P16: A distortion-weighted glimpse-based intelligibility metric for modified and synthetic speech

Yan Tang[1] and Martin Cooke[2,1] Cassia Valentini-Botinhao[3]

[1] Language and Speech Laboratory, University of the Basque Country, SP
[2] Ikerbasque (Basque Science Foundation), SP
[3] The Centre for Speech Technology Research, University of Edinburgh, UK
y.tang@laslab.org

Techniques which modify clean speech with the aim of increasing speech intelligibility in noise have emerged in recent years (e.g., Tang and Cooke, 2010; Cooke et al., in press). While listening tests provide a final validation of the success of these approaches, objective intelligibility models (OIMs) capable of accurate predictions can be very useful during algorithm development, permitting 'closed-loop' optimisation of modification techniques. However, most existing OIMs were not designed with modified speech in mind, and as a consequence tend to perform poorly with this type of speech material (Tang & Cooke, 2011). Synthetic speech is also a problem for OIMs (Valentini-Botinhao et al., 2011).

In this study we propose a new OIM for speech in noise, motivated by both the energetic masking and distorting effect of noise on speech, and test its predictive power for natural and synthetic speech, both unmodified and modified. The OIM combines elements of two classes of intelligibility model, those based on masked audibility such as the glimpse proportion metric (Cooke, 2006), and those which employ a measure of similarity between internal representations of speech and speech+noise (e.g., Christiansen et al., 2010; Taal et al, 2010). In essence, the proposed OIM is computed as follows: using an auditory-model based time-frequency representation, the proportion of time frames which survive energetic masking in each frequency region is weighted by the cross-correlation of the temporal envelopes of clean and noisy speech in that frequency region; these weighted proportions are then summed across frequency, further weighted to reflect absolute audibility and log-compressed.

The new measure was evaluated by comparison with subjective scores on 3 challenging datasets containing both modified and unmodified speech, composed of natural (Tang & Cooke, 2011), synthetic (Valentini-Botinhao et al., 2011), and natural+synthetic signals (Cooke et al., in press) respectively. Correlation coefficients of 0.73, 0.87 and 0.90 were obtained, which in all cases are higher than those produced by methods based solely on masked audibility or distortion alone. While still some way from the predictive power of OIMs for unmodified speech, the proposed approach represents a worthwhile step towards an effective objective measure for the intelligibility of modified or synthetic speech.

References

Y.Tang and M.Cooke (2010). "Energy reallocation strategies for speech enhancement in known noise conditions," in Proc. Interspeech, pp. 1636-1639.

M. Cooke, C. Mayo, C. Valentini-Botinhao, Y. Stylianou, B. Sauert and Y. Tang (in press). "Evaluating the intelligibility benefit of speech modifications in known noise conditions", Speech Communication.

M. Cooke (2006). "A glimpsing model of speech perception in noise," J. Acoust. Soc. Am., vol. 119, no. 3, pp. 1562-1573.

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