2014]; speech location, i.e. location uncertainty increased the pupil dilation response [Koelewijn et al., 2015]; syntactic complexity i.e. pupil dilations increased with syntactic complexity [Wendt et al., 2016]; and spectral resolution i.e. pupil dilation greater with degradation in spectral resolution [Winn et al., 2015].

Other measures that have been used for measuring listening effort are functional magnetic resonance imaging [Wild et al., 2012], electroencephalography [Obleser et al., 2012], heart rate variability [Seeman and Sims, 2015], skin conductance [Mackersie and Cones, 2011; Seeman and Sims, 2015] and electromyographic activity [Mackersie and Cones, 2011]. A speech understanding task that requires more effort to complete leads to higher values for those measures: i.e. increased activity in the left inferior frontal gyrus, alpha power, heart rate variability, skin conductance and electromyographic activity. For instance, Mackersie and Cones [2011] obtained psychophysiological recordings (heart rate, skin conductance, skin temperature and electromyographic activity) during speech perception tasks with intelligibility close to ceiling, but with varying task demands involving digit presentation to one or both ears. Higher levels of mean skin conductance and electromyographic activity were observed when task demand increased.

2.5 Listener preferences

When the speech quality is considered as unidimensional it can be interpreted as the listener's preference. The listener's preference is the overall listener's experience when listening to a speech signal, which can be affected by aspects of speech such as intelligibility, listening effort, pleasantness, naturalness and loudness. These factors can vary with the listening environment

[Kean et al., 2015; Walton et al., 2016], hearing impairment [Buyens et al., 2014; Shirley et al., 2017], as well as having an individual component [Walton et al., 2016].

2.5.1 Subjective measures

A commonly used method for exploring listener preferences is by asking the listener to make judgements about the presented stimuli using a rating scale. The effect of noise on speech rate has been studied in Adams and Moore [2009], and for listening conditions encountered in real life (i.e. non-degraded, reverberation, bandpass filtered, and low-pass filtered conditions) in Moore et al. [2007]. Participants judged the rates of the presented target sentence using an equal-interval 5-step scale from too slow to too fast i.e. 'too slow', 'slow, but ok', 'preferred', 'fast, but ok', and 'too fast'. They indicated their choices by clicking on an icon on the computer monitor using the mouse.

Many studies have used the paired-comparison paradigm to compare listening preferences. Listener preferences for different speaking styles have been determined, e.g. for oral reading and spontaneous speaking tasks [Lass and Prater, 1973] and for prose speech to children [Lass and Fultz, 1976; Leeper and Thomas, 1978]. The listeners' task was to choose which of the two speech rates in the pair they preferred to listen to. Brons et al. [2013] tested different hearing aid noise-reduction systems to determine whether noise-reduction systems differ perceptually, and which factors underlie the overall preference of individual listeners. The investigators used a paired-comparison rating to measure overall preference. Listeners were asked which of the fragments they would prefer for prolonged listening. There were 7 possible answers, ranging from 'A is much more natural/much less annoying/much better' to 'B is much more natural/much less annoying/much better'. The 7 choice categories were derived from the comparison category rating method described in ITU-T P.800 [ITU-T, 1996]. Listeners had the option to indicate no difference between A and B, and were allowed to listen to the fragments as often as they liked before answering the question. Boymans and Dreschler [2000] collected the subjective preferences for 4 different hearing aid settings. Listeners stated which program they preferred if they had to listen to speech in 'this condition' through the whole day.

Other studies have used a combination of the two aforementioned techniques, i.e. paired-comparison paradigm and rating-scale. The listener first indicates which of the two options is preferable to listen to (pair-comparison paradigm), and then answers the question of how much this option is preferred (rating-scale). Using this combination, previous studies have determined listener preferences for noise-reduction algorithms or settings [Ricketts and Hornsby, 2005; Luts et al., 2010]. Ricketts and Hornsby [2005] investigated the effect of digital noise reduction processing on aided speech recognition and sound quality measures. Listeners had to tell the investigator to switch to 'one' or 'two' as often as they liked and then report which setting they preferred. Once a setting was preferred, listeners had to specify the 'strength of preference' on a scale of 1 to 10, how strongly they preferred that particular setting over the other. The preference scale was provided as a visual marker to aid their decision on how strongly they preferred one setting over another. A preference of 1 was indicated as 'no or very little preference', and 10 indicated a 'very strong preference'. In Luts et al. [2010], listeners had to assess the algorithms' performance. Each algorithm was compared to the unprocessed condition. Subjects could listen to the algorithms for as long as needed and could toggle between the algorithms as often as they wanted. After indicating a preference, the subject had to rate how much better the preferred algorithm was compared to the other one. This rating can be

interpreted as the confidence of the subjective preference judgments. The outcome of the test is the amount of preference for an algorithm over the unprocessed condition. The preference score varied between 'very much worse' scored as -5 and 'very much better' scored as $+5$. Listeners did not have the option of equal preference (0).

2.5.2 Objective measures

It is a common procedure for the experimenters to choose *a priori* fixed conditions for testing the speech stimuli. An alternative approach is the listener-driven preferences technique, which allows the listeners to modify speech properties in real time using adjustment tools. This technique allows the listeners to tune the target speech stimuli to find the preferred value. Some studies have used Lexicon Varispeech [Lee, 1972], a tool for demonstrating real-time modification of speech properties [Lee, 1972]. In Riensche et al. [1979], the effect of age and sex on the preferred listening rate of speech was investigated. Listeners were presented with a reading of a prose passage and were allowed to adjust a Varispeech I time compressor/expander to yield their preferred listening rate. Wingfield and Ducharme [1999] allowed younger and older adults to adjust the speech rate of time-compressed and time-expanded speech passages of low and high predictability. Stimuli were presented through a Lexicon Varispeech II compressor/expander and listeners could control the rate using a knob to speed up the speech or slow it down. Participants were instructed to adjust the rate to the point where they felt that the passage could be understood and accurately recalled. There was no time constraint; however, all the listeners were able to find the desired rate by approximately halfway through the passage. Audio loudness preferences in realistic environments were investigated by Kean et al. [2015] via a USB knob controller. Turning the knob clockwise raised the volume and turning it counter-clockwise lowered it. Listeners were instructed to adjust the volume for the most pleasing playback effect. They could modify the speech at any time during the playbacks and leave the level in place if they were satisfied with the sound. The influence of environmental noise on audio preferences was also tested by Walton et al. [2016] in a study that simulated mobile audio listening. Listeners were able to adjust the background–foreground balance and the overall level through a virtual interface. In order to avoid any influence on the listeners' judgments, no visual feedback for the adjustments was provided.

Virtual adjustment devices have also been used to allow participants to actively control speech rate [Novak III et al., 2014; Novak III and Kenyon, 2018]. Listeners could control the rate of audio playback in real time with an on-screen slide bar. Listeners had to adjust the speech until they found the value that was best for understanding the target speech. They did not have any guidance as to what settings might be 'best'. Torcoli et al. [2017] introduced the Adjustment/Satisfaction Test, a user-adjustable system complemented by a user satisfaction assessment, applied to the evaluation of dialogue enhancement. Listeners controlled the relative level of speech with a knob and scored their satisfaction level using a rating scale. A virtual knob was used in Zhang and Shen [2019] to allow listeners to modify a local signal-to-noise ratio criterion for retaining or removing time-frequency regions of speech. Listeners could increase the local criterion value by turning the knob clockwise and decrease it by turning the knob counter-clockwise. Instructions were to 'adjust the knob so that the speech would be the clearest and easy to listen to for a long time'.

2.6 Summary

Investigating the mechanisms involved when a talker naturally modifies speech under different conditions so as to facilitate the listener, and the processes involved when a listener perceives speech, may provide insights into the development of speech enhancement algorithms. Such algorithms find applications in public address systems, hearing aids and telephony. Since the target audience of such algorithms consists of humans, it is important that the speech modifications satisfy them. Therefore, the most accurate evaluation of an algorithm is both through word task performance and evaluation of supra-intelligibility factors. The remainder of this thesis explores intelligibility and supra-intelligibility aspects of speech for different speech types and distinct speech characteristics.

Chapter 3

Native and non-native listening effort for different speech types

3.1 Introduction

[1]Listening to speech is not always performed in ideal conditions. An additional obstacle faces a listener who has to communicate in a second language (L2). Many people move to a foreign country to study or work, and collaboration between people who are native in different languages is widespread. The task becomes more demanding when the listener does not have the advantage of seeing the talker and consequently has to rely only on audio cues. Thus, it is of interest to investigate how different speech types affect non-native listeners. Their performance in challenging conditions has already been explored [Meador et al., 2000; Weiss and Dempsey, 2008] (for a review see Garcia Lecumberri et al. [2010]), as has the effect of modified speech styles on non-native intelligibility, with Cooke and Garcia Lecumberri [2016] concluding that listening in an L2 is more detrimental although native and non-native listeners display a similar pattern across speech types. However, far less emphasis has been placed on investigating the effort required to understand distinct speech types. Exposure to conditions that require a listener to exert substantial effort and the engagement of additional cognitive resources may lead to long term fatigue. Such conditions might be degraded source signals, sound transmission interference or limitations of the receiver (see review in Mattys et al. [2012]). This chapter examines listening effort for distinct speech types under conditions of additive noise for native and non-native listeners.

Previous studies have explored the impact of different speech types on listening effort. Borghini and Hazan [2020] examined the impact of conversational and clear speaking styles on listening effort in the presence of 8-talker babble noise using an SRT procedure. Apart from the expected finding that listeners tolerated a lower SNR for clear compared to plain style sentences, both the mean and peak pupil dilations were greater for plain speech, suggestive of a greater listening effort. Koch and Janse [2016] conducted an eye-tracking experiment to explore the effect of speech rate on spoken word recognition, using conversational materials with a natural variation in speech rate. While listeners exhibited longer response times for fast speech, there was no speech rate effect on the pupil response.

[1]Portions of the work described in this chapter were published as a paper in Interspeech 2018 proceedings [Simantiraki et al., 2018].

The impact of synthetic speech on listening effort has also been investigated, revealing that pupil dilation is sensitive to speech quality. Three studies by Govender and colleagues examined the differences between natural speech and four speech synthesis approaches of differing sophistication, namely Hybrid, Unit Selection, Hidden Markov Model (HMM) and Low-Quality HMM, all drawn from the Blizzard Challenge 2011 [King and Karaiskos, 2011]. Govender and King [2018b] tested synthetic speech in noise-free conditions using a dual-task paradigm, finding that synthetic speech led to slower reaction times (suggesting a higher cognitive load) as speech quality decreased. Again using stimuli in quiet, Govender and King [2018a] found greater pupil dilation for synthetic speech compared to its naturally-produced counterpart. Masking noise also led to an increase in pupil dilation for synthetic speech [Govender et al., 2019].

These findings collectively indicate that forms of speech that differ from canonical 'plain' speech have the potential to affect a listener's experience, either by reducing effort in the case of clear speech, or increasing it for synthetic speech. One hypothesis is that listening effort is influenced by naturalness. Synthetic speech, particularly that produced by less sophisticated approaches, is clearly unnatural, and the finding that effort reduces for more recent state-of-the-art synthesis techniques which are mainly distinguished by their degree of naturalness supports a potential inverse relationship between naturalness and effort. The primary goal of the current study is to further explore this relationship by examining pupil responses to both plain and synthetic speech as well as to two additional speech forms known to improve intelligibility in masked conditions but which differ in naturalness, namely Lombard speech and algorithmically-enhanced speech. Lombard speech refers to speech that results when a talker is exposed to sufficiently intense noise while speaking. Many studies have shown Lombard speech to be substantially more intelligible than plain speech when presented at the same SNR [Pittman and Wiley, 2001; Marcoux et al., 2022]. Of the many forms of algorithmically-modified speech, one of the most successful in enhancing intelligibility is SSDRC [Zorila et al., 2012], an approach which involves both spectral modification and dynamic range compression. SSDRC produced the largest gains in an international evaluation of modification techniques [Cooke et al., 2013b]. We hypothesise that in masked conditions in which Lombard speech and SSDRC are at ceiling levels of intelligibility, if naturalness is a key factor in listening effort we will see smaller peak pupil dilation for Lombard speech than for SSDRC.

Research on listening effort mainly deals with native normal-hearing or hearing-impaired listeners. Far less investigated is the listening effort exerted by non-native listeners. Speech perception by non-native listeners is one input factor that contributes to increased listening effort [Framework for Effortful Listening, FUEL; Pichora-Fuller et al., 2016]. Previous studies have shown that non-native listeners have a greater pupil response compared to native listeners when perceiving words [Schmidtke, 2014] or sentences in babble noise with equated intelligibility for the two groups [Borghini and Hazan, 2018]. In Borghini and Hazan [2020], listening effort for clear and plain speech and semantic plausibility in the presence of babble noise were studied while the SNR was adapted to obtain an intelligibility score of 50%. Their results showed that both native and non-native listeners exert less effort when listening to the naturally-enhanced speech type. In Peng and Wang [2019], the listening effort of native and non-native listeners was evaluated objectively using a dual-task of speech comprehension and an adaptive pursuit rotor (i.e. try to track a small disc on a turntable) and subjectively using the NASA task load index questionnaire [Hart and Staveland, 1988]. They tested several combinations of background noise level and reverberation time and results indicated that the reported effort of non-native listeners

was higher than that of native listeners when comprehending speech in adverse conditions. It is currently an open question as to whether listening effort follows a similar pattern i.e. do non-native listeners also suffer a disproportionate increase in effort when processing challenging forms of speech when compared to the effort experience by native listeners? The current study addresses this question by comparing the patterns of intelligibility in masked conditions with those of effort as revealed by pupil responses for both native and non-native listeners.

Two experiments explored the effect of four distinct speech types (plain, synthetic, Lombard, SSDRC) on subjectively-reported effort, pupil responses, and intelligibility. Both experiments asked listeners to process English sentences in three levels of speech-shaped noise. Expt. I (sec. 3.2) involved a cohort of native English listeners, while in Expt. II (sec. 3.3) a group of Spanish listeners processed the same materials at a more favourable set of SNRs. In this chapter, growth curve analysis (GCA) was used for analysing the pupil data since the entire time-course of the data is taken into account resulting in more meaningful information [Mirman, 2014].

The research questions for this chapter are: does listening effort vary for different speech types; does listening effort exerted by native and non-native listeners pattern differently across listening conditions?

3.2 Experiment I: Impact of different speech types on listening effort for native listeners

3.2.1 Methods

Participants

Participants (N=26, 6 males) were young normal-hearing native British English speakers (age range: $18-24$, mean=20.5, $S.D$ 1.8). Participants were requested not to wear glasses and eye makeup. All had hearing levels better than 25 dB in both ears as determined by pure-tone audiometric screening in the range $125-8000$ Hz. Listeners were paid on completion of the experiment. Technical problems during recording led to the exclusion of data from two participants.

Speech and masker materials

The Harvard sentence lists [Rothauser et al., 1969] provided the basis for the four distinct speech styles tested in the current study. Harvard sentences typically contain 7-9 words, of which 5 are preselected as keywords for scoring purposes. The speech signals used in the current study constitute a subset of the speech material used in for an international challenge in intelligibility-enhancing speech modifications [Cooke et al., 2013a].

- In the *plain* condition, sentences were produced in quiet conditions by a British English male talker who was asked to speak normally.

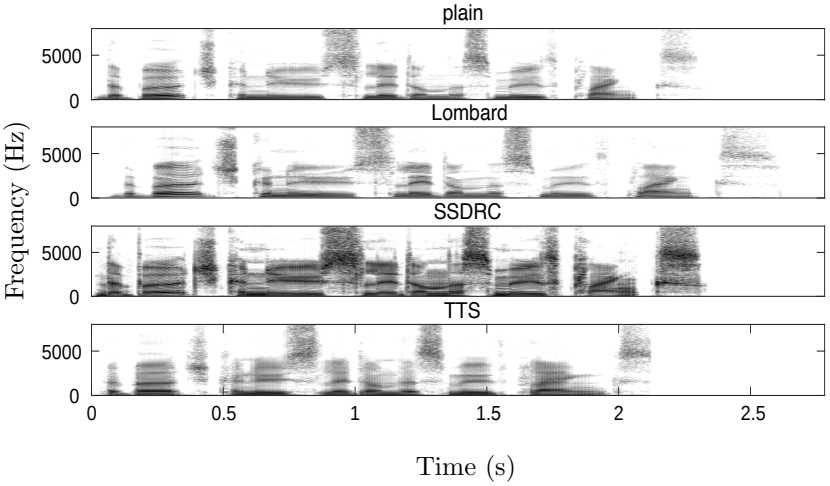
- The same talker also produced sentences in the presence of a speech-shaped noise (SSN) masker, resulting in the *Lombard* condition. Lombard speech is a natural form of modified speech with clear differences from speech produced in quiet. For example, Lombard

speech normally results in a flatter spectral tilt, increased fundamental frequency, and some segments exhibit longer durations [Summers et al., 1988].

- A *modified speech* condition was created by applying the Spectral Shaping and Dynamic Range Compression (SSDRC) method [Zorila et al., 2012] to the plain speech sentences. The SSDRC algorithm incorporates ideas from both Lombard and clear speech styles, and has been shown to produce significant intelligibility gains [Cooke et al., 2013b].

- Finally, a *synthetic speech* style was generated using hidden Markov model text-to-speech (TTS) synthesis. The system employed [Yamagishi et al., 2009] was capable of adapting to individual speakers, so in addition to the (orthographic) sentence text, the TTS system was also provided with additional speech material from the talker who produced the *plain* and *Lombard* sentences.

Figure 3.1 shows example spectrograms for the same sentence in each of the four types and provides values for duration, spectral tilt and mean fundamental frequency.



	dur.	**tilt**	f_0
plain	2.05	-5.26	105
Lombard	2.28	-3.69	136
SSDRC	2.05	-1.96	106
TTS	1.93	-5.30	106

Figure 3.1: *Spectrograms of the phrase 'The birch canoe slid on the smooth planks' for each of the speech types tested in the current study, along with corresponding values for duration (dur.) in seconds, spectral tilt (tilt) in dB/octave, and mean fundamental frequency (f_0) in Hz.*

Experimental stimuli were created by mixing sentences in each of the four styles with a speech-shaped masking noise at each of three SNRs: -1, -3 and -5 dB, resulting in 12 condition blocks. These SNRs were chosen on the basis of recommendations in a study by Ohlenforst et al. [2017] to avoid values that are too high (low noise) and likely to be effortless for participants, or too low (high noise), potentially leading participants to expend less effort due to the perceived level of difficult of the task. In pilot tests, the three SNRs chosen produced intelligibility levels both near to and below ceiling.

Procedure

Maskers started two seconds prior to the onset of each sentence, and stopped three seconds after sentence offset. Pupil data from the 1-second interval immediately preceding the onset of the sentence was used for calibration (see sec. 3.2.1). Speech-plus-noise mixtures were created by rescaling the speech signal to achieve the desired SNR in the region where it overlapped

with the masker. The resulting mixtures were normalised to have the same root mean square level, and 20 ms half-Hamming ramps applied to the start and end to reduce onset and offset transients.

Listeners heard 12 blocks of stimuli, one block for each combination of speech type and SNR. Each block consisted of 15 target sentences that were used for scoring, preceded by 5 familiarisation sentences. None of the 180 (15 x 12) sentences heard by any given listener were repeated. Block order was balanced across listeners using a Latin square design, and sentence order within blocks was randomised. Before starting the experiment, participants were able to adjust the volume to a comfortable listening level.

The experiment took place in a sound proof studio at the University of Edinburgh. Pupil data was collected using the remote EyeLink 1000 eye-tracker with sampling frequency 500 Hz and the pupil size was measured in terms of pupil area (number of black pixels) while participants listened to sentences through Sennheiser HD-380 pro headphones.

Participants were seated in front of a computer screen with a white background and a black cross in the middle (Fig. 3.2). Participants were instructed to look at the black cross while listening to the stimuli. At the end of the trial the cross became red and participants had to repeat verbally the words they had heard (Fig. 3.3).

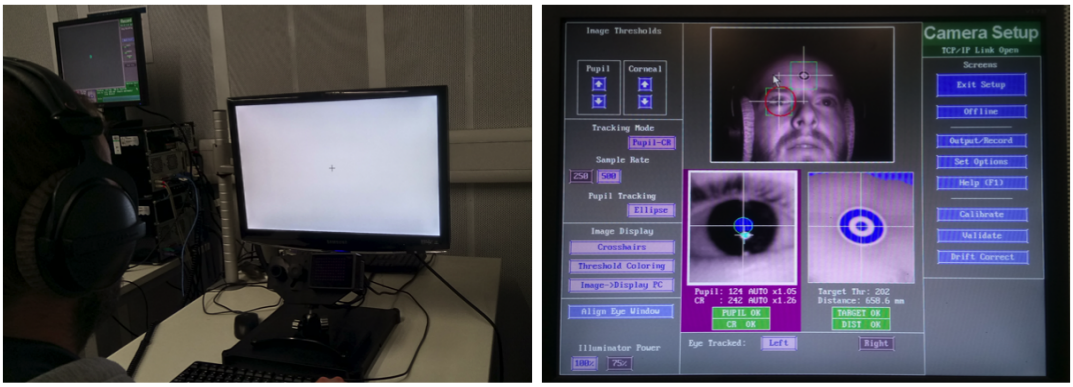


Figure 3.2: *Experimental set up. The left image shows a listener during the task while the right shows the experimenter's monitor.*

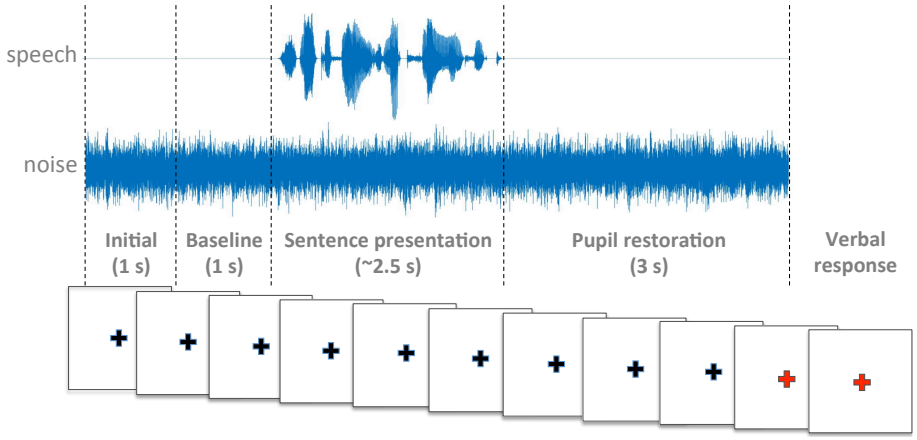


Figure 3.3: *Schematic representation of the experimental procedure.*

On completion of each block, participants answered the question 'How much effort did it

take to listen and understand the sentences in this block?' using a numeric rating scale from 0 (no effort) to 10 (very effortful). The experiment was split into two parts of approximately 30 minutes each, with an intervening 5-minute break.

Calibration

Similar practices to those suggested in Winn et al. [2018] were used in the processing of raw pupillary responses and in discarding trials. Pupil data from the left eye was used. Pupil area data was first downsampled to 50 Hz and converted to pupil diameter, and the following signal-cleaning procedure (designed, for example, to detect blinks) was applied. For each trial, cases where the pupil size was more than two standard deviations lower than the overall mean pupil size were considered as missing values. Trials with more than 15% missing values, as well as participants with fewer than 80%valid trials, were excluded from the analyses. For valid trials, any missing values were linearly-interpolated using data in a window that covered the interval from 5 samples prior to the missing value, to 8 samples after the missing value. Following signal-cleaning, pupil traces were calibrated following Wagner et al. [2015]:

$$ERPD = \frac{observation - baseline}{baseline} * 100 \tag{3.1}$$

where ERPD is the event-related pupil dilation, *observation* is the uncalibrated pupil diameter and *baseline* is the mean pupil diameter during the one second interval preceding the onset of the speech. Finally, pupil data were smoothed using a 5-point moving average filter. Figure 3.4 provides an example of the uncalibrated pupil area (upper plot) and calibrated pupil diameter (lower plot). A visual inspection for artefacts led to the exclusion of around 11% of the pupil data, based on a criterion on removing blocks with fewer than two-thirds of trials correct.

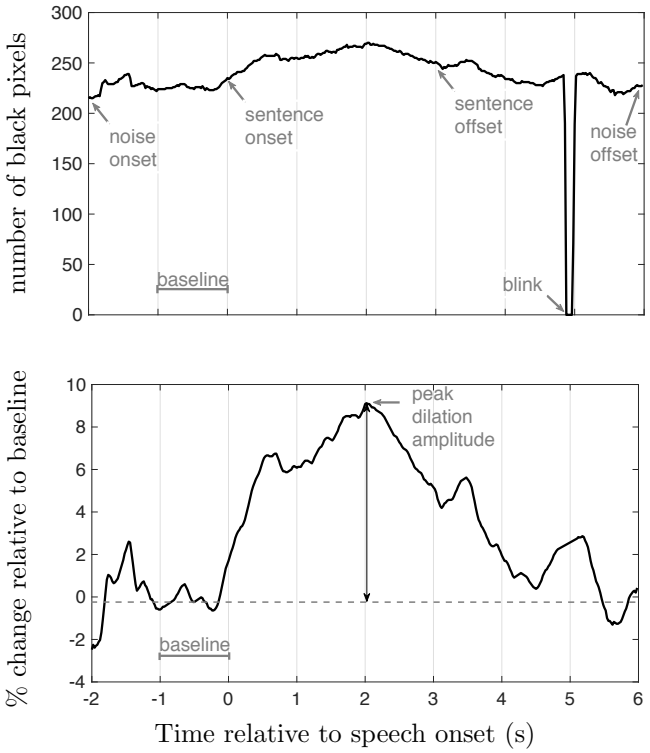


Figure 3.4: *Pupil size variation during a single trial. Upper: uncalibrated pupil area; lower: calibrated pupil diameter. Times are relative to sentence onset at 0 s.*

Statistical analyses

Statistical analyses were carried out in *R* (R Core Team [2021] version 3.3.3). The time-course of pupil dilation for each SNR was modelled using growth curve analysis [GCA; Mirman, 2014]. A third-order polynomial was used for modelling the data within the time window of 0 s (speech onset) until 4.5 s after speech onset while the peak pupil dilation was included. An order 3 polynomial was chosen after observing the average pupil dilation responses of all participants and conditions. The time terms can be interpreted as follows: the intercept as the overall mean pupil dilation, the linear term as the overall rate of pupil dilation, the quadratic term as the shape of peak, and the cubic term as the falling slope of the curve. Model selection started with the complete model, which is the time-course of the pupil dilation data as a third-order orthogonal polynomial. It included as fixed effects both speech type (*speech_type*) and intelligibility for the three orthogonal terms and as random factors the intercept and the three orthogonal terms per participant. Intelligibility scores were computed for each sentence with 5 -total keywords included in a sentence- to be the maximum score. The model fit was evaluated using model comparisons with anova. Improvements in model fit were evaluated using the log-likelihood ratio, which is distributed as χ^2 with degrees of freedom equal to the number of parameters added. Statistical significance (*p-values*) for individual parameter estimates was assessed using the normal approximation (i.e. treating the *t-value* as a *z-value*). The intelligibility factor did not improve the model and thus it was removed. The *lme4* package (Bates et al. [2015] version 1.1-15) was used for fitting a linear mixed-effect model to the data.

To model the relationship of speech conditions (i.e. SNR and speech type) with both intelligibility and subjective listening effort ratings, linear mixed effects analysis was used. As fixed effects, the SNR and speech type were added into the model, with random intercepts for participants. *P-values* were obtained by likelihood ratio tests of the full model (interaction between the fixed effect terms), the model with only the SNR as fixed factor, and the model without the interaction term. Intelligibility score percentages were converted to rationalised arcsine units in order to make them more suitable for statistical analysis [Studebaker, 1985].

Post-hoc comparisons used least-squares means [*emmeans* package; Lenth, 2021], with Tukey adjustment for multiple comparisons. Additionally, repeated-measures correlation between intelligibility scores and subjective listening effort ratings were performed via the *rmcorr* package [Bakdash and Marusich, 2017].

Trials for which listeners did not perceive any word correctly were excluded from the analysis. The sentence plus noise combination may was too hard, which can lead the listeners giving up or to exert a great amount of effort without succeeding in the task. Pupil dilation is influenced by speech intelligibility and when intelligibility is close to 0 the pupil dilation is small [Zekveld and Kramer, 2014; Ohlenforst et al., 2017]. Thus, to avoid such pupil size behaviours, these trials were excluded. For the −1 dB SNR condition, 9.7% of the trials were excluded, for the −3 dB SNR 15.6%, and for the −5 dB SNR 30.3% (for more details see Table A.1 in appendix).

3.2.2 Results

Pupil dilation

Figure 3.5 depicts the ERPD of the raw data averaged across participants for each speech style and SNR. For the most favourable SNR, TTS shows the greatest change in pupil dilation over the baseline, followed by plain speech. A similar trend is seen at the intermediate SNR. For the

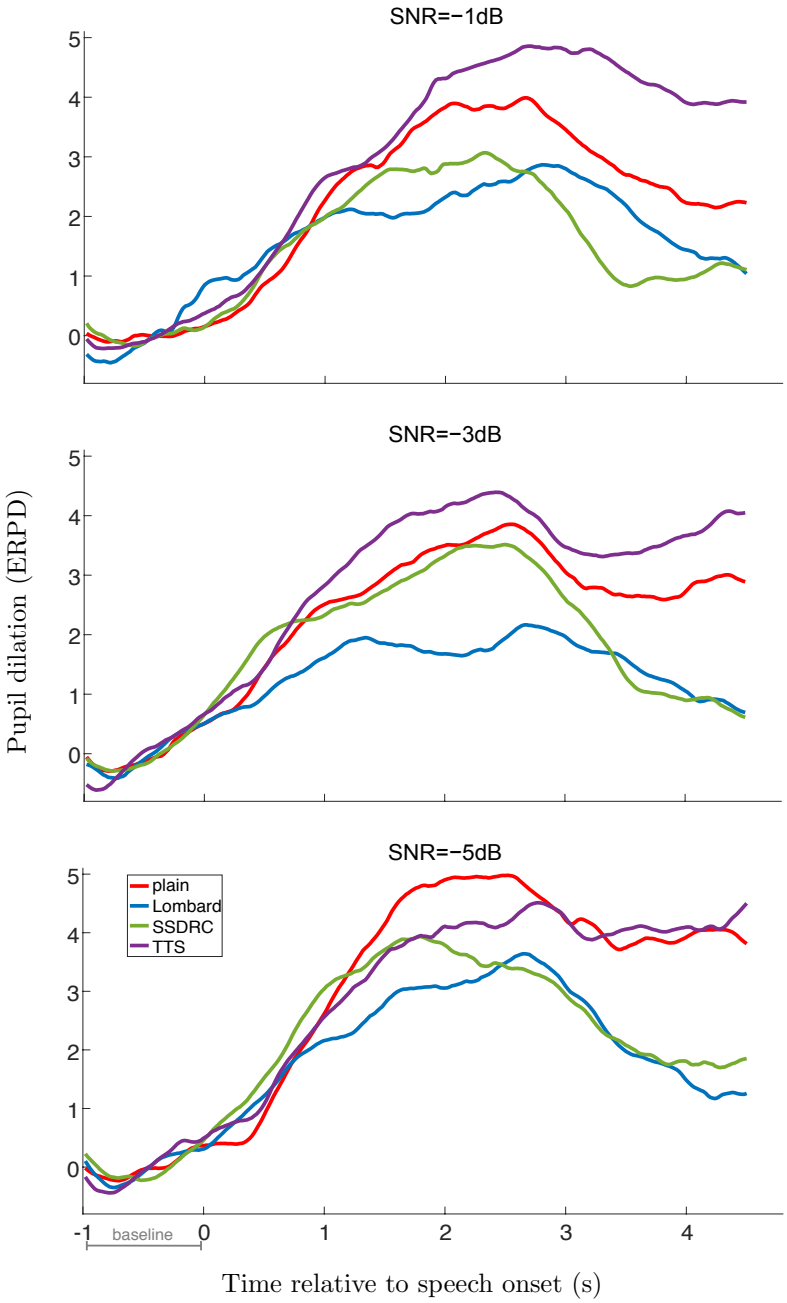


Figure 3.5: *Mean pupil size increase over baseline as a function of speech type. Noise starts 1 s before the baseline onset as shows Fig. 3.4.*

adverse SNR plain speech exhibits the largest relative increase in pupil size. Lombard speech generally results in the lowest ERPD at each noise level.

The best-fitted model was the following (Figs. 3.6 - 3.8 show the best-fitted model on the pupil data for each speech type and SNR).

$$ERPD \sim (time1 + time2 + time3) * speech\_type + (time1 + time2 + time3|participant) \quad (3.2)$$

with $time1$, $time2$, $time3$ being the 3 orthogonal terms, $speech\_type$ the 4 tested speech types, and $participant$ the participant id. Table 3.2 shows the estimates of each polynomial term and speech type for the different SNRs and Table 3.3 the interpretation of the GCA results as a

function of the polynomial term and SNR.

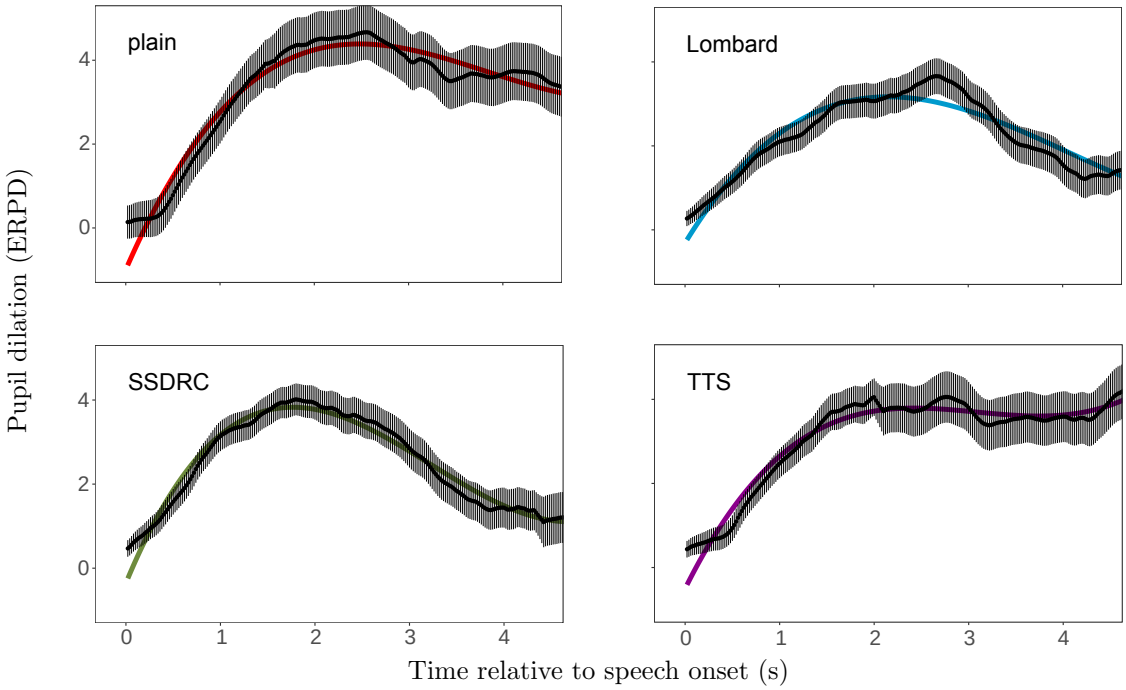


Figure 3.6: *Mean pupil size over time (black dots) with grey error bars to denote the* ± 1 *SE. The solid line shows the fitted model for the* -5 *dB SNR.*

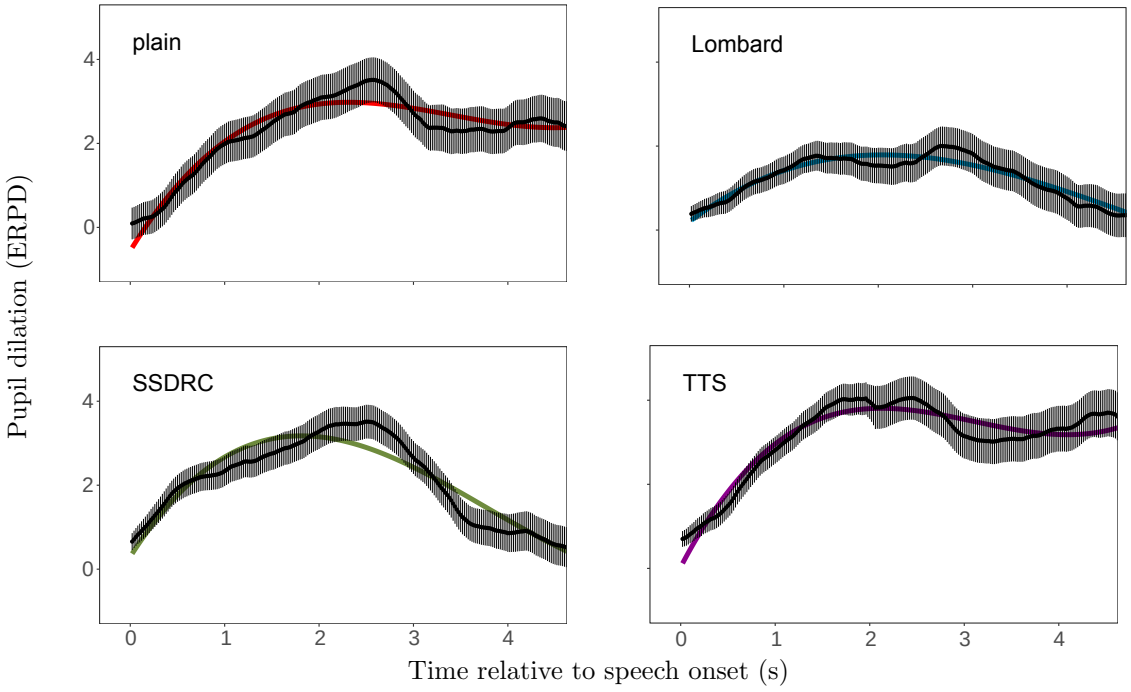


Figure 3.7: *As Fig. 3.6 but for the* -3 *dB SNR.*

Intelligibility scores

The mean percentage of correct words repeated by participants for the different speech types and SNRs is shown in Fig. 3.9 (right panel). A ceiling effect can be observed for the Lombard

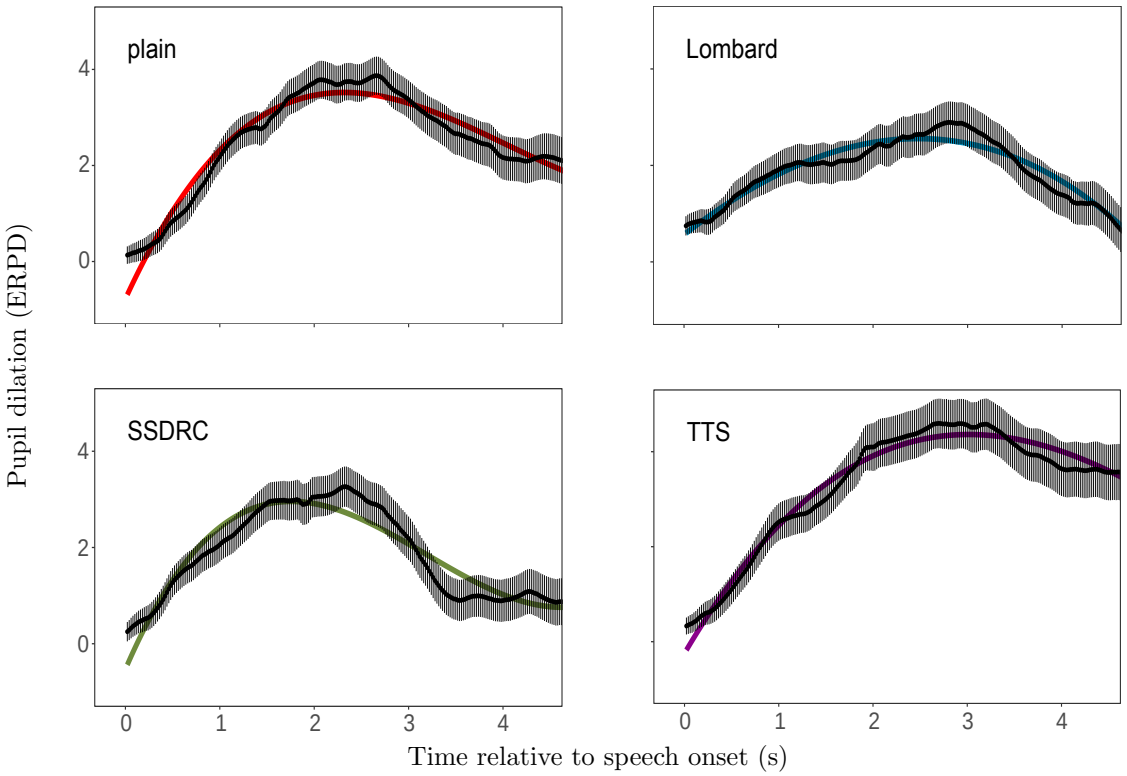


Figure 3.8: *As Fig. 3.6 but for the −1 dB SNR.*

Speech type	**-1**	**-3**	**-5**
Intercept:plain	2.45 (0.30)	2.35 (0.36)	3.03 (0.41)
Intercept:Lombard	-0.97 (0.04)*	-1.20 (0.04)*	-0.91 (0.04)*
Intercept:SSDRC	-0.84 (0.04)*	-0.60 (0.04)*	-0.83 (0.05)*
Intercept:TTS	0.81 (0.04)*	0.80 (0.04)*	-0.22 (0.04)*
time1:plain	4.74 (2.63)	6.95 (2.67)	10.16 (3.57)
time1:Lombard	-6.47 (0.61)*	-9.67 (0.63)*	-10.41 (0.68)*
time1:SSDRC	-9.20 (0.59)*	-14.36 (0.66)*	-14.17 (0.70)*
time1:TTS	8.06 (0.59)	0.11 (0.62)	0.56 (0.70)
time2:plain	-14.05 (2.61)	-8.82 (2.01)	-13.53 (2.76)
time2:Lombard	5.35 (0.61)*	2.11 (0.63)*	0.98 (0.68)
time2:SSDRC	3.82 (0.59)*	-3.11 (0.66)*	0.71 (0.70)
time2:TTS	0.83 (0.59)	0.12 (0.62)	8.09 (0.70)*
time3:plain	4.56 (1.53)	4.03 (1.26)	4.79 (1.86)
time3:Lombard	-6.43 (0.60)*	-2.33 (0.63)*	-1.17 (0.68)
time3:SSDRC	1.79 (0.59)*	0.18 (0.66)	1.74 (0.70)*
time3:TTS	-3.02 (0.59)*	2.22 (0.62)*	1.66 (0.70)*

Table 3.2: *Summary of estimates of intercept and orthogonal polynomial time terms (time*1*, time*2*, time*3*) with plain speech as baseline for the different SNRs. The standard error is shown in parentheses and the asterisk indicates those conditions significantly different from baseline.*

style for the most favourable SNR, and for SSDRC for all SNRs. As expected, intelligibility decreased with increasing noise level.

Statistical analysis also verified that as noise level decreased, intelligibility scores increased [$p < 0.05$] except for SSDRC which was not statistically different for any of the SNRs, and for Lombard speech which was not statistically different between -1 and -3 dB SNR. Speech type comparisons revealed that the intelligibility differed significantly at each of the 3 SNRs

Term	Interpretation	Order	-1	-3	-5
Intercept	overall mean pupil dilation	greater to lower	TTS = plain ≠ SSDRC = Lombard	TTS ≠ plain ≠ SSDRC ≠ Lombard	plain ≠ TTS ≠ SSDRC ≠ Lombard
Linear	overall pupil dilation rate	steeper to flatter	SSDRC ≠ Lombard ≠ plain ≠ TTS	SSDRC ≠ Lombard ≠ plain = TTS	SSDRC ≠ Lombard ≠ plain = TTS
Quadratic	shape of peak (height and width of the curve)	sharper to flatter	TTS = plain ≠ SSDRC ≠ Lombard	SSDRC ≠ plain = TTS ≠ Lombard	plain = SSDRC = Lombard ≠ TTS
Cubic	falling slope	faster to slower	Lombard ≠ TTS ≠ plain ≠ SSDRC	Lombard ≠ plain = SSDRC ≠ TTS	Lombard = plain ≠ TTS = SSDRC

Table 3.3: *Interpretation of each polynomial term and results as a function of SNR. Results are ordered based on the 3^{rd} column. The symbol '=' signifies that the speech types were not statistically significant different and '≠' the opposite.*

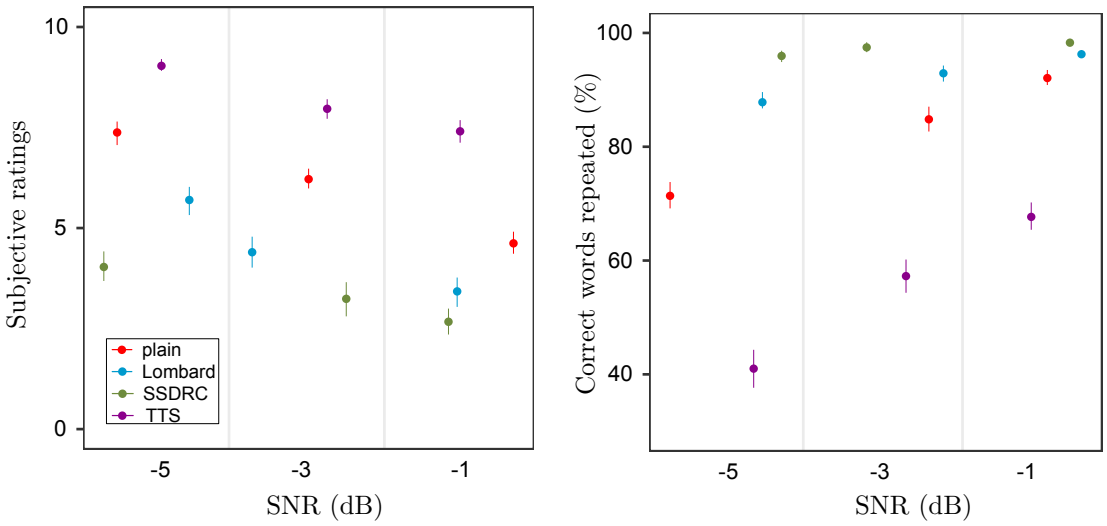


Figure 3.9: *Left plot: mean subjective listening effort ratings from 0 (no effort) to 10 (very effortful). Right plot: mean intelligibility scores. Error bars denote ±1 standard error.*

[$p < 0.01$] (only for −1 dB SNR, scores were not statistically different between Lombard and SSDRC and a marginal difference [$p = 0.07$] between Lombard and plain was obtained). The ranking was the same for all SNRs with performance order from highest to lowest to be SSDRC, Lombard, plain, and finally TTS.

Subjective listening effort ratings

Mean subjective ratings for the different speech types across the 3 SNRs are depicted in Fig. 3.9 (left panel), revealing an unambiguous ranking of speech types. Synthetic speech was considered the most effortful style and SSDRC the least. Subjective effort of all speech types increased with increasing SNR. As for the intelligibility scores, subjective listening effort ratings showed a clear ranking of effort across speech types that is the inverse of the intelligibility scores.

Statistical analysis showed that as noise level increased, each speech type was rated as significantly more effortful [$p < 0.05$] except for SSDRC which was rated similarly for all conditions (apart from −1 and −5 dB SNR [$p < 0.01$]), while ratings for Lombard and TTS were similar at −1 and −3 dB SNR. Also marginal differences were obtained between −3 and −5 dB SNR for plain [$p = 0.06$] and TTS [$p = 0.09$]. Independent of the SNR, the order of the subjective ratings was from the least to the most effortful: SSDRC, Lombard, plain, and TTS with scores for Lombard and SSDRC not to differ at −1 dB SNR.

Additionally, correlation tests verified the negative correlation of intelligibility and subjective ratings [$r = -0.87$].

3.2.3 Interim discussion

This experiment explored the effort that native English listeners exert when listening to different speech types. Participants listened to sentences in the presence of a masker at 3 noise levels. Effort was estimated both by participants' own ratings and by the physiological measure of pupil size change. At all SNRs listeners found synthetic speech to be both the least intelligible and subjectively the most effortful to process, while the converse was the case for algorithmically-modified speech. Listeners ranked Lombard speech as both more intelligible and less effortful than plain speech. Pupil size changes displayed similar tendencies as subjective effort ratings but showed a more complex pattern that varied with SNR. Two differences between the outcomes from pupil size and subjective ratings stand out: (i) while listeners rated SSDRC as the least effortful speech style, pupil size was always smallest in the Lombard speech condition, at all SNRs; (ii) synthetic speech produced the largest effort ratings at all SNRs but pupil size for plain speech was larger in the more adverse condition.

Comparing the results for the naturally produced speech types, plain speech had higher ERPD and sharper peak than Lombard speech (apart from -5 dB SNR at which they did not differ). The peak of the pupil dilation has been widely used for estimating objectively mental effort i.e. the higher the peak value the higher the effort (e.g. in Zekveld et al. [2010]; Koelewijn et al. [2012]; Zekveld and Kramer [2014]). Thus, listeners might have been engaged more for plain speech than for Lombard speech. Additionally, the pupil diameter for plain speech reached its peak with slower rate to that for Lombard speech. It is shown that the time that the pupil diameter reaches its peak (also called peak latency) increases with decreasing speech intelligibility [Zekveld et al., 2011]. Furthermore, subjective and objective measures showed that plain speech requires more effort compared to SSDRC. This outcome can be driven by both the intelligibility gains and the listeners' preferences. Indeed, a study by Tang et al. [2018] found that listeners preferred SSDRC-modified speech over plain speech at low SNRs.

For the naturally and artificially enhanced speech types, the overall lower pupil response for Lombard speech shows that naturally enhanced speech is less cognitively demanding. A possible explanation for this might be the slightly higher speech clarity when listening to SSDRC. The listener might perceive more phonemes when listening to SSDRC, thus having to piece together more phonemes might require the investment of greater effort. Another explanation for perceiving Lombard speech as less effortful compared to SSDRC may be the listeners' expectations for the speech in noise. SSDRC had lower f_0 and was shorter in duration than Lombard speech which is the speaking style that talkers adopt in noise. Finally, for all conditions, the pupil response for SSDRC reached its peak faster. This means that for the artificially enhanced speech type, listeners needed to engage their attention quicker very likely to overcome the acoustic cues imposed by an unusual speech type.

For all conditions, synthetic speech was the least intelligible and for the intermediate noise levels, the most effortful. Previous studies have shown that pupillary responses are related to intelligibility performance [Zekveld and Kramer, 2014]. In line with this, the low intelligibility scores here for this speech type led to extra processing load in performing the task compared to the other speech types. Finally, pupillary responses can also be influenced by naturalness (more prominent in less noise) i.e. less natural speaking style may result in more processing effort.

In Govender et al. [2019], among other tasks, listeners were asked to score the naturalness of the speaking styles heard and they reported lower naturalness for synthetic speech (including hidden Markov model TTS synthesis) compared to natural speech. Only for the most adverse condition in which the intelligibility score for the synthetic speech was approximately 40%, the ERPD of plain speech was higher than that of TTS. Listening effort typically is maximised for speech-in-noise tasks at intelligibility levels of around 50% while for conditions of lower intelligibility the effort declines [Zekveld and Kramer, 2014; Ohlenforst et al., 2017; Wu et al., 2016].

3.3 Experiment II: Impact of different speech types on listening effort for non-native listeners

This experiment was conducted to explore the listening effort that non-native listeners exert when listening to stimuli under conditions similar to those that native listeners experienced in Expt. I (sec. 3.2). Differences from the Expt. I (sec. 3.2.1) are presented below.

