Université du Québec
Institut National de la Recherche Scientifique
Centre - Énergie Matériaux Télécommunications

A non-intrusive objective speech intelligibility metric tailored for cochlear implant users in complex listening environments

by

João Felipe Santos

A thesis submitted in fulfillment of the requirements
for the degree of
Master of Science (M.Sc.)
in Telecommunications

December 2, 2013

Evaluation committee

Internal evaluator and committee president:	Prof. Sofiène Affes
External evaluator:	Prof. Márcio Holsbach Costa (Universidade Federal de Santa Catarina)
Research advisor:	Prof. Tiago H. Falk

Abstract

In this work, we propose a speech intelligibility metric tailored for CI users, termed SRMR-CI. This metric is based on the speech-to-reverberation modulation energy (SRMR) measure, which has been shown to correlate well with speech quality and intelligibility for normal hearing listeners. Our proposal extends SRMR to improve its correlation to CI speech intelligibility based on three main modifications. First, we replaced the acoustic filterbank used in SRMR by a CI-inspired filterbank, similar to the one used in the speech coding strategy present in Nucleus CI devices. Adjustment of the filterbank alone led to improvements of 21% in linear correlation, 6% in rank correlation, and 14% in sigmoidal-mapped correlation between the objective and subjective scores when compared to the original SRMR. The second modification was an adjustment to the modulation frequency range used by the metric to compute energies in the modulation spectrum, in order to reduce the metric's sensitivity to the fundamental frequency of speech. Lastly, we propose a modulation energy thresholding scheme to account for variability related to speech content. These two additional steps can bring additional gains of 19% in correlations and a relative decrease of 50% in the root-mean-square error of the metric predictions.

We evaluated the proposed metric and several state-of-the-art metrics using subjective cochlear implant listening ratings for speech in noisy, reverberant, and non-linearly enhanced conditions. The proposed metric showed performance in par with intrusive metrics under environmental distortion conditions and outperformed all the benchmark metrics when assessing the intelligibility of non-linearly processed speech. A non-intrusive metric such as SRMR-CI is a valuable resource for developing and validating new cochlear implant devices and enhancement algorithms, and can potentially be used in devices to adjust system parameters on the go.

Acknowledgements

First of all, I would like to express my gratitude to my supervisor, Dr. Tiago H. Falk, whose expertise added considerably to my graduate experience. Without his support, advice, great ideas, and attention to detail, this work would not have been possible.

I would like to extend my thanks to the members of the evaluation committee, Prof. Sofiène Affes and Prof. Márcio Holsbach Costa, as well as to the INRS-EMT staff and faculty. This research was made possible by funding from the Natural Sciences and Engineering Research Council of Canada (NSERC).

I would also like to thank the collaborators/co-authors of our papers: Dr. Oldooz Hazrati and Prof. Philip C. Loizou (in memoriam), who have provided access to a database of subjective listening experiments with cochlear implant users, as well as the implementation for their metric, and expert advice during earlier stages of this work. I had also the opportunity of working with Stefano Cosentino while he visited our lab. He has also provided suggestions and the implementations for many of the benchmark metrics used in this work.

The Multimedia/Multimodal Signal Analysis and Enhancement (MuSAE) Lab is a wonderful environment for research and not only because we have the resources and expertise, but because of the great people that are part of it. I feel honoured to be part of such a team. Thanks, everybody!

A big thank you goes also to my friends and family, particularly to those that are far now.

Finally, I would like to dedicate this thesis to the person that chose to be by my side no matter where or when, whose love, support, and encouragement were essential to each part of this journey: my wife. Ciana, I owe you all of this. Thank you, and let's see what happens next!

Contents

List of Figures

List of Tables

Abbreviations

AI	Articulation Index
ACE	Advanced Combination Encoder
CI	Cochlear Implant
CIS	Continuous Interleaving Sampling
CSII	Coherence-based Speech Intelligibility Index
ERB	Equivalent Rectangular Bandwidth
IRM	Ideal Reverberant Masking
ModA	Average Modulation-spectrum Area
MTF	Modulation Transfer Function
NCM	Normalized Covariance Metric
NH	Normal Hearing
RMSE	Root Mean Square Error
RSD	Relative Standard Deviation
RT60	Reverberation time
SNR	Signal-to-noise ratio
SRMR	Speech-to-Reverberation Modulation Energy Ratio
SRMR-CI	Speech-to-Reverberation Modulation Energy Ratio tailored to Cochlear Implant devices
STI	Speech Transmission Index

Symbols

ρ	Pearson's correlation coefficient
ρ_{spear}	Spearman rank correlation
ρ_{sig}	Pearson's correlation coefficient of the sigmoidal mapping

Chapter 1

Introduction

According to the Canadian Hard of Hearing Association, 10% of the population in Canada has some degree of hearing loss [1]. In seniors, this problem is even more pervasive: more than half of Canadians over the age of 65 will experience late deafening caused either by the cumulative effect of aging on hearing, noise exposure, diseases such as diabetes and hypertension, or a combination of factors. Hearing loss can lead to depression, dissatisfaction with life, reduced functional and cognitive health, and withdrawal from social activities. The authors of a report by the Quebec Network for Research on Aging predict that by 2015, the number of people age 65 and over in Quebec will surpass the number of people age 0 to 19 [2]. Quebec has one of the most rapidly aging societies in the world. The 2003 Canadian Community Health Survey shows that 7% of all seniors in the province have reported hearing problems [3].

Hearing assistive technologies play a big role on improving communication and quality of life for people living with hearing loss. In most cases (mild to moderate sensorineural hearing loss), hearing aids are sufficient to improve communication. Individuals suffering from profound hearing loss in both ears usually require cochlear implants (CI), which stimulate the cochlea directly with electrical pulses.

Since the focus of this thesis is on the development of an objective intelligibility metric for CI, the remainder of this chapter will present the background on CI hearing and intelligibility assessment. The chapter concludes with a detailed description of the thesis contributions.

1.1 Speech intelligibility in cochlear implant users

In persons with normal hearing, the pressure waves captured at the outer ear flow through the ear canal and reach the middle ear, where they make the eardrum vibrate and transmit this vibration to tiny ossicular bones [4]. These bones then transmit this vibration to the oval window of the cochlea, situated at the inner ear. The cochlea is a spiral shaped tube filled with a gelatinous fluid, where a structure called basilar membrane transforms the mechanical vibration into electrical impulses by exciting the so-called hair cells, which are auditory receptors connected to the auditory nerve. The basilar membrane shape and tautness vary along its length, which causes its frequency response to change accordingly. Each location at the basilar membrane has a different characteristic frequency, and the space-time patterns of hair cell firings are interpreted by the brain as a coding of different sounds. Damage to hair cells is the main cause of hearing loss. In persons with profound hearing loss, usually there is a significant loss of hair cells and/or auditory neurons, which makes it impossible for the cochlea to excite the auditory nerve fibers [5].

Considering that there was no significant damage to the auditory nerve fibers, the auditory neurons may be excited directly through electrical stimulation. A cochlear implant explores this possibility by using an array of electrodes to stimulate the cochlea with electrical pulses at different regions, thus corresponding to different frequencies of the acoustic signal [6]. These devices usually consist of five components, as depicted in Figure 1.1:

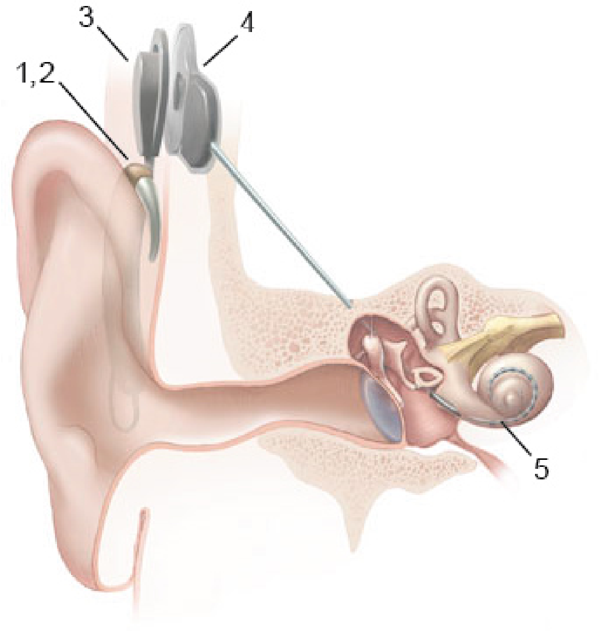


Figure 1.1: Illustration of a cochlear implant [Source: National Institute on Deafness and Other Communication Disorders at the American National Institutes of Health (public domain)].

1. **Microphone(s)** For capturing the audio signal from the environment where the user is located.

2. **Speech processor** Responsible for speech coding and enhancement.

3. **Transmitter** Transmits power and the coded sound signal across the skin via a radio-frequency link.

4. **Receiver and stimulator** Receives and converts the signal received from the speech processor to electrical pulses.

5. **Electrode array** Array consisting of 4 to 22 electrodes that stimulate the auditory nerve fibers at the cochlea.

Electrical stimulation must be able to encode both acoustic signal characteristics such as amplitude and frequency. CIs use a speech processing strategy for determining how to represent these characteristics as a series of electrical

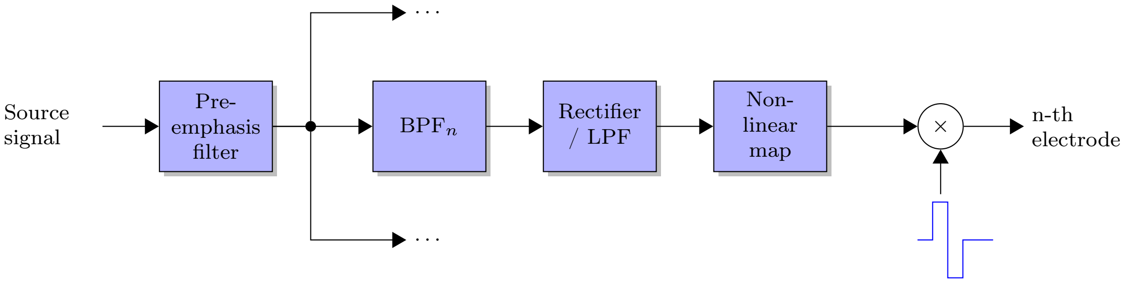


Figure 1.2: Block diagram of the CIS speech coding strategy.

pulses to different regions of the cochlea. These strategies have to take into account particularities of electrically evoked hearing, such as the narrower dynamic range, sparser frequency representation, and electrode interaction (vector summation of electric fields from different electrodes) [7]. A simple processing strategy that is still used in today's devices is continuous interleaved sampling (CIS), for which a simplified block diagram is shown in Figure 1.2. The audio signal passes through a pre-emphasis filter (that attenuates strong components in speech below 1.2 kHz), and then passes through a filterbank composed of band pass filters ($BPF_1, BPF_2, \ldots, BPF_N$, where N is the number of electrodes), its envelope is detected by a rectifier and low-pass filtered, compressed by a non-linear mapping function, and modulated by interleaved pulses (with stimulation rates of approximately 1000 pulses/s/electrode), in a way no simultaneous pulses are transmitted to different electrodes. In this case, even though no electrodes are excited at the same time, all the electrodes in the electrode array are excited with higher or lower energy depending on the subband envelope. Some other strategies, such as the spectral peak (SPEAK) and the advanced combination encoder (ACE), only excite the subset of the electrodes corresponding to the bands with the highest amplitude.

Despite the technological advances of the recent years, cochlear implant users still face many limitations when in noisy and reverberant environments. Even in environments with low reverberation times, like small rooms, reverberation greatly reduces speech intelligibility [8]. In most daily settings, both noise and reverberation are combined, resulting in poor levels of speech understanding [9]. Reverberation and noise distort important speech envelope modulation information, which is the main cue used by speech processing strategies in CI to encode the speech signal. These distortions make it extremely challenging for CI users to perceive e.g., pitch modulations, formant transitions, timbre, and word/syllable boundaries [10–12], introduce unwanted masking effects [8, 13, 14], and cause poor sound localization [15]. For example, recent research has shown that in clean conditions, an average 84% intelligibility can be achieved. This drastically decreases to 20% in a reverberant room (reverberation time of 1.0 s) and to 17% in a noisy room (signal-to-noise ratio, SNR $=$ 0 dB) [9, 16]. To overcome these limitations and to improve speech intelligibility in everyday environments, recent research has focused on the development of speech enhancement algorithms, such as noise suppression, channel selection, and dereverberation (e.g., [17, 18]).

Assessing the quality of CI devices and newly-developed enhancement algorithms, as well as the effects of environmental distortions on them, usually requires subjective listening tests. These can be performed either with CI users or by presenting vocoded speech (to simulate CI processing) to normal hearing (NH) listeners. However, such tests are expensive and time-consuming, which hinders their use during product development cycles. Objective speech quality and intelligibility metrics are an alternative to subjective listening tests and enable fast and repeatable assessment, while also allowing for on-the-fly measurements directly on the device. The latter is especially important in case speech quality/intelligibility is to be monitored and used to adapt parameters of a speech enhancement system in real time. The next section will provide a brief description of objective speech intelligibility measurement; a more detailed presentation will be given in Section 2.3.

1.2 Objective speech intelligibility assessment

Objective intelligibility (or quality) metrics can be broadly classified as intrusive (also known as double-ended or full-reference) or non-intrusive (single-ended or no-reference) depending on the need for a reference clean signal or not, respectively [19]. Intrusive metrics have the advantage of being able to assess directly the amount and type of distortion in a corrupted signal. While both can be used during the development of an enhancement algorithm or for evaluation/comparison of different CI devices, intrusive metrics cannot be used in practical real-time applications, as in this case a reference clean signal is not available. Since non-intrusive metrics do not require a clean reference signal, it is possible to apply them to quantitatively characterize the intelligibility gains achieved with a blind speech enhancement algorithm (e.g., dereverberation) directly on the device. They also enable the development of intelligibility-aware enhancement algorithms, which could adjust CI device parameters in real time taking into consideration the current intelligibility settings imposed by environmental effects (such as background noise and reverberation levels). However, when compared to intrusive metrics, non-intrusive algorithms have a stronger dependency on the speech material used on the measurements, as they do not have access to a reference signal to serve as a normalization factor. As a consequence, non-intrusive metrics commonly present higher variability in their predictions. Reducing this variability is an important step in the development of reliable non-intrusive metrics that are sensitive only to environmental/processing characteristics (e.g, speech level, background noise, and room acoustics) but not on the content of the speech signal itself. As detailed below, this is one of the mean objectives of this thesis.

1.3 Thesis objectives and contributions

In this work, we investigate the performance of both intrusive and non-intrusive state-of-the-art objective metrics and compare their performance with subjective intelligibility scores obtained from CI users via listening tests. We also propose a new non-intrusive measure by refining a previously-proposed metric termed SRMR (speech-to-reverberation modulation energy ratio) [20] to emulate CI hearing precepts and reduce its variability. We show that the investigated intrusive metrics achieve reliable performance under environmental distortions, but their performance is degraded when trying to assess intelligibility of non-linearly processed/enhanced speech. Moreover, the proposed CI-inspired metric, in turn, is shown to outperform all other benchmarks and achieves results in line with the intrusive metrics, but with the advantage of not requiring a clean reference signal. The proposed metric also shows improved performance with processed speech signals when compared to other state-of-the-art metrics.

Part of the work described in this thesis has either already been published or is currently under review for publication in conference proceedings and scientific journals. In [21], we evaluated the performance of intrusive objective metrics for CI speech intelligibility assessment. In [22], an initial version of the SRMR-CI metrics was presented. This version did not take into account the variability reduction procedures that are presented in this thesis, and its performance was not tested with enhanced speech signals. An article describing the variability reduction procedures and performance comparison when taking enhanced signals into account was submitted recently to the IEEE Transactions on Audio, Speech, and Language Processing [23].

1.4 Outline

This work is organized as follows. In this chapter, we introduced the problem of speech intelligibility assessment in CI users and described the work objectives and contributions. In Chapter 2 we discuss the state-of-the-art at the problem of objective speech intelligibility and quality assessment, the importance of temporal envelope cues for speech intelligibility, and how these cues can be measured by investigating a modulation spectral representation of the speech signal. In Chapter 3 we present the proposed refinements to the SRMR metric that aim to emulate CI hearing precepts and reduce its variability related to the speech content. We then compare the performance obtained with the proposed measure to several intrusive and non-intrusive state-of-the-art speech quality and intelligibility metrics, and discuss its advantages and shortcomings. We finally conclude the thesis with a summary of the results and some final considerations in Chapter 5.

Chapter 2

Speech Intelligibility Assessment

Speech intelligibility (SI) is a measure of how much of the information in a speech signal is perceived when the signal is subject to the effects of a transmission system. It has been traditionally measured by presenting words corrupted by a given environmental condition (such as noise or reverberation) for identification at several distortion levels [24]. Words are usually presented outside of a sentence context or within nonsense sentences, in order to avoid a bias coming from listener's predictions. Metrics for SI usually relate the relative amount of words recognized (in %) under different condition levels. Such procedure has some limitations, such as the already mentioned listener prediction bias: both signal characteristics and the subjective response are taken into account. It is possible to mitigate this issue by using multiple words/sentences per condition and, depending on the application of the measurement, multiple listeners under the same conditions.

French and Steinberg [25] summarize the factors that affect speech intelligibility as belonging to 4 different types:

1. intensity of the speech received by the ear at each frequency;

2. electrical/acoustical characteristics of the instruments (for example, a handset or hearing instrument) intervening between talker and listener;
3. conditions under which the communication takes place (e.g., background noise, reverberation time);
4. behavior of the talker and listener as modified by the characteristics of the communication system and by the conditions under which it is used (e.g., vocal effort and speech rate adjustment).

However, determining such factors separately requires complete knowledge of the speech transmission channel, talker, and listener characteristics. As this information is not always available, objective intelligibility metrics mainly consider how much perceptually important features of speech signals, such as spectral and temporal cues, are preserved in received speech. The following section focuses on one specific perceptual effect termed frequency masking.

2.1 Quantifying the effect of frequency masking on speech intelligibility: the articulation index

The perception of a sound may be affected by the presence of another sound in a phenomenon called auditory masking [4]. Two simultaneous sounds of different frequencies that are played at the same time may be perceived either as two separate sounds or a combined sound, due to the filtering that occurs in the cochlea. Cochlear processing can be modelled as a filterbank consisting of bandpass filters whose frequency responses correspond to the tuning curves of auditory neurons. Each of these bands is called a critical band. Whenever two competing sounds have a difference in frequencies small enough to be in the same critical band, the sound with the lower energy is masked by the higher energy sound; this is called frequency masking. Background noise is known to mask lower energy components of phonemes, causing confusions in determining primarily place of articulation (but also voicing and stop-fricative confusion).

To take into account the effect of frequency masking in speech intelligibility, French and Steinberg [25] proposed the use of the so-called Articulation Index (AI) as an objective metric of the effective proportion of the normal speech signal information available for a listener. The AI is based mainly on spectral cues and is computed by measuring the SNRs between an idealized speech spectrum and the masker spectrum of the noise in octave or one-third-octave spectrum bands (central frequencies between 270 – 5600 Hz). These SNR levels are then multiplied by weighting factors that reflect the relative importance of each band to speech intelligibility and averaged, as follows:

$$\mathrm{AI} = \frac{1}{N}\sum_{j=1}^{N} w_j \frac{\min(\mathrm{SNR}(j), 30)}{30} \tag{2.1}$$

where N is the number of bands and w_j, $j = 1 \ldots N$ are the band weights. The SNR values for each band are limited to a maximum of 30 dB. The resulting index can be mapped to various measures of speech intelligibility, such as the percentage of syllables, words, or sentences understood [26].

A limitation of the AI is that it considers that distortions are either additive noise or signal attenuation and can be computed separately for each band. This approach ignores the effect of convolutional distortions, such as reverberation. While corrections can be applied to the AI score depending on whether the noise is steady-state or not, reverberation time, and whether visual cues are present or not, these corrections are very simple. As an example, the effect of reverberation is taken into account by subtracting a value that depends solely on the reverberation time from the AI score [26].

2.2 Importance of temporal envelope cues for speech intelligibility

Any real signal $x(t)$ may be expressed into an analytic representation, which is a complex representation without negative frequency components. Such representation is given by

$$x_a(t) = x(t) + jH[x(t)] \tag{2.2}$$

where H is called the Hilbert transform and is defined as

$$H[x(t)] = \text{p.v.} \int_{-\infty}^{\infty} \frac{s(t-\tau)}{\pi\tau}\, d\tau \tag{2.3}$$

where *p.v.* is the Cauchy principle value of the integral [27]. The usefulness of this representation is given by the fact that the magnitude of the analytic signal is the envelope of $x(t)$, while its argument corresponds to the instantaneous phase of $x(t)$. Figure 2.1 shows an example of this decomposition. There is evidence that amplitude modulations represent an important feature of communication sounds for animals and humans: in [28], the authors show that regions in the auditory cortex of cats were tuned to specific modulation frequencies. Therefore, an audio signal representation that explicitly separates the signal into its envelope and fine structure gives us access to this feature, as emphasized by Plomp [29].

2.2.1 Perceptual studies on the effect of temporal envelope filtering to speech intelligibility

Drullman et al. [30] investigated the effect of envelope smearing on speech reception of sentences in noise and on phoneme identification in NH listeners. They decomposed the speech signal into a series of frequency bands and low-pass filtered the amplitude envelope of each subband at cutoff frequencies of 0, 0.5, 2, 4,

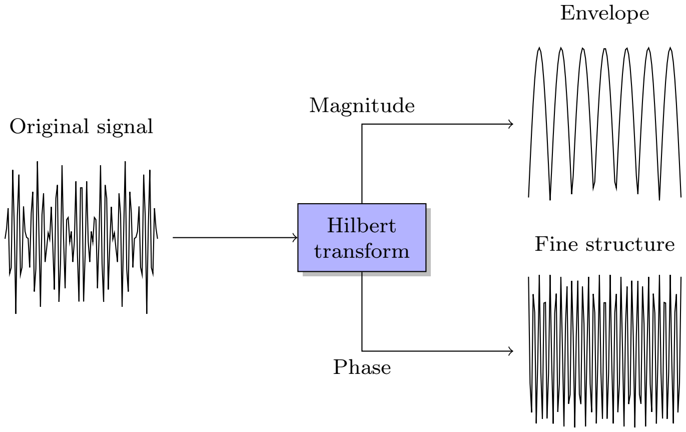


FIGURE 2.1: Decomposition of a signal into its temporal envelope and fine structure.

8, 16, 32, and 64 Hz. Their results show that filtering the amplitude envelopes at low cutoff frequencies (0 to 2 Hz) results in a severe reduction in sentence intelligibility. They also show that modulation frequencies above 16 Hz have a marginal contribution to sentence intelligibility. Temporal envelope smearing affected consonants, especially stops, more significantly than vowels. The effect of reducing slow temporal modulations was studied in [31]. Temporal envelopes were high-pass filtered at cutoff frequencies of 1, 2, 4, 8, 16, 32, 64, or 128 Hz, or completely flattened. In this case, results showed that sentence intelligibility was only affected with narrow-band processing for cutoff frequencies above 64 Hz, and no reduction of sentence intelligibility when only amplitude variations below 4 Hz were reduced. Reducing slow temporal modulations also affected vowel and consonant recognition. Similar results were shown in [32], where the authors performed perceptual experiments with Japanese syllables with filtered time trajectories of spectral envelopes. They showed that speech intelligibility is not significantly impaired as long as the filtered spectral components have a rate of change faster than 1 Hz when high-pass filtered, slower than 24 Hz when low-pass filtered, and between 1 and 16 Hz when band-pass filtered. They suppose changes outside this range are due to nonlinguistic aspects of the speech signal and can be suppressed without

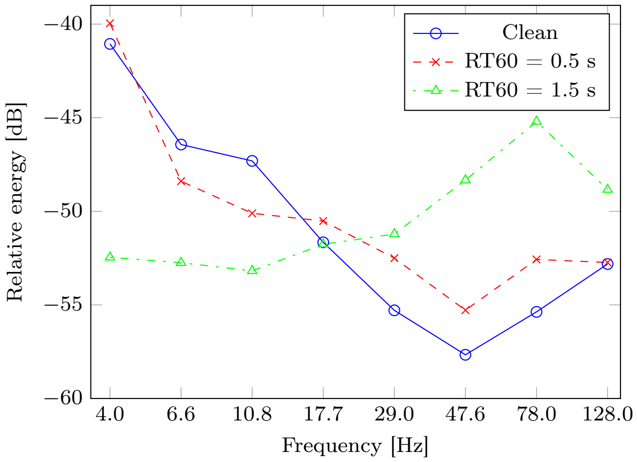


FIGURE 2.2: Modulation spectra at the 1105 Hz band for clean speech (◯) and reverberant speech with RT60 = 0.5 s (×) and 1.5 s (△).

impairing intelligibility. In [33], the effects of fine structure cues for speech intelligibility were investigated. Results show that speech signals with intact amplitude envelopes and fine structure replaced with noise retained perfect intelligibility, while signals with intact fine structure and random temporal envelopes had an average intelligibility score of 17%.

