A SPEECH PREPROCESSING STRATEGY FOR INTELLIGIBILITY IMPROVEMENT IN NOISE BASED ON A PERCEPTUAL DISTORTION MEASURE

Cees H. Taal, Richard C. Hendriks and Richard Heusdens

Signal and Information Processing Lab, Delft University of Technology, the Netherlands
2628 CD Delft, the Netherlands
Email: {c.h.taal, r.c.hendriks, r.heusdens}@tudelft.nl

ABSTRACT

A speech pre-processing algorithm is presented to improve the speech intelligibility in noise for the near-end listener. The algorithm improves the intelligibility by optimally redistributing the speech energy over time and frequency for a perceptual distortion measure, which is based on a spectro-temporal auditory model. In contrast to spectral-only models, short-time information is taken into account. As a consequence, the algorithm is more sensitive to transient regions, which will therefore receive more amplification compared to stationary vowels. It is known from literature that changing the vowel-transient energy ratio is beneficial for improving speech-intelligibility in noise. Objective intelligibility prediction results show that the proposed method has higher speech intelligibility in noise compared to two other reference methods, without modifying the global speech energy.

Index Terms— Near-end speech enhancement, intelligibility improvement, transients

1. INTRODUCTION

An important goal in speech-communication systems is to transmit a speech signal, such that it is correctly understood by the receiver. Examples can be found in the field of telephony and public address systems. Unfortunately, the speech intelligibility can be harmed due to background noise. While a decrease in speech intelligibility can be annoying in a telephone conversation, it could be potentially dangerous in the context of, for example, a voice alarm in a fire detection system. As illustrated in Fig. 1, the speech intelligibility for the near-end listener can be affected by background noise from both sides of the communication channel. That is, the noise can come from both the *far end* and the *near end*. In order to eliminate the negative impact of the far-end noise, one would typically apply a single-channel noise-reduction algorithm (see [1] for an overview). However, the speech can also be pre-processed before playback in order to become more intelligible in presence of the near-end background noise, which is the focus in this work.

To improve the speech intelligibility in a noisy environment one obvious solution would be to increase the playback level. However, at a certain point increasing the playback level may not be possible anymore due to loudspeaker limitations. Moreover, unpleasant playback levels may be reached which are close to the threshold of pain. An alternative approach would be to fix the speech energy and redistribute energy within the speech signal over time and frequency. For example, it is well-known that transient parts of speech, e.g., consonants, play an important role in speech intelligibility [2], while their energy is relatively low compared to vowels and therefore more vulnerable to noise. As a consequence, many strategies change the energy ratio between the vowels and consonants which leads to an improvement of speech intelligibility in noise [3, 4, 5]. However, these strategies are applied independent of the near-end noise, while for certain applications knowledge of the noise statistics are available and can be exploited. More recently, Sauert and Vary proposed several algorithms [6, 7], which take into account the noise. These methods improve objective speech intelligibility as predicted by the speech intelligibility index (SII) [8]. However, these methods only change the spectrum of the speech and do not use some type of consonant detection strategy. Therefore, the benefits from the earlier mentioned transient-enhancement strategies may not be present in this method.

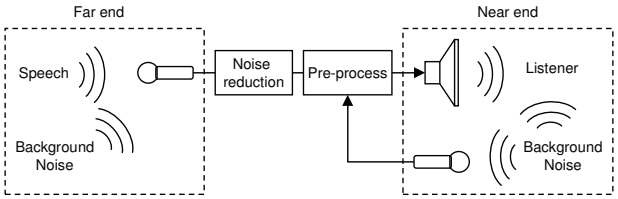

Fig. 1. Application scenario of intelligibility improvement for the near-end listener.

In this work we present a method where the speech energy is optimally re-distributed as a function of the near-end noise, relevant for a perceptual distortion measure. We assume that the recorded noisy speech from the near-end is of good quality and intelligibility. This work contributes due to the fact that the considered distortion measure is based on a spectro-*temporal* auditory model in contrast to a spectral-only model as in [6, 7]. Therefore, the proposed method is more sensitive to transient regions which will receive more amplification compared to vowels.

