



**HAL**  
open science

## Combining spectral and temporal modification techniques for speech intelligibility enhancement

Martin Cooke, Vincent Aubanel, María Luisa García Lecumberri

► **To cite this version:**

Martin Cooke, Vincent Aubanel, María Luisa García Lecumberri. Combining spectral and temporal modification techniques for speech intelligibility enhancement. *Computer Speech and Language*, 2019, 55, pp.26-39. 10.1016/j.csl.2018.10.003 . hal-02067420

**HAL Id: hal-02067420**

**<https://hal.science/hal-02067420v1>**

Submitted on 15 Mar 2019

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

## Accepted Manuscript

Combining spectral and temporal modification techniques for speech intelligibility enhancement

Martin Cooke, Vincent Aubanel, María Luisa García Lecumberri

PII: S0885-2308(18)30067-6  
DOI: <https://doi.org/10.1016/j.csl.2018.10.003>  
Reference: YCSLA 958



To appear in: *Computer Speech & Language*

Received date: 2 March 2018  
Revised date: 10 September 2018  
Accepted date: 26 October 2018

Please cite this article as: Martin Cooke, Vincent Aubanel, María Luisa García Lecumberri, Combining spectral and temporal modification techniques for speech intelligibility enhancement, *Computer Speech & Language* (2018), doi: <https://doi.org/10.1016/j.csl.2018.10.003>

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

**Highlights**

- spectral and temporal modification techniques combine synergistically
- for Spanish sentences, error rates are reduced by a factor of 3 compared to unmodified speech
- all phonemes benefit from spectral and temporal modification
- a glimpsing model predicts listener performance with a correlation of 0.96
- Cochlear-Scaled Entropy does not improve the performance of a retiming algorithm

# Combining spectral and temporal modification techniques for speech intelligibility enhancement

Martin Cooke<sup>a,b</sup>, Vincent Aubanel<sup>c</sup>, María Luisa García Lecumberri<sup>b</sup>

<sup>a</sup>*Ikerbasque (Basque Science Foundation)*

<sup>b</sup>*Language and Speech Laboratory, Universidad del País Vasco, 01006 Vitoria, Spain*

<sup>c</sup>*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-

---

\*Corresponding author

*Email address:* [m.cooke@ikerbasque.org](mailto:m.cooke@ikerbasque.org) (Martin Cooke)

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

---

## 1 **1. Introduction**

2 Speech can be altered prior to presentation in noisy environments in such  
3 a way as to increase its intelligibility compared to unmodified speech [e.g.,  
4 1, 2, 3, 4]. Speech modification can lead to substantial gains: in an exten-  
5 sive evaluation of modification techniques known as the Hurricane Challenge  
6 [5], in which speech level was constrained to be constant pre- and post-  
7 modification, the most successful approaches produced gains equivalent to  
8 boosting the level of ‘plain’ unmodified speech by more than 5 dB.

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

17 In contrast, relatively few modification approaches perform temporal mod-  
18 ifications on the speech signal. Here, the term ‘temporal modification’ refers

19 to retiming, i.e., changes to the temporal distribution of information-bearing  
20 speech elements. Such changes might involve altering the duration of speech  
21 segments [e.g., 13], or inserting pauses [e.g., 14] to effect a shift in their lo-  
22 cation. We have recently demonstrated that speech retiming is beneficial in  
23 the presence of temporally-modulated maskers, with gains ranging from 9  
24 percentage points for linearly-elongated speech to 16 percentage points for  
25 non-linearly retimed speech [15]. Note that while the aforementioned DRC  
26 stage in the SSDRC algorithm has the effect of changing the temporal dis-  
27 tribution of energy, the timing of the underlying speech segments remains  
28 unaltered.

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

39 While it is not clear *a priori* what effect the combination of the two classes  
40 of modification approach will have on intelligibility, there are some reasons to  
41 expect additional gains from applying retiming to spectrally-modified speech.  
42 Spectral and temporal dimensions are to some extent independent in con-  
43 veying information in speech. Place of articulation variations within each

44 manner class are reflected mainly in changes to the speech spectrum, while  
45 cues to distinct manner classes additionally possess a strong temporal com-  
46 ponent. Both classes of modification technique aim to augment intelligibility  
47 by increasing the likelihood that energetically-weaker portions of speech es-  
48 cape masking, but they achieve this in distinct ways. Spectral approaches  
49 operate by boosting the energy of weaker signal elements at the expense of  
50 stronger regions. Temporal techniques do not alter the level of the speech  
51 itself, but aim to shift weaker regions in time to locations where the masker  
52 is less intense. In both cases the goal is to increase the signal-to-noise ratio  
53 (SNR) of fainter speech segments.

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

66 Logically, a third possibility is that spectral and temporal modifications  
67 will combine antagonistically. Both classes of technique introduce distor-  
68 tions to the natural speech signal which are clearly evident when modified

69 speech is presented in the absence of a masker. For example, informal listen-  
70 ing to SSDRC-modified speech gives the impression that weak fricatives are  
71 overly-prominent, while for GCReTime the stretched or contracted segment  
72 durations can sound less than natural. Indeed, segment duration is explicitly  
73 contrastive in some languages, and can convey cues to adjacent phonemes  
74 in other languages where duration is not overtly contrastive (for example,  
75 the length of a vowel preceding an obstruent influences the perception of the  
76 consonant's phonological voicing status in English). In such cases, speech  
77 with artificially-modified segment durations might be less intelligible than  
78 unmodified speech.

79 In fact, there is evidence from formal listening tests that both SSDRC and  
80 GCReTime introduce distortions that can lead to a reduction in intelligibility  
81 and/or naturalness. SSDRC leads to lower quality ratings in quiet than  
82 unmodified speech, and only part of the reduction is due to the DRC element  
83 [18]. In a separate study, when SSDRC-modified speech was presented in  
84 noise-free conditions to non-native listeners (for whom scores are well below  
85 ceiling levels), keyword scores in sentences dropped relative to an unmodified  
86 speech condition [19]. Similarly, GCReTimed speech presented in stationary  
87 speech-shaped noise was substantially less intelligible than unmodified speech  
88 [15], indicating that when taken out of context – in this case the modulated  
89 masker being replaced by a stationary masker – local changes to the duration  
90 of speech segments have a negative effect on intelligibility. It is possible  
91 that the dual distortions expected to be present when spectral and temporal  
92 modifications are combined will lead to a net reduction in intelligibility.

93 The current study was carried out to determine which of the three pos-



94 sibilities raised above hold. Listeners identified unmodified sentences and  
95 sentences that had undergone spectral modification (SSDRC), temporal al-  
96 teration (GCReTime) or spectro-temporal modification (SSDRC followed by  
97 GCReTime). Sentences were presented mixed at two SNRs with a temporally-  
98 fluctuating competing speech masker. Section 2 describes the listening ex-  
99 periment, whose results are presented in section 3.1. Additional analyses  
100 of segmental errors and a quantification of energetic masking are given in  
101 sections 3.2 and 4 respectively.