These results show the importance of preserving slow temporal envelope modulation for speech intelligibility. Temporal envelopes from clean speech contain frequencies ranging from 2-20 Hz with spectral peaks at approximately 4 Hz, corresponding to the syllabic rate of spoken speech [32]. Environmental distortions such as reverberation cause envelope smearing, as the addition of late reflections increases the amount of energy at higher modulation frequencies. This effect can be seen in Figure 2.2, which shows the average modulation spectrum of the 1105 Hz band temporal envelope for clean speech and reverberant speech with reverberation times equal to 0.5 s and 1.5 s. The energies were computed by using the modulation filterbank described later in this chapter (Section 2.3.2.2).

2.2.2 The modulation transfer function and the speech intelligibility index

Houtgast and Steeneken [34] proposed that information about modulation frequencies in the temporal envelope of speech should be analyzed as a function of frequency for octave bands of speech. The output of each frequency band is then squared and low-pass filtered (with a cutoff frequency of 30 Hz) to obtain the intensity envelope for that band. This envelope is then analyzed with a set of one-third octave band-pass filters. By multiplying each band-filtered output by $\sqrt{2}$ and dividing by the long-term average value of the intensity envelope, the so-called modulation indices for speech components present in each acoustic octave band are found. These indices are then used to express how the modulations present in the temporal envelope are preserved by means of the temporal modulation transfer function (TMTF). In enclosures, the transfer of speech signals is affected by both the direct transmission path and all its reflections coming from the wall. These reflections will have different times of arrival at the listener, which results in attenuation of fast modulations. Ambient noise also plays a role in reducing modulation indices, but independently from modulation frequency. The modulation transfer function when we have a listener at a large distance from both the speaker and the noise source in a room with a diffuse indirect sound field is given by:

$$m(F) = \frac{1}{\left(1 + \frac{I_N}{I_O}\right)\sqrt{1 + 0.207F^2T^2}} \tag{2.4}$$

where $m(F)$ is the modulation index for the modulation frequency F (in Hz), T is the reverberation time in seconds, I_N is the intensity of the noise at the listener's position and I_O the mean intensity of the speaker's voice at the listener's position. The modulation indices can then be used to find the so-called speech transmission index (STI), which is related to the effect of a transmission system on speech intelligibility. A full MTF analysis is computed by using seven octave bands (center frequencies from 125 Hz to 8 kHz), with 14 modulation indices computed for each band (1/3-octave analysis from 0.63 – 12.5 Hz). Each modulation index m is converted into a corresponding apparent signal-to-noise ratio (SNR), termed $\mathrm{SNR}^{\mathrm{app}}$ by the following relationship:

$$\mathrm{SNR}^{\mathrm{app}} = 10 \log[m/(1-m)][dB] \tag{2.5}$$

These values are limited to a 30 dB range by truncating values greater than +15 dB or smaller than -15 dB with the following function:

$$\mathrm{clamp}(x, a, b) = \begin{cases} a & \text{if } x \leq a \\ x & \text{if } a < x \leq b \\ b & \text{if } x > b \end{cases} \tag{2.6}$$

where x is the $\mathrm{SNR}^{\mathrm{app}}$ value, $a = -15\mathrm{dB}$ and $b = +15\mathrm{dB}$. The apparent SNR values derived from each octave-band specific MTF are then averaged, resulting in ${\mathrm{SNR}^{\mathrm{app}}}_k$ $(k = 1, \ldots, 7)$ and an overall mean is computed taking into account octave-band specific weighting factors w_k:

$$\overline{\mathrm{SNR}^{\mathrm{app}}} = \sum_{k=1}^{7} w_k \, {\mathrm{SNR}^{\mathrm{app}}}_k \tag{2.7}$$

where $w_k = [0.13, 0.14, 0.11, 0.12, 0.19, 0.17, 0.14]$. Finally, the overall mean is converted into the STI, which reflects an apparent signal-to-noise ratio averaged over the acoustic and modulation frequencies:

$$\mathrm{STI} = \frac{\overline{\mathrm{SNR}^{\mathrm{app}}} + 15}{30} \tag{2.8}$$

It should be noted that this calculation requires complete knowledge of the environmental distortions (noise and reverberation) in order to compute its MTF. For this reason, its measurement in real rooms usually employs known artificial signals. Some extensions have been proposed in [31] to allow using a clean reference speech signal and its distorted version to measure speech intelligibility in an intrusive manner.

2.3 Objective speech intelligibility/quality assessment methods

Perceived speech can be rated for its perceived overall quality or its intelligibility [35]. While speech quality and intelligibility are related, quality is a more subjective impression of how "good" speech sounds (regarding the presence of artifacts, distortions, etc.), and is usually rated with mean opinion scores (MOS). Speech intelligibility, on the other hand, is a measure of the accuracy with which the signal under test carries its information to the listener. Objective measures may be more correlated with one or the other (or both), depending on the relation between the features used by the metric and the distortions applied to the signals under test.

As mentioned in Chapter 1, intrusive metrics make use of a clean reference signal and compute the amount of distortion in the target signal. Speech intelligibility/quality ratings are then calculated based on a distance metric between the clean and distorted signals. Non-intrusive metrics, on the other hand, can be used when only the corrupted signal is available.

In this section, we present the state-of-the-art in objective quality and intelligibility assessment. Some of the metrics presented here, namely NCM, ModA, ANIQUE+, and SRMR, directly make use of temporal envelope information, while the others use mainly spectral information. Emphasis is given to SRMR, as it is the basis of the metric proposed in this work.

2.3.1 Intrusive metrics

2.3.1.1 Normalized Covariance Metric

The normalized covariance metric (NCM) measure estimates speech intelligibility based on the covariance between the envelopes of the clean and degraded speech signals [36–38]. Computation of NCM values depends on deriving speech temporal envelopes, via a Hilbert transform, for each of the 23 gammatone filterbank channels, which are used to emulate cochlear processing. The normalized correlation between the clean and degraded speech envelopes produces an estimate of the so-called apparent SNR given by:

$$\mathrm{SNR}_{\mathrm{app}}(k) = \left[10\ log_{10}\left(\frac{r_k^2}{1-r_k^2}\right)\right]_{[-15,15]} \tag{2.9}$$

where r_k is the correlation coefficient between the clean and degraded speech envelopes estimated in filterbank channel k, and the $[\,]_{-15,15}$ operator refers to process of limiting and mapping $\mathrm{SNR}_{\mathrm{app}}$ into the $[-15, 15]$ range. The last step consists of linearly mapping the apparent SNR to the $[0, 1]$ range using the following rule:

$$\mathrm{SNR}^{\mathrm{NCM}}_{\mathrm{final}}(k) = \frac{\max(\min(\mathrm{SNR}_{\mathrm{app}}(k), +15), -15) + 15}{30} \tag{2.10}$$

The SNR^{NCM}_{final} values are then weighted in each frequency channel according to the AI weights $W(k)$. The final NCM value is given by:

$$\mathrm{NCM} = \frac{\sum\limits_{k=1}^{K=23} W(k) \cdot \mathrm{SNR}^{\mathrm{NCM}}_{\mathrm{final}}(k)}{\sum\limits_{k=1}^{k=23} W(k)} \tag{2.11}$$

2.3.1.2 Coherence-based Speech Intelligibility Index

The coherence-based speech intelligibility index (CSII) is a spectral-based speech intelligibility measure which takes into account the coherence (e.g. similarity) of the spectral coefficients for both the degraded and clean speech signals [39, 40]. In order to compute CSII values, a short-time Fourier transform is first performed such that each time-frequency segment can be weighted by a parameter called the magnitude squared coherence (MSC). The MSC is computed between the clean and processed signals as:

$$\mathrm{MSC}(f) = \frac{P_{cr}(f)^2}{P_{cc}(f) \cdot P_{rr}(f)} \tag{2.12}$$

where f indexes a particular frequency bin, $P_{cr}(f)$ is the cross spectral density estimated between the clean (c) and the degraded speech signal (r), while $P_{cc}(f)$ and $P_{rr}(f)$ are the power spectral densities of the clean and degraded signals, respectively. The MSC values are commonly grouped into 25 frequency bands using critical pass-band filters $G(k)$ described by [41]. The MSC values are then used to estimate the channel-dependent SNR given by:

$$\mathrm{SNR}^{\mathrm{CSII}}_{\mathrm{final}}(k) = \frac{1}{N}\sum_{n=1}^{N}\left[10log_{10}\frac{\sum\limits_{k=1}^{K=25} G(k)\cdot \mathrm{MSC}(f)\cdot R(n,f)^2}{\sum\limits_{k=1}^{K=25} G(k)\cdot (1-\mathrm{MSC}(f))\cdot R(n,f)^2}\right]_{[0,1]} \tag{2.13}$$

where $R(n, f)$ is the spectrum of the degraded speech signal estimated at time frame n using a sliding Hanning window of length 30 ms (25% overlap); N indicates the total number of frames within a particular sentence. The $[\,]_{[0,1]}$ operator refers to $[-15, 15]$ dB clipping and $[0, 1]$ linear mapping. Lastly, per-band $\mathrm{SNR}^{\mathrm{CSII}}_{\mathrm{final}}$ values are weight-averaged using AI weights to form the CSII measure:

$$\mathrm{CSII} = \frac{\sum_{k=1}^{K=23} W(k) \cdot \mathrm{SNR}_{\mathrm{final}}^{\mathrm{CSII}}(k)}{\sum_{k=1}^{K=23} W(k)} \tag{2.14}$$

2.3.1.3 PESQ and oPESQ

Perceptual evaluation of speech quality (PESQ) is the International Telecommunications Union (ITU-T) P.862 Recommendation for speech quality assessment of narrow-band speech [42] with more recent developments allowing for wide-band speech to also be assessed. The algorithm is based on a sensory model that aggregates two distortion-related factors: a disturbance value (D_{ind}) and an average asymmetrical disturbance value (A_{ind}). These factors are estimated through a comparison of the clean and processed signals, both mapped to a psychoacoustically-relevant domain. The final quality rating is then given by a linear mapping with coefficients optimized using conventional telephony data (e.g., voice over Internet protocol, wireless):

$$PESQ = a_0 + a_1 \cdot D_{ind} + a_2 \cdot A_{ind}, \tag{2.15}$$

$$\text{where} \begin{cases} a_0 = 4.5 \\ a_1 = -0.1 \\ a_2 = -0.0309 \end{cases} \tag{2.16}$$

The original PESQ parameters a_0, a_1 and a_2 were obtained using speech signals representative of conventional telephony applications and did not involve reverberation-related distortions. In [11], these parameters were further optimized for reverberant speech using multiple linear regression analysis and NH-listener subjective data. The “reverberation-optimized PESQ" metric is also explored in this study and is termed oPESQ. The optimized parameters are given below:

$$\begin{cases} a_0 = 4.6 \\ a_1 = -0.5678 \\ a_2 = 0.1024 \end{cases} \tag{2.17}$$

2.3.1.4 Kullback-Leibler Divergence

The Kullback-Leibler Divergence (KLD) estimates the distance between the probability distribution functions (*pdf*) of the clean and distorted speech signals and was shown to be a reliable objective quality metric for reverberant speech [11]. The motivation behind the metric lies in the fact that the spectral and temporal smearing produced by the reverberation causes the *pdf* of reverberant speech (p_R) to be flatter than that of clean vocoded speech (p_C). The KLD is a non-negative measure which characterizes distribution similarity with values tending to zero when distributions are similar (and equals zero when $p_C = p_R$). It is given by the following integral (over the time variable t):

$$KLD = -\int p_C(t) \cdot log_{10} \frac{p_C(t)}{p_R(t)} dt \tag{2.18}$$

2.3.1.5 Frequency-Weighted Segmental Speech-to-Reverberation Ratio

The Frequency-Weighted Segmental Speech-to-Reverberation Ratio (FWSSRR) measure is obtained through estimates of the SNRs for each critical band on each time frame. The AI weighting function is then used to obtain the frequency weights for each critical band. In this study, FWSSRR was computed as:

$$FWSSRR = \frac{10}{N}\sum_{n=1}^{N}\frac{\sum\limits_{k=1}^{K=25} W(k)\cdot\log_{10}\frac{|C(n,k)|^2}{|C(n,k)-R(n,k)|^2}}{\sum\limits_{k=1}^{K=35} W(n,k)} \tag{2.19}$$

where $C(n,k)$ and $R(n,k)$ are the clean and reverberant/noisy speech signals, respectively, at time frame n and critical frequency k; $K = 25$ is the total number of critical bands, N is the number of time frames and $W(k)$ is the AI weighting function. More details about the FWSSRR measure can be found in [40].

2.3.2 Non-intrusive metrics

2.3.2.1 ModA

The so-called modulation-spectrum area (ModA) [43] measure is based on the principle that the speech signal envelope is smeared by the late reflections in a reverberant room, thus affecting the modulation spectrum of the speech signal. In order to obtain the ModA metric, the signal is first decomposed into $N(=4)$ acoustic bands (lower cutoff frequencies of 300, 775, 1375, and 3676 Hz, as in [43]); the temporal envelopes for each acoustic band are then computed using the Hilbert transform, then downsampled and grouped using a 1/3-octave filterbank with center frequencies ranging between 0.5 and 8 Hz. As in [43], 13 modulation filters are used to cover the 0.5 - 10 Hz modulation frequency range. For each acoustic frequency band, the so-called “area under the modulation spectrum" is computed (A_i) and finally averaged over all $N(=4)$ acoustic bands to obtain the ModA measure:

$$\text{ModA} = \frac{1}{N}\sum_{i=1}^{N} A_i. \tag{2.20}$$

2.3.2.2 SRMR

SRMR is a recently-proposed non-intrusive metric developed originally for reverberant and dereverberated speech and evaluated against subjective normal hearing listener data [20, 44]. Recently, promising results were also reported when evaluated against vocoded speech simulating CI hearing [45]. Computation of the SRMR metric is performed in four stages, which are depicted in Figure 2.3:

Acoustic filterbank The input signal $x(n)$ is filtered by a 23-channel gammatone filterbank which emulates cochlear processing. Filter center frequencies range from 125 Hz to approximately 8 kHz (i.e., half the sampling frequency) with bandwidths characterized by the equivalent rectangular bandwidth, ERB [46]. The outputs of this filterbank are $x_j(n), j = 1, \ldots, 23$. This stage is further detailed in Section 3.1.

Temporal envelope, windowing, and DFT Temporal envelopes $e_j(n)$ are computed for each of the $j = 1, \ldots, 23$ filterbank output signals $\hat{x}_j(n)$ using the Hilbert transform. Temporal envelopes are then windowed (256 ms frames, 64 ms frame-shifts) to create $e_j(m, n)$ (where m refers to the frame index). A discrete Fourier transform $\mathcal{F}$ is applied to obtain the so-called modulation spectral energy for each critical band $E_j(m, f) = |\mathcal{F}(e_j(m, n)^2)|$, where f indexes the modulation frequency bins.

Modulation filterbank This filterbank emulates frequency selectivity in the modulation domain [47]; this is obtained by grouping the modulation frequency bins into $k = 1, \ldots, 8$ overlapping modulation bands with centre frequencies logarithmically spaced between 4-128 Hz. The outputs of this stage are the energies for each acoustic channel j for each modulation frequency band $k = 1, \ldots, 8$ and each frame m, $E_j(m, k)$.

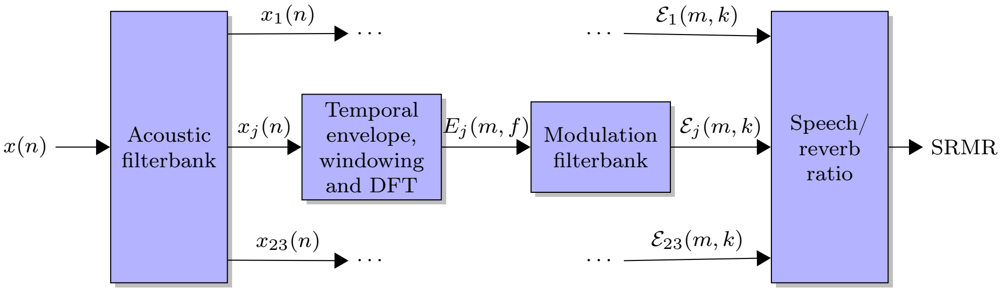


FIGURE 2.3: SRMR block diagram.

Ratio computation The SRMR value is computed as the ratio of the average modulation energy content available in the first four modulation bands ($k = 1-4$, circa 3-20 Hz, consistent with clean speech modulation content [32]) to the average modulation energy content available in the last four modulation bands ($k = 5-8$, circa 20-160 Hz).

In [20], an adaptive version of SRMR was proposed, where the ratio is computed by the following expression:

$$\mathrm{SRMR} = \frac{\sum_{k=1}^{4} \bar{\mathcal{E}}_k}{\sum_{k=5}^{K^*} \bar{\mathcal{E}}_k} \tag{2.21}$$

where $\bar{\mathcal{E}}_k$ is the average per-modulation band energy (over all frames and acoustic bands), and K^* is computed in an adaptive manner for each speech signal. Modulation frequency content in a given acoustic frequency is upper-bounded by the bandwidth of the gammatone filter that analyzed such band. K^* is chosen such that the modulation frequency bands considered for the SRMR computation have modulation frequencies up to the bandwidth of the lowest frequency gammatone filterbank for which 90% of the total modulation energy is accounted for.

2.3.2.3 ANIQUE+

ANIQUE+ [48] is a perceptual model that employs statistical learning to compute speech quality scores based on three different distortion measurement modules, namely mute distortion, non-speech, and articulation distortion modules. The mute distortion module detects unnatural mutes in the speech signals and quantifies their effects on speech quality. The non-speech module, in turn, detects and quantifies the effects of annoying non-speech activities, such as those resultant from inserting erroneous bits into a speech decoder. The articulation distortion module uses modulation spectral concepts similar to those used in both SRMR and ModA. For each critical band, ANIQUE+ computes the normalized articulation energy (average modulation energy between 2–30 Hz modulation frequencies), normalized non-articulation energy (average modulation energy for frequencies greater than 30 Hz), and the energy across the critical band. The values for these three energies are computed for all the critical bands and mapped to a frame distortion score by means of a multilayer perceptron. The frame distortion scores are aggregated, separately over active and inactive frames. The outputs from the three distortion modules are finally linearly combined to produce an overall quality score.

2.3.2.4 P.563

In 2004, ITU-T standardized the first non-intrusive speech *quality* metric for telephone-band speech applications [49, 50]. The standard algorithm estimates the quality of the tested speech signal based on three principles. First, vocal tract and linear prediction analysis is performed to detect unnaturalness in the speech signal. Second, a pseudo-reference signal is reconstructed by modifying the computed linear prediction coefficients to the vocal tract model of a typical human speaker. The pseudo-reference signal serves as input, along with the degraded speech signal, to an intrusive algorithm (similar to ITU-T P.862 [51]) to generate a basic voice quality index. Lastly, specific distortions such as noise, temporal clippings, and robotization effects (voice with metallic sounds) are characterized. The algorithm detects major distortion events in the speech signal and classifies them as belonging to one of six possible classes: high level of background noise, signal interruptions, signal-correlated noise, speech robotization, and unnatural male and female speech. Once a distortion class is found, class-specific internal parameters are mapped to an objective quality score. While P.563 was developed as an objective *quality* measure for normal hearing listeners and telephony applications, a recent study has shown promising results with P.563 as a correlate of noise-excited vocoded speech intelligibility for normal hearing listeners, but not tone-excited vocoders [45]. This could be due to the fact that P.563 has a robotization module which characterizes robotization effects, such as voice with metallic sounds. The P.563 algorithm is explored here as a correlate of speech intelligibility of CI users.

Chapter 3

SRMR metric tailored for CI users

In this chapter, we propose three modifications to the original SRMR metric. The first modification aims to emulate CI hearing precepts by replacing the acoustic filterbank by a CI-inspired filterbank. Two other modifications are proposed to reduce the impact of speaker and speech content on SRMR variability. We performed experiments using two different databases, one composed by consonant-vowel pairs and the other of complete sentences, both including samples from different speakers under clean anechoic condition, in order to understand how speech content factors affect the modulation spectrum representation. Based on the results, we propose the use of a modulation energy thresholding scheme and a narrower range of modulation filters to mitigate the variability related to speech content and fundamental frequency.

3.1 Using a CI-inspired acoustic filterbank

The acoustic filterbank used by SRMR is based on gammatone filters. The amplitude characteristic of this kind of filter has been shown to predict well human masking data. Their impulse response is given by [52]:

$$\mathrm{gt}(t) = at^{(n-1)} \exp(-2\pi bt) \cos(2\pi f_c t + \phi), \quad t > 0 \tag{3.1}$$

where a is the amplitude, b determines the duration of the impulse response (and, therefore, the bandwidth), n is the order of the filter, which determines the slope of the amplitude response skirts, and ϕ is the phase. The bandwidth of each channel and center frequency spacing are computed according to the equivalent rectangular bandwidth (ERB), which gives an approximation to the bandwidths of the human auditory filters [46]:

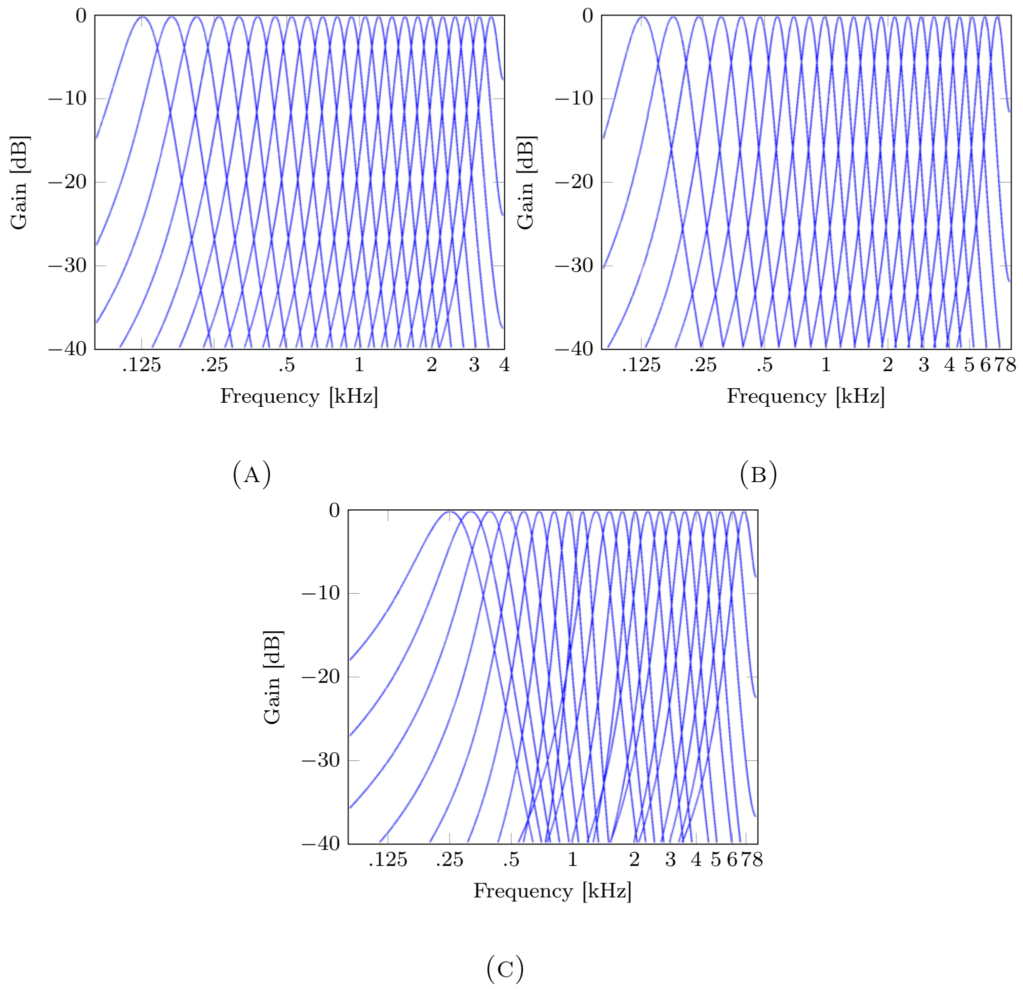
$$\mathrm{ERB}_j = \frac{f_j}{Q_{ear}} + B_{min} \tag{3.2}$$

where $Q_{ear} = 9.26449$ and $B_{min} = 24.7$ are constants corresponding to the Q factor and minimum bandwidth of human auditory filters. The center frequencies $f_j, j = 1, \ldots, 23$, in turn, are computed as:

$$f_j = -(Q_{ear} B_{min}) + e^{j \frac{-\log(f_{max} + Q_{ear} B_{min}) + \log(f_{min} + Q_{ear} B_{min})}{N}} (f_{max} + Q_{ear} B_{min}) \tag{3.3}$$

such that the ERB values are respected.

