INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Glimpse-based metrics for predicting speech intelligibility in additive noise conditions

Yan Tang[1], *Martin Cooke*[2, 3]

[1]Acoustics Research Centre, University of Salford, UK
[2]Ikerbasque (Basque Science Foundation), Bilbao, Spain
[3]Language and Speech Laboratory, Universidad del País Vasco, Vitoria, Spain
y.tang@salford.ac.uk, m.cooke@ikerbasque.org

Abstract

The glimpsing model of speech perception in noise operates by recognising those speech-dominant spectro-temporal regions, or glimpses, that survive energetic masking; hence, a speech recognition component is an integral part of the model. The current study evaluates whether a simpler family of metrics based solely on quantifying the amount of supra-threshold target speech available after energetic masking can account for subjective intelligibility. The predictive power of glimpse-based metrics is compared for natural, processed and synthetic speech in the presence of stationary and fluctuating maskers. These metrics are raw glimpse proportion, extended glimpse proportion, and two further refinements: one, FMGP, incorporates a component simulating the effect of forward masking; the other, HEGP, selects speech-dominant spectro-temporal regions with above-average energy on the noisy speech. The metrics are compared alongside a state-of-the-art non-glimpsing metric, using three large datasets of listener scores. Both FMGP and HEGP equal or improve upon the predictive power of the raw and extended metrics, with across-masker correlations ranging from 0.81–0.92; both metrics equal or exceed the state-of-the-art metric in all conditions. These outcomes suggests that easily-computed measures of unmasked, supra-threshold speech can serve as robust proxies for intelligibility across a range of speech styles and additive masking conditions.
Index Terms: speech intelligibility, modified speech, noise, objective intelligibility measures, intelligibility enhancement

1. Introduction

Speech communication often takes place under non-ideal listening conditions in which background noise and imperfect transmission channels act to lower intelligibility. Predicting the effect of adverse conditions on speech reception is an important component in many applications, such as the design of acoustic spaces and communication devices [1], the development of algorithms for speech modification and enhancement [2], and estimating intelligibility for cochlear implant users [3]. Objective intelligibility metrics (OIMs) have been investigated for nearly a century [4], with a recent focus on improving estimates in the presence of temporally-modulated maskers (e.g., [5, 6]) and for both modified [7] and synthetic speech [8].

Many OIMs have been motivated by the notion that intelligibility is closely related to the quantity of audible speech components that survive energetic masking in the auditory periphery. Early studies measuring speech intelligibility [4, 9] suggested that intelligibility is proportional to the amount of audible speech information in a number of frequency bands, leading to the articulation index [10] and the subsequent Speech Intelligibility Index (SII) [11]. More recently, the concept of masked audibility has been extended to the spectro-temporal domain in metrics such as the extended SII [5] which combines short-term SII estimates across time frames to account for the effect of fluctuating maskers.

The glimpsing model of speech perception in noise [6] also takes as its starting point the idea that masked audibility is the primary determinant of intelligibility. However, the glimpsing model was not designed as an OIM: rather than predicting average intelligibility, it provides an end-to-end model of speech processing in noise, coupling an initial stage simulating energetic masking with a subsequent speech recognition component that matches those spectro-temporal regions, or 'glimpses', deemed to have survived masking, to models for speech, using missing data techniques [12].

In principle, the glimpsing model could serve as an OIM by direct measurement of the number of items (e.g., phonemes, words) recognised correctly in the presence of noise, and subsequently correlating this figure with listeners' scores in masked conditions. However, the presence of the ASR component in the glimpsing model complicates its deployment as a practical intelligibility prediction metric: the construction of a recogniser for each task is time-consuming and requires the acquisition of training data; the recogniser may not perform at a level equivalent to listeners; intelligibility measurement using ASR is computationally-complex, precluding its use as an objective measure in closed-loop optimisation frameworks where the metric may be required to be evaluated many times. For these reasons, a number of studies (e.g., [13, 7, 14, 15, 16, 17]) have adopted the output of the initial stage of the glimpsing model – known as the 'glimpse proportion' (GP) – as a proxy for intelligibility. A recent study demonstrated that an extended GP-based metric (GP_{ext}) is capable of making reasonable intelligibility predictions [18], although its performance fell some way short of the best-performing metrics.

