Available online at www.sciencedirect.com

ScienceDirect

Computer Speech and Language 35 (2016) 73–92

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Evaluating the predictions of objective intelligibility metrics for modified and synthetic speech[☆]

Yan Tang[a,b,*], Martin Cooke[c,a], Cassia Valentini-Botinhao[d]

[a] *Language and Speech Laboratory, Universidad del País Vasco, Vitoria, Spain*
[b] *Acoustics Research Centre, University of Salford, UK*
[c] *Ikerbasque (Basque Science Foundation), Bilbao, Spain*
[d] *Centre for Speech Technology Research, University of Edinburgh, Edinburgh, UK*

Received 27 October 2014; received in revised form 3 June 2015; accepted 11 June 2015
Available online 20 June 2015

Abstract

Several modification algorithms that alter natural or synthetic speech with the goal of improving intelligibility in noise have been proposed recently. A key requirement of many modification techniques is the ability to predict intelligibility, both offline during algorithm development, and online, in order to determine the optimal modification for the current noise context. While existing objective intelligibility metrics (OIMs) have good predictive power for unmodified natural speech in stationary and fluctuating noise, little is known about their effectiveness for other forms of speech. The current study evaluated how well seven OIMs predict listener responses in three large datasets of modified and synthetic speech which together represent 396 combinations of speech modification, masker type and signal-to-noise ratio. The chief finding is a clear reduction in predictive power for most OIMs when faced with modified and synthetic speech. Modifications introducing durational changes are particularly harmful to intelligibility predictors. OIMs that measure masked audibility tend to over-estimate intelligibility in the presence of fluctuating maskers relative to stationary maskers, while OIMs that estimate the distortion caused by the masker to a clean speech prototype exhibit the reverse pattern.

Keywords: Objective intelligibility metric; Noise; Speech modifications; Synthetic speech

1. Introduction

Spoken language applications using recorded natural[1] or synthetic speech can be made more robust through algorithmic speech modification. Unlike traditional speech enhancement techniques (e.g., Hu and Loizou, 2004; Martin, 2005; Chen et al., 2006; Srinivasan et al., 2007) which focus on the noise-corrupted speech signal, the speech modification approach (e.g., Sauert and Vary, 2006; Bonardo and Zovato, 2007; Yoo et al., 2007; Brouckxon et al., 2008; Tang and Cooke, 2010) alters the clean speech signal prior to output or transmission. A recent evaluation (Cooke et al., 2013b)

[☆] This paper has been recommended for acceptance by Roger K. Moore.

[*] Corresponding author at: Language and Speech Laboratory, Universidad del País Vasco, Vitoria, Spain.
E-mail address: Y.Tang@laslab.org (Y. Tang).

[1] We use the term ‘natural’ to signify speech produced by a human talker as opposed to speech which is natural-sounding.

http://dx.doi.org/10.1016/j.csl.2015.06.002

demonstrated that speech modification can result in intelligibility gains in noise equivalent to increases of more than 5 dB in output level.

A key ingredient in the design of effective modification strategies is the estimation of listener performance at frequent intervals during the development cycle. However, while subjective intelligibility scores remain the ultimate reference, continuous behavioural testing during algorithm design is usually infeasible. An alternative is to use objective intelligibility metrics (OIMs) to predict listener scores. OIMs not only avoid the need for extensive subjective testing, but can also be used at the core of the algorithm optimisation process. A number of speech modification algorithms (e.g., Sauert and Vary, 2010a; Tang and Cooke, 2011; Taal et al., 2013; Valentini-Botinhao et al., 2014) have been developed and optimised based on maximising intelligibility predictions made by OIMs such as the Speech Intelligibility Index (SII; ANSI, 1997) or the glimpse proportion metric (GP; Cooke, 2006).

OIMs have been motivated by two distinct approaches to account for the effect of noise on speech. In addition to the aforementioned SII and GP metrics, the Articulation Index (AI; French and Steinberg, 1947; Fletcher and Galt, 1950; Kryter, 1962a,b), and the extended Speech Intelligibility Index (ESII; Rhebergen and Versfeld, 2005) focus on quantifying the *masked audibility* of speech in the presence of noise. On the other hand, techniques such as the Normalised-Covariance Measure (NCM; Holube and Kollmeier, 1996; Ma et al., 2009), the Christiansen–Pedersen–Dau metric (henceforth referred to as CPD for brevity; Christiansen et al., 2010) and the Short-Time Objective Intelligibility metric (STOI; Taal et al., 2010) correlate representations of the clean reference speech and the speech-plus-noise signal in an attempt to measure the *distortion* caused by the masker. Another distortion-based approach is the Coherence Speech Intelligibility Index (CSII) proposed by Kates and Arehart (2005). The CSII measures the similarity between clean and noisy speech using magnitude-square coherence (Carter et al., 1973; Kates, 1992) which quantifies the degree to which the output of a system is linearly related to its input.

Both audibility- and distortion-based approaches target spectro-temporal regions least affected by the noise, but differ in their assumptions. While techniques based on audibility require separated estimates of speech and noise in order to estimate masking, distortion-based OIMs assume that human listeners possess a template of the clean speech which is compared to the incoming noisy speech.

When an OIM is employed as the objective function to be maximised, the predictive accuracy of the OIM is critical in determining the validity and effectiveness of the optimisation process. Most of the OIMs mentioned above have been evaluated with recorded natural speech or speech processed by noise reduction techniques. Relatively few studies have investigated their predictive power for modified natural speech or synthetic speech in noise: most OIMs were originally proposed to predict the intelligibility of distorted natural speech, for distortions caused by additive noise together with artefacts introduced by suppression algorithms applied to the noisy speech signal.

Predicting the intelligibility impact of modification algorithms is likely to be challenging since the most successful methods (in terms of improving masked intelligibility) modify the signal in diverse domains – durational and spectral/formant – and possibly through non-linear operations. While the alterations benefit intelligibility, they may also introduce artefacts to the speech signal, leading to degraded speech quality. Nevertheless, the relation between speech intelligibility and quality is complex, and factors such as listening effort and loudness interact. Intelligibility and quality are not simply negatively or positively correlated, especially across listeners (Preminger and Tasell, 1995). For synthetic speech it might be expected that the OIMs' task is even more challenging because the natural speech reference signal is not available, i.e., distortions introduced by the text-to-speech (TTS) system cannot be taken into account. Consequently, predicting the intelligibility of poor quality synthetic speech may be even more difficult.

In two initial studies, which concerned solely the ability of OIMs to predict the masked intelligibility of modified and synthetic speech regardless of the perceptual speech quality, we observed a large reduction in the predictive accuracy of several OIMs on modified and synthetic speech relative to unmodified speech (Tang and Cooke, 2011; Valentini-Botinhao et al., 2011). The current study extends these pilots to a larger range of objective intelligibility metrics and includes behavioural data from recent extensive evaluations of 30 forms of modified and synthetic speech (Cooke et al., 2013a,b). Specifically, we evaluate the performance of one standard (SII) and six recent objective intelligibility metrics (ESII, GP, NCM, CSII, CPD, STOI) in predicting subjective intelligibility scores for both modified and synthetic speech in additive noise. The evaluation makes use of three datasets which together contain 396 combinations of speech modification, masker type and signal-to-noise ratio (SNR). The seven metrics are introduced in Section 2 while Section 3 describes the evaluation datasets. The outcome of a comparison of model predictions against behavioural data from large-scale listening tests is presented in Section 4.