3.3.1 Methods

Participants

Thirty-one normal-hearing native Spanish listeners (7 males) aged between 18 and 29 (mean age of 20.5, *S.D* 2.5 years) took part. Fifteen were monolingual in Spanish and the remaining were bilingual in Spanish and Basque. Listeners were students in the English, German, Translation and Interpretation Studies Department at the University of the Basque Country, in the second or later year of their studies. Participants reported that they did not suffer from cataracts nor diabetes, and had no known hearing problems. Additionally, they were asked not to wear hard contact lenses or eye makeup during the experiment. Participants underwent a pure tone hearing screening; all had a hearing level less than or equal to 25 dB in both ears. Listeners were paid on completion of the experiment.

Speech and masker materials

Similar stimuli to those in Expt. I (sec. 3.2) were used. For the native listeners, speech material was mixed with SSN at -1, -3, and -5 dB SNR while for the non-natives the SNRs were higher ($+20$, $+5$, and -1 dB SNR) since L2 sentence listening in adverse conditions has a more detrimental effect for the latter group of listeners [Garcia Lecumberri et al., 2010]. The -1 dB SNR was chosen as the common condition for the two groups and the most adverse one for the non-native listeners. Based on the linear model suggested in Cooke and Garcia Lecumberri [2016] (the stimuli were the same as in this study) which describes the intelligibility loss of non-native relative to native listeners, even the least intelligible speech type (TTS) was not expected to be unintelligible for the non-natives (average intelligibility score around 25%). Additionally, a low-noise condition of $+20$ dB SNR was used as baseline for evaluating the listening effort exerted for the different speech types for a near-clean SNR. Finally, an intermediate SNR of $+5$ dB was tested which corresponds to an SNR in a more realistic scenario [Pearsons et al., 1977; Smeds et al., 2015; Wu et al., 2018].

Procedure

The experiment lasted around 1 hour and 15 minutes (i.e. approximately 15 minutes more than Expt. I) with a 5 minute break in the middle. It took place in a sound proof booth at the University of the Basque Country in Vitoria-Gasteiz. Pupil data was collected using a Tobii x3-120 eye-tracker with sampling frequency of 40 Hz. The pupil size was measured in terms of pupil diameter (an estimate of the pupil size in millimetres). Participants listened to sentences through Sennheiser HD-380 pro headphones.

In addition to the task described in Expt. I (sec. 3.2.1), listeners had to score their level of competence in English for each of the skills: speaking, listening, reading and writing in a scale from 1(=beginner) to 5(=native), and also were asked to read 10 English sentences. Spoken sentences were recorded and used to rate participants' accents. For rating the degree of foreign accent, an online test (see appendix Fig. B.1) was performed and 13 native British English listeners were asked to evaluate the accent. Three evaluators with middle to high contact with Spanish, or had lived in a Spanish-speaking country for more than a month were excluded since their rating ability might have been influenced by their exposure to the Spanish language. Evaluators rated two out of the ten sentences (identical for all speakers) on a scale from 1(=native-like) to 7(=very accented). The rated sentences were drawn from the Harvard corpus (1) 'A fresh start will work such wonders' and (2) 'The club rented the rink for the fifth night'. The web test lasted approximately 5 minutes and ratings of 10 evaluators were used for the analysis.

Calibration

An identical calibration to that used for the native listeners (sec. 3.2.1) was used. Six participants were excluded from the analysis as there were trials with more than 15% of missing values and trials with artefacts (after a visual inspection on the data) that resulted in less than 80% of valid trials per participant.

Statistical analyses

Growth curve analysis was used for evaluating pupillary responses. As in the first experiment, third-order polynomials were used to model growth curves and the analysis time window started from the 0 s (speech onset) until 4.5 s after speech onset. Models were constructed by adding as separate fixed effects the speech type, intelligibility, accent ratings (median values), mean reported proficiency level in English, months lived in a foreign country, and the year of studies. As random effects, the subject id, trial number, and block number were added. Model fitting showed that only speech type as a fixed factor improved the model and thus the remaining factors were excluded from the model. In the +20 dB SNR condition, the interaction between the third polynomial term and speech type was removed due to lack of convergence. For the correlation tests, the Pearson correlation coefficient was computed for the non-repeated-measures. Specifically, comparisons between accent ratings and intelligibility, year of the studies, months that participants have spent in a foreign country, and self-reported mean English level were tested.

Trials for which listeners did not perceive any word correctly were excluded from the analysis. For the −1 dB SNR condition, 10.4% of the trials were excluded, for the +5 dB SNR 4.1%, and for the +20 dB SNR 1.4% (see appendix for more details; Table A.1).

3.3.2 Results

Pupil dilation

Figure 3.10 depicts the ERPD of the raw data averaged across-participants for each speech style and SNR. For the least noisy condition, the pupil dilates similarly for all speech types until approximately 2 s when the pupil size differs until the end of the time course of the ERPD. The divergence among the speech types increases as the noise level increases. For the most adverse condition, Lombard speech has the smallest ERPD value, followed by SSDRC, with plain and TTS having the highest ERPDs.

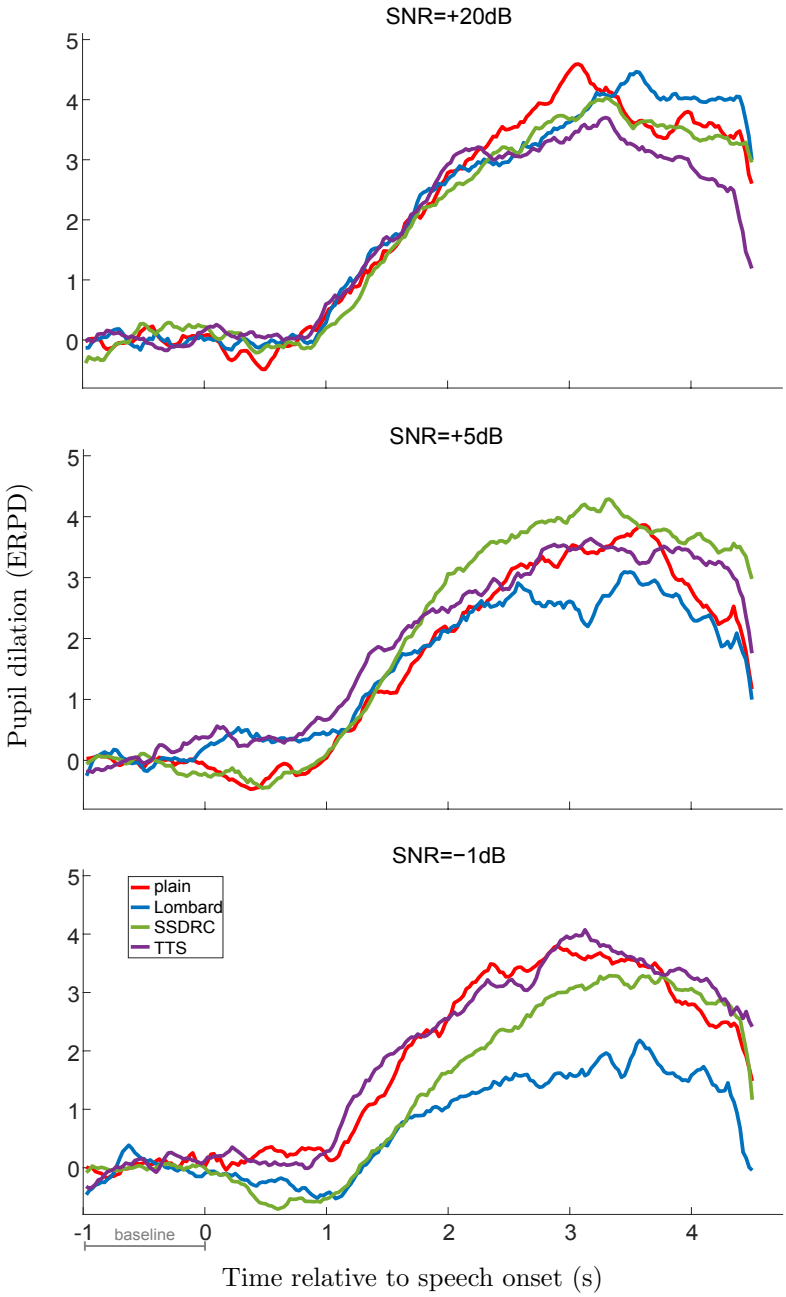


Figure 3.10: *Mean pupil size increase over baseline as a function of speech type. Noise starts 1 s before the baseline onset as shows Fig. 3.4*

The −1 dB SNR condition was identical to that presented to native listeners in Expt. I. The effect with the raw pupil data is comparable to that revealed for the non-natives (Fig. 3.5). The best-fitted model was identical to that of Expt. I (Eq. 3.2) except for the +20 dB SNR which was as follows (corresponding models can be found in Figs. 3.11 - 3.13).

$$ERPD \sim (time1 + time2 + time3) + speech\_type + time1 : speech\_type + time2 : speech\_type + (time1 + time2 + time3|participant) \quad (3.3)$$

with $time1$, $time2$, $time3$ representing the 3 orthogonal terms, *speech_type* the 4 tested speech types, and *participant* the participant id.

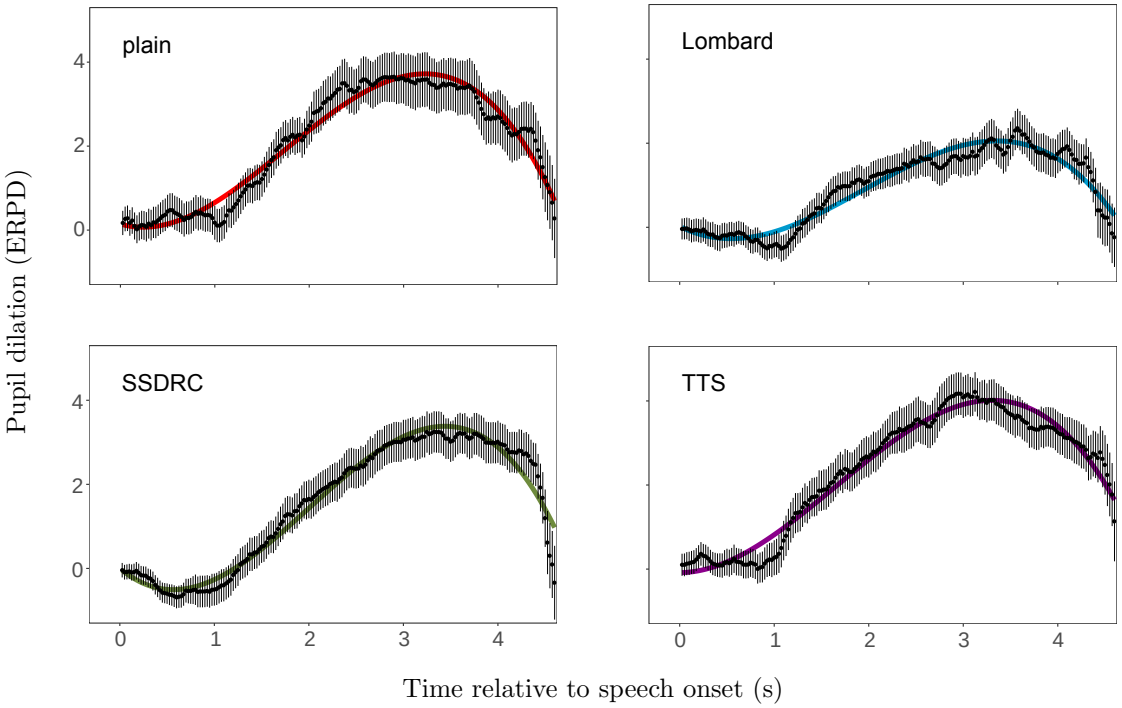


Figure 3.11: *Mean pupil size over time (black dots) with grey error bars to denote the ±1 SE. The solid line shows the fitted model for the −1 dB SNR.*

Table 3.4 shows the estimates of each polynomial term and speech type for the different SNRs and Table 3.5 the interpretation of the GCA results as a function of the polynomial term and SNR.

Intelligibility scores

The mean percentage of correct words repeated by participants for the different speech types and SNRs is shown in Fig. 3.14 (right panel). SSDRC was the most intelligible for all conditions while TTS was the least intelligible. The natural speech types were less intelligible than SSDRC. However, for the more favourable SNR, plain, Lombard, and SSDRC achieved equal scores. Intelligibility of all speech types decreased with increasing SNR.

Statistical analysis verified the visual observations. As noise level decreased intelligibility scores increased [$p < 0.001$] except for SSDRC which was statistically different only for −1 and +20 dB SNR [$p = 0.0001$]. Speech type comparisons revealed that for each of the 3 SNRs the intelligibility scores were significantly different [$p < 0.05$] except for Lombard and SSDRC at +5

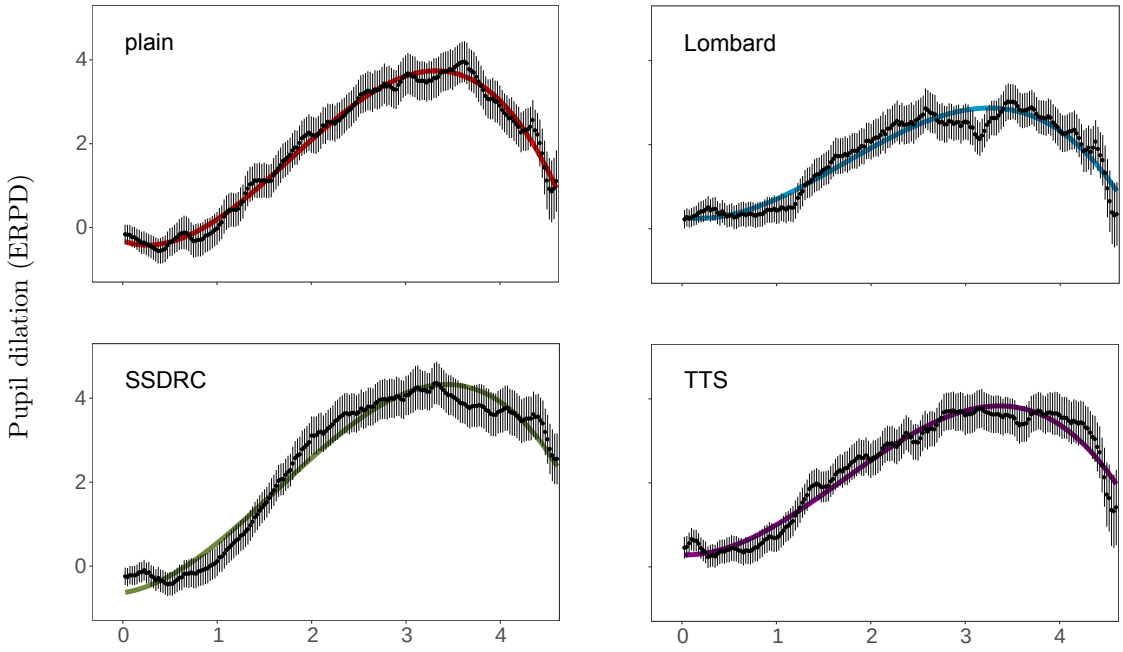


Figure 3.12: *As Fig. 3.11 but for the +5 dB SNR.*

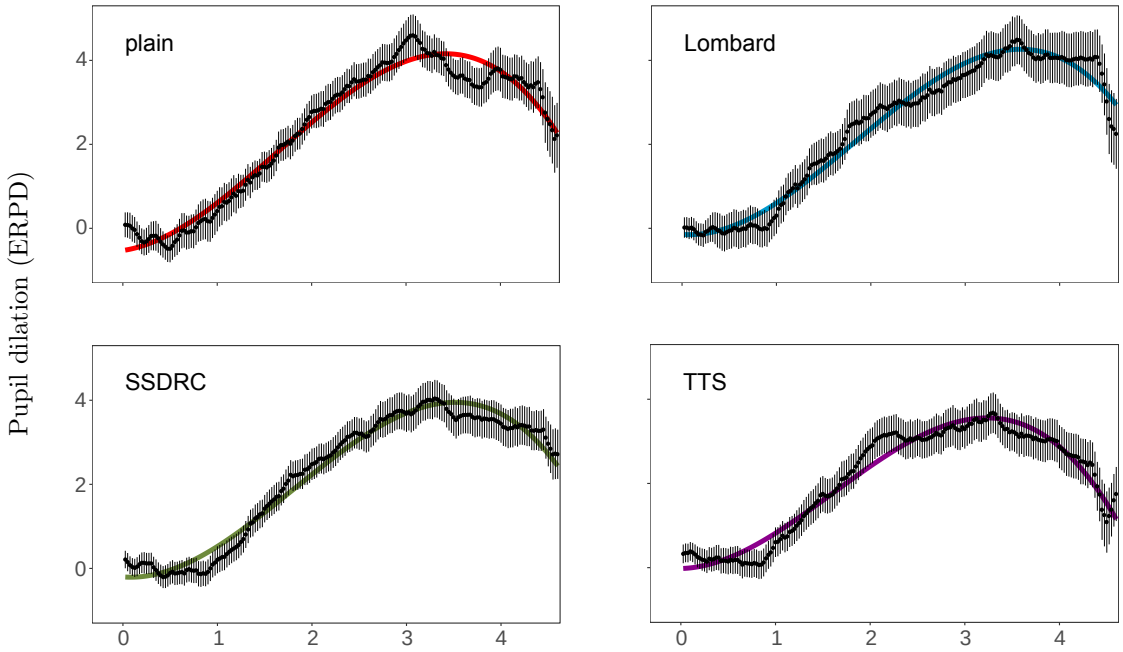


Figure 3.13: *As Fig. 3.11 but for the +20 dB SNR*

and +20 dB SNR. For the most favourable SNR, only TTS was significantly lower compared to the remaining speech types [$p < 0.0001$]. The speech type order from highest to lowest performance was identical for the mid and low SNRs; SSDRC, Lombard, plain, and finally TTS.

For the common condition with the native listeners (−1 dB SNR), non-natives produced

Speech type	+20	+5	-1
Intercept:plain	2.34 (0.56)	1.89 (0.47)	2.11 (0.42)
Intercept:Lombard	0.07 (0.04)	-0.17 (0.04)*	-1.17 (0.05)*
Intercept:SSDRC	-0.12 (0.04)*	0.50 (0.04)*	-0.58 (0.05)*
Intercept:TTS	-0.25 (0.05)*	0.44 (0.04)*	0.24 (0.05)*
time1:plain	18.61 (3.40)	15.94 (3.32)	12.82 (2.95)
time1:Lombard	1.09 (0.60)	-6.62 (0.57)*	-4.18 (0.62)*
time1:SSDRC	-0.58 (0.60)	3.87 (0.57)*	3.31 (0.62)*
time1:TTS	-6.29 (0.60)*	-2.06 (0.58)*	2.33 (0.66)*
time2:plain	-9.01 (1.81)	-9.77 (1.56)	-10.27 (1.56)
time2:Lombard	2.70 (0.60)*	2.77 (0.57)*	5.23 (0.62)*
time2:SSDRC	2.24 (0.60)*	0.44 (0.57)	3.71 (0.62)*
time2:TTS	-0.27 (0.60)	2.39 (0.58)*	1.00 (0.66)
time3:plain		-7.21 (0.92)	-7.09 (1.09)
time3:Lombard		2.65 (0.57)*	1.83 (0.62)*
time3:SSDRC		1.81 (0.57)*	-1.09 (0.62)
time3:TTS		2.38 (0.58)*	1.38 (0.66)*

Table 3.4: *Summary of estimates of intercept and orthogonal polynomial time terms (time1, time2, time3) with plain speech as baseline for the different SNRs. The standard error is shown in parentheses and the asterisk indicates the significant different conditions from baseline.*

Term	Interpretation	Order	+20	+5	-1
Intercept	overall mean pupil dilation	greater to lower	Lombard = plain ≠ SSDRC ≠ TTS	SSDRC ≠ TTS ≠ plain ≠ Lombard	TTS ≠ plain ≠ SSDRC ≠ Lombard
Linear	overall pupil dilation rate	steeper to flatter	TTS ≠ plain = Lombard = SSDRC	Lombard ≠ TTS ≠ plain ≠ SSDRC	Lombard ≠ plain ≠ SSDRC = TTS
Quadratic	shape of peak (height and width of the curve)	sharper to flatter	plain = TTS ≠ Lombard ≠ SSDRC	SSDRC = plain ≠ Lombard = TTS	plain = TTS ≠ SSDRC ≠ Lombard
Cubic	falling slope	faster to slower		plain ≠ Lombard = SSDRC = TTS	SSDRC = plain ≠ Lombard = TTS

Table 3.5: *Interpretation of each polynomial term and results as a function of SNR. Results are ordered based on the 3^{rd} column. The symbol '=' signifies that the speech types were not statistically significant different and '≠' the opposite.*

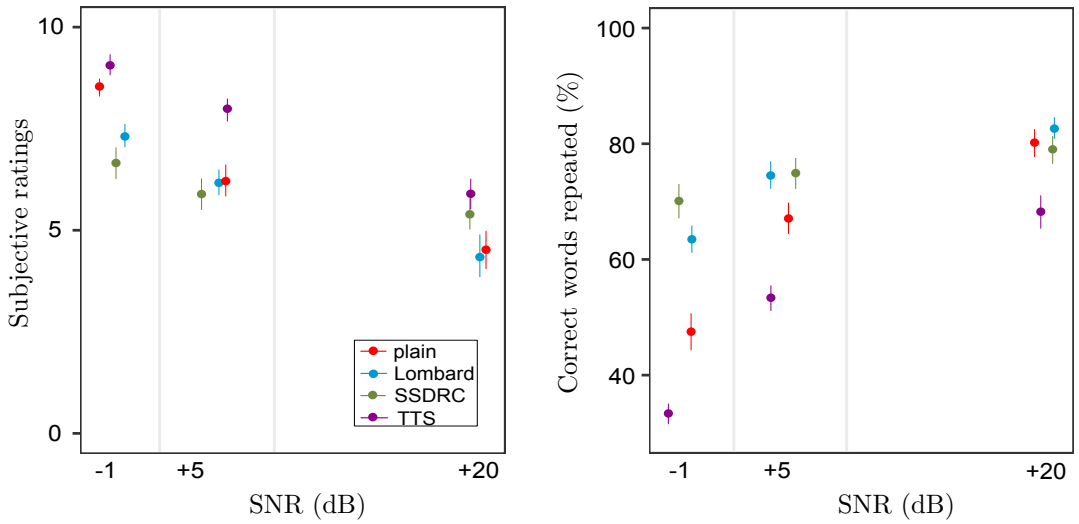


Figure 3.14: *Left plot: mean subjective listening effort ratings from 0 (no effort) to 10 (very effortful). Right plot: mean intelligibility scores. Error bars denote ±1 standard error.*

much lower intelligibility scores. The drop in intelligibility score was 28% points for SSDRC, 33% for Lombard, 44% for plain, and 34% for TTS. However, the ranking was the same. Even for the condition with the least noise (+20 dB SNR), non-natives had much lower performance than that of the native listeners at −1 dB SNR.

Subjective listening effort ratings

Mean subjective ratings for the different speech types across the 3 SNRs are depicted in Fig. 3.14 (left panel). For the adverse noise level, synthetic speech was considered as the most effortful while SSDRC and Lombard speech the least. For the positive SNRs, plain speech was as effortful as SSDRC and Lombard speech, while for the +20 dB SNR all speech types had ratings around the middle of the scale. Subjective effort of all speech types increased with increasing SNR. As for the intelligibility scores, subjective listening effort ratings showed a clear ranking across speech types that is the inverse of the intelligibility scores.

Post-hoc comparisons indicated that as noise level decreased, the subjective listening effort of all speech types decreased [$p <= 0.001$] (except for TTS and Lombard between −1 and +5 dB SNR which were not statistically different) while for SSDRC effort did not change (with marginal difference between −1 and +20 dB SNR; [$p = 0.07$]). Comparing the speech types, the reported effort for Lombard and SSDRC was similar for the different SNRs. Plain speech was reported as more effortful than SSDRC only for the −1 dB SNR [$p = 0.0001$]. TTS always reported as the most effortful [$p < 0.05$] except for the +20 dB SNR in which it was reported with the same rating as for SSDRC and for −1 dB SNR in which it was similar to that of plain speech. The correlation test showed that subjective listening effort was negatively correlated with intelligibility scores [$r = -0.71$].

Regarding the common condition with the native listeners (−1 dB SNR), non-native listeners rated all speech types as more effortful with a near-constant difference of approximately 3.9 points, except for TTS for which the difference was 1.7 points. However, the ranking was the same i.e. from the least to the most effortful it was the SSDRC, Lombard, plain, and finally TTS.

Accent ratings

To assess any relationship between participants' year of the studies, months in a foreign country, self-reported mean English level, and task performance with the mean accent ratings across judges for each participant was computed and Pearson correlation was used. Accent ratings were negatively correlated with intelligibility [$r = -0.41, p < 0.05$], year of studies [$r = -0.44, p < 0.05$], the months that participants have spent in a foreign country [$r = -0.39, p = 0.05$] and were not correlated with the self-reported mean English level. In other words, the more native-like the voice sounds, the better the listener's intelligibility performance or the higher the year in English studies or the more the months lived in a foreign country.

3.3.3 Interim discussion

Expt. II explored listening effort and intelligibility for non-native listeners. Spanish participants with a high proficiency level in English listened to English sentences in the presence of speech shaped noise at 3 noise levels. Listening effort was evaluated subjectively by asking the listeners to estimate the effort that they exerted in a continuous range from 0 to 10, and objectively using pupillometry. Additionally, intelligibility scores were computed to ensure that the task

was properly performed and to test its correlation with the listening effort measures.

For the most favourable SNR, medium or no differences among the speech types except TTS in all three measures were observed. Exploring the listening effort for a near-clean SNR which acts as a ceiling condition for the non-native listeners reveals differences that are purely related to the different types of speech and are not influenced by the masker's characteristics. In line with Rönnberg et al. [2013], little cognitive effort was reported when listening to naturally produced speech in quiet conditions. Although, the results of this study revealed that naturally produced speech types led to the greatest overall mean pupil dilation and artificial speech (i.e. SSDRC and TTS) to lower when presented at the most favourable SNR. This observation contradicts with Rönnberg et al. [2013] results possibly because they conducted experiments with native listeners. Non-native listeners may not be of benefit from similar acoustic cues to native listeners. For the SSDRC condition, listeners achieved similar intelligibility scores to the natural speech types but with slightly lower effort (flatter peak and lower overall pupil dilation). A greater speech energy concentration to higher frequencies might have made the speech comprehension easier. However, in the absence of noise, SSDRC is expected to be more detrimental than the naturally produced speech types since SSDRC processing changes the acoustic-phonetic structure of speech (e.g. formant energy modifications). Cooke and Garcia Lecumberri [2016] found a drop in intelligibility for SSDRC in quiet compared to noisy conditions. In the current study, TTS was less intelligible compared to the other speaking styles (which scored approximately 10% higher). For this speech type, listeners may have reached their highest possible intelligibility score, which could not be improved even with extra effort.

As it was expected under adverse noise levels, Lombard speech facilitated listeners' intention to understand speech compared to plain speech. This was true for all three measures in the experiment; listeners perceived speech with higher clarity, reported less effort, and pupil response was smaller (lower overall pupil dilation and flatter peak pupil dilation). This result is in line with the study by Borghini and Hazan [2020] in which non-native listeners benefited from clear speech relative to plain speech in the presence of babble noise.

Interestingly, the results for the Lombard and SSDRC speech types showed medium or no differences in intelligibility and subjective rated effort while pupil responses differed significantly. The opposite behaviour was observed for these two speech types in the most adverse and most favourable conditions. The cognitive load measured with the pupil size revealed that for the -1 dB SNR, Lombard speech was perceived with less effort (lower overall pupil dilation, steeper peak pupil dilation) and for the $+20$ dB SNR, with more effort than SSDRC. Although, SSDRC had been developed using features of Lombard speech, for the -1 dB SNR, the extra effort exerted might have resulted from features like f_0 which have not be adopted by SSDRC. Finally, for the most favourable SNR, listeners may have had to engage more for the Lombard speech. This may happen since Lombard speech is not a speech type that a listener would expect to hear in low-noise conditions.

3.4 General discussion

Previous studies have showed that non-native listeners do not perform equally well to native listeners in word identification tasks [Cooke and Garcia Lecumberri, 2016] and have to allocate a greater amount of cognitive resources compared to natives [Borghini and Hazan, 2018]. Thus, the noise levels used for the non-native listeners are lower compared to those for the native

listeners. The conditions of SNR equals to −1 dB was kept identical for both groups for evaluating the listening effort when the listeners are exposed to the same amount of acoustic-phonetic cues (second research question). The first research question has been answered for each experiment separately in the corresponding sections (3.2 and 3.3). The results demonstrated that listening effort of native and non-native listeners varies with type of speech, as judged both by participants' own ratings and by the physiological measure of pupil size change.

Intelligibility ranking from the speech types is similar for both groups.

The ranking of intelligibility derived from the speech types for the common condition is similar for both groups of listeners. For non-native listeners the loss in word recognition was around 35% (Fig. 3.9, 3.14 (right panels)). The intelligibility ranking observed is in line with that reported by Cooke and Garcia Lecumberri [2016] in which the effect of the same speech types on intelligibility in the presence of SSN was tested for non-native listeners. The results in this chapter might reflect the resistance of each speech style to the energetic masking. In line with previous studies, synthetic speech was less intelligible than natural speech [Venkatagiri, 2003; Axmear et al., 2005] and the least intelligible compared to plain, Lombard, and SSDRC speech types [Cooke and Garcia Lecumberri, 2016]. This might be a consequence of the different formant structure or/and shorter duration compared to the other three speech types which can be observed in Fig. 3.1.

Listening effort patters similarly for the different speech types for both groups.

Additionally, the listening effort revealed by pupillary responses for native and non-native listeners patterns similarly for the different speech types. More specifically, TTS and plain speech had the sharpest peak and the greatest overall mean pupil dilation, SSDRC follows, and finally Lombard speech is the one with the flattest peak and the lowest overall mean pupil dilation. Previous studies have showed that the higher the degradation of the signal, the larger the decrease in intelligibility and quality having as a result an increase in pupil dilation [Zekveld and Kramer, 2014; Koelewijn et al., 2012]. In Borghini and Hazan [2020] clear speech, which is also a naturally enhanced speech type, reduced the listening effort of both native and non-native listeners in the presence of babble noise compared to plain speech. Correspondingly, here, Lombard and SSDRC reduced listeners' effort more compared to plain and TTS. The greater duration of Lombard speech may not have contributed so to be perceived as the least effortful since Koch and Janse [2016] found that the increased speech rate does not have an effect on pupil response in young or older listeners. Lombard speech may have lessen the explicit reliance of speech understanding on working memory resulting in a lower cognitive load for both groups compared to the other speech types tested in this chapter.

Slower rise of pupil size for non-native listeners compared to natives.

A difference in pupillary responses between native and non-native listeners is that for the former group, the pupil starts to dilate just a few milliseconds after the onset of the sentence while for the latter it starts around 1 second later. This could be explained by either native listeners engage faster with the task or both groups are engaged but for the non-natives the task is harder. This was observed for all speech types. The slower rise of pupil size might signify greater cognitive demands. Several factors can contribute to the increased effort of L2 listeners such as the higher number of competing words triggered by their first language and lower proficiency level compared to natives. Thus, the greater peak pupil latency revealed for

non-native listeners might be derived from the longer processing required for comprehending speech. Pupillometry results in Borghini and Hazan [2018, 2020] showed that listening effort (measured with the mean and peak dilation) during sentence identification is higher for non-native compared to native listeners when intelligibility is equated.

Subjective ratings of listening effort are not always consistent with physiological measures.

Subjective ratings of listening effort are not always consistent with physiological measures. In both experiments, pupillary responses to synthetic speech across the different SNRs showed no clear relationship to the subjectively reported effort. Participants might have reported their opinion of their performance on the task in response to the subjective question they were asked. In Zekveld and Kramer [2014], subjective effort evaluation showed that processing load was higher for lower intelligibility, while subjective ratings and peak pupil dilation were not related to each other. This finding is also supported by other studies [Zekveld et al., 2010; Koelewijn et al., 2012] with Wendt et al. [2016] concluding that subjective ratings and pupil dilation may well represent different aspects of effort. Non-native listeners reported higher effort compared to natives for the -1 dB SNR and had much lower intelligibility scores (lower than 80% independent of the speech type). This was expected since listening to a second-language under imperfect conditions is more demanding even if the proficiency level of the listener is high. This condition was the most adverse condition for the non-native listeners and thus their ratings might be more negatively scored since they might have rated the conditions comparatively. Even though the ranking of the speech types was the same for the two groups. The cognitive load derived by pupil diameter may provide complementary information to subjective ratings and intelligibility scores.

Chapter 4

SpeechAdjuster: A tool for investigating listener preferences and speech intelligibility

4.1 Introduction

[1] In the previous chapter, one dimension of the overall listener's experience, listening effort, was investigated. A pupillometry study was conducted to estimate the required effort when listening to different speech types in noise. Results revealed a clear impact of speech type on the cognitive demands required for speech comprehension with Lombard speech to induce less demands on mental processing compared to the other speech types tested. It is of interest to examine the influence of distinct speech properties that speakers naturally modify to produce Lombard speech. In this chapter, a tool for investigating supra-intelligibility aspects of speech is introduced and in the following chapters, this tool is used to study such aspects when altering only distinct speech properties.

Intelligibility is readily measurable, but a different approach is required to capture attributes above and beyond word or sentence scores. Subjective preferences have traditionally been measured using rating scales [Moore et al., 2007; Adams and Moore, 2009; Brons et al., 2013], but these paradigms require a participant to map a large and potentially-complex subjectively-interpreted concept such as quality on to a rather artificial and usually discrete set of values such as 'very natural', 'quite natural' and the like. Furthermore, while intelligibility and subjective factors can be measured in the same task, for practical reasons these measurements are sequential and hence delayed relative to the stimulus, raising issues such as whether individual differences in working memory capacity might affect the outcome.

In this chapter, an alternative approach which attempts to avoid issues of interpretation and delayed responses is presented. The technique is to provide listeners with the ability to manipulate some (usually continuous) dimension of interest in real-time, and to select a parameter setting that they judge to be in some sense optimal. Instructions given to participants are deliberately neutral, and the emphasis is simply on the discovery of a preferred setting that allows them to recognise as many words as possible. Participants are able to spend as long

[1] SpeechAdjuster was published as a paper in Interspeech 2021 proceedings [Simantiraki and Cooke, 2021].

as necessary on exploring the available stimulus space governed by the parameter of interest. Having chosen an 'optimal' setting, listeners carry out a short task with the parameter setting frozen. For instance, the task might involve a small number of test phrases in which participants identify words in sentences.

SpeechAdjuster, an open-source, cross-platform software tool which allows the manipulation of virtually any aspect of speech, and supporting joint elicitation of listener preferences and intelligibility measures is described. It operates by precomputation of modified speech for a fixed set of points on a given modification continuum, and uses smooth, rapid, mid-utterance switching to produce the sensation of continuously-variable speech alteration.

4.2 SpeechAdjuster

For the purpose of illustration in what follows, imagine we wish to examine the possible influence of the mean fundamental frequency (f_0) of a target speech signal in the presence of background noise. Participants are provided with the means to modify mean f_0, and are instructed to use this control to adjust f_0 in such a way as to recognise as many words as possible. The task might be described as being akin to tuning a radio set to produce the best possible signal.

4.2.1 Adjustment and test phases

A SpeechAdjuster experiment consists of a sequence of trials, each of which is made up of an open-ended adjustment phase, optionally followed by a fixed-length test phase. In the adjustment phase, the listener is presented with speech material such as words, phrases or continuous speech, with or without masking noise, and the task is to explore the parameter space for the characteristic under study (e.g. mean f_0) in order to find a value which the listener considers optimal in terms of understanding as many words as possible. When the participant decides that the adjustment is complete, the endpoint value of the chosen parameter is used to generate one or more test stimuli that the user responds to, just as in a traditional speech intelligibility task. During the test phase, listeners supply responses to stimuli by typing in an input box. To avoid memory load, listeners are permitted to start typing at any point after the onset of the stimulus.

4.2.2 Virtual control of speech parameters

There are many ways to elicit continuous uni-dimensional preferences. Here, five different mechanisms were explored: (1) a pair of up/down arrow buttons; (2) a mouse wheel; (3) a normal scrollbar; (4) a scrollbar whose value returns to the midpoint when released; and (5) a virtual rotary knob. In pilot experiments, five normal-hearing adult listeners with Spanish as a native language adjusted the volume level of Spanish sentences, first in quiet and then in stationary masking noise, using each of these mechanisms in independent trials. Listeners were consistent in reporting that the rotary knob and up/down arrows were the easiest to understand and apply. Consequently, SpeechAdjuster provides the experimenter with the choice of these two forms of input (Fig. 4.1).

The selection of which mechanism is the most appropriate to use will be task-dependent. The pair of arrow buttons avoids the listener having a visual indication of the current value of the parameter (apart from feedback that the upper or lower extreme has been reached). Since

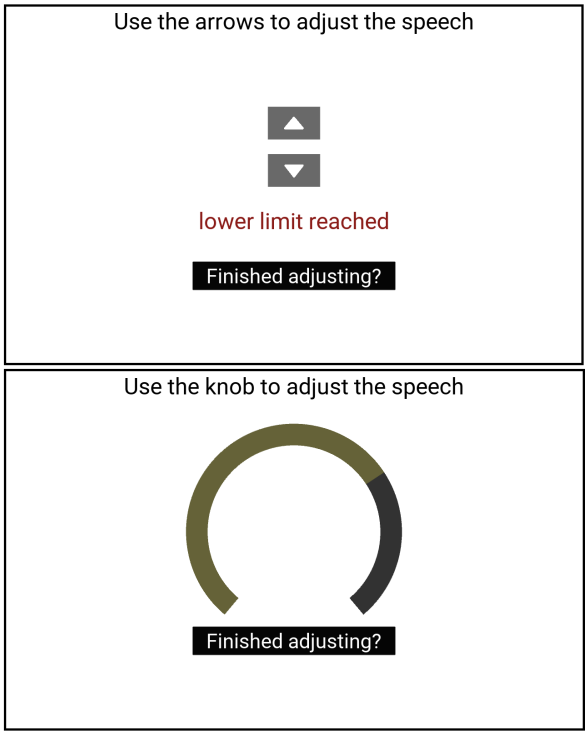


Figure 4.1: SpeechAdjuster *GUI options during the adjustment phase: (a) a pair of arrow buttons (up/down) and (b) a virtual rotary knob. During the test phase, a text input box is added.*

the listener has no indication of the parameter value at the start of each trial, they are prevented from adopting a strategy based on using the same parameter settings as on the previous trial, purely on a visual basis, since this may or may not be the most appropriate setting. On the other hand, the rotary knob can be used to simulate realistic scenarios where listeners are aware of the current parameter value and the need to adjust it from trial to trial on the basis of clear between-trial acoustic changes. To some extent, the choice will depend on how trials are blocked across conditions. For example, the choice of a mean f_0 value in the presence of stationary noise might be expected to be similar from one trial to the next if the experiment is blocked by noise type, motivating the use of up/down buttons. Conversely, if the masker changes from trial to trial, or if the masker is the same but varies in some property that is likely to interact with the target speech (e.g. a competing speech masker), the mean f_0 chosen might differ from trial to trial, in which case the visually more intuitive rotary knob would be preferred.

Figure 4.2 depicts the adjustment phase of a typical series of trials where a listener was able to control mean f_0. The plot shows all the speech modifications that a listener performed via up/down buttons during the adjustment phase for each of five trials. Initial f_0 values were chosen at random. These traces exhibit typical features of user-controlled exploration of parameter space: listeners sample the entire range during some trials, while other trials show more rapid adjustment phases and faster decisions, and overall there is a high level of consistency in the final value chosen. In this instance, listeners were not allowed to proceed to the test phase until five seconds had elapsed. This user-configurable value ensures that participants spend at least the specified time exploring the space of possible adjustments before signalling that they are ready for the test phase.

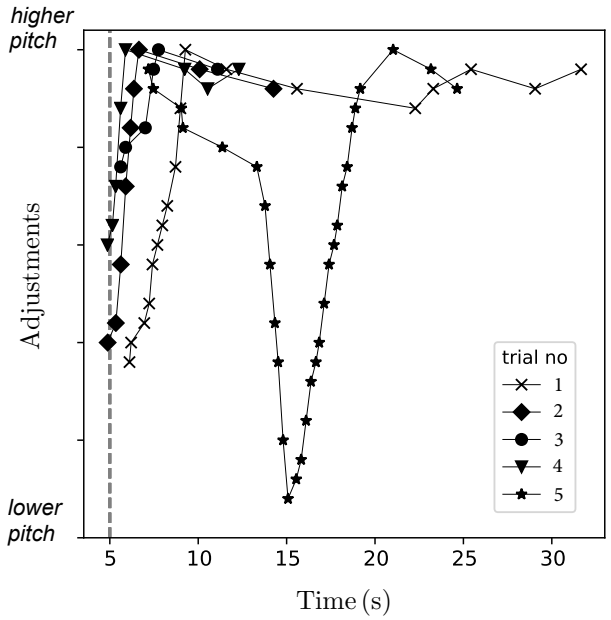


Figure 4.2: *A listener's f_0 adjustments (y-axis) across time (x-axis) for five independent trials. The vertical dotted line indicates the time point (here 5 s) when the completion button (denoted Finished adjusting? in Fig. 4.1) in the adjustment phase was activated.*

4.2.3 Stimulus preparation

SpeechAdjuster requires each stimulus (e.g. word, phrase or longer speech passage) to be precomputed at each of a range of discrete parameter values. For instance, in the case of f_0, each experimental stimulus will be processed offline to produce N different exemplars that differ only in mean f_0, with the N points along the f_0 continuum chosen by the experimenter to meet some criterion such as equal-spacing on a semitone scale. The number N of such levels is customisable and only impacts on the amount of offline storage required, and does not affect the latency of online processing. In our own experiments [Simantiraki et al., 2020; Simantiraki and Cooke, 2020] we have found that 20-25 discrete values are adequate to produce the impression of continuous change.

In online operation, all N versions of the same stimulus (e.g. the same sentence with different mean f_0) can be considered to be activated in parallel, and the user's actions control which one is actually chosen to be output by SpeechAdjuster at any given time point. In practice, the signal that the listener hears is merely the concatenation of segments. Switch-over is low latency and to minimise artefacts a short fade-out ramp is applied to the current segment and a similar fade-in ramp applied to the next segment corresponding to the new stimulus.

Figure 4.3 shows an example speech spectrogram that results as a consequence of a listener adjusting mean f_0 at several points during the utterance.

4.2.4 Configuration

The experimenter can adjust many parameters of SpeechAdjuster, allowing it to be adapted to the requirements of each listening task and to the linguistic background of the participants. Options include those that control the tool's appearance, the textual content and language of all interface components, participant instructions, inter-stimulus delays and numbers of trials. Other options control the size of audio chunks used during streaming, chosen to ensure that

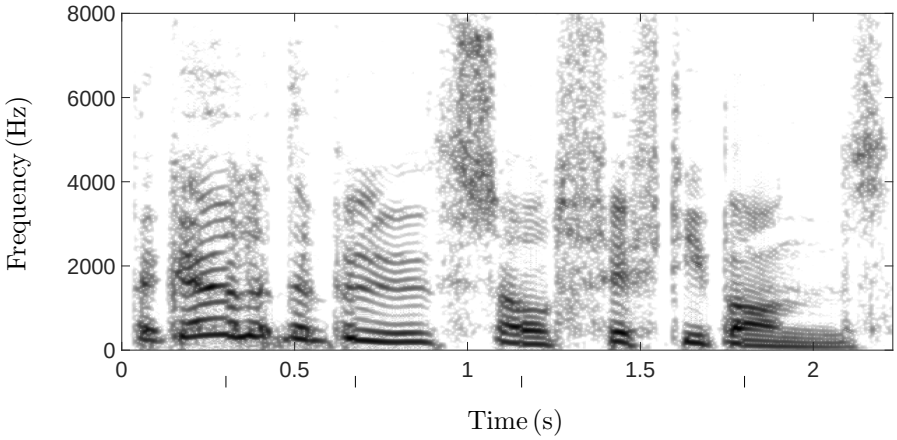


Figure 4.3: *A spectrogram of the speech sample 'The girl at the booth sold fifty bonds' that results from a listener making changes to mean f_0 at the four time instants denoted by the vertical lines. The initial mean f_0 is around* $290\,Hz$*, while the final value is* $130\,Hz$*.*

user-controlled changes are applied rapidly, but without audible artefacts. A complete list of options can be found in the user guide that is provided with the application.

4.2.5 Outputs

SpeechAdjuster collects detailed information during both the adjustment and test phases. In the former phase, the tool makes available both raw data in the form of time-stamps for all adjustments, and summary data on the initial and final parameter values and the total time taken to move to the test phase. Textual responses are collected during the test phase. SpeechAdjuster can also produce a range of figures that depict experimental outcomes. Specifically, the tool can (1) visualise the adjustments that a listener performed in a trial (as shown in Fig. 4.2); (2) produce a histogram of listeners' preferences; (3) generate box plots of listeners' choices and the time needed for the adjustments across the different experimental conditions; and (4) display a two-dimensional heatmap showing each listener's preferences for each of the tested phrases.

An example of the use of data produced by SpeechAdjuster on listeners' mean f_0 preferences coupled with intelligibility scores is shown in Fig. 4.4. This figure illustrates that preferences tap into information over and above intelligibility: in this case, the proportion of words identified correctly is at or near ceiling across the entire range of mean f_0 values, but listeners express a clear preference for values at or above the mean f_0 of the original (unmodified) speech material.

4.2.6 Implementation, platforms and availability

SpeechAdjuster is open-source software with a GNU General Public License v3.0. SpeechAdjuster is written using the Python programming language and makes use of cross-platform libraries, specifically Kivy [Virbel et al., 2011] for the graphical user interface and PyAudio [Pham, 2006] for audio streaming. Consequently, SpeechAdjuster can be used on Windows, OSX and Linux variants.

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

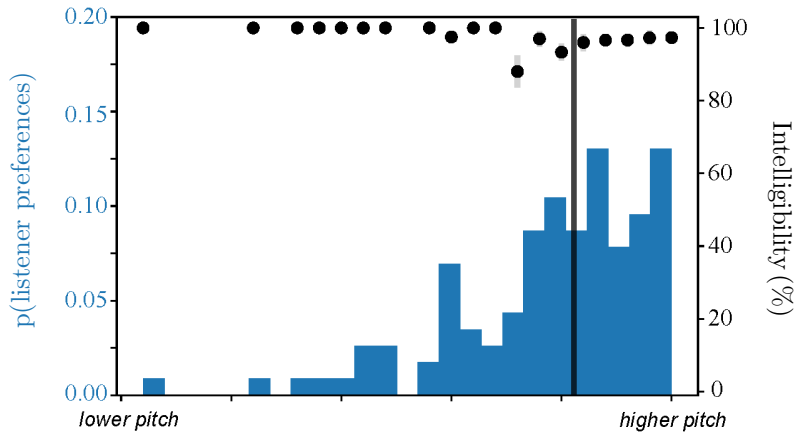


Figure 4.4: *Probability of each mean f_0 value (histogram, left axis), along with the percentage of words recalled correctly (black dots, right axis). Error bars represent ± 1 standard error. The vertical line corresponds to the mean f_0 of the original speech.*

4.3 Applications

SpeechAdjuster has been used to investigate the effect of changes in speech rate [Simantiraki and Cooke, 2020] and spectral energy reallocation, including spectral tilt modifications [Simantiraki et al., 2020]. Precomputation of stimuli permits many types of speech transformation, of arbitrary complexity, to be investigated. Examples of more complex processes include gradations in degree of foreign accent, emotional valency, or more general voice morphing. SpeechAdjuster could also be used to explore user preferences in some dimension of interest in speech synthesis, or to choose between families of synthesis algorithms. Other applications include the determination of optimal parameters in audio engineering in which the level of one audio signal is reduced by the presence of another signal [Torcoli et al., 2019] or of the proper balance between intelligibility and supra-intelligibility aspects of speech important for near-end listening enhancement algorithms [Chermaz and King, 2020].

In addition to testing listeners' preferences directly, SpeechAdjuster can help in the selection of starting parameters for conventional listening experiments with fixed conditions, and has been used in this manner in experiments on distorted speech involving sine-wave and noise-vocoded speech generation. Precomputation also allows for the possibility of experimental screening and modification of stimuli to enable artefact-removal. Figure 4.5 shows an example for three types of distorted speech.

4.4 Limitations

While SpeechAdjuster supports the elicitation of listeners' preferences and generates information that is clearly complementary to intelligibility scores (e.g. Fig. 4.4), it does so in a holistic manner, and consequently is unable to say anything about the weighting of individual factors that influence listeners' preferences, which may be due to listening effort, naturalness, pleasantness, attractiveness, familiarity, distortion or other quality-related considerations, and which can be expected to show substantial inter-participant variability.

One limitation of the current version of the tool is that speech transformations cannot involve nonlinear modifications to the length of speech constituents, as would be the case most obvi-

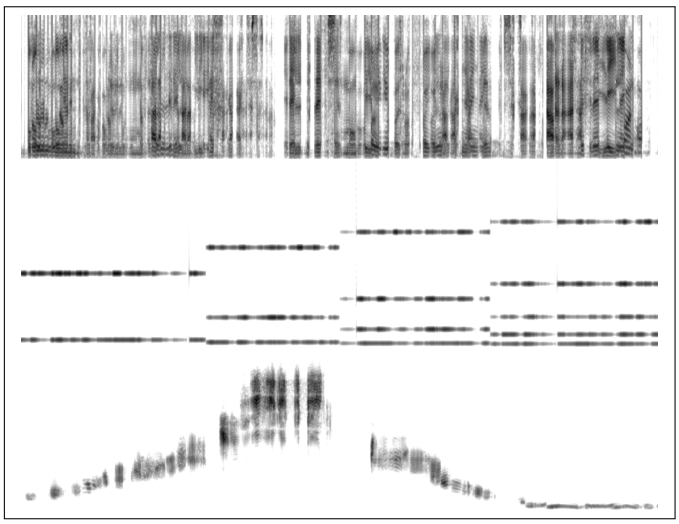


Figure 4.5: *Examples of the use of* SpeechAdjuster *to explore the parameter space of three forms of distorted speech. The top panel shows locally time-reversed speech where the user was able to adjust the size of the window within which reversal took place. The middle panel depicts the output of a tone-vocoder where listeners changed the number of vocoder channels from 2 to 5 as the utterance progressed. The lower panel shows the result of filtering speech through a narrow spectral slit. In this case the user was able to control the centre frequency of the filter.*

ously in speech rate variation, but could also occur in modifications involve mapping between speech styles such as plain, clear or Lombard speech that typically involve changes in segment durations. However, linear elongation has been tested in Simantiraki et al. [2020] in which speech rate modifications were applied on single words and the changes while tuning speech were applied from the next word onwards. In principle, while linear speech rate variations are straightforward to implement, nonlinear changes will require some form of segment annotation to ensure that durational changes are applied at the correct time-points when switching from one parameter value to the next.

Chapter 5

Listener preferences - Speech rate

5.1 Introduction

[1] Previous studies have shown that listeners are sensitive to the perception of speech rate [Kidd, 1989; Smith et al., 1989; Dilley and McAuley, 2008; Peelle and Davis, 2012]. Fast speech has been found to disrupt intelligibility of both natural [Fairbanks and Kodman, 1957; Versfeld and Dreschler, 2002] and synthetic speech [Lebeter and Saunders, 2010; Valentini-Botinhao et al., 2014]. However, there is little agreement on whether a slower speech rate benefits intelligibility [Adams and Moore, 2009; Adams et al., 2012; Nejime and Moore, 1998; Cooke et al., 2014b; Cooke and Aubanel, 2017]. A recent study [Cooke and Aubanel, 2017] found no intelligibility gains for linearly-elongated speech when presented in stationary noise, but significant gains for the same speech in the presence of both competing speech or speech-shaped noise whose envelope was modulated by that of the competing speech. However, it was unclear whether the benefit was due to the net availability of more phonetic information due to the dips in the masker, or to a difference in modulation rates between the target and masker speech. This chapter focuses on speech rate that is expected to impact both intelligibility and listener preferences.