The purpose of the current study is to evaluate two alternative extensions to the GP_{ext} metric. One extension, FMGP, described in section 2.3, incorporates a model of forward masking. The other, HEGP (section 2.4), explores the use of a high-energy subset of glimpses. The new metrics are evaluated alongside the raw and extended GP (reviewed in sections 2.1 and 2.2), together with a high-performing reference metric, CPD [19], using three large datasets of subjective intelligibility results from listeners exposed to speech in noise, containing plain, modified and synthetic speech styles (section 3).

http://dx.doi.org/10.21437/Interspeech.2016-14

2. Glimpsing metrics

2.1. Glimpse proportion – GP

The glimpse proportion (GP) [6] is defined as the proportion of time-frequency regions in modelled auditory excitation patterns whose local SNR exceeds a threshold α dB:

$$GP = \frac{1}{TF}\sum_{f=1}^{F}\sum_{t=1}^{T}\mathcal{H}[\,S_f(t) > (N_f(t)+\alpha)\,] \qquad (1)$$

T and F are the number of time frames and frequency channels, $S_f(t)$ and $N_f(t)$ denote spectro-temporal excitation patterns (STEPs) for speech and noise at time t and frequency f, and $\mathcal{H}[\cdot]$ is the unit step function which counts the number of 'glimpses' meeting the local masked audibility criterion α (set to 3 dB, a value which produced a high listener-model correlation in [6]). STEP computation proceeds by passing speech $s(t)$ and masker $n(t)$ waveforms independently through a 34-channel gammatone filterbank [20]. Filters are linearly spaced on the equivalent rectangle bandwidth scale [21] with centre frequencies from 100 to 7500 Hz. The instantaneous Hilbert envelope $e_f(t)$ at the output of each filter f is computed and smoothed by a leaky integrator with an 8 ms time constant [22], followed by downsampling to 100 Hz and log-compression.

2.2. Extended glimpse proportion – GP_{ext}

GP_{ext} [18] augments GP by (1) ensuring potential glimpses are above the threshold of audibility; (2) accommodating speech rate changes; and (3) applying a compressive transformation and converting to an index in the range 0-1.

Threshold of audibility. To prevent the inclusion of inaudible speech-dominant regions, a hearing threshold is incorporated into the glimpsing metric. Speech and masker signals at the output of each gammatone filter are multiplied by a frequency-dependent gain W_f defined by [23]. The glimpsing criterion is then adjusted (Eq. 2) to constrain potential glimpses to exceed both the local SNR α and the hearing level (HL; set to 25 dB).

$$W_f S_f(t) > max(W_f N_f(t)+\alpha, HL) \qquad (2)$$

Speech rate change compensation. Speeded-up or time-compressed speech can lead to intelligibility losses (e.g., [24, 25, 26]). Evidence for intelligibility gains from slower speech is mixed, with some studies finding benefits (e.g., [27, 28]) and others suggesting a lack of effect (e.g. [29, 30]). To model these durational effects, GP is weighted by a measure of speech rate, $1/\lambda$, where λ is the speed-up factor, defined as the ratio of the duration of the unmodified speech compared to that of the modified speech. Note that since this weighting could potentially lead to a GP exceeding 1, the weighted value is capped at unity.

Compressive transformation to OIM index. This stage accounts for the finding that listeners' performance reaches ceiling levels for a GP well below unity [13]. The quasi-log function v defined in Eq. 3 is applied to compress glimpse values. The offset δ is set to a small value to prevent log of zero issues in situations where no glimpses survive; the expression in the denominator restricts the metric to the range [0-1]. Other compressive functions of the form x^γ produce similar results.

$$v(x) = \frac{\log(1+x/\delta)}{\log(1+1/\delta)}, \quad \delta = 0.01 \qquad (3)$$

The GP_{ext} metric is summarised in Eq. 4

$$GP_{ext} = v\Bigg[\min\Bigg(\frac{1}{\lambda}\frac{1}{TF}\sum_{f=1}^{F}\sum_{t=1}^{T}\mathcal{H}[\,W_f S_f(t) > max(W_f N_f(t)+\alpha, HL)],1\Bigg)\Bigg] \qquad (4)$$

2.3. GP_{ext} with forward masking – FMGP

GP and GP_{ext} model simultaneous masking, but do not take into non-simultaneous masking, of which forward masking (FM) is the most important form for listeners [31, 32, 33]. FM reduces the sensitivity of the auditory response following an intense component in a given frequency region. Thus, earlier portions of a given signal can affect the same signal in later epochs, and other signals (including noise) can mask each other. Incorporating a forward masking model into an intelligibility metric can improve its predictive power (as in [5]).