While this filterbank has been shown to perform well for the original SRMR metric, which was designed for NH listeners, it does not necessarily reflect CI listening characteristics. Since their cochlear processing is compromised, CI users rely on the speech processor for encoding audio information in subbands. Therefore, we updated the acoustic filterbank to simulate characteristics of CI devices. Our approach was to emulate the filterbank used by the ACE and CIS speech coding strategies, which uses a mel-like spacing between bands instead of the ERB-spacing [53]. The decision to base our filterbank on the one used by these strategies was that these were the strategies used by all participants in our listening test.



(A)

(B)

(C)

FIGURE 3.1: Comparison of acoustic filterbanks: (A) ERB-spaced filterbank used in the original SRMR metric (8 kHz), (B) ERB-spaced filterbank for a sampling frequency of 16 kHz sampling frequency, (C) the CI-inspired filterbank.

Different filterbanks based on the characteristics of other strategies and devices can be also used, but this is left for future work. Figure 3.1 shows a comparison between three different acoustic filterbanks:

(A) the filterbank used in the original SRMR metric (ERB-spaced, $f_{min} = 125$ Hz, $f_{max} = 4$ kHz, J = 23);

(B) the same as (A), but using $f_{max} = 8$ kHz, which is the value to be used for computing SRMR from files with a sampling rate of 16 kHz;

(C) the CI-inspired filterbank.

The main difference between the Nucleus filterbank and the one used by SRMR are the number of bands, the range covered by the filterbanks and the filter bandwidths. The filterbank used in the ACE and CIS strategies has 22 filters, its first center frequency at 250 Hz, and the frequency spacing follows a mel-like scale instead of the ERB. In our CI-inspired filterbank, bandpass filters are still computed using Equation 3.1. In order to differentiate between the SRMR metric computed using the original filterbank and the CI-inspired filterbank, the latter will be called SRMR-CI henceforth. In Chapter 4, we show that using SRMR-CI leads to improvements in correlations with speech intelligibility for CI users.

3.2 Detecting sources of variability in the modulation spectrum

Features that represent temporal information of speech, as the temporal envelope structure, have been shown to be useful for automatic speech recognition (ASR) systems. In [54], the authors show that the modulation spectrogram, i.e., a representation of the modulation spectrum in different acoustic sub-bands, can be used as the single feature set in an ASR system, leading to word error rates of 8.5% under clean condition. This result shows that the modulation spectrum is sensitive to speech content, such as different phonemes. SRMR and SRMR-CI, on the other hand, are based on longer temporal modulation energy integration than the modulation spectrogram proposed by Kingsbury et al. [54]: 256 ms windows against 8 ms windows zero-padded to 32 ms. SRMR and SRMR-CI use long frames and frame shifts of 64 ms in order to obtain appropriate resolution for low-frequency modulation frequencies around 4 Hz. The larger frames for SRMR mean that the modulation spectrum representation used by both metrics is less sensitive to speech content than the modulation spectrogram, as we are not interested in the actual speech content but solely on the effect of the signal operating environment on the spectrum. Notwithstanding, SRMR has been shown to have high inter-speaker variability [55, 56]. The experiments described herein also show that the SRMR-CI metric is sensitive to different phonemes and fundamental frequency (F_0).

3.2.1 Experiments with consonant-vowel pairs and sentences

To investigate the effect of speech variability on the SRMR-CI metric, we processed speech data from two different databases. The first database consisted of consonant-vowel pairs (CVs), and contained 1,728 samples [57]. Four talkers (2 males and 2 females) recorded 8 tokens for each of 18 consonants in 3 vowel contexts. Files were provided at a sampling rate of 16 kHz (16 bits per sample). Table 3.1 shows the list of vowels and consonants in the CVs categorized by manner of articulation.

Table 3.1: List of consonants and vowels used in the CV analysis

Vowels	Fricatives	Stops	Nasals	Affricates
ɑ i u	s z ʃ ʒ	p t k	m n	ʧ ʤ
	f v θ ð	b d g		

To evaluate SRMR-CI variability over sentences we used a subset of the TIMIT corpus consisting of 160 anechoic, noise-free sentence recordings from 8 male and 8 female native English speakers [58]. The sentences were distributed in PCM-encoded WAV format, single channel, 16 kHz sampling rate and 16 bit samples. Each speaker recorded 10 phonetically rich sentences. Of the 160 samples used, 130 were recordings of different sentences.

We computed the SRMR-CI metric for all the CV pairs and sentences. In the case of the CV pairs, the samples were grouped using two different categories: by vowel and by manner of articulation (see Table 3.1 for details on the categories). The results were summarized in the form of box plots. Figures 3.2 and 3.3 show

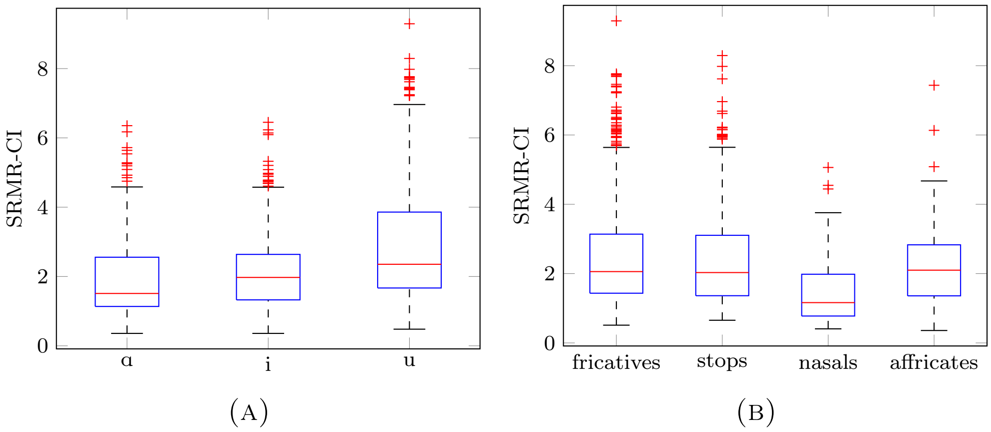


Figure 3.2: Boxplots for the original SRMR-CI scores of the CV pairs grouped by (A) vowel and (B) manner of articulation.

the results for the CV pairs and sentences, respectively. The boxes in the box-and-whiskers plots indicate the 25th, median, and 75th percentiles, the whiskers extend to approximately ±2.7 standard deviations, and all the points outside this range are considered outliers (shown as crosses). Results for the CV pairs (Figure 3.2) indicate that different SRMR means can be identified when we separated CV groups by vowel (one-way ANOVA for each pair of vowels, $p < 0.001$). As shown in Figure 3.2a, CV pairs with the vowel /ɑ/ have lower SRMR-CI means and standard deviation, followed by /i/ and /u/, with the latter showing the distribution with the largest spread. Results are not as consistent for all manner of articulation pairs (Figure 3.2b), but at least one of the groups (nasals) has different sample mean from all other groups (one-way ANOVA considering all four groups, $p < 0.001$). Intra-group relative standard deviation (RSD, standard deviation over mean) for all groups is between 47 and 61%. Considering all CV samples together (i.e., all CV pairs from the four talkers), the RSD found was 58%.

On the other hand, the SRMR-CI values for the samples on the sentence database had a RSD of 47%; its distribution is depicted by the leftmost boxplot in Figure 3.3. Additionally, we grouped sentences depending on the speaker's gender in order to assess the effects of pitch frequency on SRMR-CI variability. Figure 3.4 shows the SRMR-CI distributions for sentences by male ($F_0 = 151.4 \pm 29.7$ Hz) and female

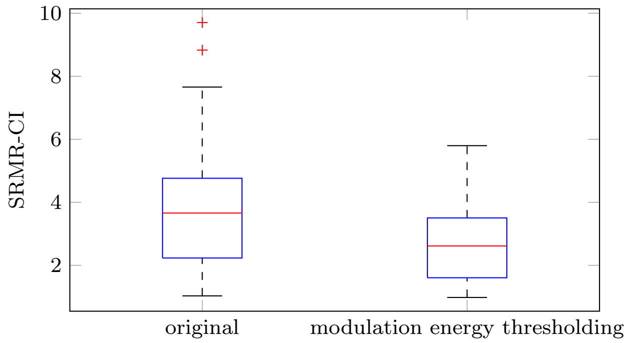


FIGURE 3.3: Boxplots for the clean sentences using the original SRMR-CI and the proposed energy thresholding scheme.

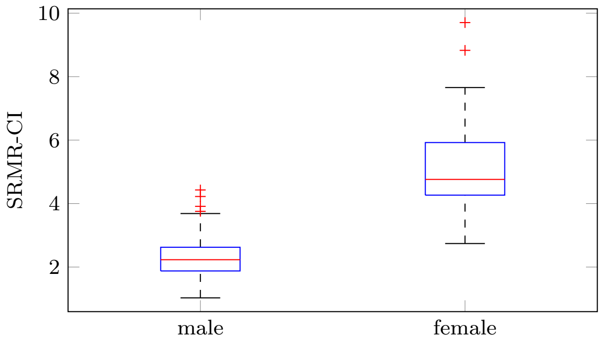


FIGURE 3.4: Boxplots for the SRMR-CI values for TIMIT sentences by male and female speakers.

($F_0 = 211.6 \pm 31.6$ Hz) speakers, where significant (ANOVA, p < 0.001) difference in average SRMR-CI scores was observed based on speaker gender. To confirm that this result was caused by a difference in pitch frequency, we computed the average F_0 for all sentences (using the PEFAC pitch tracker [59]) and calculated its correlation with the SRMR-CI value; a value of 0.76 was found, pointing to a strong relationship between pitch and the speech-to-reverberation energy ratios.

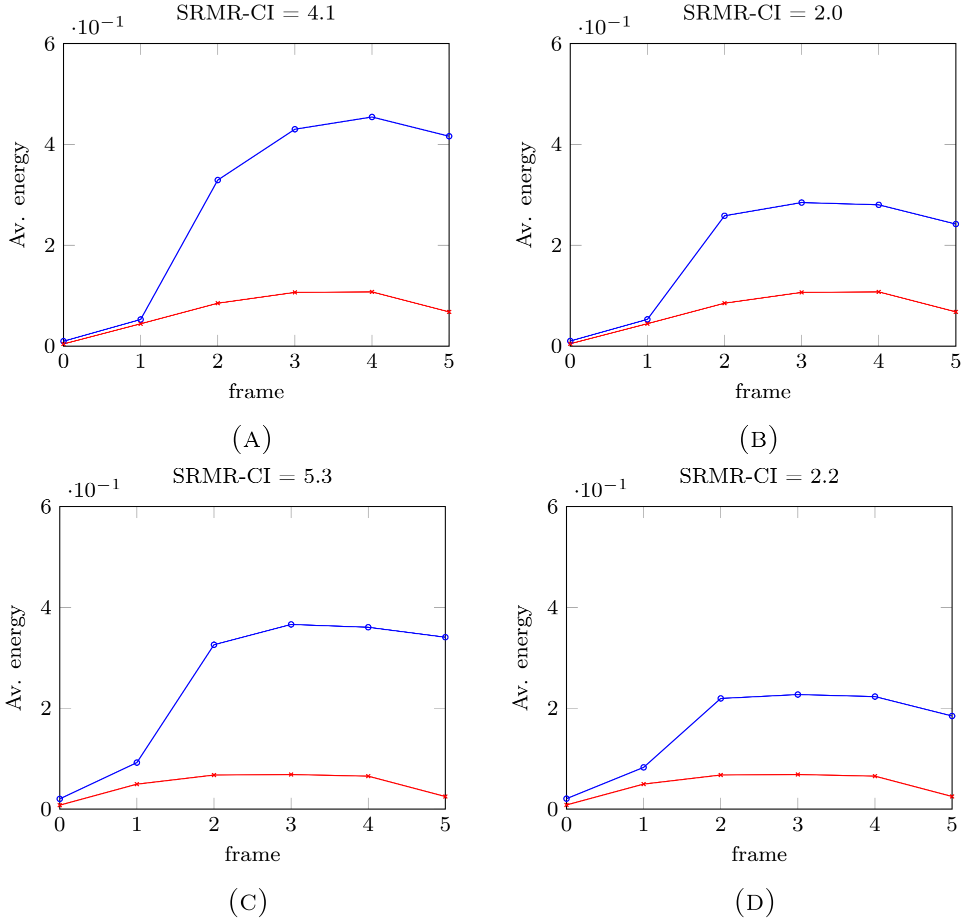
3.2.2 Reducing speech content effect by modulation spectral energy thresholding

As shown by the previously-described experiments, the modulation spectrum is sensitive to speech content, and this sensitivity results in a high variability of the SRMR-CI predictions. By inspecting the temporal evolution of the modulation energies for the CVs with larger variability, we noticed both inter- and intra-speaker variability. Figures 3.5a/3.5c and 3.6a/3.6c show the sum of the first four (marked with circles) and the last four (marked with crosses) modulation bands for the average of all 22 acoustic frequencies for each temporal frame for the same CV pair uttered by two different speakers and the same speaker, respectively. The computed SRMR-CI value (displayed in the title of each subplot) for two recordings of the same CV pair by different speakers can differ by a factor of 29%. For two instances of the same CV pair uttered by the same speaker (Figure 3.6) the difference in this example is 21%. When comparing different CV pairs, differences can be even higher, as it can be noted by looking at the SRMR-CI distributions in Figure 3.2.

In [54], the authors propose an energy thresholding method for the modulation spectrogram visualization. They use the peak value of the modulation spectrum as a reference and map all values more than 30 dB below this global peak to -30 dB. By doing this, the visualization focuses on the higher energies of the spectrogram while truncating extremely low values. As we are also interested in the relationship between energies in different parts of the spectrum and not on absolute values, we employed a similar procedure in order to reduce SRMR-CI variability. As we want to minimize the effect of phonemes that result in higher ratios than the average, we used the average peak instead of the global peak. The energy values for each of the acoustic and modulation frequencies in all the frames are limited to

$$E_j^{lim}(m,k) = \mathrm{clamp}(E_j(m,k), \bar{E}_{peak}, \frac{\bar{E}_{peak}}{1000}) \tag{3.4}$$

where $E_j^{lim}(m,k)$ is the range-limited modulation spectrum for the m-th frame at the j-th acoustic frequency band and the k-th modulation frequency band, and



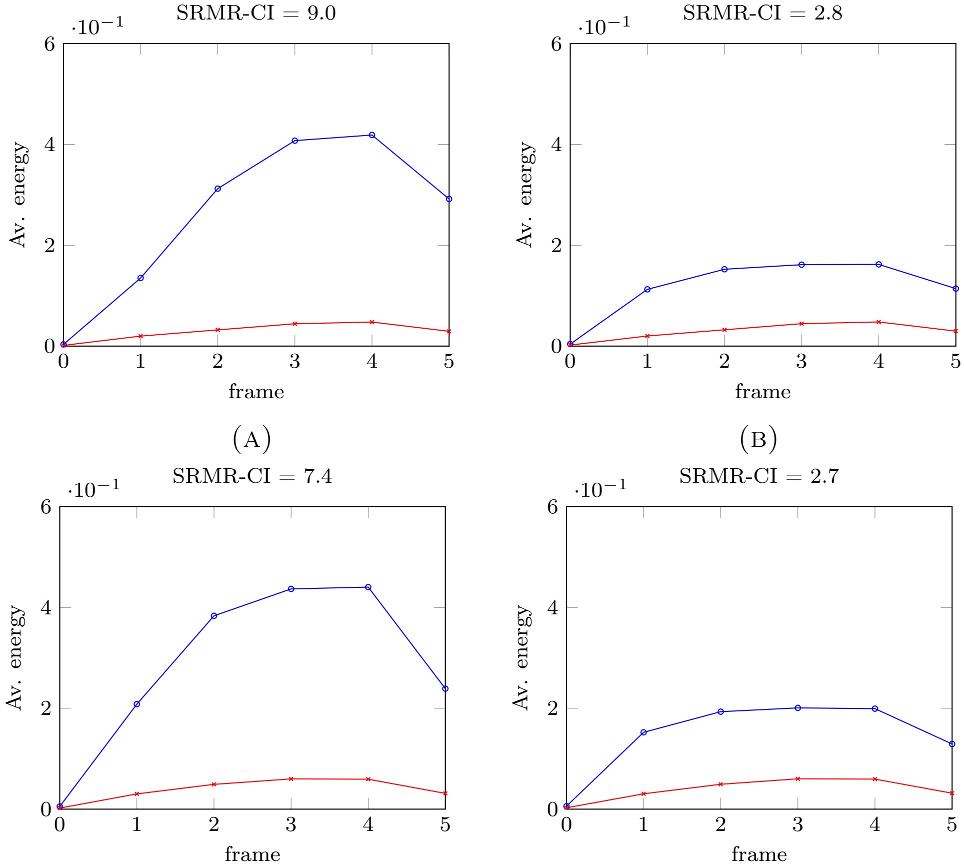
(A) (B) (C) (D)

FIGURE 3.5: Sum of the average modulation energies for the first four bands (○) and last four bands (×) for the same CV pair and two different speakers. (A) and (B) correspond to the original and limited energies for the first speaker, and (C) and (D) to the original and limited energies for the second speaker.

$$\bar{E}_{peak} = \max_{j,k}(\frac{1}{M}\sum_{m=1}^{M} E_j(m,k)) \tag{3.5}$$

is the maximum value of a single band in the long-term modulation spectra averaged over all acoustic frequencies (i.e., the largest value in the average modulation spectrum matrix). The clamp function is defined by Equation 2.6.

By using the thresholded values instead of the actual energy values to compute SRMR-CI, the variation between different speakers and different recordings of the same CV by the same speaker are reduced, as depicted in Figures 3.5b/3.5d and 3.6b/3.6d: 10% and 4%, respectively (recall that 29% and 21 % was seen with the



(A) (B) (C) (D)

FIGURE 3.6: Sum of the average modulation energies on the first four bands (○) and last four bands (×) for two different recordings of the same CV pair by a single speaker. (A) and (B) correspond to the original and limited energies for the first recording, and (C) and (D) to the original and limited energies for the second recording.

original implementation). The effects of modulation energy thresholding on the ratio distributions for the CV pairs grouped by vowel and manner of articulation are shown in Figure 3.7a and 3.7b, respectively. The RSDs for the different groups in this case were between 30% and 43% (against the 47% - 61% range that was found with the original implementation). Direct comparison between Figures 3.7 and 3.2 shows that absolute range and variability were reduced.

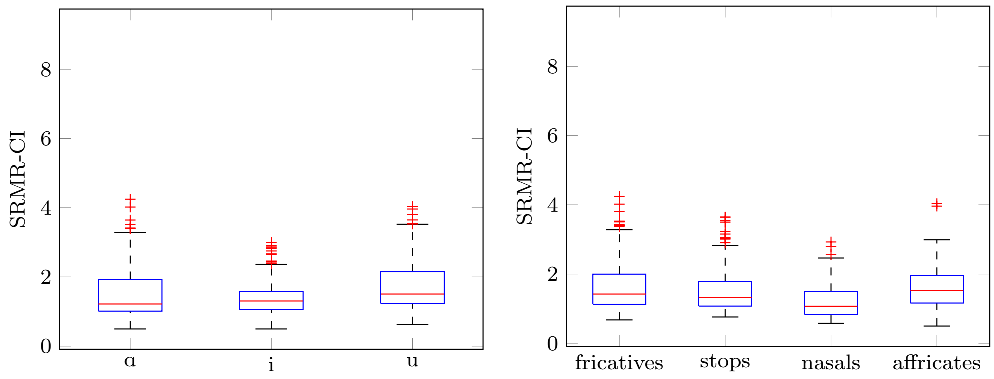


FIGURE 3.7: Boxplots for the CV pairs grouped by (A) vowel and (B) manner of articulation, using modulation energy thresholding.

3.2.3 Reducing pitch effect by reducing the range of analyzed modulation frequencies

The original SRMR metric is based on modulation energy bands with logarithmically spaced center frequencies from 4 Hz to 128 Hz. These bands are extracted from the acoustic subband envelopes by a filterbank such as the one shown in Figure 3.8, where the values $CF1$-$CF8$ are the center frequencies for each of the 8 bands. The 4 – 128 Hz range was shown useful for detecting the effects of reverberation on the speech signal [20]. As shown in our experiments with clean sentences, however, the energy levels in this range are highly correlated with the fundamental frequency of speech. In fact, the energy envelope of the speech signal has been shown to have structure with periodicity equal to its fundamental frequency [60].

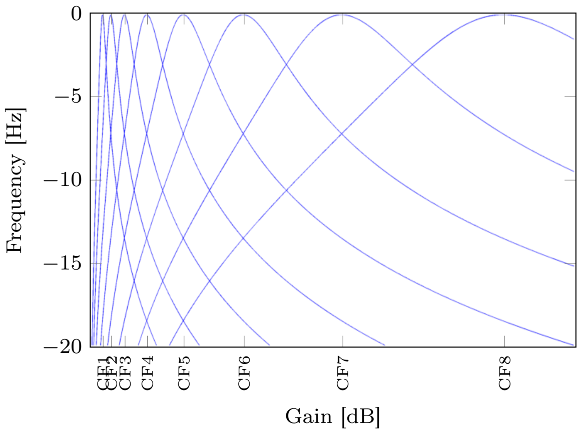


FIGURE 3.8: Modulation filterbank frequency response.

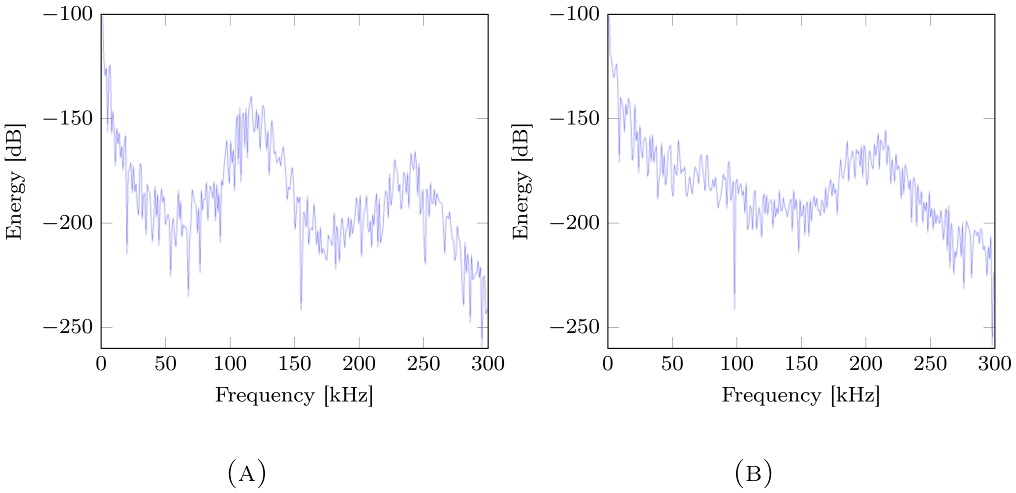


Figure 3.9: Periodograms for the envelope of the first acoustic subband from sentences uttered by (A) a male and (B) a female speaker.

While the results in [60] are based on "full-band" envelopes, we explore here if this behaviour is also observed in the acoustic subbands. To this end, we estimated the spectral density of the subband envelopes (using the CI-inspired acoustic filterbank) by computing their periodograms for each of the recordings in our sentence database. The envelopes were downsampled to a sampling rate of 600 Hz in order to be able to detect the effect of pitch up to 300 Hz. Then, for each periodogram, we computed the frequency corresponding to the first energy peak after DC. The correlation between this frequency and the average pitch was greater than 0.8 for 13 of the 22 subbands, and greater than 0.6 for 18 of the 22 subbands. The RMSE between the subbands' envelope spectral peak frequencies and the average fundamental frequency for the full-band speech signal was between 25 and 76 Hz. Figure 3.9 shows the periodograms for the envelope of the first band of the CI-inspired filterbank ($f_c = 250$ Hz, bandwidth $= 100$ Hz) for a sentence uttered by a male (A) and a female (B). The sentence uttered by the male speaker had a mean pitch of 147 Hz, and the spectrum of its envelope has a peak around 120 Hz (and another peak, probably from the first harmonic, at approximately 240 Hz). The sentence uttered by the female, on the other hand, had a mean pitch of 223 Hz, and the spectrum of its envelope shows a peak around 200 Hz.

