

An Improved Speech Processing Strategy for Cochlear Implants Based on objective measures for predicting speech intelligibility

G B Pavan Kumar

Electronics and Communication Engineering
Andhra University
INDIA

Prof P Mallikarjuna Rao

Electronics and Communication Engineering
Andhra University
India

Abstract— The purpose of this study was to improve the speech processing strategy for cochlear implants (CIs). A speech pre-processing algorithm is presented to improve the speech intelligibility in noise. The algorithm improves the intelligibility by optimally redistributing the speech energy over time and frequency for a perceptual distortion measure, the algorithm is more sensitive to transient regions. Two objective intelligibility predictors are applied before and after processing without modifying the global speech energy. Kalman filter is used to calculate estimated errors

Keywords—algorithm for perceptual distortion; Methods for speech intelligibility prediction ;STOI; coherence SII; Kalman filter;

1. INTRODUCTION

COCHLEAR implant (CI) is an auditory neural prosthesis for restoring hearing function in patients with sensori neural hearing loss. Hearing restoration is achieved by electrically stimulating the auditory nerve, and the electrical stimulation pulse parameters are derived from incoming speech-by-speech processors contained within the CI devices. Essentially, the speech processing strategy of the CI mimics the basic function of the peripheral auditory system. Most modern devices utilize a filter bank for frequency decomposition of incoming speech, which is a simplification of the frequency decomposition function of a biological cochlea, i.e., the place coding (tonotopy) of auditory information. A simple linear band pass filter bank is used for most CI devices.

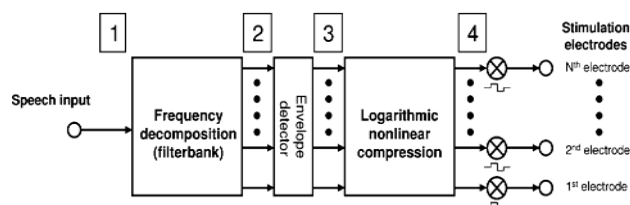


Fig. 1 general structure of the speech processor in a CI.

The structure was originally motivated by the place coding (tonotopy) of the basilar membrane. Incoming speech is first decomposed into multiple channels with different

frequency ranges. The relative strengths of multiple channels are obtained from envelope detectors, and the envelopes of sub-bands are used to modulate the amplitudes of stimulus pulses

To date, the performance of the CI has been significantly improved over time with the development of various speech processing strategies. Successful speech perceptions in quiet environments are possible for most recipients, but diminished CI performance occurs in noisy conditions.

The major purpose of this study was to develop a novel speech processing strategy by a method where the speech energy is optimally re-distributed as a function of the near-end noise, relevant for a perceptual distortion measure and to improve the Intelligibility based on two objective intelligibility Methods the first method is the short-time objective intelligibility (STOI) measure [13] and the second measure is the coherence speech intelligibility index (CSII). The estimation of the driving noise variance and of the additive noise variance are handled after a preliminary Kalman filtering

2. SPEECH PRE-PROCESSING ALGORITHM

Let x denote a time-domain signal representing clean speech and $x + \epsilon$ a noisy version, where ϵ represents background noise. The distortion measure considered in this work, denoted by $D(x, \epsilon)$, 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 $D(x, \epsilon)$ is minimized subject to the constraint that the energy of the modified speech remains unchanged.

a) The 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-timeframe (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 (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 are 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 $|\epsilon_{m,i}|^2 * h_s$. 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,

$$d(x_{m,i}, \epsilon_{m,i}) = \sum_n \frac{(|\epsilon_{m,i}|^2 * h_s)(n)}{(|x_{m,i}|^2 * h_s)(n)} \quad (1)$$

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

$$D(x, \epsilon) = \sum_{m,i} d(x_{m,i}, \epsilon_{m,i}) \quad (2)$$

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. 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 25dB range of the TF-unit with maximum energy within that particular frequency band. The noise is assumed to be a stochastic process denoted by $\epsilon_{m,i}$ 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_2 -norm, the problem can then be formalized as follows, 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.