Here, an inner hair cell (IHC) model [34] was used to simulate the FM effect. The neural response to a stationary sound at the level of the auditory nerve shows a distinct onset followed by a decay in activity, and is thought to provide in part the neurophysiological basis for psychophysical forward masking. The simulated IHC output in frequency band f, denoted $IHC_f^{[s+n]}(t)$, is computed from the mixture waveform $[s+n](t)$ following envelope extraction. Subsequently, all IHC peaks are identified, with time-locations denoted $peaks_f$. To identify and remove 'masked glimpses' of the speech target, the following rule is applied: for each peak location in $peaks_f$, if there is a putative glimpse (defined by Eq. 2) at that time, then this is treated as a non-masked glimpse. Thus: (a) if the peak coincides with the onset of a glimpse, the entire glimpse is treated as non-masked; (b) if the peak occurs some way through a glimpse, only the part of the glimpse subsequent to the peak is retained; and (c) if no part of the glimpse contains a peak, the glimpse is regarded as masked. The glimpse definition for FMGP then becomes, by extension of Eq. 2:

$$[W_f S_f(t) > max(W_f N_f(t)+\alpha, HL)] \wedge [\neg FM_f(t)] \qquad (5)$$

where the expression $FM_f(t)$ indicates that a glimpse in channel f at time frame t is masked according to the aforementioned rule. The FMGP metric simply replaces the input to $\mathcal{H}[\cdot]$ in Eq. 4 with the above expression Eq. 5.

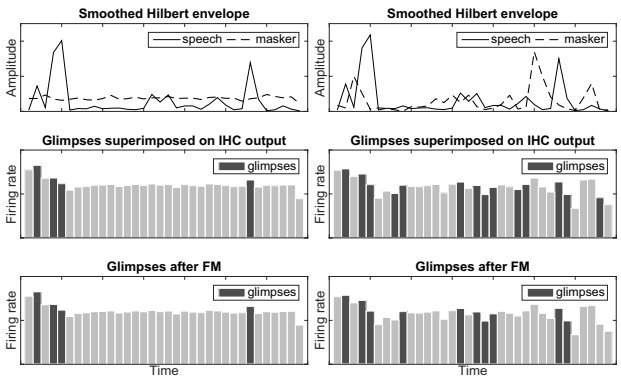

Figure 1: *An illustration of the forward masking effect in a single frequency band in the presence of a stationary (SSN, left) and a fluctuating masker (CS, right). For the purposes of illustration the sampling rate in this figure has been reduced 7-fold.*

Fig. 1 illustrates the forward masking effect. The upper panels show the response to speech and masker in a single frequency band prior to the IHC stage, represented by S_f and N_f respectively. Epochs where the response to speech exceeds that to the masker (i.e., without the FM component) by a certain amount α are shown in the middle panels, superimposed upon the IHC output in response to the mixture of speech and noise, $IHC_f^{[s+n]}$. The lower panel indicates the glimpses that remain after FM, demonstrating a greater reduction for competing speech (CS; right) than in speech-shaped noise (SSN; left). This is due to the increased likelihood of a target speech glimpse being preceded by a masker-related peak in the IHC output, which acts to mask any IHC activity peak related to the target signal. An example can be seen towards the end of the middle panel (right) of Fig. 1, where a potential speech glimpse is masked due to the presence of the preceding masker peak.

2.4. GP_{ext} with high-energy glimpses – HEGP

The study of GP_{ext} in [18] found that even when SNRs are adjusted to equalise intelligibility, more glimpses survive a competing speech (CS) masker than a stationary speech-shaped masker (SSN). This outcome suggests that some of the extra glimpses in CS may not fully contribute to the intelligibility gain, and that identifying those extra glimpses might increase the predictive performance of a glimpsing metric.

Motivated by a desire to isolate the effects of additive noise, peak clipping, and centre clipping, the Coherence Speech Intelligibility Index (CSII) [35] classifies speech frames into three levels according to the RMS energy of the frame. Since vowels are mostly high energy sounds, vowel-consonant transitions are neither high nor low in energy, and many consonants are low in energy, these speech components can be affected by noise to different degrees [35]. CSII uses a linearly-weighted sum of the mean SIIs at the three levels. The CPD metric [19] adopts a similar procedure but uses only the high-energy frames.