## 102 **2. Experiment: perception of unmodified and modified sentences** 103 **in a fluctuating masker**

### 104 *2.1. Speech and masker materials*

105 Speech material came from the Sharvard corpus [20], a collection of Span-  
106 ish sentences equivalent to the English language Harvard corpus [21]. Shar-  
107 vard sentences are moderately predictable and contain five keywords used for  
108 estimating intelligibility. The first sentence of the corpus is “Coge las hojas  
109 y las quemas todas en el fuego” [“Collect the leaves and burn them all in the  
110 fire”] (keywords underlined). The Sharvard corpus consists of 700 sentences  
111 spoken by one male and one female talker. Sentences have 31 phonemes on  
112 average (range: 20–43, std. dev. = 4). Sentences are grouped into lists of  
113 10, and each list has a phoneme frequency distribution equivalent to that of  
114 spoken Spanish. For the current experiment the first 24 lists (240 sentences)  
115 spoken by the male talker formed the basis for the target speech material.

116 The masker was competing speech spoken by a single female talker read-  
117 ing material from the Albayzin Spanish sentence corpus [22] from which

118 between-sentence pauses had been removed. The use of a masking talker  
119 with different gender from that of the target talker minimised informational  
120 masking effects, enabling a focus on a reduction in energetic masking that  
121 the speech modification algorithms were designed to promote.

122 Speech and noise stimuli were downsampled to 16 kHz prior to presenta-  
123 tion.

## 124 *2.2. Unmodified and modified speech conditions*

125 In addition to an unmodified speech condition, denoted PLAIN, listeners  
126 heard sentences processed by four speech modification algorithms, SPECT,  
127 TEMP, TEMP\* and SPECT+TEMP whose characteristics are described be-  
128 low.

### 129 *2.2.1. SPECT*

130 The class of spectral modification algorithms is represented by the SS-  
131 DRC algorithm [6]. This algorithm applies multi-stage spectral modification  
132 followed by dynamic range compression [23]. The first spectral stage consists  
133 of formant enhancement whose degree is adaptive and depends on an esti-  
134 mate of the probability of voicing. The second stage applies preemphasis,  
135 again adaptively. A third non-adaptive spectral weighting is also used to  
136 prevent attenuation of high frequencies. The result of spectral shaping forms  
137 the input to two stages of compression. The first ‘dynamic’ stage involves  
138 signal envelope compression with a 2 ms release time constant and almost  
139 instantaneous attack time constant. This is followed by static amplitude  
140 compression with the 0 dB reference level set to 0.3 times the peak of the  
141 signal envelope. SSDRC requires no knowledge of the masker, nor does it

142 modify speech duration overall or locally.

### 143 2.2.2. TEMP

144 Temporal modifications were carried out by the GCReTime algorithm  
145 [13, 15]. GCReTime finds the optimal sequence of local expansions and con-  
146 tractions of the target speech signal that jointly maximise an objective func-  
147 tion in the presence of a fluctuating masker. In GCReTime, the objective  
148 function minimises energetic masking, estimated using glimpse proportion  
149 [24] while simultaneously maximising a measure of speech information as  
150 provided by the cochlear-scaled entropy metric [CSE; 25]. The objective  
151 function is maximised using dynamic programming, and the subsequent du-  
152 rational modifications are carried out using the WSOLA algorithm [26]. The  
153 Appendix of [15] provides a detailed description of the GCReTime algorithm.

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

### 162 2.2.3. TEMP\*

163 A simpler form of temporal modification was also tested. TEMP\* is equiv-  
164 alent to TEMP but with the omission of the cochlear-scaled entropy com-  
165 ponent i.e. temporal modification via retiming is based solely on minimising

166 energetic masking.  $\text{TEMP}^*$  measures the effect of a pure temporal mod-  
 167 ification without the additional factor of retiming based on maximizing the  
 168 audibility of high-information regions of the signal.

#### 169 2.2.4. SPECT+TEMP

170 The SPECT+TEMP algorithm combines SPECT with TEMP. Specifically,  
 171 sentences from the SPECT condition were subsequently processed by the  
 172 TEMP algorithm. This order of operation was chosen because of the require-  
 173 ment to estimate glimpses as part of the GCReTime algorithm. If SSDRC  
 174 were to be applied in a stage subsequent to GCReTime, the glimpses which  
 175 contributed to retiming would be likely to be quite different from those fol-  
 176 lowing application of the SSDRC algorithm.

177 Figure 1 shows spectrograms for an example sentence from Sharvard in  
 178 unmodified form (PLAIN) and after processing by each of the four modi-  
 179 fication algorithms, along with the competing speech masker used for this  
 180 specific speech-in-noise stimulus. Some of the aforementioned characteris-  
 181 tics of the spectral and temporal manipulation algorithms are evident in this  
 182 figure. Spectrally-modified speech (SPECT, SPECT+TEMP) shows increased  
 183 energy at mid and high frequencies compared to the PLAIN and TEMP meth-  
 184 ods. This is particularly apparent for the fricative /x/ in the word ‘hojas’  
 185 (location A in figure 1). For SPECT there is no change in duration, while  
 186 the methods involving retiming (TEMP,  $\text{TEMP}^*$ , SPECT+TEMP) all result  
 187 in a similar modest expansion in the time domain. The two retiming-only  
 188 approaches show very clear differences, indicating that the presence or ab-  
 189 sence of CSE in the objective function which underlies retiming does have  
 190 a significant effect on the modified speech. For example, the entire middle

