Effects of linear and nonlinear speech rate changes on speech intelligibility in stationary and fluctuating maskers

Martin Cooke, Vincent Aubanel

Effects of linear and nonlinear speech rate changes on speech intelligibility in stationary and fluctuating maskers

Martin Cooke*

Language and Speech Laboratory, Universidad del País Vasco, 01006 Vitoria, Spain

Vincent Aubanel

University of Grenoble Alpes, CNRS, GIPSA-lab, Grenoble, France

(Dated: May 2, 2017)

Abstract

Algorithmic modifications to the durational structure of speech designed to avoid intervals of intense masking lead to increases in intelligibility, but the basis for such gains is not clear. The current study addressed the possibility that the reduced information load produced by speech rate slowing might explain some or all of the benefits of durational modifications. The study also investigated the influence of masker stationarity on the effectiveness of durational changes. Listeners identified keywords in sentences that had undergone linear and nonlinear speech rate changes resulting in overall temporal lengthening in the presence of stationary and fluctuating maskers. Relative to unmodified speech, a slower speech rate produced no intelligibility gains for the stationary masker, suggesting that a reduction in information rate does not underlie intelligibility benefits of durationally-modified speech. However, both linear and nonlinear modifications led to substantial intelligibility increases in fluctuating noise. One possibility is that overall increases in speech duration provide no new phonetic information in stationary masking conditions, but that temporal fluctuations in the background increase the likelihood of glimpsing additional salient speech cues. Alternatively, listeners may have benefitted from an increase in the difference in speech rates between the target and background.

PACS numbers: 43.71.Gv,43.72.Dv

* m.cooke@ikerbasque.org; also at Ikerbasque (Basque Science Foundation).

I. INTRODUCTION

In speech communication scenarios involving the output of natural or synthetic speech, the likelihood of correct message reception in noisy environments can be improved by modifying the speech signal prior to output [e.g., 9, 23, 40, 42, 46, 49, 50, 52]. Such approaches are highly-effective: a recent evaluation of 18 algorithms demonstrated gains equivalent to increasing signal-to-noise ratio (SNR) by 5.1 dB for natural speech and by 5.6 dB for synthetic speech [The Hurricane Challenge; 14].

Most speech modification techniques operate by reallocating energy in time and frequency under a constant input-output RMS energy constraint. Energy reallocation aims to enhance intelligibility by manipulating the spectro-temporal pattern of local SNR, enabling weaker regions to rise above the masker at the expense of portions of the signal whose local SNR is already sufficiently high. Different approaches have variously transferred energy from voiced to voiceless regions [42], boosted some regions of the spectrum at the expense of others [47], enhanced formants [9], increased the amplitude modulation depth of the mid-frequencies [25], or employed dynamic range compression [7] which has the effect of transferring energy from intense to weaker temporal epochs [41, 51]. [13] provides a review of human and algorithmic speech modifications.

An alternative to spectro-temporal energy reallocation is the modification of segment or sub-segmental durations. Altered speaking styles such as Lombard speech [e.g., 24, 36, 45], clear speech [35, 48], speech directed at infants [e.g., 21] and speech produced at a distance [19], exhibit durational changes, usually resulting in slower speech, both overall and at the level of individual speech segments. Many of these forms of speech have been found to be more intelligible than unmodified plain speech [17, 26, 35, 37, 43]. While acoustic changes to features such as intensity, spectral tilt and prosody are present in altered speech styles and may play a role in increased intelligibility, it is natural to consider whether durational changes contribute to the improvement. Durational manipulations can be expected to be particularly effective in the presence of fluctuating maskers where the opportunity arises to shift phonetic information in time to regions where the masking source is less intense.

Approaches employing durational modifications are less common than those exploiting spectro-temporal energy reallocation. Of the 14 natural speech modification approaches submitted to the aforementioned Hurricane Challenge [14], only three involved significant

durational changes. Tellingly, two of these three produced the largest enhancements in the fluctuating masker condition of the Challenge at moderate and adverse SNRs. Indeed, gains of up to 4.4 dB resulted from an approach [GCRetime; 3] that modified durational information only, indicating that alterations to segment durations alone can be a valuable strategy for maskers containing low-frequency temporal modulations.

However, the basis for the intelligibility enhancements produced by durational changes is currently unclear. It is possible that listeners are able to take advantage of the reduced information rate of slower speech rate rather than the intended energetic masking release produced by shifting information in time. Evidence for intelligibility benefits of speech rate slowing is mixed. While studies by Adams and colleagues [1, 2] have demonstrated intelligibility increases for slow speech in masking noise, no such effect was observed under conditions of simulated hearing loss by [32] nor when linear and nonlinear durational changes observed in Lombard speech were mapped on to plain speech [15]. Intriguingly, while the latter studies used stationary maskers, the sentence material used by [2] and [1] was mixed with four-talker babble, leading to the possibility that the temporal modulation characteristics of the masker played a role in the different outcomes.

One goal of the current study was to determine whether a slower speech rate *per se* contributes to the intelligibility increases observed in durationally-modified speech. Keyword scores in utterances that had been linearly-elongated were compared with those for utterances whose duration was locally-modified in a way designed to minimise energetic masking. If a reduced information rate is responsible for intelligibility gains of durationally-modified speech, we predict that such gains would be observed in linearly-elongated speech, since durational modifications in this case are independent from masker fluctuations.

The current study also addressed the issue of whether the intelligibility of durationally-modified speech is affected by the properties of the masker. Utterances were presented in stationary noise and two forms of fluctuating noise: competing speech, and speech-modulated noise with temporal envelope fluctuations matching those of competing speech.

II. EXPERIMENT 1: DURATIONAL MODIFICATIONS IN STATIONARY AND FLUCTUATING MASKERS

A. Durational modifications

Listeners heard sentences which were either unmodified (PLAIN), linearly-stretched (ELONGATED) or nonlinearly-modified (RETIMED). All durational modifications were carried out using the WSOLA algorithm [16] via a sequence of time-scale factors. In the elongation condition, a constant time-scale factor was used, while in the retiming case the time-scale factor sequence was derived from the GCRetime algorithm described in [3] and summarised below.

GCRetime is a general-purpose algorithm that takes a pair of acoustic signals and outputs a retimed version of one of them based on the result of optimising a user-defined criterion operating on a comparison of the two input signals. In the current context, the input to GCRetime is a target speech signal and a masker, and the output is a retimed speech signal which maximises a local distance function whose goal to promote the audibility of information-bearing parts of the speech in the presence of the masker. The distance function is maximised using dynamic programming, the end result being a retiming path which defines a sequence of expansions and contractions of the target speech signal. The process is illustrated in Figure 1. Here, the masker (shown at the top of the figure) is a competing speech signal. The target speech and the modified (retimed) speech are drawn on the left and bottom edges of the figure respectively. The unmodified PLAIN condition corresponds to the diagonal path.