2. PROPOSED SPEECH PRE-PROCESSING ALGORITHM

Let x denote a time-domain signal representing clean speech and $x + \varepsilon$ a noisy version, where ε represents background noise. The distortion measure considered in this work, denoted by $D(x, \varepsilon)$, will inform us about the audibility of ε in the presence of x. Hence, a lower D value implies less audible noise and therefore more audible speech. Our goal is to adjust the speech signal x such that $D(x, \varepsilon)$ is minimized subject to the constraint that the energy of the modified speech remains unchanged.

First, in Section 2.1 more details will be given about the consid-

This research is supported by the Oticon Foundation and the Dutch Technology Foundation STW.

ered distortion measure, after which in Section 2.2 we will formalize and solve the constrained optimization problem. In Section 2.3 some properties of the algorithm are revealed.

2.1. Perceptual Distortion Measure

The perceptual distortion measure is based on the work from [9], which takes into account a spectro-*temporal* auditory model and therefore also considers the temporal envelope within a short-time frame (20-40 ms), in contrast to spectral-only models. As a consequence, the distortion measure is more sensitive to transients, which are of importance for speech intelligibility.

First, a time-frequency (TF) decomposition is performed on the speech and noise by segmenting into short-time (32 ms), 50% overlapping Hann-windowed frames. Then, a simple auditory model is applied to each short-time frame, which consists of an auditory filter bank followed by the absolute squared and low-pass filtering per band, in order to extract a temporal envelope. Here, the filter bank resembles the properties of the basilar membrane in the cochlea, while the envelope extraction stage is used as a crude model of the hair-cell transduction in the auditory system.

Let h_i denote the impulse response of the i^{th} auditory filter and x_m the m^{th} short-time frame of the clean speech. Their linear convolution is denoted by $x_{i,m} = x_m * h_i$. Subsequently, the temporal envelope is defined by $|x_{m,i}|^2 * h_s$, where h_s represents the smoothing low-pass filter. Similar definitions hold for $|\varepsilon_{m,i}|^2 * h_s$. The cutoff frequency of the low-pass filter determines the sensitivity of the model towards temporal fluctuations within a short-time frame[1]. The audibility of the noise in presence of the speech, within one TF-unit, is determined by a per-sample noise-to-signal ratio [9]. By summing these ratios over time, an intermediate distortion measure for one TF-unit is obtained denoted by lower-case d. That is,

$$d\left(x_{m,i}, \varepsilon_{m,i}\right) = \sum_{n} \frac{\left(|\varepsilon_{m,i}|^2 * h_s\right)(n)}{\left(|x_{m,i}|^2 * h_s\right)(n)}, \tag{1}$$

where n denotes the time index running over all samples within one short-time frame. The distortion measure for the complete signal is then obtained by summing all the individual distortion outcomes over time and frequency, which gives,

$$D\left(x, \varepsilon\right) = \sum_{m,i} d\left(x_{m,i}, \varepsilon_{m,i}\right). \tag{2}$$

2.2. Power-Constrained Speech-Audibility Optimization

To improve the speech audibility in noise, we minimize Eq. (2) by applying a gain function α which redistributes the speech energy, i.e., $\alpha_{m,i} x_{m,i}$, where $\alpha_{m,i} \geq 0$. Only TF-units are modified where speech is present. This is done in order to prevent that a large amount of energy would be redistributed to speech-absent regions. We consider a TF-unit to be speech-active, when its energy is within a 25 dB range of the TF-unit with maximum energy within that particular frequency band. The noise is assumed to be a stochastic process denoted by $\mathcal{E}_{m,i}$ and the speech deterministic (recall that the speech signal is known in the near-end enhancement application). Hence, we minimize for the expected value of the distortion measure. Let $\mathcal{L}$ denote the set of speech-active TF-units and $\|\cdot\|$ the ℓ_2-norm, the problem can then be formalized as follows,

