A Low-Complexity Spectro-Temporal Distortion Measure for Audio Intelligibility Improvement

CH. KRISHNA1
PG Scholar, Vaagdevi College of Engineering, Warangal, AP-INDIA. E-mail: chilukakrishna@gmail.com.

V. SABITHA2
Asst Prof, Vaagdevi College of Engineering, Warangal, AP-INDIA.

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 other reference methods, without modifying the global speech energy. The Kalman filter has been the subject of extensive research and application, particularly in the area of autonomous or assisted navigation. The Kalman filter is a mathematical power tool that is playing an increasingly important role in computer graphics as we include sensing of the real world in our systems. Electrical engineers are picturing the Kalman filter as a design tool for speech, whereby it can estimate and resolve errors that are contained in speech after passing through a distorted channel. Due to this motivating fact, there are many ways a Kalman filter can be tuned to suit engineering applications such as network telephony and even satellite phone conferencing.

Keywords: Near-End Speech Enhancement, Intelligibility Improvement, Transients, Kalman Filter.

I. 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. 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.

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

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 improvement. Therefore, the algorithm redistributes the speech energy by taking into account the short-time information, which is more sensitive to transient regions. The objective of this work is to present a low-complexity spectro-temporal distortion measure for improving speech intelligibility in near-end listeners.
role in speech intelligibility, while their energy is relatively low compared to vowels and therefore more vulnerable to noise. This research is supported by the Oticon Foundation and the Dutch Technology. As a consequence many strategies change the energy ratio between the vowels and consonants which leads to an improvement of speech intelligibility in noise. However, these strategies are applied independent of the near-end noise, while for certain applications knowledge of the noise statistics is available and can be exploited. More recently, Sauert and Vary proposed several algorithms, which take into account the noise. These methods improve objective speech intelligibility as predicted by the speech intelligibility index (SII). 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.

In order to understand these processes, both human and machine speech has to be studied carefully on the structures and functions of spoken language: how we produce and perceive it and how speech technology may assist us in communication. Therefore in this thesis, we will be looking more into speech processing with the aid of an interesting technology known as the Kalman Filter. Presently, this technique is not widely used in the field of signal processing, however it is a potential nominee to be considered. New filtering techniques are also of consideration to the success of speech processing. One of the common adaptive filtering techniques that are applied to speech is the Wiener filter. This filter is capable of estimating errors however at only very slow computations. On the other hand, the Kalman filter suppresses this disadvantage.

The Kalman filter as a design tool for speech, whereby it can estimate and resolve errors that are contained in speech after passing through a distorted channel. The fact that preserving information, which is contained in speech, is of extreme importance, the availability of signal filters such as the Kalman filter is of great importance we will be looking more into speech processing with the aid of an interesting technology known as the Kalman Filter. Presently, this technique is not widely used in the field of signal processing, however it is a potential nominee to be considered. More details on Speech Processing and the Kalman filter will be explained in the later chapters of this thesis report.

II. PROPOSED SPEECH PRE-PROCESSING ALGORITHM

Let \( x \) denote a time-domain signal representing clean speech and \( x + \varepsilon \) a noisy version, where \( \varepsilon \) represents background noise. The distortion measure considered in this work, denoted by \( D(x, \varepsilon) \), will inform us about the audibility of \( \varepsilon \) 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.

![Fig2. Proposed Speech Pre-Processing Algorithm](image)

First, in Section 2.1 more details will be given about the considered 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 2.4 implementation of kalman filter.

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, time-frequency (TF) decomposition is performed on the speech and noise by segmenting into short-time (24 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 \( x_{m,i} = x_m * h_s \). Subsequently, the temporal envelope is defined by \( |x_m|^2 * h_s \), where \( h_s \) represents the smoothing low-pass filter. Similar definitions hold for \( |e_{m,i}|^2 * h_s \). The cutoff frequency of the low-pass filter determines the sensitivity of the model towards temporal fluctuations within a short-time frame. The audibility of the noise in presence of the speech, within one TF-unit, is determined by a per-sample noise-to-signal ratio. By summing these ratios over time, an intermediate distortion measure for one TF-unit is obtained denoted by lower-case \( d \). That is, 
\[
d(x_{m,i}, e_{m,i}) = \sum_n \left( \frac{|e_{m,i}|^2 + h_s}{|x_{m,i}|^2 + h_s} \right) (n),
\]
(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,
\[
D(x, e) = \sum_{m,i} d(x_{m,i}, e_{m,i}).
\]
(2)

