Subjective and objective evaluation of speech intelligibility enhancement under constant energy and duration constraints

Yan Tang[1], *Martin Cooke*[2]

[1]Language and Speech Laboratory, Universidad del País Vasco, Spain
[2]Ikerbasque (Basque Science Foundation)

y.tang@laslab.org, m.cooke@ikerbasque.org

Abstract

Speakers appear to adopt strategies to improve speech intelligibility for interlocutors in adverse acoustic conditions. Generated speech, whether synthetic, recorded or live, may also benefit from context-sensitive modifications in challenging situations. The current study measured the effect on intelligibility of six spectral and temporal modifications operating under global constraints of constant input-output energy and duration. Reallocation of energy from mid-frequency regions with high local SNR produced the largest intelligibility benefits, while other approaches such as pause insertion or maintenance of a constant segmental SNR actually led to a deterioration in intelligibility. Listener scores correlated only moderately well with recent objective intelligibility estimators, suggesting that further development of intelligibility models is required to improve predictions for modified speech.

Index Terms: speech intelligibility, objective measures, energy reallocation

1. Introduction

Human speakers seem to employ a number of strategies whose goal is to maintain speech intelligibility while talking under adverse acoustic environments. Besides spectral modifications based on well-known Lombard effects (e.g. increasing F0 and output intensity, reducing spectral tilt [1, 2]), there is increasing evidence that speakers make use of temporal changes such as pause insertion, keyword repetition, segment lengthening and avoidance of overlap with background sources (e.g. [3]) when faced with challenging communication conditions. Given the wider application of speech output technologies in a range of environments such as public transport hubs where other sound sources are active, it is of interest to explore whether context-sensitive modifications to generated speech can bring about increases in speech intelligibility.

In [4] we proposed several modification strategies to improve speech intelligibility based on the idea of speech energy reallocation. Some of these techniques (reviewed in section 2) are predicted on the basis of an evaluation conducted with a number of intelligibility models to lead to very significant intelligibility gains for listeners. The primary purpose of the current study was to measure the real gains (or otherwise) provided by speech modifications proposed in [4] and in addition to evaluate the effectiveness of pause insertion, both alone and in combination with another modification strategy.

The use of objective intelligibility models leads to a fast turnaround time during algorithm development. Indeed, one approach to finding useful speech modifications is to optimise directly intelligibility predictions derived from such models. However, while traditional objective measures of intelligibility such as the Articulation Index [5], the Speech Intelligibility Index [6] and the Speech Transmission Index [7] show good correlations with subjective scores in a subset of noisy or reverberant environments, they work less well for commonly-occurring conditions such as fluctuating noise [8]. More recent approaches to intelligibility prediction (e.g. [9, 10, 11]) are based on a fine-scale analysis of the spectro-temporal properties of both speech and masker signals and cope better with nonstationary backgrounds. Even so, very little work has been carried out using modified speech. A further goal of the current study was to compare subjective and objective intelligibility for the range of speech modifications proposed in [4].

2. Speech modifications

Six modification approaches were evaluated in the current study:

`SegSNR`: This modification applies a time-varying gain to each frame in time domain by setting the segmental SNR in each time interval to equal the global SNR. After energy normalisation to meet a constant input-output energy constraint, the effect is to reallocate speech energy across time. A window size of 50 ms with 50% overlap provides a reasonable tradeoff of objective intelligibility and quality [4].

`ChanSNR`: Here, the SNR is each frequency channel is set to the global SNR. This is a form of time-invariant spectral filtering which results in spectral reallocation of energy. Spectral processing employed by a 55-channel gammatone filterbank, where the number of channels was determined from the asymptote of an objective speech quality measure [4].

`LocalSNR`: In this extreme form of SNR equalisation, the energy of each time-frequency region is modified to equal the global SNR. Time-frequency regions were defined based on the window size and filterbank from `SegSNR` and `ChanSNR` respectively.