$$\min_{\alpha_{m,i}, \{m,i\}\in\mathcal{L}} \sum_{\{m,i\}\in\mathcal{L}} E[d\left(\alpha_{m,i} x_{m,i}, \mathcal{E}_{m,i}\right)] \quad \text{s.t.} \sum_{\{m,i\}\in\mathcal{L}} \|\alpha_{m,i} x_{m,i}\|^2 = r, \tag{3}$$

where $r = \sum_{\{m,i\}\in\mathcal{L}} \|x_{m,i}\|^2$ relates to the power constraint. By using the method of Lagrange multipliers we introduce the following cost function,

$$J = \sum_{\{m,i\}\in\mathcal{L}} E[d\left(\alpha_{m,i} x_{m,i}, \mathcal{E}_{m,i}\right)] + \lambda \left(\sum_{\{m,i\}\in\mathcal{L}} \|\alpha_{m,i} x_{m,i}\|^2 - r \right). \tag{4}$$

Due to the linearity of the convolution in Eq. (1) and the assumption that $\alpha \geq 0$ we have that $d\left(\alpha x, y\right) = d\left(x, y\right)/\alpha^2$. Therefore, we have to solve the following set of equations for α for minimizing Eq. (4),

$$\begin{array}{rcl} \frac{\partial J}{\partial \alpha_{m,i}} & = & -2\frac{E\left[d\left(x_{m,i}, \mathcal{E}_{m,i}\right)\right]}{\alpha_{m,i}^3} + \lambda 2 \alpha_{m,i} \|x_{m,i}\|^2 = 0 \\ \frac{\partial J}{\partial \lambda} & = & \sum\limits_{\{m,i\}\in\mathcal{L}} \alpha_{m,i}^2 \|x_{m,i}\|^2 - r = 0 \end{array} \tag{5}$$

The solution is given by,

$$\alpha_{m,i}^2 = \frac{r \beta_{m,i}^2}{\sum\limits_{\{m',i'\}\in\mathcal{L}} \beta_{m',i'}^2 \|x_{m',i'}\|^2}, \tag{6}$$

where,

$$\beta_{m,i} = \left(\frac{E\left[d\left(x_{m,i}, \mathcal{E}_{m,i}\right)\right]}{\|x_{m,i}\|^2} \right)^{1/4}. \tag{7}$$

In order to determine α we have to evaluate the expected value $E\left[d\left(x_{m,i}, \mathcal{E}_{m,i}\right)\right]$, which can be expressed as follows,

$$E\left[d\left(x_{m,i}, \mathcal{E}_{m,i}\right)\right] = \sum_{n} \frac{\left(E\left[|\mathcal{E}_{m,i}|^2\right] * h_s\right)(n)}{\left(|x_{m,i}|^2 * h_s\right)(n)}, \tag{8}$$

To simplify, we assume that the power-spectral density of the noise within the frequency range of an (relatively narrow) auditory band is constant, i.e., has a 'flat' spectrum. As a consequence, the noise within an auditory band can be modeled by $\mathcal{E}_{m,i} = (w_m N_{m,i}) * h_i$, where w_m denotes the window function and $N_{m,i}$ represents a zero mean, i.i.d. stochastic process with variance $E\left[N_{m,i}^2(n)\right] = \sigma_{m,i}^2, \forall n$. By combining this statistical model and the numerator of Eq. (8) we have,

$$\begin{array}{rcl} E\left[|\mathcal{E}_{m,i}|^2(n)\right] & = & E\left[\left| \sum\limits_{k} h_i(k)\, w_m(n-k)\, N_{m,i}(n-k) \right|^2 \right] \\ & = & \sum\limits_{k} h_i^2(k)\, w_m^2(n-k)\, E\left[N_{m,i}^2(n-k)\right] \\ & = & \left(h_i^2 * w_m^2\right)(n)\, \sigma_{m,i}^2. \end{array} \tag{9}$$

Here $\sigma_{m,i}^2$ is estimated with the noise PSD estimator from [10] by taking the average PSD within an auditory band.

As a final step, an exponential smoother is applied to $\alpha_{m,i}$ in order to prevent 'musical noise' which may negatively effect the speech quality[2],

[1] the envelopes for the auditory filters with low center frequencies are already low-pass signals, therefore for complexity reasons these low-pass filters may be discarded.

