Combining spectral and temporal modification techniques for speech intelligibility enhancement

Martin Cooke[a,b,*], Vincent Aubanel[c], María Luisa García Lecumberri[b]

[a] *Ikerbasque (Basque Science Foundation)*
[b] *Language and Speech Laboratory, Universidad del País Vasco, 01006 Vitoria, Spain*
[c] *University of Grenoble Alpes, Centre National de la Recherche Scientifique, GIPSA-lab, Grenoble, France*

Abstract

Modifying clean speech prior to output in noisy conditions can lead to substantial intelligibility gains. Most algorithms operate by redistributing energy across the signal, leaving the timing of the underlying speech sounds intact. Other techniques do alter the timing of speech relative to the masker. Both classes of approach – spectral and temporal – lead to a reduction in energetic masking. The current study examines how their combination affects intelligibility. Arguments can be made for both synergy and redundancy, and the presence of distortions introduced by both spectral and temporal approaches might even lead to an antagonistic combination. A cohort of native Spanish listeners identified keywords in sentences in unmodified form and following spectral, temporal and spectro-temporal modification, in the presence of a fluctuating masker. Errors in the spectro-temporal condition were substantially lower than following spectral or temporal modification alone, with a three-fold reduction compared to unmodified speech. Spectro-temporal gains were observed for all phonemes. A glimpse-based model of energetic masking incorporating speech rate changes predicts intelligibility (r=.96), and a glimpsing analysis provides further insights into the distinct mechanisms through which spectral and temporal approaches lead to a release from energetic masking.

Keywords: speech modification, intelligibility, retiming, glimpsing

*Corresponding author
Email address: m.cooke@ikerbasque.org (Martin Cooke)

1. Introduction

Speech can be altered prior to presentation in noisy environments in such a way as to increase its intelligibility compared to unmodified speech [e.g., 1, 2, 3, 4]. Speech modification can lead to substantial gains: in an extensive evaluation of modification techniques known as the Hurricane Challenge [5], in which speech level was constrained to be constant pre- and post-modification, the most successful approaches produced gains equivalent to boosting the level of 'plain' unmodified speech by more than 5 dB.

Many algorithms proposed for speech modification operate by redistributing speech energy across the spectrum, either locally or from earlier or later portions of the signal. The Spectral Shaping and Dynamic Range Compression (SSDRC) method proposed by Zorila et al. [6] is an example of the energy redistribution approach. SSDRC incorporates a stage of spectral shaping reflecting properties of both clear speech [e.g., 7, 8, 9] and Lombard speech [e.g., 10, 11, 12], followed by dynamic range compression (DRC) which has the effect of transferring energy from more to less energetic epochs.

In contrast, relatively few modification approaches perform temporal modifications on the speech signal. Here, the term 'temporal modification' refers to retiming, i.e., changes to the temporal distribution of information-bearing speech elements. Such changes might involve altering the duration of speech segments [e.g., 13], or inserting pauses [e.g., 14] to effect a shift in their location. We have recently demonstrated that speech retiming is beneficial in the presence of temporally-modulated maskers, with gains ranging from 9 percentage points for linearly-elongated speech to 16 percentage points for non-linearly retimed speech [15]. Note that while the aforementioned DRC stage in the SSDRC algorithm has the effect of changing the temporal distribution of energy, the timing of the underlying speech segments remains unaltered.

For brevity in what follows, we will use the terms 'spectral' and 'temporal' to distinguish those techniques that leave the timing of information in the speech signal intact from those that modify the timing. The purpose of the current study is to examine whether the already substantial intelligibility benefits from spectral modification can be further increased via temporal modification algorithms. We chose the SSDRC and GCRetime [13] techniques to represent spectral and temporal modifications respectively due to their high level of intelligibility gains in the Hurricane evaluation. The current study tested the performance of the two algorithms alone and in combination using a common speech-in-noise task and listener cohort.

While it is not clear *a priori* what effect the combination of the two classes of modification approach will have on intelligibility, there are some reasons to expect additional gains from applying retiming to spectrally-modified speech. Spectral and temporal dimensions are to some extent independent in conveying information in speech. Place of articulation variations within each manner class are reflected mainly in changes to the speech spectrum, while cues to distinct manner classes additionally possess a strong temporal component. Both classes of modification technique aim to augment intelligibility by increasing the likelihood that energetically-weaker portions of speech escape masking, but they achieve this in distinct ways. Spectral approaches operate by boosting the energy of weaker signal elements at the expense of stronger regions. Temporal techniques do not alter the level of the speech itself, but aim to shift weaker regions in time to locations where the masker is less intense. In both cases the goal is to increase the signal-to-noise ratio (SNR) of fainter speech segments.

However, there are also reasons to question the hypothesis that spectral and temporal modifications will combine synergistically. The notion that spectral and temporal features in speech act in an orthogonal manner in cueing phoneme judgements is an oversimplification. It has long been known that spectral and temporal cues interact in determining the identity of speech segments [e.g., 16, 17]. There is also the possibility that the modifications produced by each technique, even though arrived at by different means, end up boosting the same weak signal elements, leading to a redundant combination. In support of this hypothesis, the gains observed for the best-performing spectral and temporal entries to the aforementioned Hurricane Challenge were very similar in the modulated masker condition, at 16 and 18 percentage points respectively.

Logically, a third possibility is that spectral and temporal modifications will combine antagonistically. Both classes of technique introduce distortions to the natural speech signal which are clearly evident when modified speech is presented in the absence of a masker. For example, informal listening to SSDRC-modified speech gives the impression that weak fricatives are overly-prominent, while for GCRetime the stretched or contracted segment durations can sound less than natural. Indeed, segment duration is explicitly contrastive in some languages, and can convey cues to adjacent phonemes in other languages where duration is not overtly contrastive (for example, the length of a vowel preceding an obstruent influences the perception of the consonant's phonological voicing status in English). In such cases, speech with artificially-modified segment durations might be less intelligible than unmodified speech.

In fact, there is evidence from formal listening tests that both SSDRC and GCRetime introduce distortions than can lead to a reduction in intelligibility and/or naturalness. SSDRC leads to lower quality ratings in quiet than unmodified speech, and only part of the reduction is due to the DRC element [18]. In a separate study, when SSDRC-modified speech was presented in noise-free conditions to non-native listeners (for whom scores are well below ceiling levels), keyword scores in sentences dropped relative to an unmodified speech condition [19]. Similarly, GCRetimed speech presented in stationary speech-shaped noise was substantially less intelligible than unmodified speech [15], indicating that when taken out of context – in this case the modulated masker being replaced by a stationary masker – local changes to the duration of speech segments have a negative effect on intelligibility. It is possible that the dual distortions expected to be present when spectral and temporal modifications are combined will lead to a net reduction in intelligibility.