### 2.2 Power-Constrained Speech-Audibility Optimization

To improve the speech audibility in noise, we minimize Eq. (2) by applying a gain function \( \alpha \) which redistributes the speech energy, i.e., \( \alpha_s \frac{x_m}{x_m} \), where \( \alpha_s \geq 0 \). 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 \( e_{m,i} \) 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 \( L \) denote the set of speech-active TF-units and \( \| \cdot \| \) the \( l_1 \)-norm, the problem can then be formalized as, 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.

\[
\min_{\alpha_{m,i} \in L, \| \cdot \|} \sum_{m,i} E[d(\alpha_{m,i} x_{m,i}, e_{m,i})] \leq \sum_{m,i} \| \alpha_{m,i} x_{m,i} \|^2 = r
\]
(3)

Where 'r' relates to the power constraint.

By using the method of Lagrange multipliers we introduce the following cost function,
\[
J = \sum_{m,i} E[d(\alpha_{m,i} x_{m,i}, e_{m,i})] + \lambda \sum_{m,i} \| \alpha_{m,i} x_{m,i} \|^2 - r
\]
(4)

Due to the linearity of the convolution in Eq. (1) and the assumption that \( \alpha \geq 0 \) we have that \( d(\alpha x, y) = d(x, y) / \alpha^2 \). Therefore, we have to solve the following set of equations for \( \alpha \) for minimizing Eq. (4),
\[
\frac{\partial J}{\partial \alpha_{m,i}} = -2 E[d(\alpha_{m,i} x_{m,i}, e_{m,i})] + \lambda 2 \alpha_{m,i} \| x_{m,i} \|^2 = 0
\]
\[
\frac{\partial J}{\partial \lambda} = \sum_{m,i} \alpha_{m,i} \| x_{m,i} \|^2 - r = 0
\]
(5)

The solution is given by
\[
\alpha_{m,i} = \left( \sum_{m,i} \| x_{m,i} \|^2 \right)^{-1} \left( \sum_{m,i} \| x_{m,i} \|^2 \right)\beta_{m,i}
\]
(6)

Where,
\[
\beta_{m,i} = \left( E[d(\alpha_{m,i} x_{m,i}, e_{m,i})] \right)^{1/4}
\]
(7)

In order to determine \( \alpha \) we have to evaluate the expected value \( E[d(\alpha_{m,i} x_{m,i}, e_{m,i})] \), which can be expressed as follows,
\[
E[d(\alpha_{m,i} x_{m,i}, e_{m,i})] = \sum_n \left( \frac{E[|e_{m,i}|^2 + h_s]}{|x_{m,i}|^2 + h_s} \right) (n)
\]
(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 \( e_{m,i} = w_m N_{m,i} + h_s \), where \( w_m \) denotes the window function and \( N_{m,i} \) represents a zero mean, i.i.d. stochastic process with \( \sigma_{N_{m,i}}^2 \) given.

By combining this statistical model and the numerator of Eq. (8) we have,
\[
E[|e_{m,i}|^2] = E \left[ \sum_{k} w_m^2 (n-k) N_{m,i} (n-k)^2 \right] = \sum_{k} w_m^2 (n-k) E[N_{m,i}^2 (n-k)^2] = \sigma_{N_{m,i}}^2.
\]
(9)

Here \( \sigma_{N_{m,i}}^2 \) 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 affect the speech quality,2

\[
\hat{\alpha}_{m,i} = \left( 1 - \gamma \right) \alpha_{m,i} + \gamma \hat{\alpha}_{m-1,i}
\]
(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 DWT-domain with real-valued, even-symmetric frequency responses.3 For the filter bank the approach as presented in 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 Henning-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% quintile range is shown for all speech and noise short-time DWT 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 remains 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 intelligibility. 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.

Fig. 3. 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. 4. 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.

