Speech Communication 83 (2016) 64–80

Contents lists available at ScienceDirect

Speech Communication

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

Phase modification for increasing the intelligibility of telephone speech in near-end noise conditions – evaluation of two methods

Emma Jokinen[a,*], Hannu Pulakka[b], Paavo Alku[a]

[a] Department of Signal Processing and Acoustics, Aalto University, PO Box 13000, FI-00076 Aalto, Finland
[b] Nokia Technologies, Visiokatu 3, FI-33720 Tampere, Finland

ARTICLE INFO

Article history:
Available online 8 August 2016

Keywords:
Phase modification
Listening effort
Loudness
Intelligibility enhancement
Telephone speech

ABSTRACT

In this study, two intelligibility-increasing post-processing methods based on the modification of the phase spectrum of speech are proposed for near-end noise conditions. One of the algorithms aims to reduce the dynamic range of the signal and take advantage of the energy gain resulting from amplitude normalization to increase the loudness, while the other algorithm is designed to sharpen the high-amplitude peaks in the time-domain signal generated by the periodic glottal excitation to make the speech sound more clear. Both methods are based on first modifying only the phase spectrum, after which the time-domain signal is computed using the inverse Fourier transform. Finally, the time-domain signal is amplitude normalized by scaling its sample values so that they occupy the original amplitude range of the processed frame. The performance of the proposed methods was evaluated by first comparing them to unprocessed speech using objective quality measures as well as subjective loudness and listening preference tests. Based on the results of these evaluations, the phase-modification methods were further compared to unprocessed speech and dynamic range compression using subjective word-error rate and quality tests. Both narrowband and wideband speech from several talkers were included in both evaluations. Both of the methods were able to increase loudness in some bandwidth conditions as well as outperform unprocessed speech and dynamic range compression in terms of intelligibility in high-noise levels. Both of the methods were rated lower in quality than unprocessed speech in clean conditions. In background noise, however, where intelligibility enhancement algorithms are mostly used, both methods achieved similar results to unprocessed speech in terms of listening preference in some of the bandwidth conditions tested.

1. Introduction

Post-processing methods are used in mobile phones to improve the quality and intelligibility of the received speech signal, which can be affected by several different sources of degradation. For instance, the speech signal can be corrupted by environmental noise both at the transmitting and the receiving side of the communication channel, which are referred to as far-end and near-end noise scenarios, respectively. In addition, speech coding, bandwidth limitation, and transmission errors can reduce both quality and intelligibility. Post-processing has been traditionally motivated mainly by its ability to provide quality improvements (Chen and Gersho, 1995). Speech intelligibility, however, becomes the most important factor when communication takes place under severe noise conditions. Even under moderate noise conditions, where the listener can fully understand the received speech signal, the amount of cognitive processing and attention required for understanding the spoken message against the background noise, that is to say, the listening effort, can be great and make parallel tasks more difficult (Van Engen and Peelle, 2014). In the near-end noise scenario, which has attracted increasing interest in recent years (e.g. Cooke et al., 2013; 2014), and which is also the focus of this study, it is assumed that the decoded speech is free of noise, except for distortion resulting from coding and bandwidth limitations. This assumption differs from that used in speech enhancement, where the goal is to remove noise or reverberation from a speech signal that has already been corrupted. With *speech intelligibility enhancement*, the post-processing done at the receiving device aims to enhance the acoustical voice cues, thus increasing the speech signal's loudness and intelligibility over the environmental noise in the listener's surroundings. Typically, the energy of the signal is

* Corresponding author. Fax: +358 9 460 224.
E-mail addresses: emma.jokinen@aalto.fi (E. Jokinen), hannu.pulakka@nokia.com (H. Pulakka), paavo.alku@aalto.fi (P. Alku).

http://dx.doi.org/10.1016/j.specom.2016.08.001
0167-6393/

constrained to be the same before and after the intelligibility enhancement has been applied.

For the near-end noise scenario, several intelligibility enhancement methods have been developed based on the optimization of objective measures. While the objective measures used, such as the speech intelligibility index (SII) (ANSI S3-5.1997, 1997) or glimpse proportion (GP) (Cooke, 2006), cannot completely capture the complexity of human speech understanding, they are known to correlate with subjective intelligibility and can therefore be used to analyze how different acoustical cues affect intelligibility. For instance, Sauert and Vary (2010) used the SII to determine optimal gains for the sub-bands of the unprocessed speech signal by maximizing the objective intelligibility with constrained audio power. Taal et al. (2013) improved this approach further by utilizing a more accurate approximation of the SII at low signal-to-noise ratios (SNRs), and they obtained improved results in subjective evaluations. Similar intelligibility results were obtained by Kleijn and Hendriks (2015) using an optimization criterion based on the mutual information shared between the original message and the received symbols. Interestingly, their model leads to a measure that is related to the SII, but that can be additionally used to model observation and interpretation noise. Schepker et al. (2013) combined the SII optimization scheme with adaptive dynamic range compression (DRC), and their study also showed intelligibility improvements over unprocessed speech. When studying a scenario where speech corruption resulted from both near-end noise and reverberation, Hendriks et al. (2015) used an approximation of the short-time SII to find the optimal frequency weightings and showed improvement over previously proposed methods. Tang and Cooke (2011) used SNR to locate time-frequency bins that require enhancement and later extended this method to different noise types by utilizing the GP measure (Tang and Cooke, 2012). Although they obtained positive results with the latter method, the optimization of the objective measure was done off-line and it is therefore not directly suitable for real-time applications. Petkov et al. (2013) maximized the likelihood of noisy speech given a statistical model of clean speech to derive the optimal band-energy gains. Again, this technique is applicable only for a restricted domain because it calls for a transcription of the noisy speech, which is not generally available.