`SelectBoost`: This modification is motivated by the idea that boosting time-frequency regions where the *a priori* local SNR is already high may be a wasteful use of energy, and that similarly boosting those with very low prior SNR may require too much energy expenditure to be worthwhile. Optimisation of objective intelligibility (based on [9]) demonstrated that `SelectBoost` is most effective when applied to frequency channels in the range 1800-7500 Hz with 20 dB boosting on regions with local SNR $<$ +5 dB. The strategy appears to be less effective for lower frequencies probably due to the rapid energy

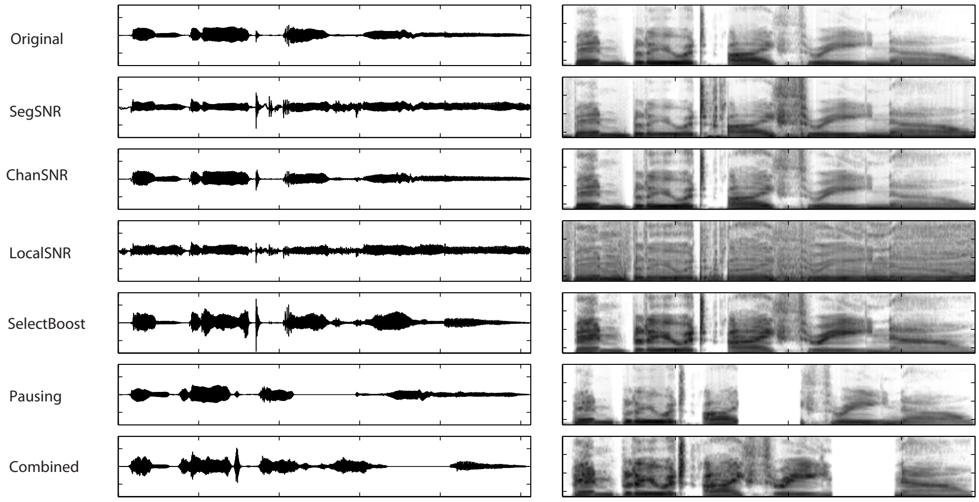

Figure 1: *Waveforms and spectrograms of original (top) modified speech for the utterance "Bin blue at I 3 now". Modifications were produced in response to a speech-shaped noise masker added at a global SNR of -6 dB.*

(and hence local SNR) changes caused by motion of resolved harmonics.

Pausing: Insertion of pauses in speech can ameliorate the impact of regions of high noise level, via an avoidance strategy. The class of approaches based on avoiding a temporal clash between masker and speech is large and includes, for example, insertion of real or filled pauses, elongation of vowel segments, and the use of false starts with repetition of information-bearing elements. Here we examine the simplest approach, that of inserting a single unfilled pause at a critical point (a word boundary) in the speech signal. The determination of which word boundary to insert a pause at is based on optimising objective intelligibility predictions using [9]. The constant energy and duration constraints interact in the pausing strategy to produce two effects which may have opposing influences on intelligibility. First, the remaining speech must be reproduced at a faster rate (here, implemented using PVoc [12]) to meet the duration constraint. At the same time, since no energy is expended in the pause, extra energy per unit time is available when the speech is present. Pilot tests suggested that 300 ms pauses led to a reasonable tradeoff between the opposing effects.

Combined: Objective intelligibility measures from our previous study [4] suggested that SelectBoost is the most effective modification, so it is of interest to explore how it interacts with the Pausing strategy. Here, the strategies were combined in sequence, with boosting applied after pause insertion.

Figure 1 presents an example of waveforms and corresponding spectrograms of original and modified speech.

3. Listening tests

Listeners identified letter and digit keywords from simple sentences such as "Place red at E 4 again" drawn from the Grid Corpus [13]. Sentences were presented in a background of speech-shaped noise (SSN) or speech-modulated noise (SMN) at two SNRs (-6, -9 dB), values chosen on the basis of pilot tests. However, these tests revealed that the LocalSNR modification led to very low scores. Consequently, the SNRs for the LocalSNR condition were increased to 0 and 3 dB.

Speech and noise material was processed by each of the six strategies leading to 28 test conditions (original + 6 modifications x 2 noise levels x 2 noise types), with 50 sentences in each condition. To avoid biases due to possible intrinsic sentence intelligibility differences (i.e. in quiet) across the 28 conditions, 28 different sentence subsets were constructed for *each* condition, producing a total of 784 subsets.

34 native British English speakers from the University of Edinburgh were paid participants in the listening tests. Each listener was assigned to one of 28 sets of sentences such that they heard a different subset in each condition. Stimuli were presented under computer control through Beyerdynamic DT770 PRO headphones in a sound-attenuating booth. Listeners heard each sentence once and responded using letter and digit keys on a computer keyboard. The entire test required about 1.5 hours including a short practice session and hearing screening. Participants completed the test over two sessions. Data from 6 participants was excluded based on high hearing thresholds at two or more frequencies, while outlier analysis led to the removal of a further 4 listeners. Data presented below are mean keywords correct scores from the remaining 24 listeners.