Traditionally, in experiments, a small number of experimenter-chosen rates is used [e.g. in Nejime and Moore, 1998]. However, a different approach in which listeners are allowed to control the stimuli rate has been also used. In Zhao [1997], listeners could modify the auditory speech rate by clicking the on-screen buttons 'Faster' and 'Slower' for making speech faster or slower, respectively. In Piquado et al. [2012], the presentation of narratives interrupted at periodic intervals and participants were allowed to pause before initiating the next segment. Both studies concluded that when the listeners controlled for the stimuli, speech comprehension improved. Novak III and Kenyon [2018] used an on-line speech dilation technique [Novak III et al., 2014] to allow listeners to modify speech rate in real-time using an on-screen slider bar. In that study, listeners were asked to fine-tune the rate of a speech signal in varying levels of background noise. They tested listener intelligibility of the preferred and unmodified speech rates and showed that as noise level increased, decreased speech rates were preferred.

The study in this chapter extends the previous research by investigating the impact of different masker types and masker modulation rates on listener preferences, while also allowing listeners to change speech rate in both directions (faster and slower). As per Cooke and Aubanel

[1] Portions of the work described in this chapter were published as a paper in Interspeech 2020 proceedings [Simantiraki and Cooke, 2020].

[2017] results for the speech-shaped noise, it is hypothesised that if listeners' choices are based on intelligibility, speech rate might not vary with regard to the noise level since slower speech rate was not beneficial to intelligibility. Also it is hypothesised that for lower variations of the speech-modulated masker, listeners will choose faster speech rates. Elongated masker's 'valleys' may function as the listening in the dips phenomenon (i.e. pull out and stitch together speech pieces from momentary masker's dips). The intelligibility gains of this phenomenon has been previously studied [Miller and Licklider, 1950; Peters et al., 1998; Füllgrabe et al., 2006]. The target speech was linearly-elongated so the durational modifications to be independent from masker fluctuations. This is in contrast to local modifications of duration for minimising energetic masking for which apriori knowledge of the masker is necessary. Listeners are able to control the speech rate using the SpeechAdjuster (see chapter 4). Listeners changed speech rate in an open-ended adjustment phase, followed by a fixed length test phase (sec. 5.3). In separate conditions, listeners adjusted and identified speech in quiet, speech-shaped noise, and speech-modulated noise (sec. 5.2).

The research questions addressed in this chapter are: for the stationary noise, does speech rate vary with regard to the noise level; for the fluctuating noise, do listeners choose a different speech rate than that of the masker?

5.2 Methods

5.2.1 Listeners

Eighteen native Spanish listeners (15 females) aged 18-23 (mean age of 19.9 years; SD = 1.4 years) participated in the experiment. All passed an audiological screening with a hearing level better than 25 dB at frequencies in the range $125 - 8000$ Hz in both ears. Listeners were paid 20 euros for their participation.

5.2.2 Stimuli

Speech material

The speech material was drawn from an open source Spanish-words corpus [Tóth et al., 2015]. This corpus consists of 3968 high frequency Spanish words, spoken by four talkers, two male and two female. The words were in a read speaking style and each one consisted of up to three syllables (e.g. 'abierta' which means 'open' in English). For the experiment, the words uttered by one of the female talkers were used.

Speech morphing

Linear elongation/compression was employed and all rate morphing was carried out using TANDEM-STRAIGHT [Kawahara et al., 2008], a version of the STRAIGHT vocoder [Kawahara et al., 1999]. The TANDEM-STRAIGHT framework deconstructs an input speech into three parameters, modifies, and reconstructs a speech signal based on the source-filter model.

Twenty-two different speech rates were available for listeners to choose covering the range from 2.5 times slower than the original to 2.5 times faster, with speech rates located at equally-spaced points on a multiplicative inverse scale. Since the target speech material was read speech and hence not as fast as casual speech, we determined on the basis of informal listening tests

to deploy 15 steps with the speech rate faster than the original, one at the original rate, and 6 with rates slower than the original. Words were independently normalised to have an equal root mean square level after rate modifications, on a 20 ms half-Hamming ramps were applied to reduce onset/offset transients.

Stimuli

Stimuli were presented in quiet and in 8 additive noise masking conditions: speech-shaped noise (SSN) at SNRs of 0, +6 and +12 dB, and speech-modulated noise (SMN) for 5 envelope modulation rates, mixed with speech at +6 dB SNR. Maskers were based on concatenated Spanish sentences from the Sharvard corpus [Aubanel et al., 2014], spoken by a female talker. Maskers were unrelated to the target speech stimuli described above. The SSN masker resulted from passing random uniform noise through a filter with the long-term spectrum of the concatenated sentences. The SMN masker was generated by multiplying the SSN masker by the instantaneous envelope of the concatenated sentences. In addition to the original rate, envelope modulation rates 1.4 and 2.5 times faster, and 1.7 and 2.5 times slower than the original were tested (rates of 1.4 and 1.7 correspond to equidistant steps from the original rate on the 22-point scale used here). In the masked conditions SNRs were computed by concatenating stimulus words without gaps.

5.3 Procedure

For the experiment, SpeechAdjuster was used. The experiment was blocked into the 9 conditions described above. Across participants, block order was counterbalanced using a Latin square design. Each block contained 22 trials. A trial started with a speech rate from the 22 available randomly permuted. The trial consisted of an adjustment phase followed by a test phase. During the adjustment phase, words were presented in a randomised order with 500 ms of intervening silence. Participants were instructed to choose an optimal value of speech rate that allowed them to recognise as many words as possible. Participants were able to control speech rate using the up/down arrow keys to speed up or slow down respectively. Changes in speech rate was applied to the next and subsequent words. The adjustment phase continued for as long as participants required (which averaged 7.04 s; $SD = 5.59$ s). Having finished their choice of speech rate, participants were able to proceed to the test phase by clicking an on-screen button. Participants were not allowed to proceed to the test phase until at least five seconds of the adjustment phase had elapsed. In the test phase, participants were presented with words spoken at the speech rate chosen in the adjustment phase, under the same experimental condition as they had experienced during the adjustment phase. Participants had to identify words during the test phase and type them into an on-screen text input box. During the test phase of a single trial, five test words were presented consecutively and across conditions no word was repeated. In total, listeners responded to 990 unique words (9 conditions x 22 trials x 5 test items) during the experiment.

Prior to the main experiment, listeners were given written guidance which encouraged them to think of the task in the same way as choosing an appropriate volume for the television: too quiet makes it difficult to understand the words, and too loud is uncomfortable. They then underwent a familiarisation phase consisting of 5 trials in each of Quiet, SSN and SMN (at a single SNR for the masked conditions). The entire experiment lasted around two hours, and

participants were able to take a break between each block. All instructions provided to the listeners were given in Spanish. Stimuli were presented through Sennheiser HD380 headphones at a fixed presentation level while listeners were seated in sound-attenuating booths in a purpose-build speech perception laboratory at the University of the Basque Country (Alava Campus).

For evaluating intelligibility, scores were computed based on the number of keywords correctly recalled in each trial (5 test words). Prior to scoring, all accents over vowels were removed. For example, both 'árbol' and 'arbol' were considered as correct responses for the word 'árbol'.

5.4 Results

Speech rate preferences

Figure 5.1 plots speech rate preferences, intelligibility and the time spent in the adjustment phase for the 9 conditions. In quiet, listeners preferred to listen to speech 1.2 times faster than the original rate, and at 97.7%, word scores were close to ceiling. Listeners spent 5.4 s during the adjustment phase in this condition, close to the 5 s minimum permitted. Compared to quiet, in noise listeners selected slower speech rates and spent longer on adjustment. Even at the various 'optimal' speech rates (i.e. those most-frequently selected) for the different masking conditions, listeners did not achieve intelligibility scores as high as those in the Quiet condition. Within each masker type (SMN, SSN) adjustment time was longer for conditions resulting in lower intelligibility. However this was not true across masker types; for example, less time was spent adjusting in the 0 dB SNR condition for the SSN masker than in some of the SMN conditions even though intelligibility was substantially lower in the 0 dB SNR SSN condition.

Separate one-way within-subjects ANOVAs conducted to compare the effect of condition on each of the three measurements indicated significant main effects on speech rate preferences [$F(8, 136) = 9.1, p < 0.001, \eta^2 = 0.09$], adjustment time [$F(8, 136) = 6.1, p < 0.001, \eta^2 = 0.14$] and intelligibility [$F(8, 136) = 124.7, p < 0.001, \eta^2 = 0.83$]. Post-hoc comparisons using the Tukey HSD test indicated that all conditions, relative to the Quiet baseline, resulted in significantly lower intelligibility, longer adjustment time (except for SSN at +6 and +12 dB SNR), and slower preferred speech rate (except SMN at the 2.5 times slower modulation rate). For the SSN conditions, these comparisons showed increasingly higher intelligibility with increasing SNR and for the adverse noise level significantly higher adjustment time and slower rate compared to the less noisy conditions. For the SMN conditions, the two slower masker modulation rates led to significantly higher intelligibility than the faster modulation rates. Significantly less time was needed to adjust speech rate when the masker's modulation rate was different than the original. Adjustment time in the face of modulated maskers was longer than for the stationary maskers apart from the least adverse SSN condition. Finally, listeners preferred significantly faster speech when the masker modulation rate was slow, with a tendency towards the converse when the masker modulation rate was fast.

Listener preferences and intelligibility

The probability with which each of the 22 permitted speech rate values was preferred by listeners, along with the percentage of keywords correctly recalled at that speech rate, is presented in Fig. 5.2 for the Quiet and SSN conditions and in Fig. 5.3 for the Quiet and SMN conditions (the Quiet condition is repeated for convenience). At the most adverse SSN condition (0 dB SNR) it

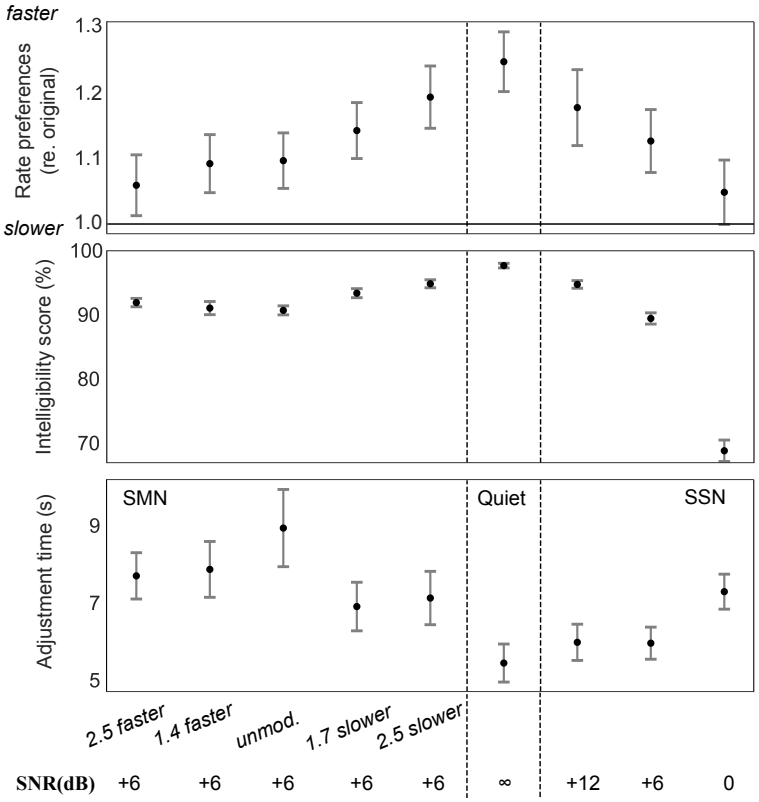


Figure 5.1: *Speech rate preferences (upper plot), intelligibility scores (middle plot) and adjustment time (lower plot) are depicted. Dashed vertical lines separate the SMN, Quiet, and SSN conditions. The solid horizontal line in the upper plot indicates the original speech rate. Error bars represent ± one standard error.*

is notable that those listeners who chose faster rates produced lower intelligibility scores, but in general intelligibility scores were relatively uniform across speech rates, though not necessarily at ceiling levels. Nevertheless, listeners showed distinct preferences for certain speech rates as manifested by the relatively sharply-peaked preference distributions.

Inter-rater reliability was determined using two-way intra-class correlation [McGraw and Wong, 1996] to assess the degree that listeners provided consistency in their mean speech rate choices across conditions, using the *icc* function of *irr* package in R. The resulting intra-class correlation value of 0.879 was in the 'excellent' range [Cicchetti, 1994], indicating that listeners had a high degree of agreement in selecting preferred speech rates.

Initial speech rate preference versus adjustment time

Repeated-measures correlation via the *rmcorr* package in R [Bakdash and Marusich, 2017] showed that the initial speech rate value of each trial was positively correlated with the final preference (i.e. a fast initial rate tended to lead to fast speech at the end of adjustment, and vice versa) [$p < 0.001$] and negatively correlated with adjustment time [$p < 0.001$]. In other words, when the initial speech rate was far from the 'optimal' value (considered as the mean value chosen by the cohort of listeners at the end of adjustment), listeners tended not to tune it all the way to this value.

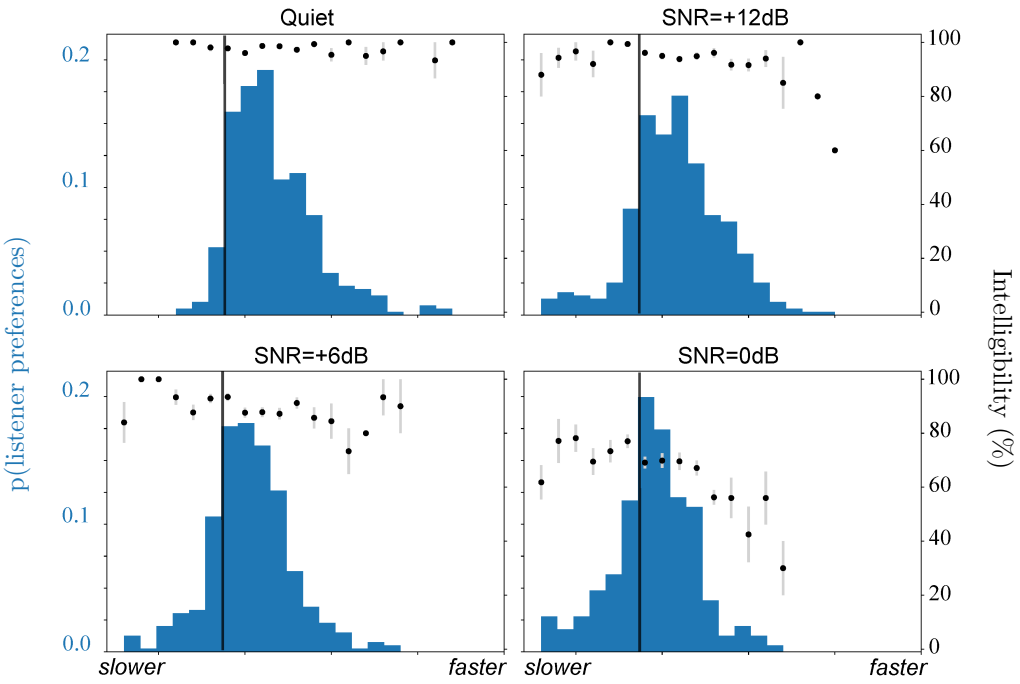


Figure 5.2: *Probability of each speech rate value preference (histogram, left axis) for the Quiet and SSN conditions, along with the percentage of words recalled correctly (black dots, right axis). Error bars represent ± one standard error. The black line denotes the step that corresponds to the speech rate of the original speech signal.*

Speech adjustments across trials

Figures 5.4 and 5.5 show the results of the three measurements (intelligibility, listener preferences, adjustment time) as a function of trial presentation order for the SMN and SSN maskers, respectively (the Quiet condition is repeated for convenience). For each participant, speech rate preferences and adjustment time were standardized (using z-score) across all conditions, in order to reduce individual trends such as slow/fast responders or always choose one response side i.e. low variability of chosen speech rate. For the Quiet condition, Spearman's rank-order correlation showed a negative correlation between trial number and adjustment time and between trial number and preferred rate [for both $r_S = -0.1$, $p < 0.05$]. This indicates that participants responded faster and chose faster rates when they were more familiar with the task while intelligibility was at ceiling regardless of trial number. For the SMN conditions, only the adjustment time of the unmodified and 2.5 times faster modulation rate conditions decreased monotonically with increased trial number [$r_S = -0.1$, $p <= 0.05$]. For the 2.5 and 1.7 times slower modulation rate conditions, listeners preferred slightly slower speech rates with increased trial number [$r_S = 0.1$, $p <= 0.05$] while for the 2.5 times faster modulation rate, slightly faster [$r_S = -0.1$, $p = 0.05$]. The results for the speech rate preferences across conditions showed that listeners preferred faster speech when the masker modulation rate was slow and the opposite when the modulation rate was fast. Combining this finding with the trial order results reveals that listeners decided to choose less extreme steps when they became more familiar with the task. For the SSN, the only monotonic relationship was revealed for the 0 dB SNR condition in which listeners' performance was slightly improved [$r_S = 0.1$, $p = 0.05$] with increased trial number.

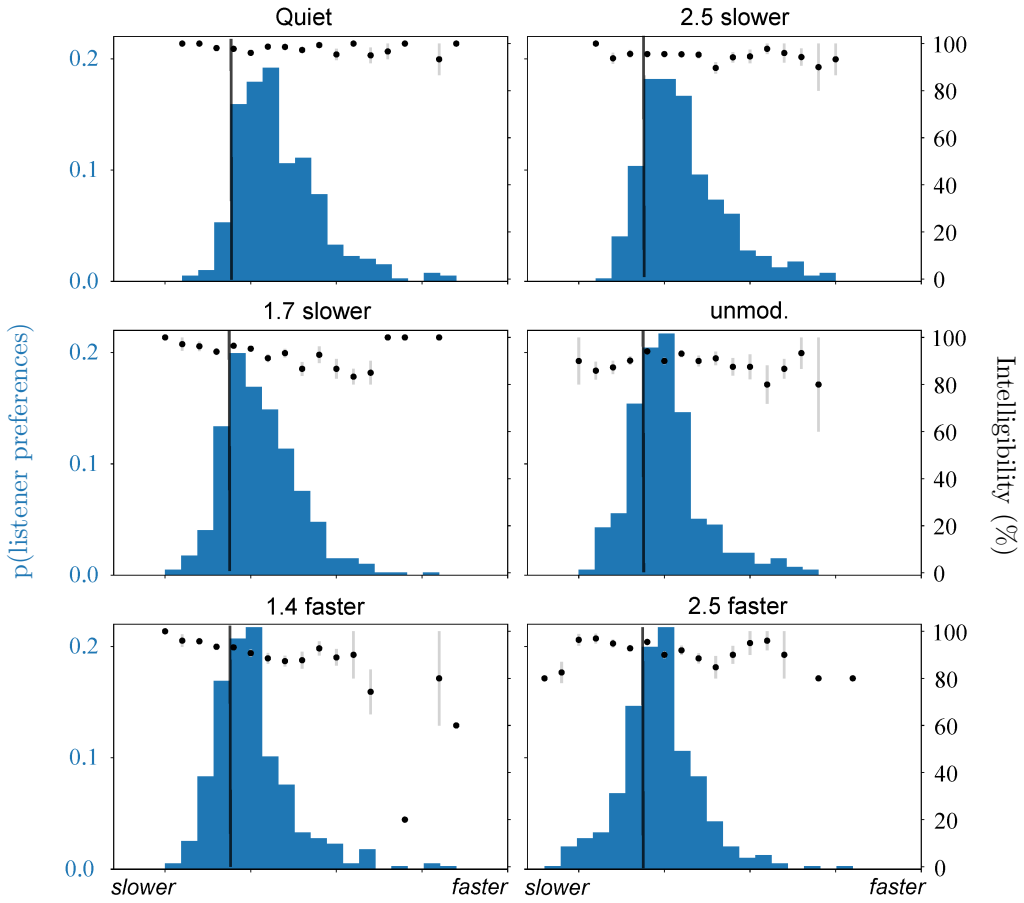


Figure 5.3: *As Fig. 5.2 but for the Quiet and SMN conditions.*

5.5 Discussion

In this chapter, listeners' speech rate preferences in stationary and temporally-modulated noise were explored. Findings revealed distinct speech rate preferences that showed up even with intelligibility at ceiling. Such preferences reveal supra-intelligibility aspects of speech rate, suggesting that in stationary noise listeners prefer a slower speech rate as noise level increases, while for fluctuating noise they prefer faster speech when masker modulation rate is slow and vice versa. These findings are in line with the real-time speech rate modification study of Novak III and Kenyon [2018], whose listeners chose to expand speech at adverse SNRs in the face of a 4-talker babble masker even though such preferences did not improve intelligibility. My findings also revealed that the preferred speech rate in Quiet was faster than any of the masking conditions. This might be due to the simple procedure of adapting to faster speaking rates in noise-free conditions [Adank and Janse, 2009].

RQ1: Does speech rate vary with regard to the noise level?

For stationary noise, it has been argued [Cooke and Aubanel, 2017] that linear elongation of speech resulting from a slower speech rate is not beneficial to intelligibility because it merely leads to elongation of those 'glimpses' of speech that escape masking rather than revealing additional speech information. The findings in this chapter support this claim: in general, intelligibility did not improve for listeners who chose slower speech (Fig. 5.2). The fact that listeners preferred slower rates in more adverse stationary noise might indicate a desire to reduce

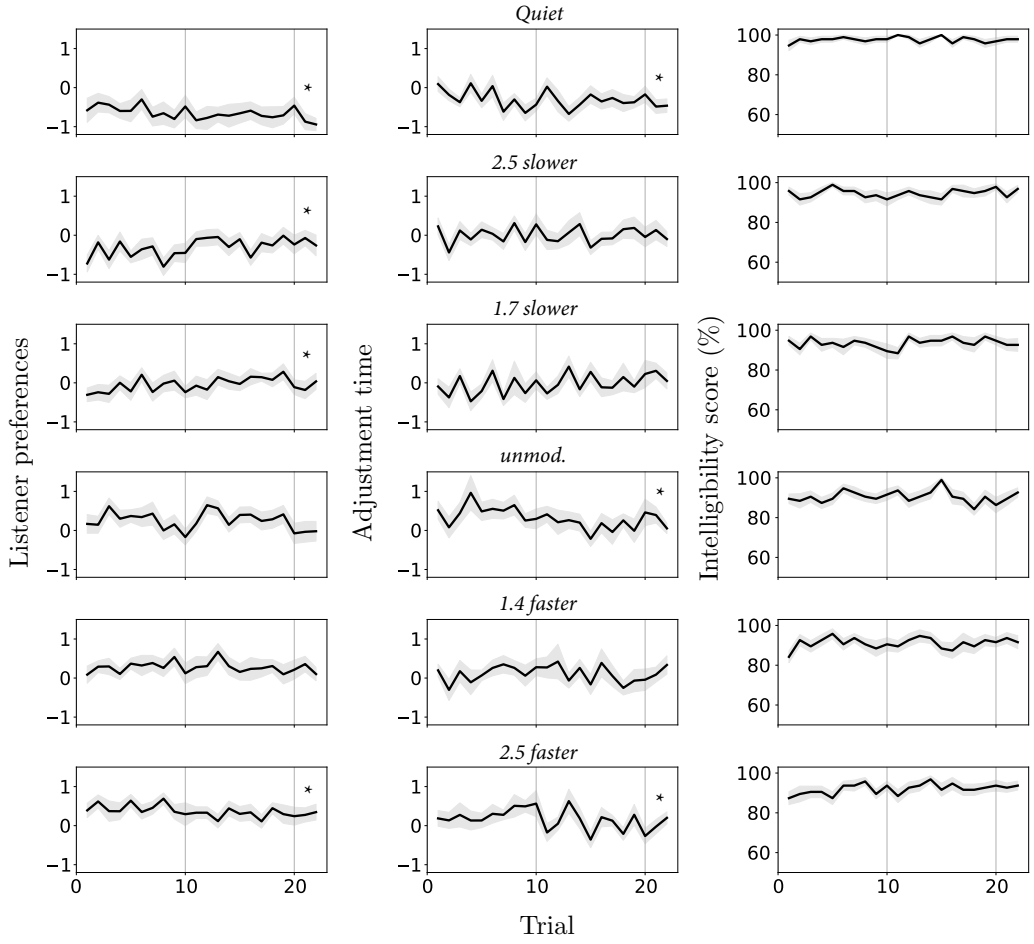


Figure 5.4: *The average of the z-score normalization of the listeners' preferences and adjustment time and the average of the percentage of the intelligibility scores across trials. Each row shows the results of the different masker's modulation rates of the SMN. Shaded areas correspond to the ± one standard error. Asterisks denote the significant correlation between the measurement and the trial number.*

listening effort since the listener has more time to process speech, or could reflect an attempt to reproduce typical speech rates experienced by participants in real-world noisy conditions which are characterised by slower speech [Tartter et al., 1993]. It is supported by Novak III and Kenyon [2018] who found a clear listener preference for decreased rates of speech as noise increased while degraded performance was revealed relative to unmodified speech in the same conditions. Thus, listener preferences criteria are above and beyond word recognition.

RQ2: Do listeners choose a different speech rate than that of the masker?

Concerning the speech-modulated noise conditions, listeners tended to prefer a target speech rate that contrasted with that of the masker. When speech is interrupted, listeners have the ability to track the speech and piece together the phoneme pieces in order to understand the speech [Miller and Licklider, 1950; Howard-Jones and Rosen, 1993]. Previous studies have showed that performance in steady-state noise is lower than in a single competing talker [Miller, 1947; Festen and Plomp, 1990]. It is attributed that the benefit in the latter condition is due to the dips which allows the listener to recognise the target speech supported by the following finding: at +6 dB SNR the intelligibility score for the SSN was lower than that for all the

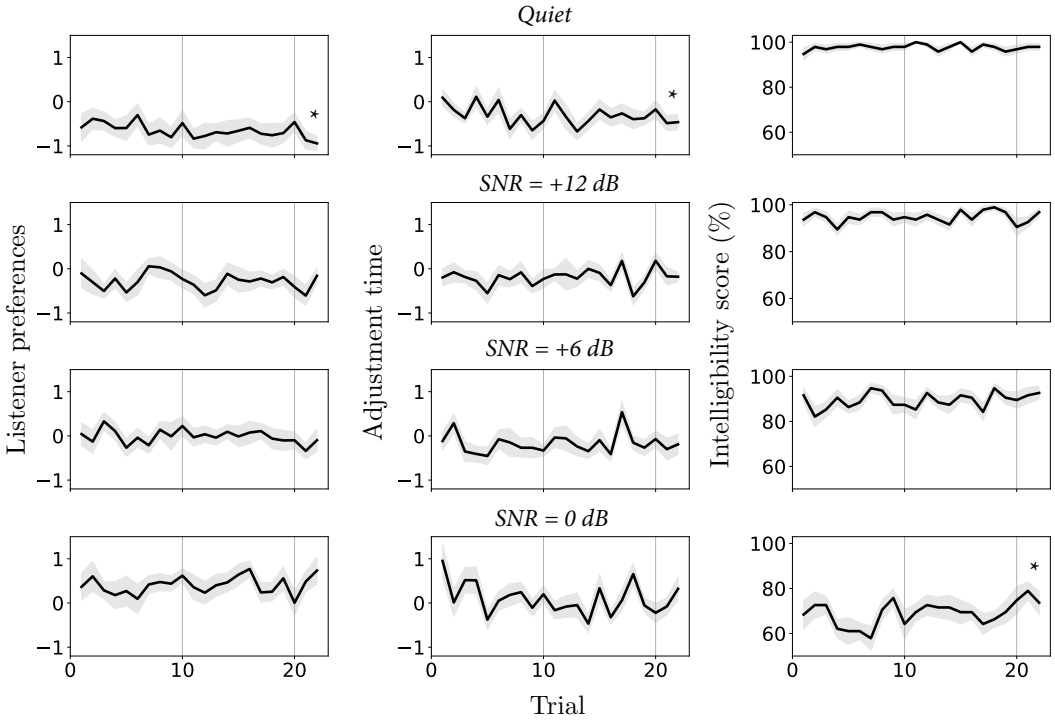


Figure 5.5: *As Fig. 5.4 but for the SSN masker.*

SMN variations. Additionally, Cooke and Aubanel [2017] found that in fluctuating masking condition, the artificial decrease of speaking rate is beneficial since by elongating the utterance, the amount of spectral glimpses increases. The effect of the acoustic and lexical factors on speech intelligibility vary with regard to the masker's modulations [Fogerty et al., 2021]. Additionally, differences in modulation rate might act as a cue for segregating the two signals, a possibility supported by that intelligibility improves when the fluctuation rates of target speech and the background speech masker [Gordon-Salant and Fitzgibbons, 2004] are mismatched. A contrast in modulation rates might help in a number of ways. One is to allow a target to be tracked through time by sequential grouping of those speech fragments with similar rates. Indeed, some listeners in Novak III and Kenyon [2018] reported that their speech rate choices helped them to track the target speaker. A complementary possibility is that listeners manipulate speech rate in order to promote energetic masking release. For example, a faster rate potentially allows more evidence of the target to 'fit' in the longer temporal dips of a masker with slow modulations. There is some evidence that talkers adopt such a strategy when 'listening-while-speaking' [Cooke and Lu, 2010]. Apart from the faster speech rates, the potential longer 'glimpses' for the slower modulation rates of the masker benefited the listeners in terms of word recognition and response time (i.e. higher intelligibility scores and faster adjustment times).

Listeners spent more time adjusting the target speech rate in the presence of temporally-modulated noise.

The modulated nature of the masker allows varying amounts of target speech energy to be audible at different points, and it is possible that this causes additional cognitive load for the listener and makes it harder to predict when to listen. In Larsby et al. [2005], more perceived effort was reported for noises with a high degree of temporal variation at a relatively high SNR (+10 dB). In Krueger et al. [2017], at low SNRs listeners rated stationary masking as more

effortful than fluctuating noise, but the difference between the two types of masker was reduced or eliminated with increasing SNR, leading to the suggestion that peaks in the fluctuating masker might have a negative impact on listeners in less noisy conditions. Finally, for the Quiet condition, listeners spent the least time for adjusting the speech rate. This can be explained from the fact that listeners use effective cognitive strategies which allow them to adapt quickly to time-constrained speech [Dupoux and Green, 1997]. Also Koch and Janse [2016] examined the effect of speech rate on listening effort in quiet condition and results revealed that differences in speaking rate did not affect listening effort using pupil dilation measures.

Preferred rate was faster than the original speech for all conditions.

Finally, we note that although listeners preferred reduced speech rates in adverse conditions, in all conditions the mean rate chosen was faster than the original speech (Fig. 5.1). This is most likely due to the use of read speech, which is typically somewhat slower than normal or casual speech [Koopmans-van Beinum, 1991]. The findings of this study can be relevant for speech enhancement algorithms to improve speech based on speech rate preferences for different masking conditions.

Chapter 6

Listener preferences - Fundamental frequency

6.1 Introduction

Speakers often adopt different speaking styles when communicating in order to adapt their voice to the current situation. Such situations include talking in a noisy environment and communicating with a specific group of interlocutors (e.g. speech directed to infants, the hearing-impaired or non-native listeners). Among other features that a talker changes in such situations are features related to prosody e.g. fundamental frequency (f_0). Prosody plays an important role in communication [Cutler et al., 1997; Wagner and Watson, 2010], since it can reveal the talker's intention and emotions, while it also functions as an attentional cue and allows salient information to be emphasised. A higher f_0 is observed with increasing vocal intensity [Summers et al., 1988; Bond and Moore, 1990], or when speaking to infants [Bradlow et al., 2003] or pets [Burnham et al., 2002], while no such increase has been found when speaking to non-native listeners [Uther et al., 2007]. Though f_0 characteristics vary across natural speaking styles [Boril and Pollak, 2005; Mayo et al., 2012] and amongst different speakers [Bradlow et al., 1996; Barker and Cooke, 2007], intelligibility benefits related to f_0 are currently unclear.

Many studies have focused on the impact of f_0 on intelligibility. Previous research has shown that, under quiet and stationary noise conditions, changes in mean f_0 alone do not facilitate intelligibility [Assmann et al., 2002; Lu and Cooke, 2009a]. However, the absence of f_0 variation can lead to poorer intelligibility [Wingfield et al., 1984; Laures and Weismer, 1999; Laures and Bunton, 2003; Watson and Schlauch, 2008]. Different studies have shown that the advantage of f_0 is more evident in the presence of a competing talker [Bird and Darwin, 1998; Assmann, 1999], since f_0 modifications benefit the segregation of target and background sources. Increases in mean f_0 difference between target and competing talker lead to improvements in identification accuracy [Brokx and Nooteboom, 1982; Bird and Darwin, 1998; Assmann, 1999]. Varying f_0 with time might also help intelligibility, since momentary differences in f_0 occur [Bregman, 1990]. However, in Assmann [1999], f_0 variation did not result in intelligibility improvements.

Rather less research has looked at listener preferences. In Assmann et al. [2006], listeners judged the naturalness of frequency-shifted (f_0 and formant frequency shifts) sentences in quiet by adjusting a graphical slider on a range from highly unnatural to definitely natural. Listeners judged sentences as more natural when f_0 and formant frequencies were 'matched' (low f_0

with low formant frequency values and vice versa), similar to natural speech. In Assmann and Nearey [2007], listeners had to adjust f_0 and formant frequencies to their preferred levels using a self-paced adjustment procedure. Listeners adjusted either mean f_0 or mean formant frequencies of vowel triads (vowel triads were formed by concatenating the vowels /i/, /a/, /u/ extracted from /hVd/ words, e.g. hod) by moving the computer mouse to the left for lower f_0 (or formant frequencies) and to the right for higher. Listeners were asked to search for the most natural-sounding voice. The results suggested that listeners preferred an f_0 and a pattern of formant frequencies similar to those of natural speech.

In this chapter, we test two main hypotheses. The first hypothesis is that if naturalness is a critical factor for the listener preferences [Assmann et al., 2006; Assmann and Nearey, 2007], listeners will choose the f_0 features to be close to those of the original speech. The second hypothesis is that in the presence of competing speech listeners will choose the f_0 features to differ from the maskers since this can lead to improvements in identification accuracy [Brokx and Nooteboom, 1982; Bird and Darwin, 1998; Assmann, 1999]. This chapter describes how SpeechAdjuster was used to explore listener preferences in a choice of f_0 characteristics. Two experiments were conducted to investigate listeners' mean f_0 and f_0 variation preferences in energetic (Expt. I sec. 6.2) and informational (Expt. II sec. 6.3) maskers. This extends the work of Assmann and Nearey [2007] to preferences in noise using sentence material (meaningful sentences instead of vowel triads).

The main questions addressed in the current chapter are: (1) do listeners choose different modifications of f_0 features (mean and variation) for different conditions; and (2) do listeners make their choices based on aspects beyond intelligibility?

6.2 Experiment I: Listeners' f_0 preferences for speech presented in conditions of energetic masking

6.2.1 Methods

Listeners

Seventeen Greek monolingual listeners (10 female) participated in the experiment. All were young adults in the age range of $19-33$ (mean 24.2 years; *S.D.* 3.8 years). Listeners reported no known hearing problems. Thirteen of the participants reported good to excellent knowledge of English and nine reported extensive music studies. An incentive of 10 euros was given for participation.

Stimuli

Sentence material

A Greek corpus [Sfakianaki, 2019] provided sentence material for the experiments. The corpus consists of 720 semi-predictable sentences in modern Greek with a similar level of difficulty to that of the original English Harvard sentences [Rothauser et al., 1969]. From this point on, the corpus will be referred as GrHarvard. The number of words in a sentence varies from 5 to 9. Each sentence contains exactly 5 keywords. For the sentence design, meaningful words resembling everyday language were used. An example is 'Θα **κόψω** το **φρούτο** σε **τρία** **ίσα**

μέρη.' ('I will cut the fruit into three equal pieces.'); the keywords are indicated with bold letters.

Speech material

A 31-year-old native Greek male talker was recruited to read the complete GrHarvard corpus. The talker was asked to read each sentence at a normal speaking rate and was able to repeat any utterance if necessary. The talker's original mean f_0 (computed using all the voiced segments of the 720 sentences) was around 130 Hz (*S.D.* 20 Hz).

The recordings took place in a sound studio at the Speech Signal Processing Laboratory, University of Crete, in Heraklion, Greece. The sentences were recorded using Pro Tools 12 software with an RME Fireface 400 recorder. A Neumann KMS104 handheld vocal condenser microphone (cardioid directional polar pattern) was placed on a desktop microphone stand, on a table at a fixed distance of 15 cm from the talker's mouth. The recordings were made at a sampling rate of 44.1 kHz. Sentences were segmented using an amplitude-based pause detector based on the normalised envelope of the signal. The algorithm's effectiveness and the quality of the recordings were screened manually. More specifically, signals were checked for clipping, if utterances were properly split, and if all utterances were of the same speaking style. In cases of recordings with issues, the utterances were recorded again. The recorded sentences had a mean duration of 2.8 s (*S.D.* 0.3 s). For the experiments, spoken phrases were downsampled to 16 kHz and a 20 ms half-Hamming ramp was applied at the beginning and end of each recording. Finally, each stimulus was normalised to the same root-mean-square level.

Stimulus preparation

In the experiment, listeners were allowed to perform modifications on the mean f_0 for half of the trials, while for the remaining trials they were allowed to modify the f_0 variation of the target speech. The talker's original f_0 was modified ($f_0{}'$) using the following formula (eq. 6.1).

$$f_0{}' = \frac{f_0 - \mu}{\sigma} \cdot \sigma' + \mu' \tag{6.1}$$

where μ and σ are the mean and standard deviation of f_0, respectively. The desired mean and standard deviation of $f_0{}'$ are denoted by μ' and σ'. Changes in mean f_0 were performed with a simple shift in the entire contour, keeping the f_0 variation constant ($\sigma' = \sigma$). Similarly, when changing the f_0 variation, mean f_0 was kept constant ($\mu' = \mu$) as described below. Pitch modification was performed using PSOLA [Charpentier and Stella, 1986].

For both features tested (mean f_0 or f_0 variation), there were 25 available steps amongst which the participant could choose. The same number of steps was used so that the participants would not be aware which feature was being tested. Each block consisted of trials testing both features. Previous research has showed that listeners prefer f_0 values close to the original voice [Assmann and Nearey, 2007]. Using exponential growth, listeners were provided with more modification options close to the original pitch. Thus, the increments to the talker's original mean f_0 values followed an exponential curve, as shown by function 6.2.

$$\mu' = \mu + (m + k * (1 + r)^t) \tag{6.2}$$

where $m = -65$ is a correction term, $k = 250$ is the starting value and $r = -0.1$ is the growth rate of the values as $t = [0 : 1 : 24]$ changes in discrete intervals with 25 steps. The terms m and k were chosen so that the mean f_0 of the specific talker would not attain values lower than 80 Hz or very high values that would result in a greatly unnatural voice. The upper plot of Fig. 6.1 shows the f_0 contour of an utterance for each of the 25 steps.

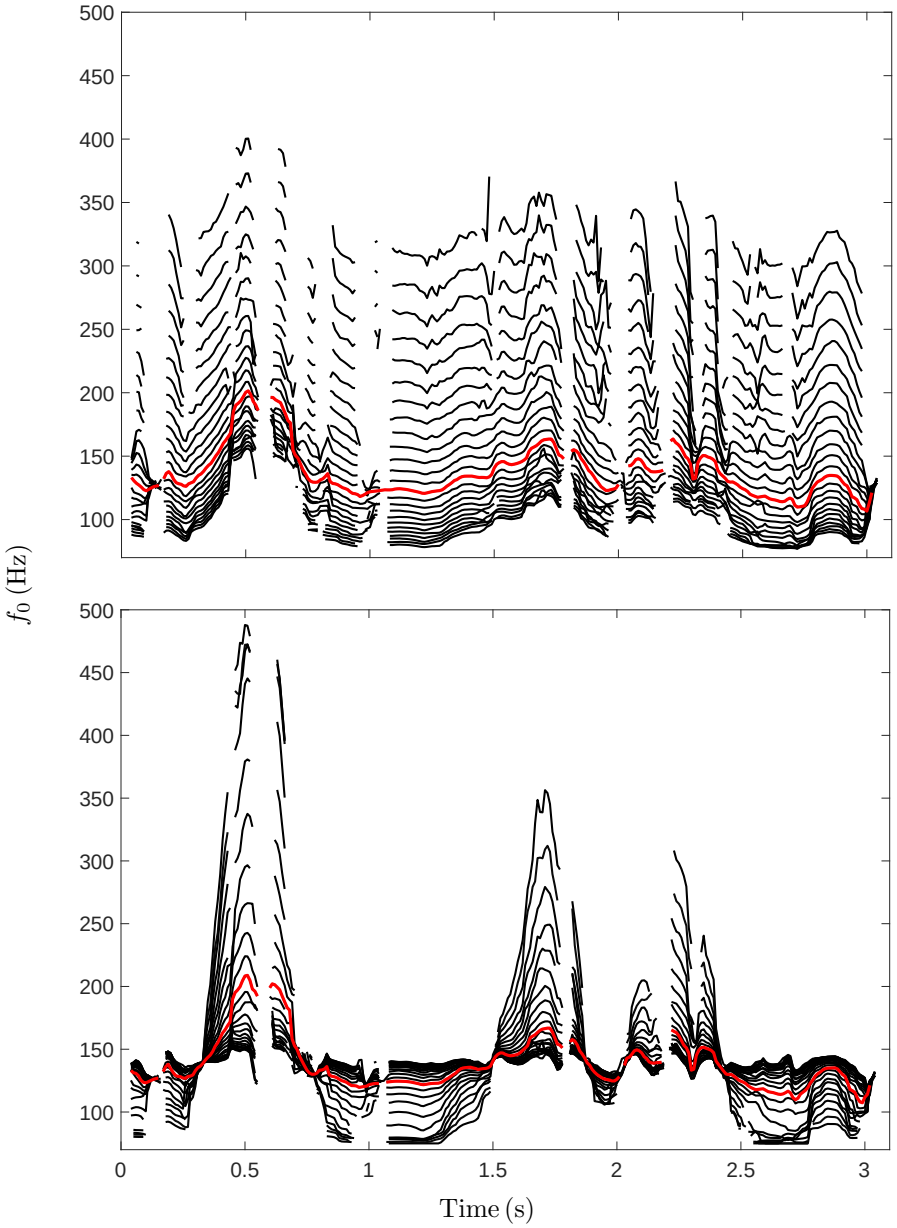


Figure 6.1: *f_0 contours of the sentence 'Οι φιγούρες από χαρτόνι φάνηκαν πίσω από τον μπερντέ'. The mean f_0 (upper plot) and f_0 variation (lower plot) modifications of the 25 steps are depicted. The red lines denote the original mean f_0 value (step= 14) and to the original f_0 variation value (step= 11)*

The changes in f_0 variation were derived as a consequence of stretching the f_0 range. The f_0 range was expanded based on the following function (eq. 6.3).

$$\sigma' = \sigma \cdot (\delta + k * (1 + r)^t) \tag{6.3}$$

where $\delta = 1e-6$ is an offset set to a small non-zero value to prevent f_0 variation going to zero, $k = 10$ is the starting value and $r = -0.2$ is the growth rate of the values as $t = [0:1:24]$ changes in discrete intervals with 25 steps. The lower plot of Fig. 6.1 shows the f_0 contour of an utterance for each of the 25 steps. Any f_0 values lower than 75 Hz or higher than 500 Hz were mapped linearly in the range $[75, 80]$ or $[450, 500]$, respectively. All f_0 modifications and values selected for the formulas were chosen based on pilot experiments to ensure that the speech produced did not have any audible artefacts or sound unnatural.

The sentences were presented in quiet, or mixed with speech-shaped noise (SSN) at -3, 0 and $+3$ dB SNR. The masker was generated by filtering random uniform noise with the long-term spectrum of the 720 concatenated sentences (without gaps) of the GrHarvard corpus. The desired SNRs were obtained by rescaling the noise. From the GrHarvard corpus, the sentence IDs used in this experiment were $350-575$ for the adjustment phase, $576-656$ for the test phase, and $714-720$ for the practice session.

Energetic masking measures

In order to examine the impact of f_0 modifications on energetic masking, the extended glimpsing model was used [Tang and Cooke, 2016]. The extended glimpsing model computes the glimpses, i.e. spectro-temporal regions, where the target energy exceeds the masker energy and augments the original glimpsing model by taking into account the absolute hearing level and durational changes, and by compressing the output values into the range $[0-1]$ (Fig. 6.2).

A new measure was introduced for determining the glimpse distribution of an utterance across frequencies (DGAF). Specifically, DGAF is the mean of glimpses across the time series of an utterance for each different frequency band (Fig. 6.3), i.e. a form of 'glimpse spectrum'. This new representation gives the overall spectral picture of the masked speech signal and provides more information than just the spectral tilt of the glimpses. It can be useful for speech enhancement algorithms that do not perform enhancement at the phoneme level, such as the Automatic Sound Engineer [Chermaz and King, 2020]. Specifically, it provides insights into which speech spectral bands need to be enhanced under different noise conditions, or how the energy could be redistributed in cases where it is concentrated in bands that are not perceptually important for the listener.

Speech stimuli were normalised so the total energy before each modification to be equal to the total energy after the modification i.e. same root-mean-square level. In other words, every speech modification represents a tradeoff between the effect of the parameter being modified (e.g. mean f_0) and the effect on local SNR change in time-frequency. Thus, the choice reveals the best 'compromise' step for the listener. The DGAF measure can help in interpreting this 'compromise' by computing the mean spectral glimpses that have been affected by the energy reallocation after speech normalisation.

Statistical analysis tools

Since not all the data in the different conditions were normally distributed, non-parametric statistical tests were used. All tests were performed in Python using functions of the *stats.scipy* library (shown in parentheses below). Differences among the experimental conditions were tested using the rank-based, Kruskal–Wallis H-test (*kruskal*). Post-hoc comparisons were performed using Dunn's test (*posthoc_dunn* is part of the *scikit_posthocs* library). For testing whether f_0

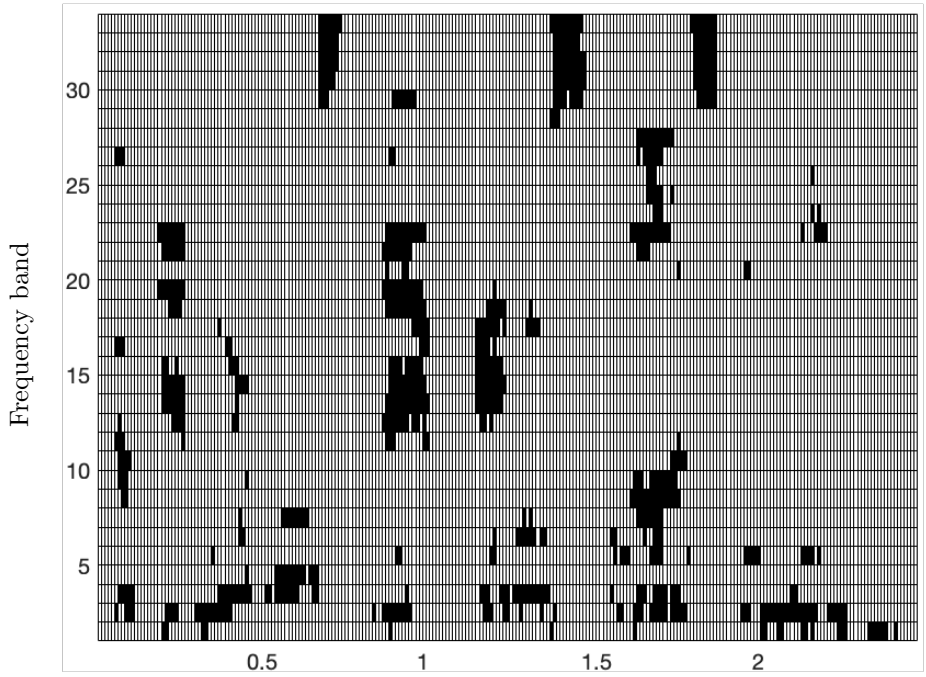


Figure 6.2: *Black areas denote the speech energy that is predicted to have survived energetic masking (glimpses). The glimpses are plotted for the phrase 'Το κανό γλιστρά πάνω στις λείες σανίδες' ('The canoe slides on the smooth planks') in time (x-axis) and across 34 frequency bands (34 equivalent rectangular bandwidths, ERB-rate scale, with filterbank frequency started at 75 Hz, y-axis). The masker was SSN at −3 dB SNR.*

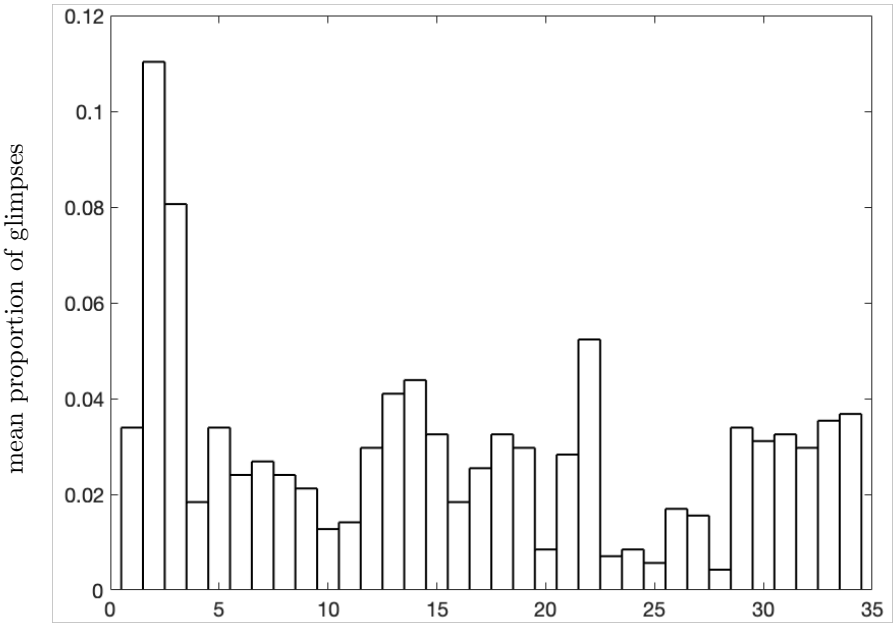


Figure 6.3: *Mean proportion of glimpses (y-axis) across the 34 frequency bands (x-axis) for the phrase used in Fig. 6.2. The generated distribution is the glimpse distribution of an utterance across frequencies (DGAF).*

preferences differed significantly from critical values (e.g. the talker's original mean f_0), the one-sample Wilcoxon signed rank test (*wilcoxon*) was used. Holm's correction [Holm, 1979] was used to adjust $p - values$ for multiple comparisons. Finally, the Kullback–Leibler divergence [Kullback and Leibler, 1951] was used for measuring the distance between two probability distributions (*entropy*).

6.2.2 Procedure

Each of the two experiments was divided into 4 blocks by condition (quiet and masked at 3 SNRs), with each block containing 5 trials in which listeners were allowed to modify mean f_0, and 5 trials for modifying f_0 variation. The presentation order of the 10 trials was random. Each trial consisted of an adjustment phase followed by a test phase. In the adjustment phase, phrases (with a 0.5 s gap between sentences) were presented in random order. Participants had to listen to at least 5 s of speech before proceeding to the test phase, but could listen to as much speech during the adjustment phase as desired. In the test phase, intelligibility was evaluated with a speech perception task using the f_0 value chosen at the end of the adjustment phase. Participants listened to two sentences separately during the test phase (the average of which was used for the statistical analysis). Participants typed what they heard into an on-screen text box. The tested phrases were presented only once. Prior to the experiment, all participants underwent a task familiarisation phase consisting of 3 trials, 1 in quiet and 2 in noise.