The GCRetime local distance function $D(i, j)$ is defined on a grid of points i, j corresponding to the ith frame of the target speech signal s and the jth frame of the masker m. The local distance function for all possible pairs (i, j) is a matrix, shown as a grayscale image in Fig. 1, where darker regions depict higher values of the function. The local distance function is composed of two components quantifying (1) the masked audibility of the speech signal in frame i in the presence of the masker at frame j, and (2) the informativeness of the speech signal in the vicinity of frame i. The first of these components is operationalised using the glimpse proportion [12], while the second makes use of cochlear-scaled entropy [CSE; 44]; together these components are reflected in the name 'GCRetime'.

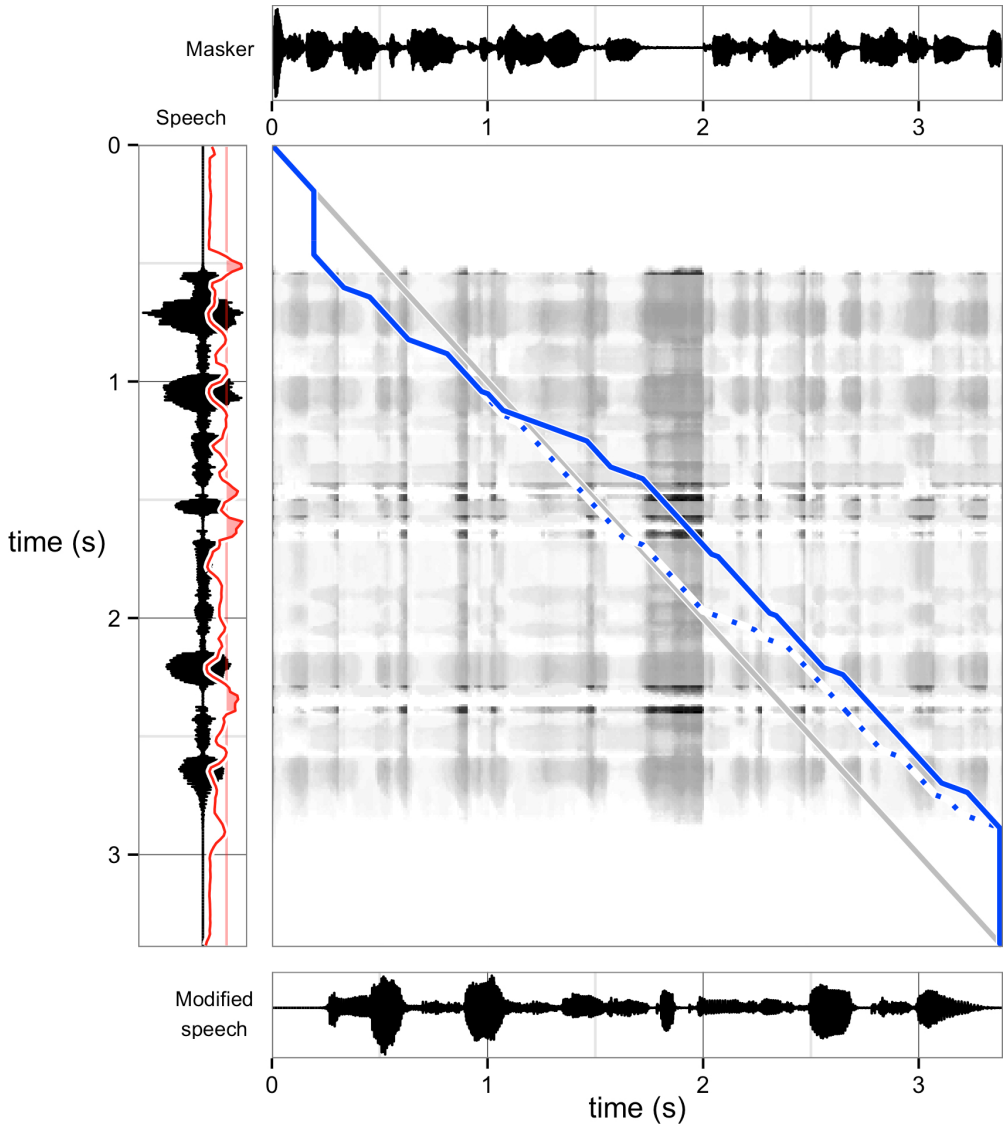

FIG. 1. *An illustration of sentence retiming in the face of a competing speech masker. The grayscale image depicts the value of the local cost function (Eq. A1) for all possible pairs of frames of the target and masker. The solid line shows the minimum cost retiming path using the glimpse proportion (GP) and cochlear-scaled entropy (CSE) components while the dotted line shows the path for the GP component alone. The red curve indicates the value of the CSE weighting defined by Eq. A3; the pink vertical line in the left panel indicates the value of the threshold used to select high-CSE regions (see Appendix). [color online]*

CSE captures localised spectral change and has been shown to predict intelligibility better than consonants, vowels or consonant-vowel/vowel-consonant transitions when tested using a noise-replacement paradigm [44]. Taken together, these two components ensure that the distance function takes on high values when the speech signal is not masked and when it is undergoing a period of rapid change. For example, the dark vertical band in the period

immediately preceding the 2 s point in the masker is due to the low level of the masker in that interval, and the darker horizontal strips within this band correspond to those portions of the target speech with a high CSE value. The path that maximises the global distance passes through some of these regions, effectively ensuring that potentially high-information-value transients in the target speech are shifted in time to regions where the masker is less intense.

Appendix A describes the computation of the GCRetime local distance function D in more detail.

B. Speech and masker materials

Utterances were drawn from the phonemically-balanced Sharvard corpus [5] which consists of Spanish sentences designed to be equivalent in difficulty to the Harvard sentences [38]. Each Sharvard sentence contains five keywords used for scoring; an example (keywords underlined) is "Llene el frasco de cristal con cola densa" ("Fill the glass flask with thick glue"). Spectrograms of this utterance in each of the three styles PLAIN, ELONGATED and RETIMED are shown in Fig. 2. Renditions of the first 243 Sharvard utterances read by a native Spanish male talker were used in Experiment 1; this figure includes sentences used as practice items.

Maskers were constructed using speech material from a native Spanish female talker who read sentences from the Albayzin corpus [30]. Inter-sentence pauses were removed and sentences concatenated to produce a signal of 13.83 minutes duration, sufficient to ensure that no masker fragment was repeated in any speech-plus-masker mixture. Successive non-overlapping fragments from this signal were used for the competing speech masking condition (CS). A speech-shaped noise (SSN) masker was constructed by passing white noise through a filter with a long-term spectrum matching that of the female talker. Each CS fragment had a matched speech-modulated noise (SMN) signal formed by multiplying the short-term temporal envelope of the CS fragment with a portion of the SSN signal selected at random. All speech and masker materials were sampled at 16 kHz.