The current study was carried out to determine which of the three possibilities raised above hold. Listeners identified unmodified sentences and sentences that had undergone spectral modification (SSDRC), temporal alteration (GCRetime) or spectro-temporal modification (SSDRC followed by GCRetime). Sentences were presented mixed at two SNRs with a temporally-fluctuating competing speech masker. Section 2 describes the listening experiment, whose results are presented in section 3.1. Additional analyses of segmental errors and a quantification of energetic masking are given in sections 3.2 and 4 respectively.

2. Experiment: perception of unmodified and modified sentences in a fluctuating masker

2.1. Speech and masker materials

Speech material came from the Sharvard corpus [20], a collection of Spanish sentences equivalent to the English language Harvard corpus [21]. Sharvard sentences are moderately predictable and contain five keywords used for estimating intelligibility. The first sentence of the corpus is "Coge las hojas y las quemas todas en el fuego" ["Collect the leaves and burn them all in the fire"] (keywords underlined). The Sharvard corpus consists of 700 sentences spoken by one male and one female talker. Sentences have 31 phonemes on average (range: 20–43, std. dev. = 4). Sentences are grouped into lists of 10, and each list has a phoneme frequency distribution equivalent to that of spoken Spanish. For the current experiment the first 24 lists (240 sentences) spoken by the male talker formed the basis for the target speech material.

The masker was competing speech spoken by a single female talker reading material from the Albayzin Spanish sentence corpus [22] from which between-sentence pauses had been removed. The use of a masking talker with different gender from that of the target talker minimised informational masking effects, enabling a focus on a reduction in energetic masking that the speech modification algorithms were designed to promote.

Speech and noise stimuli were downsampled to 16 kHz prior to presentation.

2.2. Unmodified and modified speech conditions

In addition to an unmodified speech condition, denoted Plain, listeners heard sentences processed by four speech modification algorithms, Spect, Temp, Temp* and Spect+Temp whose characteristics are described below.

2.2.1. Spect

The class of spectral modification algorithms is represented by the SSDRC algorithm [6]. This algorithm applies multi-stage spectral modification followed by dynamic range compression [23]. The first spectral stage consists of formant enhancement whose degree is adaptive and depends on an estimate of the probability of voicing. The second stage applies preemphasis, again adaptively. A third non-adaptive spectral weighting is also used to prevent attenuation of high frequencies. The result of spectral shaping forms the input to two stages of compression. The first 'dynamic' stage involves signal envelope compression with a 2 ms release time constant and almost instantaneous attack time constant. This is followed by static amplitude compression with the 0 dB reference level set to 0.3 times the peak of the signal envelope. SSDRC requires no knowledge of the masker, nor does it modify speech duration overall or locally.

2.2.2. Temp

Temporal modifications were carried out by the GCRetime algorithm [13, 15]. GCRetime finds the optimal sequence of local expansions and contractions of the target speech signal that jointly maximise an objective function in the presence of a fluctuating masker. In GCRetime, the objective function minimises energetic masking, estimated using glimpse proportion [24] while simultaneously maximising a measure of speech information as provided by the cochlear-scaled entropy metric [CSE; 25]. The objective function is maximised using dynamic programming, and the subsequent durational modifications are carried out using the WSOLA algorithm [26]. The Appendix of [15] provides a detailed description of the GCRetime algorithm.

Note that GCRetime in normal operation is not a general-purpose speech modification approach since it exploits knowledge of the instantaneous masker spectrum in a local time window centred on the current sample of the incoming speech signal. In practice this limits its applicability to scenarios such as retiming of remote multi-party conversations where a short delay can be imposed on both the output speech and masker. In spite of this limitation we chose GCRetime in order to estimate the best-case potential for combined spectro-temporal retiming relative to the chosen objective metric.

2.2.3. TEMP*

A simpler form of temporal modification was also tested. TEMP* is equivalent to TEMP but with the omission of the cochlear-scaled entropy component i.e. temporal modification via retiming is based solely on minimising energetic masking. TEMP* measures the effect of a pure temporal modification without the additional factor of retiming based on maximing the audibility of high-information regions of the signal.

2.2.4. SPECT+TEMP

The SPECT+TEMP algorithm combines SPECT with TEMP. Specifically, sentences from the SPECT condition were subsequently processed by the TEMP algorithm. This order of operation was chosen because of the requirement to estimate glimpses as part of the GCRetime algorithm. If SSDRC were to be applied in a stage subsequent to GCRetime, the glimpses which contributed to retiming would be likely to be quite different from those following application of the SSDRC algorithm.

Figure 1 shows spectrograms for an example sentence from Sharvard in unmodified form (PLAIN) and after processing by each of the four modification algorithms, along with the competing speech masker used for this specific speech-in-noise stimulus. Some of the aforementioned characteristics of the spectral and temporal manipulation algorithms are evident in this figure. Spectrally-modified speech (SPECT, SPECT+TEMP) shows increased