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

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

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

Due to the linearity of the convolution in Eq. (1), we have to solve the following set of equations for α for minimizing. (4),

$$\frac{\partial J}{\partial \alpha_{m,i}} = -2 \frac{E[d(x_{m,i}, \epsilon_{m,i})]}{\alpha_{m,i}^3} + \lambda 2 \alpha_{m,i} \|x_{m,i}\|^2 = 0$$

$$\frac{\partial J}{\partial \lambda} = \sum_{\{m,i\} \in L} \alpha_{m,i}^2 \|x_{m,i}\|^2 - r = 0 \quad (5)$$

The solution is given by

$$\alpha_{m,i}^2 = \frac{r \beta_{m,i}^2}{\sum_{\{m',i'\} \in L} \beta_{m',i'}^2 \|x_{m',i'}\|^2} \quad (6)$$

where,

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

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

$$E[d(x_{m,i}, \epsilon_{m,i})] = \sum_n \frac{E[|\epsilon_{m,i}|^2 * h_s](n)}{(|x_{m,i}|^2 * h_s)(n)} \quad (8)$$

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,

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

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. 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 filters 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.

3 METHODS FOR SPEECH INTELLIGIBILITY PREDICTION

Existing objective speech-intelligibility measures are suitable for several types of degradation, however, it turns out that they are less appropriate for methods where noisy speech is processed by a time frequency (TF) weighting, e.g., noise reduction and speech separation. In this paper, we present an objective intelligibility measure, which shows high correlation ($\rho=0.95$) with the intelligibility of both noisy, and TF-weighted noisy speech.

The proposed method shows significantly better performance than three other, more sophisticated, objective measures. Furthermore, it is based on an intermediate intelligibility measure for short-time (approximately 400ms) TF-regions, and uses a simple DFT-based TF-decomposition.

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).

4 SHORT-TIME OBJECTIVE INTELLIGIBILITY (STOI)

One of the first OIMs was developed at AT&T Bell Labs by French and Steinberg in 1947, currently known as the articulation index (AI). AI evolved to the speech-intelligibility index (SII), and has been standardized in 1997 under ANSI 3.5-1997. Later, the speech transmission index (STI) was proposed, which, in contrast

to AI, is also able to predict the intelligibility of various simple nonlinear degradations, e.g. clipping.

The majority of recent published models are still based on the fundamentals of AI, and STI (see for an overview of STI-based measures). Although the just mentioned OIMs are suitable for several types of degradation (e.g., additive noise, reverberation, filtering, clipping), it turns out that they are less appropriate for methods where noisy speech is processed by a time-frequency (TF) weighting.

This includes single-microphone speech-enhancement algorithms but also speech separation techniques like ideal time frequency segregation (ITFS), where typically a binary TF-weighting is used. STI and various STI-based measures predict an intelligibility improvement when spectral subtraction is applied. This is not in line with the results of listening experiments in literature, where it is reported that general single-microphone

Speech-enhancement algorithms are not able to improve the intelligibility

of noisy speech. Furthermore, OIMs like the coherence SII [5] and a covariance-based STI procedure, both show low correlation with the intelligibility of ITFS-processed speech. Only recently, two different OIMs are proposed which indicate promising results for ITFS-processed speech

Existing objective speech-intelligibility measures are suitable for several types of degradation, however, it turns out that they are less appropriate for methods where noisy speech is processed by a time frequency (TF) weighting, e.g., noise reduction and speech separation.

To analyze the effect of certain signal degradations on the speech-intelligibility in more detail, the OIM must be of a simple structure, i.e., transparent. However, some OIMs are based on a large amount of parameters which are extensively trained for a certain dataset. This makes these measures less transparent, and therefore less appropriate for these evaluative purposes. Moreover, OIM's are often a function of long-term statistics of entire speech signals and do not use an intermediate measure for local short-time TF regions. With these measures it is difficult to see the effect of a time-frequency localized signal-degradation on the speech intelligibility

In this method, we present an objective intelligibility measure, which shows high correlation ($\rho=0.95$) with the intelligibility of both noisy, and TF-weighted noisy speech. The proposed method shows significantly better performance than three other, more sophisticated, objective measures. Furthermore, it is based on an intermediate intelligibility measure for short-time (approximately 400 ms) TF-regions, and uses a simple DFT-based TF-decomposition.

The proposed method is a function of the clean and processed speech, denoted by x and y , respectively. The model is designed for a sample-rate of 10000 Hz, in order to cover the relevant frequency range for speech-intelligibility. Any signals at other sample-rates should be resampled. Furthermore, it is assumed that the clean and the processed signal are both time-aligned.