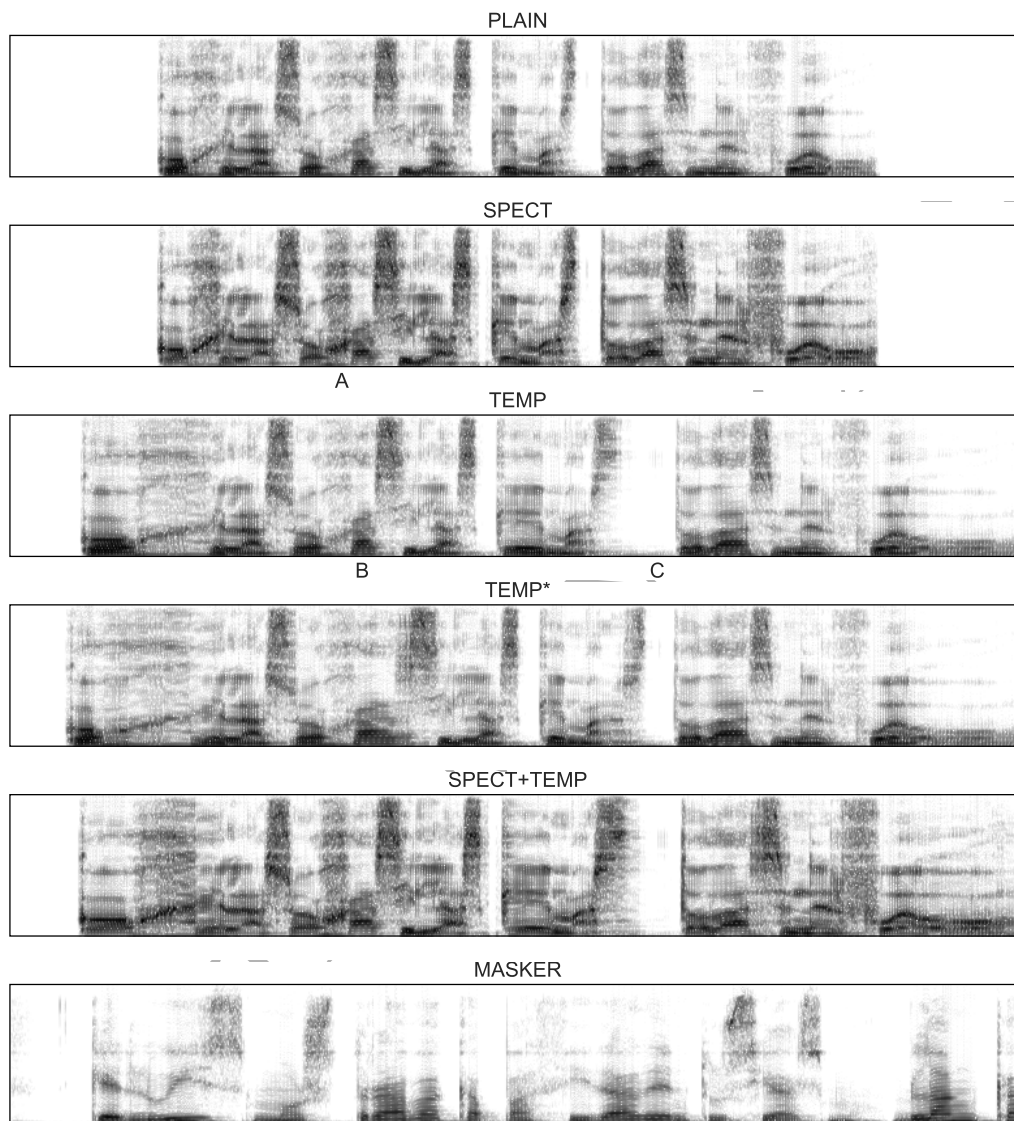


Figure 1: Spectrograms of unmodified (PLAIN) and modified speech for the utterance “Coge las hojas y las quemas todas en el fuego”. The masker used in this example is shown at the base of the figure. The frequency range is 0-8 kHz and the duration of the masker is 3.44s. Events at locations A-C are described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

191 portion of the sentence (from location B to C in figure 1) follows a different  
192 retiming path for TEMP and TEMP\*.

### 193 2.3. *Speech-in-noise mixtures*

194 Stimuli for the experiment consisted of plain and modified utterances  
195 mixed with the competing speech masker at one of two SNRs (-14 and -19  
196 dB) chosen in pilot tests to produce mean keyword identification rates of  
197 around 70% and 35% respectively in the PLAIN condition. These SNRs are  
198 denoted ‘moderate’ and ‘adverse’. The adverse SNR was chosen due to the  
199 possibility of ceiling effects arising from the modified speech in the moderate  
200 SNR condition. Sentences were centrally-embedded in the masker and the  
201 SNR computed over the region of overlap. For the PLAIN and SPECT condi-  
202 tions, the lead and lag time of the masker was 0.5 s. For the three remaining  
203 conditions which involved retiming where some overall durational modifica-  
204 tion was permitted, the speech-masker overlap time was increased. For these  
205 conditions the masker led the speech by 0.2 s, and the lag time varied, depen-  
206 dent upon the overall retiming expansion. The speech-plus-noise waveform  
207 duration was identical in all conditions with a mean value of 3.35 s (std. dev.  
208 0.28 s). The complete set of 240 utterances was processed by each of the  
209 four modification algorithms at both SNRs, leading to a total of 2400 stimuli  
210 ( $240 \times 5 \text{ conditions} \times 2 \text{ SNRs}$ ). Each listener heard a 240-member subset  
211 of these stimuli (see section 2.5 for details of stimulus and condition order  
212 balancing).

213 *2.4. Participants*

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

221 *2.5. Procedure*

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

237 a custom-built Matlab application. Each of the two blocks required just over  
238 21 minutes to complete on average.

### 239 *2.6. Postprocessing*

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

## 246 **3. Results**

### 247 *3.1. Keyword identification scores*

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

256 Figure 2 shows keyword scores (upper panel) and gains over the PLAIN base-  
257 line (lower). The pattern of scores for each SNR is similar, with larger gains  
258 at the more adverse SNR. Focusing on the adverse SNR, from a baseline of  
259 around 38% in the PLAIN condition, spectral modification alone produced