Table 3.2: Correlations between individual modulation band energies and F_0, and SRMR-CI and F_0

Band	Center frequency of last channel ($CF8$) [Hz]						
	30	40	45	50	55	64	128
1	0.16	0.16	0.16	0.16	0.16	0.16	0.19
2	0.19	0.19	0.19	0.20	0.20	0.20	0.23
3	0.24	0.26	0.26	0.27	0.27	0.28	0.27
4	0.28	0.30	0.30	0.31	0.31	0.32	0.26
5	0.31	0.32	0.32	0.32	0.32	0.31	0.17
6	0.32	0.30	0.28	0.26	0.24	0.18	-0.23
7	0.28	0.18	0.11	0.02	-0.07	-0.22	-0.70
8	0.16	-0.14	-0.27	-0.38	-0.46	-0.57	-0.87
SRMR-CI	-0.12	-0.05	-0.004	0.05	0.12	0.24	0.76

These results show that for the majority of the acoustic subband envelopes, energy density is higher around the fundamental frequency of the original speech signal. This explains why clean speech with lower pitch (e.g., speech from males compared to females) resulted in lower SRMR-CI values. SRMR analyzes a modulation spectral range up to 128 Hz, which will be affected by modulation spectral peaks caused mainly by the fundamental frequency of the speech signal.

In order to avoid this side effect, we experimented using narrower modulation frequency ranges. We tested the following values for the central frequency of the last analysis filterbank ($CF8$): 30, 40, 45, 50, 55, and 64 Hz (half of the original center frequency for the last modulation band). All other center frequencies were also adjusted such that the logarithmic spacing was kept as in the original filterbank. Table 3.2 shows the correlations between the average fundamental frequency of the sentences and the energy level at the 8 modulation filterbank bands. The results for the original bandwidth (4 – 128 Hz) are also reported. It can be noticed that for $CF8 \geq 40$, the last modulation band has negative correlation with F_0. This inverse relationship extends to the sixth and seventh bands as the range gets broader. The correlation between SRMR-CI and F_0 is shown at the bottom line. As modulation frequency bandwidth is reduced, the correlation between pitch and

SRMR-CI decreases and has its trend reverted. From these results, we see that using modulation frequency ranges narrower than 4 – 55 Hz is useful for decorrelating F_0 and SRMR-CI. In the next chapter, we will evaluate the effects of the proposed normalization schemes on speech intelligibility prediction.

3.3 Normalized intrusive SRMR-CI metric

For comparison with other intrusive metrics, we also propose a simple intrusive metric based on SRMR-CI. This metric is computed by using the SRMR-CI value of the clean original signal for normalization, as follows:

$$\text{SRMR-CI}_{\text{intr}} = \frac{\text{SRMR-CI}}{\text{SRMR-CI}_{\text{clean}}}, \tag{3.6}$$

where $\text{SRMR-CI}_{\text{clean}}$ is the SRMR-CI value for the file's clean speech counterpart. While the advantage of using SRMR-CI lies in its non-intrusive characteristic, we wanted to investigate whether the clean file score serves as a normalization factor in order to reduce variability and how it compares to using the strategies proposed previously in this chapter.

Chapter 4

Experimental Results

In this chapter, we present a performance comparison between the different SRMR-CI implementations and the different benchmark metrics presented in Chapter 2. We compared the objective metrics with speech intelligibility scores obtained from a subjective listening test performed with CI users at the University of Texas at Dallas. Here, we describe briefly the data collection procedure and the distortion conditions that were tested; the interested reader is referred to [9, 61] for more details on the subjective test. The experiments have two objectives: first, we wanted to evaluate whether the proposed metric, SRMR-CI, leads to improved correlation with subjective CI intelligibility when compared to the original SRMR metric, and which modulation filterbank range leads to optimal results. Second, we wanted to compare the performance of the proposed metric with that of existing state-of-the-art speech quality and intelligibility metrics.

4.1 Experimental setup

4.1.1 Participants

Eleven adult CI users were recruited to participate in the subjective intelligibility experiments. The participants were all native speakers of American English with post-lingual deafness and had an average age of 64 years (±8.9). Participants consented and were paid for their participation. The interested reader is referred to reference [9] for specific demographic details of the participants. All participants had a minimum one-year experience using their device routinely, with the majority being bilaterally implanted for over 6 years. Three of the eleven participants used an 'ESpirit 3G' device, six a 'Freedom' device, and two a 'Nucleus 5' device; all devices are developed by Cochlear Ltd. For consistency, all participants were temporarily fitted with a SPEAR3 research processor, programmed with the advanced combination encoder (ACE) strategy [62] with parameters matching the individual CI user's clinical settings.

4.1.2 Speech material and subjective testing

The speech material presented to the participants consisted of the IEEE sentence corpus [63], which contains sentences with 7 to 12 words, organized in 72 lists of 10 sentences each. The sentences were produced by a male speaker and recorded in anechoic conditions. The sentences were equalized to the same root mean square value of 65 dB. The sampling frequency used for recording was 25 kHz and the speech files were down-sampled to 16 kHz for this experiment. The effects of reverberation and additive noise were introduced via digital simulation.

Room impulse responses (RIR) obtained experimentally were convolved with the clean speech signals [64, 65] to generate reverberant speech with approximate reverberation times (RT60) of 0.3, 0.6, 0.8, and 1 s. The first three RIRs [64] were obtained using a Tannoy CPA5 loudspeaker inside a rectangular reverberant room with dimensions 10.06 m × 6.65 m × 3.4 m (length × width × height), and a total volume of 227.5 m^3. The overall reverberant characteristics of the room were altered by hanging absorptive panels from hooks mounted on the walls close to the ceiling. The source-to-microphone distance was beyond the critical distance (at 5.5 m). The RIR with RT60 = 1.0 s, in turn, was obtained using a CORTEX MKII manikin artificial head and a single-cone loudspeaker (FOSTEX 6301 B) with 10 cm diameter in a 5.5 m × 4.5 m × 3.1 m room without any absorptive panels [65]. The loudspeaker was placed at 0° azimuth in the frontal plane at a 1.25 m distance from the head. All RIRs were measured biterally, but only one of the responses was used to generate the reverberant stimuli.

Speech-shaped noise was also added to the anechoic and the above mentioned reverberant signals to generate the noise-only and noise-plus-reverberation conditions, respectively. Noise was added at a signal-to-noise-ratio (SNR) of -5, 0, 5 and 10 dB for the anechoic samples and 5 and 10 dB for the reverberant samples. For the noise plus reverberation condition, the reverberant signals served as reference for SNR computation.

Participants were also tested using sentences enhanced using an ideal reverberant masking (IRM) strategy as described in [61]. These sentences were under reverberant conditions with RT60s of 0.6s, 0.8s, and 1.0s, and all of the noise plus reverberation conditions described above. The IRM algorithm was configured to use either 2 or 3 different thresholds for each condition, producing a total of 18 non-linearly enhanced conditions.

Two different sentence lists (20 sentences) were used for each of the above mentioned conditions. The volume of the presented sentences was adjusted by the individual listeners to a comfortable level prior to the beginning of the experiment and then kept constant throughout the experimental protocol. To maintain consistency across all participants, speech stimuli were presented unilaterally (to the ear with the highest performance for bilateral users). Listeners were instructed to repeat all identifiable words and per-participant intelligibility scores were calculated as the ratio of the number of correctly identified words to the total number of presented words.

4.1.3 Objective metrics

MATLAB and C/C++ implementations of the benchmark metrics were used to evaluate their performance with the CI intelligibility database. For NCM, CSII, PESQ, oPESQ, FWSSRR, and KLD, we used MATLAB implementations provided by Cosentino et al. [45]. The original MATLAB code for ModA was provided by Chen et al. [43]. We used reference implementations in C/C++ for ANIQUE+ and P.563 provided by the American National Standards Institute (ANSI) and ITU-T, respectively [48, 50]. Our SRMR-CI implementation was based on the original SRMR MATLAB code by Falk et al. [20], to which the modifications proposed in Chapter 3 were added.

Different versions of the SRMR-CI were tested in order to evaluate the effect of each of the proposed modifications separately and the optimal modulation frequency range. First, we tested only the replacement of the auditory filterbank by the CI-inspired one. Then, the following modulation frequency ranges were evaluated: $4 - 64$ Hz, $4 - 50$ Hz, $4 - 40$ Hz, and $4 - 30$ Hz. Finally, modulation energy thresholding was tested with the abovementioned modulation frequency ranges. For the SRMR-CI-intr metric, the scores for the clean signal were computed with the same SRMR-CI configuration as the distorted signal.

4.1.4 Performance criteria

In order to reduce inter- and intra-subject variability, we used the per-condition averages of the intelligibility scores found during subjective listening tests as the true scores. Averages were computed for all 20 sentences for each participant, and the results were then averaged over all the participants for each distortion condition. We evaluated performance in two different cases: first, only conditions including samples from environmental distortions were considered. A total of 13 conditions were evaluated: one clean, 4 noise-only, 4 reverberation-only, and 4 noise-plus-reverberation conditions. For the second test, we added the 18 IRM-enhanced conditions described previously to the conditions used in the first test in order to evaluate how the metrics performed with non-linearly processed/enhanced files. This second test is important, as it signals how well the investigated measures would fare at a potential intelligibility-aware enhancement algorithm.

To measure how well metrics are able to map to the true intelligibility scores, we used four different performance criteria. The Pearson correlation coefficient (ρ) was used to measure linear relationships between the objective and subjective scores. Spearman rank correlation (ρ_{spear}) was used to assess the ranking capability of the objective metrics (i.e., lower intelligibility scores should lead to lower metric scores, without expecting any kind of parameterized relationship between objective and subjective scores). Finally, motivated by Plomp's work [66], we computed also the Pearson correlation coefficient of a sigmoidal function mapping between objective scores and the intelligibility scale. This function is given by:

$$Y = \frac{1}{1 + e^{-(\alpha_1 X - \alpha_2)}} \times 100\% \tag{4.1}$$

where α_1 and α_2 are the fitting parameters, X is the objective metric, and Y is the subjective intelligibility score. The fitting is done via a generalized linear model regression between the subjective scores and the objective scores (normalized to the range $[0, 1]$). To force all mapped functions to start from the origin, we added a point at $(0, 0)$ with weight 100 to the points to be fitted, while attributing weight 1 to all the other points. Based on this mapping function, we also computed the root mean square error (RMSE) between predicted and true intelligibility scores. Additionally, to evaluate variability of the predicted scores in each condition, we computed the standard deviation for each of the conditions. To summarize variability ratings for each metric, we use the average RSD (expressed in %) for all conditions. The average RSD is a measure of spread and is used to assess the precision of a given metric.

4.2 Results

4.2.1 Environmental distortion conditions

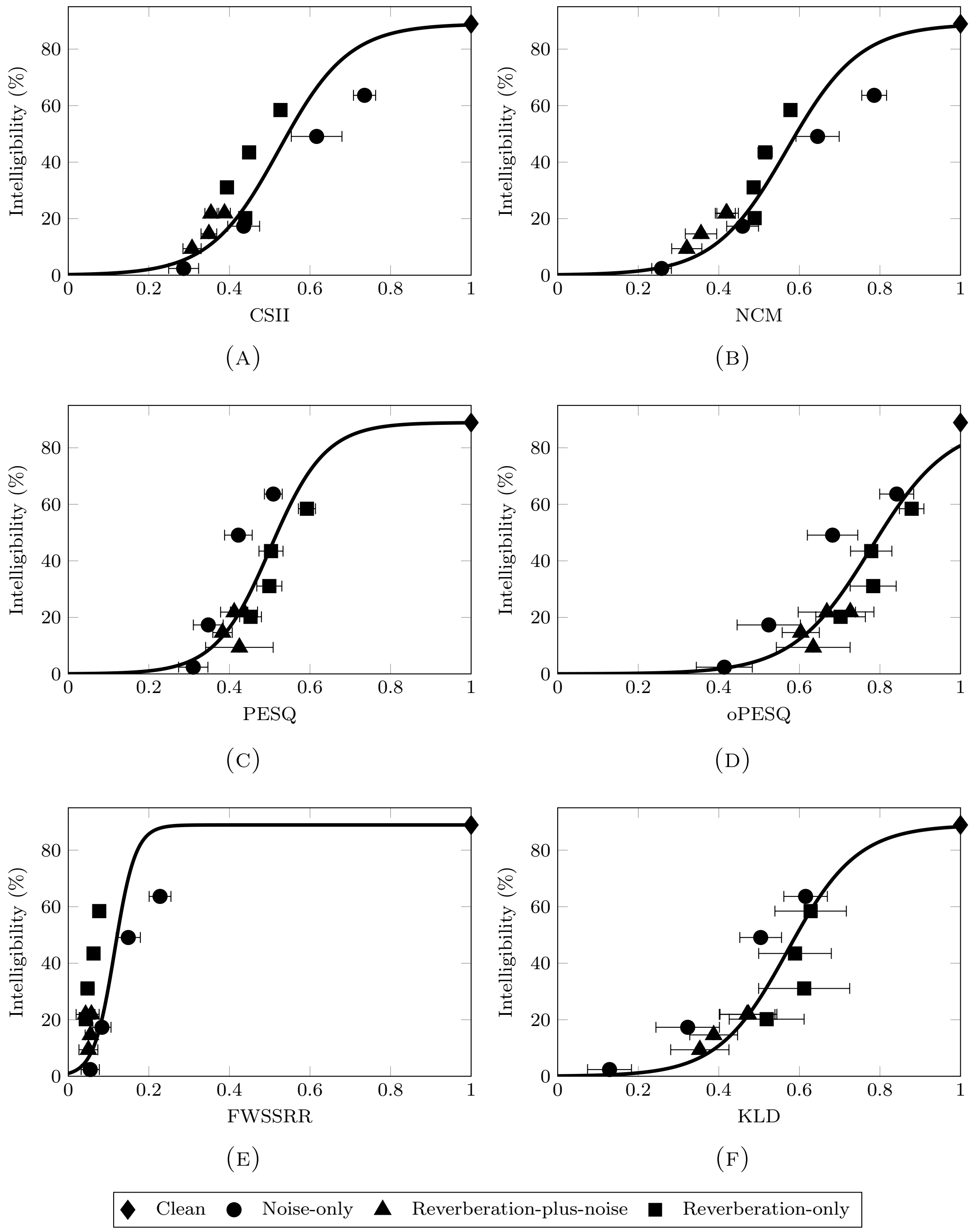
Table 4.1 summarizes the results obtained for the environmental distortion conditions (clean, noise-only, reverberation-only, and noise-plus-reverberation). NCM obtained the best performance among intrusive metrics, followed by CSII. FWSSRR and KLD showed higher variability than other intrusive metrics, on par with non-intrusive ones. P.563 had the best performance among the non-intrusive benchmark metrics. ANIQUE+, in turn, showed the poorest performance and highest variability of all evaluated metrics. The sigmoidal fits for intrusive and non-intrusive benchmark metrics can be seen in Figures 4.1 and 4.2, respectively, where different markers represent different condition types and horizontal errorbars represent one standard deviation of the condition scores. Respective coefficients α_1 and α_2 from Eq. 4.1 for the fitted functions can be found in Appendix A.

The use of the CI-inspired filterbank in SRMR-CI alone led to improvements of 0.02–0.04 in correlations. Figure 4.3) shows a comparison of both metrics. Adjusting the modulation frequency range led to additional improvements in performance, while reducing significantly RMSE and RSD%, reaching optimal correlations values for the $CF1$-$CF8$ ranges 4-40 Hz and 4-30 Hz. As shown in Figure 4.4, reducing the modulation frequency range translated scores for distorted files horizontally to the right on the SRMR-CI scale, narrowing the distance between them and the clean condition.

Modulation energy thresholding, in turn, significantly reduced RMSE and RSD% (with relative reductions of approximately 50%), even after the improvements seen with narrower modulation frequency ranges, even though it led to a small reduction in ρ_{spear} as some ranks were slightly reverted, as shown in Figure 4.5. Lastly, SRMR-CI-intr showed correlations and RMSE similar to its non-intrusive counterpart. However, RSD% values are similar to the proposed non-intrusive version with modulation energy thresholding, thus signalling the advantages obtained with the proposed changes. For practical applications, the proposed non-intrusive measure is more appealing, as it does not require access to the original clean speech file.

Table 4.1: Performance comparison for the environmental distortion conditions. In the SRMR-CI cases, the frequency range between parentheses represent the CF1-CF8 values investigated.

Metric	ρ	ρ_{spear}	ρ_{sig}	RMSE	RSD%
Intrusive benchmark metrics					
CSII	0.93	0.91	0.93	10.57	5.08
NCM	0.96	0.93	0.96	8.56	5.87
PESQ	0.85	0.83	0.89	11.02	7.58
oPESQ	0.88	0.90	0.91	9.80	8.68
FWSSRR	0.75	0.64	0.84	19.43	27.95
KLD	0.90	0.91	0.92	10.09	16.70
Non-intrusive benchmark metrics					
ModA	0.82	0.76	0.82	15.66	15.15
ANIQUE+	0.72	0.70	0.71	19.70	30.50
P.563	0.89	0.88	0.89	12.52	19.00
SRMR-based metrics (non-intrusive)					
SRMR	0.93	0.89	0.94	12.66	21.39
SRMR-CI (4-128 Hz)	0.96	0.93	0.95	11.07	22.29
SRMR-CI (4-64 Hz)	0.96	0.99	0.97	11.60	20.22
SRMR-CI (4-50 Hz)	0.98	0.99	0.98	8.86	17.09
SRMR-CI (4-40 Hz)	0.98	0.99	0.98	7.78	16.52
SRMR-CI (4-30 Hz)	0.98	0.97	0.98	6.27	15.83
SRMR-CI (energy thr., 4-64 Hz)	0.97	0.96	0.97	9.04	10.50
SRMR-CI (energy thr., 4-50 Hz)	0.98	0.96	0.98	8.18	10.12
SRMR-CI (energy thr., 4-40 Hz)	0.98	0.96	0.98	7.24	9.79
SRMR-CI (energy thr., 4-30 Hz)	0.98	0.97	0.98	6.05	9.35
SRMR-based metrics (intrusive)					
SRMR-CI-intr (4-64 Hz)	0.98	0.97	0.98	9.55	16.23
SRMR-CI-intr (4-50 Hz)	0.98	0.98	0.98	8.79	15.94
SRMR-CI-intr (4-40 Hz)	0.98	0.97	0.98	7.90	15.36
SRMR-CI-intr (4-30 Hz)	0.97	0.97	0.98	6.78	14.50
SRMR-CI-intr (energy thr., 4-64 Hz)	0.98	0.96	0.97	8.82	11.57
SRMR-CI-intr (energy thr., 4-50 Hz)	0.98	0.96	0.98	8.12	11.25
SRMR-CI-intr (energy thr., 4-40 Hz)	0.99	0.96	0.98	7.28	10.73
SRMR-CI-intr (energy thr., 4-30 Hz)	0.98	0.97	0.98	6.23	9.77



(A)

(B)

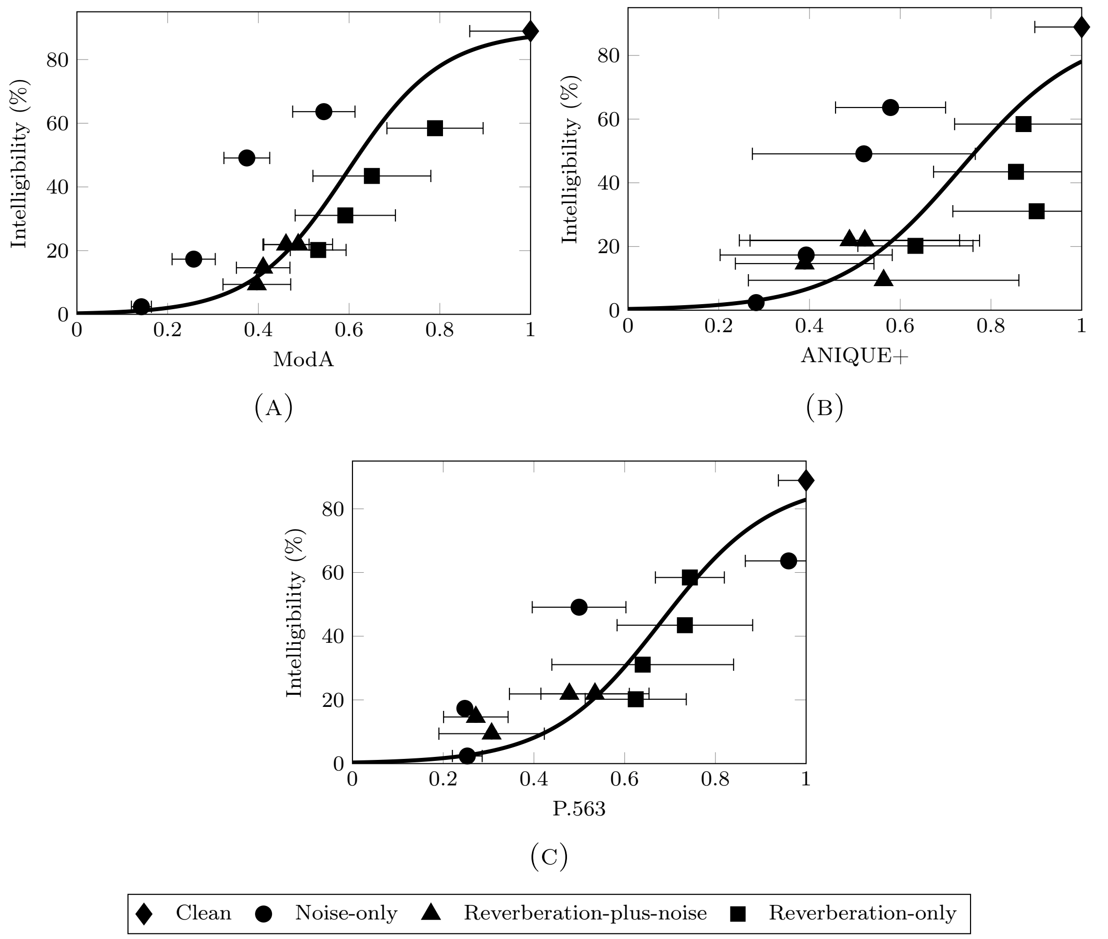
(C)

(D)

(E)

(F)

Figure 4.1: Scatterplots and sigmoidal mappings for the benchmark intrusive metrics (for environmental distortion conditions): (A) CSII, (B) NCM, (C) PESQ, (D) oPESQ, (E) FWSSRR, and (F) KLD.

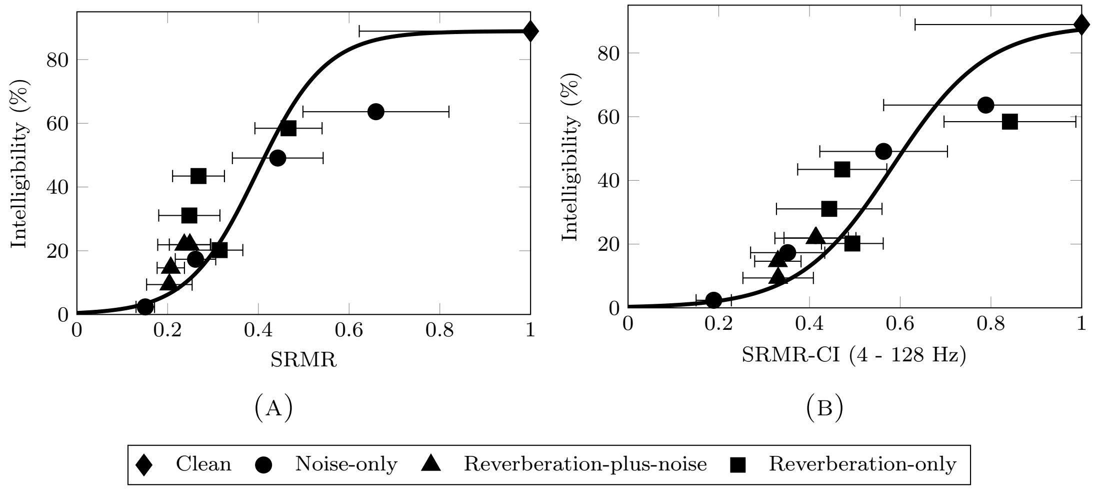


(A)

(B)

(C)

Figure 4.2: Scatterplots and sigmoidal mappings for the non-intrusive metrics (for environmental distortion conditions): (A) ModA, (B) ANIQUE+, (C) P.563.



(A)

(B)

Figure 4.3: Scatterplots and sigmoidal mappings for (a) the original SRMR implementation, (b) SRMR-CI (4 – 128 hz modulation frequency range).