For the experiment, SpeechAdjuster and instructions similar to the ones described in chapter 4 were used. Listeners were asked to tune the speech in real time until they could recognise as many words as possible. Real-time changes could be made using the up/down keys on the keyboard while listening to sentences, so listeners would not to be influenced by their previous choices. The task was explained as akin to choosing an appropriate volume for the television: too quiet makes comprehension difficult, while too loud leads to discomfort. All information was provided to the listeners orally and written in Greek. As explained in chapter 4, the pair of arrows option does not give to the listener any visual indication of the feature step changes. The only indication appears when an extreme step is reached, informing the listener with an onscreen message.

A MacBook Air computer was used to run the SpeechAdjuster software. 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 would be a comfortable level of listening. Across participants, block order was counterbalanced using a Latin square design. Experiments on average lasted around one hour and participants could have a short break at the end of each block. The experiments took place in a sound-proof room at the Speech Signal Processing Laboratory, University of Crete, in Heraklion, Greece.

To evaluate intelligibility, scores were computed based on the number of keywords correctly recalled in each trial (2 test phrases x 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. Thus, keywords were considered as correct if all a word's letters were matched.

6.2.3 Results

Listener f_0 preferences

Figures 6.4 and 6.5 plot f_0 preferences, intelligibility scores and the time spent in the adjustment

phase for the 4 conditions (Quiet and SSN at 3 SNRs). For the mean f_0 modifications (Fig. 6.4), listeners did not show any particular trend across the different conditions, while as noise level increased, intelligibility scores decreased and listeners needed more time in the adjustment phase. For the f_0 variation feature (Fig. 6.5), similar results are observed. However, it can be observed that by controlling the f_0 variation listeners were able to achieve higher intelligibility score at 0 dB SNR (almost 100% correct responses) compared to the mean f_0 modifications. For both f_0 features, listeners in general preferred speech with lower than the original mean f_0 and f_0 variation and spent on average more than 20 s during the adjustment phase in all conditions.

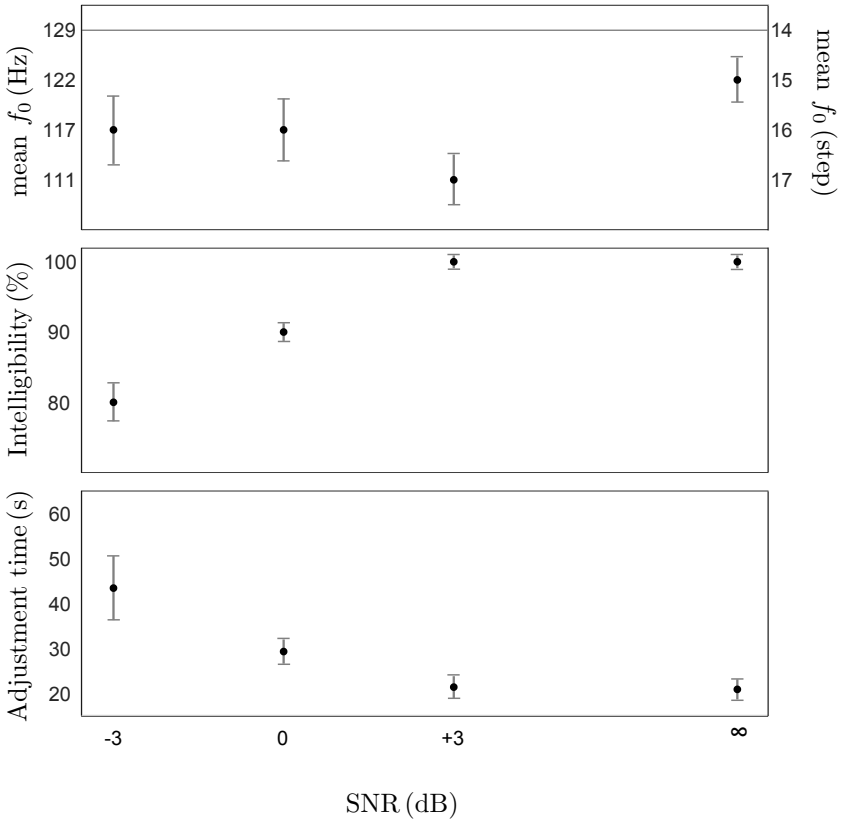


Figure 6.4: *Median values (black dots) of mean f_0 preferences (upper plot), intelligibility scores (middle plot) and adjustment time (lower plot) for the different conditions are depicted. The horizontal line in the upper plot indicates the original mean f_0 value. The error bars represent ± standard error of median. For the upper plot, statistics were computed based on the steps (right axis) and values on the left axis show the f_0 values corresponding to the steps.*

A rank-based, Kruskal–Wallis H-test was conducted to compare the effect of conditions on each of the three measurements and each of the two tested features. Results indicated significant main effects only for adjustment time (mean f_0 [$H = 53.90, p < 0.001$]; f_0 variation [$H = 64.11, p < 0.001$]) and intelligibility (mean f_0 [$H = 67.15, p < 0.001$]; f_0 variation [$H = 82.71, p < 0.001$]). Post-hoc pairwise comparisons for both f_0 features indicated that adjustment times were significantly different for the different conditions, except for Quiet and +3 dB SNR. This was also true for the intelligibility scores for the mean f_0 feature. Finally, for the f_0 variation feature, only the intelligibility at −3 dB was significantly different from the other SNRs.

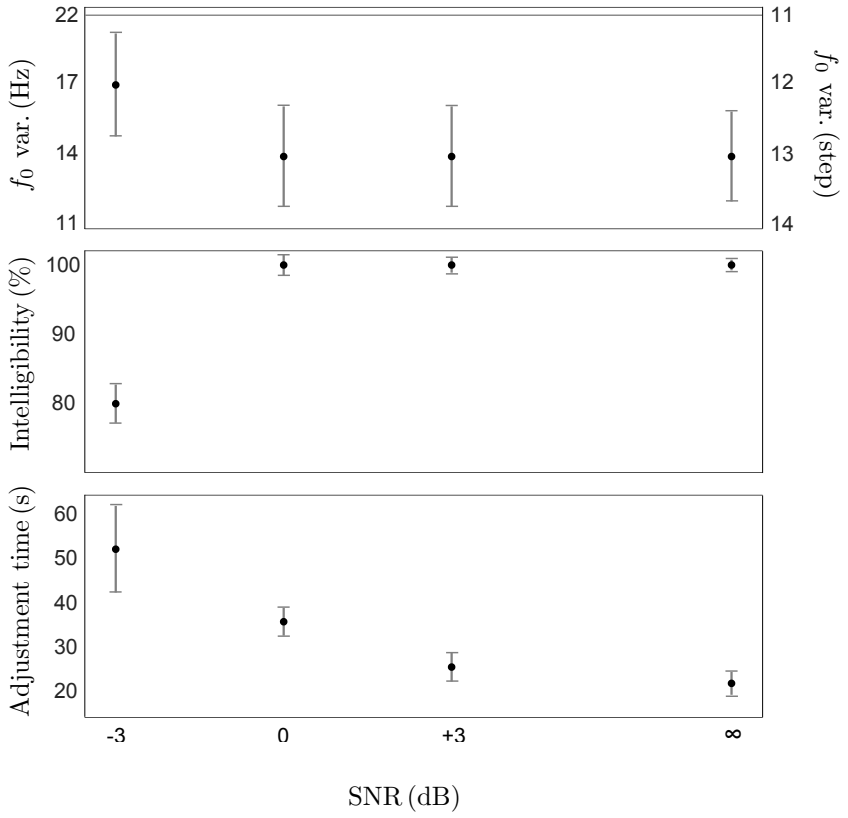


Figure 6.5: *As Fig. 6.4 but for the f_0 variation feature.*

Distribution of listener preferences vs intelligibility

The probability with which each of the 25 permitted mean f_0 and f_0 variation steps were preferred by listeners, along with the percentage of keywords correctly recalled, are presented in Fig. 6.6 and 6.7, respectively. In quiet, listeners had distinct f_0 preferences, even though intelligibility was always at ceiling. For the most adverse condition (-3 dB SNR), it is noticeable that those listeners who chose mean f_0 or f_0 variation close to the original had poorer intelligibility compared to the rest. For mean f_0, poorer intelligibility was also found for choices lower than the original values. However, in general most listeners preferred a mean f_0 slightly lower than the original.the original values. However, in general most listeners preferred a mean f_0 slightly lower than the original.

A one-sided, one-sample Wilcoxon test was performed to test whether the talker's original f_0 mean and variation were significantly lower compared to that preferred by listeners for each different condition (in total 8 tests performed; 2 features x 4 conditions). In all conditions except for f_0 variation at -3 dB SNR, the preferred f_0 steps were significantly greater than the talker's original f_0 values (mean f_0 in Quiet [$T = 2003, p < 0.001$]; $+3$ dB SNR [$T = 2357, p < 0.001$]; 0 dB SNR [$T = 2255, p < 0.001$]; -3 dB SNR [$T = 2478, p < 0.001$] and f_0 variation in Quiet [$T = 2390, p < 0.001$]; $+3$ dB SNR [$T = 2345, p < 0.001$]; 0 dB SNR [$T = 2384, p < 0.001$]).

f_0 choices and energetic masking

Figures 6.8 and 6.9 show the spectral areas where the speech energy survives energetic masking for the different noise conditions and f_0 feature steps. To better understand the importance

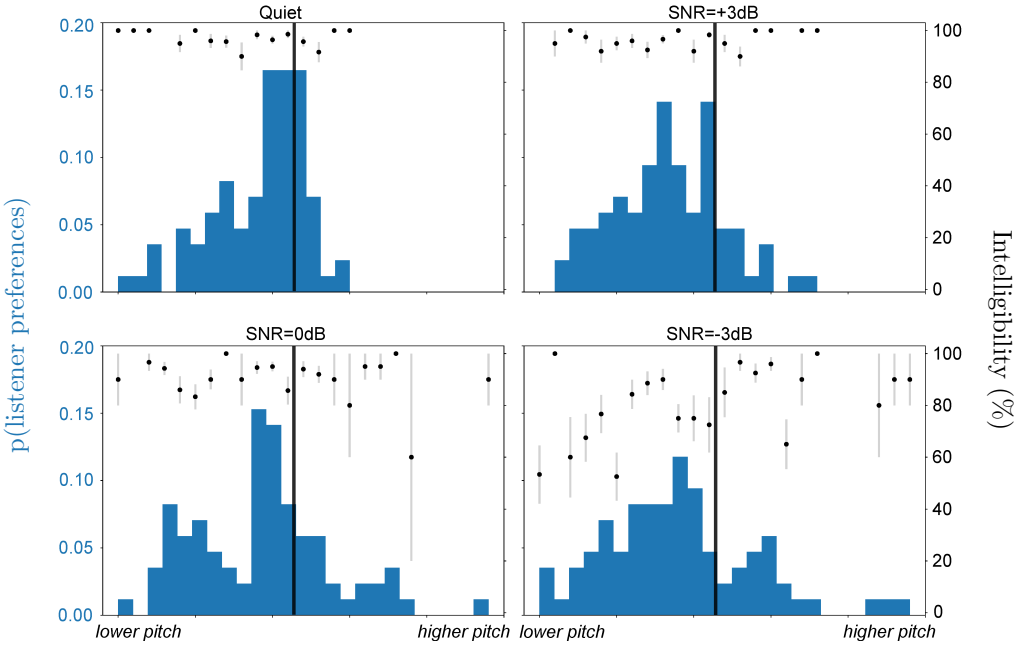


Figure 6.6: *Probability of each mean f_0 value (histogram, left axis), along with the percentage of words recalled correctly (black dots, right axis). The error bars represent ± standard error. The black vertical line denotes the step that corresponds to the original mean f_0.*

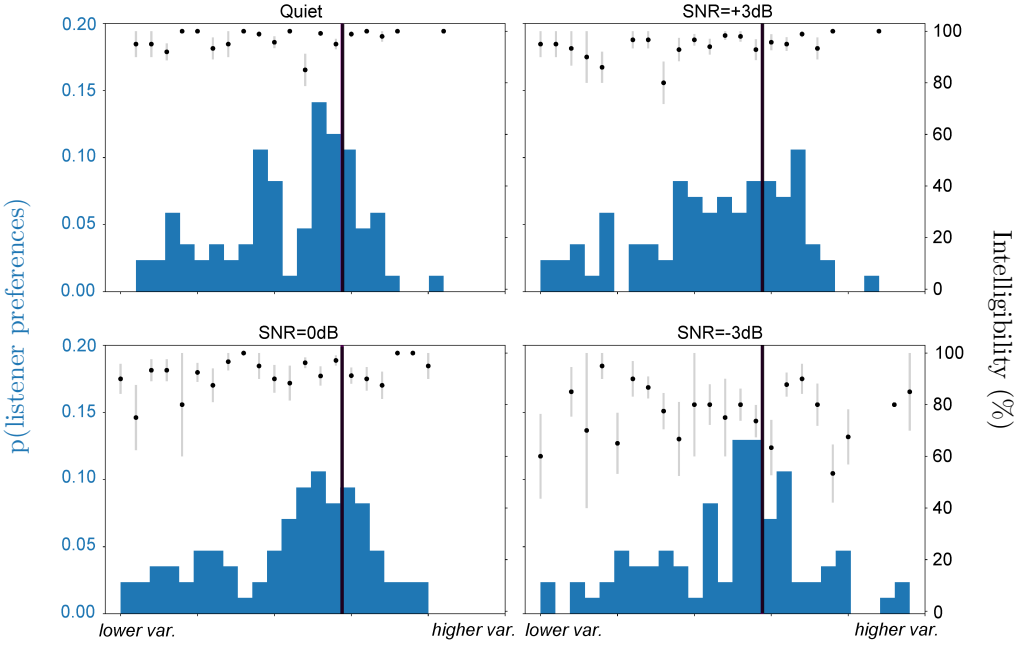


Figure 6.7: *As Fig. 6.6 but for the f_0 variation feature.*

of the spectral energy distribution in relation to the listener's preferences, the DGAF measure, described in sec. 6.2.1 'Energetic masking measures' was used. Heat maps with the DGAF of the different feature steps were computed, which allows easier comparisons and may provide insights for the interpretation of listener preferences. In the plots, the black colour denotes that glimpses are more concentrated at those frequencies while white denotes the opposite.

The masker in this experiment was SSN, generated with the long-term speech spectrum of the target talker. Thus, most of its energy is concentrated in frequencies below 1000 Hz. As

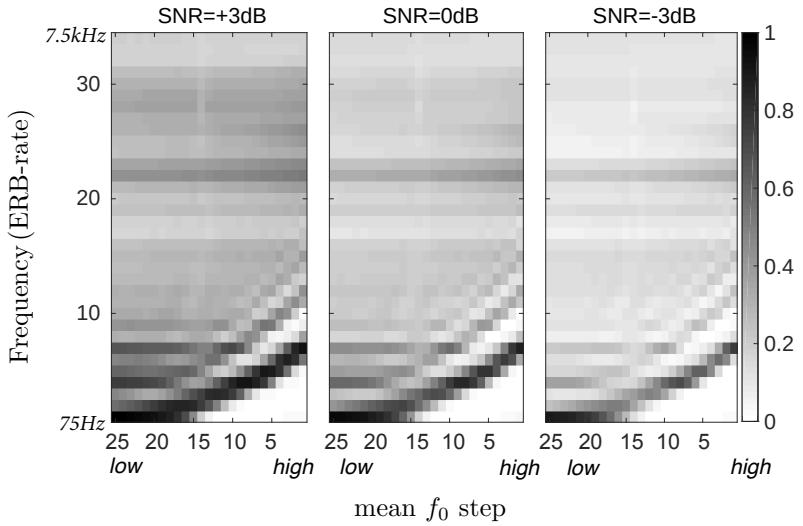


Figure 6.8: *The sum of the DGAF (see Fig. 6.3) of the 80 utterances from the test phase was computed for the ERB-rate scale (in total 34; y-axis) and for each of the 25 mean f_0 steps (x-axis). Plots are normalised with the maximum value derived from the 3 conditions (subplots). The colour bar denotes the mean amount of glimpses of all the utterances normalised with the maximum value. The black colour implies that in this frequency area a greater amount of speech energy exceeds energetic masking compared to the remaining areas, while the white colour means the opposite. low and high on the x-axis denote the lowest and highest pitch, respectively. The y-axis, shows the highest and lowest frequencies that correspond to the respective ERB-rates.*

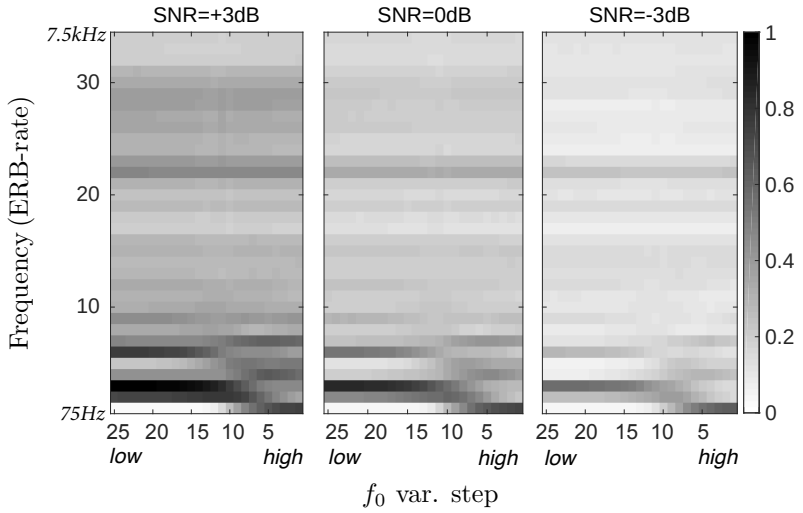


Figure 6.9: *As Fig. 6.8 but for the f_0 variation feature. low and high on the x-axis denote the lower and higher f_0 variation, respectively.*

expected, DGAF heat maps reveal that for lower pitch values, the target speech energy that survives masking is greater for lower frequencies, while the energy in low frequencies decreases as pitch increases (Fig. 6.8). Additionally, in Fig. 6.8, for the different steps it can be observed that the energy that escapes masking is the greatest close to each step's pitch value. This is true for all steps of SNR +3 dB, while for the SNRs for which the noise level is greater than or equal to the target speech level, this phenomenon declines for ERB-rates close to the target talker's mean f_0. This corresponds to the ERB-rates 2 (106 Hz) and 3 (141 Hz), since the talker's mean f_0 was around 130 Hz (step 14).

		mean f_0		f_0 variation	
		DGAF (ERB-rate 2)	GP_{ext}	DGAF (ERB-rate 2)	GP_{ext}
SNR(dB)	−3	0.64	2.71	1.75	2.14
	0	0.92	3.28	0.32	0.34
	+3	0.43	1.46	2.04	2.13

Table 6.1: *The symmetric Kullback–Leibler Divergence (KLD) derived from a comparison between the listener preferences distribution with DGAF of ERB-rate 2 distribution and GP_{ext} distribution for the different SNRs. The lower the KLD value, the closer the two distributions are. If the KLD value equals zero, the two distributions are identical.*

For the f_0 variation feature (Fig. 6.9), at low frequencies the number of values relative to glimpses decreases (white regions) when f_0 variation decreases (steps $11 - 25$ or f_0 variation values $21.5 - 0.94$ Hz). Step 11 was the original f_0 variation. The plots show that, for higher variations, the DGAF increases for ERB-rate 1 (75 Hz) and decreases for ERB-rates 2 and 3 (frequencies around 106 and 141 Hz, which are also the ERB-rates closest to the original pitch), while the opposite happens for lower variations. This was expected, since high f_0 variability results in having more frequency components with low energy, in contrast to low variability, which makes the peaks from harmonics more prominent. For the most adverse noise level, listeners might have chosen the f_0 variation of step 12, closest to the original, which allows speech energy to escape.

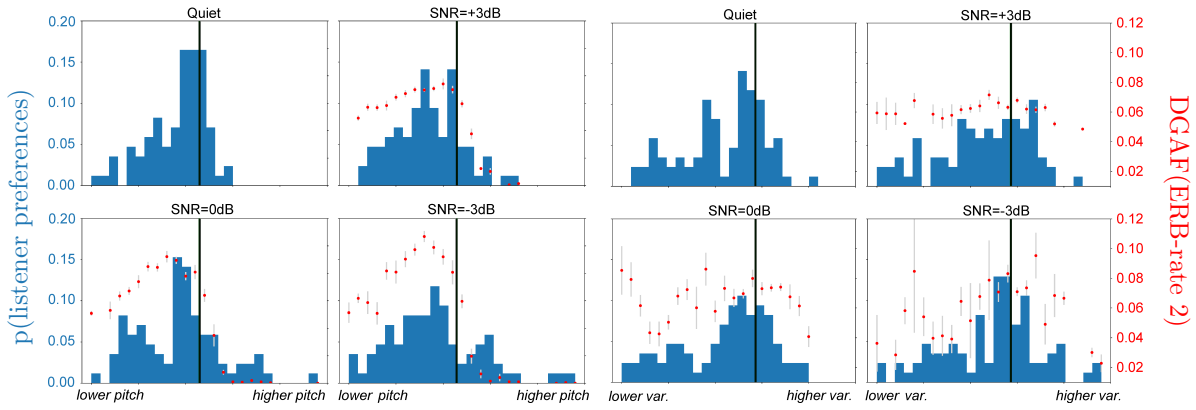
For both features, as the noise level increases, the black areas on the plots are fewer, implying that a higher amount of target speech energy is masked by noise. For both features and for all conditions, DGAFs in channels above the 15^{th} ERB-rate (or 970 Hz) do not vary much with regard to f_0 modifications, implying that frequencies related to intelligibility ($1000 - 3000$ Hz) do not contribute to energetic masking release with the f_0 modifications. The above observations suggest that the participants chose steps where more of the spectral energy of the speech escapes than in the original steps. However, the chosen steps are close to the original pitch.

Modelling f_0 preferences

For modelling the listener preferences, the DGAF closest to the listeners' preferred step was used (mean f_0 step 16 or 117 Hz). The ERB-rate closest to the preferred step was the 2^{nd} (106 Hz). The DGAF was computed only for this channel and can be seen in Fig. 6.10a and 6.10b for mean f_0 and f_0 variation features, respectively. The extended glimpse proportion (GP_{ext}) metric for the same utterances was also computed and plotted across the listeners' preferences (Fig. 6.11a and 6.11b). GP_{ext} is an objective measure of energetic masking and a good predictor of intelligibility [Tang and Cooke, 2016]. It can be observed that the DGAF for the ERB-rate of frequency 106 Hz can describe listener preferences more precisely compared to GP_{ext}. The symmetric Kullback–Leibler Divergence (KLD) test validated this observation (Table 6.1). DGAF and GP_{ext} were computed using the actual sentences and noise segments heard by listeners in the test phase of the experiment.

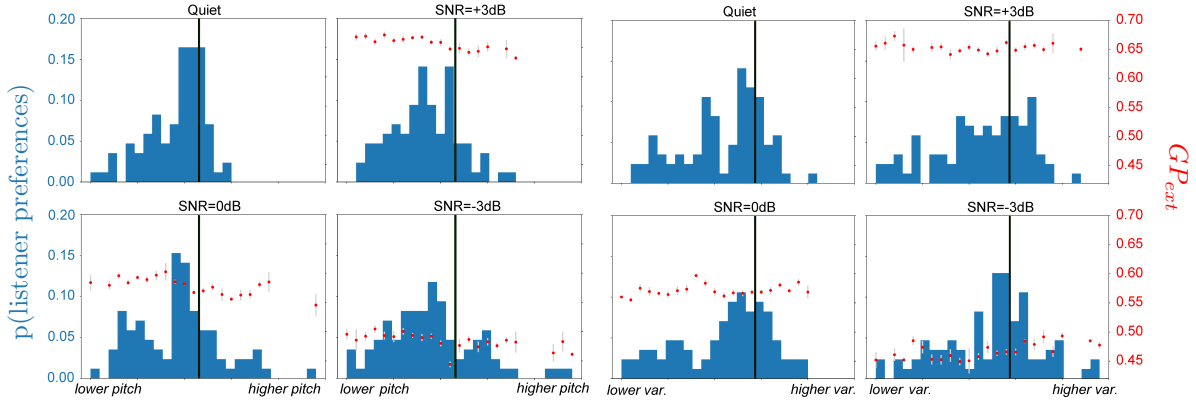
6.2.4 Interim discussion

In Expt. I, SpeechAdjuster was used to investigate listener preferences for f_0 in the presence of stationary noise. The masker in this experiment was generated to have the same long-term spectrum as the target speech, implying a high degree of energetic masking. The results reveal



(a) mean f_0 (b) f_0 variation

Figure 6.10: *Probability of each f_0 value (histogram, left axis), along with the DGAF of the 2^{nd} ERB-rate or 106 Hz (red dots, right axis). The error bars represent ± standard error. The black vertical line denotes the step that corresponds to the original value.*



(a) mean f_0 (b) f_0 variation

Figure 6.11: *Probability of each f_0 value (histogram, left axis), along with the glimpses (red dots, right axis). The error bars represent ± standard error. The black vertical line denotes the step that corresponds to the original f_0.*

which long-term spectral changes caused by f_0 modifications were preferred by the listeners. Our findings show that listeners always preferred mean f_0 and f_0 variation lower than those of the original speech; they chose similar mean f_0 values regardless of the condition; and they had a tendency to choose higher levels of f_0 variation for the most adverse condition compared to quiet. Additionally, as noise level increased, intelligibility decreased and listeners spent more time choosing their preferred f_0 value.

Listeners showed a preference for similar mean f_0 values, regardless of the condition. An explanation might be that in stationary noise such modifications do not facilitate intelligibility. In Lu and Cooke [2009a], mean f_0 increments similar to those found in Lombard speech were tested. The authors reported that the amount of energy moved to higher frequencies when the mean f_0 increase was small and insufficient to facilitate intelligibility. However, Barker and Cooke [2007] showed a positive correlation between mean f_0 and intelligibility for females for sentences presented in SSN, while there was marginal evidence that male speakers produced more intelligible speech with low mean f_0. The latter marginal outcome might explain our

finding that listeners preferred in general to lower the mean f_0. Additionally, this finding is supported by Ryalls and Lieberman [1982] and Assmann and Nearey [2008], who found that very high mean f_0 values lead to poor vowel identification compared to lower f_0 values, since vowel identification in this condition might be influenced by the sparse sampling of the harmonic spectrum. The impact of a sparse or dense sampling of the harmonic spectrum on different mean f_0 values for the different noise levels is observed in Fig. 6.8. The sparse sampling (higher pitch values) compared to the dense sampling (lower pitch values) resulted in a small number of regions with higher energy of the speech than that of the masker.

By controlling f_0 variation, listeners maintained speech understanding at high levels, while there was a tendency to choose speech with higher f_0 variation in the most severe condition compared to quiet (although this was not statistically significant). Laures and Bunton [2003] examined the effect of a flattened f_0 contour on the intelligibility of speech in white and babble noise. Consistent with our finding, they showed that the lack of f_0 variation has a significant impact on overall speech intelligibility. Additionally, Watson and Schlauch [2008] examined f_0 variation in white noise resulting in poorer intelligibility for speech with flattened f_0 compared to more variable unmodified f_0. Our findings show that listeners did not just prefer those f_0 variation values which lead to higher intelligibility, compared to what the extended glimpsing model would predict (Fig. 6.11b).

6.3 Experiment II: Listeners' f_0 preferences for speech in the presence of competing speech

6.3.1 Methods

Listeners

Twenty-three Greek monolingual listeners (4 female) were recruited. They were all young adults in the age range of $19-27$ years (mean 20.9 years; $S.D.$ 2.3 years). Listeners reported no known hearing problems. All but one participant reported good to excellent knowledge of English. Additionally, 9 of the 23 listeners reported extensive music studies. An incentive of 10 euros was given for participation. Two of the participants had also participated in Expt. I; however, the experiments were conducted around 6 months apart and a different set of test phrases was used.

Stimuli

Sentences were presented in quiet, or mixed with competing speech (CS) at -10, -6, -3 dB SNR. The masker was generated by concatenating all the 720 phrases with a gap of 0.5 s between them. The desired SNRs were obtained by rescaling the noise. The speech material used in this experiment was drawn from the same corpus described in Expt. I (sec. 6.2.1), but for this experiment a different set of phrases was used, apart from the phrases in the practice session. More specifically, the phrase IDs were $90-500$ for the adjustment phase and $1-81$ for the test phase.

As in Expt. I, listeners in each trial were allowed to modify either mean f_0 or f_0 variation. For the mean f_0 modifications, 25 steps were available to listeners, which corresponded to the first 25 half semitones starting from 75 Hz. Figure 6.12 shows the f_0 contour of an utterance for each of the 25 steps. The f_0 variation steps were computed as for Expt. I (sec. 6.2.1).

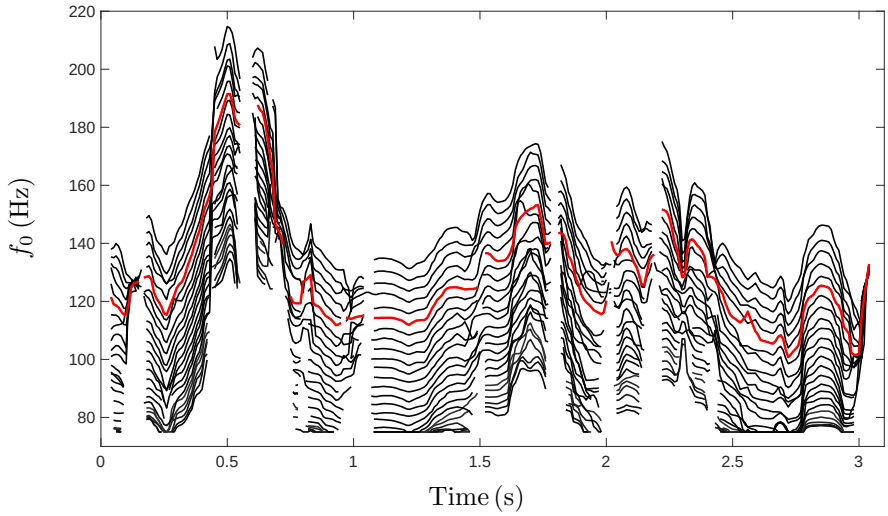


Figure 6.12: *f_0 contours of the sentence'Οι φιγούρες από χαρτόνι φάνηκαν πίσω από τον μπερντέ'. The 25 mean f_0 steps are depicted. The red line denotes the original mean f_0 value (step = 20).*

6.3.2 Procedure

The procedure used in this experiment was identical to that in Expt. I (6.2.2).

6.3.3 Results

Listener f_0 preferences

Figures 6.13 and 6.14 plot f_0 preferences, intelligibility scores and the time spent in the adjustment phase for the 4 conditions (Quiet and CS at 3 SNRs). For the mean f_0 modifications (Fig. 6.13), listeners preferred lower mean f_0 values in noise compared to Quiet conditions. As the noise level increased, intelligibility scores decreased and more time was needed in the adjustment phase. Similar results were observed for the f_0 variation feature (Fig. 6.14), with the only difference that in noise, listeners preferred slightly higher f_0 variations. In noise, listeners preferred speech with lower than the original mean f_0 value and in quiet almost equal to the original, while for the f_0 variation the opposite was seen. As in Expt. I, listeners spent on average more than 20 s during the adjustment phase in all conditions.

A rank-based, Kruskal–Wallis H-test was conducted to compare the effect of condition on each of the three measurements and each of the two tested features. Results indicated significant main effects for adjustment time (mean f_0 [$H = 31.57, p < 0.001$], f_0 variation [$H = 58.82, p < 0.001$]), intelligibility (mean f_0 [$H = 143.83, p < 0.001$], f_0 variation [$H = 180.90, p < 0.001$]) and for the preferred step (mean f_0 [$H = 70.95, p < 0.001$]; f_0 variation [$H = 19.30, p < 0.001$]). Post-hoc pairwise comparisons for the mean f_0 feature indicated that all the pairs of noise levels for intelligibility scores and for the adjustment time were significantly different, except for intelligibility at -3 dB SNR and adjustment time at -6 dB SNR, which were statistically different only from those in Quiet. For the preferred steps, only the Quiet condition differed significantly from all noise conditions. Regarding the f_0 variation, results indicated that the adjustment time and preferred steps of only the Quiet condition were significantly different from the remaining conditions. Intelligibility scores were statistically different for all pairs of conditions except for -6 and -3 dB SNR.

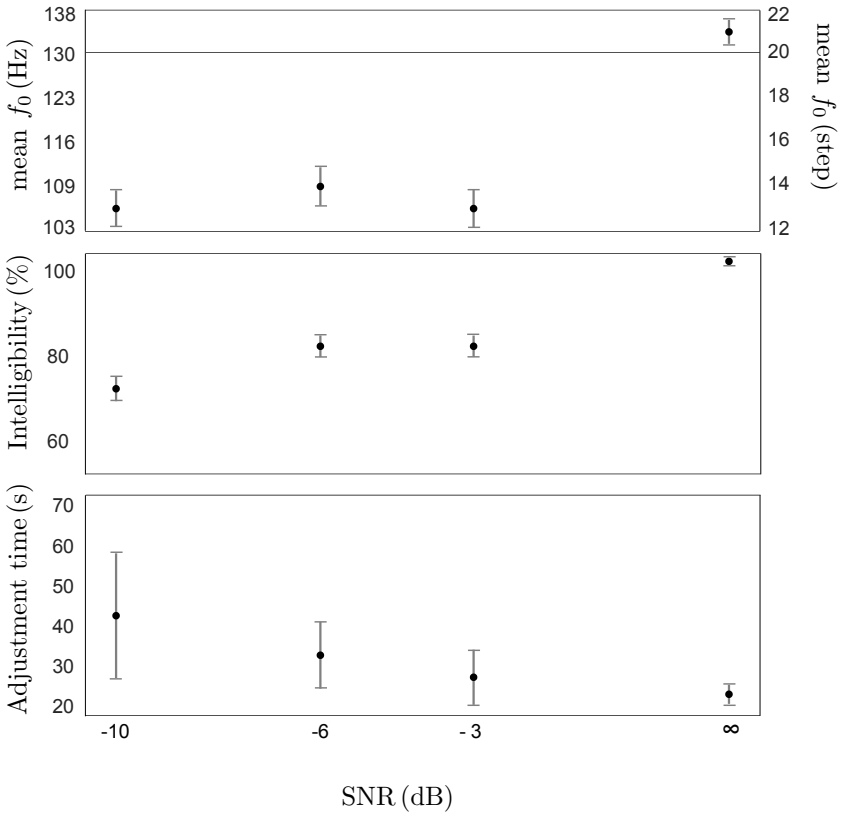


Figure 6.13: *Median values (black dots) of mean f_0 preferences (upper plot), intelligibility scores (middle plot) and adjustment time (lower plot) for the different conditions are depicted. The horizontal line in the upper plot indicates the original mean f_0 value. The error bars represent ± standard error of median. For the upper plot, statistics were computed based on the steps (right axis), while values on the left axis show the corresponding to the steps f_0 values.*

Distribution of listener preferences vs intelligibility

The probability with which each of the 25 permitted mean f_0 and f_0 variation levels were preferred by listeners along with the percentage of keywords correctly recalled are presented in Fig. 6.15 and 6.16, respectively. In Quiet, listeners had distinct f_0 preferences even though intelligibility was always at ceiling. For the conditions in noise, it can be observed that those listeners who chose mean f_0 or f_0 variation close to original had poorer intelligibility. Except for the quiet condition, listeners' preferences are widely spread across the available modification levels especially for the mean f_0 feature.

A one-sided, one-sample Wilcoxon test was used to test whether the preferred steps were significantly different from the original speech step. For the mean f_0 feature, results showed that only in noise the preferred steps were significantly lower (lower pitch) than the original (-3 dB SNR [$T = 598, p < 0.001$]; -6 dB SNR [$T = 681, p < 0.001$]; -10 dB SNR [$T = 452, p < 0.001$]). For the f_0 variation, results showed that only in Quiet listeners preferred significantly lower f_0 variation (higher step) than the original [$T = 4258, p < 0.001$].

f_0 choices and energetic masking

To examine the impact of f_0 modifications on energetic masking, the same procedure as in Expt.

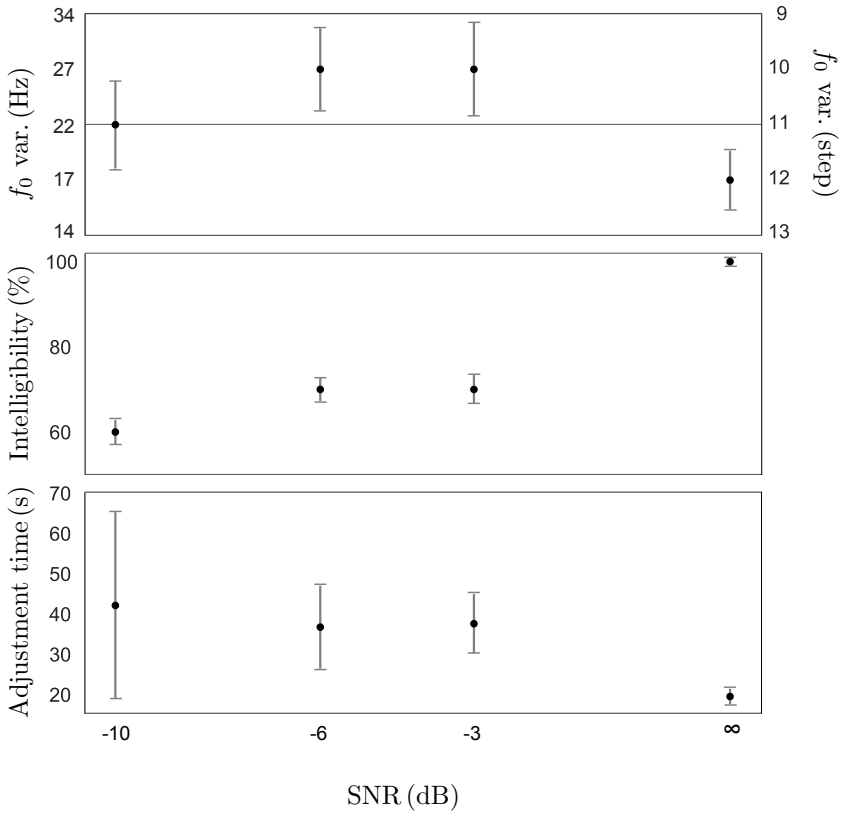


Figure 6.14: *As Fig. 6.13 but for the f_0 variation feature.*

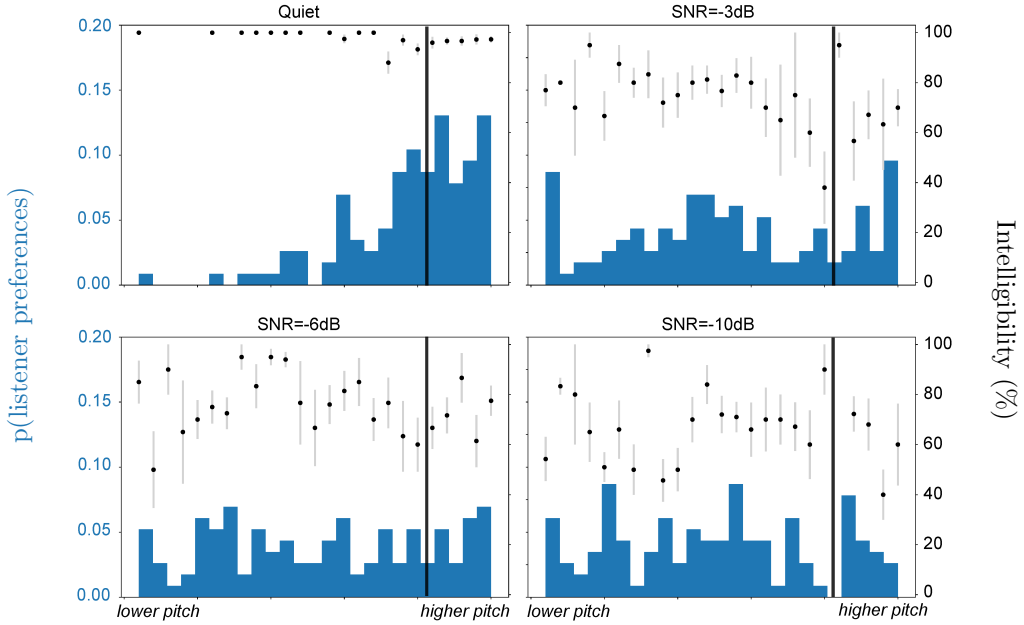


Figure 6.15: *Probability of each mean f_0 value (histogram, left axis), along with the percentage of words recalled correctly (black dots, right axis). The error bars represent ± standard error. The black vertical line denotes the step that corresponds to the original mean f_0.*

I was followed. Figures 6.17 and 6.18 show the sum of the DGAF values of all utterances in the test phase for the mean f_0 and f_0 variation features, respectively. In line with SSN, it is observed that for the competing talker masker, DGAF in channels above the 15^{th} (or 970 Hz) do not

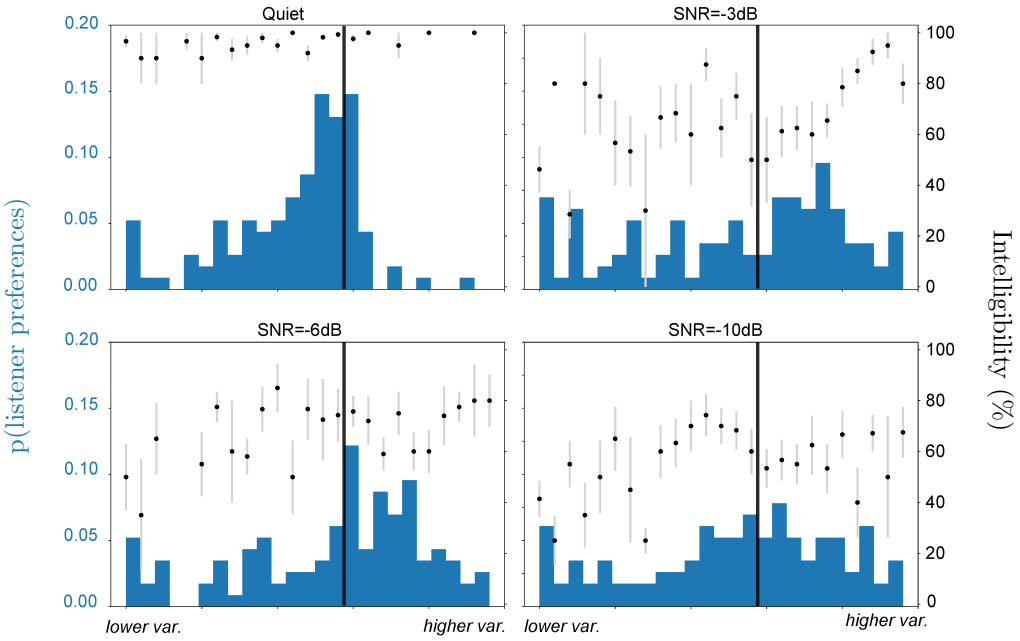


Figure 6.16: *As Fig. 6.15 but for the f_0 variation feature.*

vary much with regard to f_0 modifications, implying that the frequencies related to intelligibility ($1000 - 3000$ Hz) do not contribute to energetic masking release with these f_0 modifications. The competing talker's mean f_0 was around 130 Hz and the closest ERB channel to this is the 3^{rd} one (141 Hz). As expected, in Fig. 6.17, it can be observed that the values in this channel are smaller compared to the remaining channels for all noise levels. The difference diminishes under less noisy conditions.

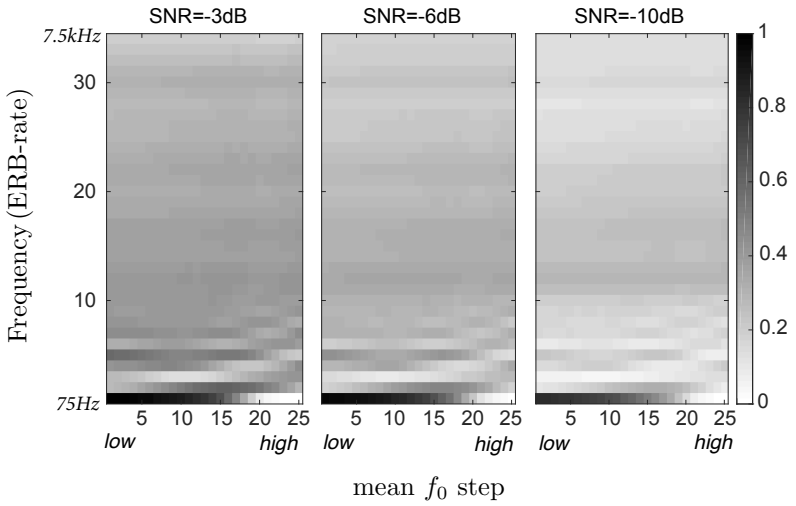


Figure 6.17: *As Fig. 6.8. The sum of the DGAF (see Fig. 6.3) of the 80 utterances from the test phase was computed for the ERB-rate scale (in total 34; y-axis) and for each of the 25 mean f_0 steps (x-axis).*

The closest mean f_0 step to the original is the 20^{th} (approx. 130 Hz). The participants chose a mean f_0 step where more spectral energy of the speech is escaped compared to that escaped for the original pitch.

For the f_0 variation feature, listeners chose steps with higher variability for the masked

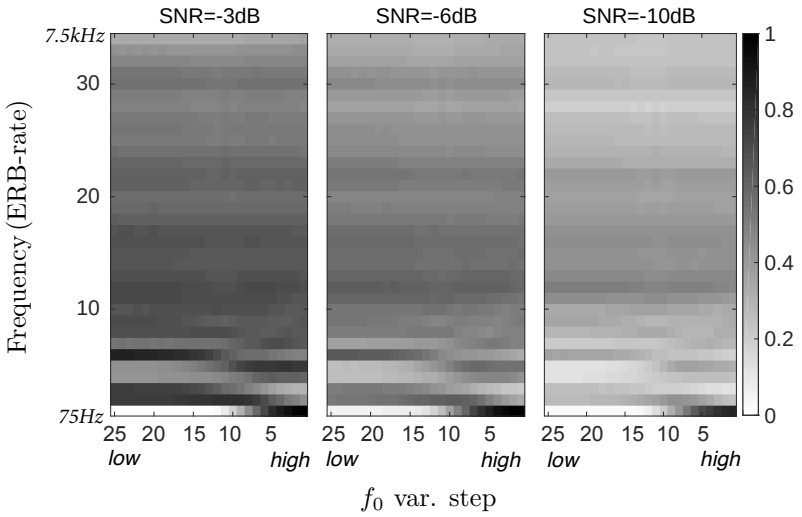


Figure 6.18: *As Fig. 6.17 but for the f_0 variation feature. low and high on the x-axis denote the lower and higher f_0 variation, respectively.*

conditions (step around 11, Fig. 6.14) than for the Quiet condition. In Fig. 6.18, it can be seen that from step 11 and for higher f_0 variation values (steps $1-11$), the overall glimpses are higher compared to those of lower f_0 variation (steps $12-25$). However, listeners did not choose different values for the different conditions.

Modelling f_0 preferences

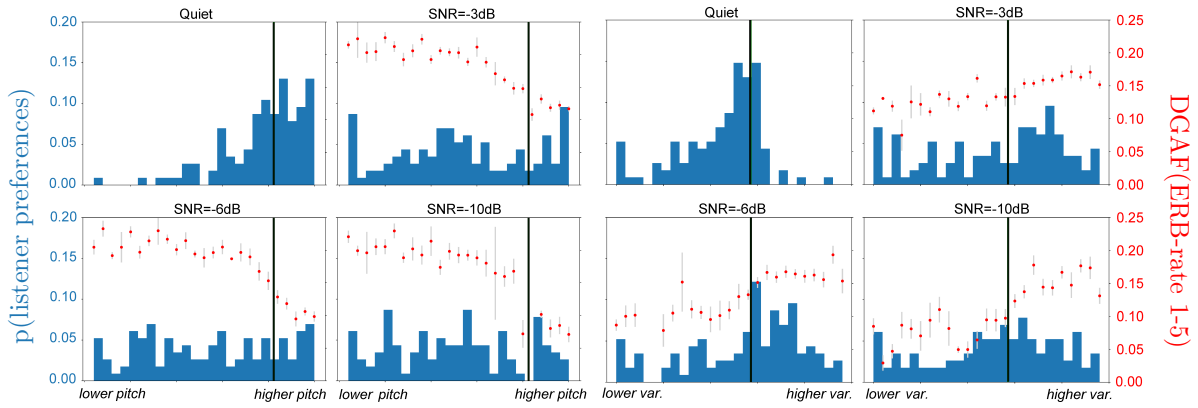
From the heat maps, the DGAF of ERB-rate 1 (75 Hz) to 5 (221 Hz) varies greatly with regard to the f_0 modifications; thus, these ERB-rates were used for modelling the listener preferences. The DGAF was computed for ERB-rates and can be seen in Fig. 6.19a and 6.19b for the mean f_0 and f_0 variation, respectively. The GP_{ext} metric for the same utterances was also computed and plotted across the listener preferences (Fig. 6.20a and 6.20b). It can be observed that the DGAF values for ERB-rates $1-5$ ($75-221$ Hz) can describe listener preferences more precisely compared to GP_{ext}. The symmetric KLD test validated this observation (Table 6.2). DGAF and GP_{ext} were computed using the actual sentences and noise segments heard by listeners in the test phase of the experiment.

		mean f_0		f_0 variation	
		DGAF(ERB-rate 1-5)	GP_{ext}	DGAF(ERB-rate 1-5)	GP_{ext}
SNR(dB)	-10	0.84	1.06	0.28	0.22
	-6	0.32	0.23	1.28	1.86
	-3	0.38	0.33	0.36	0.39

Table 6.2: *The symmetric Kullback–Leibler Divergence (KLD) derived from the comparison between the listener preferences distribution with DGAF of $1-5$ ERB-rate distribution and GP_{ext} distribution for the different SNRs. The lower the KLD value, the closer the two distributions. If it equals zero, the two distributions are identical.*

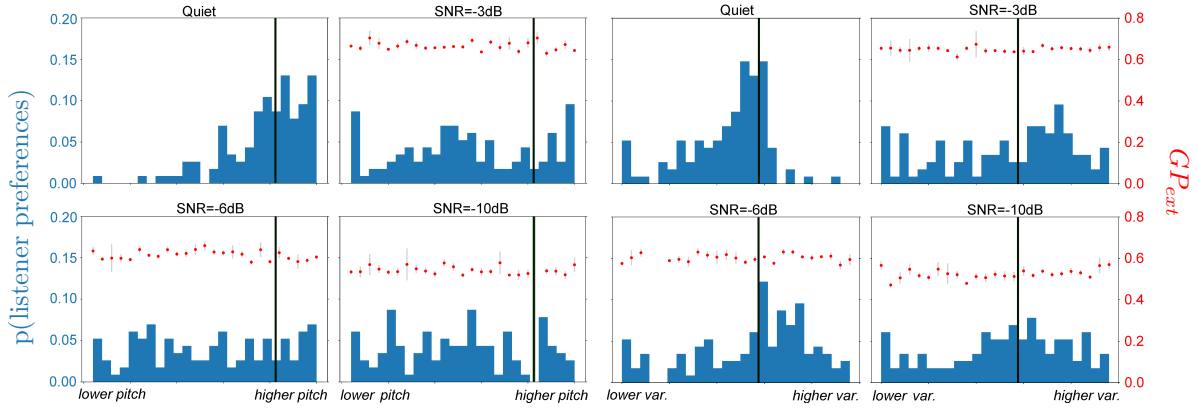
6.3.4 Interim discussion

In Expt. II, SpeechAdjuster was used to investigate listener preferences for f_0 in the presence of a competing talker. The informational masking in this experiment was extreme, since the



(a) mean f_0 (b) f_0 variation

Figure 6.19: *Probability of each f_0 value (histogram, left axis), along with the DGAF of the $1-5$ ERB-rates (red dots, right axis). The error bars represent ± standard error. The black vertical line denotes the step that corresponds to the original f_0.*



(a) mean f_0 (b) f_0 variation

Figure 6.20: *Probability of each f_0 value (histogram, left axis), along with the glimpses (red dots, right axis). The error bars represent ± standard error. The black vertical line denotes the step that corresponds to the original f_0.*

same talker was used for both target and masking voices. Brungart [2001] showed that highest intelligibility of an utterance can be achieved when the target and masking utterances are spoken by different-sex talkers and the least intelligible when the talker and masker are spoken by the same individual. Thus, it was expected that listeners would choose the target talker's f_0 characteristics to be highly different from the masker's (towards the female pitch region) in order to be able to disentangle the two talkers more easily. Contrary to our expectations, our results revealed that listeners did not prefer values higher than the original pitch for the masked conditions. As this observation holds for both experiments of this chapter, it will be addressed further in the general discussion below. Specifically, our main findings are that listeners preferred speech with a mean f_0 similar to the original in quiet, but lower in noise; a lower f_0 variation for speech in Quiet compared to noise; while only for the Quiet condition was the preferred f_0 variation value lower than that of the original speech.