[2] The energy of the signals is normalized to account for the small possible error introduced due to the exponential smoother

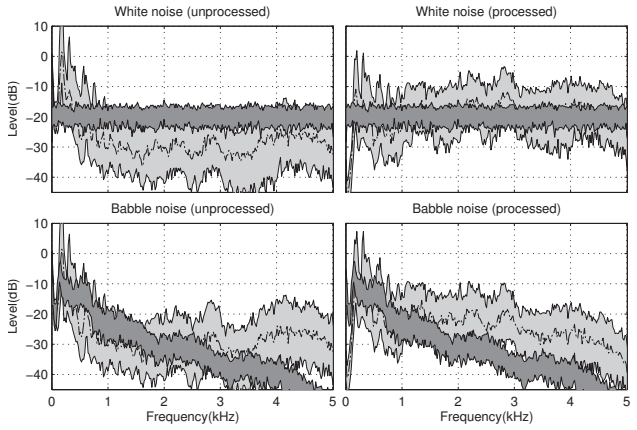

Fig. 2. 25%-75% quantile range of noise (dark-gray) and speech (light-gray) log-spectral magnitudes before and after processing for white noise and babble noise.

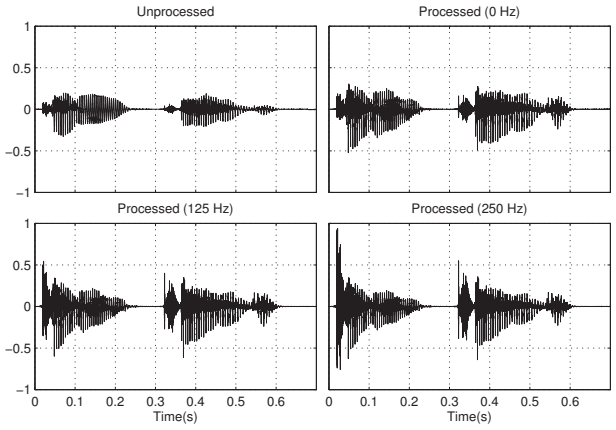

Fig. 3. Unprocessed and processed speech for the Dutch excerpt '*Tom tekent*' ('*Tom draws*' in English) for three different auditory model cutoff frequencies. Notice that the consonant '*t*' is automatically amplified when the cutoff frequency is increased.

$$\hat{\alpha}_{m,i} = (1-\gamma)\,\alpha_{m,i} + \gamma\hat{\alpha}_{m-1,i}, \tag{10}$$

where $\gamma = 0.9$.

To reduce complexity, the filter bank and the low-pass filter are applied by means of a point-wise multiplication in the DFT-domain with real-valued, even-symmetric frequency responses[3]. For the filter bank the approach as presented in [11] is used and for the low-pass filter the magnitude response of a one-pole low-pass filter is used. A total amount of 40 ERB-spaced filters are considered between 150 and 5000 Hz. Furthermore, the speech signal is reconstructed by addition of the scaled TF-units where a square-root Hann-window is used for analysis/synthesis.

2.3. Algorithm Analysis

Fig. 2 illustrates the effect of the proposed algorithm in the frequency domain for white noise and babble noise. Here the 25%-75% quantile range is shown for all speech and noise short-time DFT magnitudes for one sentence, denoted by the light and dark area, respectively. The top row shows the results for white noise and the bottom row for babble noise, before (left) and after (right) processing. Note that the energy before and after processing remaines unchanged. Overall it can be observed that the speech audibility is clearly improved for both noise types over frequency. To accomplish this, the algorithm gives the speech more or less the average spectral shape of the noise. It is known from literature that this type of frequency shaping of the speech signal indeed improves intelligiblity [6]. However, rather than a heuristic choice this is a direct result of the optimal derivations from the previous section which take into account the power constraint.