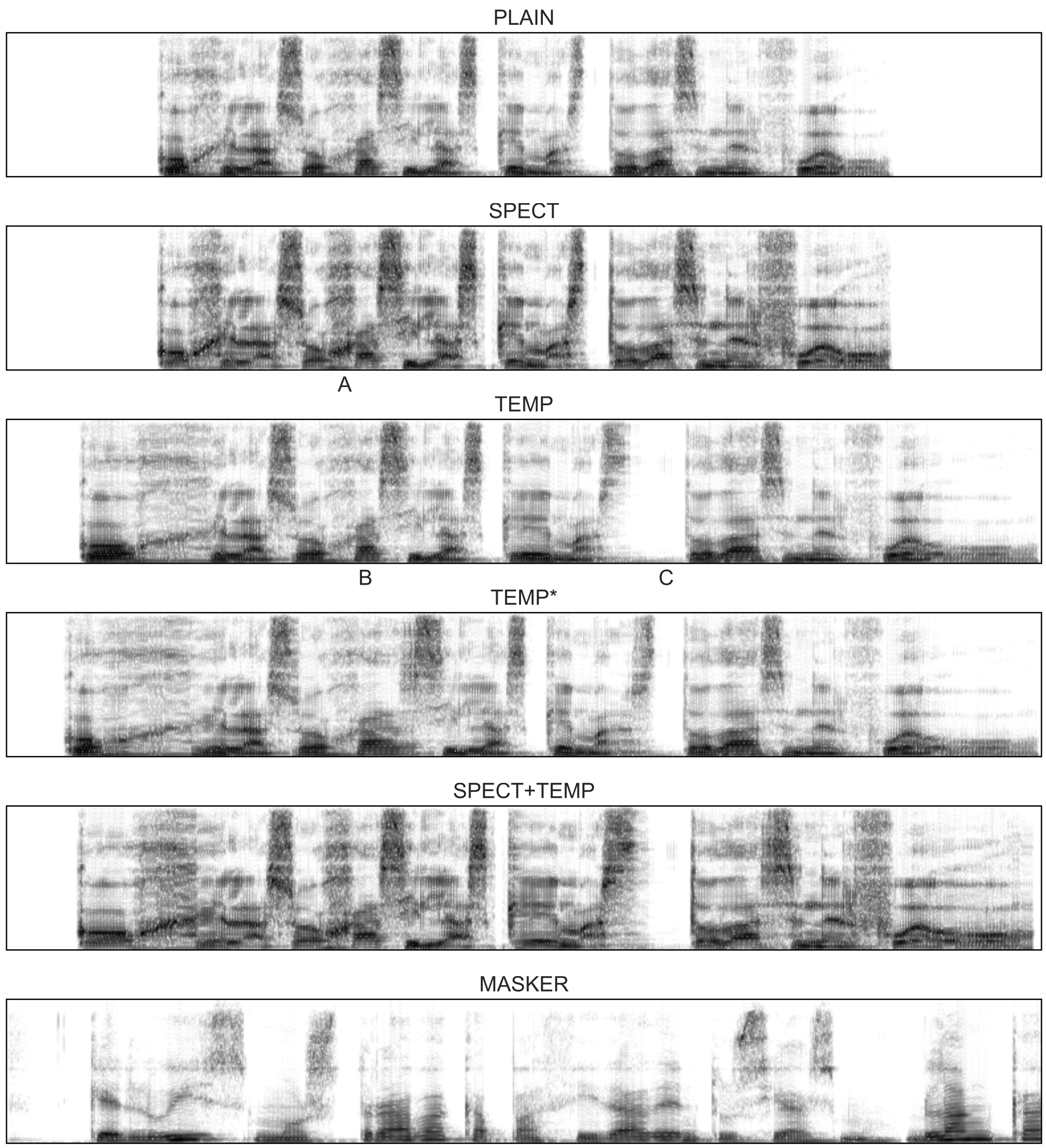

Figure 1: *Spectrograms of unmodified (*Plain*) and modified speech for the utterance "Coge las hojas y las quemas todas en el fuego". The masker used in this example is shown at the base of the figure. The frequency range is 0-8 kHz and the duration of the masker is 3.44s. Events at locations A-C are described in the text.*

energy at mid and high frequencies compared to the Plain and Temp methods. This is particularly apparent for the fricative /x/ in the word 'hojas' (location A in figure 1). For Spect there is no change in duration, while the methods involving retiming (Temp, Temp*, Spect+Temp) all result in a similar modest expansion in the time domain. The two retiming-only approaches show very clear differences, indicating that the presence or absence of CSE in the objective function which underlies retiming does have a significant effect on the modified speech. For example, the entire middle portion of the sentence (from location B to C in figure 1) follows a different retiming path for Temp and Temp*.

2.3. Speech-in-noise mixtures

Stimuli for the experiment consisted of plain and modified utterances mixed with the competing speech masker at one of two SNRs (-14 and -19 dB) chosen in pilot tests to produce mean keyword identification rates of around 70 % and 35 % respectively in the Plain condition. These SNRs are denoted 'moderate' and 'adverse'. The adverse SNR was chosen due to the possibility of ceiling effects arising from the modified speech in the moderate SNR condition. Sentences were centrally-embedded in the masker and the SNR computed over the region of overlap. For the Plain and Spect conditions, the lead and lag time of the masker was 0.5 s. For the three remaining conditions which involved retiming where some overall durational modification was permitted, the speech-masker overlap time was increased. For these conditions the masker led the speech by 0.2 s, and the lag time varied, dependent upon the overall retiming expansion. The speech-plus-noise waveform duration was identical in all conditions with a mean value of 3.35 s (std. dev. 0.28 s). The complete set of 240 utterances was processed by each of the four modification algorithms at both SNRs, leading to a total of 2400 stimuli (240 $\times$ 5 conditions $\times$ 2 SNRs). Each listener heard a 240-member subset of these stimuli (see section 2.5 for details of stimulus and condition order balancing).

2.4. Participants

Twenty-two listeners (18 female; mean age 20.7, std. dev. 4.1) participated in the experiment. All were either monolingual in Spanish or bilingual in Spanish and Basque. All listeners received hearing screening via an Interacoustics AS608 audiometer; all had normal hearing thresholds i.e. less than 20 dB hearing level over the range 125-8000 Hz. Listeners were paid for taking part. Ethics permission for the experiment was obtained under the University of the Basque Country Ethics Procedure.

2.5. Procedure

Stimuli were divided into two blocks, one for each SNR. Block order was balanced across participants. Within each block listeners heard 120 sentences, 24 for each of the 5 experimental conditions. Sentence presentation order was randomised within each block. Sentences and conditions were balanced across listeners to ensure that no listener heard the same sentence more than once in any condition and each sentence/condition pair was heard by a similar number of listeners (either 2 or 3, mean 2.2). Listeners were told that they would hear a mixture of a female voice and a less intensive male voice, and were instructed to type all the words they understood spoken by the male talker. Listeners were familiarised with the task via a short practice session consisting of 7 utterances drawn from the unused part of the Sharvard corpus. Listeners were seated in a sound-attenuating studio in the Phonetics Laboratory at the University of the Basque Country. Stimuli were presented at a level in the range 71-72 dB(A) through Sennheiser HD 380 pro headphones. Participants typed their responses into an onscreen text box in a custom-built Matlab application. Each of the two blocks required just over 21 minutes to complete on average.

2.6. Postprocessing

Listeners' text responses were processed prior to keyword scoring. First, diacritics indicating vowel stress were removed (e.g., á was replaced by a) since not all participants keyed in the stress symbol in all cases. Second, all non-alphabetic characters (e.g., punctuation symbols) were removed. Finally, words not present in the Spanish phonetic dictionary HAPLO [27] were removed.

3. Results

3.1. Keyword identification scores

Intelligibility is expressed as the percentage of keywords identified correctly across all sentences in each condition. Per-listener mean scores were computed from the 120 keywords (5 per sentence) heard by listeners in each of the 10 combinations of SNR and speech modification condition. Percentages were converted into rationalised arcsine units [RAU; 28] for statistical analysis. However, since all statistical outcomes were identical for RAU scores and percentages, the latter are used for ease of exposition in the following section.