Regarding the listeners' preference of lower pitch in noise, it is known that when the pitch

of two simultaneous speakers differs, then they can be perceptually separated and recognised separately. A tone can be perceived as separate when a low harmonic in a complex is mistuned by more than 3% of the harmonic frequency [Moore et al., 1986]. It has been shown that when two vowels are presented at the same time with different f_0s, listeners identification accuracy is improved [Scheffers, 1983]. Furthermore, previous studies have shown that increasing the difference in pitch between target and background talkers helps to distinguish them, resulting in increased intelligibility [Assmann, 1999; Brokx and Nooteboom, 1982]. Even for listeners who chose the original pitch, intelligibility was not lost (Fig. 6.15 at -3 and -6 dB SNR). This can be explained by the fact that natural mean f_0 variations exist in the concurrent speech streams, even if both come from the same talker, which might have facilitated the listener. In our study, most of the listeners chose a mean f_0 greater than 2 semitones below the original (Fig. 6.13 upper plot). Darwin et al. [2003] tested how increases in f_0 affect intelligibility when listening to two sentences uttered by the same talker. They concluded that a difference in f_0 greater than 2 semitones results in systematic improvements in performance.

Regarding the f_0 variation results, in noise higher variation was chosen compared to the Quiet condition. Increasing the target talker's f_0 variation has been found to facilitate intelligibility for several reasons. First, momentary differences in f_0 help segregate the two sources [Bregman, 1990]. Coherent f_0 modulation makes it easier to track a voice over time, since the speech is perceived as coming from a single speaker [Nooteboom et al., 1978]. When both target and competing speech derive from the same talker, by changing the target talker's f_0 variation, the perceptual fusion and crossing pitch contours are reduced. Finally, enhanced pitch prosody can help in speech recognition that is distorted by competing speech. Our results revealed that for higher f_0 variations, the energetic masking release in low frequencies is higher (Fig. 6.18).

6.4 General discussion

The aim of the experiments in this chapter was to explore listener preferences in the choice of f_0 in quiet, stationary noise and competing speech. The impact of f_0 on intelligibility for different maskers has been studied [Assmann, 1999; Lu and Cooke, 2009a], while it is not clear if only intelligibility as a factor is adequate to cover listeners' preferences on speech. For our experiments, native-Greek listeners were recruited and SpeechAdjuster was used to collect their responses. The stationary noise in Expt. I was generated using the same long-term spectrum as the competing talker in Expt. II. Thus, the differences in listeners' choices for the common SNR condition (-3 dB) in the second experiment derive from the informational masking imposed by the competing talker. For an SNR of -3 dB, the intelligibility of speech masked by speech from the same talker was expected to be substantially poorer than when the masking used speech-shaped noise. However, listeners' f_0 choices might have helped them to acquire almost the same intelligibility scores for both maskers (median score for both maskers was 80%). The preferred mean f_0 step for the competing speech (106 Hz) condition was slightly lower than that for the SSN (117 Hz). However, the two experiments for the mean f_0 feature cannot be compared directly, since the mechanism for computing the mean f_0 steps was different in each experiment. The same applies to the f_0 variation feature. Listeners achieved quite similar intelligibility scores for both maskers (median score for SSN was 80% and for CS 70%) when choosing speech with a slightly more variable f_0 for the competing speech (the difference in variability was around 4.3 Hz). Specifically, for the stationary noise (Expt. I), listeners chose

similar mean f_0 and f_0 variation values, lower than the original, regardless of the condition. For the competing talker (Expt. II), listeners preferred speech with a similar mean f_0 to the original in Quiet, but a lower value in noise. For the f_0 variation, lower values were chosen for speech in Quiet compared to noise. The stimuli under the Quiet condition for the f_0 variation feature were identical in both experiments, and prompted similar responses. In general, results showed distinct preferences, even in conditions with intelligibility at ceiling, while for some conditions listeners' adjustments helped them to maintain their intelligibility (i.e. for the f_0 variation feature at 0 dB SNR for SSN, and for both features at -6 dB SNR for CS). Even though the trend for the preferred f_0 values was similar in both experiments; preferences in noise were more evident in Expt. II compared to Expt. I. Finally, our findings are expected to be language independent, since intrinsic f_0 patterns are generally consistent across non-tonal languages (for f_0 vowel patterns [Whalen and Levitt, 1995]).

Listeners did not prefer to increase the mean f_0 with higher noise levels (Fig. 6.4, 6.13), as happens in Lombard speech, but instead they preferred speech with a lower f_0 for both masker types. This finding supports the idea that the increase in f_0 observed in Lombard speech might be a by-product of hyper-articulated speech (passive result of raising subglottal pressure [Gramming et al., 1988]), leading neither to increased intelligibility, nor to being preferred by listeners. Intelligibility benefits in Lombard speech might derive from other factors, such as a flatter spectral tilt (flatter tilt in noise is also preferred by listeners, see chapter 7 and Simantiraki et al. [2020]), changes in consonant-vowel energy ratio, and formant frequencies. Even though preferences were not greatly different from the original pitch, there was a general tendency to choose values lower than the original. A model based on the glimpsing model [Tang and Cooke, 2016] was introduced to test the impact of speech with denser or sparser harmonics on energetic masking. When the mean f_0 is lowered, the number of harmonics increases and becomes denser, thereby affecting the energetic masking by increasing the amount of speech that escapes from masking at low frequencies, as was also revealed in our results (Fig. 6.8, 6.17). It is known that harmonics are unresolved if they fall into the same equivalent rectangular bandwidth of the auditory filter [Plack and Oxenham, 2005]. Thus, the steps of which more glimpses exist at low frequencies are more likely to be preferred by listeners. Our results revealed that listeners' f_0 preferences can be described precisely in terms of the number of glimpses relative to the total of the utterance's glimpses at low frequencies (Table 6.1 and 6.2).

Even though our results (Fig. 6.6, 6.7, 6.15, 6.16) show that there is a drop in intelligibility for those listeners who chose f_0 values close to the original, the majority of listeners might have based their choices on criteria apart from intelligibility, such as naturalness. Since the target talker was male (with mean f_0 around 130 Hz), lowering the pitch might have sounded more appropriate given his formant frequencies. Previous studies of listeners' f_0 preferences have shown that if f_0 is 'matched' with formants, then speech is considered as more natural [Assmann et al., 2006] and is preferred by listeners [Assmann and Nearey, 2007], regardless of whether it is lower or higher than the original pitch.

Chapter 7

Listener preferences - Spectral energy reallocation

7.1 Introduction

Amongst other speech properties that talkers naturally modify to promote audibility is the speech energy in mid and high frequencies. Under various circumstances, such as when talkers produce clear speech [Krause and Braida, 2004], Lombard speech [Summers et al., 1988; Junqua, 1993] or speech at a distance from the listener [Liénard and Benedetto, 1999], the level of spectral components at higher frequencies is greater compared to that in conversational speech. In Lombard speech the enhanced energy in the $1000 - 4000$ Hz region comes as a side-effect of the enhancement of the higher formants' amplitude and the flatter spectral slope [Garnier and Henrich, 2014]. In Lu and Cooke [2008], an overall shift in the centre of gravity of energy from lower to higher frequencies was found for speech produced by competing talkers, babble, and stationary noise.

Although the observed energy reallocation strategies might be a passive effect of a talker's increased vocal effort [Lu and Cooke, 2009b; Garnier and Henrich, 2014], such shifts in spectral energy are effective in enhancing intelligibility in noise [Skowronski and Harris, 2006]. Increased energy in the $1000 - 3000$ Hz range was found to be one of the factors that makes clear speech more intelligible than conversational speech [Krause and Braida, 2004].

Previous studies have explored the impact of different spectral energy reallocation methods on intelligibility. Lu and Cooke [2009a] tested spectral tilt modifications and found that spectral tilt flattening can lead to intelligibility gains in the presence of noise. In Tang and Cooke [2010], five energy reallocation strategies were tested (3 based on equalising local SNRs to a fixed global SNR, and 2 strategies of energy modification to a subset of frequency channels or changes based on the local SNR). The investigators found that increasing the SNR of specific frequency bands led to a large increase in intelligibility, but accompanied by a significant reduction in speech quality. In Tang and Cooke [2012], modifications of energy reallocation by adding energy to frequencies that are less likely to be masked resulted in intelligibility improvements in noise.

This chapter extends the previous research on intelligibility by investigating the impact of enhancing or attenuating the energy of different frequency components on listener preferences and intelligibility. Listeners were able to reallocate the speech energy by adjusting [1] spectral tilt, [2] energy of certain spectral bands, [3] cut-off frequency of a high-pass filter, [4] cut-

off frequency of a low-pass filter, [5] bandwidth of a band-pass filter, and [6] location of the frequency band to enhance (Fig. 7.1). The features tested were divided up into two experiments. A listener had the option to participate in only one or both experiments. In the first experiment (Expt. I, sec. 7.2) listeners were allowed to adjust features [1] and [2]. In this experiment, the magnitude of specific bands was changing. In the second experiment (Expt. II, sec. 7.3) features [3] to [6] were tested, allowing the listeners to enhance different frequencies while maintaining the band magnitude. Spectral modifications were carried out via SpeechAdjuster (chapter 4) and tested for speech in quiet and in three levels of speech-shaped noise, with the constraint to maintain the overall energy unchanged. Similar practices used in both experiments are described in the first sections: i.e. speech material (subsection in sec. 7.2.1), statistical tools (subsection in sec. 7.2.1), and procedure (sec. 7.2.2). Listener demographics, stimuli preparation and results are described separately for each experiment.

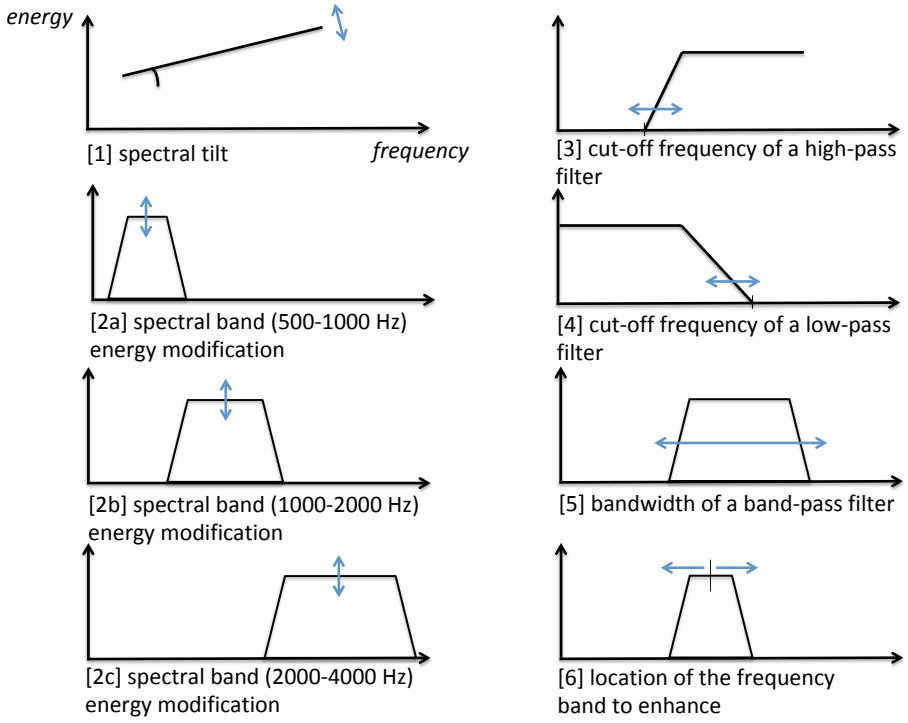


Figure 7.1: *A visual-summary of the features tested in this chapter. The allowed user variations are indicated with blue arrows.*

In noise, it is expected that listeners will adjust speech in order to increase intelligibility. One assumption, therefore, is that their choices will lead to enhancement of mid-frequencies and spectral tilt flattening, so that as noise level increases, the enhancement will be greater and the tilt flatter. The first research question is the following: do listeners choose to reallocate speech spectral energy so to maximise intelligibility? Listener preferences might be also influenced by other factors beyond intelligibility. Conditions in which intelligibility is almost constant, such as the noise-free condition, should be investigated. Thus, the second research question is: do listeners' preferences show patterns that are independent of intelligibility? Finally, the third research question is: do listeners' preferences change in challenging conditions? Listeners' preferences may change in a way similar to Lombard speech, where the level of background noise affects speech characteristics related to the spectral energy.

7.2 Experiment I: Effects of spectral tilt and spectral band energy modifications on listeners' preferences and intelligibility

7.2.1 Methods

Listeners

Thirty-five native Spanish listeners (30 females) aged between 18 and 34 years (mean 20.1; SD 2.6) were recruited. All listeners passed an audiological screening with a hearing level better than 25 dB at frequencies in octave steps in the range $125 - 8000$ Hz in both ears. Listeners were paid 10 euros for their participation.

Stimuli

Speech material

Speech stimuli were drawn from the Sharvard Corpus [Aubanel et al., 2014]. It consisted of Spanish sentences spoken by one male and one female native Spanish talker at a normal speaking rate. The level of difficulty of this corpus is similar to that of the original English Harvard Corpus. Each sentence contains 5 keywords: e.g. 'El **color gris está muy** de **moda**' ('The gray color is very fashionable'); keywords are indicated in bold. For the experiment, a male voice was used as the target speech and a female voice for generating the maskers.

Stimulus preparation

Two speech features were tested: changes consisting of modifications to spectral tilt (Fig. 7.1[1]) and spectral band energy (Fig. 7.1[2a]-[2c]). The latter feature has three variations.

Spectral tilt. For spectral tilt modifications (Fig. 7.1[1]), pre-emphasis and de-emphasis filters were used, in order to enhance or attenuate the energy in the higher frequencies, respectively. Changes in spectral tilt were achieved by filtering the speech signal with a digital filter (*filter* function in Matlab 2016b), using the rational transfer function $H(z) = 1 - \lambda z^{-1}$ for pre-emphasis and $H(z) = \frac{1}{1-\lambda z^{-1}}$ for de-emphasis. The λ coefficients, for both the pre-emphasis and the de-emphasis filter, were drawn linearly from the range $[0.2, 1]$. In total, 23 steps were constructed, corresponding to tilts in the range $[-10.85, 0.59]$ dB/octave. Eleven steps were constructed with spectral tilt steeper than the original, 1 with the original spectral tilt, and 11 with spectral tilt flatter than the original. The original spectral tilt value was -5.24 dB/octave. To measure spectral tilt, the speech spectral energy in octave bands was computed and a first degree polynomial was fitted to the data. The first coefficient of the polynomial was used to express the tilt in dB per octave. Figure 7.2 shows spectrograms of a phrase with the original and two extreme spectral tilts. It can be seen that, for the steepest tilt, the speech energy is up to 4000 Hz, while for the shallowest tilt it spans the full spectrum.

Spectral band energy modifications. For the spectral band energy modifications (Fig. 7.1[2a]-[2c]), three bands were chosen that correspond to the frequency ranges of $500-1000$, $1000-2000$,

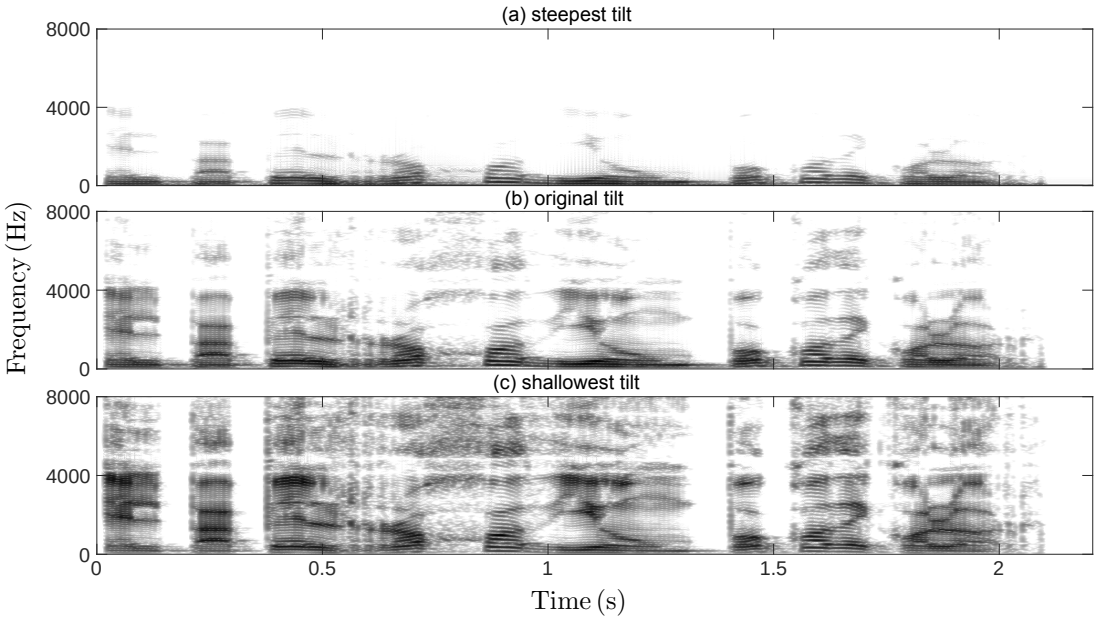


Figure 7.2: *Spectrograms of the phrase 'El papel rojo dio un poco de color' for the spectral tilt modifications.*

and $2000 - 4000$ Hz. In any single trial, the listener was able to modify the energy of one of the bands. In total, 21 spectral band energy steps were assigned using an exponential growth function (eq. 7.1).

$$magnitude_{dB} = m + k * (1 + r)^t) \tag{7.1}$$

where $m = 15$ is a correction term, $k = -100$ is the starting value of the exponential growth and $r = -0.2$ is the growth rate of the values, as $t = [0 : 1 : 20]$ represents discrete intervals in 21 steps. The terms m and k were chosen so that the band that corresponds to the greatest energy attenuation contains the minimum energy and the greatest energy enhancement the maximum energy, without causing clipping to the output speech. Exponential growth was used so the number of steps are almost equally distributed for enhanced and attenuated speech.

In a frame-by-frame analysis (window of 30 ms), a Fourier transform was performed. The amplitude spectra were multiplied using a custom filter that enhances/attenuates the desired frequencies and retains the remaining ones. Then, the Fourier coefficients were reconstructed from the modified amplitude and the original phase and the filtered signal was generated using the inverse Fourier transform. Figure 7.3 shows the spectrograms of a phrase with the original and the two extreme magnitude steps for the band $1000 - 2000$ Hz. We can see that for the greatest attenuation, there is almost no energy in the band $1000 - 2000$ Hz.

Stimuli were presented in quiet and in 3 additive noise conditions using a speech-shaped noise (SSN) masker at SNRs of -6, -3 and 0 dB. The masker was generated by filtering random uniform noise with the long-term spectrum of the 700 concatenated sentences of the female talker of the Sharvard corpus, without gaps. The desired SNRs were obtained by rescaling the noise. Figure 7.4 shows the long-term spectrum of the 3 maskers. From the Spanish corpus, the sentence IDs used in this experiment were $1 - 380$ for the adjustment phase, $381 - 541$ for the test phase, and for the practice session 10 random unique phrases were drawn from the adjustment phase set. The amplitude of each sentence was normalised using a fixed root-mean-square

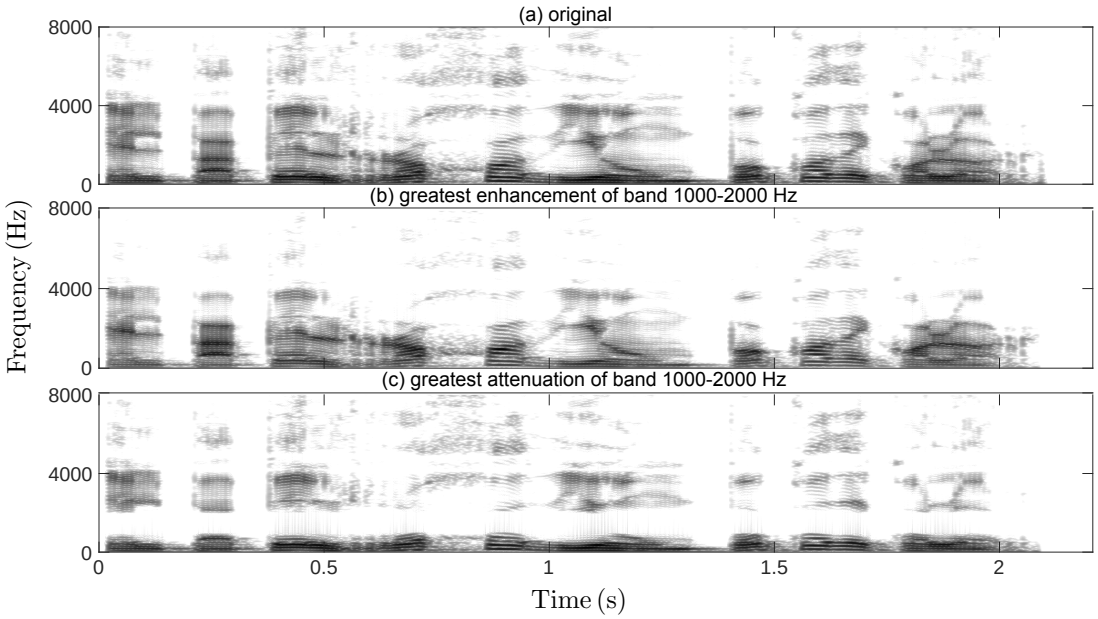


Figure 7.3: *An example of the spectrograms of the phrase 'El papel rojo dio un poco de color' after modifying the magnitude of band* $1000 - 2000$ *Hz.*

criterion so as to ensure that each sentence had approximately the same presentation level as every other sentence.

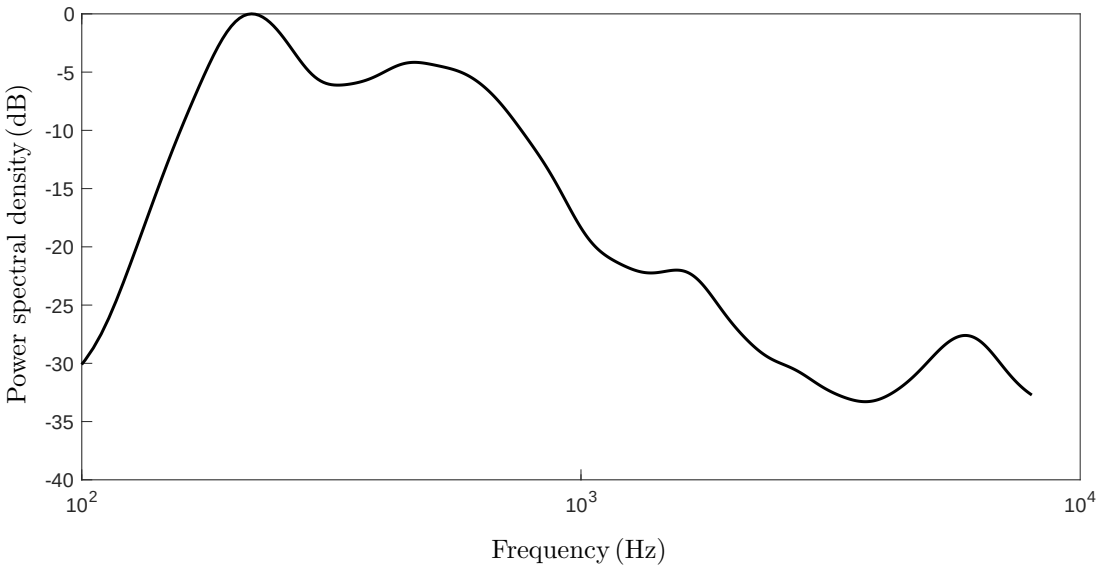


Figure 7.4: *Long-term average spectrum of the concatenated sentences of the Sharvard corpus uttered by a female.*

Statistical analysis tools

Since not all the data in the different conditions were normally distributed, non-parametric statistical tests were used and the median values (robust statistic) of the data were analysed. All tests were performed in Python using functions of the *stats.scipy* and *scikit_posthocs* libraries (showed in parentheses below). Differences among the experimental conditions were tested using the rank-based, Kruskal–Wallis H-test (*kruskal*). Post-hoc comparisons were per-

formed using Dunn's test (*posthoc_dunn*), and the Holm correction [Holm, 1979] was followed to counteract the problem of multiple comparisons. Finally, Kullback–Leibler divergence was computed to compare the distributions derived from listener preferences and intelligibility measures (*entropy*).

7.2.2 Procedure

The experiment was divided into 4 blocks according to condition (Quiet and SSN at 3 SNRs), with each block containing 5 trials in which listeners were able to modify one of 4 parameters (spectral tilt and energy of 3 spectral bands). The presentation order of the 20 trials was random. The procedure was similar to that described in chapter 6. 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), starting at a random feature value. The term 'feature' refers to the spectral tilt and the 3 spectral band modifications. Participants had to listen to at least 5 s of speech before proceeding to the test phase, but could listen to as much speech during the adjustment phase as desired. In the test phase, intelligibility was evaluated by a speech perception task using the value of the feature as chosen at the end of the adjustment phase. Participants listened to 2 sentences separately and had to type what they heard into an on-screen text box after each sentence presentation. Prior to the experiment, all the participants underwent a task familiarisation phase consisting of 5 trials, 2 in quiet and 3 in noise.

The real-time modifications technique and the instructions used in this experiment were similar to those described in chapter 4. Listeners were asked to tune the speech in real-time until they could recognise as many words as possible. Real-time changes could be made by using the up/down keys while listening to sentences. The task was explained as akin to choosing an appropriate volume for a television: too quiet makes comprehension difficult, while too loud leads to discomfort.

A balanced Latin square design was used for block ordering across participants. Stimuli were presented through Sennheiser HD380 headphones at a fixed presentation level, different for each condition (approx. 84, 79, 77 and 75 dB SPL for the -6, -3, 0 dB SNR and quiet conditions, respectively). For the calibration, a Brüel and Kjaer type 4153 artificial ear and a Brüel and Kjaer type 2260 sound level meter were used. Listeners were seated in a sound-attenuating booth in a purpose-built speech perception laboratory at the University of the Basque Country.

7.2.3 Results

Spectral energy reallocation preferences

Figures 7.5 and 7.6 show the median spectral energy reallocation preferences, intelligibility scores, and the time spent in the adjustment phase for the 4 conditions (Quiet and SSN at 3 SNRs). For the spectral tilt modifications (Fig. 7.5), in all conditions, listeners chose to reduce tilt with respect to the original. As the noise level increased, listeners preferred speech with progressively more energy at higher frequencies (resulting in a preference for a flatter spectral tilt), while intelligibility scores were below ceiling only for the most adverse condition (-6 dB SNR), and the time for adjusting the speech increased. For spectral band energy modifications (Fig. 7.6), in noise listeners chose to attenuate the energy in band $500-1000$ Hz and to enhance bands $1000-2000$ Hz and $2000-4000$ Hz, while for the Quiet condition, participants chose

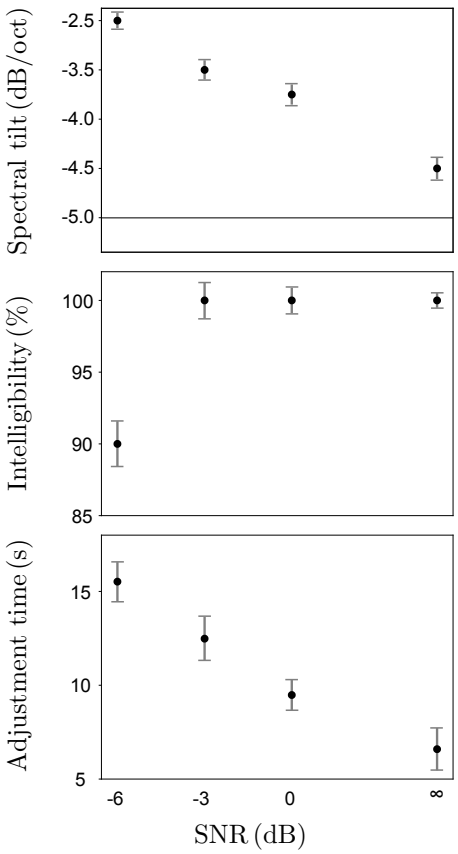


Figure 7.5: *Median values (black dots) for spectral tilt preferences (upper plot), intelligibility scores (middle plot) and adjustment time (lower plot) for the different conditions. The horizontal line in the upper plot indicates the original spectral tilt value. The error bars represent the ± one standard error of median.*

speech with spectral energy allocated as in the original speech. Intelligibility scores were at ceiling when the target speech had equal or greater level than the noise. For the negative SNRs, the enhancement of the $2000-4000$ Hz band helped listeners to achieve an intelligibility score greater or equal to 88% (similar to those for the spectral tilt modifications), while the manipulation of the other bands resulted in lower scores and required more adjustment time. For all features, the time needed to find the appropriate step increased with the noise level. For band $1000-2000$ Hz, listeners preferred identical magnitude for both -3 and 0 dB SNR. However, for the -3 dB, listeners needed more time compared to that for 0 dB and their intelligibility scores were poorer. Apart from the spectral tilt modifications, in which the spectral slope of a phrase directly changes, the slope also changes as a result of modifying the spectral band energy. The slopes for all the test phrases at each of the steps for the 3 spectral band energy modifications were computed (as described in sec. 7.2.1 Stimulus preparation) and plotted in Fig. 7.7.

As expected, when the band $500-1000$ Hz is enhanced, corresponding to a higher step, the slope becomes steeper (i.e. more negative), while for the remaining variations the opposite is true. For the adverse noise conditions, listeners chose to attenuate the band $500-1000$ Hz, which implies an enhancement of the remaining bands, although they did not choose the lowest possible step. Spectral tilt and band $2000-4000$ Hz features showed similar intelligibility scores (Fig. 7.5 and Fig. 7.6, respectively); thus, the frequency components greater than 4000 Hz

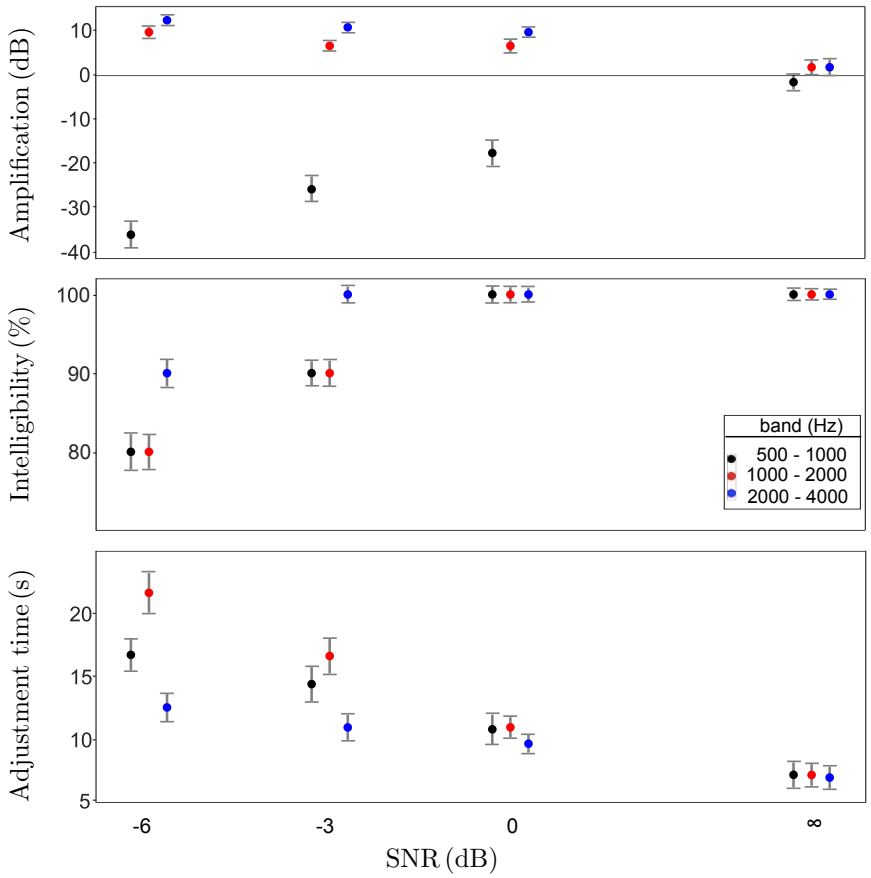


Figure 7.6: *As Fig. 7.5 but for the spectral band energy modifications.*

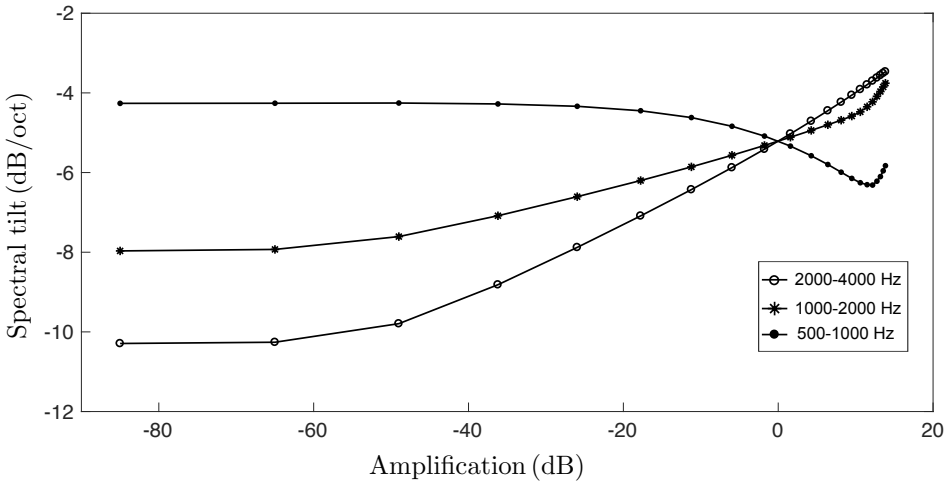


Figure 7.7: *Mean spectral tilts of all the test phrases at each of the steps (markers correspond to the 25 modification steps) for the spectral band energy modifications. The standard error at each data point was around* 0.03 *(for clarity, not shown).*

did not contribute to the increased intelligibility. Additionally, for the spectral tilt, listeners chose speech with slope at 0 dB (Fig. 7.5), which is the same as the one chosen for the band $2000 - 4000$ Hz at -6 dB (Fig. 7.6). This result signifies the high importance of this band in relation to both the performance and the listener preferences in adverse noise levels.

A rank-based, Kruskal–Wallis H-test was conducted to compare the effect of conditions on each of the three measurements (preferences, intelligibility scores, and adjustment time) and each of the 4 tested modifications. Results indicated significant main effects for all measurements [$p < 0.001$] (Table 7.1). Post-hoc pairwise comparisons for the spectral tilt feature indicated that all measurements were significantly different for the different conditions (except for the adjustment time at -3 and -6 dB SNR and the preferences at 0 and -3 dB SNR). For the spectral band energy modifications, results indicated that all measurements were significantly different for the different conditions, with the following exceptions: the preferred steps at -3 and -6 dB SNR for band $500 - 1000$ Hz; the preferred steps at 0 and -3 dB SNR for band $1000 - 2000$ Hz and the adjustment time at the SNR pairs of -3, -6 and -3, 0; and the intelligibility scores at -3, 0 for band $2000 - 4000$ Hz.

	adjustment time	intelligibility	preferred step
sp. tilt	114.98	99.69	172.05
0.5-1 kHz	130.27	149.31	183.45
1-2 kHz	177.57	212.96	56.33
2-4 kHz	55.19	97.46	175.07

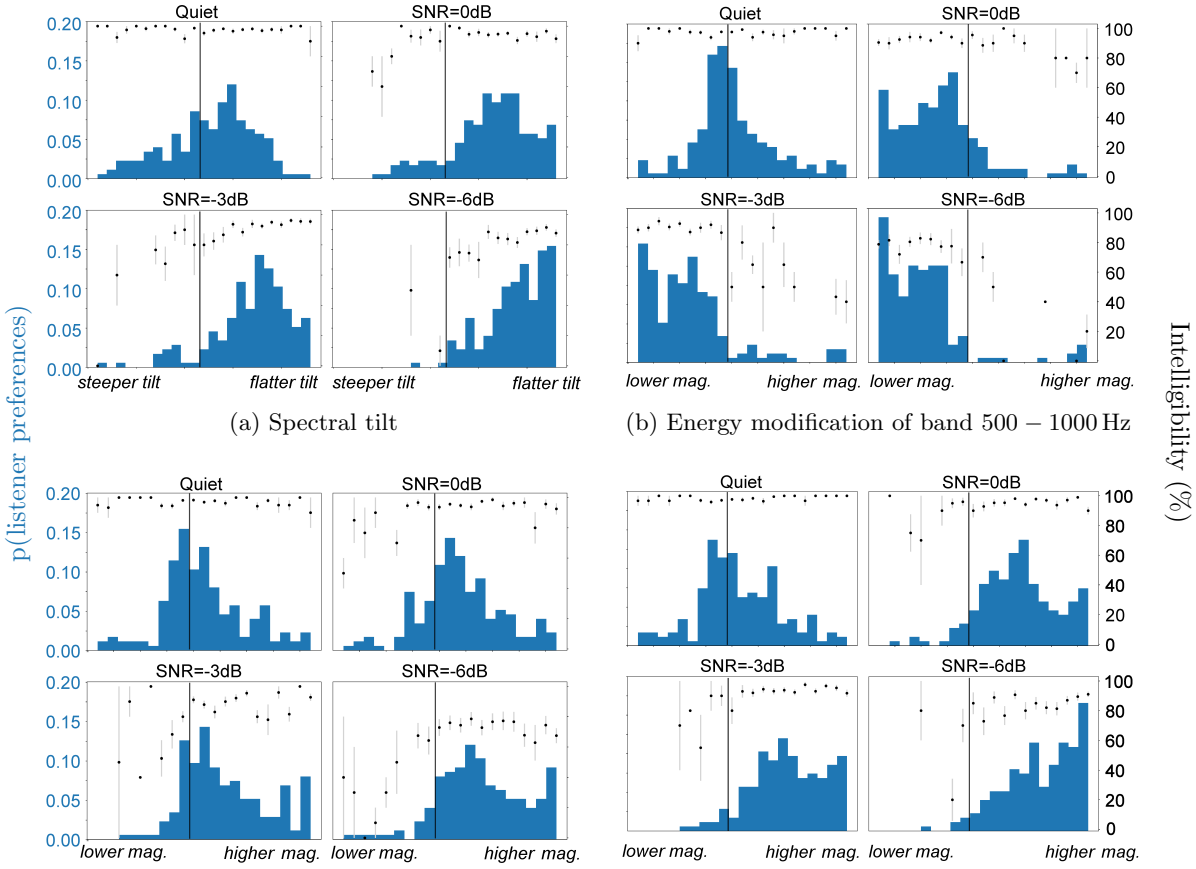
Table 7.1: *H-values of the Kruskal-Wallis statistical test for the different features and measures.*

Effects of spectral modifications on listener preferences and intelligibility

Figures 7.8a-7.8d show the probability that each value of spectral tilt (Fig. 7.8a) or of the specific band's energy (Fig. 7.8b-7.8d) is preferred by listeners, along with the percentage of keywords identified correctly as a function of noise level. Under all conditions, listeners showed distinct spectral tilt preferences, even when intelligibility is at or near ceiling performance. With increasing SNR, the number of steps with intelligibility scores at ceiling increases and listener preferences occupy a wider range. One feature of these results is the poorer intelligibility score obtained by those listeners who preferred to listen to speech with a steeper spectral tilt, or with more enhanced magnitude of band $500 - 1000$ Hz, or with more attenuated magnitude of bands $1000 - 2000$ Hz and $2000 - 4000$ Hz compared to the original speech, especially under the more adverse conditions. Listener preferences clearly reveal information beyond that captured by intelligibility scores. For some conditions, it can be observed that the median value does not represent the step that was mostly preferred. This can be seen in the histograms at -6 dB SNR for band $500 - 1000$ Hz modifications, where listeners preferred the greatest attenuation available, while for band $2000 - 4000$ Hz they preferred the lowest.

Energy reallocation choices and energetic masking

Figures 7.9a-7.9d show the overall speech energy that survives the energetic masking under the different conditions as computed by the DGAF measure, described in sec. 6.2.1 'Energetic masking measures'. It can be observed that listeners are aiming for a similar DGAF pattern at each SNR. Considering our results (preferred step approx. 16, 18, 21 for 0, -3, and -6 dB SNR, respectively), listeners seem to seek for a tilt close to the original, but at the same time the tilt with enough glimpses in the range $2000 - 4000$ Hz. Figure 7.9b shows that by enhancing the band $500 - 1000$ Hz (higher steps), important frequency regions for speech perception in noise are deactivated, while the opposite is true when band $500 - 1000$ Hz is attenuated (lower steps).



(a) Spectral tilt

(b) Energy modification of band 500 – 1000 Hz

(c) Energy modification of band 1000 – 2000 Hz

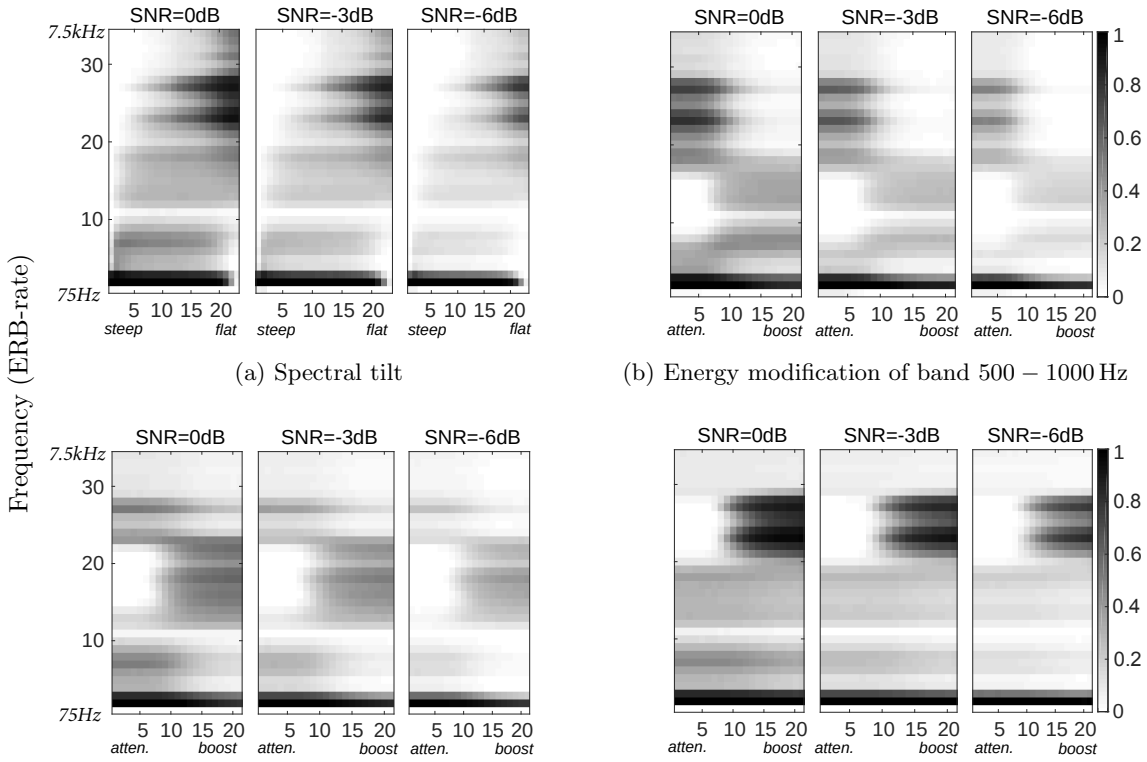
(d) Energy modification of band 2000 – 4000 Hz

Figure 7.8: *Probability of each step (histogram, left axis), along with the percentage of words recalled correctly (black dots, right axis). The error bars represent ± one standard error. The vertical line denotes the step that corresponds to the original speech.*

Listeners chose steps approx. 6, 5, 4 for 0, −3 and −6 dB SNR, respectively, which suggests that for the less noisy conditions they preferred to listen to speech with band 500 – 1000 Hz neither fully activated or deactivated, while for the more adverse condition the enhancement of the 22 – 28 ERB-rate was of high importance. For band 1000 – 2000 Hz (steps approx. 12, 12, 14 for 0, −3 and −6 dB SNR, respectively) listeners preferred to enhance the 16 – 22 ERB-rate (1000 – 2000 Hz), while some glimpses also remained in the upper and lower frequencies. The same can be observed for band 2000 – 4000 Hz, but for the 22 – 28 ERB-rate (preferred steps approx. 14, 15, 17 for 0, −3 and −6 dB SNR, respectively). The poorer intelligibility scores of band 500 – 1000 Hz and 1000 – 2000 Hz compared to those of band 2000 – 4000 Hz might be explained, since the boost of the 2000 – 4000 Hz range was greater for band 2000 – 4000 Hz variation. However, the preferred steps were not only those that corresponded to the maximum enhancement of the region 2000 – 4000 Hz. It is possible that listeners adjusted speech so as to find a balance between their performance and speech naturalness.

Modelling energy reallocation preferences

Figure 7.10a-7.10d show the listener preferences for the different feature values along with the glimpses (GP_{ext}). It can be observed that for some conditions the glimpses follow the



(a) Spectral tilt

(b) Energy modification of band 500 − 1000 Hz

(c) Energy modification of band 1000 − 2000 Hz

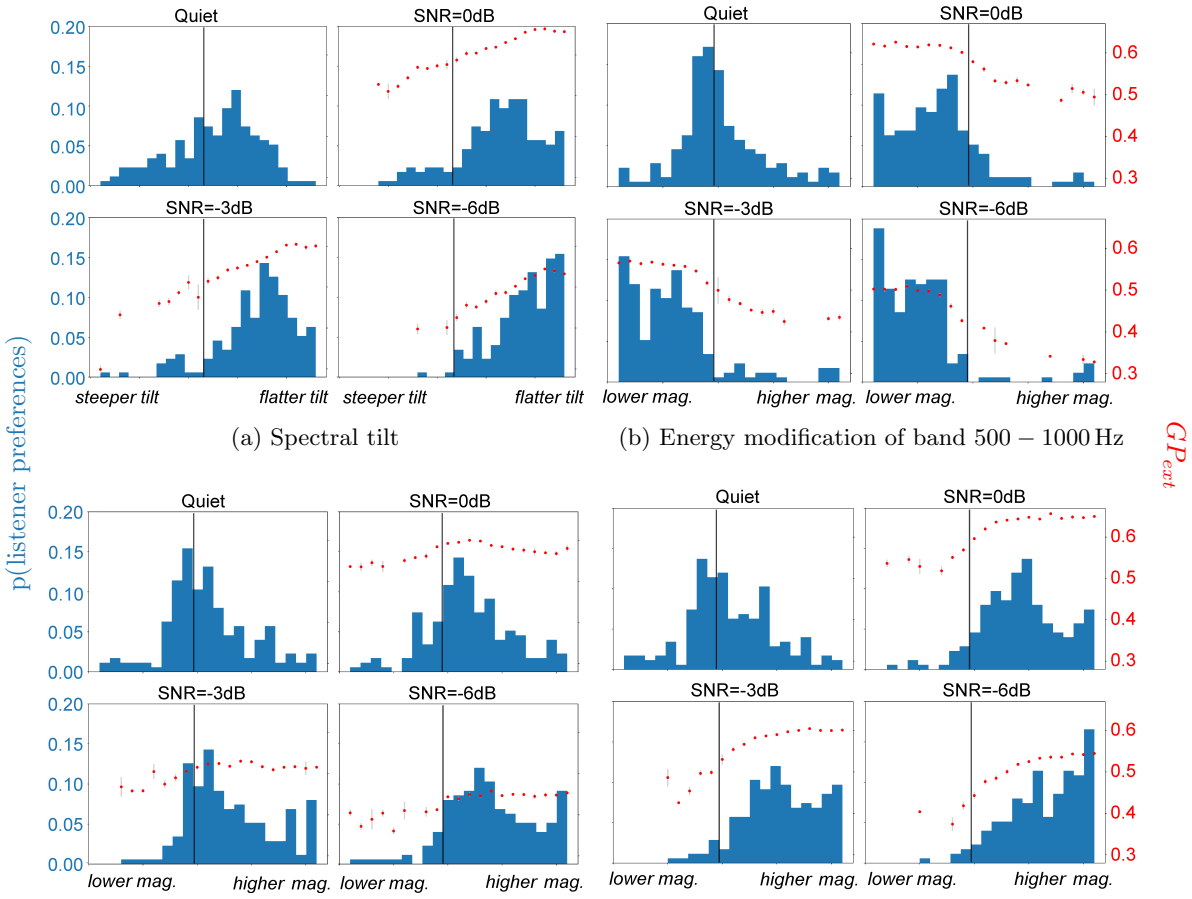
(d) Energy modification of band 1000 − 4000 Hz

Figure 7.9: *Speech energy that survived the energetic masking across frequencies (34 ERB-rate, y-axis) and for each step (x-axis). For each step, the sum of the DGAF (see Fig. 6.3) of 160 utterances from the test phase in the experiment was computed and normalised with the total number of glimpses for all conditions. Black colour denotes that the concentration of speech energy exceeding energetic masking is high.*

distribution of the listener preferences (e.g. Fig. 7.10b), although this is not true everywhere (e.g. Fig. 7.10d). A similar observation can be made for the actual intelligibility scores in Fig. 7.8. Table 7.2 shows the Kullback–Leibler Divergence (KLD) results for comparing the distributions derived from listeners' performance, preferences and glimpse proportion (GP_{ext}) for the different conditions under investigation. As expected, the results showed that the performance and GP_{ext} distributions were similar (symmetric KLD close to zero), whereas each differed to a much greater extent from the distribution of listeners' preferences (higher symmetric KLD values). This suggests that listener preferences encompass information that is not limited to intelligibility.

7.2.4 Interim discussion

In Expt. I, listeners' preferences for spectral tilt and energy in certain spectral bands (500 − 1000, 1000 − 2000, and 2000 − 4000 Hz) in additive speech-shaped noise were investigated. The results reveal whether any change in the long-term speech spectrum, caused by these features, is preferred by listeners. Regardless of the feature, our findings show that listeners preferred similar values to those of unmodified speech for the Quiet condition, while in adverse noise levels they chose to enhance mid-frequencies. As the noise level increased, listeners chose flatter tilt or enhanced energy at higher frequencies, while intelligibility decreased and the adjustment time



(a) Spectral tilt

(b) Energy modification of band $500 - 1000$ Hz

(c) Energy modification of band $1000 - 2000$ Hz

(d) Energy modification of band $2000 - 4000$ Hz

Figure 7.10: *Probability of each feature value (histogram, left axis), along with the glimpses (red dots, right axis). Glimpses computed for the 160 utterances of the test phase. The error bars represent ± one standard error. The vertical line denotes the step that corresponds to the original speech.*

increased. Listeners in noise achieved the highest intelligibility (approx. 90% for -6 dB SNR and 100% up to -3 dB SNR) when they manipulated the spectral tilt and the energy of band $2000 - 4000$ Hz. For these features, listeners also needed less time to find the appropriate step, compared to the other features. Our results showed that the target speech energy in frequency components greater than 4000 Hz did not contribute to increasing intelligibility.