Figure 2 shows listener scores for the 28 conditions grouped by more and less intense noise level (LocalSNR is also included in this plot but not subject to across-modification comparisons since the SNRs used were different than in the other cases). In the stationary masker condition (SSN) at the more adverse noise level, SelectBoost and Combined lead to very large improvements in intelligibility (38.4 and 24.6 percentage points respectively) while SegSNR produced more modest gains. However, both ChanSNR and in particular Pausing have a negative impact relative to the unmodified speech. Similar findings are seen at the less adverse noise level although here there is some evidence that ChanSNR is beneficial rather than harmful. A somewhat different pattern of results is visible for the modulated masker (SMN) at the adverse noise level. While SelectBoost continues to produce reasonable gains, the other approaches have little or negative impact. SegSNR in particular which showed reasonable benefits for stationary noise

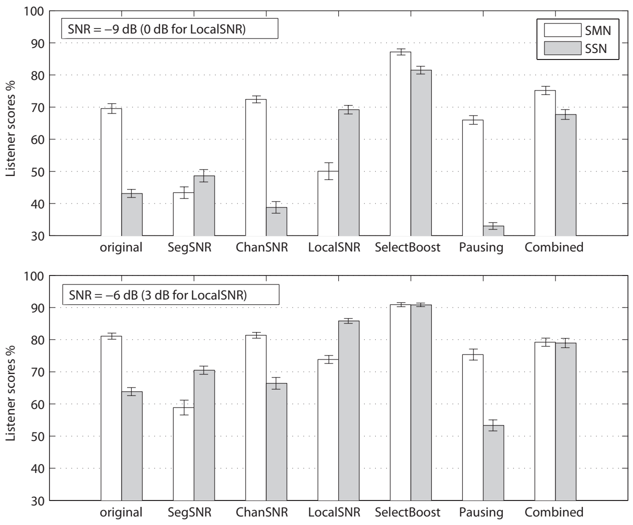

Figure 2: *Subjective scores for original and modified speech in intense (upper) and less intense (lower) noise conditions. Error bars indicate ± 1 standard error.*

suffered a 26.2 percentage points reduction in intelligibility in the face of a modulated masker at the same SNR. The strategy of pause insertion was always harmful, particularly for SSN. Any overall benefits of the `Combined` strategy therefore appear to be due solely to `SelectBoost`.

A three-way repeated measure ANOVA with within-subjects factors of noise type, SNR level and modification strategy supported visual impressions and indicated a significant interaction [F(6, 138) = 12.17, $p < 0.001$] among the 3 factors, as well as significant bi-factor interactions, confirming that the intelligibility pattern of the proposed modifications varies with noise type and SNR. Post-hoc pairwise comparisons with Bonferroni correction for multiple comparisons indicated `SelectBoost` outperformed all other approaches for each of the four noise type/SNR combinations [all $p < 0.001$].

4. Objective measures

Listener scores were compared with quantitative predictions made by three recent objective models of speech intelligibility that have demonstrated high correlations ($\rho = 0.95$ and above) in subjective perceptual tests results, albeit for different test material and noise conditions:

1. The *glimpse proportion* (GP) metric is obtained by calculating the percentage of spectro-temporal regions in modelled auditory excitation patterns whose local SNR exceeds some threshold (here set at +7 dB). The glimpse proportion is one output of the full glimpsing model [9] and is intended to reflect the local audibility of speech in noise.
2. The *Dau model* [14] starts from a detailed treatment of signal processing in the human auditory system, and estimates speech intelligibility by the average (across-frames) cross-correlation of internal representations of processed and reference signals.
3. The *Short-term objective intelligibility* (STOI) metric [11] decomposes signals into time-frequency regions, followed by energy clipping and normalisation. Intelligibility predictions are based on mean cross-correlations between processed and reference signals across time-frequency cells.