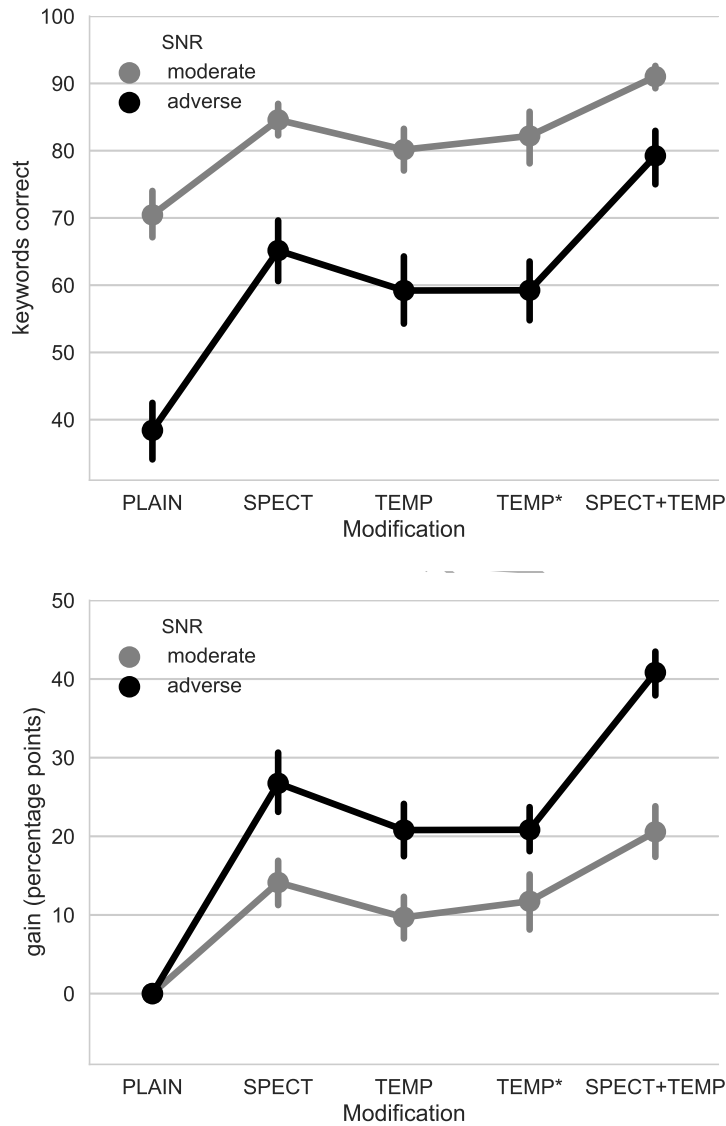


Figure 2: Upper: Percentage of keywords recognised correctly as a function of modification technique and SNR. Lower: Gains in percentage points over unmodified speech. Error bars represent 95% confidence intervals.

260 a gain of nearly 27 percentage points (p.p.), while both temporal modifica-  
 261 tion techniques led to gains of nearly 21 p.p. The combination of spectral  
 262 and temporal modifications resulted in a gain of 41 p.p., corresponding to  
 263 a keyword score of 79%, a near three-fold reduction in error rate over the  
 264 PLAIN baseline (62% errors vs. 21% errors). The moderate SNR led to a 21  
 265 p.p. gain from spectro-temporal modification, corresponding to an error rate  
 266 reduction factor of 3.3. Spectral modifications were generally more successful  
 267 than temporal modification. Both temporal modification algorithms led to  
 268 similar gains.

269 A repeated-measures ANOVA on gains with factors of SNR and mod-  
 270 ification condition confirms clear effects of both SNR [ $F(1, 21) = 46, p <$   
 271  $0.001, \eta^2 = 0.44$ ], modification [ $F(3, 63) = 75, p < 0.001, \eta^2 = 0.40$ ], to-  
 272 gether with a small but significant interaction between the two [ $F(3, 63) =$   
 273  $11.4, p < 0.001, \eta^2 = 0.07$ ] due to the more limited potential for gains from  
 274 the SPECT+TEMP modification approach at the moderate SNR. Based on  
 275 a Fisher's Least Significant Difference of 2.9 p.p., spectro-temporal gains ex-  
 276 ceeded those seen in all other processing conditions. Gains in the SPECT con-  
 277 dition were greater than the two temporal conditions at the adverse SNR.  
 278 However, SPECT and TEMP\* produced equivalent gains in the moderate  
 279 SNR condition.

280 The two temporal modification conditions produced statistically-equivalent  
 281 gains. The lack of a significant benefit in using a component motivated by  
 282 cochlear-scaled entropy [25] in retiming, demonstrated by the equivalence of  
 283 scores in the TEMP and TEMP\* conditions, is consistent with recent findings  
 284 reported in [29] and [30], where it was observed that the 'entropy' element of

285 cochlear-scaled entropy is not the main determinant of which speech regions  
286 are important for intelligibility.

### 287 3.2. Phoneme scores

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

298 Average phoneme hit rates (not shown) follow the same pattern as the  
299 keyword scores presented in section 3.1 but from a higher baseline, rang-  
300 ing from 47% for PLAIN speech in the low SNR condition to 95% for the  
301 SPECT+TEMP modification in the moderate SNR condition. Figure 3 de-  
302 picts per-phoneme recognition rates for consonants (upper panels) and vowels  
303 (lower panels). While baseline scores in the PLAIN condition differ across  
304 individual consonants and vowels, the striking feature of this figure is the  
305 near-uniform ranking of temporal, spectral and spectro-temporal modifica-  
306 tion methods across phonemes. At the more adverse SNR, spectral modifica-  
307 tion is more beneficial than temporal modification for nearly all consonants.  
308 Likewise, the combination of spectral and temporal modification clearly out-  
309 performs spectral modification for each individual consonant. The picture

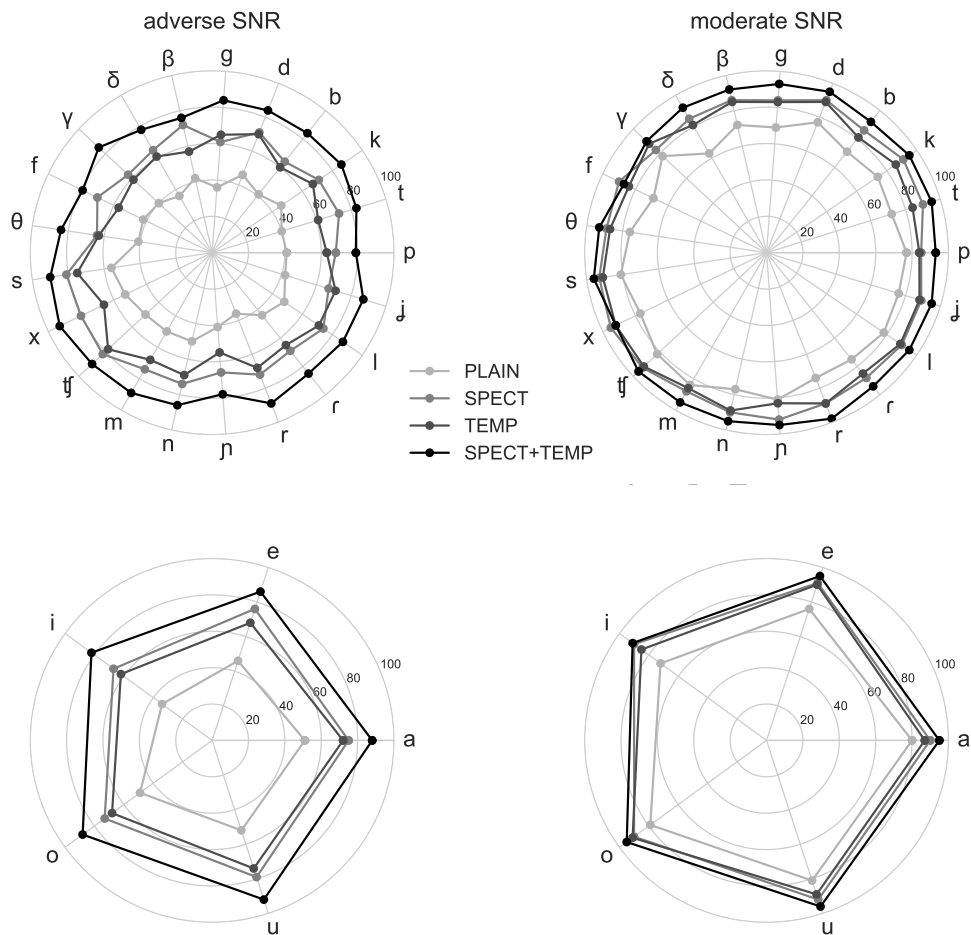


Figure 3: *Identification rates (percentage correct) for individual consonants (top) and vowels (bottom).*

310 is similar for vowels at both SNRs. At the moderate SNR there is less of a  
 311 clear separation between the spectral and temporal techniques with respect  
 312 to consonant scores, but the proximity of scores to ceiling levels precludes  
 313 deeper analysis.