Listeners appeared to make their choices in a way that maximised intelligibility (Fig. 7.10). In line with Lu and Cooke [2009a], it was found that listeners prefer flatter spectral tilts or increased energy in higher frequencies, which also facilitate intelligibility when noise increases, as more speech information survives the energetic masking – i.e. more glimpses [Cooke, 2006] – and thus speech perception performance increases. Under all conditions, the extended glimpsing model follows the distribution of listener preferences well; however, the proportions of the glimpses at ceiling span in a wider region than the peak of the listener preferences. Our experiment also confirmed that the glimpsing model is a good predictor of intelligibility (Table 7.2). However, it failed to predict speech aspects beyond that. This observation strengthens our hypothesis for the second research question of this chapter, that listeners' preferences encompass

feature	SNR(dB)	Performance vs preferences	Performance vs GP_{ext}	GP_{ext} vs preferences
sp. tilt	−6	0.905	0.074	1.814
	−3	3.395	0.064	3.203
	0	0.489	0.009	0.476
0.5-1 kHz	−6	2.278	0.263	4.107
	−3	2.973	0.038	3.436
	0	2.540	0.006	2.363
1-2 kHz	−6	1.496	0.188	1.563
	−3	0.679	0.036	0.715
	0	1.233	0.014	1.457
2-4 kHz	−6	0.961	0.039	1.506
	−3	0.478	0.007	0.468
	0	1.613	0.007	1.437

Table 7.2: *The symmetric Kullback–Leibler Divergence (KLD) derived from the comparison between the following pairs of distributions for the different SNRs: listener preferences and performance, listener preferences and GP_{ext}, and GP_{ext} and listener preferences. The lower the KLD value, the closer the two distributions are. If KLD equals zero, the two distributions are identical.*

information beyond intelligibility.

Another criterion for listeners' choices might be naturalness. Neutral speaking style in quiet is characterised by less flat spectral tilt compared to the speech that a human produces in noise (i.e. Lombard speech) [Summers et al., 1988]. Furthermore, Lombard speech is affected by the level of background noise, with speakers producing speech with a flatter spectral tilt at higher noise levels [Summers et al., 1988; Tartter et al., 1993; Varadarajan and Hansen, 2006]. However, it is not clear what would be the listeners' preferences for other than SSN. It would be interesting to examine whether, for a noise with higher energy at higher frequencies, listeners would choose to shift the spectral energy downwards to avoid noise overlap with the target speech, and therefore to increase the number of glimpses, or whether, as in this experiment, they would prefer to shift the spectral energy in a way influenced by how humans speak naturally in response to high-pass filtered noise [Lu and Cooke, 2009b] or broadband noise [Garnier and Henrich, 2014].

7.3 Experiment II: Effect of frequency bands on listeners' preferences.

A second experiment was conducted in order to test four more spectral energy reallocation techniques. Listeners were allowed to adjust the cut-off frequency of a high-pass filter, the cut-off frequency of a low-pass filter, the bandwidth of a band-pass filter, and which frequency band to enhance (Fig. 7.1[3-6]). To complement Expt. I, in which magnitude modifications were investigated, this experiment explored the frequency areas that listeners would choose to enhance under different noise conditions.

7.3.1 Methods

Listeners

Thirty-seven native Spanish listeners (32 females) aged between 18 and 34 (mean 20.1 years;

SD 2.6 years) participated in this experiment. All listeners passed an audiological screening with a hearing level better than 25 dB, at frequencies in octave steps in the range $125 - 8000$ Hz in both ears. Listeners were paid 10 euros for their participation. Some of the listeners also participated in Expt. I (sec. 7.2).

Stimuli

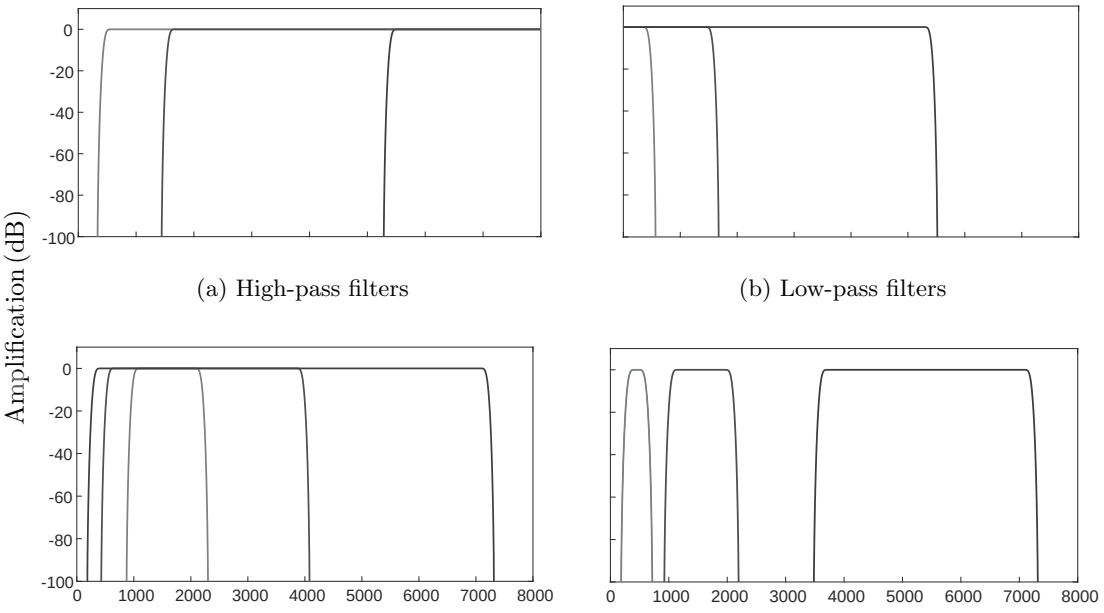
Speech material

The speech material used in this experiment was taken from the same source as in Expt. I (7.2.1), but a different set of test phrases was used.

Stimulus preparation

Four speech features were tested: the cut-off frequency of a high-pass filter, the cut-off frequency of a low-pass filter, the bandwidth of a band-pass filter, and the enhancement of a frequency band using a sliding band-pass filter. For each feature, a filter in the frequency domain was designed and applied to the phrases using the *fir1* and *filter* functions in Matlab 2016b, respectively. An FIR filter with a Chebyshev window of 100 dB of ripple was designed. The Chebyshev filter was used as it has a very good amplitude response.

For the first two features, the middle frequency of 25 log-spaced frequency bands from 300 to 7200 Hz was used as the cut-off frequency of the filter, as shown in Fig. 7.11a for the high-pass filter and 7.11b for the low-pass filter, resulting in 25 different steps. The original speech



(a) High-pass filters

(b) Low-pass filters

(c) Band-pass filters having $989 - 2184$ Hz as central frequency band

(d) Log-spaced band-pass filters from 300 to 7200 Hz.

Figure 7.11: *The filters applied for the two extreme and the middle (step=13) steps of each feature.*

was that of the 1^{st} step for the high-pass filtering and the 25^{th} step for the low-pass. For the

band-pass case, 51 log-spaced frequency bands, starting from 300 Hz and up to 7200 Hz, were computed. The middle band ($989 - 2184$ Hz) was the lowest in bandwidth, corresponding to the 1^{st} step. For each step the spectral components of two bands were added, one before and one after the central band. A total of 25 steps were constructed, with the final (original speech) having 300 Hz and 7200 Hz as cut-off frequencies (Fig. 7.11c). For the last feature, 25 log-spaced band-pass filters from 300 to 7200 Hz were constructed (Fig. 7.11d). None of the steps was close to the original speech. Figure 7.12 shows the spectrograms of a phrase for the original and two extreme steps for the bandwidth of the band-pass filter. The corresponding spectrograms of the remaining features can be found in Appendix D (Fig. D.1-D.3). From the Spanish corpus, the sentence IDs used in this experiment were $1 - 380$ for the adjustment phase, $541 - 701$ for the test phase, while for the practice session 4 random unique phrases were drawn from the adjustment phase set.

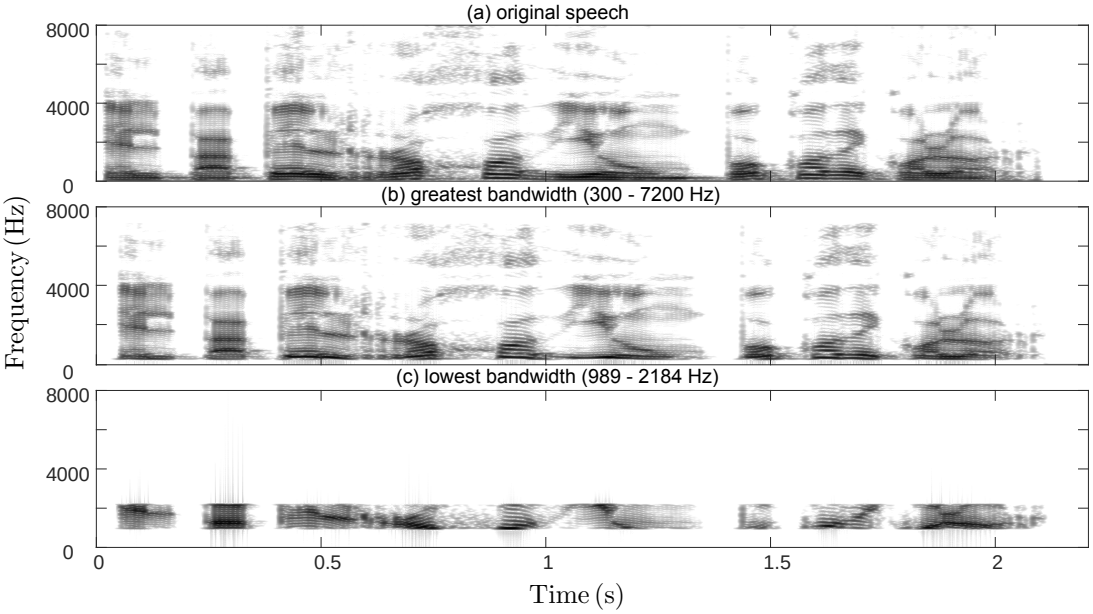


Figure 7.12: *Spectrograms of the phrase 'El papel rojo dio un poco de color' for the different bandwidth band-pass filters. The middle plot corresponds to the phrase with the spectral energy concentrated in the frequency area with the greatest bandwidth, and the bottom plot to that with the lowest bandwidth.*

7.3.2 Procedure

The procedure used in this experiment was identical to that in Expt. I (7.2.2).

7.3.3 Results

Spectral energy allocation preferences

Figures 7.13 and 7.14 show the spectral energy allocation preferences, intelligibility scores, and the time spent in the adjustment phase for the 4 conditions (Quiet, and SSN at 3 SNRs). For the high-pass filter (Fig. 7.13), as noise level increased, listeners preferred higher cut-off frequencies, while even for the quiet condition listeners did not choose speech energy allocated as in the original speech. Furthermore, for SNRs below 0 dB listeners' performance decreased and

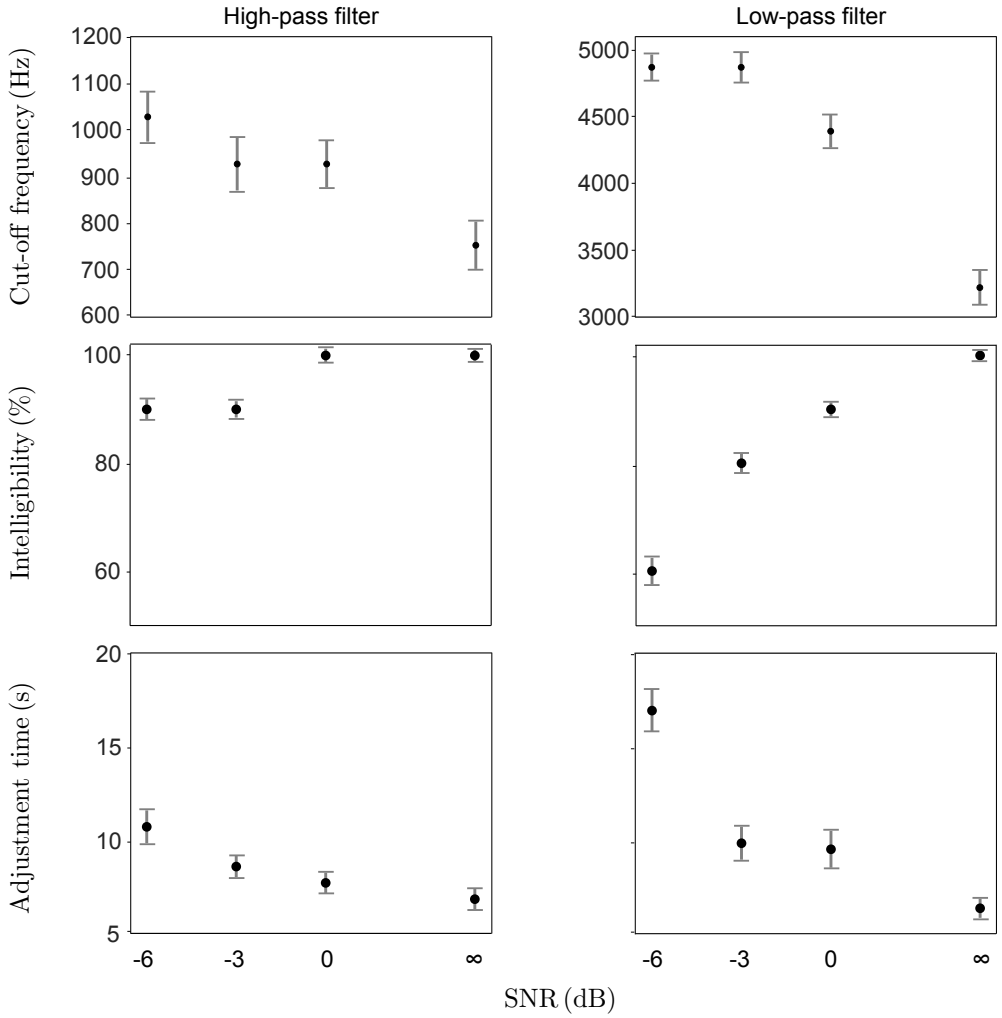


Figure 7.13: *Median values (black dots) for listener preferences (upper plot), intelligibility scores (middle plot) and adjustment time (lower plot) for the different conditions. The left column corresponds to the results for the high-pass filter and the right column those for the low-pass filter. The horizontal lines in the upper plots indicate the unmodified speech. The error bars represent ± one standard error of the median.*

adjustment time increased. For the low-pass filter (Fig. 7.13), similar results can be observed. However, intelligibility scores, apart from the Quiet condition, were much lower compared to the speech choices for the high-pass filter, achieving only 60% at the −6 dB SNR level compared to 90% for the high-pass filter. For the bandwidth of the band-pass filter (Fig. 7.14), listeners preferred to enhance frequencies in the range of $450 - 4500$ Hz for the low-noise and noise-free conditions, while a slightly wider frequency area (around $400 - 5000$ Hz) was preferred under more adverse conditions. Intelligibility score decreased with increasing noise level; however, the score was neither as low as it was for the low-pass filter nor as high as that achieved for the high-pass filter. Finally, for the sliding band filter (Fig. 7.14), as the noise level increased listeners chose to enhance bands located within higher frequency ranges than those preferred under the Quiet condition. Intelligibility scores decreased as the noise level increased, with the maximum score of 70% being achieved under the Quiet condition. Intelligibility scores were generally lower compared to those achieved when the other features were adjusted. For all features, the adjustment time increased with increasing noise level, and the sliding band filter

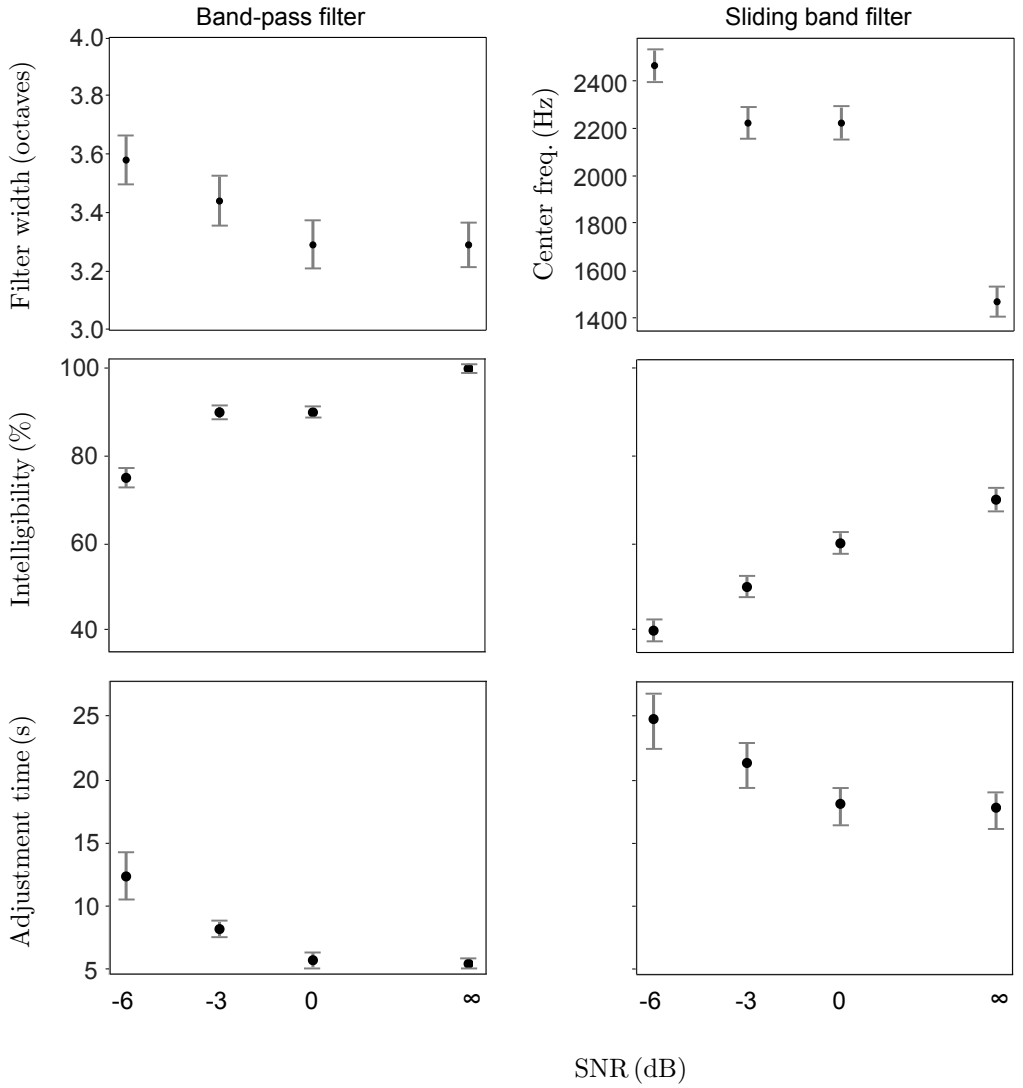


Figure 7.14: *As Fig. 7.13 but for the bandwidth of the band-pass filter and the sliding band. For the bandwidth of the band-pass filter, the centre frequency is around* 1450 *Hz.*

required a substantially longer time, even for the Quiet condition (around 15 s, while for the remaining features the minimum time permitted, around 5 s, was sufficient).

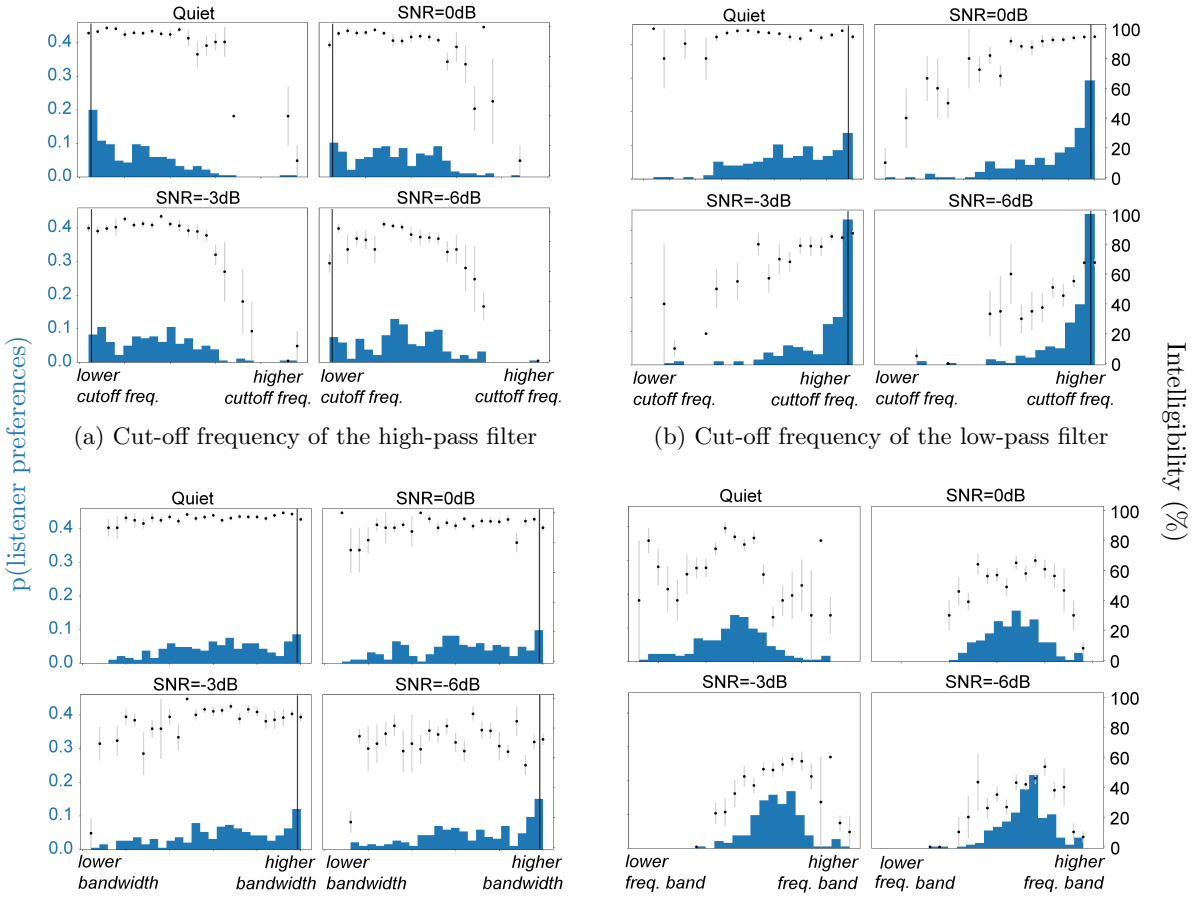
A rank-based, Kruskal–Wallis H-test was conducted to compare the effect of condition on each of the three measurements and each of the tested features. The results indicated significant main effects for all measurements [$p < 0.001$], except for listeners' preferences for the bandwidth band-pass filter [$p = 0.1$]. The results of the statistical analysis can be found in Table 7.3. Post-hoc pairwise comparisons for the high-pass filter indicated that listener preferences differed significantly only between Quiet and each of the remaining conditions, intelligibility scores differed according to the different conditions, and adjustment time was significantly different for all pairs of conditions except for −3 and 0, 0 and Quiet. For the low-pass filter, all pairs of conditions differed significantly for the three measurements (except for the listener preferences between −3 and −6 dB SNR, and adjustment time between 0 and −3 dB SNR). For the band-pass filter, intelligibility and adjustment time differed for all pairs of conditions. Finally, for the sliding band filter, preferences and intelligibility scores differed significantly for the different conditions, while adjustment time differed only for the pairs −6 and 0, −6 and Quiet.

	adjustment time	intelligibility	preferred step
hpf	49.72	87.21	43.26
lpf	162.01	288.73	100.56
bpf	148.39	209.63	ns
sbf	25.32	107.19	202.27

Table 7.3: *H-values of the Kruskal–Wallis statistical test for the different features and measures. hpf, lpf, bpf, and sbf stand for the high-pass, low-pass, band-pass, and sliding band filters, respectively. ns represents a statistically non-significant difference.*

Effects of spectral modifications on listener preferences and intelligibility

Figures 7.15a-7.15d show the probability with which each feature value was preferred by listeners, along with the percentage of keywords perceived correctly. One feature of these results is the poorer intelligibility score obtained by those listeners who preferred to listen to speech filtered by very high cut-off frequencies for the high-pass filter, or very low cut-off frequencies for the low-pass filter, even under the Quiet condition. For the low-pass filter, intelligibility was higher for those listeners who preferred higher cut-off frequencies. Figure 7.15c shows



(a) Cut-off frequency of the high-pass filter

(b) Cut-off frequency of the low-pass filter

(c) Filter's bandwidth

(d) Sliding band filter

Figure 7.15: *Probability of each step (histogram, left axis), along with the percentage of words recalled correctly (black dots, right axis). The error bars represent ± one standard error.*

that listeners' choices were spread over a wide area for all conditions. Although intelligibility was high for all listeners' choices in Quiet, under more adverse conditions there was a drop in intelligibility for those listeners who chose to listen to speech with energy concentrated more in the area around $1000 - 2000$ Hz. In Fig. 7.15d, preferences are seen to be concentrated around a few frequencies for all conditions, following the intelligibility distribution. For some conditions, it can be observed that the median value does not represent the step that was most preferred. This is apparent from the histogram of the high-pass filter for the Quiet condition, which shows that the most preferred step was the lowest one (Fig. 7.15a; cut-off frequency of 450 Hz) while the median preferred value was approx. 755 Hz (Fig. 7.13). A similar association was also observed in the low-pass filter histograms, where the highest step (cut-off frequency of 5400 Hz) was the most preferred under all conditions, but more so as the noise level increased (Fig. 7.15b). The median preferences were approx. 4869 Hz for -6 and -3 dB SNR, 4390 Hz for 0 dB SNR and 3218 Hz for Quiet (Fig. 7.13). Finally, for the bandwidth of the band-pass filter the median value again did not correspond to the step that was most preferred. The most preferred bandwidth was the widest one ($300 - 7200$ Hz), and the number of people choosing this step increased with the noise level (Fig. 7.15c). The median preferences were around $447 - 4838$ Hz for all conditions (Fig. 7.14).

Energy reallocation choices and energetic masking

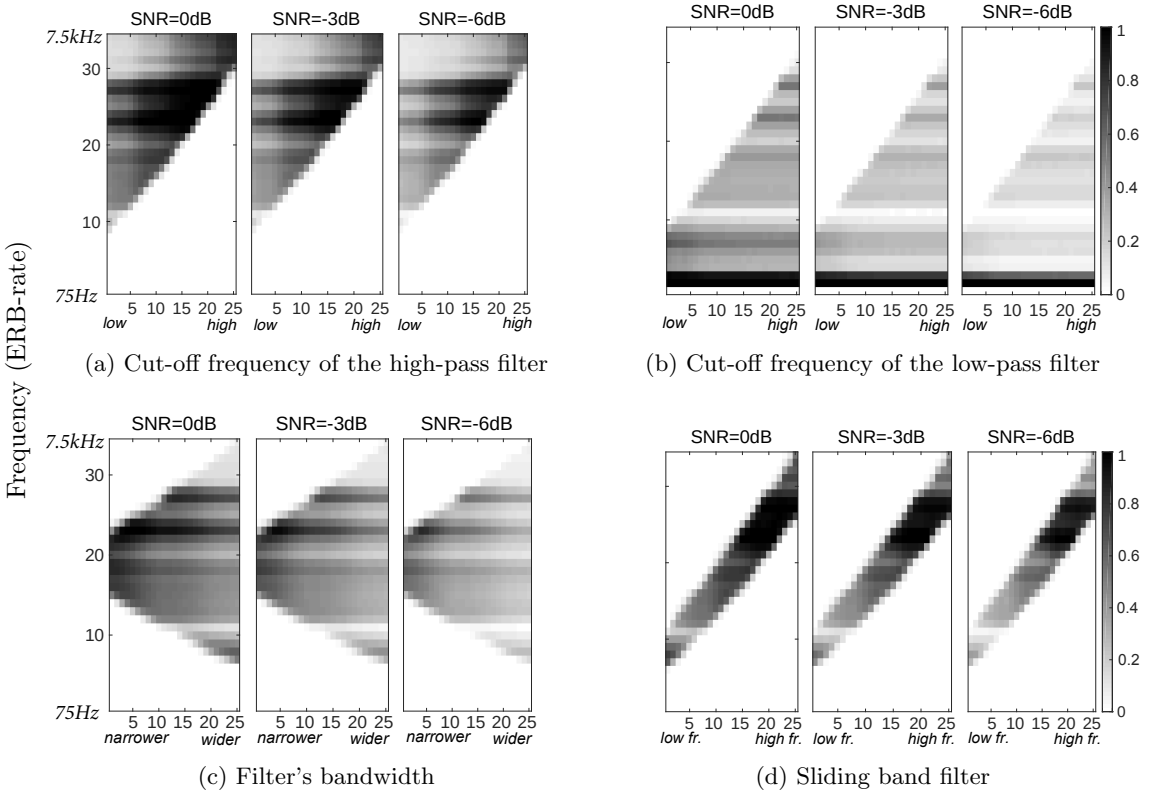
Figures 7.16a-7.16d show the speech energy that survives energetic masking for the different conditions. For all features, as noise level increases fewer areas survive the masking. As in Expt. I, listeners chose steps for which the speech energy in the range $2000 - 4000$ Hz ($22 - 28$ ERB-rate) was high and glimpses outside this range also exist. The bands chosen by listeners corresponded to frequencies higher than those dominated by noise ($2 - 16$ ERB-rate) and focused on frequencies important for speech perception.

Modelling energy allocation preferences

Figure 7.17a-7.17d show the listener preferences for the different feature values along with the glimpses (GP_{ext}). It can be observed that for some conditions the glimpses follow the listeners' preferences distribution (e.g. 7.17d); however, this is not true everywhere (e.g. 7.17c). For the sliding band filter at 0 dB SNR, the peak of glimpses (band of $2145 - 4291$ Hz) is not identical to that of preferences (band of $1572 - 3145$ Hz), while for higher noise levels the peaks coincide. Table 7.4 shows the KLD results from comparing the distributions derived from listeners' performance, preferences and glimpse proportion (GP_{ext}) for the different conditions. As in Expt. I, listeners' performance and GP_{ext} distributions were close to each other (symmetric KLD close to zero) and differed less (smaller symmetric KLD value) compared to the difference between them and the listeners' preferences distribution.

7.3.4 Interim discussion

In Expt. II, listeners' preferences in relation to frequency band enhancement were investigated. Specifically, the experimental features were the cut-off frequencies of high- and low-pass filters, the bandwidth and the cut-off frequencies of band-pass filters. To model listener preferences, the extended glimpsing model was used. Results showed that the model described the preferences



(a) Cut-off frequency of the high-pass filter

(b) Cut-off frequency of the low-pass filter

(c) Filter's bandwidth

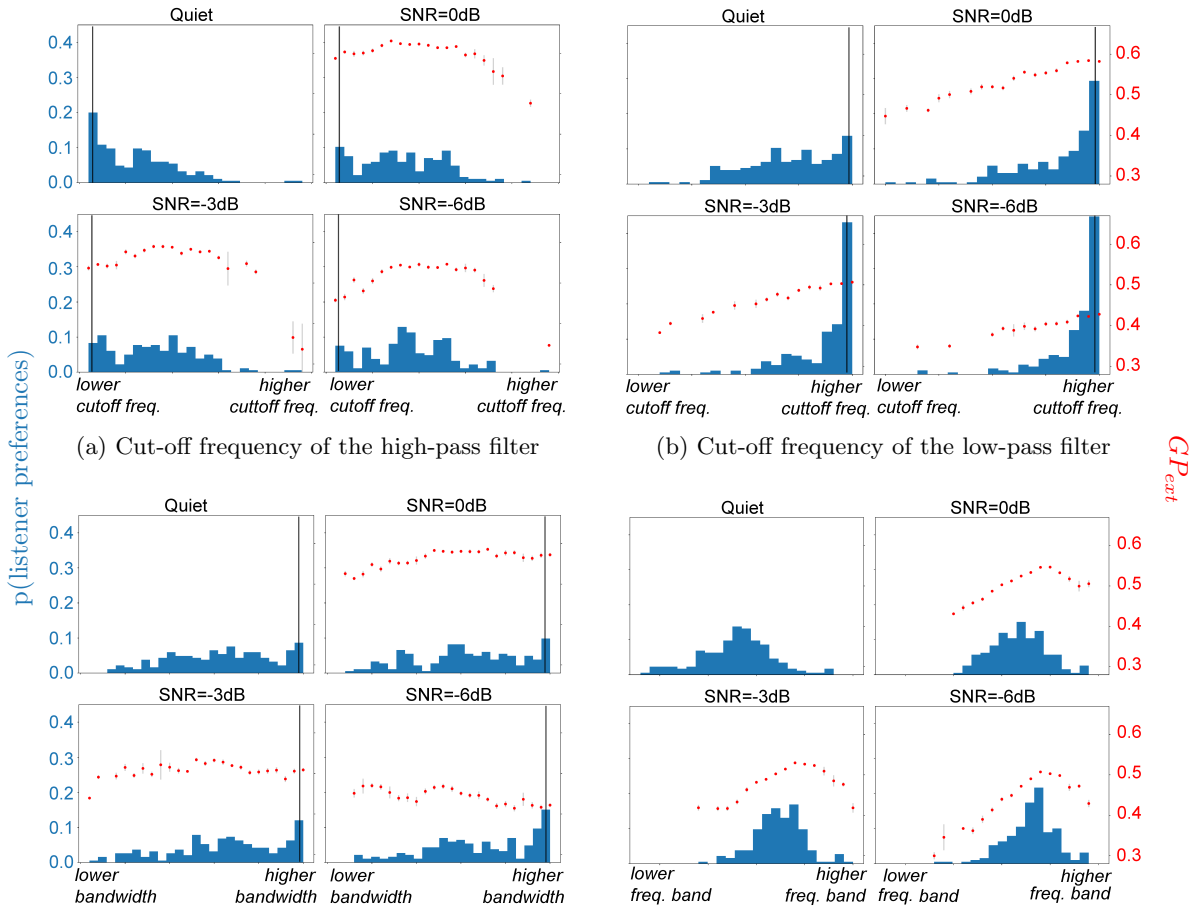
(d) Sliding band filter

Figure 7.16: *Speech energy that survived the energetic masking across frequencies (*34 *ERB-rate, y-axis) and for each of the 25 steps (x-axis). For each step, the sum of the DGAF (see Fig. 6.3) of 160 utterances from the test phase in the experiment was computed. lcf and hcf on the x-axis denote the lower and higher cut-off frequencies, respectively.*

feature	SNR(dB)	Performance vs preferences	Performance vs GP_{ext}	GP_{ext} vs preferences
hpf	−6	2.722	0.105	3.486
	−3	1.230	0.167	2.962
	0	1.042	0.075	1.471
lpf	−6	4.542	0.298	5.663
	−3	3.136	0.124	4.946
	0	2.994	0.079	3.241
bpf	−6	0.541	0.049	0.534
	−3	1.062	0.053	1.220
	0	0.432	0.004	0.418
sbf	−6	1.061	0.399	1.685
	−3	1.304	0.249	1.795
	0	0.370	0.116	0.588

Table 7.4: *The symmetric Kullback–Leibler Divergence (KLD) derived from comparisons between the following pairs of distributions for the different SNRs: listener preferences and performance, listener preferences and GP_{ext}, and GP_{ext} and listener preferences. The lower the KLD value, the closer the two distributions are. If KLD equals zero, the two distributions are identical. hpf, lpf, bpf, and sbf stand for the high-pass, low-pass, band-pass, and sliding band filters, respectively.*

well (Fig. 7.17). As expected, one of the listeners' choices in noise was to enhance the amount of speech spectral energy that escapes the energetic masking. Possible reasons for the listeners' preferences are discussed below.



(a) Cut-off frequency of the high-pass filter

(b) Cut-off frequency of the low-pass filter

(c) Filter's bandwidth

(d) Sliding band filter

Figure 7.17: *Probability of each step (histogram, left axis), along with the glimpses (red dots, right axis) for the different features. Glimpses computed for the 160 utterances of the test phase. The error bars represent ± one standard error.*

In our experiment, listener preferences for high-pass filtered speech resulted in high task performance. In Quiet, listeners had a distinct preference for the lowest available cut-off frequency (450 Hz), even though intelligibility was at ceiling for a wider range of steps. Task performance was lower for those listeners who chose cut-off frequencies above approximately 1700 Hz (Fig. 7.15a). This is in line with the results in McClurg [2018], where in quiet, intelligibility was high when utterances were high-pass filtered at 700 Hz (the lowest tested cut-off frequency) and significantly declined at 1973 Hz, while when speech was high-pass filtered above 3000 Hz it became unintelligible.

For the intermediate conditions (Fig. 7.15a at 0 and −3 dB SNR), listener preferences are unclear, i.e. there is no dominant peak. The probability mass of the preferences is concentrated in the area where intelligibility is at ceiling. However, for the most adverse noise level, listeners chose to enhance more the speech energy in frequency ranges above the energetic masker (most preferred cut-off frequency around 1030 Hz). Not only did this choice allow frequencies higher than 1030 Hz to pass, but also the overall energy level of this region was higher compared to those for lower cut-off frequencies. Listeners may have made a compromise between these two effects. Several studies have shown the relevance of shifts in the spectral energy distribution

to the mid-frequency region, especially above 1000 Hz, in relation to intelligibility. Such shifts are also observed in Lombard and clear speech [Krause and Braida, 2004; Skowronski and Harris, 2006; Lu and Cooke, 2009b]. Niederjohn and Grotelueschen [1976] suggested high-pass filtering followed by amplitude compression for enhancing F2 and F3 formants while supressing F1. Their results revealed a significant effect on intelligibility in white noise. Furthermore, in Godoy and Stylianou [2012], intelligibility gains were reported when the frequency region $1000 - 4000$ Hz was boosted while maintaining the overall energy of the signal.

Unlike high-pass filtering, which reduces the low-order harmonics that are important for pitch perception, low-pass filtering retains this information. For the low-pass filtered speech, listeners chose higher cut-off frequencies in noise (5400 Hz) compared to Quiet (3218 Hz). Low-pass filtering has been widely used in speech perception research, since it has been shown that almost all cues of speech intelligibility are contained within the low frequency region of the speech spectrum, up to 4000 Hz [Fletcher and Galt, 1950; French and Steinberg, 1947], which is consistent with our findings (listeners in Quiet chose cut-off frequencies lower than 4000 Hz and achieved close to 100% intelligibility). Low-pass filtering retains lower frequency acoustic energy including the tonal quality of the speech which maintains prosodic features such as pitch range, intonation contour and rhythm. Frota et al. [2002] showed that prosodic differences can be perceptually detectable within the low 400 Hz frequency region.

For the bandwidth modifications, in contrast to what we would expect, listeners did not choose to enhance mid-frequencies more in noise than in Quiet. The preferred range in all conditions was $300 - 7200$ Hz (the highest modification step). One explanation is that the energy level in this band is higher than that of the original plain speech (because of RMS normalisation), resulting in more glimpsing areas. However, as Fig. 7.17c shows, narrower bands allow more speech energy to escape masking. Listeners may have made their choice so that the output speech sounded less artificial than if they had chosen a narrower band. In telephony a range comparable to this is used (adaptive multi-rate wideband of $50 - 7000$ Hz [3GPP TS 26.190, 2005]). This range enhances the clarity, resulting in greater intelligibility and better talker recognition, while the speech also sounds more natural, expressing emotions more precisely.

7.4 General discussion

In this chapter, the impact of spectral energy redistribution on listener preferences and intelligibility in quiet and speech-shaped noise at three different noise levels (-6, -3, and 0 dB SNR) was explored. In total, 6 features were tested (Fig. 7.1): namely, spectral tilt, energy of certain spectral bands, cut-off frequencies of high- and low-pass filters, bandwidth of a band-pass filter, and frequency band to enhance. The tested features were split up into two experiments. For the experiments, native Spanish listeners were recruited and SpeechAdjuster was used to collect their responses.

RQ1: Do listeners choose to reallocate speech spectral energy so to maximise intelligibility?

Overall, the listeners' choices demonstrated that their main priority was to maximise intelligibility. For all conditions, listener preferences were in the range where intelligibility was at ceiling (as estimated by the extended glimpsing model; Fig. 7.10 and 7.17). In noise, listeners

chose to shift the speech energy (under a regime of constant energy overall) in frequencies above 1000 Hz. This mechanism reduces the effect of energetic masking that speech-shaped noise imposes and thus, the spectral energy is enhanced in the range where the ear is more sensitive. Previous studies have shown that enhancing mid and higher frequencies improves intelligibility. Krause and Braida [2004] showed that the shift of spectral energy to high frequencies helps to increase the intelligibility of clear speech relative to conversational speech in the presence of speech-shaped noise. Tang and Cooke [2018] investigated speech modification strategies based on reallocating energy statically across the spectrum, using masker-specific spectral weightings. The RMS level before and after the modifications was kept constant. They concluded that generic spectral weighting patterns that boost energy above 1000 Hz are beneficial for maskers with a speech-shaped long-term spectrum. Finally, in Niederjohn and Grotelueschen [1976], high-pass filtering was used to enhance the energy in high frequencies and supress the first formant in white noise. The enhancement of the second formant relative to the first may have led to improvements in intelligibility.

RQ2: Do listeners' preferences show patterns that are independent of intelligibility?

Listeners may have made their choices so as to decrease processing demands, and hence to reduce listening effort. Lexical retrieval can be facilitated by a greater number of glimpses, i.e. more acoustic information available, so the acoustic mismatches between the target speech and the mental representation are reduced [Rönnberg et al., 2013]. A study by Borghini and Hazan supports this speculation [Borghini and Hazan, 2020]. They conducted a pupillometry experiment to test the impact of clear versus plain speech on listening effort. Listeners were presented with the stimuli in babble noise while intelligibility level was equated (SNR was individually adjusted to target a 50% intelligibility level). Pupil data revealed that clear speech requires less listening effort. One of the characteristics that differentiates clear from plain speech is greater energy in the mid frequencies. This allows more energy to escape the babble noise masking and thus might have contributed to the reduced listening effort.

Listeners also may have chosen to maintain the speech quality. In quiet, listeners did not choose to reallocate spectral energy. One explanation could be that the closer the average spectrum to the original speech, the more natural the speech sounds. Moore and Tan [2003] determined how the perceived naturalness of speech is affected by several spectral distortions in quiet. Their results revealed that spectral tilt modifications degrade naturalness, especially when they are applied over the whole frequency range. On the other hand, according to our experiments, in noise listeners shifted the energy to higher frequencies. High-frequency regions include cues that may contribute to the perception of sound quality. In Gabrielsson et al. [1988], the effect of different frequency responses on speech quality was evaluated. The most preferred system for all tested conditions (quiet and in noise at +10 dB SNR) was characterised by a flat response at lower frequencies (below 1000 Hz) and a 6 dB/octave increase above that (1000 − 4000 Hz). Such an increase at higher frequencies led to an improvement in brightness, nearness (sounds close to the listener), spaciousness (sounds open and spacious), clarity (sounds clear, distinct and pure), and total impression (an overall judgement of how good the listener thinks the reproduction is), and a decrease in softness (sounds soft and gentle).

Another supra-intelligibility aspect of speech related to listener preferences may be speech familiarity. Both in quiet and in noise, the preferred spectral energy modifications are similar to those that a talker naturally produces under corresponding conditions. Specifically, in quiet,

listeners adjusted speech to be close to original/plain speech, which is the speaking style used under real-life noise-free conditions. In noise, listeners chose to enhance mid and higher frequencies, as in Lombard and clear speech types. Lombard speech produced by male speakers has slightly less energy in the $0-1000\,\text{Hz}$ region compared to speech produced in quiet [Garnier and Henrich, 2014], while in noise male talkers increase the energy between $2000-4000\,\text{Hz}$ [Junqua, 1993; Castellanos et al., 1996]. In Garnier and Henrich [2014], the speech spectrum in noise was significantly enhanced in the regions $1000-2000\,\text{Hz}$ and $2000-4000\,\text{Hz}$, while energy in frequencies above $4000\,\text{Hz}$ was lower compared to conversational speech in quiet. Such enhancements are also observed in opera singers ('singing formant') and in stage actors ('actor's formant'), who enhance the speech energy around $1000-4000\,\text{Hz}$ for projecting their voice over a distance [Bele, 2006].

RQ3: Do listeners' preferences change under challenging conditions?

As noise level increases, listeners preferred a greater increase in the energy at higher frequencies, e.g. by choosing a flatter spectral tilt. Listeners chose the feature values that do not severely attenuate pitch and harmonic information. Specifically, they did not choose the flattest spectral tilt options, for which the normalised DGAF shows a high energy attenuation of low frequencies. For instance, this can be observed in the listeners' choices for the spectral tilt feature under intermediate conditions (Fig. 7.9a). In previous studies, algorithms that boost mid and high frequencies, sacrificing energy below $1000\,\text{Hz}$, were judged as poorer in quality compared to others [Gabrielsson et al., 1988; Tang et al., 2018]. Another explanation could be that the available speech cues in the low frequencies also benefit the match between the acoustic signal and the existing representations in the long-term memory, reducing cognitive processing. Finally, for the spectral tilt feature at only the most adverse noise level, listeners preferred to increase intelligibility at the expense of the energy in low frequencies. Under this condition the steps for which intelligibility was at ceiling were limited (Fig. 7.10a). Thus, this outcome supports the hypothesis that listeners' priority was to maximise intelligibility.

To sum up, once intelligibility was no longer an issue, listeners based their choices on other criteria. For the conditions where intelligibility was at ceiling for a range of steps, preferences were not uniformly distributed across those steps. In almost all the cases, a peak emerged, orientated towards the enhancement of lower frequencies. This observation supports our earlier speculations that listeners may attempt not to harm the speech quality and to reduce listening effort.

Chapter 8

Conclusions

8.1 Summary

Artificially enhanced speech is not always satisfying for the listener. Speech enhancement algorithms focus mainly on improving intelligibility, while speech aspects beyond intelligibility that also have an impact on the listener have been far less thoroughly investigated. The topic of this thesis was inspired by the fact that the 'optimal' quality for user-centric applications can only be estimated with respect to the listener, and the purpose of this research was to test the impact of different speech types and of distinct speech characteristics specifically on supra-intelligibility aspects of speech. First, measures of listening effort were collected for different speaking styles and comparisons were drawn between native and non-native listeners (chapter 3). Then, a tool named SpeechAdjuster (chapter 4) was introduced and implemented to investigate intelligibility and supra-intelligibility aspects of speech. SpeechAdjuster was used in a series of experiments that were conducted to collect, along with intelligibility scores, listeners' preferences for distinct speech parameters that were selected to be features that talkers naturally modify in noisy conditions. The features and conditions tested were the speech rate in stationary and temporally-modulated noise (chapter 5), the fundamental frequency in the presence of stationary noise and competing speech (chapter 6), and other features that led to spectral energy reallocation in stationary noise (chapter 7). Finally, a measure was introduced for determining the glimpse distribution of an utterance across frequencies (DGAF) (chapter 6).

8.2 Outcomes

8.2.1 Innovations

- A tool, called SpeechAdjuster, was developed that allows the manipulation of almost any aspect of speech and supports joint elicitation of listener preferences and intelligibility measures (chapter 4).

- The DGAF measure was introduced to determine the glimpse distribution of an utterance across frequencies, i.e. the mean of glimpses across the time series of an utterance for each different frequency band (chapter 6).

8.2.2 Main findings

- Listeners showed distinct preferences for the tested speech features, revealing aspects of speech beyond intelligibility. Specifically, for constant intelligibility:
  - In quiet, listeners adjusted speech to be close to the original (i.e. plain) speech.
  - For modulated maskers, listeners preferred speech rates modulated in a way that contrasted to those of the masker (chapter 5).
  - For stationary noise, listeners preferred a slower speech rate as the noise level increased (chapter 5).
  - Regardless of the noise level, listeners chose a slightly lower mean fundamental frequency compared to the original (chapter 6).
  - For stationary noise, listeners preferred to reallocate the speech energy, choosing settings that enhanced the energy in the lower frequencies (chapter 7).
- The more cognitively demanding the task was, the greater the adjustment time the listeners needed.
  - In quiet, they needed around the lowest permitted time (5 s).
  - In noise, they needed progressively more time depending on the noise level.
  - They needed different times for the different masker types (e.g. more time for speech-modulated noise compared to speech-shaped noise for the same SNR; chapter 5).
  - They needed different times for features that cause different acoustical or phonological distortions (e.g. the sliding band filter required the greatest time; chapter 7).
- Subjective ratings of effort were correlated with intelligibility, while they were not always consistent with the pupillary responses (chapter 3).
- A clear impact of speech type on the cognitive demands required for speech comprehension was apparent. Pupil dilation determined that Lombard speech imposes smaller demands on mental processing compared to plain, TTS, and artificially enhanced speech (chapter 3).

The aforementioned findings are discussed in the next section in an attempt to interpret listener preferences.

8.3 Interpreting listener preferences

Results revealed clear preferences for the different speech features and conditions tested. Listeners tended to opt for feature values that would improve intelligibility, as well as speech aspects beyond that once understanding the speech was no longer an issue. Listeners were instructed to tune the speech until they could recognise as many words as possible, without any additional information on how to make the adjustments. For all the listener preference experiments, the total energy of the stimuli before and after the modifications was kept constant. Thus, listener preferences reflected the best compromise between the effect of the parameter being modified and the effect on local SNR change in time-frequency. An estimate of the local SNR can be

indicated by the proposed DGAF measure. Tang and Cooke have investigated the gains in intelligibility and the effects on quality when local SNRs are enhanced using several strategies [Tang and Cooke, 2010, 2011]. However, an understanding of the beneficial effects on local SNR that made listeners in our experiments choose the specific feature values will require further analysis.

Listener preferences revealed an attempt to reduce cognitive demands.

According to the Ease of Language Understanding model [ELU; Rönnberg et al., 2013], in ideal conditions, speech understanding is an implicit, automated, and effortless process, while distorted speech (e.g. noisy conditions, signal processing, hearing loss) is detrimental to this process. To make sense of a distorted speech signal, the top-down cognitive analysis is enhanced [Gatehouse, 1990; Pichora-Fuller et al., 1995; Wingfield, 1996] and explicit cognitive processing is triggered, requiring more cognitive resources. The limited capacity of the working memory [Kahneman, 1973] makes such a task effortful. The degree of explicit processing needed for speech understanding is positively linked to effort [Rönnberg et al., 2019]. In our experiments, under quiet conditions, listeners may have chosen the original (i.e. plain) speech, as the distortions caused by the speech processing for the remaining options may have led to phonological mismatches with the expected mental representation. In noise, listeners chose feature values in such a way as to overcome the energetic masking by avoiding it, either in time, i.e. choosing a target speech rate that contrasts with that of the speech modulated noise masker (chapter 5), or spectrally, i.e. reallocating spectral energy when speech is masked with speech-shaped noise (chapter 7). These preferences may have derived from the listeners' desire to reduce listening effort. In noise, part of the acoustic information is masked, resulting in decreased audibility. Missing or incomplete segments of the target speech lead to a mismatch with the stored lexical representation; thus, the acoustic signal requires more perceptual processing to interpret speech. This process results in greater listening effort [Winn and Teece, 2021]. Listeners under the speech-shaped noise condition may have selected slower speaking rates as the noise level increased, in order to extend the time available for processing speech (chapter 5).