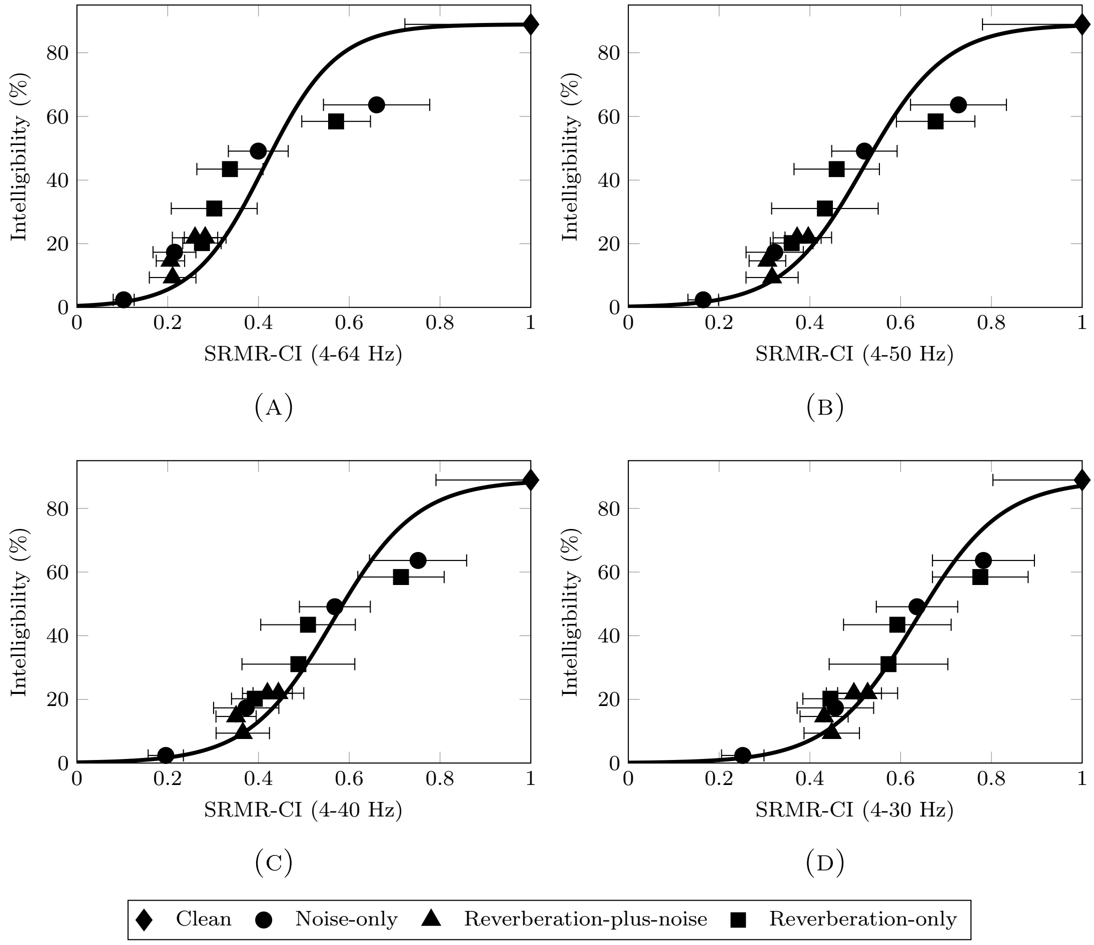
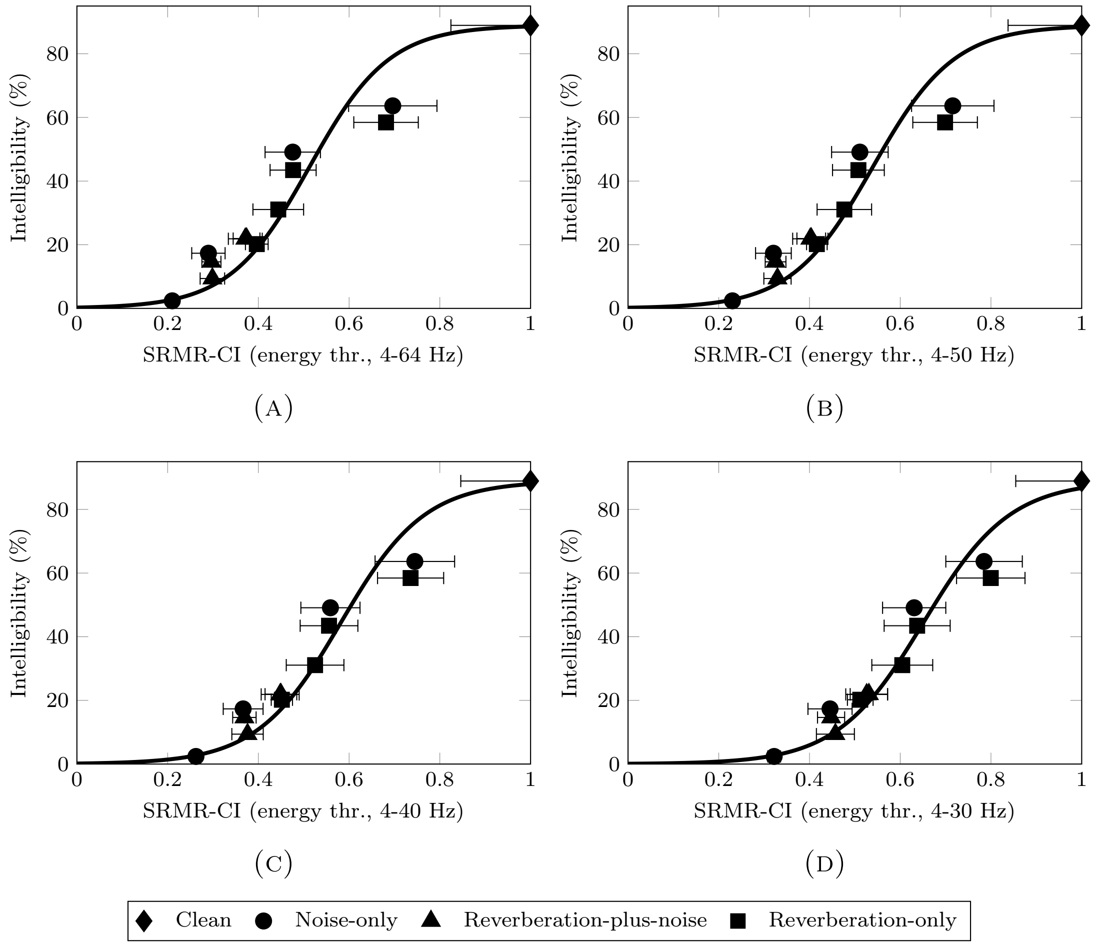


FIGURE 4.4: Scatterplots and sigmoidal mappings for SRMR-CI using different modulation frequency ranges: (A) 4–64 Hz, (B) 4–50 Hz, (C) 4–40 Hz, (D) 4–30 Hz.



(A)

(B)

(C)

(D)

FIGURE 4.5: Scatterplots and sigmoidal mappings for SRMR-CI using the proposed modulation energy thresholding scheme and different modulation frequency ranges: (A) 4–64 Hz, (B) 4–50 Hz, (C) 4–40 Hz, (D) 4–30 Hz.

4.2.2 Environmental distortion and enhanced conditions

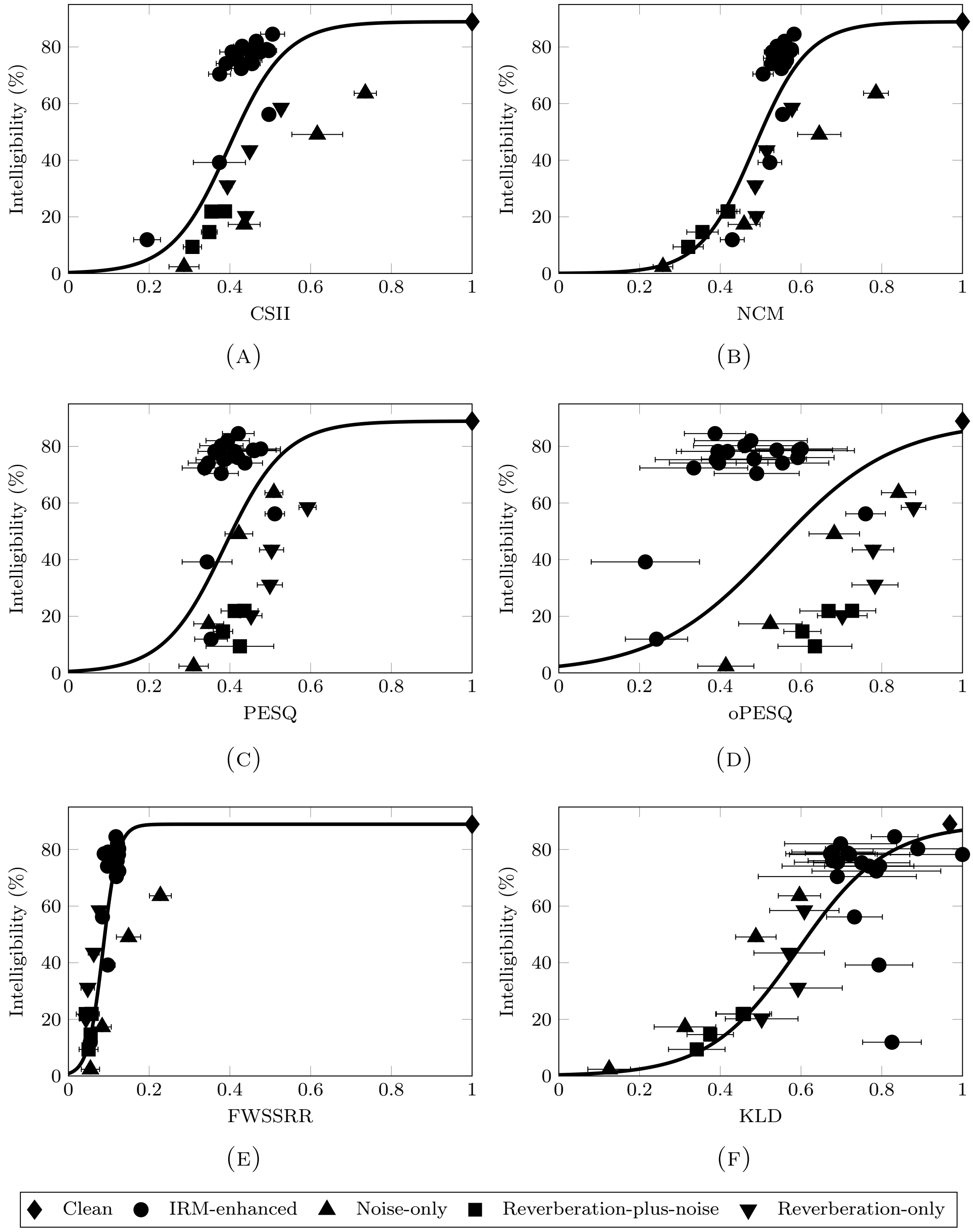
Results for the tests including the IRM-enhanced conditions are presented in Table 4.2. Compared to the results presented in Section 4.2.1, the mappings from objective to subjective scores were poorer for all metrics. As we can note in Figure 4.6, most intrusive metrics had a ranking issue with the enhanced files, as their scores were similar to those found for conditions with lower subjective ratings. The lowest ρ_{spear} values were found for PESQ and oPESQ. From all intrusive metrics, the higher correlations were found for NCM and KLD; NCM, however, shows much lower variability per condition. Non-intrusive benchmark metrics did not have a performance decrease as significant as intrusive metrics. ModA had higher ρ, ρ_{sig} and lower RSD%, while ANIQUE+ showed better ranking and slightly lower RMSE. Scatterplots and the respective sigmoidal fits can be seen in Figure 4.7, were we can note that while there are some ranking problems between IRM-enhanced and environmental distortion conditions, they are not as expressive as those seen in Figure 4.6 for intrusive files.

The SRMR-CI metrics resulted in the best performances for this test. Results followed a trend similar to that found for non-enhanced files. The gains, however, with the proposed SRMR-CI and modified SRMR-CI measures relative to the original SRMR were more significant than in the previous test (Section 4.2.1). For example, the use of the CI-inspired filterbank (see Figure 4.8) led to an increase of up to 21% in ρ. Moreover, narrowing the range of analyzed modulation frequencies consistently increased ρ and ρ_{sig} and reduced both RMSE and RSD% (see Figure 4.9). Using the proposed modulation energy thresholding scheme led to additional improvements in both correlations, RMSE and RSD%. As we can see in Figure 4.10d, the adjusted SRMR-CI metric resulted in an improved fit to subjective data. Compared to the original SRMR metric, the proposed metric shows a gain of 40% in ρ, 25% in ρ_{spear}, 34% in ρ_{sig}, and relative reductions of 50% in both

RMSE and RSD%. As in the previous test, no significant differences were found between the SRMR-CI-intr metric and the proposed non-intrusive counterpart, thus suggesting that the SRMR-CI metric does not benefit from the additional "normalization" provided by using SRMR-CI scores from the clean signals.

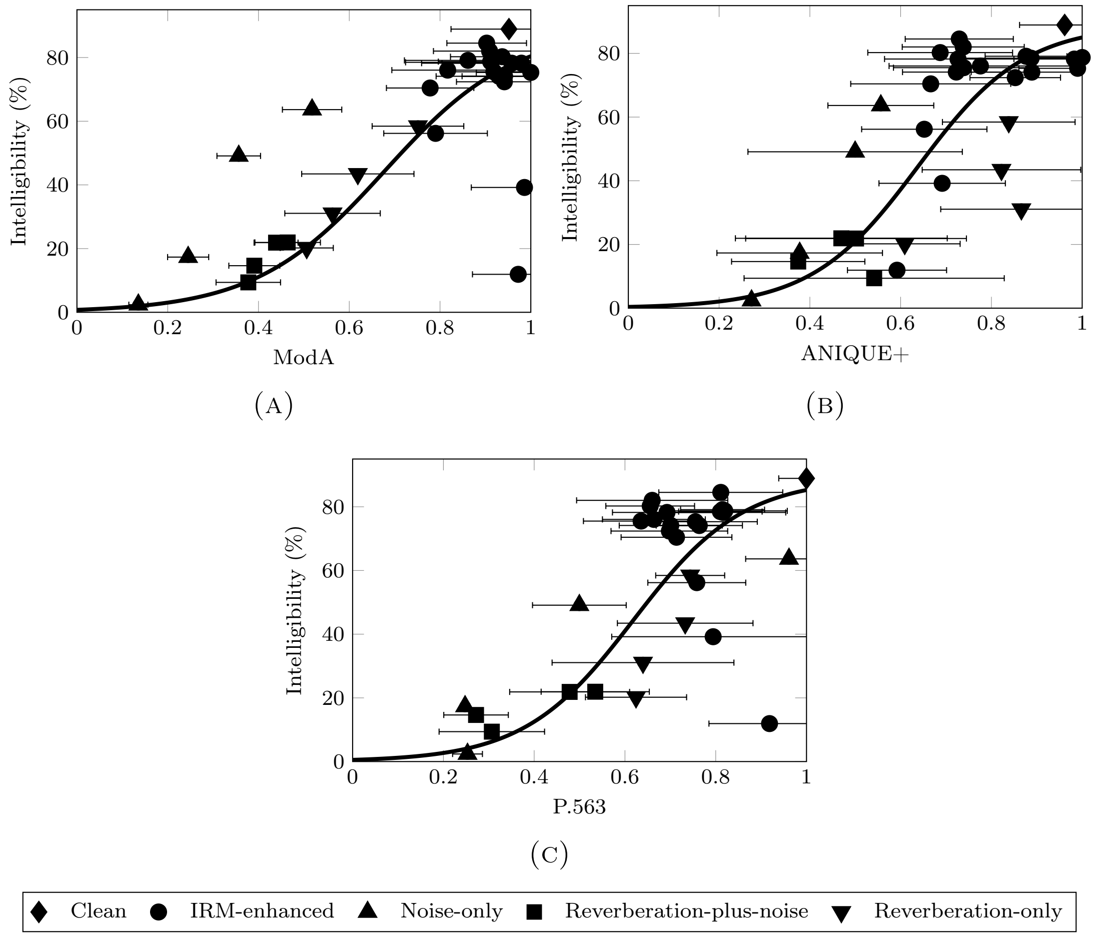
Table 4.2: Performance comparison for the environmental distortion and IRM-enhanced conditions. In the SRMR-CI cases, the frequency range between parentheses represent the CF1-CF8 values investigated.

Metric	ρ	ρ_{spear}	ρ_{sig}	RMSE	RSD%
Intrusive benchmark metrics					
CSII	0.51	0.60	0.62	20.87	5.96
NCM	0.68	0.74	0.84	14.47	4.59
PESQ	0.26	0.19	0.25	28.62	10.33
oPESQ	-0.06	-0.17	0.06	37.76	19.53
FWSSRR	0.37	0.72	0.83	15.39	17.56
KLD	0.76	0.66	0.79	17.36	16.78
Non-intrusive benchmark metrics					
ModA	0.78	0.59	0.79	18.55	14.29
ANIQUE+	0.75	0.69	0.77	18.05	23.22
P.563	0.66	0.55	0.69	20.66	17.87
SRMR-based metrics (non-intrusive)					
SRMR	0.49	0.53	0.61	21.50	20.64
SRMR-CI (4-128 Hz)	0.70	0.59	0.75	19.00	20.96
SRMR-CI (4-64 Hz)	0.73	0.71	0.83	16.19	20.14
SRMR-CI (4-50 Hz)	0.80	0.76	0.89	12.36	16.82
SRMR-CI (4-40 Hz)	0.83	0.75	0.91	11.00	16.42
SRMR-CI (4-30 Hz)	0.86	0.77	0.93	9.87	16.10
SRMR-CI (energy thr., 4-64 Hz)	0.82	0.76	0.89	12.59	11.60
SRMR-CI (energy thr., 4-50 Hz)	0.84	0.78	0.91	11.06	11.21
SRMR-CI (energy thr., 4-40 Hz)	0.86	0.76	0.93	9.87	10.83
SRMR-CI (energy thr., 4-30 Hz)	0.89	0.78	0.95	8.80	10.30
SRMR-based metrics (intrusive)					
SRMR-CI-intr (4-64 Hz)	0.77	0.72	0.83	15.46	18.91
SRMR-CI-intr (4-50 Hz)	0.80	0.76	0.87	13.47	18.49
SRMR-CI-intr (4-40 Hz)	0.83	0.76	0.89	12.02	17.92
SRMR-CI-intr (4-30 Hz)	0.86	0.75	0.91	10.96	17.14
SRMR-CI-intr (energy thr., 4-64 Hz)	0.82	0.78	0.88	13.17	13.06
SRMR-CI-intr (energy thr., 4-50 Hz)	0.85	0.79	0.91	11.23	12.63
SRMR-CI-intr (energy thr., 4-40 Hz)	0.88	0.77	0.93	9.88	11.98
SRMR-CI-intr (energy thr., 4-30 Hz)	0.91	0.77	0.95	8.74	10.93



(A) (B) (C) (D) (E) (F)

Figure 4.6: Scatterplots and sigmoidal mappings for the benchmark intrusive metrics (including IRM-enhanced files): (A) CSII, (B) NCM, (C) PESQ, (D) oPESQ, (E) FWSSRR, and (F) KLD.

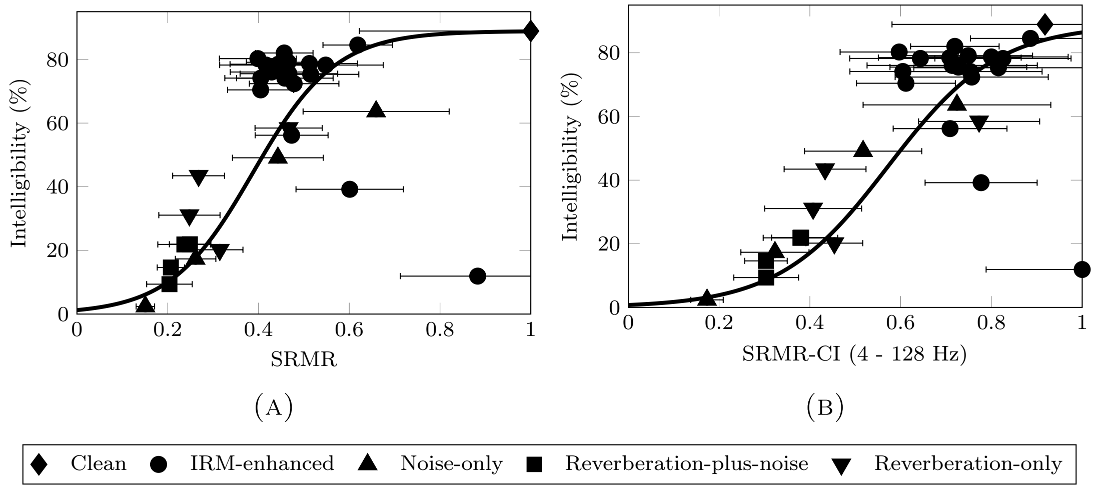


(A)

(B)

(C)

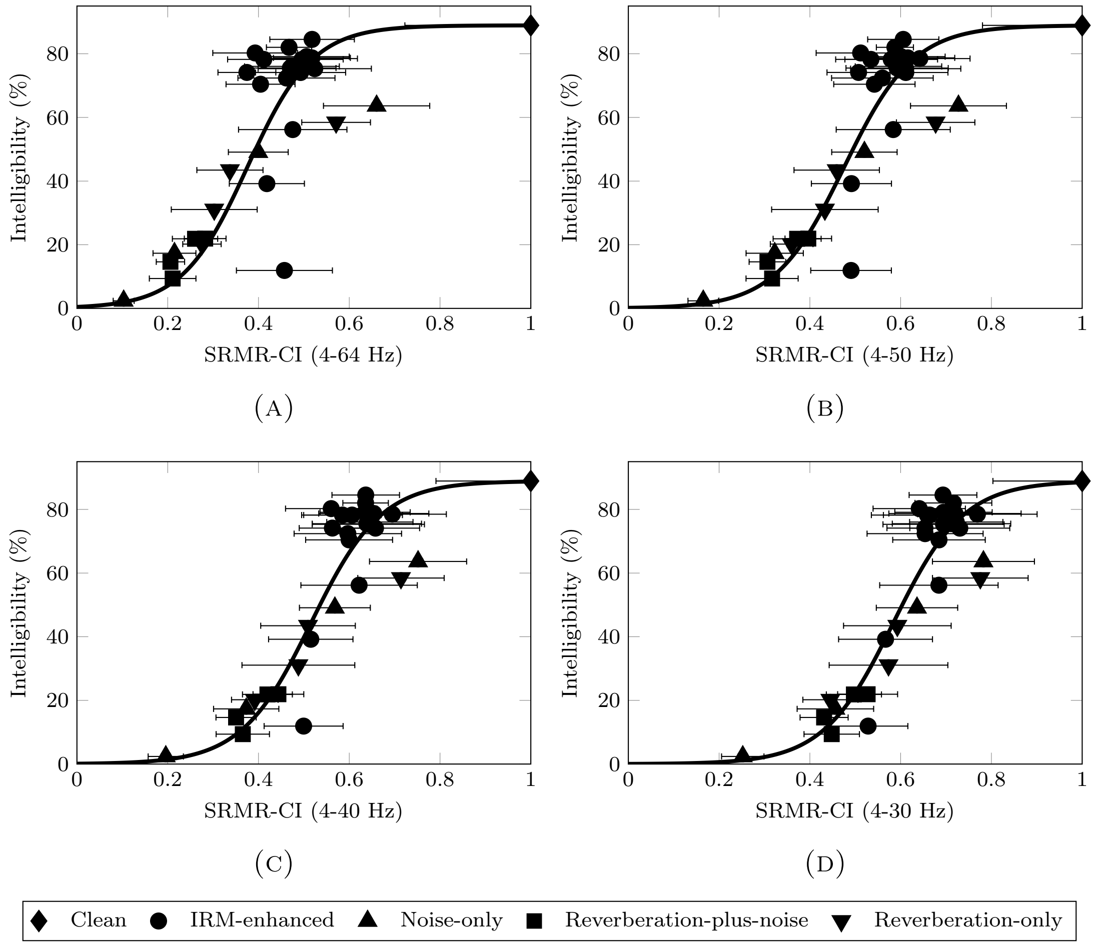
Figure 4.7: Scatterplots and sigmoidal mappings for the non-intrusive metrics (including enhanced files): (A) ModA, (B) ANIQUE+, (C) P.563.



(A)

(B)

Figure 4.8: Scatterplots and sigmoidal mappings (including enhanced files) for (A) the original SRMR implementation, (B) SRMR-CI (4 – 128 Hz modulation frequency range).



(A)

(B)

(C)

(D)

Figure 4.9: Scatterplots and sigmoidal mappings (including enhanced files) for SRMR-CI using different modulation frequency ranges: (A) 4–64 Hz, (B) 4–50 Hz, (C) 4–40 Hz, (D) 4–30 Hz.

