Enhancement of speech intelligibility in near-end noise conditions with phase modification

Emma Jokinen\textsuperscript{1,2}, Marko Takanen\textsuperscript{1}, Hannu Pulakka\textsuperscript{3}, Paavo Alku\textsuperscript{1}

\textsuperscript{1}Aalto University, Department of Signal Processing and Acoustics, Espoo, Finland
\textsuperscript{2}International Audio Laboratories Erlangen, Friedrich-Alexander Universität (FAU), Germany
\textsuperscript{3}Lumia Audio Technology, Microsoft, Tampere, Finland

\texttt{emma.jokinen@aalto.fi}

Abstract

Post-processing methods can be used in mobile communications to improve the intelligibility of speech in adverse near-end noise conditions. This study proposes a phase spectrum modification method for intelligibility enhancement in mobile devices. The method first modifies the phase spectrum of speech in order to reduce the amplitude range of the signal. The amplitude level is then equalized to that of the unprocessed speech hence resulting in amplification of signal energy. The performance of the proposed method was evaluated in comparison to a known post-processing method with an objective intelligibility metric in multiple noise conditions. Results indicate that the proposed post-processing method improves speech intelligibility over the reference method.

Index Terms: Intelligence enhancement, phase modification, telephone speech

1. Introduction

Mobile phones are used in increasingly difficult background noise conditions where the corrupting noise may be on the sending or receiving side of the communication channel. If the noise corruption takes place at the receiving side, referred to as the near-end noise scenario, post-processing can be used in the receiving mobile device to enhance the intelligibility of speech. This scenario assumes that the decoded speech is free of noise, except for distortion due to quantization and coding, and the post-processing aims to enhance the acoustical cues of speech and thus, to increase the prominence of speech over the environmental noise in the listener’s surroundings.

Several intelligibility enhancement methods based on the optimization of objective measures have been developed for near-end noise scenarios. The utilized measures, such as the speech intelligibility index (SII) \cite{1} or the glimpse proportion (GP) \cite{2}, are known to correlate with subjective intelligibility. For instance, the SII was used in \cite{3}, where optimal gains for the sub-bands of the unprocessed speech signal were determined by maximizing the objective intelligibility with constrained audio power. This approach was further enhanced by utilizing a more accurate approximation of the SII \cite{4} and by combining it with adaptive dynamic range compression (DRC) \cite{5}. In \cite{6}, frequency regions with low signal-to-noise ratios (SNRs) were enhanced and this method was later extended to different noise types by utilizing the GP measure in \cite{7}, where an offline optimization of the objective measure was conducted. In \cite{8}, the maximization of the likelihood of noisy speech given a statistical model of clean speech was utilized to derive the optimal band-energy gains. This technique, however, calls for a transcription of the noisy speech which makes the algorithm applicable for a restricted domain only.

In post-filtering, an adaptive filter is used to reallocate energy in the frequency domain from perceptually less important regions to those which are more relevant in terms of quality or intelligibility. Traditionally, post-filtering has been employed in improving the perceptual quality by utilizing a filter that emphasizes spectral peaks and attenuates spectral valleys where the level of the quantization noise surpasses the speech level \cite{9, 10}. In intelligibility enhancement, the traditional post-filter may be replaced with a high-pass type filter that effectively attenuates the low frequencies, where most of the noise energy usually is, and enhances the frequencies between 1 kHz to 4 kHz, resulting in increased speech intelligibility \cite{11, 12, 13, 14}.

More sophisticated post-filtering algorithms emulate phenomena that occur in natural conversations when humans are trying to overcome communication barriers, such as background noise. Such natural phenomena include, e.g., the Lombard effect which is observed when talkers modify their speaking style to make their speech more intelligible in the presence of environmental noise \cite{15}. The Lombard effect corresponds to multiple modifications to the speech signal, such as increased vocal intensity and fundamental frequency \((F_0)\) and changed formant frequencies, longer word durations, and decreased spectral tilt. In \cite{16}, energy reallocation was utilized to transfer energy from voiced sounds to unvoiced utterances. In \cite{17}, adaptive spectral shaping, aimed at sharpening the formants and reducing the spectral tilt, was combined with DRC. The approach was shown to improve both objective and subjective intelligibility but is partially based on long-term energy normalization which is not suitable for real-time processing. In \cite{18, 19}, the spectral tilt reduction and formant sharpening were combined in a post-filtering method that was shown to improve intelligibility in various noise conditions.