314 We also examined changes in segment durations relative to the PLAIN base-

315 line in the retimed condition TEMP as well as the SPECT condition. Dura-  
316 tions were obtained by aligning sentences to their phoneme transcriptions  
317 using the Montreal Forced Aligner [31] which uses triphone-based hidden  
318 Markov models (HMMs). To avoid any bias from aligning modified speech  
319 using models trained on PLAIN speech, a separate set of HMMs was trained  
320 for each modification using all sentences for that condition.

321 Changes in consonant and vowel durations as a result of retiming, along-  
322 side those from the SPECT condition, are shown in Figure 4, expressed as  
323 percentage increases relative to the PLAIN baseline. As expected, changes in  
324 the SPECT condition are small; any variations from the 0% baseline (i.e., no  
325 increase in duration) stem from the fact that a separate set of HMMs was  
326 trained in each condition, leading to slight phoneme alignment differences. In  
327 contrast, individual consonants show significant changes in the TEMP condi-  
328 tion, the majority falling in the range of 20-40% expansion. No clear pattern  
329 linked to manner or place of articulation is evident. However, the voice-  
330 less plosives /p, t, k/ and the affricate /tʃ/ show least expansion. These  
331 are the only phonemes in Spanish with significant silent intervals (note that  
332 Spanish voiced plosives, when not realised as approximants, have at most a  
333 brief period of occlusion [32]). It seems likely that the expansion of sounds  
334 consisting largely of near-silence is not favoured by the criterion of maxim-  
335 ing glimpsing opportunities embodied in the GCRetime algorithm. Overall,  
336 vowel durations increase proportionally less than those of consonants, proba-  
337 bly because their higher energy produces less of a need for masker-avoidance  
338 via retiming. Durational changes were not correlated with intelligibility gains  
339 at either SNR [adverse SNR: Pearson  $r = -0.01, p = .97$ ; moderate SNR:

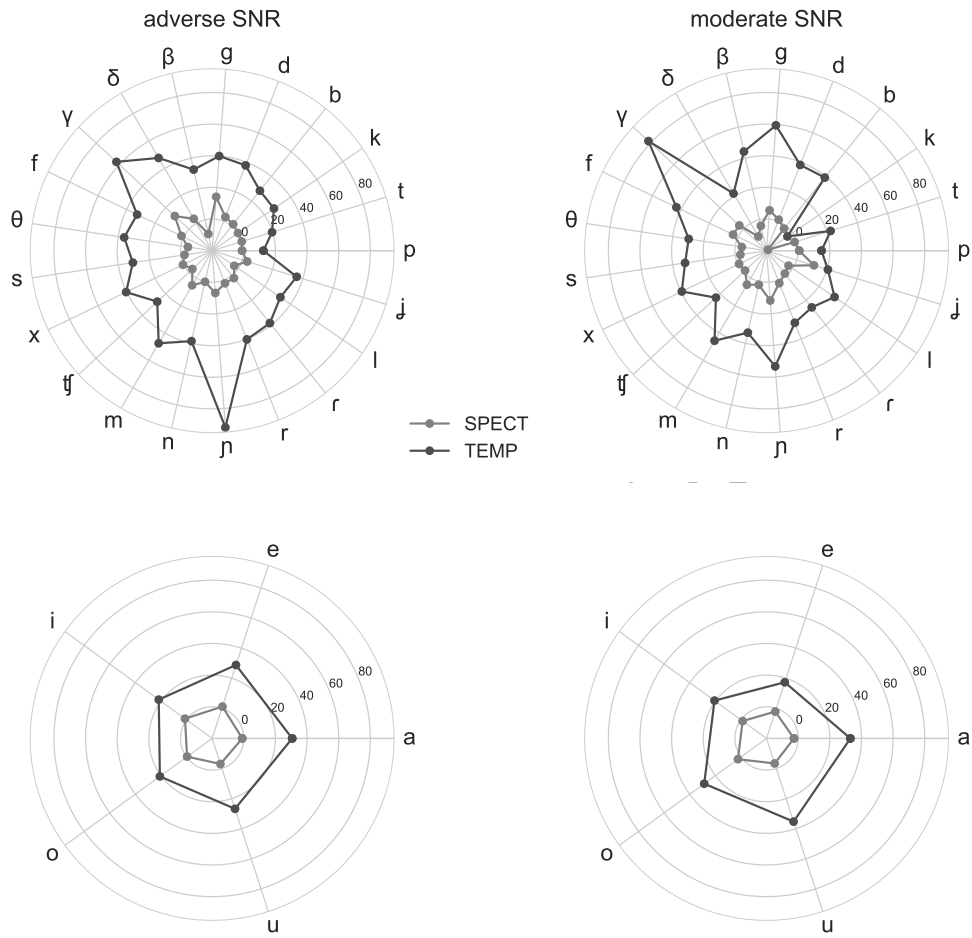


Figure 4: *Relative increases in duration, expressed in percentages, for the TEMP and SPECT conditions. The SPECT condition is included as a reference to indicate the scale of variations due to the forced alignment procedure (see text).*

<sup>340</sup>  $r = -0.31, p = .17$ ].

### 341 3.3. Independent gains?

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

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

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

## 351 4. Energetic masking

352 To explore the basis for intelligibility improvements, an analysis of ener-  
 353 getic masking was carried out using a glimpsing metric. Glimpsing measures  
 354 the degree to which a target signal exceeds the masker in time and frequency,  
 355 computed using an auditorily-inspired signal representation. Glimpse pro-  
 356 portion (GP) is the output of the initial stage of the glimpsing model of  
 357 speech perception [24] and has been used as proxy for energetic masking  
 358 in objective intelligibility metrics in applications involving speech synthesis  
 359 [e.g., 33], speech broadcasting [34], and estimation of binaural speech intelli-  
 360 gibility [35]. The starting point for GP computation is an auditory ratemap, a  
 361 time-frequency-energy representation of the speech and masker signals. The  
 362 ratemap is computed by passing the signal through a 55-channel gammatone