The average PLAIN sentence duration was 2.34 s (s.d. 0.29 s). To allow for overall elongation, maskers were constructed to have a duration 0.8 s longer than the target speech utterances they were paired with. Sentences in the ELONGATED and RETIMED conditions

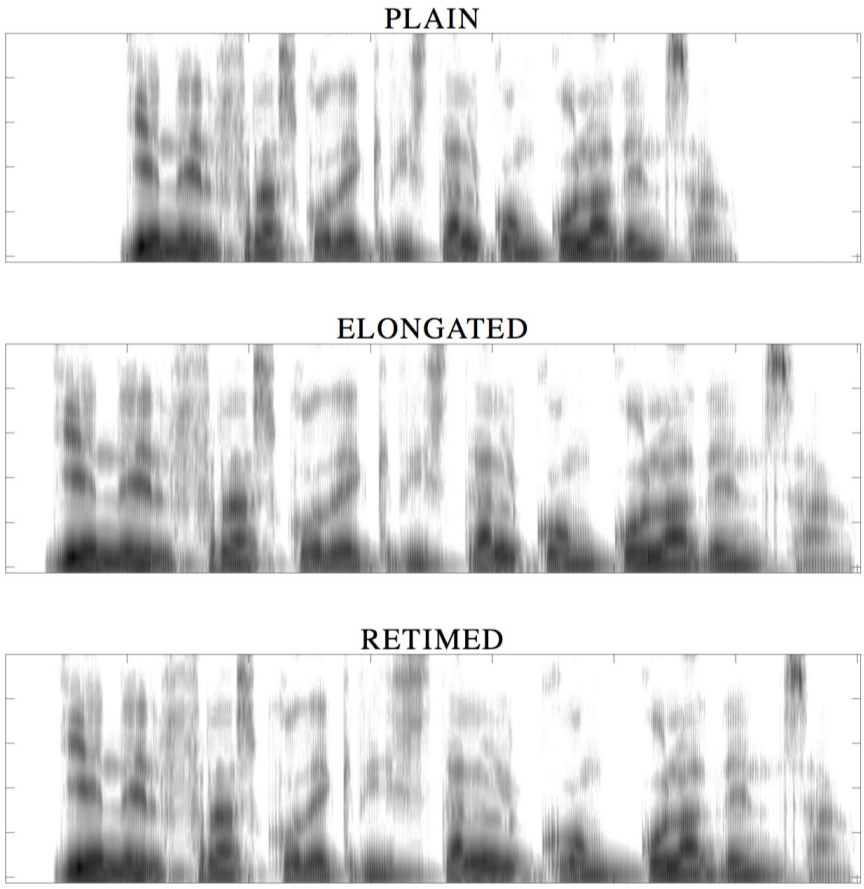

FIG. 2. *Example spectrograms of the utterance "Llene el frasco de cristal con cola densa" in each of the three durational modification conditions.*

were 24-55% longer than their PLAIN counterparts (mean 34%, s.d. 4%). Each RETIMED sentence had a duration that was 97.4-99.2% of the equivalent ELONGATED sentence (mean 98.6%). In Experiment 1, regardless of the masker (CS, SMN or SSN), retiming was carried out using the CS masker.

C. Participants

Eighteen native Spanish speakers (10 female) with a mean age of 22.3 years (s.d.=3.8) took part in the experiment. Speakers were either monolingual in Spanish or bilingual in Spanish and Basque. All listeners had normal hearing thresholds (< 20 dB HL) in the range of 125 Hz to 8 kHz, as tested with an Interacoustics AS608 screening audiometer. Listeners were paid for their participation. Ethics permission was obtained following the University of the Basque Country ethics procedure.

D. Procedure

Listeners heard a total of 234 utterances made up of 26 sentences in each of the nine conditions resulting from the combination of the PLAIN, ELONGATED and RETIMED manipulations with the three maskers (CS, SMN and SSN). The SNR for the SSN masking conditions was set to -6.5 dB, a value which led to a 50% keyword score for the male talker in [5]. Since competing speech is a less effective masker than SSN when presented at the same SNR, the SNR for the CS masker was set following pilot tests to -17 dB, while similar tests indicated the need for an intermediate SNR for the SMN case of -12 dB.

To avoid sentence subset effects due to possible differing intrinsic intelligibilities of the speech material, the three speech processing conditions (PLAIN, ELONGATED, RETIMED) were applied to the complete set of 234 utterances. Listeners were assigned to subsets of sentences in such a way as to ensure that each sentence in each processing condition was heard the same number of times across listeners, and that each listener heard each sentence exactly once. Speech-plus-noise stimuli were blocked by masker type; within each block listeners heard equal numbers of sentences from each of the three processing conditions in a randomised order. Immediately prior to each block of 78 sentences, listeners responded to 3 unscored practice stimuli designed to familiarise them with the type of masker. Practice sentences did not occur elsewhere in the main experiment. Presentation order of the three blocks was balanced across listeners.

The listening experiment was conducted in a sound-attenuated studio in the Phonetics Laboratory at the University of the Basque Country, Spain. Speech-plus-noise stimuli were delivered diotically at a presentation level in the range 70.8-71.7 dB(A) through Sennheiser HD 380 pro headphones. Listeners received on-screen instructions prior to each block. The experiment ran under computer control using a custom MATLAB program. The experiment was self-paced: following each stimulus presentation participants typed their answer into a text box, after which the next stimulus was presented. On average, listeners required 47 minutes (s.d.= 7) to complete the three blocks.

E. Postprocessing

Listener responses were scored automatically based on the number of keywords identified correctly in each sentence. Vowel stress marks were removed prior to scoring, so that, for example, both 'mas' and 'más' were considered to be correct responses for the word 'más'. A total of 130 keywords were scored (26 sentences times 5 keywords per sentence) in each of the nine conditions. Scores were expressed as percentages of keywords identified correctly in each condition. Since none of the scores lay outside the range 23-83%, raw (untransformed) percentages were used in subsequent statistical analyses.

F. Results

Keyword scores for the PLAIN speech condition were 46.1%, 37.8% and 51.6% for the CS, SMN and SSN maskers respectively.

Figure 3 plots changes in scores over the PLAIN baseline for ELONGATED and RETIMED sentences in the three maskers. Elongation led to a small gain in keyword scores of 3.0 percentage points (p.p.) in the SSN condition. Substantially larger gains of 8.3 and 9.0 p.p. were observed in the two temporally-modulated masking conditions CS and SMN respectively.

Retimed speech produced a larger spread of differences over the PLAIN baseline across the three maskers. In stationary noise, retiming was highly detrimental to intelligibility, producing a loss of 14.9 p.p. compared to unmodified speech. For the modulated noise masker (SMN) the gain of 10.3 p.p. was similar to that seen for the ELONGATED condition. However, with a gain of 16.3 p.p., retimed utterances were substantially more intelligible than their elongated counterparts in the competing speech masker.