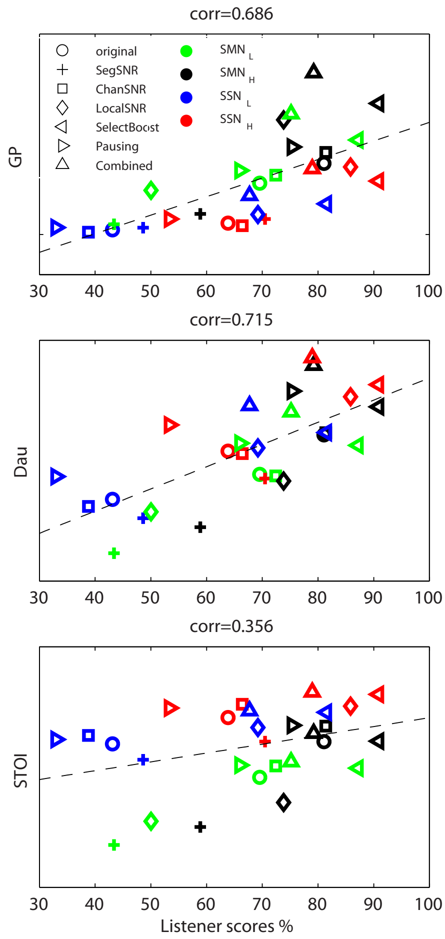

Figure 3: *Comparisons of correlations between subjective identification rates and GP, Dau and STOI.*

Subjective-objective comparisons for speech processed by the proposed modifications are shown in Figure 3 for the GP, Dau and STOI measures. In general, correlation coefficients for modified speech are somewhat lower than those reported in the literature for unprocessed speech in noise, with GP ($\rho = 0.686$) and Dau ($\rho = 0.715$) providing a better match than the STOI measure ($\rho = 0.356$).

All three measures tended to overestimate listener performance in the strategies involving pause insertion, suggesting that the presence of significant amounts of time-shifting/compression is harmful to the predictive power of current models, which may fail to take into account the disruptive effect of pause insertion on speech rhythm for listeners. The models also display different biases for stationary and fluctuating noise: glimpse proportion tends to overpredict listener performance for modulated noise and underpredict the intelligibility of speech in speech-shaped noise, while STOI shows the opposite pattern. The Dau model is more balanced in this respect. Further work is needed to clarify the basis for bias differences between the models.

5. Discussion

Speech modification strategies which reallocate energy in time and frequency can benefit intelligibility in both stationary and fluctuating noises. The `SelectBoost` approach here led to a notable absolute improvement in keyword identification rates of between 9.8 and 38.4 percentage points across noise types and SNRs. `SelectBoost` works by channelling speech energy into those regions with a local SNR of less than 5 dB in the frequency region above 1800 Hz, parameter values chosen on the basis of extensive optimisation of predicted objective intelligibility with genetic algorithms. The overall effect is to increase the number and extent of just-audible regions in the centre of the perceptually-important second formant zone. It is worth noting that `SelectBoost` parameters are likely to depend on noise statistics and may take different values if subjective rather than objective measures were to be used during optimisation.

Not all strategies led to intelligibility gains, and several harmed performance on the task. The three approaches based on shifting energy to equalise segmental, channel and spectro-temporal SNRs (`SegSNR ChanSNR` and `LocalSNR` respectively) showed no benefits and often led to modified speech which was significantly less intelligible for listeners. This was particularly the case for SNR equalisation in the temporal and spectro-temporal domains. One possibility is that the use of SNR equalisation when global SNRs are negative has the effect of spreading energy so that more regions are adversely affected by the noise. This behaviour is most acute in the case of `LocalSNR` which indeed showed extremely low levels of intelligibility at SNRs of -6 dB in pilot tests. Conversely, these findings suggest that an effective strategy to combat high noise levels may be to do the reverse, by focusing energy more sparsely in the signal.

Masking noise type influenced the pattern of results. At any given SNR, stationary noise was a more effective masker than fluctuating noise (as reported by [15]). However, the modulated noise benefit was decreased substantially for `SelectBoost` and reversed in the case of `LocalSNR` and `SegSNR`. The latter two strategies have in common their reassignment of energy in the time domain, which can be considered as introducing new noise-correlated envelope modulations into the speech signal. It is possible that noise modulations interfere with the processing of speech modulations under these conditions [16].

Pause insertion also did not help, even though it might have been expected to be beneficial to intelligibility due both to the avoidance of regions of high noise and by utilising the excess energy created by applying a constant overall duration constraint. It appears that listeners were affected adversely by the need to compress speech in the non-pause regions to meet this constraint. Another possibility is that pauses disrupted listeners expectations of *when* to listen for keywords. Further studies are needed to clarify why listeners suffered from the presence of pauses and to explore whether pausing at more appropriate junctures might be beneficial, as well as the effect of other segmental and prosodic-scale modifications such as elongation, filled pauses and repetitions.