First, a TF-representation is obtained by segmenting both signals into 50% overlapping, Hanning-windowed frames with a length of 256 samples, where each frame is zero-padded up to 512 samples and Fourier transformed. Then, an one-third octave band analysis is performed by grouping DFT-bins. In total 15 one-third octave bands are used, where the lowest center frequency is set equal to 150Hz.

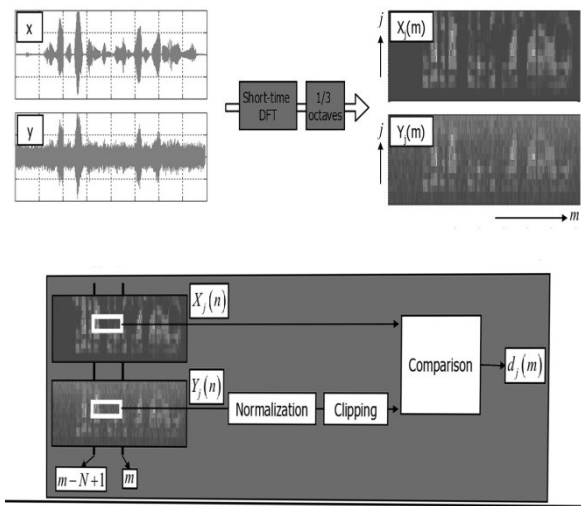


Fig 2 STOI process schematic representation

Let $\hat{x}(k, m)$ denote the k th DFT-bin of the m th frame of the clean speech. The norm of the j th one-third octave band, referred to as a

$$X_j(m) = \sqrt{\sum_{k=k_1(j)}^{k_2(j)-1} |\hat{x}(k, m)|^2} \quad (11)$$

Where k_1 and k_2 denote the one-third octave band edges, which are rounded to the nearest DFT-bin. The TF-representation of the processed speech is obtained similarly, and will be denoted by $Y_j(m)$. The intermediate intelligibility measure for one TF-unit, say $d_j(m)$, depends on a region of N consecutive TF-units from both $X_j(n)$ and $Y_j(n)$, where $n \in M$ and $=\{m-N+1, (m-N+2), \dots, m-1, m\}$. First, a local normalization procedure is applied, by scaling all the TF-units from $Y_j(n)$ with a factor

$$\alpha = \left(\frac{\sum_n X_j(n)^2}{\sum_n Y_j(n)^2} \right)^{1/2}$$

Such that its energy equals the clean speech energy, within that TF-region. Then, $\alpha Y_j(n)$ is clipped in order to lower bound the Signal-to-distortion ratio (SDR), which we define as,

$$SDR_j(n) = 10 \log_{10} \left(\frac{X_j(n)^2}{(\alpha Y_j(n) - X_j(n))^2} \right) \quad (12)$$

Hence

$$Y' = \max(\min(\alpha Y, X + 10^{-\beta/20} X), X - 10^{-\beta/20} X) \quad (13)$$

Where Y' represents the normalized and clipped TF-unit and β denotes The lower SDR bound. The frame and one-third octave band indices are omitted for notational convenience. The intermediate intelligibility measure is defined as an estimate of the linear correlation coefficient between the clean and modified processed TF-units,

$$d_j(m) = \frac{\sum_n \left(X_j(n) - \frac{1}{N} \sum_l X_j(l) \right) \left(Y'(n) - \frac{1}{N} \sum_l Y'(l) \right)}{\sqrt{\sum_n \left(X_j(n) - \frac{1}{N} \sum_l X_j(l) \right)^2} \sqrt{\sum_n \left(Y'(n) - \frac{1}{N} \sum_l Y'(l) \right)^2}} \quad (14)$$

Where $l \in M$. Finally, the eventual OIM is simply given by the average of the intermediate intelligibility measure over all bands and frames,

$$d = \frac{1}{JM} \sum_{j,m} d_j(m) \quad (15)$$

Where M represents the total number of frames and J the number of One-third octave bands. In our experiments, we used different values of $N \in [20, 30, 40, 50, 60]$ and $\beta \in [-\infty, -30, -20, -15, -10]$. Maximum correlation is obtained with $\beta = -15$ and $N = 30$, which means that the intermediate measure depends on speech information from the last ≈ 400 ms

5 COHERENCE SPEECH INTELLIGIBILITY INDEX (CSII)

Other extensions to the SII measure were proposed by Kates and Arehart(2005) for predicting the intelligibility of peak-clipping and center-clipping distortions in the speech signal, such as those found in hearing aids. The modified index, called the CSII index, used the base form of the SII procedure, but with the SNR estimate replaced by the signal to-distortion ratio, which was computed using the coherence function between the input and processed signals.