An ANOVA of changes-over-baseline scores with within-subjects factors of modification method (ELONGATED, RETIMED) and masker (CS, SMN, SSN) demonstrated a clear interaction in the effect of modifications and maskers [$F(2, 34) = 37.7, p < .001, \eta^2 = 0.28, MSE = 43.1$], with significant main effects of both modification type [$F(1, 17) = 10.8, p < .01, \eta^2 = 0.03, MSE = 20.8$] and masker [$F(2, 34) = 21.2, p < .001, \eta^2 = 0.46, MSE = 164.4$]. Based on a Fisher's Least Significant Difference (LSD) of 4.4 p.p., gains for ELONGATED speech in the two modulated maskers were equivalent, while ELONGATED speech was statistically-

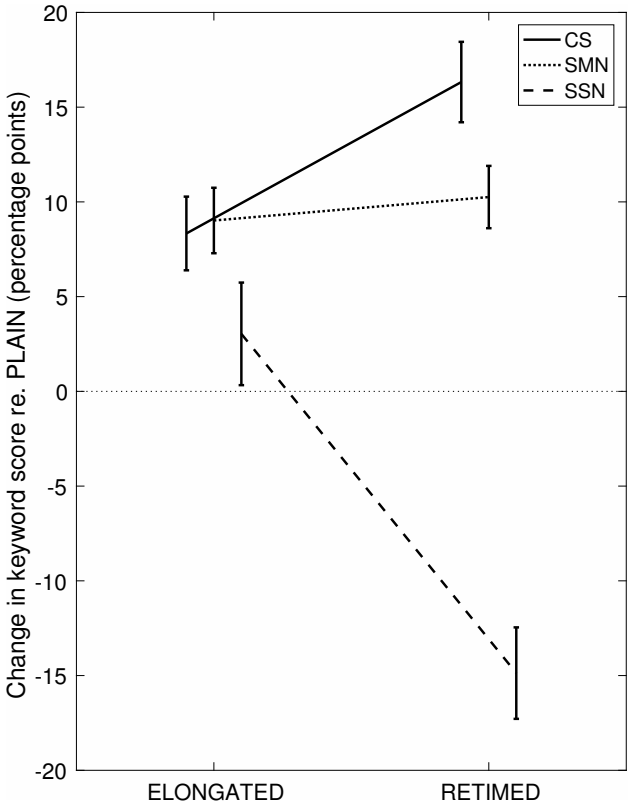

FIG. 3. *Changes in mean keywords correct relative to* PLAIN *speech for* ELONGATED *and* RETIMED *utterances in the presence of competing speech (CS), speech-modulated noise (SMN) and speech shaped noise (SSN) maskers. Error bars here and in Fig. 5 represent* ± 1 *standard error.*

equivalent to the PLAIN baseline for the SSN masker. Changes in keyword scores for RETIMED speech were significantly different in the three masking conditions.

G. Interim discussion

A strategy of retiming speech information by shifting the waveform nonlinearly in time to attenuate the effect of intense masker epochs has previously been shown to produce substantial intelligibility gains in the presence of fluctuating maskers [3]. Experiment 1 confirms the effectiveness of algorithmic speech retiming, and extends this finding to speech material in a different language: the 16.3 p.p. gain produced for Spanish sentences in the RETIMED condition of the current study in the CS masker condition at an SNR of -17 dB is consistent with the improvements of 16 and 18 p.p. observed in [3] for English sentences at SNRs of -14 and -21 dB in the equivalent masking condition of that study.

Elongation of speech had a negligible impact on intelligibility for stationary maskers,

suggesting that a slower speech rate in itself is not responsible for the gains observed when speech is retimed. In contrast, for fluctuating maskers, elongation led to a clear increase in intelligibility. This finding goes some way to explaining the discrepancies among earlier studies on the effectiveness of a slower speech rate in noise. As noted in the Introduction, while [32] and [15] failed to find an intelligibility benefit of slower speech when presented in a stationary masker, [1] reported a beneficial effect of a slower speech rate in four-talker babble, a type of masker that shows a greater temporal modulation depth than that of a purely stationary noise. The issue of how a fluctuating masker might promote intelligibility increases for elongated speech is addressed in the General Discussion.

One intriguing finding is the observation of substantially larger gains produced by retiming in the CS condition than in the SMN condition. This outcome would be unexpected if the gains in a fluctuating masker were derived solely from shifting speech information in time to avoid more intense masker intervals. However, since the glimpse proportion analysis underlying GCRetime operates in the spectro-temporal domain, a retiming path produced by a CS masker is not necessarily the same as that produced in response to a SMN masker in spite of the latter having the temporal modulations of the former. Compared to SMN, CS contains some variation in the spectrum across time due to its spectral fine structure of peaks and dips, and it is possible that the retiming path suggested by the GCRetime algorithm is able to take advantage of the glimpsing opportunities afforded in both the spectral and temporal domains. Consequently, the retiming path for CS may be suboptimal for SMN, and vice versa (see Fig. 4). The fact that a competing speech signal was used for retiming in Experiment 1 may have favoured retimed speech when presented in a competing speech masker. To test this hypothesis, a second experiment examined the role of the retiming masker using matched and mismatched retiming maskers.

III. EXPERIMENT 2: ROLE OF THE RETIMING MASKER

A. Listeners

A new cohort of twenty-one normal hearing paid native Spanish speakers (16 female) with a mean age of 20.0 years (s.d.=1.5) and the same profile as the participants of Experiment 1 took part in Experiment 2. Results from one participant who treated the competing speech

masker as the target in a number of conditions were excluded.

B. Materials and methods

Raw speech materials, maskers and SNRs were the same as those used in Experiment 1. The ELONGATED condition was not tested. Instead, listeners heard utterances in unmodified PLAIN form and in two distinct retiming conditions. In one retiming condition (CS RETIMED) the durational modifications were based on GCRetime using a CS masker, while in the other retiming condition (SMN RETIMED) the modifications results from the counterpart SMN masker. In this way, listeners heard sentences retimed by a matched or unmatched masker. The nine experimental conditions (3 modifications x 3 maskers) were presented to listeners using the blocking and balancing procedure of Experiment 1 as described in Sec. II D.

Figure 4 illustrates the matched/mismatched retiming procedure for the case where retimed speech was presented in the SMN masker. A comparison of the speech retimed by the SMN masker (matched condition, second panel) and that retimed by the CS masker (mismatched condition, fourth panel) shows that although the SMN masker is derived from the CS masker, the retimed speech in the matched and mismatched conditions display different temporal structures.

C. Results

Mean keywords correct scores for the PLAIN speech condition were 49.6%, 42.5% and 55.8% for the CS, SMN and SSN presentation maskers respectively. Figure 5 plots changes in scores over the PLAIN baseline for sentences retimed using the CS and SMN maskers for each of the three presentation maskers. Changes over baseline for the CS RETIMED conditions were similar to those observed in the equivalent conditions of Experiment 1 (CS, 16.3 vs. 16.5 p.p.; SMN 10.2 vs. 10.8; SSN -14.9 vs -12.5), confirming the findings of the first experiment with a different listener cohort.