Fig. 5. From a statistical point of view, many signals such as speech exhibit large amounts of correlation. From the perspective of coding or filtering, this correlation can be put to good use. The all pole, or autoregressive (AR), signal model is often used for speech. From Crisafulli et al, the AR signal model is introduced as:

\[
y_k = \frac{1}{1 - \sum_{i=1}^{N} a_i z^{-i}} w_k
\]

Equation (11) can also be written in this form as shown below:

\[
y_k = a_1 y_{k-1} + a_2 y_{k-2} + \ldots + a_N y_{k-N} + w_k
\]

Where,
A Low-Complexity Spectro-Temporal Distortion Measure for Audio Intelligibility Improvement

In order to apply Kalman filtering to the speech expression shown above, it must be expressed in state space form as

\[
H_k = X_{k-1} + W_k
\]  

\[
y_k = \mathbf{g} H_k
\]  

Where

\[
X = \begin{bmatrix}
a_1 & a_2 & \cdots & a_{N-1} & a_N \\
1 & 0 & \cdots & 0 & 0 \\
0 & 1 & \cdots & 0 & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & \cdots & 1 & 0
\end{bmatrix}
\]

\[
Y_k = \begin{bmatrix}
y_k \\
y_{k-1} \\
y_{k-2} \\
\vdots \\
y_{k-N+1}
\end{bmatrix}
\]

\[
W_k = \begin{bmatrix}
w_k \\
0 \\
0 \\
\vdots \\
0
\end{bmatrix}
\]

\[
g = [1 \ 0 \ \cdots \ 0]
\]

\[
W_k = [\begin{bmatrix}
1 & \cdots & 0 \\
0 & \cdots & 0 \\
\vdots & \ddots & \vdots \\
0 & \cdots & 1
\end{bmatrix}]
\]

\[
K_k = P_{k-1}H_k^T[ H_k^T P_k H_k + R ]^{-1}
\]

With

\[
K_k \quad \text{is the Kalman Gain Matrix,}
\]

\[
P_{k-1} \quad \text{is the a priori error covariance matrix,}
\]

\[
R \quad \text{is Measurement noise covariance.}
\]

Thereafter the reconstructed speech signal, \( Y_k \) after Kalman filtering will be formed in a manner similar to (4.2):

\[
Y_k = a_1 Y_{k-1} + a_2 Y_{k-2} + \ldots + a_N Y_{k-N} + w_k
\]

The flow chart for the simulation of this project is illustrated in the below figure.

---

**Fig. 6:** Flow chart
III. EXPERIMENTAL EVALUATION

To evaluate the performance of the proposed (PROP) method and compare it to several reference methods, speech is degraded with 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. For all experiments a sample rate of 16000 Hz is used. A comparison is made with one other algorithm. That is, the method of maximal power transfer proposed by Sauert et al. (SAU) which applies a TF-dependent gain function and takes into account the noise. Secondly, our results are compared with the method 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 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 and the second measure is the coherence speech intelligibility index (CSII). Both measures can predict the intelligibility of noisy speech and various nonlinear speech degradations. The Kalman filter was successful in removing most of the noise from noisy speech signal. The results are shown in Fig. 11 and 12, 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 the reference methods for all noise types.

Fig 7. Pure Input Speech Signal.

Fig 8. Input Speech Signal Spectrogram.

Fig 9. Noise Speech Signal.

Fig 9a. Noise Speech Signal Spectrum White noise.
A Low-Complexity Spectro-Temporal Distortion Measure for Audio Intelligibility Improvement

Fig10. STOI intelligibility predictions for the proposed Method.

Fig.10. STOI intelligibility predictions for the proposed method (PROP), the unprocessed noisy speech (UN), a method based on voiced/unvoiced Energy redistribution [4]. A higher STOI-score denotes better intelligibility.

Fig11. CSII intelligibility predictions for the proposed method.

Fig.11. CSII intelligibility predictions for the proposed method (PROP), the unprocessed noisy speech (UN), a method based on voiced/unvoiced Energy redistribution [4]. A higher CSII-score denotes better intelligibility.

IV. CONCLUSION

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. Without losing intelligibility. Filtering, smoothing, noise extraction are possible and Effective Numerical simulations provide good results. Classification of series of structured objects (folding of proteins, music,
etc. Specification of the matrices involved Bayesian learning of the model Noise model corresponding to white noise in feature space the Kalman filter helped to reduce or eliminate most of the noise. Objective intelligibility prediction results show that with the proposed algorithm, the SNR can be lowered 3-5 dBs without losing intelligibility.

V. SOFTWARE DESCRIPTION

MATLAB 7.5: MATLAB® is a high-performance language for technical computing. It integrates computation, visualization, and programming in an easy-to-use environment where problems and solutions are expressed in familiar mathematical notation

VI. REFERENCES