The more cognitively demanding a condition is, the more time the listener may need to find the 'optimal' value. Listeners in our experiments were under no time constraint while performing the task; thus, they could spend as much time as necessary on perceptual learning. The perceptual system is capable of recalibrating speech processes and adapting to distortions that the speech imposes [Samuel and Kraljic, 2009]. In quiet, speech understanding is supposed to be effortless; thus, the listeners in these experiments needed close to the shortest permitted time to finalise their choice. For increasing noise levels, the processing demands increase and the adjustment time became progressively higher. For instance, for the spectral tilt feature at -3 and $0\,\mathrm{dB}$ SNR, the listeners achieved almost equal intelligibility scores, whereas they needed more time to adjust the speech under the more adverse conditions (chapter 7). The extra processing demands required in conditions with more noise were reflected in pupillary activity, a known measure of listening effort in which peak pupil dilation increases with noise level [Ohlenforst et al., 2017]. Finally, apart from the additional cognitive demands that the higher noise level imposes, different types of masker also had different impacts on effort [Brungart et al., 2013]. For the experiment in chapter 5, listeners spent more time adjusting the speech rate in the presence of temporally-modulated noise compared to stationary noise. The modulated nature of the masker may impose an additional cognitive load on the listener.

Listeners may have tried to maintain the speech quality.

Listeners chose feature values that do not have a negative impact on naturalness. Specifically, listeners chose a similar fundamental frequency to the original speech, since a simple shift in fundamental frequency without the appropriate formant adjustments has a negative effect on naturalness [Assmann et al., 2006]. In line with our findings, a simple shift in fundamental frequency was not preferred by the listeners in Assmann and Nearey [2007]. Moreover, in noise, listeners in our experiments did not choose the options that entailed extreme attenuation of pitch and harmonic information. In previous studies, the quality of algorithms that boost mid and high frequencies, sacrificing energy below 1000 Hz, was judged as poorer compared to others [Gabrielsson et al., 1988; Tang et al., 2018].

Listeners may have made their choices based on what they find familiar.

In Tang et al. [2018], listeners in quiet conditions with intelligibility at ceiling preferred plain speech over modified speech, while plain speech was rated to have better quality. Listeners were not given any specific reference for the quality assessment, but one explanation is that they applied a consistent quality standard based on their experience. In our experiments, in noisy conditions, listeners chose speech features similar to those naturally produced by speakers in noise. Specifically, listeners preferred slower speech rates and flatter spectral tilts, while for higher noise levels the effect was greater [Tartter et al., 1993]. Speech in noise is also characterised by an increase in pitch, which was not observed in our results. The deterioration in quality may have discouraged the listeners from selecting a higher pitch.

Do all supra-intelligibility aspects of speech culminate in listening effort?

Listener choices influenced by speech quality and familiarity may well lead to a reduction in listening effort. Speech distortions that lead to poorer speech quality may require the investment of higher cognitive resources. It has been shown that, for constant intelligibility, changes in signal quality, such as increased spectral resolution in a cochlear implant simulation, can result in decreased listening effort measured using the dual-task paradigm [Pals et al., 2013]. Additionally, naturally produced speech types that listeners are accustomed to hearing under specific conditions have been shown to be less cognitively demanding. In Borghini and Hazan [2020], for the same level of speech intelligibility, the cognitive effort increased when attending to plain instead of clear speech in the presence of babble noise. Furthermore, this study found (chapter 3) that Lombard speech was the least effortful compared to plain and artificial speech types in the presence of speech-shaped noise. Speech characteristics of clear and Lombard speech were preferred in our experiments under noisy conditions (i.e. slower speech rate, flatter spectral tilt). However, the distinct speech features that led to the reduction in listening effort should be further investigated.

8.4 Potential research directions

One direction may be to develop a model to predict listener preferences. Such a model could be useful in the optimisation of speech enrichment algorithms. In near-end listening enhancement algorithms, objective intelligibility metrics are usually used to optimise the algorithm's parameters. However, an objective intelligibility metric cannot really distinguish between conditions of ceiling intelligibility, which can only be understood in terms of listeners' preferences. SpeechAdjuster can be valuable for determining the optimal parameters in audio engi-

neering, in which the level of one audio signal is reduced by the presence of another signal, or for achieving the proper balance between intelligibility and supra-intelligibility aspects of speech important for near-end listening enhancement algorithms. In the development of such a predictive model, the DGAF measure can be useful to indicate the preferred spectral profile.

One limitation of this thesis is that the relationship between listener preferences and supra-intelligibility factors was not investigated. Only indirect comparisons were drawn between listener preferences and listening effort, quality and familiarity. Thus, another research direction could be to extend the relationship between the outcomes from tools such as SpeechAdjuster to other measures. For instance, it would be of interest to compare SpeechAdjuster-elicited preferences with listening effort metrics based on pupillometry [Winn et al., 2018], EEG [Sauseng and Klimesch, 2008; Obleser et al., 2012] or self-reports [Gatehouse and Noble, 2004; Rudner et al., 2012].

8.5 Endpiece

The overarching objective of this thesis is to determine whether listeners exhibit supra-intelligibility preferences when they are given the power to manipulate distinct speech properties. The collective outcome of the experiments described here suggests that the answer is yes. Listeners in noise chose to elongate speech, chose spectral modifications that cause the least possible damage to the lower frequencies, and selected slightly lower fundamental frequencies compared to the original speech. Dissecting the relationship between listener preferences and quality, naturalness and cognitive effort is a fruitful area for future research.

Appendix A

Trial exclusions in chapter 3

Table A.1 depicts the percentage of trials with which native (left table) and non-native (right table) listeners did not perceive any word correctly for each speech type and SNR.

		Speech type			
		SSDRC	plain	TTS	Lombard
SNR	−5	0.3%	6.1%	23.3%	0.6%
	−3	0.3%	1.7%	13.3%	0.3%
	−1	0%	0.8%	8.3%	0.6%

		Speech type			
		SSDRC	plain	TTS	Lombard
SNR	−1	0.5%	3.1%	5.6%	1.2%
	+5	0.3%	1.3%	2.3%	0.2%
	+20	0.1%	0.1%	1.1%	0.1%

Table A.1: *Percentage of trials with 0 words recalled correctly by native (left table) and non-native (right table) listeners.*

Appendix B

Accent evaluation - web test in chapter 3

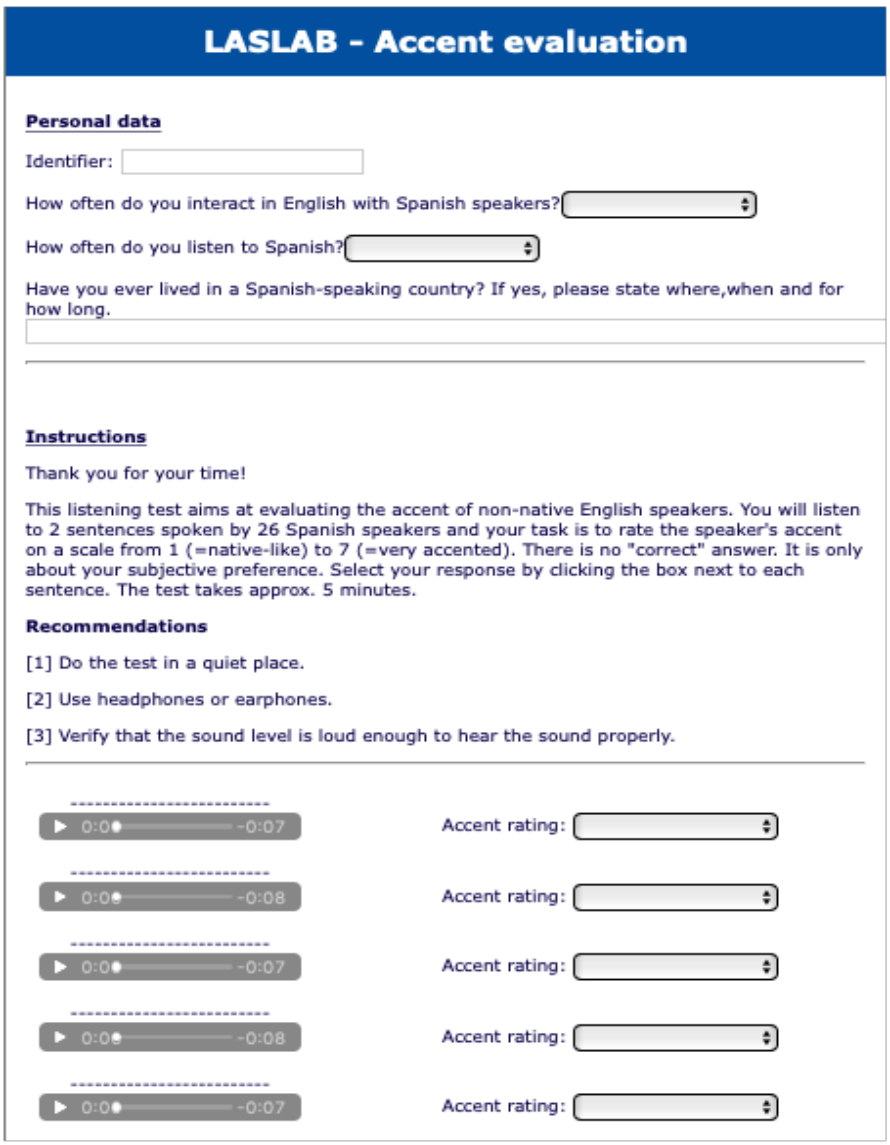


Figure B.1: *Online test with which native British English listeners evaluated the accent of the non-native listeners.*

Appendix C

Individual differences in chapter 6

C.1 Experiment I

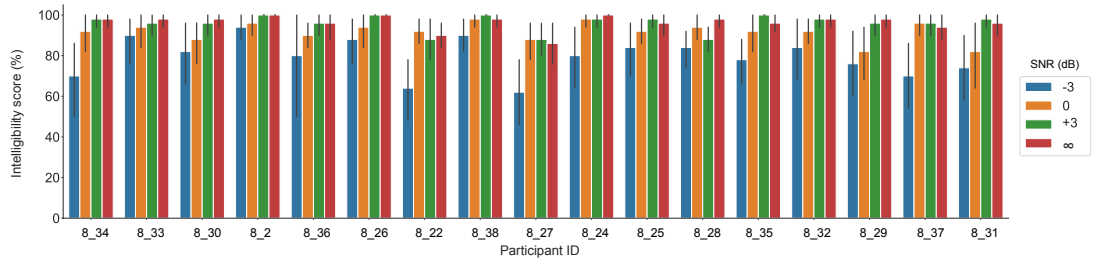


Figure C.1: Bar plot of the mean f_0 intelligibility scores for each listener. The error bars represent the 95% confidence intervals. Ordering on x-axis is based on the listeners' mean f_0 preferences at -3 dB SNR.

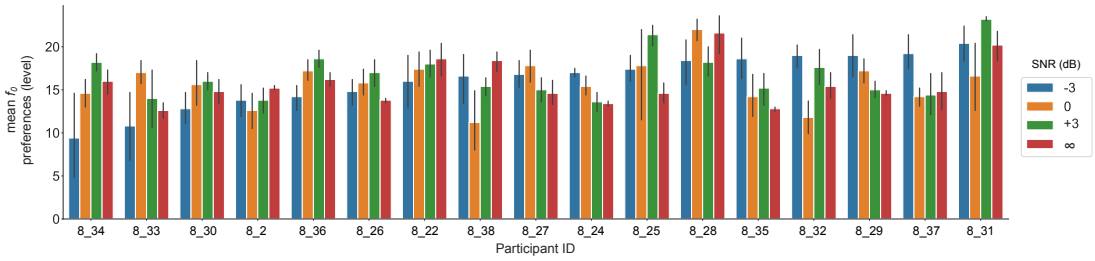


Figure C.2: Similar to C.1 but for the mean f_0 preferences.

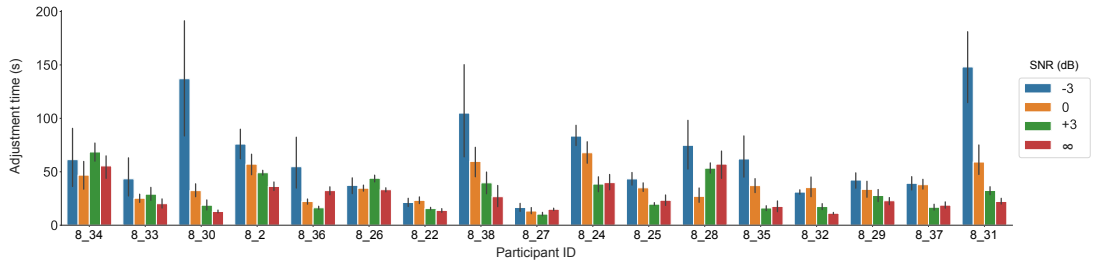


Figure C.3: Similar to C.1 but for the time required for choosing the preferred mean f_0.

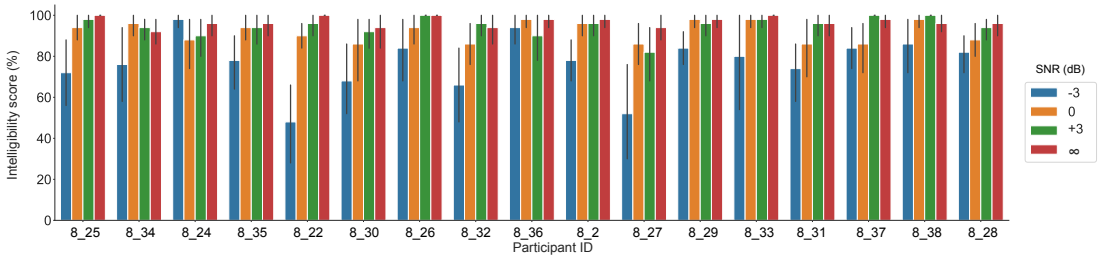


Figure C.4: Bar plot of the intelligibility scores achieved with the preferred f_0 variation for each listener. The error bars represent the 95% confidence intervals. Ordering on x-axis is based on the listeners' f_0 variation preferences at -3 dB SNR.

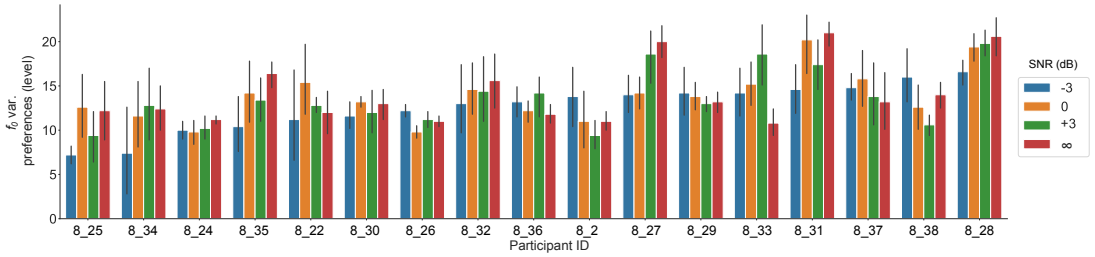


Figure C.5: Similar to C.4 but for the f_0 variation preferences.

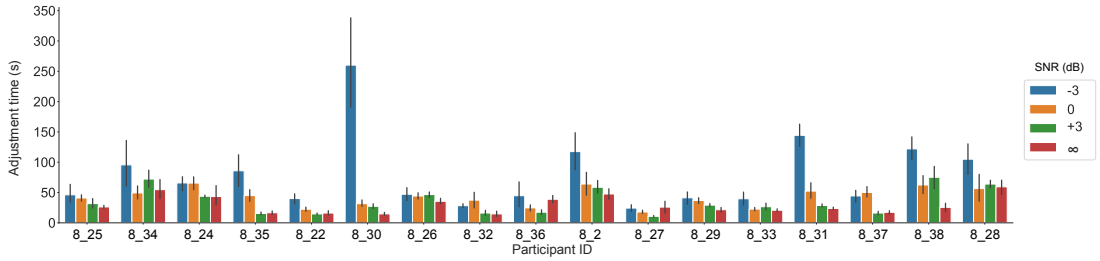


Figure C.6: Similar to C.4 but for the time required for choosing the preferred f_0 variation.

Figures C.1-C.6 show the results of intelligibility, listener preferences and adjustment time of both f_0 features for each different listener and condition. It can be observed that for both features, listener preferences vary across listeners and conditions. In general, listeners seem to have different f_0 preferences regarding the noise level except the participant with ID 8_2 who preferred almost the same mean f_0 values regardless the noise level while he/she indicated different f_0 variation preferences for the different conditions. In contrast, participant with ID

8_29 preferred the same f_0 variation, while different mean f_0 values. In general, listeners spent more adjustment time in more adverse conditions with some listeners spending more than 3 minutes (e.g. 8_30, 8_31) without this observation to indicate greater intelligibility achievements. Finally, intelligibility scores were generally poorer for conditions with greater noise level with some listeners performing better than others (e.g. 8_2, 8_38) regardless condition. No difference was observed for those participants who reported extensive music studies.

C.2 Experiment II

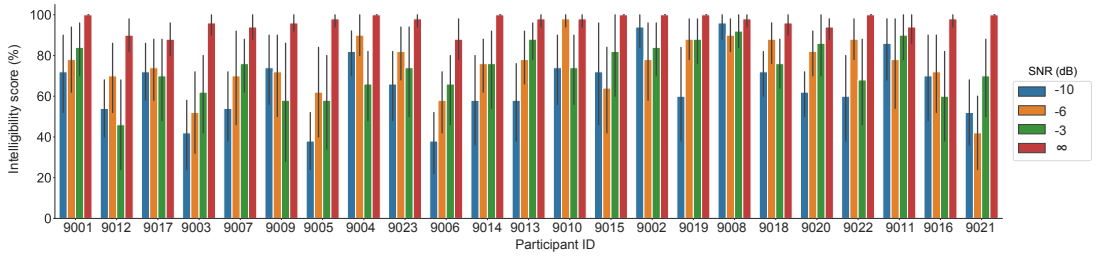


Figure C.7: Bar plot of the intelligibility scores achieved with the preferred mean f_0 for each listener. The error bars represent the 95% confidence intervals. Ordering on x-axis is based on the listeners' mean f_0 preferences at -10 dB SNR.

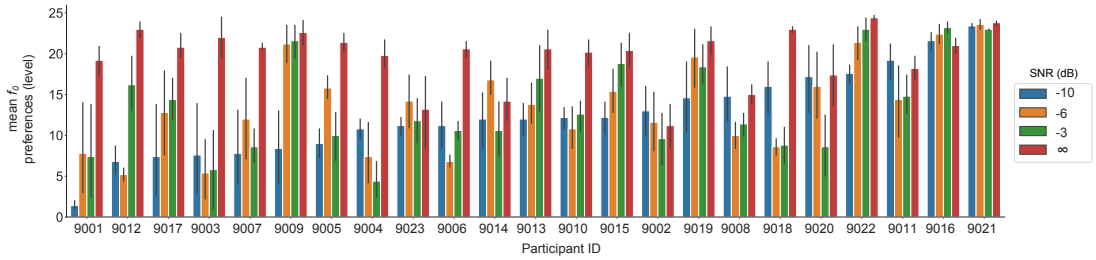


Figure C.8: Similar to C.7 but for the mean f_0 preferences.

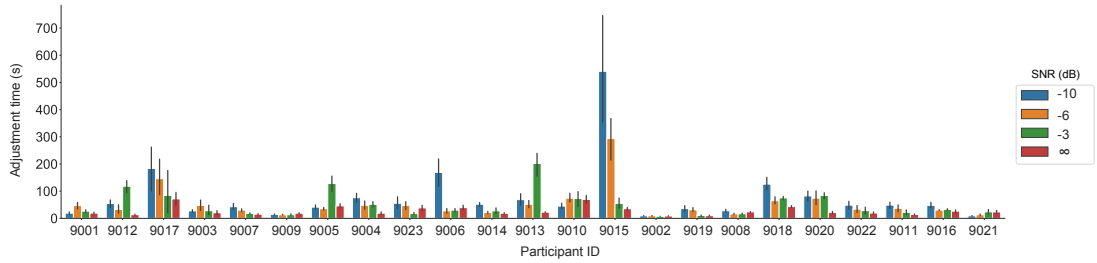


Figure C.9: Similar to C.7 but for the time required for choosing the preferred mean f_0.

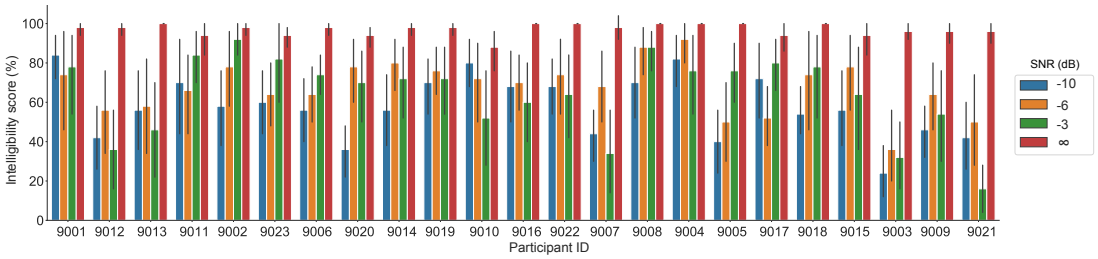


Figure C.10: Bar plot of the intelligibility scores achieved with the preferred f_0 variation for each listener. The error bars represent the 95% confidence intervals. Ordering on x-axis is based on the listeners' f_0 variation preferences at -10 dB SNR.

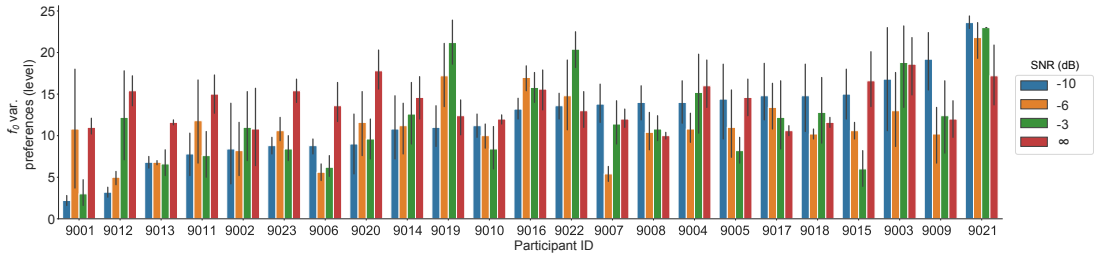


Figure C.11: Similar to C.10 but for the f_0 variation preferences.

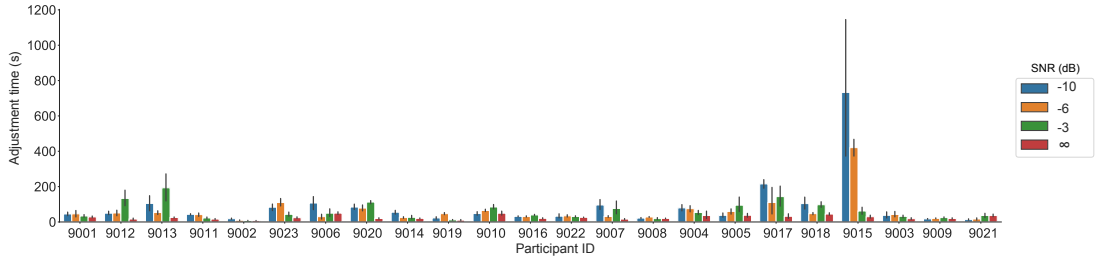


Figure C.12: Similar to C.10 but for the time required for choosing the preferred f_0 variation.

Figures C.7-C.12 show the results of intelligibility, listener preferences and adjustment time of both f_0 features for each different listener and condition. Intelligibility scores vary, with some listeners to achieve higher scores compared to others regardless the condition (e.g. 9011 for the mean f_0 and 9001 for the f_0 variation) and some other listeners to perform poorly (e.g. 9003 for the mean and f_0 variation). This might be related to the listeners' choices since listener 9011 overall preferred higher pitch values compared to listener 9003. Similar results showed for f_0 variation feature where listener 9001 preferred overall higher f_0 variation resulting to better performance compared to listener 9003 who preferred lower f_0 variation. Regarding the adjustment time (Fig. C.9 and C.12) there were some participants who needed more than 4 minutes for adjusting speech in adverse conditions (listener 9015 for both f_0 features and listener 9017 for mean f_0 only). No difference was observed for those participants who reported extensive music studies.

Appendix D

Spectrograms of features tested in chapter 7

D.1 Experiment II

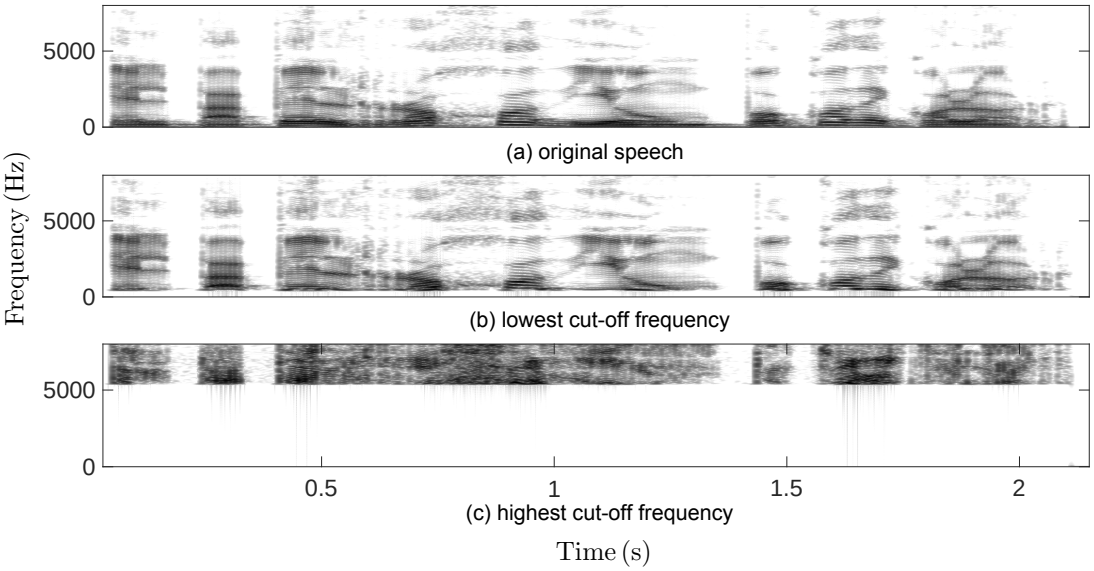


Figure D.1: *Spectrograms of the phrase 'El papel rojo dio un poco de color'. The upper plot corresponds to the phrase of the original speech, the middle to the phrase after applying the high-pass filter with the lowest cut-off frequency, and the lower plot to that with the highest cut-off frequency.*

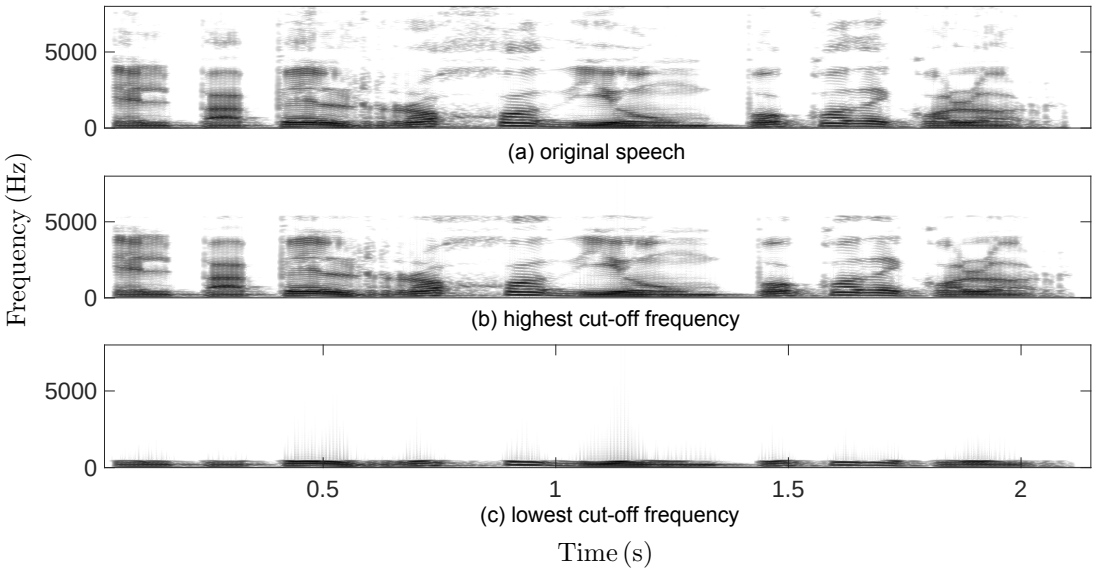


Figure D.2: *As Fig. D.1 but for the low-pass filter.*

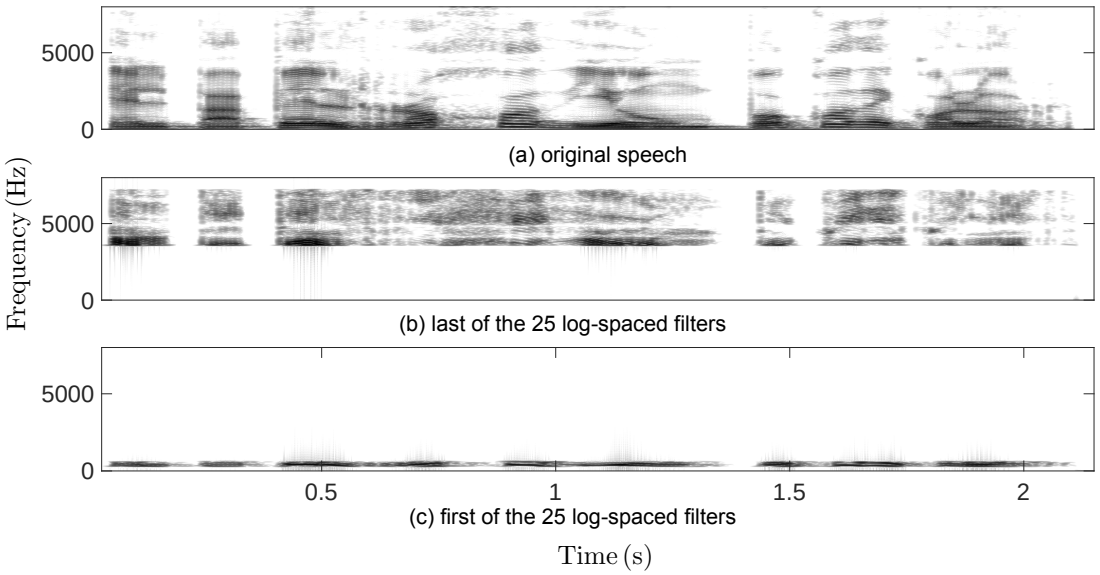


Figure D.3: *As Fig. D.1 but for the sliding band-pass filter with middle plot corresponding to the phrase filtered with the first of the* 25 *log-spaced filters and the lower plot to the phrase with the last of the filters.*

Bibliography

3GPP TS 26.190. Adaptive Multi-Rate-Wideband (AMR-WB) Speech Codec, Transcoding Functions. 3rd Generation Partnership Project, 2005. Valbonne, France, version 6.1.1.

E. M. Adams and R. E. Moore. Effects of speech rate, background noise, and simulated hearing loss on speech rate judgment and speech intelligibility in young listeners. *Journal of the American Academy of Audiology*, 20:28–39, 2009.

E. M. Adams, S. Gordon-Hickey, H. Morlas, and R. Moore. Effect of Rate-Alteration on Speech Perception in Noise in Older Adults With Normal Hearing and Hearing Impairment. *American Journal of Audiology*, 21(1):22–32, 2012.

P. Adank and E. Janse. Perceptual learning of time-compressed and natural fast speech. *The Journal of the Acoustical Society of America*, 126(5):2649–2659, 2009.

P. F. Assmann. Fundamental frequency and the intelligibility of competing voices, 1999.

P. F. Assmann and T. M. Nearey. Relationship between fundamental and formant frequencies in voice preference. *The Journal of the Acoustical Society of America*, 122(2):EL35–EL43, 2007.

P. F. Assmann and T. M. Nearey. Identification of frequency-shifted vowels. *The Journal of the Acoustical Society of America*, 124(5):3203–3212, 2008.

P. F. Assmann, T. M. Nearey, and J. M. Scott. Modeling the perception of frequency-shifted vowels. *International Conference on Spoken Language Processing-2002*, pages 425–428, 2002.

P. F. Assmann, S. Dembling, and T. M. Nearey. Effects of frequency shifts on perceived naturalness and gender information in speech. *INTERSPEECH-2006*, pages 889–892, 2006.

V. Aubanel, M. L. G. Lecumberri, and M. Cooke. The Sharvard Corpus: A phonemically-balanced Spanish sentence resource for audiology. *International Journal of Audiology*, 53(9): 633–638, 2014.

E. Axmear, J. Reichle, M. Alamsaputra, K. Kohnert, K. Drager, and K. Sellnow. Synthesized Speech Intelligibility in Sentences. *Language, Speech, and Hearing Services in Schools*, 36(3): 244–250, 2005.

J. Z. Bakdash and L. R. Marusich. Repeated Measures Correlation. *Frontiers in Psychology*, 8:456, 2017.

J. Barker and M. Cooke. Modelling speaker intelligibility in noise. *Speech Commun.*, 49, 2007.

D. Bates, M. Mächler, B. Bolker, and S. Walker. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1):1–48, 2015.

J. G. Beerends and J. A. Stemerdink. A Perceptual Speech-Quality Measure Based on a Psychoacoustic Sound Representation. *Journal of the Audio Engineering Society*, 42(3):115–123, 1994.

J. G. Beerends, A. P. Hekstra, A. W. Rix, and M. P. Hollier. Perceptual Evaluation of Speech Quality (PESQ) The New ITU Standard for End-to-End Speech Quality Assessment Part II: Psychoacoustic Model. *Journal of the Audio Engineering Society*, 50(10):765–778, 2002.

J. G. Beerends, C. Schmidmer, J. Berger, M. Obermann, R. Ullmann, J. Pomy, and M. Keyhl. Perceptual objective listening quality assessment (polqa), the third generation itu-t standard for end-to-end speech quality measurement part i–temporal alignment. 61(6):366–384, 2013.

I. V. Bele. The Speaker's Formant. *Journal of Voice*, 20(4):555 – 578, 2006.

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.*, pages 263–269, 1998.

Z. S. Bond and T. J. Moore. A note on loud and Lombard speech. *International Conference on Spoken Language Processing*, pages 969–972, 1990.

G. Borghini and V. Hazan. Listening Effort During Sentence Processing Is Increased for Non-native Listeners: A Pupillometry Study. *Frontiers in Neuroscience*, 12:152, 2018.

G. Borghini and V. Hazan. Effects of acoustic and semantic cues on listening effort during native and non-native speech perception. *The Journal of the Acoustical Society of America*, 147(6):3783–3794, 2020.

H. Boril and P. Pollak. Design and collection of Czech Lombard database. *International Conference on Spoken Language Processing*, pages 1577–1580, 2005.

M. Boymans and W. Dreschler. Field trials using a digital hearing aid with active noise reduction and dual-microphone directionality. *Audiology*, 39(5):260–268, 2000.

A. R. Bradlow, G. M. Torretta, and D. B. Pisoni. Intelligibility of normal speech I: Global and fine-grained acoustic-phonetic talker characteristics. *Speech Communication*, 20(3):255–272, 1996.

A. R. Bradlow, N. Kraus, and E. Hayes. Speaking Clearly for Children With Learning Disabilities. *Journal of Speech, Language, and Hearing Research*, 46(1):80–97, 2003.

A. S. Bregman. Auditory Scene Analysis. *Cambridge, MIT Press*, 1990.

D. E. Broadbent. *Perception and communication*. Pergamon Press, 1958.

J. P. L. Brokx and S. G. Nooteboom. Intonation and the perceptual separation of simultaneous voices. *Journal of Phonetics*, 10(1):23–36, 1982.

I. Brons, R. Houben, and W. A. Dreschler. Perceptual Effects of Noise Reduction With Respect to Personal Preference, Speech Intelligibility, and Listening Effort. *Ear and Hearing*, 34:29–41, 2013.

D. Brungart, N. Iyer, E. R. Thompson, B. D. Simpson, S. Gordon-Salant, J. Schurman, C. Vogel, and K. Grant. Interactions between listening effort and masker type on the energetic and informational masking of speech stimuli. *Proceedings of Meetings on Acoustics*, 19(1):060146, 2013.

D. S. Brungart. Informational and energetic masking effects in the perception of two simultaneous talkers. *The Journal of the Acoustical Society of America*, 109(3):1101–1109, 2001.

J. Brunskog, A. C. Gade, G. P. Bellester, and L. R. Calbo. Increase in voice level and speaker comfort in lecture rooms. *The Journal of the Acoustical Society of America*, 125(4):2072–2082, 2009.

D. Burnham, C. Kitamura, and U. Vollmer-Conna. What's new, pussycat? On talking to babies and animals. *Science*, 296:1435, 2002.

W. Buyens, B. van Dijk, M. Moonen, and J. Wouters. Music mixing preferences of cochlear implant recipients: a pilot study. *Int J Audiol.*, 53(5):294–301, 2014.

P. J. Carolan, A. Heinrich, K. J. Munro, and R. E. Millman. Quantifying the Effects of Motivation on Listening Effort: A Systematic Review and Meta-Analysis. *Trends in Hearing*, 26, 2022.

A. Castellanos, J. Benedí, and F. Casacuberta. An analysis of general acoustic-phonetic features for Spanish speech produced with the Lombard effect. *Speech Communication*, 20(1):23–35, 1996.

F. Charpentier and M. Stella. Diphone synthesis using an overlap-add technique for speech waveforms concatenation. In *ICASSP '86. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 11, pages 2015–2018, 1986.

C. Chermaz and S. King. A Sound Engineering Approach to Near End Listening Enhancement. In *Proc. Interspeech*, pages 1356–1360, 2020.

D. V. Cicchetti. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, 6(4):284–290, 1994.

M. Cooke. Glimpsing speech. *Journal of Phonetics*, 31(3):579–584, 2003.

M. Cooke. A glimpsing model of speech perception in noise. *The Journal of the Acoustical Society of America*, 119(3):1562–1573, 2006.

M. Cooke and V. Aubanel. Effects of linear and nonlinear speech rate changes on speech intelligibility in stationary and fluctuating maskers. *The Journal of the Acoustical Society of America*, 141(6):4126–4135, 2017.

M. Cooke and M. L. Garcia Lecumberri. The Effects of Modified Speech Styles on Intelligibility for Non-Native Listeners. In *Interspeech 2016*, pages 868–872, 2016.

M. Cooke and Y. Lu. Spectral and temporal changes to speech produced in the presence of energetic and informational maskers. *The Journal of the Acoustical Society of America*, 128 (4):2059–2069, 2010.

M. Cooke, C. Mayo, and C. Valentini-Botinhao. Intelligibility-enhancing speech modifications: the Hurricane Challenge. In *Proc. Interspeech*, pages 3552–3556, 2013a.

M. Cooke, C. Mayo, C. Valentini-Botinhao, Y. Stylianou, B. Sauert, and Y. Tang. Evaluating the intelligibility benefit of speech modifications in known noise conditions. *Speech Communication*, 55(4):572–585, 2013b.

M. Cooke, S. King, M. Garnier, and V. Aubanel. The listening talker: A review of human and algorithmic context-induced modifications of speech. *Computer Speech & Language*, 28: 543–571, 2014a.

M. Cooke, C. Mayo, and J. Villegas. The contribution of durational and spectral changes to the Lombard speech intelligibility benefit. *The Journal of the Acoustical Society of America*, 135(2):874–883, 2014b.

A. Cutler, D. Dahan, and W. van Donselaar. Prosody in the Comprehension of Spoken Language: A Literature Review. *Language and Speech*, 40(2):141–201, 1997.

C. J. Darwin, D. S. Brungart, and B. D. Simpson. Effects of fundamental frequency and vocal-tract length changes on attention to one of two simultaneous talkers. *The Journal of the Acoustical Society of America*, 114(5):2913–2922, 2003.

P. Dawes, K. J. Munro, S. Kalluri, and B. Edwards. Acclimatization to Hearing Aids. *Ear and Hearing*, 35(2):203–212, 2014.

R. L. Diehl. Acoustic and auditory phonetics: the adaptive design of speech sound systems. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 363:965–978, 2008.

L. C. Dilley and J. D. McAuley. Distal prosodic context affects word segmentation and lexical processing. *Journal of Memory and Language*, 59(3):294–311, 2008.

E. Dupoux and K. Green. Perceptual adjustment to highly compressed speech: Effects of talker and rate changes. In *Journal of Experimental Psychology: Human Perception and Performance*, volume 23, pages 914–927, 1997.

G. Fairbanks and F. Kodman. Word intelligibility as a function of time compression. *The Journal of the Acoustical Society of America*, 29:636–644, 1957.

T. H. Falk, V. Parsa, J. F. Santos, K. Arehart, O. Hazrati, R. Huber, J. M. Kates, and S. Scollie. Objective Quality and Intelligibility Prediction for Users of Assistive Listening Devices: Advantages and limitations of existing tools. *IEEE Signal Processing Magazine*, 32 (2):114–124, 2015.

A. Fernald and C. Mazzie. Prosody and focus in speech to infants and adults. *Developmental Psychology*, 27(2):209–221, 1991.

J. M. Festen and R. Plomp. Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing. *The Journal of the Acoustical Society of America*, 88(4):1725–1736, 1990.

M. Fitzpatrick, J. Kim, and C. Davis. The effect of seeing the interlocutor on auditory and visual speech production in noise. *Speech Communication*, 74:37–51, 2015.

H. Fletcher and R. H. Galt. The Perception of Speech and Its Relation to Telephony. *The Journal of the Acoustical Society of America*, 22(2):89–151, 1950.

D. Fogerty, J. B. Ahlstrom, and J. R. Dubno. Glimpsing keywords across sentences in noise: A microstructural analysis of acoustic, lexical, and listener factors. *The Journal of the Acoustical Society of America*, 150(3):1979–1996, 2021.

N. R. French and J. C. Steinberg. Factors Governing the Intelligibility of Speech Sounds. *The Journal of the Acoustical Society of America*, 19(1):90–119, 1947.

S. Frota, M. Vigário, and F. Martins. Language Discrimination and Rhythm Classes: Evidence from Portuguese. pages 319–322, 2002.

C. Füllgrabe, F. Berthommier, and C. Lorenzi. Masking release for consonant features in temporally fluctuating background noise. *Hearing Research*, 211(1):74–84, 2006.

A. Gabrielsson, B. Schenkman, and B. Hagerman. The Effects of Different Frequency Responses on Sound Quality Judgments and Speech Intelligibility. *Journal of Speech, Language, and Hearing Research*, 31(2):166–177, 1988.

A. Gabrielsson, B. Hagerman, T. Bech-Kristensen, and G. Lundberg. Perceived sound quality of reproductions with different frequency responses and sound levels. *The Journal of the Acoustical Society of America*, 88(3):1359–1366, 1990.

J.-P. Gagné, J. Besser, and U. Lemke. Behavioral Assessment of Listening Effort Using a Dual-Task Paradigm: A Review. *Trends in Hearing*, 21, 2017.

M. L. Garcia Lecumberri, M. Cooke, and A. Cutler. Non-native speech perception in adverse conditions: A review. *Speech Communication*, 52(11–12):864–886, 2010.

M. Garnier and N. Henrich. Speaking in noise: How does the Lombard effect improve acoustic contrasts between speech and ambient noise? *Computer Speech & Language*, 28(2):580–597, 2014.

S. Gatehouse. The role of non-auditory factors in measured and self-reported disability. *Acta Otolaryngol Suppl*, 476:249–256, 1990.

S. Gatehouse and J. Gordon. Response times to speech stimuli as measures of benefit from amplification. *British Journal of Audiology*, 24(1):63–68, 1990.

S. Gatehouse and W. Noble. The speech, spatial and qualities of hearing scale (SSQ). *Int J Audiol*, 43:85–99, 2004.

E. Godoy and Y. Stylianou. Unsupervised acoustic analyses of normal and lombard speech, with spectral envelope transformation to improve intelligibility. *Interspeech*, pages 1472–1475, 2012.

S. Gordon-Salant and P. J. Fitzgibbons. Effects of stimulus and noise rate variability on speech perception by younger and older adults. *The Journal of the Acoustical Society of America*, 115(4):1808–1817, 2004.

P. A. Gosselin and J.-P. Gagné. Older Adults Expend More Listening Effort Than Young Adults Recognizing Speech in Noise. *Journal of Speech, Language, and Hearing Research*, 54 (3):944–958, 2011.

A. Govender and S. King. Using Pupillometry to Measure the Cognitive Load of Synthetic Speech. In *Proc. Interspeech*, pages 2838–2842, 2018a.

A. Govender and S. King. Measuring the Cognitive Load of Synthetic Speech Using a Dual Task Paradigm. In *Proc. Interspeech*, pages 2843–2847, 2018b.

A. Govender, A. E. Wagner, and S. King. Using Pupil Dilation to Measure Cognitive Load When Listening to Text-to-Speech in Quiet and in Noise. In *Proc. Interspeech 2019*, pages 1551–1555, 2019.

P. Gramming, J. Sundberg, S. Ternström, R. Leanderson, and W. H. Perkins. Relationship between changes in voice pitch and loudness. *Journal of Voice*, 2(2):118–126, 1988.

L. M. Guijo and A. C. V. Cardoso. Physiological methods as indexes of listening effort measurement: an integrative literature review. *Revista Cefac*, 20:541–549, 2018.

J. H. L. Hansen and B. L. Pellom. An effective quality evaluation protocol for speech enhancement algorithms. In *International Conference on Spoken Language Processing*, volume 7, pages 2819–2822, 1998.

S. G. Hart and L. E. Staveland. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In *Human Mental Workload*, volume 52 of *Advances in Psychology*, pages 139–183. 1988.

M. L. Hawley, R. Y. Litovsky, and J. F. Culling. The benefit of binaural hearing in a cocktail party: Effect of location and type of interferer. *The Journal of the Acoustical Society of America*, 115(2):833–843, 2004.

V. Hazan and R. Baker. Acoustic-phonetic characteristics of speech produced with communicative intent to counter adverse listening conditions. *The Journal of the Acoustical Society of America*, 130(4):2139–2152, 2011.

C. B. Hicks and A. M. Tharpe. Listening Effort and Fatigue in School-Age Children With and Without Hearing Loss. *Journal of Speech, Language, and Hearing Research*, 45(3):573–584, 2002.

S. Holm. A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics*, 6:65–70, 1979.

R. Houben, M. van Doorn-Bierman, and W. A. Dreschler. Using response time to speech as a measure for listening effort. *International Journal of Audiology*, 52(11):753–761, 2013.

C. S. Howard, K. J. Munro, and C. J. Plack. Listening effort at signal-to-noise ratios that are typical of the school classroom. *International Journal of Audiology*, 49(12):928–932, 2010.

P. A. Howard-Jones and S. Rosen. Uncomodulated glimpsing in "checkerboard" noise. *The Journal of the Acoustical Society of America*, 93(5):2915–2922, 1993.

Y. Hu and P. Loizou. Subjective Comparison of Speech Enhancement Algorithms. In *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, volume 1, pages I–I, 2006.

Y. Hu and P. C. Loizou. Subjective comparison and evaluation of speech enhancement algorithms. *Speech Communication*, 49(7):588–601, 2007.

Y. Hu and P. C. Loizou. Evaluation of Objective Quality Measures for Speech Enhancement. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(1):229–238, 2008.

ITU-T. Methods for subjective determination of transmission quality. ITU-T Recommendation, 1996. P. 800.

ITU-T. Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs. ITU-T Recommendation, 2001. P. 862.

ITU-T. Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm. ITU-T Recommendation, 2003. P. 835.

ITU-T. Wideband extension to recommendation P.862 for the assessment of wideband telephone networks and speech codecs. ITU-T Recommendation, 2005. P. 862.2.

ITU-T. Perceptual Objective Listening Quality Assessment. ITU-T Recommendation, 2011. P. 863.

J. Johnson, J. Xu, R. Cox, and P. Pendergraft. A Comparison of Two Methods for Measuring Listening Effort As Part of an Audiologic Test Battery. *American Journal of Audiology*, 24 (3):419–431, 2015.

J. C. Junqua. The Lombard reflex and its role on human listeners and automatic speech recognizers. *The Journal of the Acoustical Society of America*, 93(1):510–524, 1993.

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

M. Karjalainen. A new auditory model for the evaluation of sound quality of audio systems. In *International Conference on Acoustics, Speech and Signal Processing*, volume 10, pages 608–611, 1985.

H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné. Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds. *Speech Communication*, 27(3–4): 187–207, 1999.

H. Kawahara, M. Morise, T. Takahashi, R. Nisimura, T. Irino, and H. Banno. Tandem-STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0, and aperiodicity estimation. In *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3933–3936, 2008.

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

J. R. Kerlin, A. J. Shahin, and L. M. Miller. Attentional Gain Control of Ongoing Cortical Speech Representations in a "Cocktail Party". *Journal of Neuroscience*, 30(2):620–628, 2010.

G. R. Kidd. Articulatory-rate context effects in phoneme identification. *Journal of Experimental Psychology: Human Perception and Performance*, 15(4):736–748, 1989.

S. King and V. Karaiskos. The Blizzard Challenge 2011. 2011.

D. Klatt. Prediction of perceived phonetic distance from critical-band spectra: A first step. In *International Conference on Acoustics, Speech and Signal Processing*, volume 7, pages 1278–1281, 1982.

X. Koch and E. Janse. Speech rate effects on the processing of conversational speech across the adult life span. *The Journal of the Acoustical Society of America*, 139(4):1618–1636, 2016.

T. Koelewijn, A. A. Zekveld, J. M. Festen, and S. E. Kramer. Pupil Dilation Uncovers Extra Listening Effort in the Presence of a Single-Talker Masker. *Ear and hearing*, 33(2):291–300, 2012.

T. Koelewijn, A. A. Zekveld, J. M. Festen, and S. E. Kramer. The influence of informational masking on speech perception and pupil response in adults with hearing impairment. *Journal of the Acoustical Society of America*, 135(3):1596–1606, 2014.

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

F. Koopmans-van Beinum. Spectro-temporal reduction and expansion in spontaneous speech and read text: focus words versus non-focus words. In *Phonetics and Phonology of Speaking Styles*, 1991.

J. C. Krause and L. D. Braida. Acoustic properties of naturally produced clear speech at normal speaking rates. *The Journal of the Acoustical Society of America*, 115(1):362–378, 2004.

M. Krueger, M. Schulte, M. A. Zokoll, K. C. Wagener, M. Meis, T. Brand, and I. Holube. Relation Between Listening Effort and Speech Intelligibility in Noise. *American Journal of Audiology*, 26(3S):378–392, 2017.

S. Kullback and R. A. Leibler. On Information and Sufficiency. *Ann. Math. Statist.*, 22(1): 79–86, 1951.

C. Lam and C. Kitamura. Mommy, speak clearly: induced hearing loss shapes vowel hyperarticulation. *Developmental Science*, 15(2):212–221, 2012.

B. Larsby, M. Hällgren, B. Lyxell, and S. Arlinger. Cognitive performance and perceived effort in speech processing tasks: Effects of different noise backgrounds in normal-hearing and hearing-impaired subjects. *International Journal of Audiology*, 44(3):131–143, 2005.

N. J. Lass and V. A. Fultz. A Normative Study of Children's Listening Rate Preferences. *Language and Speech*, 19(2):144–149, 1976.

N. J. Lass and C. E. Prater. A Comparative Study of Listening Rate Preferences for Oral Reading and Impromptu Speaking Tasks. *Journal of Communication*, 23(1):95–102, 1973.

J. S. Laures and K. Bunton. Perceptual effects of a flattened fundamental frequency at the sentence level under different listening conditions. *Journal of Communication Disorders*, 36 (6):449–464, 2003.

J. S. Laures and G. Weismer. The Effects of a Flattened Fundamental Frequency on Intelligibility at the Sentence Level. *Journal of Speech, Language, and Hearing Research*, 42(5): 1148–1156, 1999.

J. Lebeter and S. Saunders. The effects of time compression on the comprehension of natural and synthetic speech. *Working Papers of the Linguistics Circle*, 20(1):63–81, 2010.