Most of the intelligibility enhancement methods developed so far focus solely on modifying the magnitude spectrum of speech. The magnitude spectrum is a natural choice for frequency domain intelligibility enhancement methods, because e.g., the Lombard effect is manifested in the magnitude features of the speech spectrum. The phase spectrum, however, offers a relatively unexplored domain for intelligibility enhancement. To the best of our knowledge, algorithms intended for intelligibility enhancement in a post-processing context based on phase modifications have not been proposed previously.

The purpose of this study is to introduce a phase-based approach for intelligibility enhancement and to evaluate its performance in comparison with an existing post-processing method. The proposed method is implementable in mobile devices (e.g.,...
of low delay) and takes advantage of frame-based amplitude normalization which is based on the dynamic range restrictions in mobile phones set by the amplifiers. The main idea of the introduced algorithm is to modify the phase spectrum in order to reduce the dynamic range of the speech signal without touching the magnitude spectrum. This dynamic range reduction can be utilized to conserve the battery life of a device or, as it was used in this study, to obtain an energy and intelligibility gain by using amplitude equalization after the phase modification. The performance evaluation was conducted with narrowband speech because it is still prevalent in mobile communications even though wideband speech transmission is becoming increasingly popular. Additionally, intelligibility enhancement is more important for narrowband speech which has fewer speech cues and is thus more severely affected by environmental noise.

2. Proposed method

The goal of the phase modification (PM) is to express a time domain signal on a smaller amplitude range by changing the signal’s phase spectrum while keeping the magnitude components untouched. However, the selection of the frequency components to be modified and the determination of the optimal phases for the selected components is not a trivial task. Computational techniques, such as machine learning or numerical optimization (e.g. [20], [21]), have been used in similar problems, but these methods are computationally intensive and the solution is still suboptimal because no closed form solution exists to the best of our knowledge.

The approach adopted in this paper is computationally light and operates with a short delay to accommodate the requirements set by the target application domain, mobile devices. First, the incoming speech frame is transformed to the frequency domain using the fast Fourier transform (FFT), the harmonics are estimated and the most energetic “frequency cone” (FC) is selected for the phase modification. The term “frequency cone” is used to refer to a group of consecutive frequency components centered around a harmonic peak, as depicted in Fig. 2. After the FC is located, an exhaustive search is utilized to find for the entire FC the phase shift that minimizes the amplitude of the time domain signal. By scaling the amplitude to the level of the unprocessed frame after the phase modification, the energy of the speech frame is increased and the speech stands out better from the background noise.

The incoming speech signal is processed with a 8-kHz sampling frequency in 16-ms frames which are first windowed with \( w_n = \sin(\pi n/2N) \cdot (n + 0.5) \) [22], where \( N \) is the length of the window. To reduce audible artefacts at frame borders, the same window is also applied after the processing and a 50 % overlap between consecutive frames is used. The incoming speech frames are classified either as silence or speech using frame energy. In addition, avoicing decision is made based on the gradient-index [23] to further categorize speech frames into voiced and unvoiced. Frames classified as silence or unvoiced are not processed. The processing chain for frames classified as voiced speech is depicted in Fig. 1.