While a modest correlation was obtained with the CSII index, a different version was proposed that divided the speech segments into three level regions and computed the CSII index separately for each level region. The three-level CSII index yielded higher correlations for both intelligibility and subjective quality ratings of hearing-aid type of distortions. Further testing of the CSII index is performed in the present study to examine whether it can be used to predict the intelligibility of speech corrupted by fluctuating maskers and 2 to predict the intelligibility of noise suppressed speech containing different types of non-linear distortions than those introduced by hearing aids.

The STI measure by (Steeneken and Houtgast, 1980) is based on the idea that the reduction in intelligibility caused by additive noise or reverberation distortions can be modeled in terms of the reduction in temporal envelope modulations.

The STI metric has been shown to predict successfully the effects of reverberation, room acoustics, and additive noise. It has also been validated in several languages. In its original form the STI measure used artificial signals _e.g., sine wave(modulated signals) as probe signals to assess the reduction in signal modulation in a number of frequency bands and for a range of modulation frequencies _0.6–12.5 Hz_ known to be important for speech intelligibility. When speech is subjected, however, to non-linear processes such as those introduced by dynamic envelope compression _or expansion_ in hearing aids, the STI measure fails to successfully predict speech intelligibility since the processing itself might introduce additional modulations which the STI measure interprets as increased SNR

For that reason, several modifications have been proposed to use speech or speech-like signals as probe signals in the computation of the STI measure. Despite of these modifications, several studies have reported that the speech-based STI methods fail to predict the intelligibility of nonlinearly-processed speech .several modifications were made to existing speech-based STI measures but none of these modifications were validated with intelligibility scores obtained with human listeners.

The SII and speech-based STI measures can account for linear distortions introduced by filtering and additive noise, but have not been tested extensively in conditions where in non-linear distortions might be present

The increased modulation might be interpreted as increased SNR by the STI measure. Hence, it remains unclear whether the speech-based STI measures or the SII measure can account for the type of distortions introduced by noise-suppression algorithms and to what degree they can predict speech intelligibility. It is also not known whether any of the numerous objective measures that have been proposed to predict speech quality in voice communications applications can be used to predict speech intelligibility.

An objective measure that would predict well both speech intelligibility and quality would be highly desirable in voice communication and hearing-aid applications.

The objective quality measures are primarily based on the idea that speech quality can be modeled in terms of differences in loudness between the original and processed signals

The perceptual evaluation of speech quality (PESQ) objective measure, for instance, assesses speech quality by estimating the overall loudness difference between the noise-free and processed signals. This measure has been found to predict very reliably _ the quality of telephone networks and speech codec's as well as the quality of noise-suppressed speech. Only a few studies have tested the PESQ measure in the context of predicting speech intelligibility. High correlation was reported, but it was for a relatively small number of noisy conditions which included speech processed via low-rate vocoders and speech processed binaurally via beam forming algorithms. The speech distortions introduced by noise-suppression algorithms (based on single-microphone recordings) differ, however, from those introduced by low-rate vocoders. Hence, it is not known whether the PESQ measure can predict reliably the intelligibility of noise-suppressed speech containing various forms of Non-linear distortions, such as musical noise.

OBJECTIVE MEASURES

A number of objective measures are examined in the present study for predicting the intelligibility of speech in noisy conditions. Some of the objective measures (PESQ) have been used successfully for the evaluation of speech quality while others are more appropriate for intelligibility assessment. A description of these measures along with the proposed modifications to speech-based STI and AI-based measures is given next.

THE PERCEPTUAL EVALUATION OF SPEECH QUALITY (PESQ)

Among all objective measures considered, the PESQ measure is the most complex to compute and is the one recommended by for speech quality assessment of 3.2 kHz _narrow-band_ handset telephony and narrow-band speech codec's The PESQ measure is computed as follows. The original (clean) and degraded signals are first level equalized to a standard listening level and filtered by a filter with response similar to that of a standard telephone handset. The signals are time aligned to correct for time delays, and then processed through an auditory transform to obtain the loudness spectra. The difference in loudness between the original and degraded signals is computed and averaged over time and frequency to produce the prediction of subjective quality rating. The PESQ produces a score between 1.0 and 4.5, with high values indicating better quality. High correlations ($r=0.92$) with subjective listening tests were reported by using the above PESQ measure for a large number of testing conditions taken from voice-over-internet protocol applications. High correlation ($r=0.9$) was also reported

AI-BASED MEASURES

A simplified version of the SII measure is considered in this study that operates on a frame-by-frame basis. The proposed measure differs from the traditional SII measure in many ways:

- (a) It does not require as input the listener’s threshold of hearing,
- (b) Does not account for spread of upward masking,
- (c) Does not require as input the long-term average spectrum (sound-pressure) levels of the speech and masker signals.