F. F. Lee. Time Compression and Expansion of Speech by the Sampling Method. *Journal of the Audio Engineering Society*, 20(9):738–742, 1972.

H. Leeper and C. Thomas. Young Children's Preferences for Listening Rates. *Perceptual and Motor Skills*, 47(3):891–898, 1978.

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

J. S. Liénard and M. G. D. Benedetto. Effect of vocal effort on spectral properties of vowels. *The Journal of the Acoustical Society of America*, 106(1):411–422, 1999.

P. C. Loizou. Speech quality assessment. In *Multimedia Analysis, Processing and Communications*, volume 346, pages 623–654. Springer, 2011.

Y. Lu and M. Cooke. Speech production modifications produced by competing talkers, babble, and stationary noise. *The Journal of the Acoustical Society of America*, 124(5):3261–3275, 2008.

Y. Lu and M. Cooke. The contribution of changes in F0 and spectral tilt to increased intelligibility of speech produced in noise. *Speech Communication*, 51(12):1253–1262, 2009a.

Y. Lu and M. Cooke. Speech production modifications produced in the presence of low-pass and high-pass filtered noise. *The Journal of the Acoustical Society of America*, 126(3):1495–1499, 2009b.

H. Luts, K. Eneman, J. Wouters, M. Schulte, M. Vormann, M. Buechler, N. Dillier, R. Houben, W. A. Dreschler, M. Froehlich, H. Puder, G. Grimm, V. Hohmann, A. Leijon, A. Lombard, D. Mauler, and A. Spriet. Multicenter evaluation of signal enhancement algorithms for hearing aids. *The Journal of the Acoustical Society of America*, 127(3):1491–1505, 2010.

C. L. Mackersie and H. Cones. Subjective and psychophysiological indexes of listening effort in a competing-talker task. *Journal of the American Academy of Audiology*, 22(2):113–122, 2011.

A. Macleod and Q. Summerfield. Quantifying the contribution of vision to speech perception in noise. *British Journal of Audiology*, 21(2):131–141, 1987.

K. Marcoux, M. Cooke, B. V. Tucker, and M. Ernestus. The Lombard intelligibility benefit of native and non-native speech for native and non-native listeners. *Speech Communication*, 136:53–62, 2022.

S. L. Mattys, M. H. Davis, A. R. Bradlow, and S. K. Scott. Speech recognition in adverse conditions: A review. *Language and Cognitive Processes*, 27(7–8):953–978, 2012.

C. Mayo, V. Aubanel, and M. Cooke. Effect of prosodic changes on speech intelligibility. In *Interspeech*, 2012.

M. McClurg. *Effects of high-pass filtering on perception of dialect and talker sex*. PhD thesis, The Ohio State University, 2018.

S. L. McCoy, P. A. Tun, L. C. Cox, M. Colangelo, R. A. Stewart, and A. Wingfield. Hearing Loss and Perceptual Effort: Downstream Effects on Older Adults' Memory for Speech. *The Quarterly Journal of Experimental Psychology Section A*, 58(1):22–33, 2005.

R. McGarrigle, K. J. Munro, P. Dawes, A. J. Stewart, D. R. Moore, J. G. Barry, and S. Amitay. Listening effort and fatigue: what exactly are we measuring? A British Society of Audiology Cognition in Hearing Special Interest Group 'white paper'. *Int J Audiol.*, 53(7):433–440, 2014.

K. O. McGraw and S. P. Wong. Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, 1(1):30–46, 1996.

D. Meador, J. E. Flege, and I. R. A. Mackay. Factors affecting the recognition of words in a second language. *Bilingualism: Language and Cognition*, 3(1):55–67, 2000.

G. A. Miller. The masking of speech. *Psychological Bulletin*, 44(2):105–129, 1947.

G. A. Miller and J. C. R. Licklider. The Intelligibility of Interrupted Speech. *The Journal of the Acoustical Society of America*, 22(2):167–173, 1950.

D. Mirman. *Growth Curve Analysis and Visualization Using R*. CRC Press, 2014.

B. C. Moore. Basic auditory processes involved in the analysis of speech sounds. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 363:947–963, 2008.

B. C. Moore, L. K. Tyler, and W. Marslen-Wilson. Introduction. The perception of speech: from sound to meaning. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 363:917–921, 2008.

B. C. J. Moore and C.-T. Tan. Perceived naturalness of spectrally distorted speech and music. *The Journal of the Acoustical Society of America*, 114(1):408–419, 2003.

B. C. J. Moore, B. R. Glasberg, and R. W. Peters. Thresholds for hearing mistuned partials as separate tones in harmonic complexes. *The Journal of the Acoustical Society of America*, 80(2):479–483, 1986.

R. E. Moore, E. M. Adams, P. A. Dagenais, and C. Caffee. Effects of reverberation and filtering on speech rate judgment. *International Journal of Audiology*, 46(3):154–160, 2007.

P. Mowlaee, R. Saeidi, M. G. Christensen, and R. Martin. Subjective and objective quality assessment of single-channel speech separation algorithms. In *International Conference on Acoustics, Speech and Signal Processing*, pages 69–72, 2012.

D. R. Murphy, F. I. M. Craik, K. Z. H. Li, and B. A. Schneider. Comparing the effects of aging and background noise on short-term memory performance. *Psychology and Aging*, 15 (2):323–334, 2000.

Y. Nejime and B. C. J. Moore. Evaluation of the effect of speech-rate slowing on speech intelligibility in noise using a simulation of cochlear hearing loss. *The Journal of the Acoustical Society of America*, 103(1):572–576, 1998.

R. Niederjohn and J. Grotelueschen. The enhancement of speech intelligibility in high noise levels by high-pass filtering followed by rapid amplitude compression. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 24(4):277–282, 1976.

S. G. Nooteboom, J. P. L. Brokx, and J. J. de Rooij. Contributions of prosody to speech perception. *Studies in the Perception of Language (W.J.M. Levelt and G.B. Flores d'Arcais, eds)*, pages 75–109, 1978.

J. S. Novak III and R. V. Kenyon. Effects of User Controlled Speech Rate on Intelligibility in Noisy Environments. In *Proc. Interspeech*, pages 1853–1857, 2018.

J. S. Novak III, A. Tandon, J. Leigh, and R. V. Kenyon. Networked On-Line Audio Dilation. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*, pages 255–258. Association for Computing Machinery, 2014.

J. Obleser, M. Wöstmann, N. Hellbernd, A. Wilsch, and B. Maess. Adverse listening conditions and memory load drive a common alpha oscillatory network. *J Neurosci*, 32:12376 – 12383, 2012.

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

C. Pals, A. Sarampalis, and D. Baskent. Listening effort with cochlear implant simulations. *J. Speech Hearing Language Res.*, 56:1075–1084, 2013.

K. Pearsons, R. Bennett, and F. S. Speech levels in various noise environments. *Washington, DC: U.S. Environmental Protection Agency (Report No. EPA-600/1-77-025)*, 1977.

J. E. Peelle and M. H. Davis. Neural Oscillations Carry Speech Rhythm through to Comprehension. *Frontiers in psychology*, 3:320–320, 2012.

D. Pelegrín-García, B. Smits, J. Brunskog, and C.-H. Jeong. Vocal effort with changing talker-to-listener distance in different acoustic environments. *The Journal of the Acoustical Society of America*, 129(4):1981–1990, 2011.

Z. E. Peng and L. M. Wang. Listening Effort by Native and Nonnative Listeners Due to Noise, Reverberation, and Talker Foreign Accent During English Speech Perception. *J Speech Lang Hear Res*, 62(4):1068–1081, 2019.

R. W. Peters, B. C. J. Moore, and T. Baer. Speech reception thresholds in noise with and without spectral and temporal dips for hearing-impaired and normally hearing people. *The Journal of the Acoustical Society of America*, 103(1):577–587, 1998.

H. Pham. PyAudio: Python Bindings for PortAudio, 2006. URL `https://pypi.org/project/PyAudio/`.

M. A. Picheny, N. I. Durlach, and L. D. Braida. Speaking clearly for the hard of hearing I: Intelligibility differences between clear and conversational speech. *J. Speech Hear. Res.*, 28: 96–103, 1985.

M. A. Picheny, N. I. Durlach, and L. D. Braida. Speaking clearly for the hard of hearing II: Acoustic characteristics of clear and conversational speech. *Journal of Speech, Language, and Hearing Research*, 29(4):434–446, 1986.

M. Pichora-Fuller, B. Schneider, and M. Daneman. How young and old adults listen to and remember speech in noise. *The Journal of the Acoustical Society of America*, 97(1):593–608, 1995.

M. Pichora-Fuller, S. Kramer, M. Eckert, B. Edwards, B. Hornsby, L. Humes, U. Lemke, T. Lunner, M. Matthen, C. Mackersie, G. Naylor, N. Phillips, M. Richter, M. Rudner, M. Sommers, K. Tremblay, and A. Wingfield. Hearing Impairment and Cognitive Energy: The Framework for Understanding Effortful Listening (FUEL). *Ear and Hearing*, 37:5S–27S, 2016.

E. M. Picou, T. A. Ricketts, and B. W. Y. Hornsby. Visual Cues and Listening Effort: Individual Variability. *Journal of Speech, Language, and Hearing Research*, 54(5):1416–1430, 2011.

T. Piquado, J. I. Benichov, H. Brownell, and A. Wingfield. The hidden effect of hearing acuity on speech recall, and compensatory effects of self-paced listening. *International Journal of Audiology*, 51(8):576–583, 2012.

D. B. Pisoni, L. M. Manous, and M. J. Dedina. Comprehension of natural and synthetic speech: effects of predictability on the verification of sentences controlled for intelligibility. *Computer Speech & Language*, 2(3):303–320, 1987.

A. L. Pittman and T. L. Wiley. Recognition of Speech Produced in Noise. *Journal of Speech, Language, and Hearing Research*, 44(3):487–496, 2001.

C. J. Plack and A. J. Oxenham. *The Psychophysics of Pitch*, pages 7–55. Springer New York, 2005.

J. Preminger and D. V. Tasell. Quantifying the relation between speech quality and speech intelligibility. *Journal of Speech, Language and Hearing Research*, 38(3):714–725, 1995.

S. Quackenbush, T. Barnwell, and M. Clements. Objective Measures of Speech Quality. In *Englewood Cliffs, NJ: Prentice-Hall*, 1988.

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

J. Rennies, H. Schepker, C. Valentini-Botinhao, and M. Cooke. Intelligibility-Enhancing Speech Modifications - The Hurricane Challenge 2.0. In *Proc. Interspeech 2020*, pages 1341–1345, 2020.

T. Ricketts and B. Hornsby. Sound quality measures for speech in noise through a commercial hearing aid implementing digital noise reduction. *Journal of the American Academy of Audiology*, 16(5):270–277, 2005.

L. Riensche, G. Lawson, D. Beasley, and L. Smith. Age and sex differences on preferred listening rates for speech. *The Journal of auditory research*, 19(2):91–94, 1979.

A. Rix and M. Hollier. The perceptual analysis measurement system for robust end-to-end speech quality assessment. In *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 3, pages 1515–1518, 2000.

A. Rix, M. Hollier, Hekstra, and J. A., Beerends. Perceptual evaluation of speech quality (PESQ): The new ITU standard for end-to-end speech quality assessment. Part I. Time-delay compensation. 50:755–764, 2002.

G. Robinson, J. Casali, E. Berger, L. Royster, D. Driscoll, J. Royster, and M. Layne. Chapter 14: Speech Communications and Signal Detection in Noise. 2002.

T. Rohdenburg, V. Hohmann, and B. Kollmeier. Objective perceptual quality measures for the evaluation of noise reduction schemes. In *Proc. 9th Int. Workshop Acoust. Echo Noise Control*, pages 169–172, 2005.

J. Rönnberg. Cognition in the hearing impaired and deaf as a bridge between signal and dialogue: a framework and a model. *International Journal of Audiology*, 42(sup1):68–76, 2003.

J. Rönnberg, T. Lunner, A. Zekveld, P. Sörqvist, H. Danielsson, B. Lyxell, O. Dahlström, C. Signoret, S. Stenfelt, M. Pichora-Fuller, and M. Rudner. The Ease of Language Understanding (ELU) model: theoretical, empirical, and clinical advances. *Front Syst Neurosci.*, 7 (31), 2013.

J. Rönnberg, E. Holmer, and M. Rudner. Cognitive hearing science and ease of language understanding. *International Journal of Audiology*, 58(5):247–261, 2019.

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*, 17:225–246, 1969.

M. Rudner, T. Lunner, T. Behrens, E. S. Thorén, and J. Rönnberg. Working Memory Capacity May Influence Perceived Effort during Aided Speech Recognition in Noise. *Journal of the American Academy of Audiology*, 23(8):577–589, 2012.

J. H. Ryalls and P. Lieberman. Fundamental frequency and vowel perception. *The Journal of the Acoustical Society of America*, 72(5):1631–1634, 1982.

A. G. Samuel and T. Kraljic. Perceptual learning for speech. *Attention, Perception, & Psychophysics*, 71(6):1207–1218, 2009.

J. Sankowska, M. L. G. Lecumberri, and M. Cooke. Interaction of intrinsic vowel and consonant durational correlates with foreigner directed speech. *Poznań Studies in Contemporary Linguistics*, 47(1):109–109, 2011.

A. Sarampalis, S. Kalluri, B. Edwards, and E. Hafter. Objective measures of listening effort: Effects of background noise and noise reduction. *Journal of Speech, Language, and Hearing Research*, 52(5):1230–1240, 2009.

P. Sauseng and W. Klimesch. What does phase information of oscillatory brain activity tell us about cognitive processes? *Neuroscience & Biobehavioral Reviews*, 32(5):1001–1013, 2008.

M. T. Scheffers. *Sifting vowels. Auditory pitch analysis and sound segregation.* PhD thesis, 1983. Doctoral dissertation, University of Groningen, The Netherlands.

H. Schepker, J. Rennies, and S. Doclo. Speech-in-noise enhancement using amplification and dynamic range compression controlled by the speech intelligibility index. *The Journal of the Acoustical Society of America*, 138(5):2692–2706, 2015.

J. Schmidtke. Second language experience modulates word retrieval effort in bilinguals: evidence from pupillometry. *Frontiers in Psychology*, 5:137, 2014.

S. Seeman and R. Sims. Comparison of Psychophysiological and Dual-Task Measures of Listening Effort. *Journal of Speech, Language, and Hearing Research*, 58(6):1781–1792, 2015.

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.

B. G. Shirley, M. Meadows, F. Malak, J. S. Woodcock, and A. Tidball. Personalized Object-Based Audio for Hearing Impaired TV Viewers. *J Audio Eng Soc.*, 65(4):293–303, 2017.

O. Simantiraki and M. Cooke. Exploring Listeners' Speech Rate Preferences. In *Proc. Interspeech*, pages 1346–1350, 2020.

O. Simantiraki and M. Cooke. SpeechAdjuster: A Tool for Investigating Listener Preferences and Speech Intelligibility. In *Proc. Interspeech*, pages 1718–1722, 2021.

O. Simantiraki, M. Cooke, and S. King. Impact of Different Speech Types on Listening Effort. In *Proc. Interspeech*, pages 2267–2271, 2018.

O. Simantiraki, M. Cooke, and Y. Pantazis. Effects of Spectral Tilt on Listeners' Preferences And Intelligibility. In *International Conference on Acoustics, Speech and Signal Processing*, pages 6254–6258, 2020.

M. D. Skowronski and J. G. Harris. Applied principles of clear and Lombard speech for automated intelligibility enhancement in noisy environments. *Speech Communiceation*, 48(5): 549–558, 2006.

K. Smeds, F. Wolters, and M. Rung. Estimation of signal-to-noise ratios in realistic sound scenarios. *Journal of the American Academy of Audiology*, 26:183–196, 2015.

M. R. Smith, A. Cutler, S. Butterfield, and I. Nimmo-Smith. The Perception of Rhythm and Word Boundaries in Noise-Masked Speech. *Journal of Speech, Language, and Hearing Research*, 32(4):912–920, 1989.

M. S. Sommers and D. Phelps. Listening Effort in Younger and Older Adults: A Comparison of Auditory-Only and Auditory-Visual Presentations. *Ear and Hearing*, 37:62S–68S, 2016.

J. F. Strand, V. A. Brown, M. B. Merchant, H. E. Brown, and J. Smith. Measuring Listening Effort: Convergent Validity, Sensitivity, and Links With Cognitive and Personality Measures. *Journal of Speech, Language, and Hearing Research*, 61(6):1463–1486, 2018.

G. A. Studebaker. A rationalized arcsine transform. *Journal of Speech, Language, and Hearing Research*, 28(3):455–462, 1985.

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. *The Journal of the Acoustical Society of America*, 84(3):917–928, 1988.

C. H. Taal and J. Jensen. SII-based speech preprocessing for intelligibility improvement in noise. In *INTERSPEECH*, pages 3582–3586, 2013.

C. H. Taal, R. C. Hendriks, and R. Heusdens. Speech energy redistribution for intelligibility improvement in noise based on a perceptual distortion measure. *Computer Speech & Language*, 28(4):858–872, 2014.

Y. Tang and M. Cooke. Energy reallocation strategies for speech enhancement in known noise conditions. In *Proc. Interspeech 2010*, pages 1636–1639, 2010.

Y. Tang and M. Cooke. Subjective and objective evaluation of speech intelligibility enhancement under constant energy and duration constraints. In *Proc. Interspeech 2011*, pages 345–348, 2011.

Y. Tang and M. Cooke. Optimised spectral weightings for noise-dependent speech intelligibility enhancement. In *INTERSPEECH*, 2012.

Y. Tang and M. Cooke. Glimpse-Based Metrics for Predicting Speech Intelligibility in Additive Noise Conditions. In *Interspeech 2016*, pages 2488–2492, 2016.

Y. Tang and M. Cooke. Learning static spectral weightings for speech intelligibility enhancement in noise. *Computer Speech & Language*, 49:1–16, 2018.

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.*, 1(1), 2018.

V. C. Tartter, H. Gomes, and E. Litwin. Some acoustic effects of listening to noise on speech production. *The Journal of the Acoustical Society of America*, 94(4):2437–2440, 1993.

M. Torcoli, J. Herre, J. Paulus, C. Uhle, H. Fuchs, and O. Hellmuth. The Adjustment/Satisfaction Test (A/ST) for the Subjective Evaluation of Dialogue Enhancement. In *Proc. of 143rd Audio Engineering Society Conv.*, 2017.

M. Torcoli, A. Freke-Morin, J. Paulus, C. Simon, and B. Shirley. Preferred Levels for Background Ducking to Produce Esthetically Pleasing Audio for TV with Clear Speech. *J Audio Eng Soc.*, 67(12):1003–1011, 2019.

M. A. Tóth, M. L. G. Lecumberri, Y. Tang, and M. Cooke. A corpus of noise-induced word misperceptions for Spanish. *The Journal of the Acoustical Society of America*, 137(2):EL184–EL189, 2015.

J. Tribolet, P. Noll, B. McDermott, and R. Crochiere. A study of complexity and quality of speech waveform coders. In *International Conference on Acoustics, Speech and Signal Processing*, volume 3, pages 586–590, 1978.

M. Uther, M. A. Knoll, and D. Burnham. Do you speak E-NG-L-I-SH? A comparison of foreigner- and infant-directed speech. *Speech Communication*, 49:2–7, 2007.

C. Valentini-Botinhao, M. Toman, M. Pucher, D. Schabus, and J. Yamagishi. Intelligibility Analysis of Fast Synthesized Speech. In *Proc. Interspeech*, pages 2922–2926, 2014.

T. E. M. van Esch, B. Kollmeier, M. Vormann, J. Lyzenga, T. Houtgast, M. Hällgren, B. Larsby, S. P. Athalye, M. E. Lutman, and W. A. Dreschler. Evaluation of the preliminary auditory profile test battery in an international multi-centre study. *International Journal of Audiology*, 52(5):305–321, 2013.

V. S. Varadarajan and J. H. L. Hansen. Analysis of lombard effect under different types and levels of noise with application to in-set speaker ID systems. In *Proc. Interspeech*, 2006.

H. S. Venkatagiri. Segmental intelligibility of four currently used text-to-speech synthesis methods. *The Journal of the Acoustical Society of America*, 113(4):2095–2104, 2003.

N. J. Versfeld and W. A. Dreschler. The relationship between the intelligibility of time-compressed speech and speech in noise in young and elderly listeners. *The Journal of the Acoustical Society of America*, 111(1):401–408, 2002.

M. Virbel, T. E. Hansen, and O. Lobunets. Kivy - A Framework for Rapid Creation of Innovative User Interfaces. In *Mensch & Computer Workshopband*, pages 69–73, 2011.

S. Voran. Objective estimation of perceived speech quality. I. Development of the measuring normalizing block technique. *IEEE Transactions on Speech and Audio Processing*, 7(4):371–382, 1999.

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

M. Wagner and D. G. Watson. Experimental and theoretical advances in prosody: A review. *Language and Cognitive Processes*, 25(7–9):905–945, 2010.

T. Walton, M. Evans, D. Kirk, and F. Melchior. Does Environmental Noise Influence Preference of Background-Foreground Audio Balance? In *Audio Engineering Society Conv. 141*, 2016.

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*, 17(4):348–355, 2008.

D. Weiss and J. J. Dempsey. Performance of Bilingual Speakers on the English and Spanish Versions of the Hearing in Noise Test (HINT). *Journal of the American Academy of Audiology*, 19(1):5–17, 2008.

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

D. Whalen and A. G. Levitt. The universality of intrinsic F0 of vowels. *Journal of Phonetics*, 23(3):349–366, 1995.

C. J. Wild, A. Yusuf, D. E. Wilson, J. E. Peelle, M. H. Davis, and I. S. Johnsrude. Effortful listening: The processing of degraded speech depends critically on attention. *J Neurosci*, 32: 14010–14021, 2012.

A. Wingfield. Cognitive factors in auditory performance: context, speed of processing, and constraints of memory. *Journal of the American Academy of Audiology*, 7(3):175–182, 1996.

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*, 54B(3):P199–P202, 1999.

A. Wingfield, L. Lombardi, and S. Sokol. Prosodic Features and the Intelligibility of Accelerated Speech. *Journal of Speech, Language, and Hearing Research*, 27(1):128–134, 1984.

M. B. Winn and K. H. Teece. Listening Effort Is Not the Same as Speech Intelligibility Score. *Trends in Hearing*, 25, 2021.

M. B. Winn, J. R. Edwards, and R. Y. Litovsky. The impact of auditory spectral resolution on listening effort revealed by pupil dilation. *Ear and Hearing*, 36(4):153–165, 2015.

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*, 22, 2018.

World Health Organization. Regional Office for Europe. Burden of disease from environmental noise: quantification of healthy life years lost in Europe. *World Health Organization. Regional Office for Europe.*, 2011.

Y. H. Wu, E. Stangl, X. Zhang, J. Perkins, and E. Eilers. Psychometric functions of dual-task paradigms for measuring listening effort. *Ear and Hearing*, 37(6):660–670, 2016.

Y.-H. Wu, E. Stangl, O. Chipara, S. S. Hasan, A. Welhaven, and J. Oleson. Characteristics of Real-World Signal to Noise Ratios and Speech Listening Situations of Older Adults With Mild to Moderate Hearing Loss. *Ear and Hearing*, 39(2), 2018.

J. Yamagishi, T. Nose, H. Zen, Z. H. Ling, T. Toda, K. Tokuda, S. King, and S. Renals. Robust speaker-adaptive HMM-based text-to-speech synthesis. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(6):1208–1230, 2009.

L. Zampini, M. Fasolo, and L. D'Odorico. Characteristics of maternal input to children with Down syndrome: A comparison with vocabulary size and chronological age-matched groups. *First Language*, 32(3):324–342, 2012.

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

A. A. Zekveld and S. E. Kramer. Cognitive processing load across a wide range of listening conditions: Insights from pupillometry. *Psychophysiology*, 51(3):277–284, 2014.

A. A. Zekveld, S. E. Kramer, and J. M. Festen. Pupil response as an indication of effortful listening: The influence of sentence intelligibility. *Ear and Hearing*, 31(4):480–490, 2010.

Z. Zhang and Y. Shen. Listener Preference on the Local Criterion for Ideal Binary-Masked Speech. In *Proc Interspeech 2019*, pages 1383–1387, 2019.

Y. Zhao. The Effects of Listeners' Control of Speech Rate on Second Language Comprehension. *Applied Linguistics*, 18(1):49–68, 1997.

T. Zorila, V. Kandia, and Y. Stylianou. Speech-in-noise intelligibility improvement based on spectral shaping and dynamic range compression. In *Proc. Interspeech*, pages 635–638, 2012.

RESUMEN

En nuestra vida cotidiana estamos expuestos a una variedad de tipos de habla, tanto naturales como artificiales. Tanto los hablantes como los que desarrollan mejoras del habla intentan ayudar al oyente modificando las características del habla. Los hablantes modifican su habla cuando se exponen al ruido, produciendo un habla *lombarda*. Los anuncios de megafonía en vivo y grabados pueden contener modificaciones diseñadas para mejorar la inteligibilidad. El habla generada sintéticamente es habitual en dispositivos móviles, asistentes de voz y sistemas de consultas telefónicas.

La percepción del habla es un proceso que involucra tres pasos secuenciales; se escucha, interpreta y comprende un sonido del habla [Moore et al., 2008]. En concreto, se recibe la información auditiva; posteriormente se transforma en una señal neuronal; y finalmente, se procesa la información fonética. El procesamiento del habla es automático y sin esfuerzo cuando ocurre en condiciones ideales. En un entorno silencioso, la información del habla en cuanto a frecuencia y tiempo excede la requerida para percibir el habla con precisión por oyentes con audición normal [Moore, 2008]. En un entorno ruidoso, sin embargo, la tarea de percepción del habla se vuelve más difícil y se requiere un trabajo adicional por parte de los procesos automáticos. Una hipótesis es que hay dos tipos de procesos automáticos que están involucrados en la percepción del habla: a saber, procesos de abajo hacia arriba (*bottom-up*) y de arriba hacia abajo (*top-down*). Durante el proceso de abajo hacia arriba, se analiza la señal de señal de habla entrante, mientras que el proceso de arriba hacia abajo se basa en el conocimiento previo del oyente. El cerebro es capaz de aislar ciertas fuentes de sonido y filtrar otras (mecanismo de 'ganancia selectiva' [Kerlin et al., 2010]). En la literatura, se han sugerido varias técnicas como participantes en el mecanismo automático. Algunas de éstas son el agrupamiento y unión de partes del habla en una sola señal (es decir, agrupación auditiva [Bregman, 1990]), extracción de regiones de tiempo-frecuencia donde el habla objetivo está menos enmascarada (es decir, *glimpses* [Cooke, 2003]), o separación espacial entre la señal objetivo y el enmascarador cuando se encuentran en diferentes regiones [Hawley et al., 2004]. Finalmente, los indicios visuales también son un mecanismo útil para ayudar a distinguir fonemas en ruido [Macleod y Summerfield, 1987].

Aunque hay una serie de factores que pueden interferir con la comprensión óptima del habla, los oyentes con audición normal pueden entender el habla en condiciones severas [Diehl, 2008]. Para lograr una comunicación exitosa, los hablantes modifican naturalmente su estilo de hablar teniendo en cuenta las condiciones ambientales y a su interlocutor [para una revisión, consultar

Cooke et al., 2014a]. Tales condiciones ambientales pueden ser ruido ambiental aditivo -en el que el hablante produce la llamada habla 'lombarda' (por ejemplo, en un reunión social con múltiples hablantes) [Summers et al., 1988; Hazan y Baker, 2011]-, reverberación [Brunskog et al., 2009], o separación amplia entre hablante e interlocutor [Pelegrín et al., 2011]. Por otro lado, los tipos de habla modificados por el hablantee atendiendo a las características del interlocutor incluyen el habla dirigida a bebés [Burnham et al., 2002], a niños con discapacidades de aprendizaje [Bradlow et al., 2003], a oyentes con discapacidad auditiva [Lam and Kitamura, 2012], a no nativos [Sankowska et al., 2011], a máquinas [Mayo et al., 2012] o a mascotas [Burnham et al., 2002].

La intención del hablante es facilitar la comprensión al oyente aumentando la claridad del habla y reduciendo el esfuerzo cognitivo requerido. Esto se logra haciendo adaptaciones acústicas y lingüísticas, por separado o en combinación. Para las modificaciones acústicas en particular, un mecanismo es mejorar la audibilidad aumentando la intensidad vocal [Picheny et al., 1986; Castellanos et al., 1996; Pelegrín et al., 2011], elevando la frecuencia fundamental para desplazar el espectro a frecuencias a las que el oído es más sensible [Bond y Moore, 1990; Pittman y Wiley, 2001], realzando los sonidos sonoros en cuanto a intensidad y duración [Boril y Pollak, 2005] y reasignando energía espectro-temporal [Lu y Cooke, 2008]. Otro mecanismo es aumentar la coherencia del habla en presencia de sonidos competidores aumentando la modulación del habla [Krause y Braida, 2004; Boril y Pollak, 2005], con cambios en los dos primeros formantes [Picheny et al., 1986; Bradlow et al., 2003], o insertando pausas entre palabras [Picheny et al., 1986]. Finalmente, también se pueden aplicar modificaciones a nivel lingüístico, como usar un vocabulario más simple.

Para estudiar el efecto de diferentes tipos de habla sobre la percepción del habla, los investigadores suelen evaluar la inteligibilidad. La inteligibilidad se puede definir como el porcentaje de palabras reconocidas con precisión (tasa de reconocimiento de palabras). Los factores que pueden reducir la inteligibilidad incluyen condiciones de audición imperfectas, con o sin aspectos como el ruido ambiental o la reverberación; limitaciones del interlocutor, como no ser nativo o tener problemas de audición; y limitaciones del hablante, como el habla con acento. La inteligibilidad disminuye en función de la ratio señal-ruido, conocida como SNR (*Signal-to-Noise-Ratio*): es decir, una SNR más baja conduce a una inteligibilidad más baja. Además, la reducción de la inteligibilidad se ve afectada de manera diferente para los diferentes tipos de voz, SNR, distorsiones, enmascaradores y reverberación [Picheny et al., 1985; Picheny et al., 1986; Summers et al., 1988; Robinson et al., 2022]. Por ejemplo, se ha demostrado que el habla "alargada" aumenta la inteligibilidad en el ruido multi-hablante [Adams y Moore, 2009], mientras que en el ruido estacionario no se observaron ganancias significativas [Nejime y Moore, 1998]. Además, el aplanamiento de la inclinación espectral conduce a mejoras en la inteligibilidad en presencia de ruido, pero el aumento de la frecuencia fundamental no tiene ningún impacto [Lu y Cooke, 2009a].

La recepción correcta del mensaje es crítica en muchas situaciones. En consecuencia, se ha dedicado un gran esfuerzo a evaluar el efecto sobre la inteligibilidad que tienen diferentes estilos de habla [Cooke et al., 2013a] y cambios en diversas propiedades del habla [Nejime y Moore, 1998; Adams y Moore, 2009; Lu y Cooke, 2009a]. Los algoritmos de mejora de la escucha cercana al oyente (*near end*) pueden lograr mejoras significativas en la comprensión del habla en comparación con el habla no procesada en condiciones adversas [Taal y Jensen, 2013; Schepker et al., 2015]. Sin embargo, el habla percibida después de la mejora de la

escucha cercana al oyente podría no ser completamente satisfactoria para el oyente, ya que los algoritmos de mejora del habla comúnmente utilizados se centran principalmente en mejorar la inteligibilidad. Deben también tenerse en cuenta otros aspectos subjetivos del habla percibida, como el esfuerzo auditivo, la calidad, naturalidad, que sea grato escucharla y las preferencias generales del oyente. Para referirse a los atributos del habla más allá del reconocimiento de palabras, se utiliza el término 'supra-inteligibilidad'. El objetivo de esta tesis es doble: estudiar los aspectos de supra-inteligibilidad del habla en términos de esfuerzo auditivo y preferencias del oyente, y desarrollar una herramienta para investigar los aspectos de supra-inteligibilidad del habla.

Complementaria a la dimensión de la claridad del habla es la experiencia general del oyente, que ha sido mucho menos investigada. Escuchar puede volverse difícil, incluso cuando la inteligibilidad está al nivel más alto posible. El esfuerzo de escucha refleja los recursos cognitivos necesarios para la comprensión del habla. A veces es necesario un gran esfuerzo en situaciones con ruido de fondo, baja intensidad del habla, conexión móvil deficiente, habla con acento o alta motivación del oyente (por ejemplo, mayor dilatación máxima de la pupila para mayores recompensas; ver una revisión reciente en Carolan et al. [2022]). También se pueden encontrar variaciones en el esfuerzo de escucha entre diferentes poblaciones. Por ejemplo, los oyentes no nativos realizan un mayor esfuerzo en comparación con los oyentes nativos, incluso cuando realizan una tarea al mismo nivel [Borghini y Hazan, 2018].

Una alta asignación de recursos cognitivos impone una gran desventaja al oyente, lo que lleva a un rendimiento reducido en multitareas [Sarampalis et al., 2009], mayor sensación de escucha y/o fatiga mental, o rechazo a la vida social. En un caso más extremo [Organización Mundial de la Salud. Oficina Regional para Europa., 2011], trabajar en un entorno desagradable con anuncios frecuentes y ruidosos puede provocar problemas de salud. El esfuerzo de escucha se ha estimado utilizando medidas subjetivas como cuestionarios, métricas de comportamiento (por ejemplo, tiempo de respuesta) y medidas fisiológicas como la pupilometría (ver la revisión de McGarrigle et al. [2014]).

El esfuerzo de escucha se puede considerar como uno de los varios aspectos individuales de las preferencias del oyente. Las preferencias del oyente surgen del juicio generalizado de la percepción del habla que incluye factores como la inteligibilidad, la naturalidad y el agrado. Las preferencias de los oyentes se pueden recopilar permitiéndoles modificar las propiedades del habla utilizando herramientas de ajuste. Los oyentes están familiarizados con el concepto de modificaciones suaves de audio, como las que se usan para ajustar el volumen en la televisión y la radio. Las respuestas de los oyentes derivadas de las herramientas de ajuste del habla pueden ser precisas, ya que el habla se puede modificar finamente, en contraste con las pruebas tradicionales en las que el oyente solo tiene pocas opciones. Estudios previos han sugerido diferentes modificaciones de características del audio en tiempo real [Assmann y Nearey, 2007; Kean et al., 2015; Zhang y Shen, 2019; Novak III y Kenyon, 2018]. Se puede esperar que tales preferencias varíen según el entorno de escucha [Kean et al., 2015; Walton et al., 2016] y cualquier discapacidad auditiva [Buyens et al., 2014; Shirley et al., 2017], además de tener un componente individual [Walton et al., 2016]. Una mejor comprensión de la base de las preferencias del oyente promete aportar información para el diseño de algoritmos de modificación del habla que sean capaces tanto de aumentar la inteligibilidad como de reducir el esfuerzo auditivo, proporcionando una mejor experiencia auditiva general.

La motivación de esta tesis es dilucidar el efecto 'difícil de escuchar' (es decir, la condición

en la que el habla requiere más esfuerzo por parte del oyente) evaluando la contribución de distintos factores del habla a este efecto para diferentes condiciones de escucha. El objetivo general de esta tesis es determinar si los oyentes manifiestan preferencias de suprainteligibilidad cuando se les da la posibilidad de manipular distintas propiedades del habla que los hablantes modifican naturalmente para producir habla lombarda. La hipótesis principal es que cuando la inteligibilidad está en niveles máximos, los oyentes intentarán reducir el esfuerzo de escucha y mantener la calidad del habla. Se supone que, para las condiciones en las que se maximiza la inteligibilidad, la relación entre las preferencias del oyente y una amplia gama de valores de las características del habla (por ejemplo, la pendiente espectral) tendrá una distribución en forma de campana. Además del aspecto centrado en el usuario de las preferencias del oyente, cuando los oyentes del mismo grupo (por ejemplo, adultos jóvenes frente a adultos mayores, normo-oyentes frente a personas con discapacidad auditiva, nativos frente a no nativos) escuchan en las mismas condiciones (por ejemplo, ruido estacionario, habla competidora), se espera que valores similares de rasgos del habla exciten patrones cognitivos o auditivos similares (por ejemplo, frecuencias a las que el oyente es particularmente sensible, intensidad del habla, características prosódicas de un idioma que le son familiares).

Esta tesis describe una investigación sobre dos factores de supra-inteligibilidad: el esfuerzo de escucha y las preferencias del oyente. En primer lugar, se recopilaron medidas del esfuerzo de escucha para diferentes estilos de habla y se realizaron comparaciones entre oyentes nativos y no nativos. Se investigaron tres medidas del esfuerzo auditivo: (i) una medida objetiva de inteligibilidad, (ii) una medida fisiológica del esfuerzo auditivo (tamaño de la pupila) y (iii) los juicios subjetivos de los oyentes. Los tipos de habla examinados fueron habla simple (natural), habla producida en ruido (habla lombarda), habla mejorada para promover la inteligibilidad (SSDRC; Zorila et al. [2012]) y habla sintética (TTS; Yamagishi et al. [2009]) en presencia de ruido con forma de voz (*speech shaped noise*) a tres SNR diferentes. Se realizaron dos experimentos, uno con oyentes nativos y otro con no nativos. Para el primer experimento, se reclutaron 26 oyentes nativos de inglés británico. Los estímulos se presentaron a -5, -3 y -1 dB de SNR. Para el segundo experimento, la configuración fue idéntica a la anterior excepto por los niveles de ruido que fueron -1, +5, +20 dB de SNR. En total se reclutaron 31 oyentes españoles con alto nivel de competencia en inglés. También se realizó una prueba web paralela para que los oyentes nativos de inglés evaluaran el acento en inglés de los participantes españoles. Se recogieron las puntuaciones de los 10 evaluadores. Los resultados revelaron que las valoraciones subjetivas del esfuerzo estaban correlacionadas con la inteligibilidad, mientras que no siempre eran consistentes con las respuestas pupilares. Además, fue evidente un claro impacto del tipo de habla en las demandas cognitivas requeridas para la comprensión del habla. La dilatación de la pupila determinó que el habla lombarda impone demandas menores en el procesamiento mental en comparación con el habla simple, síntesis *text to speech* y habla mejorada artificialmente.

A continuación, se introdujo e implementó una herramienta llamada *SpeechAdjuster* para investigar los aspectos de inteligibilidad y supra inteligibilidad del habla. *SpeechAdjuster* es una herramienta de código abierto que invierte los roles de oyente y experimentador al permitir que los oyentes controlen directamente las características del habla en tiempo real. Este cambio de paradigma permite medir directamente las preferencias de los oyentes, sin recurrir a escalas de calificación. La incorporación de una fase de prueba en la que se congelan las preferencias también permite estimar la inteligibilidad dentro del mismo ensayo. El cálculo previo (*offline*) y la interpolació en línea (*online*) dentro de la herramienta permiten medir el impacto de los

cambios en prácticamente cualquier característica del habla presentada (por ejemplo, frecuencia fundamental o pendiente espectral) o característica de fondo (por ejemplo, espectro de ruido), independientemente de su complejidad.

Se realizaron varios experimentos con *SpeechAdjuster* para explorar los efectos sobre las preferencias auditivas y la inteligibilidad de las propiedades del habla que normalmente se modifican en el habla mejorada de forma natural. En primer lugar, se investigó la relación entre la velocidad del habla y las propiedades del enmascarador. El habla rápida puede reducir la inteligibilidad, pero hay poco acuerdo sobre si los oyentes se benefician de un habla más lenta en condiciones ruidosas. Dieciocho oyentes nativos de español ajustaron la velocidad del habla mientras escuchaban secuencias de palabras en silencio, en ruido estacionario con relaciones señal-ruido de 0, +6 y +12 dB, y en ruido modulado para 5 velocidades de modulación de envolvente. Después de seleccionar una velocidad preferida, los participantes identificaron las palabras presentadas a esa velocidad. En segundo lugar, se investigaron las preferencias con respecto a la frecuencia fundamental para el habla presentada en enmascaradores de energéticos y de habla competidora. Los beneficios de inteligibilidad relacionados con la frecuencia fundamental (F0) no están claros actualmente, mientras que las preferencias de F0 de los oyentes han sido poco investigadas. En este sentido, se realizaron dos experimentos para investigar las preferencias de los oyentes sobre el F0 ante enmascaradores energéticos e informativos. Para los experimentos, se reclutaron oyentes nativos griegos. El material de oraciones del corpus griego utilizado se presentó en Sfakianaki [2019]. En el primer experimento, se recogieron las preferencias de F0 de 17 oyentes en silencio y en presencia de ruido estacionario a -3, 0 y +3 dB de SNR. En el último experimento, se recogieron las preferencias de F0 de 23 oyentes en silencio y en presencia de un hablante competidor (el mismo que el hablante objetivo) a -10, -6 y -3 dB de SNR. Finalmente, se investigaron las propiedades espectrales (inclinación espectral, modificaciones de energía de banda espectral, características del filtro de frecuencia) del habla presentada en condiciones de enmascaramiento energético. El aplanamiento de la inclinación espectral ha revelado ganancias de inteligibilidad en presencia de ruido [Lu y Cooke, 2009a]. En nuestro experimento, los oyentes ajustaron las propiedades espectrales del ruido estacionario en relaciones señal/ruido de -6, -3 y 0 dB y, posteriormente, se evaluó la inteligibilidad.

Esta tesis estudia los efectos de las características del habla distintas de la intensidad de la señal en las preferencias del oyente. En todo momento, los estímulos se normalizaron para tener la misma energía cuadrática media antes y después de la modificación. Este enfoque es común, por ejemplo, en la evaluación del rendimiento de los algoritmos de mejora del habla [Cooke et al., 2013a, Rennies et al., 2020], y lleva a centrarse en las modificaciones del habla que benefician a los oyentes independientemente del simple recurso de aumentar la audibilidad elevando el nivel de la señal. Una consecuencia de la normalización es que las modificaciones del habla siempre representan el resultado conjunto resultante tanto del efecto directo del parámetro modificado en sí mismo (por ejemplo, una inclinación espectral más plana) como del efecto sobre cualquier cambio en la SNR local a lo largo del tiempo y la frecuencia debido a la normalización posterior (por ejemplo, más energía en las frecuencias medias). Para evaluar los cambios en la SNR local que resultan de la modificación de los parámetros del habla, se introdujo una nueva métrica que mide la distribución de los *glimpses* del habla a través de la frecuencia.

Los resultados revelaron que los oyentes tienen distintas preferencias por las características del habla evaluadas, lo que revela aspectos del habla más allá de la inteligibilidad. Específicamente, para una inteligibilidad constante: (1) en silencio, los oyentes ajustaron el habla para estar cerca

del habla original (es decir, simple); (2) para enmascaradores modulados, los oyentes preferían velocidades de voz moduladas de una manera que contrastaba con las del enmascarador; (3) para el ruido estacionario, los oyentes prefirieron una velocidad de habla más lenta a medida que aumentaba el nivel de ruido; (4) independientemente del nivel de ruido, los oyentes eligieron una frecuencia fundamental media ligeramente más baja en comparación con la original; (5) para el ruido estacionario, los oyentes prefirieron reasignar la energía del habla, eligiendo configuraciones que mejoraran la energía en las frecuencias más bajas. Los resultados también mostraron que cuanto más exigente cognitivamente era la tarea, mayor era el tiempo de ajuste que necesitaban los oyentes: (1) en silencio, necesitaban alrededor del tiempo permitido más bajo (5 s); (2) en ruido, necesitaron progresivamente más tiempo dependiendo del nivel de ruido; (3) necesitaban tiempos diferentes para los diferentes tipos de enmascaradores (por ejemplo, más tiempo para el ruido modulado en comparación con el ruido con forma de voz para la misma SNR; (4) necesitaban tiempos diferentes para las características que provocan distorsiones acústicas o fonológicas diferentes (por ejemplo, el filtro de banda deslizante requirió el mayor tiempo).

El estudio de la pupilometría se utilizó como investigación de referencia para comprender el esfuerzo que implican los estilos de habla que se pueden encontrar en condiciones de la vida real. Cada uno de los tipos de voz probados implica cambios de una o más de las características estudiadas en los experimentos de preferencias del oyente realizados con SpeechAdjuster. Por lo tanto, se puede obtener información sobre si los valores preferidos de las distintas características del habla han contribuido a reducir el esfuerzo de escucha requerido.

Las preferencias de los oyentes revelaron un intento de reducir las demandas cognitivas. Según el modelo de Facilidad de Comprensión del Lenguaje [ELU; Rönnberg et al., 2013], en condiciones ideales, la comprensión del habla es un proceso implícito, automatizado y sin esfuerzo, mientras que el habla distorsionada (por ejemplo, condiciones ruidosas, procesamiento de señales, pérdida de audición) es perjudicial para este proceso. Para dar sentido a una señal de voz distorsionada, se mejora el análisis cognitivo de arriba hacia abajo (*top-down*) [Gatehouse, 1990; Pichora-Fuller et al., 1995; Wingfield, 1996] y se activa el procesamiento cognitivo explícito, lo que requiere más recursos cognitivos. La capacidad limitada de la memoria de trabajo [Kahneman, 1973] hace que esta tarea sea laboriosa. El grado de procesamiento explícito necesario para la comprensión del habla está positivamente relacionado con el esfuerzo [Rönnberg et al., 2019]. En nuestros experimentos, en condiciones de silencio, los oyentes puede que hayan elegido el habla original (es decir, simple), dado que las distorsiones causadas por el procesamiento del habla para las opciones restantes pueden haber provocado desajustes fonológicos con la representación mental esperada. En ruido, los oyentes eligieron los valores de las características que superaran el enmascaramiento energético evitándolo, ya sea en el tiempo, es decir, eligiendo una velocidad de habla objetivo que contraste con la del enmascarador de ruido modulado por habla, o espectralmente, es decir, reasignando energía espectral cuando el habla se enmascara con ruido en forma de habla. Estas preferencias pueden haberse derivado del deseo de los oyentes de reducir el esfuerzo de escucha. En presencia de ruido, parte de la información acústica se enmascara, lo que reduce la audibilidad. Los segmentos del habla que faltan o están incompletos conducen a una falta de coincidencia con la representación léxica almacenada; por lo tanto, la señal acústica requiere más procesamiento perceptivo para interpretar el habla. Este proceso da como resultado un mayor esfuerzo de escucha [Winn y Teece, 2021]. Los oyentes bajo la condición de ruido en forma de voz pueden haber seleccionado veloci-

dades de habla más lentas a medida que aumentaba el nivel de ruido para aumentar el tiempo disponible para procesar el habla.

Cuanto más exigente cognitivamente sea una condición de escucha, más tiempo necesitará el oyente para encontrar el valor "óptimo". Los oyentes en nuestros experimentos no tenían limitaciones de tiempo mientras realizaban la tarea; por lo tanto, podían dedicar todo el tiempo necesario al aprendizaje perceptivo. El sistema perceptivo es capaz de recalibrar los procesos del habla y adaptarse a las distorsiones que impone el habla [Samuel y Kraljic, 2009]. En condiciones de silencio, se supone que la comprensión del habla se realiza sin esfuerzo; por lo tanto, los oyentes en estos experimentos necesitaron casi el menor tiempo permitido para finalizar su selección. Para niveles de ruido crecientes, las demandas de procesamiento aumentan y el tiempo de ajuste se hizo progresivamente más alto. Por ejemplo, para la función de inclinación espectral a -3 y 0 dB de SNR, los oyentes lograron puntuaciones de inteligibilidad casi iguales, mientras que necesitaron más tiempo para ajustar el habla en las condiciones más adversas. Las demandas de procesamiento adicionales requeridas en condiciones con más ruido se reflejaron en la actividad pupilar, una medida conocida del esfuerzo cognitivo en la que la dilatación máxima de la pupila aumenta con el nivel de ruido [Ohlenforst et al., 2017]. Finalmente, además de las demandas cognitivas adicionales que impone el mayor nivel de ruido, los diferentes tipos de enmascaradores también tuvieron diferentes impactos en el esfuerzo [Brungart et al., 2013]. En el experimento que investigaba la velocidad del habla, los oyentes dedicaron más tiempo a ajustar la velocidad del habla en presencia de ruido modulado temporalmente en comparación con el ruido estacionario. La naturaleza modulada del enmascarador puede imponer una carga cognitiva adicional al oyente.

Otro hallazgo es que los oyentes pueden haber intentado mantener la calidad del habla. Los oyentes eligieron valores característicos que no tienen un impacto negativo en la naturalidad. Específicamente, los oyentes eligieron una frecuencia fundamental similar a la del discurso original, ya que un simple cambio en la frecuencia fundamental sin los ajustes de formantes apropiados tiene un efecto negativo en la naturalidad [Assmann et al., 2006]. De acuerdo con nuestros hallazgos, los oyentes de Assmann y Nearey [2007] no preferían un simple cambio en la frecuencia fundamental. Además, en ruido, los oyentes de nuestros experimentos no eligieron las opciones que implicaban una atenuación extrema del tono y la información armónica. En estudios previos, la calidad de los algoritmos que potencian las frecuencias medias y altas, sacrificando energía por debajo de los 1000 Hz, se consideró más pobre en comparación con otros [Gabrielsson et al., 1988; Tang et al., 2018].

Los oyentes pueden haber hecho sus elecciones basándose en lo que les resulta familiar. En [Tang et al., 2018], los oyentes en condiciones tranquilas con inteligibilidad a niveles máximos prefirieron el habla simple sobre el habla modificada y el habla simple se calificó como de mejor calidad. A los oyentes no se les dio ninguna referencia específica para la evaluación de la calidad, pero una explicación a estos resultados es que aplicaron un estándar de calidad consistente basado en su experiencia. En nuestros experimentos, en condiciones ruidosas, los oyentes eligieron características del habla similares a las que producen naturalmente los hablantes en ambientes ruidosos. Específicamente, los oyentes preferían velocidades de habla más lentas e inclinaciones espectrales más planas, y para niveles de ruido más altos el efecto era mayor [Tartter et al., 1993]. El habla en ruido también se caracteriza por un aumento en el tono (F0), que no se observó en nuestros resultados. El deterioro de la calidad puede haber disuadido a los oyentes de seleccionar un tono más alto.

Finalmente, las elecciones de los oyentes influidas por la calidad del habla y la familiaridad pueden conducir a una reducción en el esfuerzo de escucha. Las distorsiones del habla que producen una peor calidad del habla pueden requerir la utilización de mayores recursos cognitivos. Se ha demostrado que, para una inteligibilidad constante, los cambios en la calidad de la señal, como una mayor resolución espectral en una simulación de implante coclear, pueden resultar en una disminución del esfuerzo auditivo medido con el paradigma de doble tarea [Pals et al., 2013]. Además, se ha demostrado que los tipos de habla producidos de forma natural que los oyentes están acostumbrados a escuchar en condiciones específicas son menos exigentes desde el punto de vista cognitivo. En Borghini y Hazan [2020], para el mismo nivel de inteligibilidad del habla, el esfuerzo cognitivo aumentó al escuchar habla simple en lugar de habla clara en entornos con habla multihablante de fondo. Además, en nuestros experimentos encontramos que el habla lombarda fue la que menos esfuerzo requirió en comparación con los tipos de habla simples y artificiales en entornos de ruido en forma de habla. Nuestros oyentes, en condiciones ruidosas, prefirieron las características del habla clara y lombarda (es decir, una velocidad de habla más lenta, una inclinación espectral más plana). Sin embargo, las distintas características del habla que condujeron a la reducción del esfuerzo de escucha deben investigarse más a fondo.

El resultado conjunto de los experimentos descritos aquí sugiere que los oyentes exhiben preferencias de suprainteligibilidad cuando se les da la posibilidad de manipular distintas propiedades del habla. Los oyentes, en presencia de ruido de fondo, eligieron alargar el habla, modificaciones espectrales que causan el menor daño posible a las frecuencias más bajas y seleccionaron frecuencias fundamentales ligeramente más bajas en comparación con el habla original. Diseccionar la relación entre las preferencias del oyente y la calidad, la naturalidad y el esfuerzo cognitivo es un área fructífera para futuras investigaciones.