First, a 128-point FFT of the speech frame is computed and divided into the magnitude and phase components. The harmonic peaks corresponding to the multiples of the fundamental frequency (\( F_0 \)) are located from the magnitude spectrum. The \( F_0 \) estimate is obtained directly from the adaptive multi-rate (AMR) decoder [24] which this study assumes for to be used in speech transmission. To compensate for the possible inaccuracies in the \( F_0 \) estimate, it is smoothed and the multiples of the smoothed estimate, \( F_n = (n + 1) \cdot F_0 \), between 300 Hz and the critical phase frequency [25] are used as initial points for the search of the harmonic peaks. The critical phase frequency, \( f_c = \kappa F_0 \), denotes the frequency below which a phase modification of a harmonic component is perceptually irrelevant. The value of \( \kappa \) is computed as

\[
\kappa = \left[ Q_{\text{out}} \left( 1 - \frac{B_{\text{min}}}{F_0} \right) - 0.5 \right] ,
\]

where \( Q_{\text{out}} = 9.26449 \) and \( B_{\text{min}} = 24.7 \). Although the critical phase frequency has been derived for strictly harmonic signals and does not necessarily apply for general speech signals, it was found to be suitable for restricting the search space to the low frequency range. To find the exact locations of the harmonic peaks in the spectrum, the local maxima of the magnitude spectrum in the
frequency regions surrounding the estimated harmonics are searched. Once the peaks have been resolved, the smoothed $F'_0$ estimate is utilized to approximate the width of the FC in the spectrum as $|F'_0 - F'_0/2, F'_0 + F'_0/2|$, where $F'_0$ is the center frequency of the estimated harmonic. The most energetic FC is selected for the phase modification.

To prevent discontinuities between consecutive frames, all subsequent phase values are updated before they are modified. That is, the respective phase changes that were made in previous frames are added to the current phase values. After this, the updated phases of the components within the selected FC are modified such that the component corresponding to the harmonic peak is delayed or advanced by $[-D_{\text{MAX}}, D_{\text{MAX}}]$ steps, where $D_{\text{MAX}} = 4$ and the step size is set to 0.2 samples. The maximum shift was selected based on informal listening where larger shifts tended to cause audible artefacts. The shift is implemented such that the components on the edges of the FC experience zero delay and the shift is linearly increased until it reaches the desired value at the harmonic peak in the center of the FC. This approach provides a smoothly changing group delay function. To select the optimal shift, $\Delta_{\text{OPT}}$, all possible shifts between $-D_{\text{MAX}}$ and $D_{\text{MAX}}$ are evaluated individually and the one producing the smallest maximum amplitude in the time domain signal is chosen. To reduce the computational load, full inverse FFT (IFFT) is not utilized in transforming the processed signal to the time domain. Instead, only the time domain signal corresponding to the modified components is computed. The majority of the frequency components remain unchanged during each round of the optimization and can be, therefore, computed once in the beginning of the frame. An example of the amplitude reduction obtained by the phase modification before normalization can be seen in Fig. 3.

After the phase modification, amplitude normalization is used as in [26]. First, the maximum amplitude of the processed frame is compared with the original frame and a scaling factor is determined. Each sample is then multiplied with an updated scaling factor $\beta_n = \alpha \beta_{n-1} + (1-\alpha) \gamma_A$, where $\gamma_A$ is the amplitude scaling factor for the frame and $\alpha = 0.9$. The energy gains, defined here as the ratio of the RMS energies of processed and unprocessed speech, obtained by the PM algorithm are demonstrated in Fig. 4.

3. Reference method

The formant equalizing post-filter (FE) was introduced by Hall and Flanagan [12] for enhancing the intelligibility of wide-band speech with a 22.05-kHz sampling frequency. Their post-processing method utilizes a fixed high-pass filter which was obtained by inverting the average amplitudes of the first two formants measured from adult male speakers. As a result, the filter attenuates the approximate region of the first formant with maximum attenuation near 360 Hz. As the filter was originally intended for wideband speech it was modified for narrowband speech using the $z$ transform given in the original paper [12].

The long-term average spectra of the PM method compared to the spectra of the reference method and to that of unprocessed speech is depicted in Fig. 5.

4. Objective evaluation

The performance of the proposed method was evaluated and compared to that of the FE post-filter in terms of an objective intelligibility metric, the SII [1]. For the evaluation, 160 Finnish speech samples with a duration of 2-3 seconds from nine speakers (4 male, 5 female) were used. For four of the speakers, speech was taken from a database recorded originally for intelligibility testing [27]. For the remaining five talkers, the data was taken from a new, unpublished database containing both normal and Lombard speech for the training of speech synthesizers. Only the normal speech was used for the evaluation.