The proposed AI-ST measure divides the signal into short 30 ms data segments, computes the AI value for each segment, and averages the segmental AI values over all frames. It can be computed as follows

$$AI - ST = \frac{1}{M} \sum_{m=0}^{M-1} \frac{\sum_{j=1}^K W(j,m) T(j,m)}{\sum_{j=1}^K W(j,m)} \tag{16}$$

Where M is the total number of data segments in the signal, $W(j, m)$ is the weight i.e., band importance function, placed on the j th frequency band, and

$$T(j, m) = \frac{SNR(j, m) + 15}{30}$$

$$SNR(j, m) = 10 \log_{10} \frac{X(j,m)^2}{D(j,m)^2} \tag{17}$$

COHERENCE-BASED MEASURES

The aim of the METHOD is to evaluate the performance of new speech-based STI measures, modified coherence-based measures, the modified Coherence-based measures and speech-based STI measures incorporating signal-specific band-importance functions yielded the highest correlations ($r=0.89-0.94$). The modified coherence measure, in particular, that only included vowel/consonant transitions and weak consonant information yielded the highest correlation ($r=0.94$) with sentence recognition scores.

To evaluate the performance of conventional objective measures originally designed to predict speech quality and to evaluate the performance of new speech-based STI measures, modified coherence-based measures (CSII), as well as AI-based measures that were designed to operate on short-term (20–30) ms intervals in realistic noisy conditions. A number of modifications to the speech-based STI, coherence-based, and AI measures are proposed and evaluated in this study.

The articulation index AI and speech-transmission index (STI) are by far the most commonly used today for predicting speech intelligibility in noisy conditions. The AI measure was further refined to produce the speech intelligibility index (SII).

The SII measure is based on the idea that the intelligibility of speech depends on the proportion of spectral information that is audible to the listener and is computed by dividing the spectrum into 20 bands

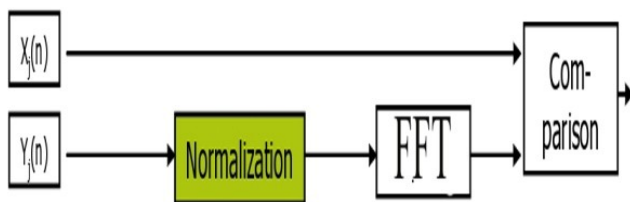


Fig 3 CSII schematic representation

The magnitude-squared coherence (MSC) function is the normalized cross-spectral density of two signals and has been used to assess distortion in hearing aids. It is computed by dividing the input (clean) and output (processed) signals in a number M of overlapping windowed

segments, computing the cross power spectrum for each segment using the FFT, and then averaging across all segments.

For M data segments (frames), the MSC at frequency bin ω is given by

$$MSC(\omega) = \frac{|\sum_{m=1}^M X_m(\omega) Y_m^*(\omega)|^2}{\sum_{m=1}^M |X_m(\omega)|^2 \sum_{m=1}^M |Y_m(\omega)|^2} \tag{18}$$

Where the asterisk denotes the complex conjugate and $X_m(\omega)$ and $Y_m(\omega)$ denote

the FFT spectra of the $x(t)$ and $y(t)$ signals, respectively, computed in the m th data segment. In our case, $x(t)$ corresponds to the clean signal and $y(t)$ corresponds to the enhanced signal. The MSC measure takes values in the range of 0–1. The averaged, across all frequency bins, MSC was used in our study as the objective measure.