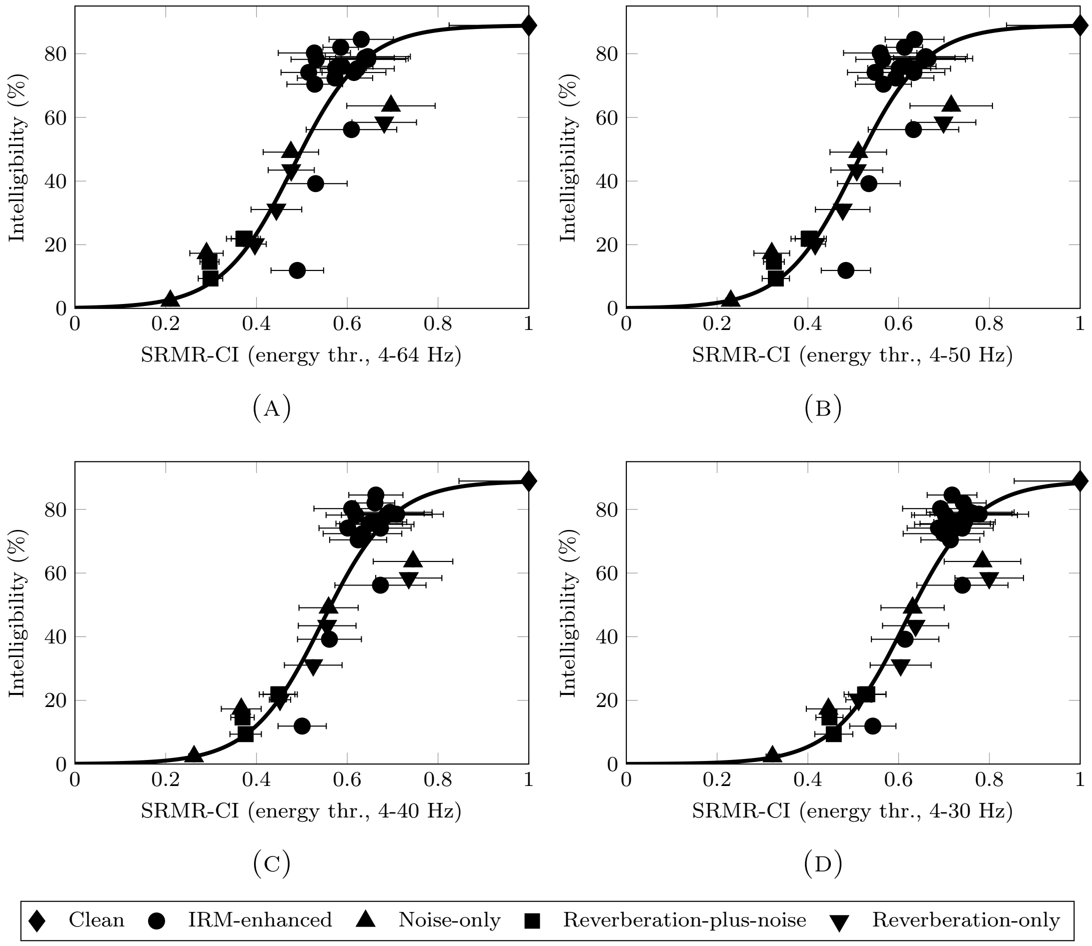


FIGURE 4.10: Scatterplots and sigmoidal mappings for SRMR-CI using the proposed modulation energy thresholding scheme and different modulation frequency ranges: (A) 4–64 Hz, (B) 4–50 Hz, (C) 4–40 Hz, (D) 4–30 Hz.

4.3 Discussion

4.3.1 Objective intelligibility measurement: importance of temporal envelope cues for CI users

Preservation of temporal envelope cues has long been regarded as an important factor in speech perception [67, 68]. This is particularly true for hearing-impaired listeners who have reduced ability to process fine temporal structure and spectral cues [69, 70]. To this end, we observed that the NCM intrusive measure, which itself is based on temporal envelope cues, outperformed the CSII measure, based on fine spectral cues (see Tables 4.1 and 4.2) in terms of the correlation and RMSE metrics. Moreover, the results obtained with both SRMR-based measures and ModA provide further evidence of the importance of temporal envelope cues for speech intelligibility prediction in cochlear implants. These findings corroborate those previously reported in the literature showing reliable intelligibility predictions for vocoded speech with NH listeners obtained by intrusive and non-intrusive measures based on temporal envelope cues [45]. Lastly, it is also known that reverberation modifies temporal envelope cues, thus severely degrades speech recognition for CI users.

4.3.2 CI-inspired metrics: Are they always better?

The original SRMR metric depends on temporal envelope cues obtained from multiple acoustic frequency bands, similar to the cues used by CI listeners. Notwithstanding, the SRMR metric mimics several normal hearing percepts, such as cochlear processing (i.e., 23-channel gammatone filterbank) and temporal envelope frequency selectivity [47]. As such, we expected that improved performance would be obtained once CI hearing percepts were incorporated into the measure.

This was indeed observed and the proposed SRMR-CI measure incorporated a CI-inspired filterbank (i.e., emulated the Nucleus mel-like filterbank) and explored an optimal modulation frequency representation by adjusting its energy and frequency ranges. The optimal modulation frequency range, besides being useful to reduce the metric variability, may also be a better match to temporal envelope frequency selectivity in cochlear implant users. With such changes to the metric, an increase of up to 5% was obtained in ρ relative to the original SRMR measure with non-enhanced files and 40% when including IRM-enhanced files.

Inspired by the improvements in correlation with the above CI-inspired metrics, we also investigated two updates to the NCM and CSII measures. More specifically, the NCM measure was updated to include the Nucleus-inspired filterbank and reduced modulation frequency ranges: 3 – 82 Hz and 3 – 38 Hz, similar to the ranges covered by the modulation filterbank used by SRMR-CI with center frequencies from 4 – 64 Hz and 4 – 30 Hz, respectively. Unlike the SRMR-CI measure, the so-called NCM-CI measure did not show any improvements in objective speech intelligibility prediction, thus suggesting that it is the ratio of low-to-high modulation frequency content that correlates with speech perception in noise, and not the entire modulation spectrum. A second update included the reduction of the dynamic range of the NCM and CSII measures from [-15,15] dB to [-5,5] dB, to mimic the limited electrical dynamic range associated with electrical stimulation. While a slight increase in ρ was observed for the NCM measure, all other performance criteria resulted in slightly lower correlation for both the NCM and CSII measures.

The ModA metric, in turn, despite being developed specifically with CI users in mind, does not emulate device characteristics, such as is proposed in this work for SRMR-CI. Instead, it uses a simplified filterbank based on four 4th-order Butterworth bandpass filters with mel-scale like center frequencies. While this setup was shown to perform well in reverberation-only conditions (e.g., in [43]), the experiments we presented have shown reduced performance in the noise-only condition, likely due to the fact that the 0.5-10 Hz modulation frequency range was significantly affected by the speech-shaped noise. As can be seen in Figures 4.2a and 4.7a, ModA scores were consistently lower for noise-only conditions when compared to reverberation-only conditions, even though some of these conditions have similar subjective scores. As such, we recommend that objective intelligibility measures tailored towards CI applications be equipped with a higher-resolution CI-inspired acoustic filterbank, such as the one used in the investigated NCM-CI and SRMR-CI measures.

4.3.3 Assessing intelligibility of enhanced speech

One desirable feature for a speech intelligibility metric tuned to CI is being able to assess intelligibility from both natural and enhanced signals, thus opening doors for intelligibility-aware speech enhancement strategies. However, enhanced speech may include unwanted artifacts that distort the speech signal differently than environment distortions. For example, artifacts may increase the energy in the lower modulation frequency bands (4–20 Hz), which would be interpreted by SRMR and ModA as speech content and therefore skew their predicted scores. Such phenomenon may be observed in Figures 4.9a and 4.7a, where all IRM-enhanced signals resulted in similar SRMR-CI and ModA scores. However, while the majority of the enhanced conditions have intelligibility scores around 60–80%, two configurations of the IRM algorithm resulted in intelligibility scores of 39% and 12%. As an example, ModA predictions for these two conditions are higher than the predicted score for the clean condition. However, the SRMR-CI version with narrower modulation frequency range and modulation energy thresholding achieved higher correlations as it was able to differentiate these conditions, leading to a better sigmoidal mapping to the subjective scores.

The effect of non-linear processing was significant on intrusive benchmark metrics, including NCM and CSII, leading to poorer correlations and higher RMSE than SRMR-CI for almost all cases. Channel selection schemes, such as the IRM algorithm, will remove any information in spectral channels dominated by reverberation, causing coherence/SNR scores for channels that were removed to decrease significantly. KLD, on the other hand, is a metric based on signal statistics and therefore "blind" to such changes, which led to a slightly higher ρ than the other intrusive metrics.

4.3.4 Speech content variability in non-intrusive metrics based on the modulation spectrum

As ModA uses only low modulation frequencies, we expected it to be less susceptible to the fundamental frequency of speech. We ran the same experiments with the CV pairs and TIMIT sentences described in Chapter 3 using ModA, and found a correlation coefficient of 0.06 between ModA intelligibility scores and average pitch. Regarding manner of articulation and vowel score variability, RSDs were between 22% and 28%, on par with the implementation of SRMR-CI using modulation energy thresholding. However, what causes the poorer performance on CI speech intelligibility under noise and reverberation is the extreme limitation on modulation spectrum range. By using only modulation frequencies up to 10 Hz, ModA is unable to detect energies in frequencies dominated by distortions, being able only to measure the decrease in energies at the modulation frequency range related to speech articulation, thus resulting in lower correlation values.

ANIQUE+, on the other hand, makes use of both low and high modulation frequencies to compute quality scores. Energies in low modulation frequencies (from 2 to 30 Hz) are used to compute the average articulation power, while the high modulation frequencies (from 30 Hz up to 256 Hz) are used to compute the so-called non-articulation power. As shown by the experiments in Chapter 3, this region is subject to energies coming from the fundamental frequency of human speech. However, unlike SRMR and SRMR-CI, the so-called articulation and non-articulation modulation energies are used as inputs to a multilayer perceptron. This step removes any pitch effect on the metric ratings, as evidenced by our experiment: for the TIMIT database, a correlation of 0.09 between ANIQUE+ with average F_0 was obtained. This metric presents a variability on par with the original SRMR-CI for the CV pairs, ranging from 30.6% to 42.8%.

4.3.5 Errors per distortion type

Table 4.3 shows the RMSE values for each distortion type, computed from the sigmoidal fit to the subjective scores including the enhanced conditions. We can see that for the SRMR-CI metric, errors in noise and enhanced conditions are reduced when we reduce modulation frequency range. When using modulation energy thresholding, errors for the noise-plus-reverberation conditions are also reduced. Performance for the speech samples corrupted by reverberation was not significantly affected by these changes overall.

Most benchmark intrusive metrics show high errors for almost all condition types, as reflected by their poor correlations. From the non-intrusive metrics, ModA showed higher error values for the noisy conditions, while P.563 and the original SRMR metric had higher errors for the enhanced condition. The noise-plus-reverberation condition had lower RMSE for almost all metrics, as this condition consistently had low subjective scores and lower subjective score variability than the other conditions.

Table 4.3: RMSE comparison for each condition type, based on the sigmoidal fit

Metric	Noise	Reverb	N+R	Enhanced
Intrusive benchmark metrics				
CSII	29.95	23.19	14.27	19.65
NCM	22.36	16.99	2.65	13.56
PESQ	14.58	35.56	36.04	28.17
oPESQ	20.87	37.40	44.78	40.02
FWSSRR	26.23	20.71	7.58	12.12
KLD	16.10	9.48	2.30	20.97
Non-intrusive benchmark metrics				
ModA	29.88	4.69	5.27	19.39
ANIQUE+	22.52	30.47	9.71	14.82
P.563	17.45	18.76	6.43	24.07
SRMR-based metrics (non-intrusive)				
SRMR	11.44	15.10	5.28	26.65
SRMR-CI (4-128 Hz)	23.33	8.68	5.76	15.33
SRMR-CI (4-64 Hz)	12.66	14.14	5.18	19.12
SRMR-CI (4-50 Hz)	12.20	12.58	3.52	13.86
SRMR-CI (4-40 Hz)	12.01	12.64	3.35	11.77
SRMR-CI (4-30 Hz)	10.75	13.49	4.33	9.88
SRMR-CI (energy thr., 4-64 Hz)	11.55	12.18	4.46	14.35
SRMR-CI (energy thr., 4-50 Hz)	11.11	12.30	3.88	12.10
SRMR-CI (energy thr., 4-40 Hz)	10.48	12.48	3.29	10.31
SRMR-CI (energy thr., 4-30 Hz)	9.03	13.07	2.68	8.71
SRMR-based metrics (intrusive)				
SRMR-CI-intr (4-64 Hz)	13.23	11.71	5.03	18.35
SRMR-CI-intr (4-50 Hz)	13.50	11.86	4.84	15.35
SRMR-CI-intr (4-40 Hz)	13.32	12.10	5.13	13.08
SRMR-CI-intr (4-30 Hz)	12.14	13.10	7.07	11.17
SRMR-CI-intr (energy thr., 4-64 Hz)	11.83	11.37	4.46	15.31
SRMR-CI-intr (energy thr., 4-50 Hz)	11.54	11.47	4.01	12.45
SRMR-CI-intr (energy thr., 4-40 Hz)	10.99	11.46	3.42	10.46
SRMR-CI-intr (energy thr., 4-30 Hz)	9.71	11.79	2.99	8.80

Chapter 5

Conclusion

5.1 Summary of research

In this work, we proposed a speech intelligibility metric tailored for CI users, termed SRMR-CI, based on a modulation spectral signal representation. This metric is based on SRMR, a metric that has been show to correlate well with speech quality and intelligibility for NH listeners, as well as room acoustics characteristics such as reverberation time.

Our proposal extends SRMR to improve its correlation to CI speech intelligibility based on three main modifications. First, we replaced the acoustic filterbank used in SRMR by a CI-inspired filterbank, similar to the one used in the ACE speech coding strategy present in Nucleus CI devices. Adjustment of the filterbank alone led to improvements of 21% in linear correlation, 6% in rank correlation, and 14% in sigmoidal-mapped correlation between the objective and subjective scores when compared to the original SRMR. The second modification was an adjustment to the modulation frequency range used by the metric to compute energies in the modulation spectrum. We showed that the range used by the original metric (4 – 128 Hz) is sensitive to variations in the fundamental frequency of speech, which led to a correlation of objective scores and average sentence pitch. We showed that reducing this bandwidth effectively eliminates this sensitivity and improves correlations with CI intelligibility. Lastly, we proposed a modulation energy thresholding scheme to account for variability in the modulation spectrum and SRMR scores related to speech content. By limiting the modulation energies to a range of 30 dB, we were able to reduce the metric variability and further improve correlations.

Compared to both intrusive and non-intrusive state-of-the-art metrics, our proposed metric achieved higher performance, especially when evaluating intelligibility from non-linearly processed/enhanced speech. Since most, if not all, CI devices currently on the market make use of advanced speech processing techniques to improve intelligibility for users, an objective speech intelligibility metric for CI users has to be able to take into account the effects of non-linear distortions. A non-intrusive metric such as SRMR-CI is a valuable resource for developing and validating new CI devices and enhancement algorithms, and can potentially be used in devices to adjust system parameters “on the fly”.

5.2 Future work

In this work, we tested SRMR-CI under several environmental conditions including different noise levels, reverberation times, and combined noise and reverberation effects. We also tested its performance on assessing intelligibility scores of a channel selection strategy designed to enhance speech signals under the effects of noise and reverberation for CI users (ideal reverberant masking, IRM). However, many different speech enhancement algorithms used in current devices are not necessarily based on binary spectral masking (e.g. spectral subtraction, acoustic beamforming), and more experiments are needed to evaluate whether the metric is able to correctly predict the intelligibility of speech subject to other non-linear processing algorithms.

Moreover, different CI users may show a large variability on speech recognition performance. SRMR-CI, while using general CI-inspired precepts, does not take individual characteristics such as device configuration/mapping into account. HASQI, a quality metric developed for hearing aids, incorporates individual characteristics such as the hearing loss levels at different frequencies into its auditory model to simulate the effects of outer and inner hair cells into the cochlear processing [71]. In the case of CI users, informations provided by clinical testing and device fitting sessions such as the dynamic range and gain for each electrode, stimulation rates, acoustic filterbank configuration, and processing strategy may be useful for improving the performance of the proposed objective metric.

Besides the advantages for device testing and development, access to a reliable non-intrusive speech intelligibility/quality metric may open doors to intelligibility- and/or quality-aware speech enhancement algorithms to improve speech-in-noise recognition for CI users. By detecting under which kind of environment the device is operating in (different types and levels of noise and reverberation), a speech enhancement algorithm can be adapted on the fly to better suit the current condition and fine-tune specific parameters in order to improve its performance.

Appendix A

Sigmoidal fit coefficients

This appendix includes the α_1 and α_2 coefficients found with the fitting described in Chapter 4. The metric scores have to be normalized to the [0,1] range by using the minimum and maximum scores for the condition means, and then applied to Equation 4.1.

TABLE A.1: Sigmoidal fit coefficients (environmental distortions only)

Metric	α_1	α_2
CSII	-4.1613	7.6904
NCM	-4.5476	7.8136
PESQ	-5.9434	11.6469
oPESQ	-7.3091	9.3711
FWSSRR	-1.7249	12.4981
KLD	-4.7876	8.2650
ModA	-3.3699	5.5082
ANIQUE+	-3.1044	4.1520
P.563	-3.3885	4.9625

Continued on next page

Metric	α_1	α_2
SRMR	-2.8516	6.7629
SRMR-CI (4-128 Hz)	-3.6014	6.0155
SRMR-CI (4-64 Hz)	-2.9328	6.7406
SRMR-CI (4-50 Hz)	-3.8574	7.2174
SRMR-CI (4-40 Hz)	-4.3280	7.5315
SRMR-CI (4-30 Hz)	-5.2461	8.2062
SRMR-CI (energy thr., 4-64 Hz)	-3.8828	7.3610
SRMR-CI (energy thr., 4-50 Hz)	-4.2013	7.6062
SRMR-CI (energy thr., 4-40 Hz)	-4.6730	7.8945
SRMR-CI (energy thr., 4-30 Hz)	-5.5934	8.5082
SRMR-CI-intr (4-64 Hz)	-3.5000	6.7650
SRMR-CI-intr (4-50 Hz)	-3.8324	6.8213
SRMR-CI-intr (4-40 Hz)	-4.2864	6.9998
SRMR-CI-intr (4-30 Hz)	-5.1453	7.4772
SRMR-CI-intr (energy thr., 4-64 Hz)	-3.8731	7.1455
SRMR-CI-intr (energy thr., 4-50 Hz)	-4.2077	7.2248
SRMR-CI-intr (energy thr., 4-40 Hz)	-4.6913	7.4430
SRMR-CI-intr (energy thr., 4-30 Hz)	-5.6132	7.9963

Table A.2: Sigmoidal fit coefficients (environmental distortions and enhanced files)

Metric	α_1	α_2
CSII	-3.2353	8.6233
NCM	-6.4088	13.3628
PESQ	-1.2112	3.9660

Continued on next page

Metric	α_1	α_2
oPESQ	0.7237	-0.4173
FWSSRR	-2.4168	31.3264
KLD	-3.5713	6.3887
ModA	-2.6867	4.4196
ANIQUE+	-3.7127	6.1017
P.563	-2.8526	5.0569
SRMR	-1.4135	4.5116
SRMR-CI (4-128 Hz)	-2.5843	5.0191
SRMR-CI (4-64 Hz)	-3.3871	9.5933
SRMR-CI (4-50 Hz)	-5.0587	10.8725
SRMR-CI (4-40 Hz)	-5.8260	11.4384
SRMR-CI (4-30 Hz)	-7.0728	12.1720
SRMR-CI (energy thr., 4-64 Hz)	-4.9128	10.4730
SRMR-CI (energy thr., 4-50 Hz)	-5.5372	11.1705
SRMR-CI (energy thr., 4-40 Hz)	-6.2666	11.6194
SRMR-CI (energy thr., 4-30 Hz)	-7.4535	12.1742
SRMR-CI-intr (4-64 Hz)	-4.0759	9.2656
SRMR-CI-intr (4-50 Hz)	-4.8025	9.8593
SRMR-CI-intr (4-40 Hz)	-5.5675	10.2756
SRMR-CI-intr (4-30 Hz)	-6.7579	10.8236
SRMR-CI-intr (energy thr., 4-64 Hz)	-4.7746	9.9115
SRMR-CI-intr (energy thr., 4-50 Hz)	-5.4417	10.3595
SRMR-CI-intr (energy thr., 4-40 Hz)	-6.1689	10.6688
SRMR-CI-intr (energy thr., 4-30 Hz)	-7.3194	11.1213

Annexe B

Synthèse

B.1 Introduction

Les technologies d'assistance à l'audition jouent un rôle important sur l'amélioration de la communication et la qualité de vie des personnes ayant une perte auditive. Dans la plupart des cas (perte auditive neurosensorielle légère à modérée), les audioprosthèses sont suffisantes pour améliorer la communication. Les personnes souffrant de déficience auditive profonde dans les deux oreilles nécessitent habituellement des implants cochléaires (IC), qui stimulent la cochlée directement avec des impulsions électriques.

Malgré les avancées technologiques de ces dernières années, les utilisateurs des IC se heurtent encore à de nombreuses limitations lorsq'ils sont éxposés à des environnements bruyants et réverbérants. Même dans des environnements à faibles temps de réverbération, comme les petites salles, la réverbération réduit considérablement l'intelligibilité de la parole [8]. Dans la plupart des environnements quotidiens, les effets du bruit et de la réverbération sont combinés, ce qui entraîne un faible niveau de compréhension de la parole [9]. La réverbération et le bruit distordent les informations importantes contenues dans la modulation de la parole, sachant que cette dernière represente une caractéristique principale pour le traitement de la parole es est utilisée dans les IC pour coder le signal de parole. Ces distorsions entravent la perception des modulations d'amplitude, des transitions de formants, du timbre et des frontières syllabiques et entre mots [10–12], introduisent des effets indésirables de masquage [8, 13, 14] et provoquent une mauvaise localisation du son [15]. Pour compenser ces limitations et pour améliorer l'intelligibilité de la parole dans l'environnement quotidien, des recherches récentes ont mis l'accent sur le développement d'algorithmes d'amélioration de la qualité des signaux de parole, comme la suppression du bruit, la sélection des canaux et déréverbération (par exemple,[17, 18]).

L'évaluation de la qualité des IC et des algorithmes d'amélioration nouvellement développés, ainsi que les effets des distorsions environnementales sur eux, nécessitent généralement des tests d'écoute subjectifs. Ceux-ci peuvent être réalisés soit avec les utilisateurs des implants ou en présentant de la parole vocodée (pour simuler le traitement du signal d'un implante cochléaire) à des auditeurs ayant une audition normale. Toutefois, ces tests sont coûteux et fastidieux, ce qui empêche leur utilisation lors du développement de produits. Les mesures de qualité et intelligibilité vocale objectives sont une alternative aux tests d'écoute subjectifs et permettent une évaluation rapide et reproductible, tout en permettant d'effectuer des mesures en temps réel directement sur l'appareil. Ce dernier point est particulièrement important dans le cas où la qualité/intelligibilité de la parole doit être contrôlée afin d'être utilisée pour adapter les paramètres d'un système d'amélioration de la qualité/intelligibilité de la parole en temps réel.

Les métriques objectives d'intelligibilité (ou qualité) peuvent être classées comme intrusives (avec un signal de parole sans bruit comme référence) ou non-intrusives (ou sans signal de référence). Métriques intrusives ont l'avantage d'être capable d'évaluer directement la quantité et le type de distorsion dans un signal corrompu.

Même si les deux classes de métriques peuvent être utilisés lors de l'élaboration d'un algorithme d'amélioration ou pour une évaluation / comparaison des différents dispositifs, les métriques intrusives ne peuvent pas être utilisées dans des applications pratiques en temps réel, car dans ce cas le signal de référence n'est pas disponible. Métriques non-intrusives ne nécessitent pas un signal de référence, il est donc possible de les appliquer pour caractériser quantitativement les gains d'intelligibilité obtenus avec un algorithme d'amélioration de la parole (par exemple, déréverbération) directement sur l'appareil. Ils permettent également le développement d'algorithmes d'amélioration « conscients » d'intelligibilité, ce qui pourrait régler les paramètres de l'appareil en temps réel en tenant compte des paramètres actuelles d'intelligibilité imposées par les effets environnementaux (tels que les bruits de fond et les niveaux de réverbération). Cependant, contrairement aux mesures intrusives, les algorithmes non-intrusives dépendent du matériel utilisé lors des mesures, car ils n'ont pas accès à un signal de référence pour servir de facteur de normalisation. Par conséquent, les mesures non-intrusives communément présentent une plus grande variabilité dans leurs prédictions. La réduction de cette variabilité est une étape importante dans l'élaboration de mesures non-intrusives fiables qui ne sont pas sensibles qu'aux caractéristiques de l'environnement / traitement (par exemple, le niveau de la parole, le bruit de fond et acoustique de la pièce), mais pas sur le contenu du signal de parole lui-même.