A within-subjects ANOVA with factors of retiming masker and presentation masker for the two fluctuating masking conditions (CS and SMN) indicated no main effect of either factor, but revealed a significant interaction between the two factors $[F(1, 19) = 6.96, p < .05, \eta^2 = 0.048, MSE = 32.9]$. Post-hoc tests based on a Fisher's LSD of 3.80 percentage

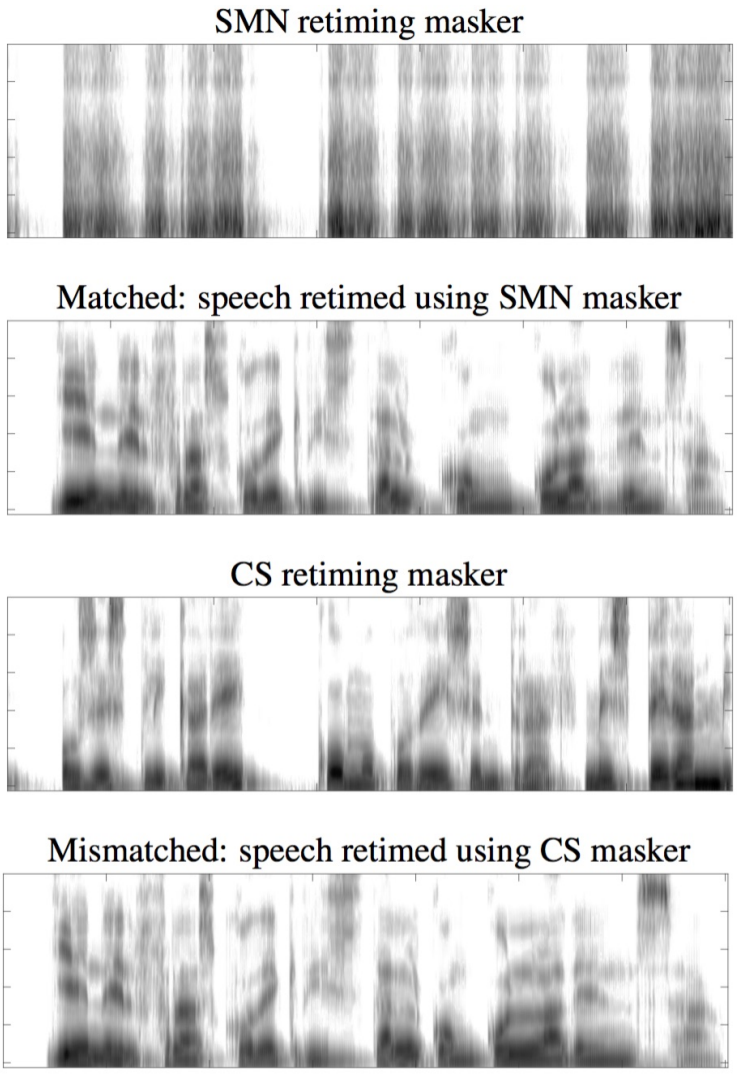

FIG. 4. *Matched and mismatched retiming for the utterance shown in Fig. 2. The top panel shows the SMN masker used to produce the retiming path which results in the utterance shown in the second panel (*SMN RETIMED*). The third panel shows the CS masker used for retiming which results in the utterance shown in the lower panel (*CS RETIMED*).*

points indicate that gains for CS-based retiming was more effective in a matched CS masker (16.5 p.p.) than in a mismatched SMN masker (10.9 p.p.). However, there was no benefit of matched masker type for SMN-based retiming, with similar gains of 13.3 and 12.2 p.p. in the matched and unmatched conditions respectively. Critically, CS RETIMED speech led to higher gains than SMN RETIMED speech when presented in a CS masker, suggesting that the specific details of the retiming path are important. Gains in the two matched conditions (i.e., CS RETIMED in CS masker and SMN RETIMED in SMN masker) did not differ statistically, the difference of 3.2 p.p. falling short of the critical LSD value.

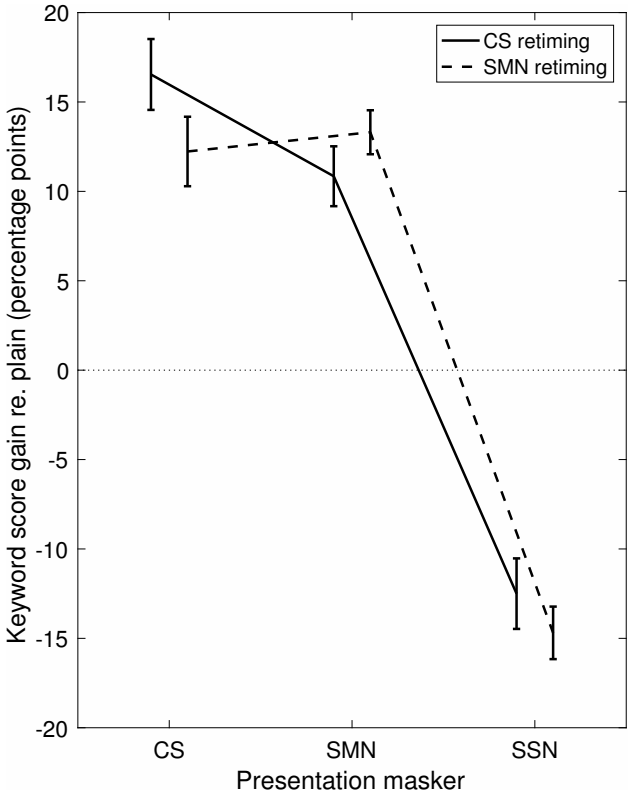

FIG. 5. *Changes in keywords correct scores relative to unmodified plain speech for utterances retimed using the CS or SMN masker, for the three presentation maskers.*

D. Interim discussion

Experiment 2 demonstrated that the benefits of retiming are affected by the relationship between the retiming masker and the presentation masker in the case of the CS masker but not for the SMN masker. This outcome suggests that there is a limit to the benefits of retiming for a temporally-modulated noise masker, while for CS there may be both temporal and spectral opportunities which are taken into account by the energetic masking model underlying the glimpse proportion calculation.

IV. GENERAL DISCUSSION

Experiment 1 addressed the primary research question of the current study by measuring the extent to which intelligibility gains are present for speech that is linearly elongated to generate the same average speech rate as that produced by retiming. The absence of a benefit of elongated speech in the presence of a stationary speech-shaped masker appears to

rule out reduced speech rate *per se* as a contributory factor.

Nevertheless, elongation led to significant increases in keyword scores in fluctuating maskers, demonstrating that the intelligibility benefits of retiming are not entirely due to the deliberate noise avoidance encapsulated in the GCRetime algorithm. This outcome suggests that reducing speech rate can be a very effective strategy for increasing intelligibility in real-life situations characterised by non-stationary sources of noise.

There are several ways in which a temporally-fluctuating masker might promote intelligibility gains for slowed speech while a stationary masker does not. One possibility is that the regions of elongated speech which escape masking by a stationary noise provide no new phonetic information. The upper two panels of Fig. 6 depict glimpses of speech in the presence of the SSN masker for the example utterance, both unmodified and elongated. It is clear that while small differences in putative glimpses exist due to fluctuations in the speech-shaped noise, the nature of the available information is largely identical in the PLAIN and ELONGATED conditions: the glimpses are simply elongated. For example, the phoneme /k/ in 'frasco' is devoid of glimpses in both cases, and the /s/ in the same word conveys the same information in the two cases. In contrast, for the competing speech masker (lower panels), temporal fluctuations in the masker increase the likelihood of observing new phonetic information in elongated speech. For example, in the PLAIN speech, there is a paucity of critical low frequency information to indicate the identity of the vowel /e/ in 'densa', while such information is present in the ELONGATED version. Of course, while some information is gained in this way, other regions of the signal are likely to be masked with a commensurate loss of information. However, we speculate that since the overall signal duration is increased in the ELONGATED case, so is the net amount of phonetic information.