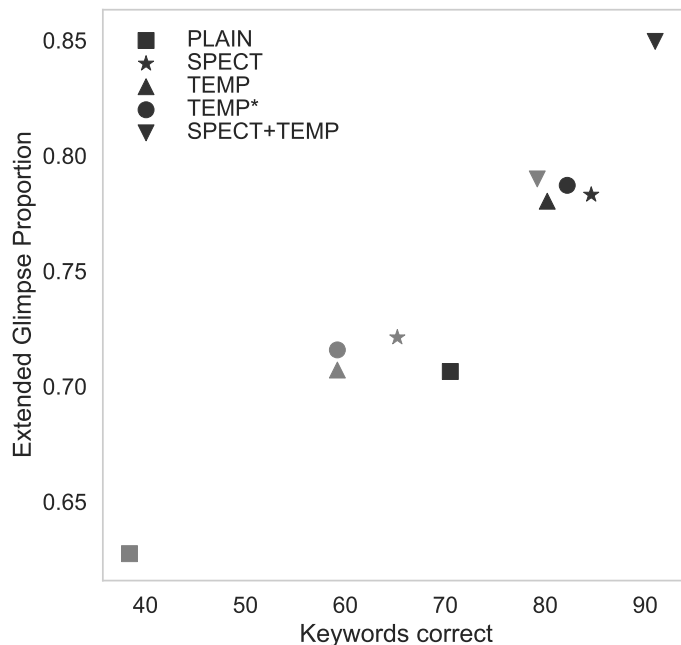


Figure 5: *Keyword scores plotted against intelligibility predictions from the extended glimpse proportion metric for the conditions of the experiment. Darker symbols come from the moderate SNR conditions.*

363 filterbank with filter centre frequencies arranged on an ERB-rate scale from  
 364 50 Hz to 8000 Hz. The instantaneous (Hilbert) envelope at the output of each  
 365 filter is smoothed with leaky integrator with time constant of 8 ms, downsam-  
 366 pled to 100 Hz and log-compressed. Ratemaps are produced independently  
 367 for speech and masker, and the proportion of time-frequency regions of the  
 368 ratemap for speech exceeding that of the masker by a local SNR threshold  
 369 (here set at 0 dB) defines the raw glimpse proportion.

370 The mean GP in each of the current set of 10 experimental conditions



371 (5 modifications including PLAIN  $\times$  2 SNRs) predicts intelligibility quite  
 372 well, with a Pearson correlation coefficient of 0.89 [ $p < .001$ ]. However, we  
 373 recently demonstrated that for a speech signal whose duration changes with  
 374 respect to a reference speech signal (in this case the PLAIN speech), better  
 375 predictions are possible using the extended GP metric, GP<sub>ext</sub> [36]. Amongst  
 376 other features, GP<sub>ext</sub> takes speech rate changes into account by weighting  
 377 glimpse proportion by a factor corresponding to the ratio of the modified  
 378 speech duration to the unmodified speech duration. For the conditions of  
 379 the current experiment, GP<sub>ext</sub> is highly-correlated with intelligibility [ $\rho =$   
 380  $.96, p < .001$ ], as shown in Figure 5. This outcome suggests that listeners'  
 381 performance in the task is dominated by peripheral energetic masking rather  
 382 than informational masking from the competing talker. Indeed, given both  
 383 the target-masker gender difference and the relatively adverse SNRs of the  
 384 current experiment, there seems little possibility that listeners were confusing  
 385 or misallocating speech material from the target and masker.

386 Continuing with the glimpse-based characterisation of the target-masker  
 387 relationship, the upper panel of Figure 6 presents marginal distributions  
 388 of raw (i.e., GP rather than GP<sub>ext</sub>) glimpse likelihoods as a function of  
 389 auditorily-scaled frequency for the adverse SNR condition (the pattern for  
 390 the moderate SNR is very similar). These 'GP spectra' are per-frequency-  
 391 channel means of GP measured across the entire corpus, for each modification  
 392 technique. The two temporal modification techniques (TEMP and TEMP\*)  
 393 produced very similar results; for clarity only TEMP is shown.

394 GP spectra reveal some clear differences between those modifications in-  
 395 volving spectral changes (SPECT and SPECT+TEMP) and the TEMP mod-

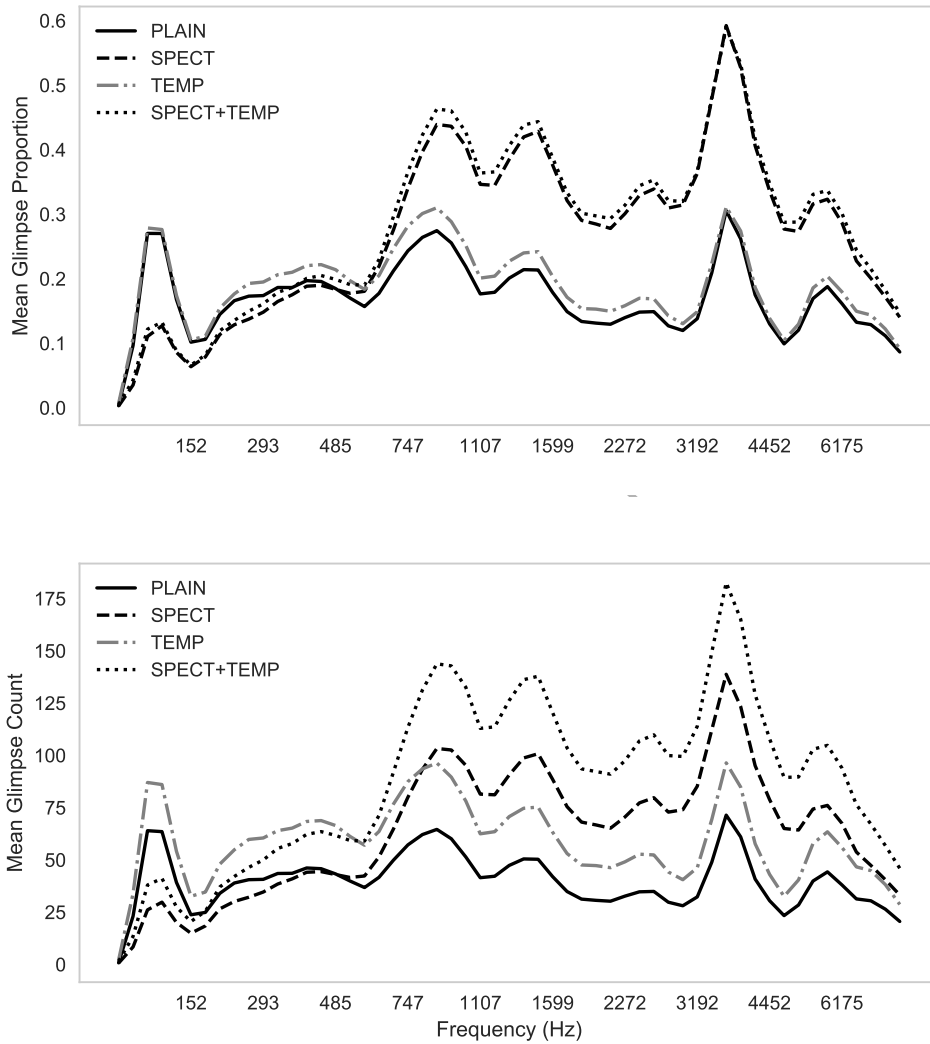


Figure 6: Mean glimpse proportion (upper panel) and mean glimpse count (lower panel) in each frequency channel for the adverse SNR condition.