Inspired by this idea, the current study investigated the effect of selecting only high energy glimpses. As a starting point, time-frequency bins which constitute glimpses in the speech-plus-noise mixture are categorised according to the relative energy Y' of the STEP of a glimpsed region in the noisy speech $Y_f(t)$ and the mean $\bar{Y}_f$ across all time frames in channel f:

$$Y'_f(t) = W_f(Y_f(t) - \bar{Y}_f) \tag{6}$$

According to the three-level criteria used in CSII, high-energy time-frequency bins are those in which Y' is 0 dB or above, those between -10 and 0 dB fall into the mid-energy level, while low-energy glimpses have Y' between -30 to -10 dB. As illustrated in Fig. 2, glimpses are available at all three energy levels across time and frequency in CS, while almost all

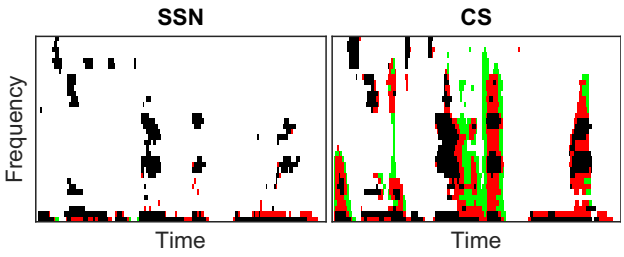

Figure 2: *Glimpses in the presence of CS at -14 dB SNR and SSN at -7 dB SNR, colour-coded by energy level classification (black: high energy; red: mid; green: low).*

are high-energy glimpses in SSN, with a small number of mid-energy glimpses at the low frequencies. Therefore, the approach adopted here is to take into account solely the contribution of the high-energy glimpses to intelligibility, replacing the input to $\mathcal{H}[\cdot]$ in Eq. 4 by Eq. 7:

$$[W_f S_f(t) > (W_f N_f(t) + \alpha)] \wedge [W_f Y_f(t) > \max(W_f \bar{Y}_f, HL)] \tag{7}$$

3. Evaluation

The ability of the four glimpse-based metrics – GP, GP_{ext}, FMGP and HEGP – to predict listeners' scores in sentence-based tasks was evaluated. Alongside the glimpse-based metrics, the performance of the CPD approach [19], the best-performing metric in a recent comparative evaluation [18], is also reported. The current evaluation uses the datasets of the evaluation in [18]. This consists of three sets of listener data which total over 396 experimental conditions, as summarised in Table 1. The three datasets, denoted NATURAL [7], TTS [8] and HURRICANE [36, 37], contain plain (i.e., natural, unmodified) speech, natural speech with algorithmic modifications applied, natural Lombard speech, and synthetic speech, with and without additional modifications. In each case listeners identified keywords in sentences presented in a range of stationary and fluctuating additive noise maskers.

Table 1: *Composition of the evaluation datasets. SMN: speech-modulated noise; BAB: speech babble. The number of each style or speech rate condition is indicated in parentheses.*

	NATURAL	TTS	HURRICANE
Styles	plain (1) modified (5)	synthetic (1) modified synthetic (4)	plain (2) Lombard (1) modified (19) synthetic (8)
Sentences	Matrix	Matrix	Harvard
Maskers	SSN, SMN	SSN, BAB Car, HighFreq	SSN, CS
Conditions	24	192	180
Listeners	24	88	314
Speech rate changes	none	faster (16) slower (32)	slower (72)

Subjective intelligibility was measured as the percentage of keywords identified correctly by listeners in each condition. Predictive performance was evaluated using the Pearson correlation coefficient ρ between subjective intelligibility and the output of the metrics, along with the error of the standard deviation of listener scores, defined as $\sigma_e = \sigma_d\sqrt{1-\rho^2}$, where σ_d is the standard deviation of listener scores per condition.

