Speech Communication 90 (2017) 15–25

Contents lists available at ScienceDirect

Speech Communication

journal homepage: www.elsevier.com/locate/specom

Modeling the effects of single-microphone noise-suppression

James M. Kates

Department of Speech Language and Hearing Sciences, University of Colorado, Boulder, CO 80309, USA

ARTICLE INFO

Article history:
Received 18 October 2016
Revised 12 April 2017
Accepted 19 April 2017
Available online 20 April 2017

Keywords:
Speech intelligibility
Speech quality
Intelligibility indices
Speech quality indices
Noise suppression
Spectral subtraction
Ideal binary mask
Hearing aids

ABSTRACT

Many different algorithms have been proposed for single-microphone noise suppression. Comparing algorithms can be difficult, however, since different studies use different speech stimuli, different types of noise, and different signal-to-noise ratios (SNRs). This paper proposes the HASPI intelligibility and HASQI quality metrics as effective tools to measure and compare the performance of noise-suppression algorithms. The stimuli are sentences presented in multi-taker babble at SNRs ranging from −20 to +40 dB. Six different noise-suppression algorithms are compared: Wiener, power, and magnitude spectral subtraction, a smoothed gain-vs-SNR approach that limits the maximum attenuation, binary mask processing, and auditory masked transform noise suppression. Exact restoration of the speech envelope and no processing provide reference conditions. Also investigated are the impact of differing amounts of knowledge about the noise signal and the effects of impaired versus normal hearing. In general, the algorithms that worked best for normal hearing were also most effective for a simulated moderate hearing loss. The results indicate that many algorithms provide high predicted intelligibility under ideal conditions where the speech and noise power are known for every segment within each frequency band, but that the performance degrades to that measured for no processing when the local noise power is replaced by the power averaged over all of the segments within the band. The upper bound to algorithm performance is shown to be exact restoration of the noisy speech envelope to match that of the noise-free signal.

1. Introduction

A large number of single-microphone algorithms have been proposed for suppressing the noise in noisy speech. This type of signal processing has a wide range of applications, from voice communication systems to hearing aids. The general procedure is to divide the signal into frequency bands, and to then modify the signal intensity in each band so that segments having a high signal-to-noise ratio (SNR) are passed unaltered while segments having a poor SNR are attenuated. The weighted segments are then summed across frequency to produce the output signal. The purpose of this paper is to systematically study the effects of noise suppression on speech intelligibility and quality; the study compares several noise suppression algorithms, different amounts of knowledge about the noise signal, and the impact of impaired vs. normal hearing. The algorithm evaluation is based on perceptual metrics for speech intelligibility and quality, which provides a unified framework for comparing the different procedures.

The benefits of noise suppression are often limited. While algorithms are typically defined in terms of the rule used to adjust the gain as a function of SNR (e.g. Wiener filter, magnitude spectral subtraction, etc.), the algorithm performance also depends on accurate estimation of the noise properties. Ideal systems, where the speech and noise powers are known exactly for each segment in each frequency band, can yield large improvements in speech intelligibility. Examples include the ideal binary mask (IBM) (Wang et al., 2008) and the ideal Wiener filter (Madhu et al., 2013). On the other hand, systems that estimate the noise power directly from the noisy speech signal show greatly reduced benefits (Hu and Loizou, 2007b; Brons et al., 2012; Hilkhuysen et al., 2012). For example, Hu and Loizou (2007b) compared several noise-suppression algorithms where a voice activity detector was used to update the noise estimate during silences in the speech. They evaluated several different types of noise at different SNRs for normal hearing (NH) listeners, and found no significant improvement in intelligibility except for the Wiener filter applied to automobile noise at an SNR of 5 dB. Hilkhuysen et al. (2012) also compared several noise-suppression algorithms using the Martin (2001) minimum statistics procedure to estimate the noise power, and in general found a small but significant reduction in intelligibility for automobile noise and multi-talker babble. A similar lack of intelligibility benefit has been observed for noise reduction systems implemented in hearing aids and tested with hearing-

E-mail address: James.Kates@colorado.edu

http://dx.doi.org/10.1016/j.specom.2017.04.004
0167-6393/

impaired (HI) listeners (Boymans and Dreschler, 2000; Alcántara et al., 2003; Bentler et al., 2008; Desjardins and Doherty, 2014).

The impact of the noise-estimation procedure is investigated in this paper. To address this issue, two types of noise estimation are considered. The first is ideal processing, which gives error-free estimates of the local noise power thus isolates the effects of the rule that assigns gain as a function of SNR. The second type replaces the segment-by-segment noise power with the noise power averaged over the duration of the stimulus. This approach produces the processing effects that would be achieved with perfect estimation of the noise power given the common assumption that the noise is stationary, with statistics that remain constant from one segment to the next. One can thus begin to address the question of which is more important: the gain rule or the noise-estimation procedure.

A second question considered in this paper is whether there might be general criteria against which noise-suppression algorithm performance can be judged. One consideration is envelope fidelity. Speech with very high intelligibility can be achieved using a noise vocoder, in which noise bands are modulated to reproduce the envelope fluctuations of the original speech signal (Shannon et al., 1995; Souza and Rosen, 2009). High intelligibility is also achieved by the ideal Wiener filter (Madhu et al., 2013), even at highly negative SNRs where the processing output essentially comprises modulated noise. In addition, several metrics have been developed that rely exclusively on envelope measurements in predicting speech intelligibility (Boldt and Ellis, 2009; Taal et al., 2011; Taghia and Martin, 2014; Falk et al., 2015). The success of these processing strategies and of the metrics reinforces the idea that intelligibility is strongly related to envelope fidelity (Kates and Arehart, 2015). To assess this concept further, perfect envelope reconstruction is used in this paper as one of the processing conditions, and noise suppression performance is compared against the results achieved for a perfect match of the processed noisy speech envelope to that of the noise-free signal.