The MSC was computed by segmenting the sentences using 30-ms duration Hamming windows with 75% overlap between adjacent frames. The use of a large frame overlap (50%) was found to reduce bias and variance in the estimate of the MSC. It should be noted that the above MSC function can be expressed as a weighted

The main difference between the MTF used in the computation of the STI measure and the MSC function is that the latter function is evaluated for all frequencies spanning the signal bandwidth, while the MTF is evaluated only for low modulation frequencies

The new measure, called coherence SII (CSII), was proposed that used the SII index as the base measure and replaced the SNR term with the signal-to-distortion ratio term, which was computed using the coherence between the input and output signals. That is, the $SNR(j,m)$ term in Eq(3) was replaced with the following expression

$$SNR_{CSII}(j, m) = 10 \log_{10} \frac{\sum_{k=1}^N G_j(\omega_k) * MSC(\omega_k) |Y_m(\omega_k)|^2}{\sum_{k=1}^N G_j(\omega_k) * [1 - MSC(\omega_k)] |Y_m(\omega_k)|^2} \tag{19}$$

$$CSII = \frac{1}{M} \sum_{m=0}^{M-1} \frac{\sum_{j=1}^K W(j,m) T_{CSII}(j,m)}{\sum_{j=1}^K W(j,m)} \tag{20}$$

CALCULATION OF ESTIMATED ERROR

In 1960, R.E. Kalman published his famous paper describing a recursive solution to the discrete-data linear filtering problem. Since that time, 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.

Although, the applications of Kalman filtering encompass many fields, its use as a tool is mainly for two purposes: *estimation* and *performance analysis of estimators*. Since the Kalman filter uses a complete description of the probability of its estimation errors in determining the optimal filtering gains

6 KALMAN FILTER

Theoretically, the Kalman Filter is an estimator for what is called the “linear quadratic problem”, which focuses on estimating the instantaneous “state” of a linear dynamic system perturbed by white noise. Statistically, this estimator is optimal with respect to any quadratic function of estimation errors.

In practice, this Kalman Filter is one of the greater discoveries in the history of statistical estimation theory and possibly the greatest discovery in the twentieth century. It has enabled mankind to do many things that could not have been done without it,

and it has become as indispensable as silicon in the makeup of many electronic systems

In a more dynamic approach, controlling of complex dynamic systems such as continuous manufacturing processes, aircraft, ships or spacecraft, are the most immediate applications of Kalman filter. In order to control a dynamic system, one needs to know what it is doing first. For these applications, it is not always possible or desirable to measure every variable that you want to control, and the Kalman filter provides a means for inferring the missing information from indirect (and noisy) measurements. Some amazing things that the Kalman filter can do is predicting the likely future courses of dynamic systems that people are not likely to control, such as the flow of rivers during flood, the trajectories of celestial bodies or the prices of traded commodities. From a practical standpoint, these are the perspectives that this section will present:

It aids mankind in solving problems; however, it does not solve any problem all by itself. This is however not a physical tool, but a mathematical one, which is made from mathematical models. In short, essentially tools for the mind. They help mental work become more efficient, just like mechanical tools, which make physical work less tedious. Additionally, it is important to understand its use and function before one can apply it effectively.

It uses a finite representation of the estimation problem, which is a finite number of variables; therefore this is the reason why it is said to be "ideally suited to digital computer implementation". However, assuming that these variables are real numbers with infinite precision, some problems do happen. This is due from the distinction between finite dimension and finite information, and the distinction between "finite" and "manageable" problem sizes. On the practical side when using Kalman filtering, the above issues must be considered along with the theory.

This is a complete characterization of the current state of knowledge of the dynamic system, including the influence of all past measurements. The reason behind why it is much more than an estimator is because it propagates the entire probability distribution of the variables it is tasked to estimate. These probability distributions are also useful for statistical analyses and the predictive design of sensor systems.

The estimation problem is modeled in a way that distinguishes between *phenomena* (what one is able to observe) and *noumena* (what is really going on). Above that, the state of knowledge about the *noumenais* that one can deduce from the *phenomena*. That state of knowledge is represented by probability distributions, which represent knowledge of the real world. Thus this cumulative processing of knowledge is considered a learning process. It is a fairly simple process, however quite effective in many applications. Probability distribution may be used in assessing its performance as a function of the "design parameters" of the following estimation systems:

- Types of sensors to be used;
- Locations and orientations of the various sensor types with respect to the system to be estimated;
- Allowable noise characteristics of the sensors;
- Pre-filtering methods for smoothing sensor noise;
- Data sampling rates for the various sensor types and
- The level of model simplification for reducing implementation requirements.