396 ification approach. For the frequency region from 700 Hz upwards, spectral  
 397 techniques achieve a glimpse proportion of nearly double that of the tem-

398 poral modification, which in turn shows only a small advantage over the  
399 PLAIN baseline. However, the inverse pattern is seen below 500 Hz, with  
400 substantially fewer glimpses available as a result of spectral modification.  
401 These patterns suggest that much of the advantage of SSDRC stems from  
402 the transfer of energy from low frequencies (the first formant region and be-  
403 low) to mid and high frequencies (F2/F3 region and above). The fact that  
404 temporal modification produces only a modest gain over the unmodified base-  
405 line in terms of raw GP suggests that the intelligibility gains stemming from  
406 TEMP and TEMP\* are not due to spectrally-based increases in glimpsing  
407 opportunities. Instead, gains presumably come from durational changes, as  
408 indicated in the duration-sensitive  $GP_{ext}$  metric. The mean GP curves for  
409 SPECT+TEMP reflect an almost identical modest gain over SPECT as those  
410 seen for TEMP over PLAIN, supporting the idea that temporal processes em-  
411 bodied in the GCReTime algorithm act to a large degree independently of  
412 spectral changes in SSDRC.

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

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

423 used to produce the temporal modification path.

## 424 5. Discussion

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

440 The current experiment provides no evidence that specific groups of sounds  
441 benefit from the spectral, temporal or spectro-temporal modification algo-  
442 rithms under test. Gains, while not uniform, were observed for all con-  
443 sonants and vowels, with a ranking that closely mirrors across-consonant  
444 mean intelligibility scores. One possible explanation arises from the nature  
445 of fluctating maskers, where the main determiner of intelligibility is the local  
446 temporal relationship between target and masker. Compared to a stationary

447 masker, where high-energy phonemes are likely to escape masking most of  
448 the time while weaker sounds are more consistently masked, in the presence  
449 of a nonstationary masker with sufficient modulation depth (as is the case for  
450 competing speech) more intense sounds will suffer masking at least some of  
451 the time; similarly, fainter sounds will escape masking some of the time. An  
452 alternative and perhaps complementary reason as to why gains are spread  
453 across all phonemes comes from the fact that the task required listeners to  
454 identify words in sentences, thereby imposing morphological, lexical, syn-  
455 tactic and to a limited extent semantic constraints on their responses. In  
456 support of this notion, almost all errors at the phoneme level were deletions:  
457 the ratio of deletions to combined insertions + substitutions rose from 3.4  
458 for SPECT+TEMP at the moderate SNR level to 9.4 for PLAIN speech at  
459 the adverse SNR. Listeners clearly preferred to delete entire words than to  
460 hypothesise alternative candidates.

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

468 Finally, we note that SSDRC and GCReTime were chosen to represent  
469 spectral and temporal modification approaches respectively, but other choices  
470 merit investigation. We recently demonstrated that uniform elongation of  
471 speech (i.e. a uniform reduction in speech rate) is also an effective strategy

472 for intelligibility enhancement in fluctuating maskers [15], producing similar  
473 gains to GCReTime in a modulated noise condition. Uniform time-stretching  
474 was applied to SSDRC as part of the ‘uwSSDRcT’ technique reported in  
475 [37], but this combination did not increase intelligibility over SSDRC in a  
476 competing talker condition. However, uwSSDRcT also contained components  
477 to expand the vowel space and enhance transients, and it is possible that these  
478 interacted negatively with time-scale expansion. Future studies are needed to  
479 clarify whether imposing a slower speech rate on spectrally-modified speech  
480 leads to additional benefits.

## 481 **6. Conclusions**

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

## 489 **Acknowledgements**

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

494 **References**

- 495 [1] M. D. Skowronski, J. G. Harris, Applied principles of clear  
496 and Lombard speech for automated intelligibility enhancement in  
497 noisy environments, *Speech Communication* 48 (5) (2006) 549–558.  
498 doi:10.1016/j.specom.2005.09.003.
- 499 [2] B. Sauert, P. Vary, Near end listening enhancement: Speech intelli-  
500 gibility improvement in noisy environments, in: *Proc. ICASSP, Toulouse,*  
501 *France, 2006*, pp. 493–496. doi:10.1109/ICASSP.2006.1660065.
- 502 [3] H. Brouckxon, W. Verhelst, B. D. Schuymer, Time and frequency de-  
503 pendent amplification for speech intelligibility enhancement in noisy en-  
504 vironments, in: *Proc. Interspeech, Vol. 9, 2008*, pp. 557–560.
- 505 [4] C. H. Taal, J. Jensen, A. Leijon, On optimal linear filtering of speech for  
506 near-end listening enhancement, *IEEE Signal Proc. Let.* 20 (3) (2013)  
507 225–228.
- 508 [5] M. Cooke, C. Mayo, C. Valentini-Botinhao, Y. Stylianou, B. Sauert,  
509 Y. Tang, Evaluating the intelligibility benefit of speech modifications in  
510 known noise conditions, *Speech Communication* 55 (2013) 572–585.
- 511 [6] T. Zorila, V. Kandia, Y. Stylianou, Speech-in-noise intelligibility im-  
512 provement based on spectral shaping and dynamic range compression,  
513 in: *Proc. Interspeech, 2012*, pp. 635–638.
- 514 [7] M. A. Picheny, N. I. Durlach, L. D. Braida, Speaking clearly for the hard  
515 of hearing. I: Intelligibility differences between clear and conversational  
516 speech, *J. Speech Hear. Res.* 28 (1985) 96–103.