The speech samples were preprocessed to resemble narrow-band telephone speech by first downsampling them to 16 kHz and then filtering with the MSIN filter [28]. The MSIN filter is a high-pass filter with a cut-off at 195 Hz designed to simulate mobile station input characteristics. After the filtering, the speech samples were downsampled to 8 kHz, encoded and decoded with the AMR codec [24] and equalized to $-26$ dB to SV56 [28, 29]. Finally, the samples were processed with one of the methods (FE and PM conditions).

Six different types of background noise were involved in the evaluation: car, factory, cafeteria, babble, white and high-pass noise. Car, factory, cafeteria and babble noise are com-
mon types of environmental distortion of low-pass characteristics. Both car and babble noise are stationary while factory and cafeteria noise are non-stationary with occasional loud clangs. White and high-pass noise were included to evaluate the algorithms’ performance with a noise power spectral density differing from the most common low-pass type. Each of the noise types was used with five different SNR levels: $-10$ dB, $-5$ dB, 0 dB, 5 dB, and 10 dB. The SII metric was computed for each speech sample by first removing silent periods from the samples, computing the SII in segments of 9.4 ms and then averaging the obtained values over the whole sample.

The obtained SII values were analyzed with a four-way analysis of variance (ANOVA) procedure using 5% significance level. The method (UN, FE, PM), the SNR level ($-10$ dB, $-5$ dB, 0 dB, 5 dB, 10 dB), and the speaker gender (male, female) were modelled as fixed factors, while the speaker (9 of which 5 female) was modelled as a random factor nested within the speaker gender. The noise type (car, factory, cafeteria, babble, white and high-pass noise) was excluded from the final model since it or its interactions with other factors did not have a significant effect on the results. A visual inspection verified the underlying ANOVA assumptions about the normality of the residuals and the normality of the random effects. The ANOVA showed that the method $[F(2,16)=383.9, p < 0.001]$, the SNR level $[F(4,32)=7329.3, p < 0.001]$, as well as the interactions between the method and the SNR level $[F(8,64)=13.3, p < 0.001]$ and between the method and the speaker gender $[F(2.6)=28.4, p < 0.01]$ affected significantly the SII values.

Subsequently, the marginal means and their 95% confidence intervals were computed to investigate the nature of the effects and the statistical significance of the findings was confirmed with the Dunnett’s T3 post-hoc test using the 5% significance level. The values shown in Fig. 6 illustrate that PM achieved at all SNR levels a higher SII score than UN and distinguished itself from FE in the worst SNR condition. The difference between FE and UN increased with SNR level, but failed to reach statistical significance. On one hand, Fig. 7 illustrates that the SII score of male speech was, on average, improved more by PM than that of female speech.

5. Conclusion

A post-processing method based on the modification of the phase spectrum (PM) was introduced and compared to an existing post-filtering approach (FE) and to unprocessed speech (UN) in terms of an objective intelligibility metric, the speech intelligibility index, in several background noise conditions. The proposed method is based on reducing the dynamic range of speech via phase modifications after which amplitude equalization can be used to achieve a gain in energy. Post-processing approaches aimed at intelligibility enhancement in the near-end noise scenario based on modifications of the phase spectrum have not been previously studied. Furthermore, the proposed algorithm has short delay and low computational complexity and is therefore suitable for implementation in mobile devices.

The objective results show that the proposed method achieved higher intelligibility scores than UN and in low SNR conditions it also outperformed FE. Based on these promising results, future studies will include subjective tests, such as in [13] and in [19], where the proposed method will be compared to other known post-processing methods, such as dynamic range reduction as in [26].

6. Acknowledgements

The research leading to these results was supported by the GETA graduate school, Nokia, the Academy of Finland (research programmes 256961 and 13251770), and the Mide/Uiart project of Aalto University.
7. References