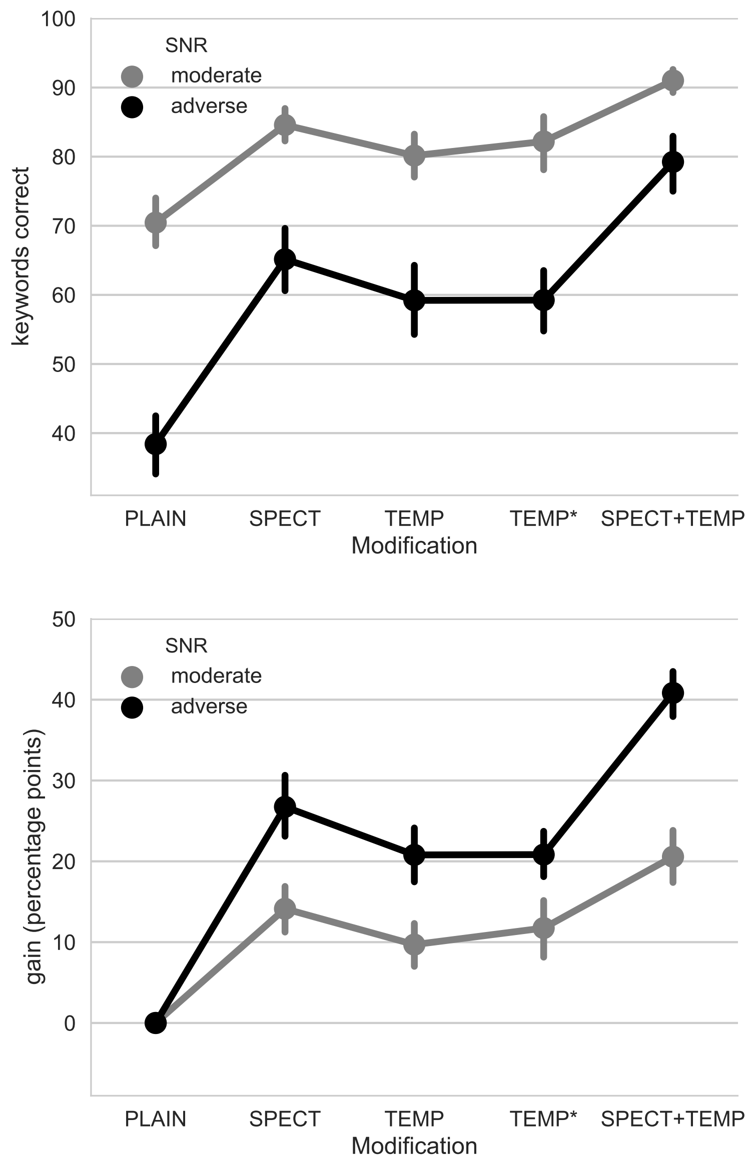

Figure 2: *Upper: Percentage of keywords recognised correctly as a function of modification technique and SNR. Lower: Gains in percentage points over unmodified speech. Error bars represent 95% confidence intervals.*

Figure 2 shows keyword scores (upper panel) and gains over the PLAIN baseline (lower). The pattern of scores for each SNR is similar, with larger gains at the more adverse SNR. Focusing on the adverse SNR, from a baseline of around 38% in the Plain condition, spectral modification alone produced a gain of nearly 27 percentage points (p.p.), while both temporal modification techniques led to gains of nearly 21 p.p. The combination of spectral and temporal modifications resulted in a gain of 41 p.p., corresponding to a keyword score of 79%, a near three-fold reduction in error rate over the Plain baseline (62% errors vs. 21% errors). The moderate SNR led to a 21 p.p. gain from spectro-temporal modification, corresponding to an error rate reduction factor of 3.3. Spectral modifications were generally more successful than temporal modification. Both temporal modification algorithms led to similar gains.

A repeated-measures ANOVA on gains with factors of SNR and modification condition confirms clear effects of both SNR [$F(1, 21) = 46, p < 0.001, \eta^2 = 0.44$], modification [$F(3, 63) = 75, p < 0.001, \eta^2 = 0.40$], together with a small but significant interaction between the two [$F(3, 63) = 11.4, p < 0.001, \eta^2 = 0.07$] due to the more limited potential for gains from the Spect+Temp modification approach at the moderate SNR. Based on a Fisher's Least Significant Difference of 2.9 p.p., spectro-temporal gains exceeded those seen in all other processing conditions. Gains in the Spect condition were greater than the two temporal conditions at the adverse SNR. However, Spect and Temp* produced equivalent gains in the moderate SNR condition.

The two temporal modification conditions produced statistically-equivalent gains. The lack of a significant benefit in using a component motivated by cochlear-scaled entropy [25] in retiming, demonstrated by the equivalence of scores in the Temp and Temp* conditions, is consistent with recent findings reported in [29] and [30], where it was observed that the 'entropy' element of cochlear-scaled entropy is not the main determinant of which speech regions are important for intelligibility.

3.2. Phoneme scores

In order to determine whether individual consonants or vowels benefitted preferentially from spectral or temporal modification, a phoneme-level analysis of listener responses to the sentence stimuli was carried out. In all, sentences contained some 163 960 phonemes, enabling robust estimation of hit rates for individual phonemes. The distribution of phonemes of the Sharvard sentences can be found in [20]. Responses were matched at the phoneme level to transcriptions of Sharvard sentences using a dynamic programming alignment algorithm. In each case the entire response rather than the keywords alone was used for matching, in order to allow for alternative word segmentations.

Average phoneme hit rates (not shown) follow the same pattern as the keyword scores presented in section 3.1 but from a higher baseline, ranging from 47% for Plain speech in the low SNR condition to 95% for the Spect+Temp modification in the moderate SNR condition. Figure 3 depicts per-phoneme recognition rates for consonants (upper panels) and vowels (lower panels). While baseline scores in the Plain condition differ across individual consonants and vowels, the striking feature of this figure is the near-uniform ranking of temporal, spectral and spectro-temporal modification methods across phonemes. At the more adverse SNR, spectral modification is more beneficial than temporal modification for nearly all consonants. Likewise, the combination of spectral and temporal modification clearly outperforms spectral modification for each individual consonant. The picture is similar for vowels at both SNRs. At the moderate SNR there is less of a clear separation between the spectral and temporal techniques with respect to consonant scores, but the proximity of scores to ceiling levels precludes deeper analysis.

We also examined changes in segment durations relative to the Plain baseline in the retimed condition Temp as well as the Spect condition. Durations were obtained by aligning sentences to their phoneme transcriptions using the Montreal Forced Aligner [31] which uses triphone-based hidden Markov models (HMMs). To avoid any bias from aligning modified speech using models trained on Plain speech, a separate set of HMMs was trained for each modification using all sentences for that condition.

Changes in consonant and vowel durations as a result of retiming, alongside those from the Spect condition, are shown in Figure 4, expressed as percentage increases relative to the Plain baseline. As expected, changes in the Spect condition are small; any variations from the 0% baseline (i.e., no increase in duration) stem from the fact that a separate set of HMMs was trained in each condition, leading to slight phoneme alignment differences. In contrast, individual consonants show significant changes in the Temp condition, the majority falling in the range of 20-40% expansion. No clear pattern linked to manner or place of articulation is evident. However, the voiceless plosives /p, t, k/ and the affricate /tʃ/ show least expansion. These are the only phonemes in Spanish with significant silent intervals (note that Spanish voiced plosives, when not realised as approximants, have at most a