Tables 2-4 show correlations between measured (i.e. listener) and predicted intelligibility for each dataset. Considering overall across-masker correlations, while raw GP produces similar correlations to CPD for the NATURAL and TTS datasets, those for the HURRICANE dataset are significantly poorer. The extended GP metric GP_{ext} leads to clear improvements for each dataset, and the two OIMs proposed here lead to further increases in correlation, substantially so for the HURRICANE con-

Table 2: *Listener-metric correlations (ρ; σ_e in parentheses) for the* NATURAL *dataset. Metrics with the highest correlations (including those that are statistically-equivalent) are highlighted.*

	Overall	SSN	SMN
GP	0.79 (0.10)	0.79	0.88
GP_{ext}	0.89 (0.07)	0.93	0.94
FMGP	0.90 (0.07)	0.92	0.93
HEGP	0.92 (0.07)	0.92	0.89
CPD	0.79 (0.10)	0.79	0.79

Table 3: *Correlations for the* TTS *dataset.*

	Overall	SSN	BAB	Car	HighFreq
GP	0.71 (0.17)	0.79	0.81	0.78	0.83
GP_{ext}	0.78 (0.15)	0.89	0.88	0.91	0.85
FMGP	0.81 (0.14)	0.91	0.89	0.91	0.86
HEGP	0.83 (0.13)	0.92	0.93	0.92	0.88
CPD	0.73 (0.17)	0.73	0.76	0.65	0.86

Table 4: *Correlations for the* HURRICANE *dataset.*

	Overall	SSN	CS
GP	0.53 (0.23)	0.84	0.83
GP_{ext}	0.66 (0.20)	0.90	0.85
FMGP	0.71 (0.19)	0.90	0.85
HEGP	0.87 (0.13)	0.90	0.86
CPD	0.83 (0.15)	0.86	0.78

ditions. Statistical comparisons using chi-squared tests on Z-transformed scores demonstrate that for the NATURAL and TTS datasets, GP_{ext}, FMGP and HEGP are statistically-equivalent, all with higher correlations than GP and CPD [all $p < .05$]. For the HURRICANE dataset, HEGP and CPD are equivalent [$Z = 1.57$, $p = .12$], higher than the remaining metrics [all $p < .001$]. Overall, the use of high energy glimpses produces the best predictions of any metric tested, for all datasets.

Fig. 3 examines in more detail the subjective-objective relationship for FMGP and HEGP, coded by masker type. The higher correlations resulting from HEGP [$Z = 5.61$, $p < .001$] come from improvements in across- rather than within-masker correlation. This is especially evident for the HURRICANE dataset where the two maskers are almost perfectly-separated by the FMGP metric. Here, within-masker correlations are identical for FMGP and HEGP, but FMGP overestimates the intelligibility of speech in the CS background.

4. Discussion

The current study was motivated by the question of how well metrics based solely on glimpse proportion – the quantity of target speech escaping energetic masking – might serve as a proxy for intelligibility, divorced from the later ASR stage of the full glimpsing model [6]. The outcome suggests that glimpse-based metrics are capable of making robust across-masker predictions of the average intelligibility of a range of speech styles in the presence of stationary and fluctuating noise.

While the refinements inherent in GP_{ext} lead to improvements over GP, metrics incorporating simulations of forward masking (FMGP) or making use of high-energy glimpses only (HEGP) produce further useful gains. HEGP in particular has its greatest impact on across-masker predictions, strikingly for the HURRICANE data where the stationary and competing

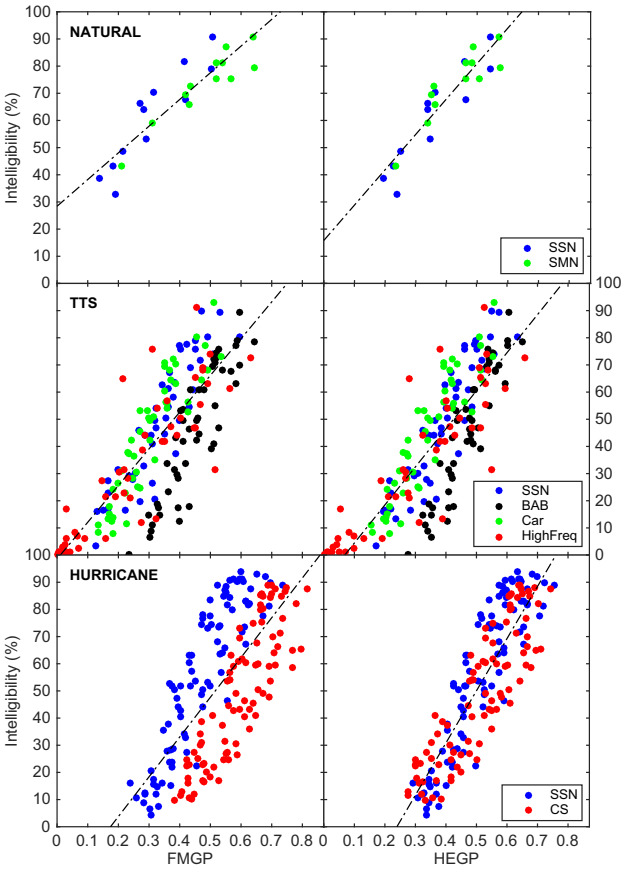

Figure 3: *Subjective intelligibility versus predictions by FMGP (left) and HEGP (right) for the three datasets, coded by masker type. The best linear fit is also shown.*