Post-filtering is a special type of post-processing in which an adaptive filter is used to reallocate energy in the frequency domain from perceptually less important regions to those regions that are more relevant from the standpoint of quality and intelligibility. Traditionally, post-filtering has been employed to improve the perceptual quality of speech by utilizing a filter that emphasizes spectral peaks and attenuates spectral valleys where the level of the quantization noise surpasses the speech level (Chen and Gersho, 1995; Grancharov et al., 2008). In intelligibility enhancement, the traditional post-filter may be replaced with a high-pass type filter. It effectively attenuates low frequencies, where most of the noise energy is usually located in the majority of noise types encountered in real life. Additionally, a high-pass type filter amplifies the level of high frequencies between 1 kHz and 4 kHz, thus resulting in increased speech intelligibility (Hall and Flanagan, 2010; Jokinen et al., 2013; 2012; Niederjohn and Grotelueschen, 1976).

Instead of using the high-pass type of post-filtering technique, it is also possible to use more advanced post-filtering algorithms that model how the human speech production mechanism functions in natural conversations when humans are trying to overcome communication barriers, such as background noise. Such natural phenomena include in particular the Lombard effect, which is observed when talkers modify their speaking style to make their speech more intelligible in the presence of environmental noise (Summers et al., 1988). The Lombard effect corresponds to multiple modifications to the speech signal, such as increased vocal intensity and fundamental frequency (F_0), changed formant frequencies, longer word durations, and decreased spectral tilt. In a study by Skowronski and Harris (2006), energy reallocation was utilized to transfer energy from voiced sounds to unvoiced utterances. Zorilă et al. (2012) took advantage of adaptive spectral shaping aimed at sharpening the formants and reducing the spectral tilt, which they then combined with DRC. While their approach was shown to improve both objective and subjective intelligibility, it is partially based on long-term energy normalization, which is not suitable for speech transmission applications where signal delay is a limiting factor. Griffin et al. (2015) proposed an additional noise reduction section preceding the spectral shaping and DRC to improve performance in situations where the received signal already contains noise. This resulted in improved intelligibility when quantified using objective measures, but the study did not involve any subjective evaluations. Jokinen et al. (2014b) combined the spectral tilt reduction and formant sharpening in a post-filtering method that improved intelligibility with subjective tests in various noise conditions.

The majority of the intelligibility enhancement methods presented above, as well as more generally in the literature, focus on explicitly modifying the magnitude spectrum of speech or changing the time domain signal using methods such as DRC. Although the perceptual effects of the phase spectrum have been studied (Kazama et al., 2010; Liu et al., 1997; Paliwal and Alsteris, 2005) and phase processing has been used in speech technology, for example, in feature extraction of speech recognition (Alsteris and Paliwal, 2005; Schlüter and Ney, 2001) and in speech enhancement (Lu and Loizou, 2008; Mowlaee and Saeidi, 2013; Paliwal et al., 2011), the phase spectrum is still a relatively unexplored topic in the field of speech intelligibility enhancement. The magnitude spectrum is a natural choice for frequency-domain intelligibility enhancement methods because, for example, the Lombard effect manifests itself in magnitude spectrum features. In contrast to the main stream magnitude domain processing, however, Jokinen et al. (2014a) recently proposed a phase-based approach for intelligibility enhancement in mobile phones. With this technique, the dynamic range of the speech signal's time-domain sample values in a frame is first reduced by modifying only the signal's phase spectrum while keeping the amplitude spectrum intact. After phase processing, the signal can be amplified so that it occupies the original range of its sample values, hence achieving a gain in speech energy that ultimately results in increased loudness and intelligibility. The results from objective intelligibility evaluations done by Jokinen et al. (2014a) were positive, but no subjective listening tests were conducted in the study. Likewise, the idea of reducing the dynamic range of the signal by modifying the phase spectrum alone has also been used in the context of audio processing (Parker and Välimäki, 2013). Notably, phase-based modifications are also taken advantage of, for instance, in peak-to-average power ratio reduction in telecommunications (Bauml et al., 1996; Wang et al., 2010) and in the synthesis of generic harmonic signals with low peak amplitudes (Horner and Beauchamp, 1996; Schroeder, 1970), but, importantly, these application domains deal with signals that are both temporally and in terms of spectral content considerably different from speech. A technique for phase dispersion of speech signals was proposed by Quatieri and McAulay (1991). This technique, combined with amplitude compression and spectral shaping, was shown to produce peak-to-average power ratio reduction, but it removes all the original phase information.

The goal of this study is to evaluate whether phase-based modifications, in contrast to more commonly exploited magnitude spectrum or time-domain modifications, can be used to enhance the intelligibility of telephone speech in near-end noise conditions. In other words, the study aims to investigate whether modifying the