- 517 [8] Z. S. Bond, T. J. Moore, A note on the acoustic-phonetic characteristics  
518 of inadvertently clear speech, *Speech Communication* 14 (4) (1994) 325–  
519 337.
- 520 [9] R. M. Uchanski, Clear speech, in: D. B. Pisoni, R. E. Remez (Eds.),  
521 *The Handbook of Speech Perception*, Blackwell, Oxford, UK, 2005, pp.  
522 207–235.
- 523 [10] J. J. Dreher, J. J. O’Neill, Effects of ambient noise on speaker intel-  
524 ligibility for words and phrases, *J. Acoust. Soc. Am.* 29 (12) (1957)  
525 1320–1323.
- 526 [11] D. B. Pisoni, R. H. Bernacki, H. C. Nusbaum, M. Yuchtman, Some  
527 acoustic-phonetic correlates of speech produced in noise, in: *ICASSP*,  
528 Tampa, Florida, 1985, pp. 1581–1584.
- 529 [12] W. V. Summers, D. B. Pisoni, R. H. Bernacki, R. I. Pedlow, M. A.  
530 Stokes, Effects of noise on speech production: Acoustic and perceptual  
531 analyses, *J. Acoust. Soc. Am.* 84 (3) (1988) 917–928.
- 532 [13] V. Aubanel, M. Cooke, Information-preserving temporal reallocation of  
533 speech in the presence of fluctuating maskers, in: *Proc. Interspeech*,  
534 Lyon, France, 2013, pp. 3592–3596.
- 535 [14] R. M. Uchanski, S. S. Choi, L. D. Braida, C. M. Reed, N. I. Durlach,  
536 *Speaking clearly for the hard of hearing IV: Further studies of the role*  
537 *of speaking rate*, *J. Speech Hear. Res.* 39 (3) (1996) 494–509.
- 538 [15] M. Cooke, V. Aubanel, Effects of linear and nonlinear speech rate



- 539 changes on speech intelligibility in stationary and fluctuating maskers,  
540 J. Acoust. Soc. Am. 141 (2017) 4126–4135.
- 541 [16] Q. Summerfield, M. Haggard, On the dissociation of spectral and tem-  
542 poral cues to the voicing distinction in initial stop consonants, J. Acoust.  
543 Soc. Am. 62 (1977) 436–448.
- 544 [17] R. V. Shannon, F.-G. Zeng, V. Kamath, J. Wygonski, M. Ekelid, Speech  
545 recognition with primarily temporal cues, Science 270 (5234) (1995) 303–  
546 304.
- 547 [18] Y. Tang, C. Arnold, T. Cox, A study on the relationship between the  
548 intelligibility and quality of algorithmically-modified speech for normal  
549 hearing listeners, Journal of Otorhinolaryngology, Hearing and Balance  
550 Medicine 1 (2017) 5.
- 551 [19] M. Cooke, M. L. García Lecumberri, The effects of modified speech  
552 styles on intelligibility for non-native listeners, in: Proc. Interspeech,  
553 2016, pp. 868–872. doi:10.21437/Interspeech.2016-41.
- 554 [20] V. Aubanel, M. L. García Lecumberri, M. Cooke, The Sharvard Corpus:  
555 A phonemically-balanced Spanish sentence resource for audiology , Int.  
556 J. Audiology 53 (2014) 633–638.
- 557 [21] E. H. Rothausser, W. D. Chapman, N. Guttman, M. H. L. Hecker, K. S.  
558 Nordby, H. R. Silbiger, G. E. Urbanek, M. Weistock, V. E. McGee, U. P.  
559 Pacht, W. D. Voiers, IEEE Recommended practice for speech quality  
560 measurements, IEEE Trans. Audio Acoust. (1969) 225–246.

- 561 [22] A. Moreno, D. Poch, A. Bonafonte, E. Lleida, J. Llisterri, J. B. Marino,  
562 C. Nadeu, Albayzín speech database: Design of the phonetic corpus, in:  
563 Eurospeech, Berlin, Germany, 1993, pp. 175–178.
- 564 [23] B. A. Blesser, Audio dynamic range compression for minimum perceived  
565 distortion, *IEEE Trans. on Audio and Electroacoustics* 17 (1) (1969) 22–  
566 32.
- 567 [24] M. Cooke, A glimpsing model of speech perception in noise, *J. Acoust.*  
568 *Soc. Am.* 119 (3) (2006) 1562–1573.
- 569 [25] C. Stilp, K. Kluender, Cochlea-scaled entropy, not consonants, vowels,  
570 or time, best predicts speech intelligibility, *P. Natl. Acad. Sci. USA*  
571 107 (27) (2010) 12387–12392.
- 572 [26] M. Demol, W. Verhelst, K. Struyve, P. Verhoeve, Efficient non-uniform  
573 time-scaling of speech with WSOLA, in: *Int. Conf. on Speech and Com-*  
574 *puters (SPECOM)*, 2005, pp. 163–166.
- 575 [27] R. Perez Ramon, Haplo: Herramienta automática de procesamiento  
576 lingüístico ortofonético, in: *Proc. Asociación Española de Lingüística*  
577 *Aplicada*, Lleida, 2012.
- 578 [28] G. Studebaker, A rationalized arcsine transform, *Journal of Speech and*  
579 *Hearing Research* 28 (1985) 455–462.
- 580 [29] A. J. Oxenham, J. E. Boucher, H. A. Kreft, Speech intelligibility is best  
581 predicted by intensity, not cochlea-scaled entropy, *J. Acoust. Soc. Am.*  
582 142 (3) (2017) EL264–EL269.

- 583 [30] V. Aubanel, M. Cooke, C. Davis, J. Kim, Temporal factors in cochlea-  
584 scaled entropy and intensity-based intelligibility predictions, *J. Acoust.*  
585 *Soc. Am.* 143 (6) (2018) EL443–EL448. doi:10.1121/1.5041468.
- 586 [31] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, M. Sonderegger, Mon-  
587 treal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi, in:  
588 *Proc. Interspeech 2017*, 2017, pp. 498–502.
- 589 [32] J. I. Hualde, *The Sounds of Spanish*, Cambridge University Press, 2005.
- 590 [33] C. Valentini-Botinhao, R. Maia, J. Yamagishi, S. King, H. Zen, Cepstral  
591 analysis based on the Glimpse proportion measure for improving the in-  
592 telligibility of HMM-based synthetic speech in noise, in: *Proc. ICASSP*,  
593 2012, pp. 3997–4000.
- 594 [34] Y. Tang, B. Fazenda, T. Cox, Automatic speech-to-background ratio se-  
595 lection to maintain speech intelligibility in broadcasts using an objective  
596 intelligibility metric, *Applied Sciences* 8 (2018) 59.
- 597 [35] Y. Tang, Q. Liu, W. Wang, T. Cox, A non-intrusive method for estimat-  
598 ing binaural speech intelligibility from noise-corrupted signals captured  
599 by a pair of microphones, *Speech Communication* 96 (2017) 116–128.
- 600 [36] Y. Tang, M. Cooke, Glimpse-based metrics for predicting speech intel-  
601 ligibility in additive noise conditions, in: *Proc. Interspeech*, 2016, pp.  
602 2488–2492. doi:10.21437/Interspeech.2016-14.
- 603 [37] E. Godoy, Y. Stylianou, Increasing speech intelligibility via spectral  
604 shaping with frequency warping and dynamic range compression plus  
605 transient enhancement, in: *Proc. Interspeech*, 2013, pp. 3572–3576.