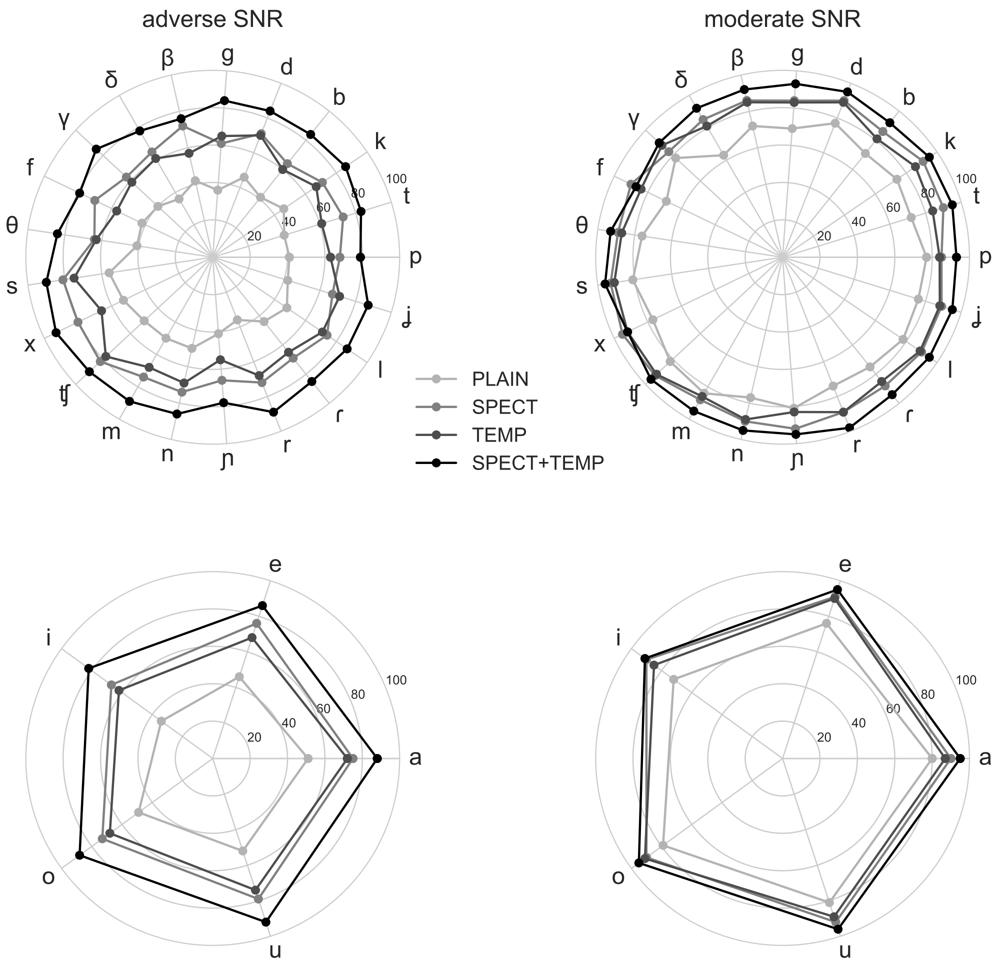

Figure 3: *Identification rates (percentage correct) for individual consonants (top) and vowels (bottom).*

brief period of occlusion [32]). It seems likely that the expansion of sounds consisting largely of near-silence is not favoured by the criterion of maximing glimpsing opportunities embodied in the GCRetime algorithm. Overall, vowel durations increase proportionally less than those of consonants, probably because their higher energy produces less of a need for masker-avoidance via retiming. Durational changes were not correlated with intelligibility gains at either SNR [adverse SNR: Pearson $r = -0.01, p = .97$; moderate SNR: $r = -0.31, p = .17$].

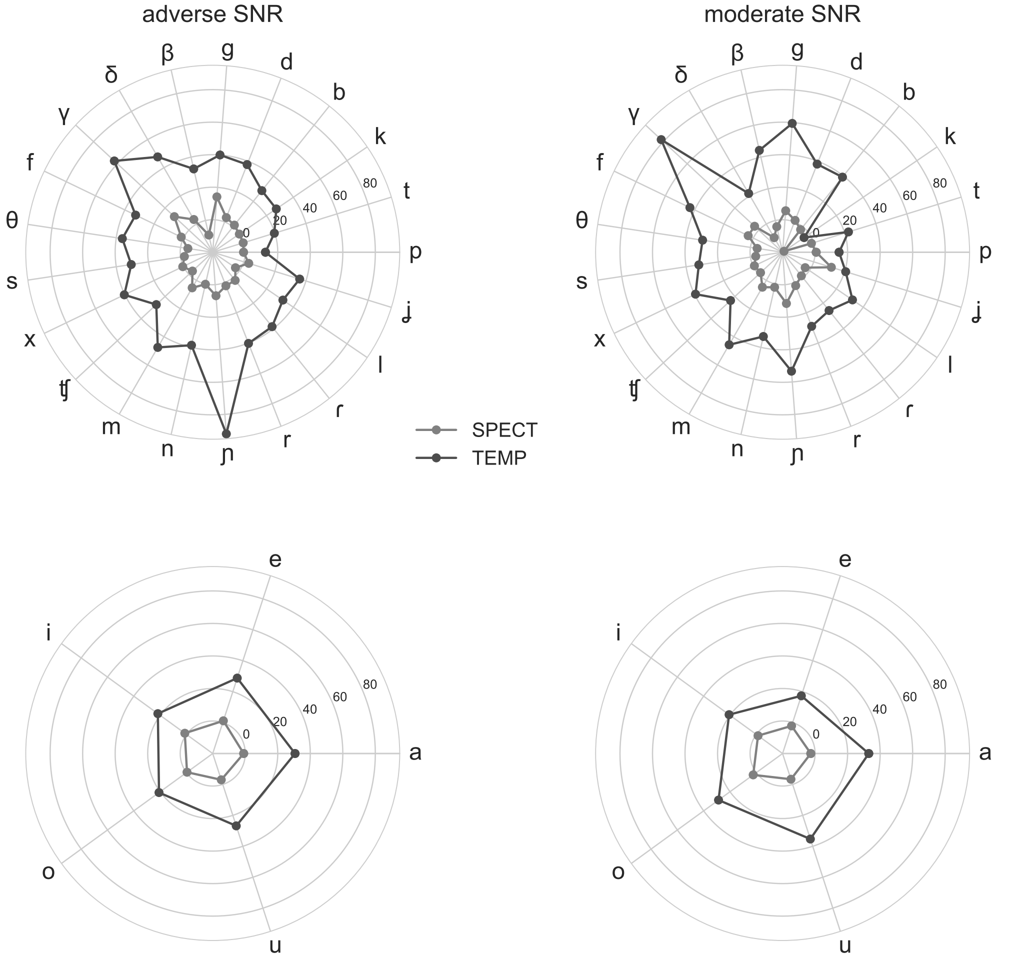

Figure 4: *Relative increases in duration, expressed in percentages, for the* TEMP *and* SPECT *conditions. The* SPECT *condition is included as a reference to indicate the scale of variations due to the forced alignment procedure (see text).*

3.3. Independent gains?

While spectral and temporal modification methods combine synergistically, the gains fall short of those that would be produced if the two methods reduced error rates independently. An assumption of independence of errors requires scores given by

$$\text{Score}_{Spect+Temp} = 1 - (1 - \text{Score}_{Temp})(1 - \text{Score}_{Spect})$$

This leads to predictions of 86% for the adverse condition (actual: 79%) and 97% at the moderate SNR (actual: 91%). An analysis at the level of phoneme hit rates rather than keywords produces similar results (91% predicted versus 85% actual for the adverse SNR, 98% predicted versus 94% actual for the moderate SNR).