Dans ce travail, nous étudions les performances des mesures objectives présentées dans l'état de l'art et comparons leurs performances avec les scores d'intelligibilité des utilisateurs des IC. Nous proposons une nouvelle mesure non-intrusive en affinant la mesure qu'on appelle SRMR pour émuler des caractéristiques de l'audition des utilisateurs des implants et réduire sa variabilité. Nous montrons que les mesures intrusives étudiées atteignent une performance fiable dans les distorsions de l'environnement, mais cette performance est dégradée en essayant d'évaluer l'intelligibilité de la parole traitée avec des algorithmes non-linéaires. La mesure proposée surpasse toutes les autres mesures de référence et obtient des résultats similaires à ceux obtenus par les mesures intrusives, mais avec l'avantage de ne pas nécessiter d'un signal de référence propre. La mesure proposée montre également une amélioration des performances avec des signaux de parole traités avec des algorithmes non-linéaires comparé à d'autres mesures existants.

Ce travail est organisé comme suit. Dans cette section, nous avons présenté le problème de l'évaluation de l'intelligibilité de la parole chez les utilisateurs des IC et décrit les objectifs du travail et les contributions. Dans la section B.2, nous discutons l'état de l'art sur le problème de l'évaluation objective de l'intelligibilité et de la qualité de la parole, l'importance des caractéristiques de l'enveloppe temporelle pour l'intelligibilité de la parole et comment ces caractéristiques peuvent être mesurées par une enquête sur une représentation spectrale de la modulation d'amplitude du signal. Dans la section B.3, nous présentons les améliorations proposées à la mesure SRMR qui visent à émuler des caractéristiques de l'audition des utilisateurs des implants et réduire sa variabilité liée au contenu de la parole. Nous procédons ensuite à une évaluation des performances de la mesure proposée et une comparaison avec des autres mesures intrusives et non intrusives de la qualité et de l'intelligibilité de la parole. Nous concluons finalement le travail avec un résumé des résultats et quelques considérations finales dans la section B.5.

B.2 Évaluation de l'intelligibilité de la parole

L'intelligibilité de la parole est une mesure de la quantité de perception d'information contenue dans un signal de parole traversant un système de transmission. Elle a été traditionnellement mesurée en présentant des mots degradés (soit par du bruit ou de la réverbération) pour l'identification à plusieurs niveaux de distorsion [24]. Les mots sont habituellement présentés en dehors du contexte de la phrase ou dans des phrases absurdes, afin d'éviter un biais provenant des prévisions de l'auditeur. Les mesures pour d'intelligibilité concernent généralement la quantité relative de mots reconnus (en %) sous différents niveaux de la condition. Cette procédure a quelques limitations, comme le biais de prédiction de l'auditeur déjà mentionné : les caractéristiques du signal et la réponse subjective sont pris en compte. Il est possible de remédier à ce problème en utilisant plusieurs mots / phrases par condition et, en fonction de l'application de la mesure, plusieurs auditeurs dans les mêmes conditions.

French and Steinberg [25] résument les facteurs qui influent sur l'intelligibilité de la parole en quatre types :

1. l'intensité de la parole reçue par l'oreille à chaque fréquence ;
2. les caractéristiques électriques / acoustiques des instruments (par exemple, un combiné ou appareil auditif) intervenant entre le locuteur et l'auditeur ;
3. les conditions dans lesquelles la communication a lieu (par exemple, le bruit de fond, le temps de réverbération) ;
4. le comportement du locuteur et de l'auditeur tel que modifié par les caractéristiques du système de communication et par les conditions dans lesquelles il est utilisé (par exemple, l'effort vocal et le réglage de la vitesse d'élocution).

Toutefois, la détermination de ces facteurs exige la connaissance complète du canal de transmission de la parole, du locuteur et des caractéristiques de l'auditeur. Comme cette information n'est pas toujours disponible, les mesures objectives d'intelligibilité considèrent surtout combien la perception des caractéristiques importantes des signaux de parole, comme caractéristiques spectrales et temporelles, sont conservées dans le signal reçu.

La perception d'un son peut être affectée par la présence d'un autre son dans un phénomène appelé masquage auditif [4]. Deux sons simultanés de fréquences différentes qui sont joués en même temps peuvent être perçus soit comme deux sons distincts ou un son combiné, en raison du filtrage qui se produit dans la cochlée. Le traitement cochléaire peut être modélisé comme un banc de filtres comprenant des filtres passe-bande dont les réponses en fréquence correspondent aux courbes d'accord des neurones auditifs. Chacune de ces bandes s'appelle une bande critique. Lorsque deux sons concurrents ont une différence de fréquences suffisamment petites pour être dans la même bande critique, le son avec l'énergie inférieure est masquée par le son d'énergie plus élevée, ce qui s'appelle masquage fréquentiel. Le bruit de fond est connue pour masquer les composantes énergétiques inférieurs de phonèmes, provoquant des confusions dans la détermination du point d'articulation, vocalisation, et entre les modes d'articulation occlusive et fricative. Pour mesurer l'effet du masquage fréquentiel sur l'intelligibilité de la parole, French and Steinberg [25] ont proposé l'utilisation de ce qu'on appelle l'indice d'articulation (AI), qui est une mesure du rapport signal sur bruit (RSB) entre des différentes bandes spectrales. Toutefois, cette mesure exige la connaissance de toutes les caractéristiques de l'environnement, du locuteur et de l'auditeur.

Plusieurs études démontrent l'importance des caractéristiques de l'enveloppe temporelle du signal de parole pour l'intelligibilité [30–33]. Les enveloppes temporelles de parole propre contiennent des fréquences allant de 2 à 20 Hz avec des pics spectraux à environ 4 Hz (correspondant au taux syllabique de la parole [32]). Les distorsions environnementales, tel que la réverbération, provoquent un étalement de l'enveloppe, du à l'ajout des dernières réflexions qui augmente la quantité d'énergie dans les hautes fréquences de modulation. Cet effet peut être vu dans la figure 2.2, qui montre le spectre de modulation moyenne d'enveloppe à la bande 1105 kHz d'un signal de parole non dégradé et des signaux réverbérants avec des temps de réverbération de 0,5 s et 1,5 s. Les énergies ont été calculées en utilisant le banc de filtres de modulation décrit plus loin dans ce section (Section B.2.2.4). Houtgast and Steeneken [34] ont proposé que des informations sur les fréquences de modulation dans l'enveloppe temporelle de la parole doivent être analysées comme une fonction de la fréquence pour les bandes d'octave de la parole. La méthode proposée, nommée indice de transmission de la parole (STI, *Speech Transmission Index*), requiert une connaissance complète de toutes les conditions de l'environnement (comme la méthode spectrale AI), mais il y a des propositions dans la littérature pour utiliser la mesure avec des signaux synthétiques d'une manière intrusive.

Dans les sections suivantes, nous présentons l'état de l'art de l'évaluation objective de la qualité et de l'intelligibilité de la parole. Certaines des mesures présentées ici, à savoir NCM, ModA, ANIQUE+ et SRMR, utilisent directement l'information de l'enveloppe temporelle, tandis que les autres utilisent principalement des informations spectrales. L'accent est mis sur la mesure SRMR, qu'est la base de la mesure proposée dans ce travail.

B.2.1 Mesures avec référence

B.2.1.1 Mesure de covariance normalisée (NCM)

La mesure NCM estime l'intelligibilité de la parole basée sur la covariance entre les enveloppes des signaux de parole non dégradés et dégradés [36–38]. Le calcul des valeurs NCM dépend de la dérivation d'enveloppes temporelles de la parole, par une transformée de Hilbert, pour chacun des 23 canaux de la banque de filtres, qui sont utilisés pour imiter le traitement cochléaire. La corrélation normalisée entre les enveloppes de parole non dégradé et dégradé produit une estimation de ce qu'on appelle RSB apparent. Ces valeurs sont limitées dans la gamme $[-15, 15]$ dB, mappées à la gamme $[0, 1]$ et sont ensuite ponderées à chaque canal de fréquence en fonction des poids de l'indice d'articulation (AI) et moyennées pour calculer la valeur finale de la mesure.

B.2.1.2 Indice de l'intelligibilité de la parole basé sur la cohérence (CSII)

Le CSII est une mesure de l'intelligibilité de la parole basée sur le spectre qui prend en compte la cohérence (ou similitude) des coefficients spectraux pour les signaux de parole dégradés et non dégradés [39, 40]. Afin de calculer les valeurs de CSII, une transformation de Fourier à court terme est d'abord effectuée de telle sorte que chaque segment temps-fréquence peut être pondéré par un paramètre appelé la cohérence d'amplitude au carré (*Magnitude Squared Coherence*, MSC). Ces valeurs de MSC sont groupés en 25 bandes de fréquence en utilisant des filtres passe-bande critiques et utilisés pour estimer le RSB pour chaque canal spectral et sa moyenne pondérée par les poids de l'indice d'articulation est calculé pour donner finalement la valeur de la mesure.

B.2.1.3 PESQ et oPESQ

PESQ est la recommandation P.862 de l'Union internationale des télécommunications (UIT-T) pour l'évaluation de la qualité vocale de la parole à bande étroite [42] avec des développements les plus récents permettant d'evaluer également la parole à large bande. L'algorithme est basé sur un modèle sensoriel qui regroupe deux des facteurs de distorsion liés : un facteur de perturbation (D_{ind}) et un facteur de perturbation asymétrique moyenne (A_{ind}). Ces facteurs sont estimés au moyen d'une comparaison des signaux non dégradés et dégradés. L'évaluation de la qualité finale est alors donnée par un mappage linéaire avec des coefficients optimisés en utilisant des données de téléphonie classiques (par exemple, la voix sur protocole Internet, sans fil). Les paramètres originaux de la mesure PESQ ont été obtenus en utilisant des signaux de parole dégradés par les canaux de transmission téléphonique et non par des réverbérations. Dans [11], ces paramètres ont été optimisés pour la voix réverbérée par analyse de régression linéaire multiple et données subjectives avec des auditeurs avec audition normale. La mesure « PESQ optimisée pour la parole réverbérée » est également explorée dans cette étude et est appelée oPESQ.

B.2.1.4 KLD

La divergence de Kullback-Leibler (KLD) estime la distance entre les fonctions de distribution de probabilité (*pdf*) des signaux de parole non dégradé et dégradé et s'est révélée comme une mesure objective de qualité fiable pour la parole réverbérée [11]. La motivation de la métrique réside dans le fait que l'étalement spectrale et temporelle produite par la réverbération rend le *pdf* de la parole réverbérée plus plate que celle de la parole propre. Le KLD est une mesure non négatif qui caractérise la similitude de distribution des valeurs tendant vers zéro lorsque les distributions sont similaires.

B.2.1.5 FWSSRR

Le rapport pondéré segmentaire de parole sur réverbération (*Frequency-Weighted Segmental Speech-to-Reverberation Ratio*) est obtenu par des estimations des rapports signal sur bruit pour chaque bande critique sur chaque fenêtre d'observation. La fonction de pondération AI est ensuite utilisée pour obtenir les poids de fréquence pour chaque bande critique. Plus de details sur la mesure FWSSRR peuvent être trouvés dans [40].

B.2.2 Mesures sans référence

B.2.2.1 P.563

En 2004, l'UIT-T a standardisé la première mesure non-intrusive de la qualité de la parole pour les applications vocales à bande téléphonique [49, 50]. L'algorithme estime la qualité du signal de parole testé basé sur trois principes. Tout d'abord, une analyse du conduit vocal et de prédiction linéaire est effectuée pour détecter les anormalités dans le signal de parole. En second lieu, un signal de pseudo-référence est reconstruit par la modification des coefficients de prédiction linéaire calculés pour le modèle du conduit vocal d'un locuteur humain typique. Le signal pseudo-référence sert d'entrée, avec le signal de parole dégradé, à un algorithme intrusif (similaire à l'UIT-T P.862 [51]) pour générer un indice de qualité de voix de base. Enfin, les distorsions spécifiques telles que le bruit, les coupures temporelles et les effets de robotisation (voix avec des bruits métalliques) sont caractérisées. Bien que le P.563 a été développé comme une mesure objective de la qualité pour les auditeurs avec une audition normale et les applications de téléphonie, une étude récente a montré des résultats prometteurs avec le P.563 comme un corrélat de l'intelligibilité de la parole vocodée pour des auditeurs avec une audition normale [45].

B.2.2.2 ModA

La mesure qu'on appelle l'aire sur le spectre de modulation (ModA) [43] est basée sur le principe selon lequel l'enveloppe du signal de parole est étalée pour les dernières réflexions dans une chambre réverbérante, affectant ainsi le spectre de modulation du signal de parole. Afin d'obtenir la valeur ModA d'un signal, le dernier est d'abord décomposé en N (= 4) bandes sonores, les enveloppes temporelles pour chaque bande sonore sont ensuite calculées en utilisant la transformée de Hilbert, puis sous-échantillonnées et groupées en utilisant un banc de filtres 1/3-octave avec des fréquences centrales comprises entre 0,5 et 8 Hz. Le spectre des enveloppes est calculé par 13 filtres de modulation qui couvrent la gamme de 0,5 à 10 Hz. Pour chaque bande de fréquence acoustique, l'aire sous le spectre de modulation est calculée et, enfin, les aires pour toutes les bandes acoustiques sont moyennées pour obtenir la mesure ModA.

B.2.2.3 ANIQUE+

ANIQUE+ [48] est un modèle perceptif qui emploie l'apprentissage statistique pour calculer des scores de qualité de la parole sur la base de trois modules de mesure de distorsion différents. Le module de distorsion muet détecte muets artificiels dans les signaux de parole et de quantifier leurs effets sur la qualité de la parole. Le module non-parole, à son tour, détecte et quantifie les effets des activités non vocalles gênantes, telles que celles résultant de l'insertion de bits erronés dans un décodeur de parole. Le module de distorsion de l'articulation utilise des concepts de la modulation spectrale similaires à ceux utilisés dans les mesures SRMR et ModA, mais au lieu d'utiliser les valeurs du spectre de modulation directement, il les fait correspondre à une valeur de distorsion par un perceptron multicouche. Les sorties des trois modules sont linéairement combinées pour générer un score de qualité globale.

B.2.2.4 SRMR

SRMR est une mesure non-intrusive qui a été récemment proposée et développée à l'origine pour évaluer la qualité et intelligibilité de la parole réverbérée et déreverbérée et testé avec des données subjectives des auditeurs avec une audition normale [20, 44]. Récemment, des résultats prometteurs ont également été signalés lors d'une évaluation en utilisant des signaux de parole vocodés pour simuler l'audition des utilisateurs des IC [45]. Le calcul de la métrique SRMR est réalisée en quatre étapes, comme suit. Tout d'abord, le signal est traité par un banc de filtres acoustiques qui émule le traitement cochléaire. La bande passante de chaque filtre est determinée par la largeur de bande rectangulaire équivalente (ERB, [46]). Des enveloppes temporelles sont calculés pour chaque sortie du banc de filtres par une transformée de Hilbert et ces enveloppes sont fenêtrées (période de 256 ms avec un pas de 64 ms). Une transformée de Fourier discrète est utilisée pour obtenir ce qu'on appelle les énergies de modulation spectrale pour chaque bande critique. Ces énergies sont groupées en 8 bandes superposées avec les fréquences centrales espacées de façon logarithmique entre 4 et 128 Hz. Enfin, la valeur de SRMR est calculée comme le rapport de l'énergie de modulation moyenne dans les quatre premières bandes de modulation (environ 3 à 20 Hz, en accord avec le contenu de la modulation de la parole) à la teneur en énergie moyenne dans les quatre derniers bandes de modulation (environ 20 – 160 Hz).

B.3 Mesure inspirée par les IC

Dans cette section, nous proposons trois modifications à la métrique SRMR originale. La première modification vise à émuler des préceptes de l'audition des utilisateurs des IC en remplaçant le banc de filtres acoustiques par un banc de filtres inspiré par celui utilisé dans un implant. Deux autres modifications sont proposées pour réduire l'impact du locuteur et du contenu de la parole sur la variabilité de la mesure SRMR. Nous avons effectué des expériences en utilisant deux bases de données différentes, l'une composée de paires de consonne-voyelle et l'autre de phrases complètes, composées des échantillons de différents locuteurs en condition anéchoïque, afin de comprendre comment les facteurs du contenu de la parole influent sur la représentation du spectre de modulation. Basé sur les résultats, nous proposons l'utilisation d'un système de seuil d'énergies de modulation et une gamme plus étroite de filtres de modulation pour atténuer la variabilité liée au contenu de la parole et de la fréquence fondamentale.

B.3.1 Le banc de filtres acoustique inspiré par les IC

L banc de filtres utilisé par la SRMR a été conçu pour émuler le traitement cochléaire pour les auditeurs avec une audition normale, donc il ne reflète pas nécessairement des caractéristiques de l'audition des utilisateurs d'un IC. Comme leur traitement cochléaire est compromise, les utilisateurs des IC s'appuient sur le processeur vocal pour coder des informations audio en sous-bandes. Par conséquent, nous avons mis à jour la banque de filtres acoustiques pour simuler les caractéristiques des appareils IC. Notre approche consistait à imiter le banc de filtres utilisé dans les stratégies de codage de la parole ACE (codeur de combinaison avancé) et CIS (échantillonage entrelacé continu), qui utilisent un espacement similaire à une échelle de Mel entre les bandes au lieu de l'espacement ERB [53]. La Figure 3.1 montre une comparaison entre le banc de filtres original utilisé par SRMR (à 8 kHz et à 16 kHz) et le banc de filtres inspiré de celui utilisée dans Nucleus. A partir de maintenant, nous appelons la version de SRMR qui utilise le banc de filtres inspiré par les IC par SRMR-CI.

B.3.2 Sources de variabilité dans le spectre de modulation

Les caractéristiques qui représentent des informations temporelles de la parole, comme la structure de l'enveloppe temporelle, sont utiles pour les systèmes de reconnaissance automatique de la parole [54], ce qui montre qu'ils sont sensibles au contenu de la parole. Bien que la SRMR utilise des fenêtres longues et des moyennes de plusieurs fenêtres, cette mesure a été démontré une forte variabilité inter-locuteur [55, 56]. Les expériences décrites ici montrent également que la mesure SRMR-CI est sensible à différents phonèmes et à la fréquence fondamentale (F_0).

B.3.2.1 Expériences avec des paires de consonne-voyelle et des phrases

Pour étudier l'effet de la variabilité de la parole sur la mesure SRMR-CI, nous avons traité des données de parole à partir de deux bases de données différentes. La première base de données est composée de paires de consonne-voyelle (CV), et contenait 1.728 échantillons [57]. Quatre locuteurs (2 hommes et 2 femmes) ont enregistré 8 jetons pour chacun des 18 consonnes dans 3 contextes vocaliques. Pour évaluer la variabilité de SRMR-CI avec des phrases, nous avons utilisé un sous-ensemble du corpus TIMIT constitué de 160 enregistrements de phrases anéchoïques et sans bruit produites par 8 hommes et 8 femmes de langue maternelle anglaise [58]. Chaque locuteur a enregistré 10 phrases phonétiquement riches. Nous avons calculé la mesure SRMR-CI pour toutes les paires de CV et des phrases. Dans le cas des paires de CV, les échantillons ont été regroupés par deux catégories différentes : par voyelle et par mode d'articulation.

Résultats pour les paires CV (Figure 3.2) indiquent que des différents moyens pour les valeurs de SRMR-CI peuvent être identifiés pour des groupes CV séparés par voyelle (ANOVA à sens unique pour chaque paire de voyelles, $p < 0,001$) et, dans le cas des modes d'articulation, au moins l'un des groupes (nasales) a une moyenne différente des restants (ANOVA à sens unique pour tous les groupes de mode d'articulation, $p < 0,001$). Considérant tous les échantillons CV ensemble (toutes les paires de CV provenant des quatre locuteurs), la valeur absolue du coefficient de variation (en anglais, *relative standard deviation*, RSD) trouvée était de 58%.

D'autre part, les valeurs SRMR-CI pour les échantillons de phrases avaient un RSD de 47% ; sa distribution est représentée par le diagramme à gauche dans la figure 3.3. En outre, nous avons regroupé les phrases selon le sexe du locuteur afin d'évaluer les effets de la fréquence de hauteur sur la variabilité de la mesure SRMR-CI. La figure 3.4 montre les distributions des valeurs de SRMR-CI pour les phrases prononcées par les hommes ($F_0 = 151,4 \pm 29.7$ Hz) et par les femmes ($F_0 = 211,6 \pm 31,6$ Hz), où une différence significative (ANOVA, $p < 0,001$) de scores moyens de SRMR-CI a été observée selon le sexe du locuteur. Pour confirmer que ce résultat a été causé par une différence de fréquence de hauteur, nous avons calculé la fréquence fondamentale moyenne (F_0) pour toutes les phrases et calculé sa corrélation avec la valeur SRMR-CI ; une valeur de 0,76 a été trouvée, montrant une forte corrélation entre la hauteur et les proportions de l'énergie de parole et réverbération.

B.3.2.2 Système de seuil d'énergies de modulation

Dans [54], les auteurs proposent une méthode de seuillage de l'énergie pour le spectrogramme de modulation. Ils utilisent la valeur de crête du spectre de modulation comme une référence et mappent toutes les valeurs de plus de 30 dB en dessous de ce sommet global à -30 dB. En faisant cela, la visualisation se concentre sur les énergies supérieures du spectrogramme en tronquant des valeurs extrêmement faibles. Comme nous nous sommes également intéressés à la relation entre les énergies dans différentes parties du spectre et non sur des valeurs absolues, nous avons utilisé une procédure similaire afin de réduire la variabilité de la mesure SRMR-CI. Comme nous voulons minimiser l'effet des phonèmes qui se traduisent par des ratios plus élevés que la moyenne, nous avons utilisée la valeur de crête de la fenêtre moyenne au lieu de la valeur de crête global.

En utilisant les valeurs seuillées à la place des valeurs réelles d'énergie pour calculer SRMR-CI, la variation entre les différents locuteurs et les différents enregistrements de la même CV par le même locuteur sont réduites, comme représenté sur les figures 3.5b/3.5d et 3.6b/3.6d : 10 % et 4 %, respectivement (rappelons que 29% et 21% ont été observées avec la mesure original). Les effets de seuillage des énergies de modulation sur les distributions des valeurs de SRMR-CI pour les paires CV groupés par voyelle et mode d'articulation sont présentés dans les figures 3.7a et 3.7b, respectivement. Les RSDs pour les différents groupes dans ce cas se situaient entre 30% et 43% (contre 47% - 61% qui ont été trouvés avec la mesure originale). La comparaison directe entre les figures 3.7 et 3.2 montre que la gamme des valeurs et la variabilité absolue ont été réduites.

B.3.2.3 Rétrécissement de la gamme d'analyse des fréquences de modulation

La gamme de fréquences de modulation entre 4-128 Hz s'est avérée utile pour détecter les effets de réverbération du signal de parole [20]. Comme nos expériences avec des phrases ont indiqué, cependant, les niveaux d'énergie dans cette gamme sont fortement corrélés avec la fréquence fondamentale de la parole. En effet, l'enveloppe de l'énergie du signal de parole a une structure dont la périodicité est égale à sa fréquence fondamentale [60]. Bien que les résultats dans [60] sont basés sur des enveloppes en « large bande », nous explorons ici si ce comportement est aussi observé dans les sous-bandes acoustiques. À cette fin, nous avons estimé la densité spectrale des enveloppes de sous-bande (en utilisant le banc de filtres acoustiques inspiré par les IC) en calculant leurs périodogrammes pour chacun des enregistrements dans notre base de données de phrases. Les enveloppes sont sous-échantillonnées à une fréquence d'échantillonnage de 600 Hz de manière à être capable de détecter l'effet des hauteurs jusqu'à 300 Hz. Ensuite, pour chacun de ces périodogrammes, nous avons calculé la fréquence correspondant à la premiere crête d'énergie après CC. La corrélation entre cette fréquence et la hauteur moyenne était supérieure à 0,8 pour 13 des 22 sous-bandes, et supérieure à 0,6 pour 18 des 22 sous-bandes. Ces résultats montrent que pour la plupart des enveloppes des sous-bandes acoustiques, la densité d'énergie est plus élevée autour de la fréquence fondamentale du signal de parole original. Cela explique pourquoi les signaux de parole à hauteurs inférieurs (par exemple, les signaux de parole des hommes comparés à ceux des femmes) ont abouti à des valeurs SRMR-CI inférieures. La mesure SRMR-CI analyse une gamme spectrale de modulation allant jusqu'à 128 Hz, celle qui sera affectée par la modulation des crêtes spectrales provoquées principalement par la fréquence fondamentale du signal de parole.