A system designer is able to assign an "error budget" to subsystems of an estimation system, which this is allowed by the analytical capability of the Kalman filter formalism. Moreover, it can trade off the budget allocations to optimize cost or other measures of performance while achieving a required level of estimation accuracy.

RELATIVE ADVANTAGES OF KALMAN FILTER

Below are some advantages of the Kalman filter, comparing with another famous filter known as the Wiener Filter, which this filter was popular before the introduction of Kalman filter. The information below is obtained from.

1. The Kalman filter algorithm is implementable on a digital computer, which this was replaced by analog circuitry for estimation and control when Kalman filter was first introduced. This implementation may be slower compared to analog filters of Wiener; however it is capable of much greater accuracy.
2. Stationary properties of the Kalman filter are not required for the deterministic dynamics or random processes. Many applications of importance include non-stationary stochastic processes.
3. The Kalman filter is compatible with state-space formulation of optimal controllers for dynamic systems. It proves useful towards the 2 properties of estimation and control for these systems.
4. The Kalman filter requires less additional mathematical preparation to learn for the modern control engineering student, compared to the Wiener filter.
5. Necessary information for mathematically sound, statistically-based decision methods for detecting and rejecting anomalous measurements are provided through the use of Kalman filter.

ESTIMATION OF PROCESS

After going through some of the introduction and advantages of using Kalman filter, we will now take a look at the process of this magnificent filter. The process commences with the addresses of a general problem of trying to estimate the state of a discrete-time controlled process that is governed by a linear stochastic difference equation:

$$\mathbf{X}_k = \mathbf{A}\mathbf{x}_{k-1} + \mathbf{B}\mathbf{u}_k + \mathbf{w}_{k-1} \quad (21)$$

with a measurement $\mathbf{Z}_k \in \mathcal{A}^m$ that is

$$\mathbf{Z}_k = \mathbf{H}\mathbf{X}_k + \mathbf{V}_k \quad (22)$$

The random variables $\mathbf{w}_k, \mathbf{v}_k$ and represent the process and measurement noise (respectively). We assume that they are independent of each other, white, and with normal probability distributions

$$\mathbf{P}(\mathbf{w}) \cong \mathbf{N}(\mathbf{0}, \mathbf{Q}) \quad (23)$$

$$\mathbf{P}(\mathbf{v}) \cong \mathbf{N}(\mathbf{0}, \mathbf{R}) \quad (24)$$

Ideally, the process noise covariance \mathbf{Q} and measurement noise covariance \mathbf{R} matrices are assumed to be constant, however in practice, they might change with each time step or measurement.

In the absence of either a driving function or process noise, the $n \times n$ matrix \mathbf{A} in the difference Equation (21) relates the state at the previous time step $k-1$ to the state at the current step k . In practice, \mathbf{A} might change with each time step, however here it is assumed constant. The $n \times 1$ matrix \mathbf{B} relates the optional control input $u \in \mathcal{R}^1$ to the state \mathbf{x} . \mathbf{H} which is a matrix in the measurement Equation (22) which relates the state to the measurement, z_k . In practice \mathbf{H} might change with each time step or measurement, however we assume it is constant.

IMPLEMENTATION OF KALMAN FILTER TO SPEECH

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:

$$Y_k = \frac{1}{1 - \sum_{i=1}^N a_i z^{-i}} \quad (25)$$

Equation (5.1) can also be written in this form as shown below:

$$y_k = a_1 y_{k-1} + a_2 y_{k-2} + \dots + a_N y_{k-n} + w_k \quad (26)$$

where,

- $k \rightarrow$ Number of iterations;
- $y_k \rightarrow$ current input speech signal sample;
- $y_{k-N} \rightarrow$ (N-1)th sample of speech signal;
- $a_N \rightarrow$ Nth Kalman filter coefficient; and
- $w_k \rightarrow$ excitation sequence (white noise).

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

$$K_k = P_{k-1} H_{k-1} [H_{k-1}^T P_{k-1} H_{k-1} + R]^{-1} \cdot H_{k-1}^T P_{k-1} H_{k-1} + Q$$

Where P_k is the posteriori error covariance matrix?