speech noises are brought into line. While the use of high-energy glimpses in HEGP is inspired by CSII [35] and CPD [19], there are several differences in the way such regions are selected and used: HEGP uses time-frequency pixels rather than time frames; the reference energy level in HEGP is frequency-dependent; in HEGP classification is performed on the noisy speech signal. These differences are strongly related to the glimpsing concept at the core of HEGP.

The incorporation of a forward-masking component into the GP_{ext} metric has a modest impact on predictive power. In practice, relatively few glimpses are removed by the non-simultaneous masking process, since the masking level starts to decay logarithmically after masker offset, lasting up to 200 ms [31, 32, 33].

It is worth noting that OIMs based solely on energetic masking cannot account for more central informational masking effects [38, 39, 40]. Nevertheless, the ability of easily-computed glimpse-based metrics to quantify the effects on intelligibility of both natural and synthetic speech of a range of maskers makes them a potentially useful tool in applications such as the development of speech modification algorithms designed to enhance speech reception in adverse listening conditions.

Acknowledgements This study was supported by the LISTA Project (http://listening-talker.org; FET-Open grant 256230), funded by the Future and Emerging Technologies programme within the 7th Framework Programme for Research of the European Commission.

5. References

[1] G. Ballou, *Handbook for sound engineers*. Taylor & Francis, 2013.

[2] C. Taal, R. C. Hendriks, and R. Heusdens, "Speech energy redistribution for intelligibility improvement in noise based on a perceptual distortion measure," *Computer Speech and Language*, vol. 28, no. 4, pp. 858–872, 2014.

[3] J. F. Santos and T. H. Falk, "Updating the SRMR-CI Metric for Improved Intelligibility Prediction for Cochlear Implant Users," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 2197–2206, 2014.

[4] H. Fletcher, "An empirical theory of telephone quality," *AT&T Internal Memorandum*, vol. 101, no. 6, 1921.

[5] K. S. Rhebergen, N. J. Versfeld, and W. A. Dreschler, "Extended speech intelligibility index for the prediction of the speech reception threshold in fluctuating noise," *J. Acoust. Soc. Am.*, vol. 120, no. 6, pp. 3988–3997, 2006.

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

[7] Y. Tang and M. Cooke, "Subjective and objective evaluation of speech intelligibility enhancement under constant energy and duration constraints," in *Proc. Interspeech*, 2011, pp. 345–348.

[8] C. Valentini-Botinhao, J. Yamagishi, and S. King, "Can objective measures predict the intelligibility of modified HMM-based synthetic speech in noise?" in *Proc. Interspeech*, Florence, Italy, 2011, pp. 1837–1840.

[9] N. R. French and J. C. Steinberg, "Factors governing the intelligibility of speech sounds," *J. Acoust. Soc. Am.*, vol. 19, no. 1, pp. 90–119, 1947.

[10] K. D. Kryter, "Methods for the calculation and use of the Articulation Index," *J. Acoust. Soc. Am.*, vol. 34, pp. 1689–1697, 1962.

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

[12] M. Cooke, P. Green, L. Josifovski, and A. Vizinho, "Robust automatic speech recognition with missing and unreliable acoustic data," *Speech Comm.*, vol. 34, pp. 267–285, 2001.

[13] J. Barker and M. Cooke, "Modelling speaker intelligibility in noise," *Speech Comm.*, vol. 49, pp. 402–417, 2007.

[14] Y. Tang and M. Cooke, "Optimised spectral weightings for noise-dependent speech intelligibility enhancement," in *Proc. Interspeech*, 2012, pp. 955–958.

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

[16] J. Villegas and M. Cooke, "Maximising objective speech intelligibility by local f0 modulation," in *Proc. Interspeech*, 2012, pp. 1704–1707.

[17] V. Aubanel and M. Cooke, "Information-preserving temporal reallocation of speech in the presence of fluctuating maskers," in *Proc. Interspeech*, 2013, pp. 3592–3596.