The cutoff frequency of the auditory model lowpass filter h_s (see Section 2.1) determines the temporal sensitivity of the distortion measure. For example, a higher cutoff frequency will result in a larger intermediate distortion value for transient signals while a cutoff of 0 Hz would equal a spectral-only distortion measure. To demonstrate the benefits of taking into account short-time information, i.e., a cutoff frequency larger than zero, its effect is shown in Fig. 3. Here, the Dutch speech excerpt '*Tom tekent*' ('*Tom draws*' in English) is degraded with white noise at -5 dB SNR (noise signal is not shown for visibility) where three cutoff frequencies are considered: 0, 125 and 250 Hz. The plots clearly show that the proposed algorithm distributes more energy to the transient regions when the cutoff frequency is increased, from which we know that this will improve speech intelligibility [4, 5]. This means that the vowel-consonant energy ratio can be adjusted automatically with only one parameter. Based on informal listening tests the cutoff frequency is set to 125 Hz.

[3]This particular choice will lead to time-domain aliasing due to circular convolution, however, the applied window function will minimize the effect of these unwanted artifacts.

3. EXPERIMENTAL EVALUATION

To evaluate the performance of the proposed (PROP) method and compare it to several reference methods, speech is degraded with babble, F16, factory and white noise for an SNR-range between -15 and 5 dB. In total, 50 random sentences from a female speaker are used from the Dutch matrix test [12]. For all experiments a sample rate of 16000 Hz is used. A comparison is made with two other algorithms. That is, the method of maximal power transfer proposed by Sauert *et. al* (SAU) [6] which applies a TF-dependent gain function and takes into account the noise. Secondly, our results are compared with the method from [4] which modifies the vowel-transient ratio. In our experiments, the energy is redistributed for a complete sentence at once (around 3 seconds). Applications for this situation would be when the speech is pre-recorded in environments where the noise is known, e.g., navigation voice in a car or safety announcements in an airplane. Note, that the delay of the proposed method can be reduced by restricting the amount of TF-units in $\mathcal{L}$ taken into account from the past. In near future research we will evaluate low-delay performance of the algorithm.

Two objective intelligibility predictors are applied before and after processing. The first method is the short-time objective intelligibility (STOI) measure [13] and the second measure is the coherence speech intelligibility index (CSII) [14]. Both measures can predict the intelligibility of noisy speech and various nonlinear speech degradations. The results are shown in Figs. 4 and 5, where the plots show that for all noise types a significant intelligibility improvement is predicted. A conclusion which is in line with informal listening tests. The proposed method shows better performance compared to

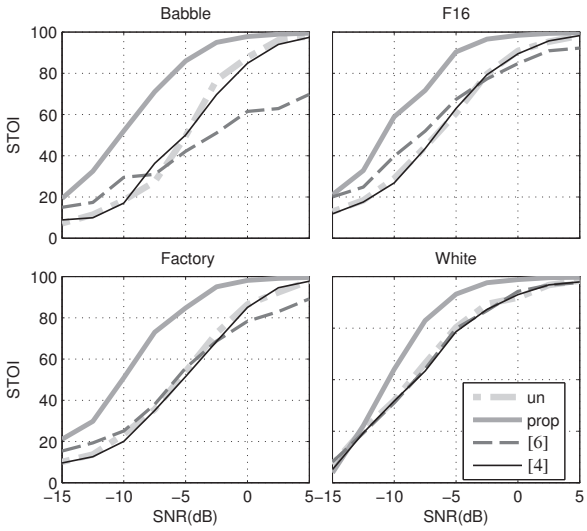

Fig. 4. STOI intelligibility predictions for the proposed method (PROP), the unprocessed noisy speech (UN), the method from Sauert *et. al* [6] and a method based on voiced/unvoiced energy redistribution [4]. A higher STOI-score denotes better intelligibility.