An alternative explanation for the observed gains in fluctuating maskers arises from the possibility that listeners are better able to separate target and background speech due to speech rate differences between the target and masker, overcoming a potential source of informational masking. This notion is supported by a study by [20] in which a cohort of young normal hearing listeners recognised more words in time-compressed sentences when the compression ratio did not match that of 12-talker background babble. The ELONGATED condition of the current study does indeed lead to an increase in speech rate differences between the target and background: the PLAIN speech was articulated at a rate of 5.8 vowels/s, comparable to the 6.6 vowels/s of the competing speech masker. Speech rate slowing

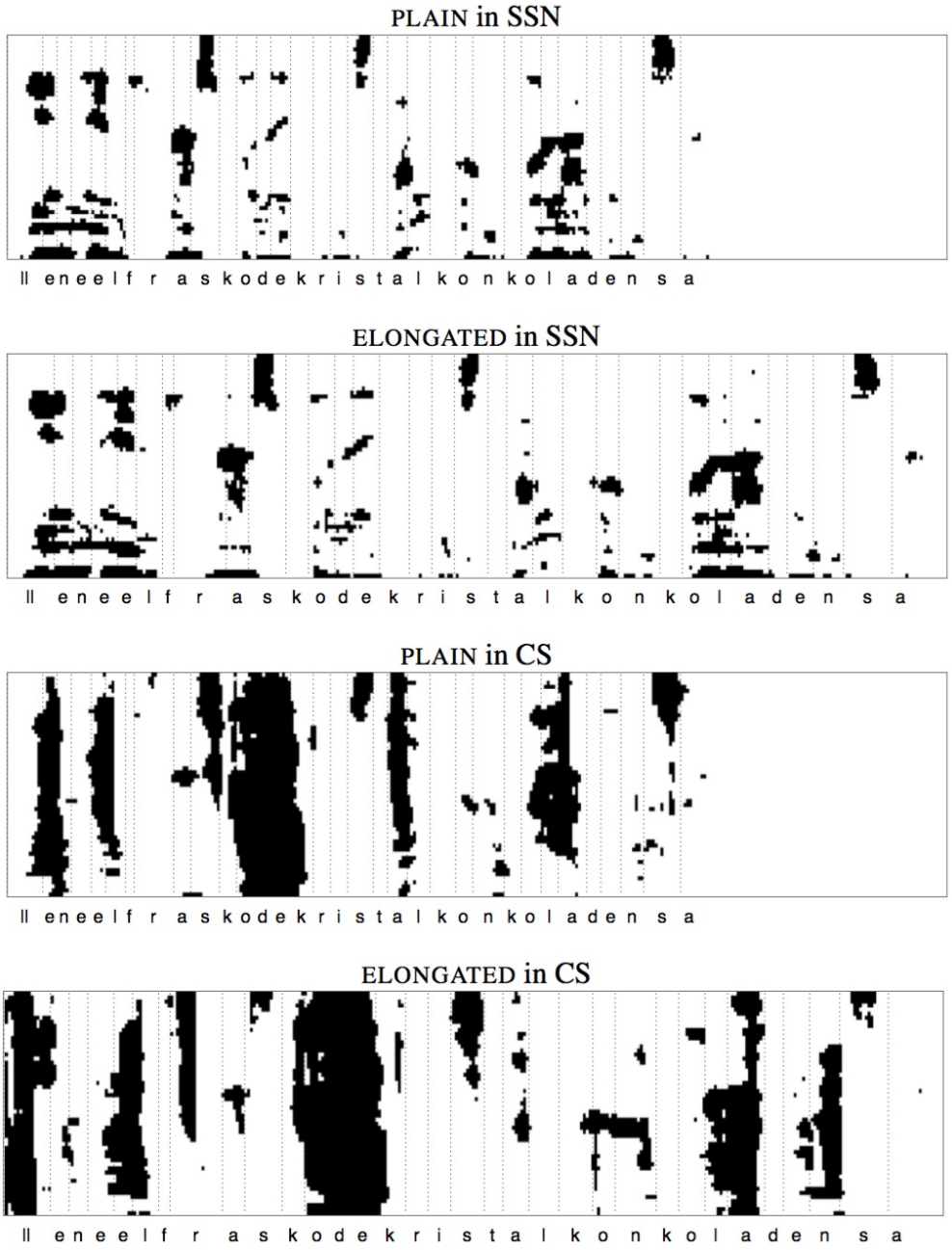

FIG. 6. *Regions of a target speech utterance of the phrase "Llene el frasco de cristal con cola densa" which are deemed to escape energetic masking according to a glimpsing model [12], for* PLAIN *and* ELONGATED *speech in the presence of a stationary masker (top two panels) and a fluctuating masker (lower two panels). A broad phoneme-level transcription is provided in each case.*

in the ELONGATED condition reduced the average speech rate to 4.4 vowels/s, increasing the target-background speech rate difference. Earlier studies with listeners [28] or models [8] have demonstrated sensitivity to differences in speech interruption rate and speech modulation rate. Further studies controlling for speech rate differences are needed to rule out their

possible contribution to the observed gains in the current study. However, since gains were also observed for the non-speech masker, it appears necessary to invoke a more generalised notion of temporal modulation rate differences that go beyond speech-on-speech informational masking, which in this case is likely to have a relatively small effect since the gender of the target and masking talkers differed [10]. A related possibility is that the presence of modulation in the masker itself imposes a cognitive burden on listeners: a slower rate of information transmission via a lower speech rate may be beneficial in reducing listening effort.

A striking and consistent outcome observed in both experiments was the substantial loss of intelligibility (amounting to 13-15 percentage points) that occurred when speech was nonlinearly-retimed in stationary noise, contrasting with no loss for linearly-elongated speech in the presence of the stationary masker. In the GCRetime algorithm, retiming takes no account of anything other than the temporal relationship between speech and masker (the glimpsing component), and speech dynamics (the CSE component). Consequently, properties such as segment duration and the local speech rate of unmodified speech are not preserved by the algorithm. Such distortions are likely both to confound listeners' expectations of when salient information is going to occur and to diminish the effectiveness of contextual cues that depend on relative durations. For Spanish, changes in relative segment durations induced by GCRetime may have interfered with phonological cues [e.g., 27] or syllabification [e.g., 22]. Intriguingly, such a tradeoff between modifications that overcome masking and those which preserve phonological integrity is also seen in naturally-produced speech in noise. [39] measured durational (vowel lengthening) cues to the voicing distinction in English plosives in plain and Lombard speech, finding a reduced contrast in the latter case. The benefits of retiming in fluctuating maskers presumably reflect a net effect of masking release and durational distortion, suggesting that even larger benefits in noise are realisable if the phonological impact of durational modifications can be minimised.