However, even with perfect envelope reconstruction, the time-varying gain will introduce amplitude-modulation distortion, and at poor SNRs the noise will introduce modifications to the signal temporal fine structure (TFS). An additional question is to therefore determine the impact of these envelope modulation and TFS changes on speech intelligibility and quality. Several studies have compared the accuracy of different perceptually-motivated metrics for predicting listener responses to noise suppression processing (Ma et al., 2009; Taal et al., 2011; Gómez et al., 2012; Kates and Arehart, 2014a,b; Falk et al., 2015), although the test conditions have been restricted in some studies to one talker gender or to a small range of SNRs such as 0 and 5 dB. In many instances good accuracy has been obtained for metrics based exclusively on envelope fidelity (Taal et al., 2011), but good accuracy has also been observed for metrics that are based on signal coherence (Ma et al., 2009; Gómez et al., 2012). Combining measures of envelope fidelity and/or coherence can give improved accuracy (Gómez et al., 2012; Kates and Arehart, 2014a,b), leading to metrics that are sensitive to the signal modifications caused by both envelope and TFS changes in the speech, and the analysis in this paper uses a pair of metrics that measure both envelope and TFS changes.

The Hearing Aid Speech Perception Index (HASPI) (Kates and Arehart, 2014a) to predict intelligibility and the Hearing Aid Speech Quality Index (HASQI) (Kates and Arehart, 2014b) are used to predict overall quality. Both metrics have been shown to be accurate for speech under a variety of noise and nonlinear processing conditions for both NH and HI listeners. HASPI produced a Pearson correlation coefficient of 0.978 when compared to listener intelligibility scores for sentences corrupted by additive noise and nonlinear processing, and a correlation coefficient of 0.978 for noisy speech processed using ideal binary mask noise suppression (Kates and Arehart, 2014a). HASQI had a correlation coefficient of 0.974 when compared to listener quality ratings for sentences corrupted by additive noise and nonlinear processing, and a correlation coefficient of 0.937 for noisy speech processed using ideal binary mask noise suppression (Kates and Arehart, 2014b). HASPI combines measurements of changes in the envelope with changes in the TFS to predict intelligibility, while HASQI combines measurements of the envelope changes, TFS changes, and changes in the signal long-term spectrum. Both indices incorporate a model of the auditory periphery (Kates, 2013) that reproduces the behavior associated with impaired as well as normal hearing, and the noise suppression approaches are evaluated for both normal hearing and a moderate hearing loss.

Six algorithms, representative of different noise assumptions and signal-processing approaches, are compared in this paper. The algorithms are Wiener filtering, power spectral subtraction, magnitude spectral subtraction, a generalized version of the Eger et al. (1984) approach, an auditory masked transform (AMT) algorithm derived by Tsoukalas et al. (1997), and a binary mask algorithm (Wang et al., 2008). The stimuli are sentences in multi-talker babble as the SNR is varied. Babble was chosen as the interference to facilitate comparison with previous work (Tsoukalas et al., 1997; Hu and Loizou, 2007a; Hu and Loizou, 2007b; Ma et al., 2009; Brons et al., 2012; Gómez et al., 2012; Hilkhuysen et al., 2012; Madhu et al., 2013; Kates and Arehart, 2014a, 2014b; Falk et al., 2015). However, instead of performing listening tests directly with normal-hearing and hearing-impaired listeners, intelligibility and speech quality indices are used to predict the listener responses. The advantage of using perceptual metrics is that the different signal-processing strategies can easily be compared under identical conditions. The interactions between processing, SNR, and hearing loss will be illustrated, along with the trade-offs between intelligibility and quality.

The remainder of this paper starts with a summary of the six noise-suppression algorithms considered in the analysis, followed by a description of the intelligibility and quality metrics used to evaluate the signal processing. The results comparing the predicted speech intelligibility and quality are then presented as a function of algorithm, noise estimation, and SNR for speech in multi-talker babble. Both normal hearing and a moderate hearing loss are considered.

2. Noise-suppression algorithms

A set of representative noise-suppression algorithms was selected for evaluation, and this paper is not intended to present an exhaustive processing comparison. The noise-suppression algorithms were all implemented using the structure shown in Fig. 1. The signals were processed through an 18-band linear-phase FIR filterbank with band center frequencies ranging from 250 to 7500 Hz. The limiting condition of the envelope restoration processing at very poor SNRs is essentially a noise vocoder, and 18 bands is more than sufficient to give nearly perfect speech intelligibility for this condition (Başkent, 2006). The sampling rate was 22,050 Hz, and each filter was 440 samples (20 ms) long. The outputs in each frequency band were segmented using a 16-ms raised cosine (von Hann) window with fifty-percent overlap. In ideal processing the speech and noise signals were processed separately through the filterbank so that the speech and noise power in each segment was known exactly. The exact noise power for each segment for each frequency band was then used to adjust the gain according to the specified noise-suppression rule. In average noise power processing, as opposed to ideal processing, the noise power in each segment was replaced by the noise power averaged over all of the segments within the frequency band.