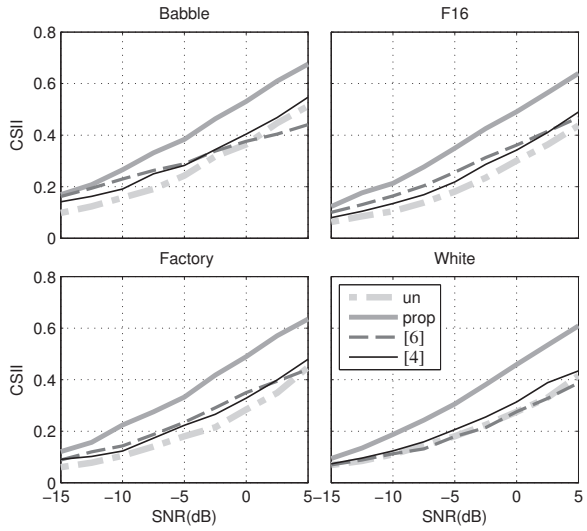

Fig. 5. CSII intelligibility predictions for the proposed method (PROP), the unprocessed noisy speech (UN), the method from Sauert *et. al* [6] and a method based on voiced/unvoiced energy redistribution [4]. A higher CSII-score denotes better intelligibility.

the reference methods for all noise types.

4. CONCLUDING REMARKS

A speech pre-processing algorithm is presented to improve the speech intelligibility in noise for the near-end listener. This was accomplished by optimally redistributing the speech energy over time and frequency based on a perceptual distortion measure. Due to the fact that the distortion measure takes into account short-time information, transient signals, which are more important for speech intelligibility than vowels, receive more amplification. Objective intelligibility prediction results show that with the proposed algorithm, the SNR can be lowered 3-5 dBs without losing intelligibility.

5. REFERENCES

[1] P. C. Loizou, *Speech enhancement: theory and practice*, CRC, Boca Raton, FL, 2007.

[2] W. Strange, J.J. Jenkins, and T.L. Johnson, "Dynamic specification of coarticulated vowels," *J. Acoust. Soc. Am.*, vol. 74, no. 3, pp. 695–705, 1983.

[3] R. Niederjohn and J. Grotelueschen, "The enhancement of speech intelligibility in high noise levels by high-pass filtering followed by rapid amplitude compression," *IEEE Trans. on Acoust., Speech, Signal Process.*, vol. 24, no. 4, pp. 277 – 282, 1976.

[4] M.D. Skowronski and J.G. Harris, "Applied principles of clear and Lombard speech for automated intelligibility enhancement in noisy environments," *Speech Communication*, vol. 48, no. 5, pp. 549–558, 2006.

[5] V. Hazan and A. Simpson, "The effect of cue-enhancement on the intelligibility of nonsense word and sentence materials presented in noise," *Speech Communication*, vol. 24, no. 3, pp. 211 – 226, 1998.

[6] B. Sauert, G. Enzner, and P. Vary, "Near end listening enhancement with strict loudspeaker output power constraining," in *Proceedings of International Workshop on Acoustic Echo and Noise Control (IWAENC)*, 2006.

[7] B. Sauert and P. Vary, "Near end listening enhancement optimized with respect to speech intelligibility index and audio power limitations," in *Proceedings of European Signal Processing Conference (EUSIPCO)*, 2010.

[8] ANSI, "Methods for calculation of the speech intelligibility index," *S3.5-1997, (American National Standards Institute, New York)*, 1997.

[9] C. H. Taal and R. Heusdens, "A low-complexity spectro-temporal based perceptual model," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2009, pp. 153–156.

[10] R. C. Hendriks, R. Heusdens, and J. Jensen, "MMSE based noise PSD tracking with low complexity," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2010, pp. 4266–4269.

[11] S. van de Par, A. Kohlrausch, R. Heusdens, J. Jensen, and S.H. Jensen, "A perceptual model for sinusoidal audio coding based on spectral integration," *EURASIP J. on Appl. Signal Processing*, vol. 2005, no. 9, pp. 1292–1304, 2005.

[12] J. Koopman, R. Houben, W. A. Dreschler, and J. Verschuure, "Development of a speech in noise test (matrix)," in *8th EFAS Congress, 10th DGA Congress*, Heidelberg, Germany, June 2007.

[13] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Trans. Audio Speech Lang. Process.*, vol. 19, no. 7, pp. 2125–2136, 2011.

[14] J. M. Kates and K. H. Arehart, "Coherence and the speech intelligibility index," *J. Acoust. Soc. Am.*, vol. 117, no. 4, pp. 2224–2237, 2005.