4. Energetic masking

To explore the basis for intelligibility improvements, an analysis of energetic masking was carried out using a glimpsing metric. Glimpsing measures the degree to which a target signal exceeds the masker in time and frequency, computed using an auditorily-inspired signal representation. Glimpse proportion (GP) is the output of the initial stage of the glimpsing model of speech perception [24] and has been used as proxy for energetic masking in objective intelligibility metrics in applications involving speech synthesis [e.g., 33], speech broadcasting [34], and estimation of binaural speech intelligibility [35]. The starting point for GP computation is an auditory ratemap, a time-frequency-energy representation of the speech and masker signals. The ratemap is computed by passing the signal through a 55-channel gammatone filterbank with filter centre frequencies arranged on an ERB-rate scale from 50 Hz to 8000 Hz. The instantaneous (Hilbert) envelope at the output of each filter is smoothed with leaky integrator with time constant of 8 ms, downsampled to 100 Hz and log-compressed. Ratemaps are produced independently for speech and masker, and the proportion of time-frequency regions of the ratemap for speech exceeding that of the masker by a local SNR threshold (here set at 0 dB) defines the raw glimpse proportion.

The mean GP in each of the current set of 10 experimental conditions (5 modifications including Plain $\times$ 2 SNRs) predicts intelligibility quite well, with a Pearson correlation coefficient of 0.89 [$p < .001$]. However, we recently demonstrated that for a speech signal whose duration changes with respect to a reference speech signal (in this case the Plain speech), better predictions are possible using the extended GP metric, $\mathrm{GP}_{\mathrm{ext}}$ [36]. Amongst other features, $\mathrm{GP}_{\mathrm{ext}}$ takes speech rate changes into account by weighting glimpse proportion by a factor corresponding to the ratio of the modified speech duration to the unmodified speech duration. For the conditions of the current experiment, $\mathrm{GP}_{\mathrm{ext}}$ is highly-correlated with intelligibility [$\rho = .96, p < .001$], as shown in Figure 5. This outcome suggests that listeners' performance in the task is dominated by peripheral energetic masking rather

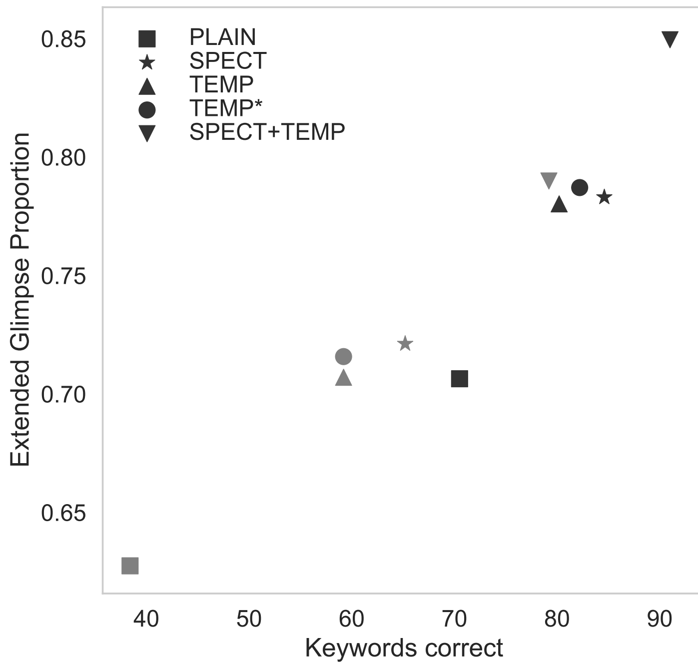

Figure 5: *Keyword scores plotted against intelligibility predictions from the extended glimpse proportion metric for the conditions of the experiment. Darker symbols come from the moderate SNR conditions.*

than informational masking from the competing talker. Indeed, given both the target-masker gender difference and the relatively adverse SNRs of the current experiment, there seems little possibility that listeners were confusing or misallocating speech material from the target and masker.

Continuing with the glimpse-based characterisation of the target-masker relationship, the upper panel of Figure 6 presents marginal distributions of raw (i.e., GP rather than $\mathrm{GP}_{\mathrm{ext}}$) glimpse likelihoods as a function of auditorily-scaled frequency for the adverse SNR condition (the pattern for the moderate SNR is very similar). These 'GP spectra' are per-frequency-channel means of GP measured across the entire corpus, for each modification technique. The two temporal modification techniques (Temp and Temp*) produced very similar results; for clarity only Temp is shown.