Distortions to the target speech might also have contributed to the observed differences in effectiveness of retiming in the presence of competing speech maskers and temporally-modulated noise, even when the influence of a matching or mismatching retiming masker was controlled for, as in Experiment 2. In a non-informational masker such as modulated noise, listeners' attentional focus is presumably directed to the target speech alone, and any departures from expected phonological forms may be noticeable, and potentially lead to

the consideration of additional competitor words. In contrast, when the masker itself contains speech, it is conceivable that the cognitive burden imposed by foreground-background separation precludes a more detailed analysis of the target signal, or a mis-attribution of retiming-based distortions to the competing speech signal. Another possibility is that any gains due to retiming outweigh losses due to mistiming of phonological features.

The outcome of the current study points to the potential of durational changes as a mechanism for improving intelligibility in noise, but also highlights the need to take the temporal properties of the masker into account, given the deficits resulting from the retiming method in the presence of stationary noise. The finding that gains are possible merely by elongating the speech signal in fluctuating maskers suggests that speech rate slowing could be a component of a simple practical strategy for boosting intelligibility. As noted in the Introduction, modified duration is not by any means the sole manifestation of natural 'altered' speaking styles, and spectral factors in particular are known to have a sizeable influence on intelligibility [15]. Spectral and durational changes are orthogonal to a large extent e.g., changes to properties such as spectral tilt can be imposed independently of durational changes.

As is typically the case when targetting the 50% correct response rate with a normal-hearing adult population, all testing was done at negative SNRs. Further work is needed to measure the efficacy of a slower speech rate at more realistic SNRs [31], as such environments have also been found to induce slower rate of speech in talkers [4]. The benefits observed in the current study of nonlinear retiming at negative SNRs may be reduced at higher SNRs; lower than expected benefits for a fluctuating masker advantage in comparison to stationary noise have consistently been observed at positive SNRs [6, 18, 33].

V. CONCLUSIONS

(i) Reductions in speech rate resulting from linear elongation of the speech signal did not lead to intelligibility increases (nor did they disrupt intelligibility) for sentences in the presence of stationary speech-shaped maskers, suggesting that intelligibility gains seen in durationally-modified speech were not due to the reduction in information rate that accompanies slower speech.

(ii) However, identical elongations produced significant intelligibility increases in fluctu-

ating maskers. One explanation is that while elongation in stationary maskers produces no new speech information, the altered pattern of glimpses in fluctuating maskers leads to the unmasking of new phonetic cues. An alternative is that slower speech enables listeners to separate target speech from the background due to greater differences in speech rate, or reduces the cognitive burden of processing speech in a modulated background. Further studies are needed to distinguish these possibilities.

(iii) Nonlinear durational modifications designed to reduce energetic masking of speech information led to larger intelligibility gains in competing speech maskers than those produced by linear elongation in spite of the distortion of phonetic integrity indicated by the reduced intelligibility of the same modifications in stationary maskers.

ACKNOWLEDGEMENTS

This work was partially funded by the "Listening Talker" project, supported by the Future and Emerging Technologies (FET) programme within the Seventh Framework Programme for Research of the European Commission, under FET-Open grant no. 256230. Author VA also acknowledges support from the FP7 FET Project "Speech Unit(e)s", grant no. 339152.

Appendix A: Computation of the GCRetime local distance function

The GCRetime local distance function (Eq. A1) is defined for each pair of time frames i of the speech signal and j of the masker as the product of two terms: (i) glimpse proportion, $GP(i, j)$, the proportion of the speech signal in frame i glimpsed in the presence of the masker in frame j (Eq. A2); and (ii) $W_{CSE}(i)$, a weighting term based on the cochlear-scaled entropy of the speech signal in frame i (Eq. A3):

$$D(i, j) = GP(i, j)W_{CSE}(i) \tag{A1}$$

Glimpse proportion

The glimpse proportion is intended to reflect the local audibility of speech in noise and is defined as the percentage of spectral regions where the modelled auditory excitation pattern

for the target speech exceeds that of the masker:

$$GP(i,j) = \frac{1}{F}\sum_{f=1}^{F}\mathcal{H}[\,S_f(i) > M_f(j)\,] \tag{A2}$$

where F is the number of frequency channels, S_f and M_f denote the excitation patterns of speech and masker in frequency channel f, and $\mathcal{H}(.)$ is the Heaviside unit step function counting the number of channels where the speech exceeds the masker. Excitation patterns are derived via a gammatone filterbank [34] using an implementation introduced by Cooke [11]. The Hilbert envelope of each gammatone filter output is computed and smoothed by a leaky integrator with a 8 ms time constant [29], downsampled and log-compressed. Here the gammatone filterbank contained $F = 32$ frequency channels spaced equally on an ERB-rate scale between 50 Hz and 7500 Hz.

Cochlear-scaled entropy

In the current study we use the concept of cochlear-scaled entropy [CSE; 44] to identify spectral regions which are changing most rapidly in order to give them greater weight in the computation of the local distance function. CSE is implemented as a locally averaged measure of spectral change across time based on excitation patterns of the target speech signal:

$$CSE(i) = \sum_{k=-\lambda/2}^{\lambda/2} d(i+k)$$

where

$$d^2(t) = \sum_{f=1}^{F}[S_f(t+1) - S_f(i)]^2$$

and λ is the number of frames over which the CSE is computed. Following [44], $\lambda = 5$, equivalent to 80 ms for the 16 ms time frames used here.

The CSE-based weighting is defined as

$$W_{\mathrm{CSE}}(i) = (w-1)\,\mathcal{H}[CSE(i) - \beta] + 1 \tag{A3}$$

where β is a threshold used to identify high-CSE regions, and w defines the degree of boosting of the CSE value. Here, values of $\beta = 0.6$ and $w = 3$ were used.

[1] Adams, E. M., Gordon-Hickey, S., Morlas, H., and Moore, R. (2012). Effect of rate-alteration on speech perception in noise in older adults with normal hearing and hearing impairment. *Am. J. Audiol.*, 21(1):22–32.

[2] Adams, E. M. and Moore, R. E. (2009). Effects of speech rate, background noise, and simulated hearing loss on speech rate judgment and speech intelligibility in young listeners. *J. Am. Acad. Audiol.*, 20:28–39.

[3] Aubanel, V. and Cooke, M. (2013). Information-preserving temporal reallocation of speech in the presence of fluctuating maskers. In *Proc. Interspeech*, pages 3592–3596, Lyon, France.

[4] Aubanel, V., Cooke, M., Villegas, J., and García Lecumberri, M. L. (2011). Conversing in the presence of a competing conversation: effects on speech production. In *Proc. of Interspeech*, pages 2833–2836, Florence, Italy.

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

[6] Bernstein, J. G. and Grant, K. W. (2009). Auditory and auditory-visual intelligibility of speech in fluctuating maskers for normal-hearing and hearing-impaired listeners. *J. Acoust. Soc. Am.*, 125:3358–3372.

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

[8] Bronkhorst, A. W., Bosman, A. J., and Smoorenburg, G. F. (1993). A model for context effects in speech recognition. *J. Acoust. Soc. Am.*, 93:499–509.

[9] Brouckxon, H., Verhelst, W., and Schuymer, B. D. (2008). Time and frequency dependent amplification for speech intelligibility enhancement in noisy environments. In *Proc. Interspeech*, volume 9, pages 557–560.