A further finding was the reduction in predictive performance of objective models of speech intelligibility when applied to modified speech, echoing recent results for modified (synthetic) speech [17]. Classical models such as [5, 6, 7] were designed to handle stationary additive noise and latterly reverberation, while the more recent models employed here were motivated by the need to cope with fluctuating noise sources. It is likely that further evolution of objective models will be required to cater for speech which has been processed to enhance intelligibility. Notwithstanding their current limitations, optimising objective functions derived from intelligibility models may be a valuable strategy for offline development of speech modification strategies whose aim is to increase the intelligibility of generated speech in challenging listening conditions.

Acknowledgements This study was supported by the LISTA Project (http://listening-talker.org), funded by the Future and Emerging Technologies programme within the 7th Framework Programme for Research of the European Commission, FET-Open grant number 256230. Listening tests were carried out when the second author was a visitor at the University of Edinburgh. We thank Rob Clark, Vasilios Karaiskos, Cassie Mayo and Cassia Valentini Botinhao at the University of Edinburgh as well as Vincent Aubanel from Ikerbasque for their help in participant recruitment and data collection.

6. References

[1] J. C. Junqua, "The Lombard reflex and its role on human listeners and automatic speech recognizers," *J. Acoust. Soc. Am.*, vol. 93, no. 1, pp. 510–524, 1993.

[2] W. V. Summers, D. B. Pisoni, R. H. Bernacki, R. I. Pedlow, and M. A. Stokes, "Effects of noise on speech production: acoustic and perceptual analyses," *J. Acoust. Soc. Am.*, vol. 84, no. 3, pp. 917–928, 1988.

[3] M. Cooke and Y. Lu, "Spectral and temporal changes to speech produced in the presence of energetic and informational maskers," *J. Acoust. Soc. Am.*, vol. 128, no. 4, pp. 2059–2069, 2010.

[4] Y. Tang and M. Cooke, "Energy reallocation strategies for speech enhancement in known noise conditions," in *Proc. Interspeech*, 2010, pp. 1636–1639.

[5] K. Kryter, "Methods for the calculation and use of the Articulation Index," *J. Acoust. Soc. Am.*, vol. 34, no. 11, pp. 1689–1697, 1962.

[6] "ANSI S3.5-1997 Methods for the calculation of the Speech Intelligibility Index," 1997.

[7] H. J. M. Steeneken and T. Houtgast, "A physical method for measuring speech-transmission quality," *J. Acoust. Soc. Am.*, vol. 67, no. 1, pp. 318–326, 1980.

[8] K. S. Rhebergen and N. J. Versfeld, "Speech Intelligibility Index-based approach to predict the speech reception threshold for sentences in fluctuating noise for normal-hearing listeners," *J. Acoust. Soc. Am.*, vol. 117, no. 4, pp. 2181–2192, 2005.

[9] M. Cooke, "A glimpsing model of speech perception in noise," *J. Acoust. Soc. Am.*, vol. 119, no. 3, pp. 1562–1573, 2006.

[10] C. Christiansen, M. S. Pedersen, and T. Dau, "Prediction of speech intelligibility based on an auditory preprocessing model," *Speech Comm.*, vol. 52, no. 7-8, pp. 678–692, 2010.

[11] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short time objective intelligibility measure for time-frequency weighted noisy speech," in *Proc. ICASSP*, 2010, pp. 4214–4217.

[12] D. Ellis, "A phase vocoder in Matlab," online `http://labrosa.ee.columbia.edu/matlab/pvoc/`, accessed on 19 Nov 2010.

[13] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *J. Acoust. Soc. Am.*, vol. 120, no. 5, pp. 2421–2424, 2006.

[14] T. Dau, D. Puschel, and A. Kohlrausch, "A quantitative model of the effective signal processing in the auditory system. I. Model structure," *J. Acoust. Soc. Am.*, vol. 99, no. 6, pp. 3615–3622, 1996.

[15] J. M. Festen and R. Plomp, "Effects of fluctuating noise and interfering speech on the speech reception threshold for impaired and normal hearing," *J. Acoust. Soc. Am.*, vol. 88, pp. 1725–1736, 1990.

[16] B. C. J. Moore and U. Jorasz, "Detection of changes in modulation depth of a target sound in the presence of other modulated sounds," *J. Acoust. Soc. Am.*, vol. 91, no. 2, pp. 1051–1061, 1992.

[17] C. Valentini-Botinhao, J. Yamagishi, and S. King, "Evaluation of objective measures for intelligibility prediction of HMM-based synthetic speech in noise," in *Proc. ICASSP2011*, 2011.