Afin d'éviter cet effet secondaire, nous avons expérimenté l'utilisation des bandes plus étroites de fréquences de modulation. Nous avons testé les valeurs suivantes pour la fréquence centrale du dernier filtre d'analyse ($CF8$) : 30, 40, 45, 50, 55 et 64 Hz. Toutes les autres fréquences centrales ont été ajustées de telle sorte que l'espacement logarithmique est maintenu comme dans le banc de filtres d'origine. Les résultats (résumées dans le tableau 3.2) ont montré que pour $CF8 \geq 40$ Hz, la dernière bande de modulation a une corrélation négative avec F_0. Cette relation inverse s'étend aux sixième et septième bandes si on utilise des gammes plus larges. En réduisant la bande passante des fréquences de modulation, la corrélation entre la hauteur et SRMR-CI diminue. Nos résultats montrent que l'utilisation des fréquences de modulation comprises en dessous de 55 Hz sont utiles pour décorréler F_0 et SRMR-CI. Dans la section suivante, nous allons évaluer les effets des systèmes de normalisation proposés sur la prédiction de l'intelligibilité de la parole.

B.3.3 Mesure normalisée avec référence

A titre de comparaison avec d'autres mesures avec référence, nous proposons aussi une métrique sans référence basée sur SRMR-CI. Cette mesure est calculée en utilisant la valeur de SRMR-CI du signal de référence pour normaliser la valeur du signal dégradé, comme suit :

$$\text{SRMR-CI}_{\text{intr}} = \frac{\text{SRMR-CI}}{\text{SRMR-CI}_{\text{clean}}}, \tag{B.1}$$

où $\text{SRMR-CI}_{\text{clean}}$ est la valeur SRMR-CI du signal de référence. Même si les avantages d'utiliser SRMR-CI sont ses caractéristiques non intrusives, nous voudrions étudier si le résultat du fichier de référence sert comme un facteur de normalisation afin de réduire la variabilité et comparer cette normalisation avec les stratégies proposées précédemment dans ce chapitre.

B.4 Résultats expérimentaux

Dans cette section, nous présentons les résultats d'expériences réalisés pour évaluer les performances des différentes implémentations SRMR-CI et les mesures de référence présentées au chapitre 2. Nous comparons les scores des métriques objectives d'intelligibilité de la parole avec les scores d'un test d'écoute subjective réalisé avec les utilisateurs IC [9, 61]. Les expériences ont deux objectifs : d'abord, nous voulons évaluer si la mesure proposée, SRMR-CI, conduit à une meilleure corrélation avec les scores subjectives d'intelligibilité comparativement à la mesure SRMR originale et quelle gamme de banc de filtres de modulation conduit à des résultats optimaux. Deuxièmement, nous voulons comparer les performances de la mesure proposée avec celui des mesures de la qualité et d'intelligibilité de la parole existantes.

B.4.1 Montage expérimental et critères de performance

B.4.1.1 Base de donnés du test d'écoute subjective avec des utilisateurs d'IC

Onze adultes utilisateurs d'IC ont été recrutés pour participer à des expériences subjectives d'intelligibilité. Les participants étaient tous des locuteurs natifs de l'anglais américain avec une surdité post-linguale et avaient un âge moyen de 64 ans (±8.9). Le lecteur intéressé se reportera à la référence [9] pour plus de détails démographiques spécifiques des participants. Tous les participants avaient une expérience d'un an minimum d'utilisation régulière de leurs implants, la majorité étant implantée bilatéralement pendant plus de 6 ans. Tous les participants ont utilisé des dispositifs développés par Cochlear Ltd. ; par souci de cohérence, ils ont été temporairement équipés d'un processeur de recherche SPEAR3, programmé avec la stratégie ACE [62] avec les paramètres correspondant à des paramètres cliniques de l'utilisateur.

Les signaux présentés aux participants sont de la base de données IEEE [63], qui contient des phrases avec 7 à 12 mots, organisées dans 72 listes de 10 phrases chacune. Les phrases ont été produites par un locuteur masculin et enregistrées dans des conditions anéchoïques. Les phrases ont été égalisées à la même valeur quadratique moyenne de 65 dB. La fréquence d'échantillonnage utilisée pour l'enregistrement est de 25 kHz et les fichiers étaient sous-échantillonné à 16 kHz pour cette expérience. Les effets de réverbération et le bruit additif ont été introduits par simulation numérique. Quatre réponses impulsionelles obtenues expérimentalement [64, 65] ont eté convoluées avec des signaux de parole propre pour générer des signaux de parole réverbérants avec des temps de réverbération de 0,3, 0,6, 0,8 et 1 s. Le bruit avec un spectre fréquentiel similaire à la parole a été ajouté aux signaux de parole sans et avec réverbération pour générer des conditions « bruit seulement » (RSB de -5, 0, 5 et 10 dB) et « bruit plus réverbération « (RSB de

5 et 10 dB). Les participants ont été testés aussi en utilisant des phrases améliorées en utilisant une stratégie de masquage réverbérant idéal (IRM) décrite dans [61]. Ces phrases ont été effectuées dans des conditions de réverbération avec des temps de réverbération de 0,6 s, 0,8 s, 1,0 s et toutes les conditions de « bruit plus réverbération » décrits ci-dessus. L'algorithme IRM a été configuré pour utiliser soit 2 ou 3 seuils différents pour chaque condition. Les auditeurs ont été invités à répéter tous les mots identifiables et les scores d'intelligibilité par participant ont été calculés comme le rapport entre le nombre de mots correctement identifiés au nombre total de mots présentés.

B.4.1.2 Critères de performance

Afin de réduire la variabilité inter-et intra-sujet, nous avons utilisé les moyennes des scores d'intelligibilité trouvés par chaque condition lors des tests d'écoute subjectifs comme les valeurs réelles. Les moyennes ont été calculées pour l'ensemble des 20 phrases pour chaque participant sur chaque condition et les résultats ont ensuite été moyennés sur l'ensemble des participants. Nous avons évalué les performances dans deux cas différents : en utilisant des échantillons avec des distorsions environnementales seulement et en utilisant tous les échantillons (en ajoutant des fichiers améliorés par l'algorithme IRM).

Pour évaluer la performance des mesures, nous avons utilisé quatre critères de performance différents : le coefficient de Pearson de corrélation (ρ), le coefficient de corrélation de Spearman (ρ_{spear}), le coefficient de corrélation de Pearson d'un mappage de fonction sigmoïde entre les scores objectifs et l'échelle d'intelligibilité et la racine carrée de l'erreur quadratique moyenne (RMSE). Pour évaluer la variabilité de chaque mesure sur chaque condition, nous avons utilisé la valeur absolue du coefficient de variation (RSD%) moyennée sur toutes les conditions.

B.4.2 Résultats et discussion

Les résultats pour les conditions avec des distorsions environnementales sont résumés dans le tableau 4.1. Les mesures NCM et P.563 ont obtenu les meilleures performances parmi les mesures intrusives et non-intrusives, respectivement. L'utilisation du banc de filtres inspiré par les IC a amélioré les corrélations de la mesure SRMR de 0.02 à 0.04. Le réglage de la gamme de fréquences de modulation a conduit à des améliorations supplémentaires de la performance, tout en réduisant considérablement RMSE et RSD%, atteignant des valeurs de corrélations optimales pour les gammes 4-40 Hz et de 4 - 30 Hz. Le seuillage des énergies de modulation, à son tour, a réduit de façon significative RMSE et RSD % (avec des réductions relatives d'environ 50%), même après les améliorations observées avec les bandes plus étroites de fréquences de modulation.

Les résultats des tests incluant les conditions améliorées par l'algorithme IRM, à leur tour, sont présentés dans le tableau 4.2. Par rapport aux résultats précédemment présentés, les corrélations étaient plus faibles pour toutes les mesures. Les mesures intrusives ont montré un problème de rang avec les fichiers améliorés, comme leurs scores étaient similaires à ceux trouvés pour les conditions avec les scores subjectifs inférieurs. De toutes les mesures intrusives, les corrélations les plus élevées ont été trouvées pour NCM et KLD ; NCM, cependant, montre moins de variabilité par condition. Les mesures de référence non-intrusives n'ont pas eu une diminution de performance aussi importante que celle constatée pour les mesures intrusives. Les mesures SRMR-CI ont abouti à des corrélations plus élevées pour ce test. Les résultats ont suivi une tendance similaire à celle observée pour les fichiers non améliorées, mais nous avons observé des gains de performance plus élevés. Par rapport à la mesure SRMR originale, la mesure proposée affiche un gain de 0,40 en ρ, 0,25 en ρ_{spear}, 0,34 en ρ_{sig}, et une réduction relative de 50% de RMSE et RSD%. Aucune différence significative n'a été trouvée entre la mesure SRMR-CI-intr et son homologue non-intrusive, suggérant ainsi que la mesure SRMR-CI ne bénéficie pas de la « normalisation » supplémentaire fournie par l'utilisation des scores SRMR-CI à partir des signaux propres.

Nos expériences ont montré que des mesures basées sur des indices temporels (NCM, ModA et SRMR-CI) ont mieux réussi que d'autres mesures pour les utilisateurs d'IC. Ces constatations corroborent celles précédemment rapportées dans la littérature montrant des prédictions fiables de l'intelligibilité de la parole vocodée avec auditeurs avec une audition normale pour des mesures basées en caractéristiques d'enveloppe temporelle [45]. Enfin, on sait aussi que la réverbération modifie les indices d'enveloppe temporelle, ce qui dégrade sévèrement la reconnaissance vocale pour les utilisateurs d'IC.

B.5 Conclusion

Dans ce travail, nous avons proposé une mesure de l'intelligibilité de la parole adaptée pour les utilisateurs des implantes cochléaires, appelé SRMR-CI. Cette mesure est basée sur SRMR, une mesure qui a des hautes corrélations avec la qualité et l'intelligibilité de la parole pour les auditeurs avec une audition normale, ainsi qu'avec les caractéristiques acoustiques telles que le temps de réverbération.

Notre proposition étend SRMR pour améliorer sa corrélation avec l'intelligibilité de la parole des utilisateurs des implants cochléaires basée sur trois principales modifications. Tout d'abord, nous avons remplacé le banc de filtres acoustiques utilisé dans la mesure SRMR par un banc de filtres similaire à celui utilisé dans la stratégie de codage de la parole utilisée dans les dispositifs Nucleus. Cette modification a conduit à des améliorations de 21% en corrélation linéaire, 6% en corrélation de rang, et 14% en corrélation d'un mappage sigmoïdale entre les scores objectifs et les scores subjectifs, comparativement à la mesure originale. La deuxième modification est un ajustement de la gamme de fréquences de modulation utilisée par la mesure pour calculer les énergies dans le spectre de modulation, afin de réduire la sensibilité de la mesure à la fréquence fondamentale de la parole. Enfin, nous proposons un schéma de seuillage des énergies de modulation pour tenir compte de la variabilité liée au contenu de la parole.

Par rapport aux mesures intrusives et non-intrusives de l'état de l'art, notre mesure a montré des meilleures performances, notamment lors de l'évaluation de l'intelligibilité de la parole traitée avec des algorithmes non-linéaires. Comme la plupart, sinon la totalité, des dispositifs IC actuellement sur le marché utilisent des techniques avancées de traitement de la parole pour améliorer l'intelligibilité pour les utilisateurs, une mesure objective de l'intelligibilité de la parole pour les utilisateurs d'IC doit être capable de prendre en compte l'effet des distorsions non-linéaires. Une mesure non-intrusive comme SRMR-CI est une ressource précieuse pour le développement et la validation de nouveaux dispositifs et des algorithmes d'amélioration et peuvent potentiellement être utilisée dans des dispositifs pour ajuster les paramètres du système en temps réel.

Liste des Figures

Liste des Tableaux

Bibliography

[1] Canadian Hard of Hearing Association. Frequently Asked Questions About Hearing Loss (booklet), 2013. URL http://www.chha.ca/documents/en/faq_about_hearing_loss_booklet.pdf.

[2] H. Bergman, P. Gaudreau, and Y. Joanette. Quebec Network for Research on Aging: Scientific Program 2008-2012, 2008. URL http://www.rqrv.com/en/document/RQRV_Scientific_program_2008-2012.pdf.

[3] W. J. Millar. Hearing problems among seniors. *Health Reports – Statistics Canada*, 16(4):49–52, June 2005.

[4] D. O'Shaughnessy. *Speech Communications: Human and Machine*. Wiley-IEEE Press, 2nd edition, November 1999.

[5] P. Loizou. Mimicking the human ear. *IEEE Signal Processing Magazine*, 15 (5):101–130, September 1998.

[6] B. S. Wilson and M. F. Dorman. Cochlear implants: current designs and future possibilities. *Journal of Rehabilitation Research and Development*, 45 (5):695–730, 2008.

[7] B. S. Wilson and M. F. Dorman. Cochlear implants: a remarkable past and a brilliant future. *Hearing Research*, 242(1):3–21, 2008.

[8] A. Nabelek. Effects of room acoustics on speech perception through hearing aids by normal hearing and hearing impaired listeners. In G. Stubebaker and

I. Hochberg, editors, *Acoustical Factors Affecting Hearing Aid Performance*, pages 15–28. Allyn and Bacon, Needham Heights, USA, 1993.

[9] O. Hazrati and P. C. Loizou. The combined effects of reverberation and noise on speech intelligibility by cochlear implant listeners. *Int. J. Audiol.*, 51(6): 437–443, February 2012.

[10] S. Drgas and M. A. Blaszak. Perception of speech in reverberant conditions using AM-FM cochlear implant simulation. *Hearing Research*, 269(1-2):162–168, 2010.

[11] K. Kokkinakis and P. C. Loizou. Evaluation of objective measures for quality assessment of reverberant speech. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2420–2423, 2011.

[12] A. Watkins and N. Holt. Effects of a complex reflection on vowel identification. *Acust.Acta Acust*, 86:532–542, 2000.

[13] A. Nabelek, T. Letowski, and F. Tucker. Reverberant overlap- and self-masking in consonant identification. *The Journal of the Acoustical Society of America*, 86:318–326, 1989.

[14] S. Poissant, N. Whitmal, and R. Freyman. Effects of reverberation and masking on speech intelligibility in cochlear implant simulations. *The Journal of the Acoustical Society of America*, 119(3):1606–1615, 2006.

[15] Y. Zheng, J. Koehnke, J. Besing, and J. Spitzer. Effects of noise and reverberation on virtual sound localization for listeners with bilateral cochlear implants. *Ear and Hearing*, 32(5):569, 2011.

[16] K. Kokkinakis and P. C. Loizou. The impact of reverberant self-masking and overlap-masking effects on speech intelligibility by cochlear implant listeners (L). *The Journal of the Acoustical Society of America*, 130:1099, 2011.

[17] K. Kokkinakis, O. Hazrati, and P. C. Loizou. A channel-selection criterion for supressing reverberation in cochlear implants. *The Journal of the Acoustical Society of America*, 129(5):3221–3232, 2011.

[18] P. C. Loizou, A. Lobo, and Y. Hu. Subspace algorithms for noise reduction in cochlear implants. *The Journal of the Acoustical Society of America*, 118: 2791, 2005.

[19] S. Moller, W. Y. Chan, N. Cote, T. H. Falk, A. Raake, and M. Waltermann. Speech quality estimation: Models and trends. *Signal Processing Magazine, IEEE*, 28(6):18–28, 2011.

[20] T. H. Falk, C. Zheng, and W.-Y. Chan. A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(7):1766–1774, September 2010.

[21] J. F. Santos, S. Cosentino, O. Hazrati, P. C. Loizou, and T. H. Falk. Performance comparison of intrusive objective speech intelligibility and quality metrics for cochlear implant users. In *Proc. InterSpeech*, pages 1724–1727, Portland, Oregon, USA, 2012.

[22] J. F. Santos, S. Cosentino, O. Hazrati, P. C. Loizou, and T. H. Falk. Objective Speech Intelligibility Measurement for Cochlear Implant Users in Complex Listening Environments. *Speech Communication*, 55(7-8):815–824, September 2013.

[23] J. F. Santos and T. H. Falk. Normalization of the SRMR-CI Metric for Improved Intelligibility Prediction for Cochlear Implant Users. *IEEE Transactions on Audio, Speech, and Language Processing*, October 2013 (submitted).

[24] D. B. Pisoni. Word identification in noise. *Language and cognitive processes*, 11(6):681–687, 1996.

[25] 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.

[26] K. D. Kryter. Methods for the calculation and use of the articulation index. *The Journal of the Acoustical Society of America*, 34(11):1689–1697, 1962.

[27] B. Boashash. Estimating and interpreting the instantaneous frequency of a signal. i. fundamentals. *Proceedings of the IEEE*, 80(4):520–538, 1992.

[28] C. E. Schreiner and J. V. Urbas. Representation of amplitude modulation in the auditory cortex of the cat. i. the anterior auditory field (AAF). *Hearing Research*, 21(3):227–241, 1986.

[29] R. Plomp. The role of modulation in hearing. In D. R. Klinke and D. R. Hartmann, editors, *HEARING - Physiological Bases and Psychophysics*, pages 270–276. Springer Berlin Heidelberg, January 1983.

[30] R. Drullman, J. M. Festen, and R. Plomp. Effect of temporal envelope smearing on speech reception. *The Journal of the Acoustical Society of America*, 95:1053, 1994.

[31] R. Drullman, J. M. Festen, and R. Plomp. Effect of reducing slow temporal modulations on speech reception. *The Journal of the Acoustical Society of America*, 95:2670, 1994.

[32] T. Arai, M. Pavel, H. Hermansky, and C. Avendano. Intelligibility of speech with filtered time trajectories of spectral envelopes. In *Proc. Fourth International Conference on Spoken Language (ICSLP)*, volume 4, pages 2490 – 2493, 1996.

[33] R. Drullman. Temporal envelope and fine structure cues for speech intelligibility. *The Journal of the Acoustical Society of America*, 97(1):585–592, 1995.

[34] T. Houtgast and H. J. M. Steeneken. A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria. *The Journal of the Acoustical Society of America*, 77(3):1069–1077, 1985.

[35] K. Kondo. *Subjective quality measurement of speech its evaluation, estimation and applications*. Springer, Berlin; New York, 2012.

[36] F. Chen and P. C. Loizou. Predicting the intelligibility of vocoded speech. *Ear and Hearing*, 32(3):331–338, 2011.

[37] R. L. Goldsworthy and J. E. Greenberg. Analysis of speech-based speech transmission index methods with implications for nonlinear operations. *The Journal of the Acoustical Society of America*, 116(6):3679, 2004.

[38] I. Holube and B. Kollmeier. Speech intelligibility prediction in hearing-impaired listeners based on a psychoacoustically motivated perception model. *The Journal of the Acoustical Society of America*, 100(3):1703–1716, 1996.

[39] J. Kates and K. Arehart. A model of speech intelligibility and quality in hearing aids. In *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 53–56. IEEE, 2005.

[40] J. Ma, Y. Hu, and P. C. Loizou. Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions. *The Journal of the Acoustical Society of America*, 125(5):3387–3405, 2009.

[41] B. C. J. Moore and B. R. Glasberg. A revision of Zwicker's loudness model. *Acta Acustica united with Acustica*, 82(2):335–345, 1996.

[42] ITU-T P.862. Perceptual evaluation of speech quality: An objective method for end-to-end speech quality assessment of narrow-band telephone network and speech coders, 2001.

[43] F. Chen, O. Hazrati, and P. C. Loizou. Predicting the intelligibility of reverberant speech for cochlear implant listeners with a non-intrusive intelligibility measure. *Biomedical Signal Processing and Control*, 8(3):311–314, 2012.

[44] T. H. Falk and W.-Y. Chan. Modulation spectral features for robust far-field speaker identification. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(1):90–100, January 2010.

[45] S. Cosentino, T. Marquardt, D. McAlpine, and T. H. Falk. Towards objective measures of speech intelligibility for cochlear implant users in reverberant environments. In *Proc. Intl Conf Information Science, Signal Process and Applications*, pages 4710–4713, Montreal, Canada, 2012.

[46] B. R. Glasberg and B. C. Moore. Derivation of auditory filter shapes from notched-noise data. *Hearing Research*, 47(1-2):103–138, 1990.

[47] S. D. Ewert and T. Dau. Characterizing frequency selectivity for envelope fluctuations. *The Journal of the Acoustical Society of America*, 108:1181, 2000.

[48] ATIS-PP-0100005.2006. Auditory non-intrusive quality estimation plus (ANIQUE+): Perceptual model for non-intrusive estimation of narrowband speech quality. Technical report, American National Standards Institute, 2006.

[49] L. Malfait, J. Berger, and M. Kastner. P.563: The ITU-T standard for single-ended speech quality assessment. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(6):1924–1934, November 2006.

[50] ITU-T P.563. Single ended method for objective speech quality assessment in narrow-band telephony applications. Technical report, Intl. Telecom Union, 2004.

[51] ITU-T P.862. Perceptual evaluation of speech quality: An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs. Technical report, Intl. Telecom Union, 2001.

[52] R. D. Patterson, I. Nimmo-Smith, J. Holdsworth, and P. Rice. An efficient auditory filterbank based on the gammatone function. In *Proc. Meeting of the IOC Speech Group on Auditory Modelling at RSRE*, volume 2, 1987.

[53] P. Loizou. Speech processing in vocoder-centric cochlear implants. In A. Moller, editor, *Cochlear and Brainstem Implants*. Karger Publishers, 2006.

[54] B. E. Kingsbury, N. Morgan, and S. Greenberg. Robust speech recognition using the modulation spectrogram. *Speech Communication*, 25(1):117–132, 1998.

[55] N. D. Gaubitch, H. W. Loellmann, M. Jeub, T. H. Falk, P. A. Naylor, P. Vary, and M. Brookes. Performance comparison of algorithms for blind reverberation time estimation from speech. In *Proc. International Workshop on Acoustic Signal Enhancement (IWAENC)*, pages 1–4, 2012.

[56] J. Schröder, T. Rohdenburg, V. Hohmann, and S. D. Ewert. Classification of reverberant acoustic situations. In *Proceedings of the International Conference on Acoustics NAG/DAGA*, pages 606–609, 2009.

[57] UCLA Speech Processing and Auditory Perception Laboratory. Consonant vowel tokens CV database, 2013. URL http://www.seas.ucla.edu/spapl/cv.html.

[58] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, and N. Dahlgren. Timit: acoustic-phonetic continuous speech corpus. Technical report, Linguistic Data Consortium, 1993. URL http://catalog.ldc.upenn.edu/LDC93S1.

[59] S. Gonzalez and M. Brookes. A pitch estimation filter robust to high levels of noise (PEFAC). In *Proc. European Signal Processing Conference (EUSIPCO)*, 2011.

[60] H. Quast, O. Schreiner, and M. R. Schroeder. Robust pitch tracking in the car environment. In *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 353–356. IEEE, 2002.

[61] O. Hazrati and P. C. Loizou. Tackling the combined effects of reverberation and masking noise using ideal channel selection. *Journal of Speech, Language, and Hearing Research*, 55(2):500–510, April 2012.

[62] A. E. Vandali, L. A. Whitford, K. L. Plant, and G. M. Clark. Speech perception as a function of electrical stimulation rate: using the Nucleus 24 cochlear implant system. *Ear and Hearing*, 21(6):608–624, December 2000.

[63] 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(3):225–246, 1969.

[64] A. C. Neuman, M. Wroblewski, J. Hajicek, and A. Rubinstein. Combined effects of noise and reverberation on speech recognition performance of normal-hearing children and adults. *Ear and Hearing*, 31(3):336–344, June 2010.

[65] T. Van Den Bogaert, S. Doclo, J. Wouters, and M. Moonen. Speech enhancement with multichannel wiener filter techniques in multimicrophone binaural hearing aids. *The Journal of the Acoustical Society of America*, 125(1):360–371, 2009.

[66] R. Plomp. A signal-to-noise ratio model for the speech-reception threshold of the hearing impaired. *Journal of Speech and Hearing Research*, 29(2):146, 1986.

[67] H. Dudley. Remaking speech. *The Journal of the Acoustical Society of America*, 11(2):169, 1939.

[68] C. Lorenzi and B. C. J. Moore. Role of temporal envelope and fine structure cues in speech perception: A review. In *Proc. International Symposium on Auditory and Audiological Research (ISAAR)*, pages 263–272, 2008.

[69] B. C. J. Moore. The role of temporal fine structure processing in pitch perception, masking, and speech perception for normal-hearing and hearing-impaired people. *JARO: Journal of the Association for Research in Otolaryngology*, 9(4):399–406, December 2008.

[70] L. Xu and B. Pfingst. Spectral and temporal cues for speech recognition: Implications for auditory prostheses. *Hearing Research*, 242:132–140, 2008.

[71] J. M. Kates and K. H. Arehart. The hearing-aid speech quality index (hasqi). *Journal of the Audio Engineering Society*, 58(5):363–381, 2010.