GP spectra reveal some clear differences between those modifications in-

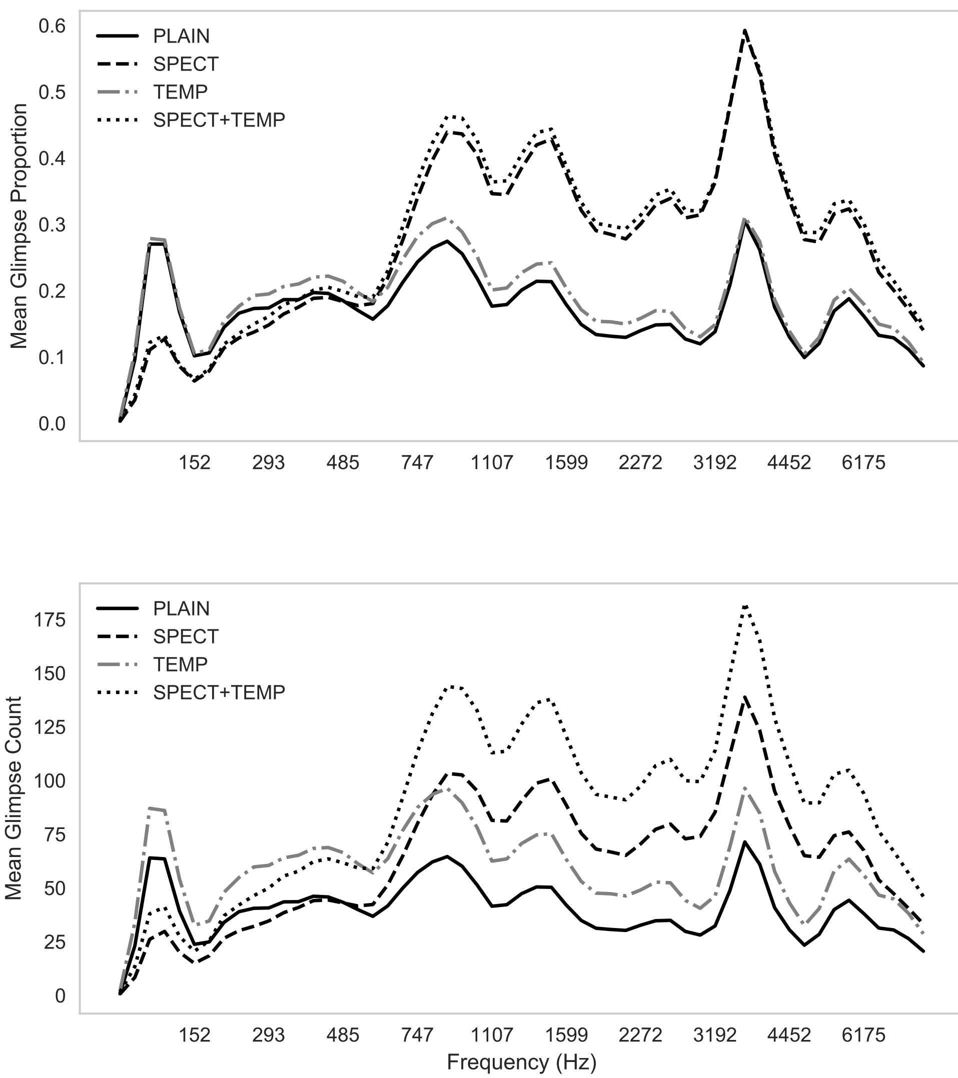

Figure 6: *Mean glimpse proportion (upper panel) and mean glimpse count (lower panel) in each frequency channel for the adverse SNR condition.*

volving spectral changes (Spect and Spect+Temp) and the Temp modification approach. For the frequency region from 700 Hz upwards, spectral techniques achieve a glimpse proportion of nearly double that of the temporal modification, which in turn shows only a small advantage over the

PLAIN baseline. However, the inverse pattern is seen below 500 Hz, with substantially fewer glimpses available as a result of spectral modification. These patterns suggest that much of the advantage of SSDRC stems from the transfer of energy from low frequencies (the first formant region and below) to mid and high frequencies (F2/F3 region and above). The fact that temporal modification produces only a modest gain over the unmodified baseline in terms of raw GP suggests that the intelligibility gains stemming from TEMP and TEMP* are not due to spectrally-based increases in glimpsing opportunities. Instead, gains presumably come from durational changes, as indicated in the duration-sensitive GP_{ext} metric. The mean GP curves for SPECT+TEMP reflect an almost identical modest gain over SPECT as those seen for TEMP over PLAIN, supporting the idea that temporal processes embodied in the GCRetime algorithm act to a large degree independently of spectral changes in SSDRC.

The lower panel of Figure 6 shows mean glimpse *counts* per channel. With this duration-sensitive measure, TEMP now shows a clear advantage over the PLAIN baseline throughout the entire frequency range. However, it is of interest to note that in spite of the augmented glimpse count for TEMP due to durational expansion, SPECT still produces a larger absolute glimpse count in the frequency region above 800 Hz.

In spite of the explanatory power of the glimpsing model in the current experiment, generalisation to other temporal modification algorithms needs to be tested, since a glimpsing metric (albeit GP and not GP_{ext}) was one component, along with cochlear-scaled entropy, of the GCRetime algorithm used to produce the temporal modification path.

5. Discussion

The main finding of the current study is that the application of a temporal modification technique to spectrally-modified speech leads to substantial additional gains over and above the sizeable improvements produced by spectral modification alone. The fact that intelligibility scores are very well predicted by the extended glimpse proportion model [36] that takes durational changes into account suggests that gains are largely due to energetic masking release rather than changes that reduce informational masking, since the glimpsing metric is based on identifying spectro-temporal regions that survive masking in the auditory periphery. With respect to energetic masking release, SSDRC exhibits a clear transfer of energy from the frequency region below 500 Hz to the mid and high frequency part of the spectrum. The loss of low frequency energy can be expected to reduce the salience of voicing cues conveyed by resolved harmonics. However, the impact of such a loss might have been relatively minor here since the contrastive role of voicing in Spanish is not great compared to languages such as English [32].

The current experiment provides no evidence that specific groups of sounds benefit from the spectral, temporal or spectro-temporal modification algorithms under test. Gains, while not uniform, were observed for all consonants and vowels, with a ranking that closely mirrors across-consonant mean intelligibility scores. One possible explanation arises from the nature of fluctating maskers, where the main determiner of intelligibility is the local temporal relationship between target and masker. Compared to a stationary masker, where high-energy phonemes are likely to escape masking most of the time while weaker sounds are more consistently masked, in the presence of a nonstationary masker with sufficient modulation depth (as is the case for competing speech) more intense sounds will suffer masking at least some of the time; similarly, fainter sounds will escape masking some of the time. An alternative and perhaps complementary reason as to why gains are spread across all phonemes comes from the fact that the task required listeners to identify words in sentences, thereby imposing morphological, lexical, syntactic and to a limited extent semantic constraints on their responses. In support of this notion, almost all errors at the phoneme level were deletions: the ratio of deletions to combined insertions + substitutions rose from 3.4 for Spect+Temp at the moderate SNR level to 9.4 for Plain speech at the adverse SNR. Listeners clearly preferred to delete entire words than to hypothesise alternative candidates.

The notion of high-level constraints on phoneme hit rates can also be invoked to explain the lack of a significant correlation between durational increases and score increases at the segmental level. Additionally, as mentioned in the introduction, changes to segment durations might have had a negative impact, but since we observe the net benefits of modification it is entirely possible that some of the positive effects of energetic masking release were counteracted by distortions to canonical forms.

Finally, we note that SSDRC and GCRetime were chosen to represent spectral and temporal modification approaches respectively, but other choices merit investigation. We recently demonstrated that uniform elongation of speech (i.e. a uniform reduction in speech rate) is also an effective strategy for intelligibility enhancement in fluctuating maskers [15], producing similar gains to GCRetime in a modulated noise condition. Uniform time-stretching was applied to SSDRC as part of the 'uwSSDRCt' technique reported in [37], but this combination did not increase intelligibility over SSDRC in a competing talker condition. However, uwSSDRCt also contained components to expand the vowel space and enhance transients, and it is possible that these interacted negatively with time-scale expansion. Future studies are needed to clarify whether imposing a slower speech rate on spectrally-modified speech leads to additional benefits.

6. Conclusions

In the current study, spectral and temporal modification techniques combined synergistically to boost the intelligibility of sentences in the presence of a fluctuating competing speech masker. While gains from spectral and temporal modification were not independent, increases in keyword scores were substantial, corresponding to a 3-fold reduction in error rates over unmodified speech. Intelligibility rates are well-predicted by a glimpse-based energetic masking metric which incorporates speech rate changes.

Acknowledgements

We thank Yannis Stylianou for providing code implementing the SSDRC algorithm. This work was supported in part by the EU Project ENRICH and by the Basque Government Consolidado grant to the Language and Speech Laboratory (LASLAB).

[1] M. D. Skowronski, J. G. Harris, Applied principles of clear and Lombard speech for automated intelligibility enhancement in noisy environments, Speech Communication 48 (5) (2006) 549–558. doi:10.1016/j.specom.2005.09.003.

[2] B. Sauert, P. Vary, Near end listening enhancement: Speech intelligibility improvement in noisy environments, in: Proc. ICASSP, Toulouse, France, 2006, pp. 493–496. doi:10.1109/ICASSP.2006.1660065.