[18] Y. Tang, M. Cooke, and C. Valentini-Botinhao, "Evaluating the predictions of objective intelligibility metrics for modified and synthetic speech," *Computer Speech and Language*, vol. 35, pp. 73–92, 2016.

[19] 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.

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

[21] B. C. J. Moore and B. R. Glasberg, "Suggested formulas for calculating auditory-filter bandwidths and excitation patterns," *J. Acoust. Soc. Am.*, vol. 74, pp. 750–753, 1983.

[22] B. C. J. Moore, B. R. Glasberg, C. J. Plack, and A. K. Biswas, "The shape of the ear's temporal window," *J. Acoust. Soc. Am.*, vol. 83, no. 7-8, pp. 1102–1116, 1988.

[23] ISO 389-7, "ISO 389-7:2006 Acoustics – Reference Zero For The Calibration Of Audiometric Equipment – Part 7: Reference Threshold Of Hearing Under Free-field And Diffuse-field Listening Conditions," 2006.

[24] G. Fairbanks and F. Kodman, "Word intelligibility as a function of time compression," *J. Acoust. Soc. Am.*, vol. 29, pp. 636–644, 1957.

[25] N. J. Versfeld and W. A. Dreschler, "The relationship between the intelligibility of time-compressed speech and speech in noise in young and elderly listeners," *J. Acoust. Soc. Am.*, vol. 111, no. 1, pp. 401–408, 2002.

[26] C. Valentini-Botinhao, M. Toman, M. Pucher, D. Schabus, and J. Yamagishi, "Intelligibility Analysis of Fast Synthesized Speech," in *Proc. Interspeech*, 2014, pp. 2922–2926.

[27] Z. S. Bond and T. J. Moore, "A note on the acoustic-phonetic characteristics of inadvertently clear speech," *Speech Comm.*, vol. 14, pp. 325–337, 1994.

[28] V. Hazan and D. Markham, "Acoustic-phonetic correlates of talker intelligibility for adults and children," *Journal of the Acoustical Society of America*, vol. 116, pp. 3108–3118, 2004.

[29] A. R. Bradlow, G. M. Torretta, and D. B. Pisoni, "Intelligibility of normal speech I: Global and fine-grained acoustic-phonetic characteristics," *Speech Communication*, vol. 20, pp. 255–272, 1996.

[30] M. Cooke, C. Mayo, and J. Villegas, "The contribution of durational and spectral changes to the Lombard speech intelligibility benefit," *J. Acoust. Soc. Am.*, vol. 135, p. 874, 2014.

[31] R. Plomp, "Rate of decay of auditory sensation," *J. Acoust. Soc. Am.*, vol. 36, pp. 277–282, 1964.

[32] L. L. Elliott, "Masking of tones before, during, and after brief silent periods in noise," *J. Acoust. Soc. Am.*, vol. 45, pp. 1277–1279, 1969.

[33] G. J. Kidd and L. L. Fetch, "Patterns of residual masking," *Hear. Res.*, vol. 5, pp. 49–67, 1981.

[34] M. Cooke, "Modelling Auditory Processing and Organisation," Ph.D. dissertation, University of Sheffield, 1993.

[35] J. Kates and K. Arehart, "Coherence and the speech intelligibility index," *J. Acoust. Soc. Am.*, vol. 117, no. 4, pp. 2224–2237, 2005.

[36] M. Cooke, C. Mayo, C. Valentini-Botinhao, Y. Stylianou, B. Sauert, and Y. Tang, "Evaluating the intelligibility benefit of speech modifications in known noise conditions," *Speech Comm.*, vol. 55, pp. 572–585, 2013.

[37] M. Cooke, C. Mayo, and C. Valentini-Botinhao, "Intelligibility-enhancing speech modifications: the Hurricane Challenge," in *Proc. Interspeech*, 2013, pp. 3552–3556.

[38] R. Carhart, T. W. Tillman, and E. S. Greetis, "Perceptual masking in multiple sound backgrounds," *J. Acoust. Soc. Am.*, vol. 45, pp. 694–703, 1969.

[39] 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, no. 4, pp. 1725–1736, 1990.

[40] D. S. Brungart, "Informational and energetic masking effects in the perception of two simultaneous talkers," *J. Acoust. Soc. Am.*, vol. 109, no. 3, pp. 1101–1109, 2001.