[10] Brungart, D. S., Simpson, B. D., Ericson, M. A., and Scott, K. R. (2001). Informational and energetic masking effects in the perception of multiple simultaneous talkers. *J. Acoust. Soc. Am.*, 100(5):2527–2538.

[11] Cooke, M. (1993). *Modelling Auditory Processing and Organisation.* Cambridge University Press.

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

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

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

[15] Cooke, M., Mayo, C., and Villegas, J. (2014b). The contribution of durational and spectral changes to the Lombard speech intelligibility benefit. *J. Acoust. Soc. Am.*, 135(2):874–883.

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

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

[18] Freyman, R. L., Balakrishnan, U., and Helfer, K. S. (2008). Spatial release from masking with noise-vocoded speech. *J. Acoust. Soc. Am.*, 124:1627–1637.

[19] Fux, T., Feng, G., and Zimpfer, V. (2012). Natural-to-shouted voice transformation for distance cues of monosyllabic consonant-vowel-consonant words. *Acta Acust. United Ac.*, 98(5):839–843.

[20] Gordon-Salant, S. and Fitzgibbons, P. J. (2004). Effects of stimulus and noise rate variability on speech perception by younger and older adults. *J. Acoust. Soc. Am.*, 115(4):1808–1817.

[21] Grieser, D. A. L. and Kuhl, P. K. (1988). Maternal speech to infants in a tonal language: Support for universal prosodic features in motherese. *Dev. Psychol.*, 24(1):14.

[22] Hualde, J. and Chitoran, I. (2003). Explaining the distribution of hiatus in Spanish and Romanian. In *Proc. Int. Conf. Phonetic Sciences*, pages 1683–1686, Barcelona.

[23] Jokinen, E., Remes, U., and Alku, P. (2016). The use of read versus conversational Lombard speech in spectral tilt modeling for intelligibility enhancement in near-end noise conditions. In *Proc. Interspeech*, pages 2771–2775.

[24] Junqua, J.-C. (1993). The Lombard reflex and its role on human listeners and automatic speech recognizers. *J. Acoust. Soc. Am.*, 93(1):510–524.

[25] Koutsogiannaki, M. and Stylianou, Y. (2016). Modulation enhancement of temporal envelopes

for increasing speech intelligibility in noise. In *Interspeech 2016*, pages 2508–2512.

[26] Lu, Y. and Cooke, M. (2008). Speech production modifications produced by competing talkers, babble, and stationary noise. *J. Acoust. Soc. Am.*, 124(5):3261–3275.

[27] Mendoza, E., Carballo, G., Cruz, A., Fresneda, M. D., Muoz, J., and Marrero, V. (2003). Temporal variability in speech segments of Spanish: context and speaker related differences. *Speech Communication*, 40:431–447.

[28] Miller, G. A. and Licklider, J. C. (1950). The intelligibility of interrupted speech. *J. Acoust. Soc. Am.*, 22:167–173.

[29] Moore, B. C. J., Glasberg, B. R., Plack, C. J., and Biswas, A. K. (1988). The shape of the ear's temporal window. *J. Acoust. Soc. Am.*, 83(7-8):1102–1116.

[30] Moreno, A., Poch, D., Bonafonte, A., Lleida, E., Llisterri, J., Marino, J. B., and Nadeu, C. (1993). Albayzín speech database: Design of the phonetic corpus. In *Eurospeech*, pages 175–178, Berlin, Germany.

[31] Naylor, G. (2016). Theoretical issues of validity in the measurement of aided speech reception threshold in noise for comparing nonlinear hearing aid systems. *J. Am. Acad. Audiology*, 27:504–514.

[32] Nejime, Y. and Moore, B. C. J. (1998). Evaluation of the effect of speech-rate slowing on speech intelligibility in noise using a simulation of cochlear hearing loss. *J. Acoust. Soc. Am.*, 103(1):572–576.

[33] Oxenham, A. J. and Simonson, A. M. (2009). Masking release for low-and high-pass-filtered speech in the presence of noise and single-talker interference. *J. Acoust. Soc. Am.*, 125:457–468.

[34] Patterson, R. D., Holdsworth, J., Nimmo-Smith, I., and Rice, P. (1988). SVOS Final Report: The Auditory Filterbank. Technical report 2341. MRC Applied Psychology Unit.

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

[36] Pisoni, D. B., Bernacki, R. H., Nusbaum, H. C., and Yuchtman, M. (1985). Some acoustic-phonetic correlates of speech produced in noise. In *ICASSP*, pages 1581–1584, Tampa, Florida.

[37] Pittman, A. L. and Wiley, T. L. (2001). Recognition of speech produced in noise. *J. Speech Lang. Hear. Res.*, 44(3):487–496.

[38] Rothauser, E. H., Chapman, W. D., Guttman, N., Hecker, M. H. L., Nordby, K. S., Silbiger,

H. R., Urbanek, G. E., Weistock, M., McGee, V. E., Pachl, U. P., and Voiers, W. D. (1969). IEEE Recommended practice for speech quality measurements. *IEEE Trans. Audio Acoust.*, pages 225–246.

[39] Sankowska, J., García Lecumberri, M. L., and Cooke, M. (2011). Interaction of intrinsic vowel and consonant durational correlates with foreigner directed speech. *Poznań Studies in Contemporary Linguistics*, 47:109–119.

[40] Sauert, B. and Vary, P. (2006). Near end listening enhancement: Speech intelligibility improvement in noisy environments. In *Proc. ICASSP*, pages 493–496, Toulouse, France.

[41] Schepker, H., Rennies, J., and Doclo, S. (2013). Improving speech intelligibility in noise by sii-dependent preprocessing using frequency-dependent amplification and dynamic range compression. In *Proc. Interspeech*, pages 3577–3581.

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

[43] Song, J. Y., Demuth, K., and Morgan, J. (2010). Effects of the acoustic properties of infant-directed speech on infant word recognition. *J. Acoust. Soc. Am.*, 128(1):389–400.

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

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

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

[47] Tang, Y. and Cooke, M. (2012). Optimised spectral weightings for noise-dependent speech intelligibility enhancement. In *Proc. Interspeech*, pages 955–958, Portland, USA.

[48] Uchanski, R. M. (2005). Clear speech. In Pisoni, D. B. and Remez, R. E., editors, *The Handbook of Speech Perception*, pages 207–235. Blackwell, Oxford, UK.

[49] Valentini-Botinhao, C., Yamagishi, J., and King, S. (2012). Mel cepstral coefficient modification based on the Glimpse Proportion measure for improving the intelligibility of HMM-generated synthetic speech in noise. In *Proc. Interspeech*, pages 631–634, Portland, USA.

[50] Yoo, S. D., Boston, J. R., El-Jaroudi, A., Li, C.-C., Durrant, J. D., Kovacyk, K., and Shaiman,

S. (2007). Speech signal modification to increase intelligibility in noisy environments. *J. Acoust. Soc. Am.*, 122(2):1138–1149.

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

[52] Zorila, T.-C. and Stylianou, Y. (2015). A fast algorithm for improved intelligibility of speech-in-noise based on frequency and time domain energy reallocation. In *Proc. Interspeech*, pages 60–64.