Available online at www.sciencedirect.com

ScienceDirect

ELSEVIER

Computer Speech & Language xxx (2017) xxx-xxx

www.elsevier.com/locate/csl

Learning static spectral weightings for speech intelligibility enhancement in noise☆

Q1 Yan Tang[a,]*, Martin Cooke[b,c]

[a] Acoustics Research Centre, University of Salford, UK
[b] Ikerbasque (Basque Science Foundation), Bilbao, Spain
Q2 [c] Language and Speech Laboratory, Universidad del País Vasco, Vitoria, Spain

Received 30 May 2017; received in revised form 3 October 2017; accepted 16 October 2017
Available online xxx

Abstract

Near-end speech enhancement works by modifying speech prior to presentation in a noisy environment, typically operating under a constraint of limited or no increase in speech level. One issue is the extent to which near-end enhancement techniques require detailed estimates of the masking environment to function effectively. The current study investigated speech modification strategies based on reallocating energy statically across the spectrum using masker-specific spectral weightings. Weighting patterns were learned offline by maximising a glimpse-based objective intelligibility metric. Keyword scores in sentences in the presence of stationary and fluctuating maskers increased, in some cases by very substantial amounts, following the application of masker- and SNR-specific spectral weighting. A second experiment using generic masker-independent spectral weightings that boosted all frequencies above 1 kHz also led to significant gains in most conditions. These findings indicate that energy-neutral spectral weighting is a highly-effective near-end speech enhancement approach that places minimal demands on detailed masker estimation.

Keywords: Speech; Intelligibility; Glimpsing; Noise; Pattern search; Spectral weighting

1. Introduction

Listening to speech in noisy or reverberant environments is both error-prone and effortful. Consequently, reducing the impact of noise via speech enhancement has been the goal of a significant research effort (e.g. Hu and Loizou, 2004; Paliwal and Alsteris, 2005; Martin, 2005; Chen et al., 2006; Srinivasan et al., 2007; Kim et al., 2009; Williamson et al., 2015). Techniques such as noise cancellation or suppression are widely used in human-machine interfaces, and in technologies such as mobile communication and noise-cancelling headphones. However, these approaches have limited use in applications such as public-address systems where listeners are not directly adjacent to the end-point of the transmission channel since, even when the speech signal is further enhanced at the listener's end, the ensuing signal may suffer further contamination in a noisy listening environment.

☆ This paper has been recommended for acceptance by Prof. R. K. Moore.

* Corresponding author.

E-mail address: y.tang@salford.ac.uk (Y. Tang).

http://dx.doi.org/10.1016/j.csl.2017.10.003
0885-2308/ 2017 Elsevier Ltd. All rights reserved.

An alternative approach is to manipulate the speech signal itself, analogous to the way human talkers adjust their speaking style in noisy conditions (e.g. Lombard, 1911; Summers et al., 1988; Junqua et al., 1998; Boril and Pollak, 2005; Cooke and Lu, 2010). Many approaches have been proposed in the last decade to increase speech intelligibility under adverse conditions by altering the clean speech signal. The concept of *near-end* listening enhancement, introduced by Sauert and Vary (2006), describes situations where the speech signal originating at the end of the transmission channel distant from the listener is modified to increase speech intelligibility for the near-end listener who is assumed to be located in a noisy environment. Techniques are generally based on raising the speech spectrum above the average noise spectrum using spectro-temporal manipulation of local signal-to-noise ratio (SNR). Bonardo and Zovato (2007) introduced a dynamic range controller to increase perceived loudness of synthetic speech while maintaining the original intensity range. Time-frequency dependent amplification was employed by Brouckxon et al. (2008) in formant-enhancement, leading to a decreased speech reception threshold in noise.

The aforementioned studies show that increasing SNR via amplification provides a clear benefit for listeners. However, the use of excessive output levels may lead to listener discomfort and stress, and sustained exposure can cause damage to hearing (Knobel and Sanche, 2006) or equipment (Sabin and Schoenike, 1998). Methods proposed in more recent studies (e.g. Yoo et al., 2007; Sauert and Vary, 2009; Tang and Cooke, 2010, 2012; Taal et al., 2014; Schepker et al., 2015) operate under a constant input−output regime for the speech signal, precluding any intelligibility gains due simply to an increase in overall SNR. Even under these constraints speech modification can be highly-effective. Extensive across-algorithm comparisons involving 26 speech modification techniques and using the same dataset for evaluation (Cooke et al., 2013a, 2013b) have shown that state-of-the-art approaches are able to boost intelligibility by an amount equivalent to increasing the gain of unmodified speech by more than 5 dB.

Objective intelligibility or quality metrics (OIMs) have been used in the design of near-end speech modification techniques based on optimising model parameters by maximising the objective metric. Sauert and Vary (2009) optimised the Speech Intelligibility Index (ANSI S3.5−1997, 1997), while the algorithm proposed by Taal et al. (2014) transferred energy to consonant-vowel transients by optimising a perceptual distortion measure developed by Taal and Heusdens (2009), leading to significant listener gains. In our previous work (Tang and Cooke, 2012), the glimpse proportion metric (Cooke, 2006) was used as the OIM in closed-loop optimisation process to derive a series of masker- and level-dependent spectral weightings. Akin to band-importance functions (Studebaker et al., 1987; Studebaker and Sherbecoe, 1991; Bell et al., 1992) which quantify the contribution of each frequency region to overall intelligibility, spectral weightings inject more energy in certain frequency bands at the expense of others, although unlike band-importance functions the weightings depend on the masker. Speech with optimised spectral weights was more intelligible than unmodified speech for both stationary and fluctuating maskers (Cooke et al., 2013b).

The current study extends Tang and Cooke (2012) in three directions. First, the optimisation process makes use of a new glimpse-based OIM recently shown to outperform the original glimpse proportion measure. The success of an optimisation strategy is limited by the accuracy of the chosen OIM. In a recent comparison (Tang and Cooke, 2016) of glimpse-based optimisation approaches alongside a state-of-the-art OIM (Christiansen et al., 2010), a metric based on high-energy glimpses led to the most accurate predictions of listener intelligibility scores across nearly 400 conditions varying in speech style, masker type and SNR. The high-energy glimpsing metric, described in Section 2, forms the basis for the optimisation approach of the current study.

Second, the effect on intelligibility of both masker-dependent and masker-independent spectral weightings is evaluated, by questioning the assumption behind many of the aforementioned modification approaches (e.g. Sauert and Vary, 2009; Tang and Cooke, 2010; Taal et al., 2014) that the background noise signal is known or capable of being accurately-estimated. In practice, noise estimation can be problematic, particularly at short time delays. Consequently, algorithms have been proposed that operate independently of knowledge of the masker. Such algorithms typically boost those speech regions or properties believed to convey salient speech information. For example, Zorila et al. (2012) demonstrated that subjective intelligibility can benefit from enhancing formant information and emphasising voicing segments while preserving high frequency components, with a further intelligibility boost produced by dynamic range compression. In another study, Jokinen et al. (2016) showed that modifying the phase spectrum of wide-band telephony speech by enhancing high-amplitude peaks caused by the glottal excitation in the time domain can also increase speech intelligibility in noise. Consequently, one of the objectives of the current study was to determine the effectiveness of spectral weightings learnt offline (Expt. 1) or based on a generic masker-independent boosting pattern (Expt. 2).

Please cite this article as: Y. Tang, M. Cooke, Learning static spectral weightings for speech intelligibility enhancement in noise, Computer Speech & Language (2017), http://dx.doi.org/10.1016/j.csl.2017.10.003