[3] H. Brouckxon, W. Verhelst, B. D. Schuymer, Time and frequency dependent amplification for speech intelligibility enhancement in noisy environments, in: Proc. Interspeech, Vol. 9, 2008, pp. 557–560.

[4] C. H. Taal, J. Jensen, A. Leijon, On optimal linear filtering of speech for near-end listening enhancement, IEEE Signal Proc. Let. 20 (3) (2013) 225–228.

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

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

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

[8] Z. S. Bond, T. J. Moore, A note on the acoustic-phonetic characteristics of inadvertently clear speech, Speech Communication 14 (4) (1994) 325–337.

[9] R. M. Uchanski, Clear speech, in: D. B. Pisoni, R. E. Remez (Eds.), The Handbook of Speech Perception, Blackwell, Oxford, UK, 2005, pp. 207–235.

[10] J. J. Dreher, J. J. O'Neill, Effects of ambient noise on speaker intelligibility for words and phrases, J. Acoust. Soc. Am. 29 (12) (1957) 1320–1323.

[11] D. B. Pisoni, R. H. Bernacki, H. C. Nusbaum, M. Yuchtman, Some acoustic-phonetic correlates of speech produced in noise, in: ICASSP, Tampa, Florida, 1985, pp. 1581–1584.

[12] W. V. Summers, D. B. Pisoni, R. H. Bernacki, R. I. Pedlow, M. A. Stokes, Effects of noise on speech production: Acoustic and perceptual analyses, J. Acoust. Soc. Am. 84 (3) (1988) 917–928.

[13] V. Aubanel, M. Cooke, Information-preserving temporal reallocation of speech in the presence of fluctuating maskers, in: Proc. Interspeech, Lyon, France, 2013, pp. 3592–3596.

[14] R. M. Uchanski, S. S. Choi, L. D. Braida, C. M. Reed, N. I. Durlach, Speaking clearly for the hard of hearing IV: Further studies of the role of speaking rate, J. Speech Hear. Res. 39 (3) (1996) 494–509.

[15] M. Cooke, V. Aubanel, Effects of linear and nonlinear speech rate changes on speech intelligibility in stationary and fluctuating maskers, J. Acoust. Soc. Am. 141 (2017) 4126–4135.

[16] Q. Summerfield, M. Haggard, On the dissociation of spectral and temporal cues to the voicing distinction in initial stop consonants, J. Acoust. Soc. Am. 62 (1977) 436–448.

[17] R. V. Shannon, F.-G. Zeng, V. Kamath, J. Wygonski, M. Ekelid, Speech recognition with primarily temporal cues, Science 270 (5234) (1995) 303–304.

[18] Y. Tang, C. Arnold, T. Cox, A study on the relationship between the intelligibility and quality of algorithmically-modified speech for normal hearing listeners, Journal of Otorhinolaryngology, Hearing and Balance Medicine 1 (2017) 5.

[19] M. Cooke, M. L. Garcia Lecumberri, The effects of modified speech styles on intelligibility for non-native listeners, in: Proc. Interspeech, 2016, pp. 868–872. doi:10.21437/Interspeech.2016-41.

[20] V. Aubanel, M. L. García Lecumberri, M. Cooke, The Sharvard Corpus: A phonemically-balanced Spanish sentence resource for audiology , Int. J. Audiology 53 (2014) 633–638.

[21] E. H. Rothauser, W. D. Chapman, N. Guttman, M. H. L. Hecker, K. S. Nordby, H. R. Silbiger, G. E. Urbanek, M. Weistock, V. E. McGee, U. P. Pachl, W. D. Voiers, IEEE Recommended practice for speech quality measurements, IEEE Trans. Audio Acoust. (1969) 225–246.

[22] A. Moreno, D. Poch, A. Bonafonte, E. Lleida, J. Llisterri, J. B. Marino, C. Nadeu, Albayzín speech database: Design of the phonetic corpus, in: Eurospeech, Berlin, Germany, 1993, pp. 175–178.

[23] B. A. Blesser, Audio dynamic range compression for minimum perceived distortion, IEEE Trans. on Audio and Electroacoustics 17 (1) (1969) 22–32.

[24] M. Cooke, A glimpsing model of speech perception in noise, J. Acoust. Soc. Am. 119 (3) (2006) 1562–1573.

[25] C. Stilp, K. Kluender, Cochlea-scaled entropy, not consonants, vowels, or time, best predicts speech intelligibility, P. Natl. Acad. Sci. USA 107 (27) (2010) 12387–12392.

[26] M. Demol, W. Verhelst, K. Struyve, P. Verhoeve, Efficient non-uniform time-scaling of speech with WSOLA, in: Int. Conf. on Speech and Computers (SPECOM), 2005, pp. 163–166.

[27] R. Perez Ramon, Haplo: Herramienta automática de procesamiento linguístico ortofonético, in: Proc. Asociación Espaola de Linguística Aplicada, Lleida, 2012.

[28] G. Studebaker, A rationalized arcsine transform, Journal of Speech and Hearing Research 28 (1985) 455–462.

[29] A. J. Oxenham, J. E. Boucher, H. A. Kreft, Speech intelligibility is best predicted by intensity, not cochlea-scaled entropy, J. Acoust. Soc. Am. 142 (3) (2017) EL264–EL269.

[30] V. Aubanel, M. Cooke, C. Davis, J. Kim, Temporal factors in cochlea-scaled entropy and intensity-based intelligibility predictions, J. Acoust. Soc. Am. 143 (6) (2018) EL443–EL448. doi:10.1121/1.5041468.

[31] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, M. Sonderegger, Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi, in: Proc. Interspeech 2017, 2017, pp. 498–502.

[32] J. I. Hualde, The Sounds of Spanish, Cambridge University Press, 2005.

[33] C. Valentini-Botinhao, R. Maia, J. Yamagishi, S. King, H. Zen, Cepstral analysis based on the Glimpse proportion measure for improving the intelligibility of HMM-based synthetic speech in noise, in: Proc. ICASSP, 2012, pp. 3997–4000.

[34] Y. Tang, B. Fazenda, T. Cox, Automatic speech-to-background ratio selection to maintain speech intelligibility in broadcasts using an objective intelligibility metric, Applied Sciences 8 (2018) 59.

[35] Y. Tang, Q. Liu, W. Wang, T. Cox, A non-intrusive method for estimating binaural speech intelligibility from noise-corrupted signals captured by a pair of microphones, Speech Communication 96 (2017) 116–128.

[36] Y. Tang, M. Cooke, Glimpse-based metrics for predicting speech intelligibility in additive noise conditions, in: Proc. Interspeech, 2016, pp. 2488–2492. doi:10.21437/Interspeech.2016-14.

[37] E. Godoy, Y. Stylianou, Increasing speech intelligibility via spectral shaping with frequency warping and dynamic range compression plus transient enhancement, in: Proc. Interspeech, 2013, pp. 3572–3576.