And

$$Q = \begin{pmatrix} 1 & 0 & \dots & \dots & 0 & 0 \\ 0 & 1 & \dots & \dots & 0 & 0 \\ \vdots & \vdots & \dots & \dots & \vdots & \vdots \\ 0 & 0 & \dots & \dots & 1 & 0 \\ 0 & 0 & \dots & \dots & 0 & 1 \end{pmatrix}$$

Thereafter the reconstructed speech signal, Y_k after Kalman filtering will be formed in a manner similar to Eq (22):

$$y_k = a_1 y_{k-1} + a_2 y_{k-2} + \dots + a_N y_{k-n} + w_k \quad (27)$$

After the calculation of gain in the form of estimated error we are eliminating to find the improved speech for the speech processing of cochlear implant patients

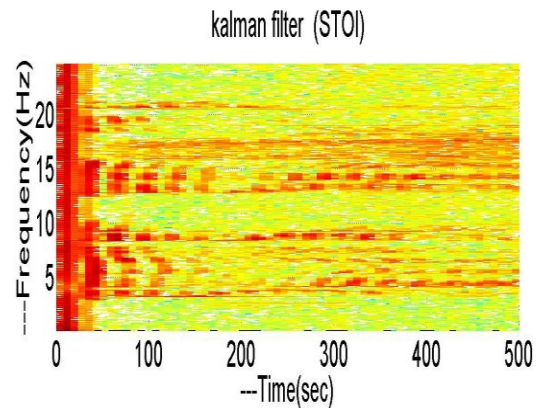
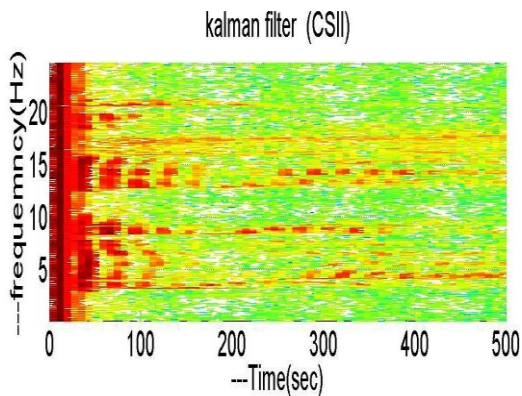
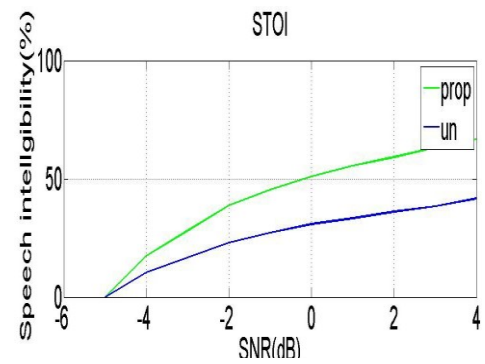
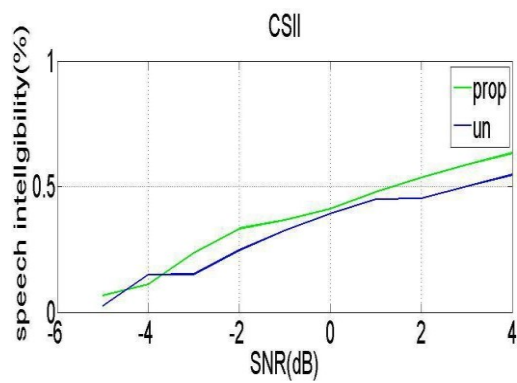


Fig 4 STOI intelligibility predictions for the proposed method (PROP), the unprocessed noisy speech (UN),

FIG 5 CSII INTELLIGIBILITY PREDICTIONS FOR THE PROPOSED METHOD (PROP), THE UNPROCESSED NOISY SPEECH (UN)

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. 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) which applies a TF-dependent gain function and takes into account the noise. Secondly, our results are compared with the method from 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 [13] 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 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 b the reference methods for all noise types.

conclusion

A speech processing strategy required for bionic ear (cochlea implant) to receive all types of audio signals for a hearing impairment patient. To improve the speech processing in noise, a filter bank contains band pass filters are taken. But diminished performance occurs in noise condition because of change in signal strength

A speech processing algorithm is presented to improve speech intelligibility accomplished by optimally redistributing the speech energy over time and frequency based on a perceptual distortion measure. Speech processing algorithm is more sensitive to transient regions, which will therefore receive more amplification compared to stationary vowels.

From the results we can observe the input signal with noise in both time domain and frequency domain.

We can observe the intelligibility of signal in two methods Objective intelligibility prediction method of Coherence speech intelligibility index (CSII) and short time objective intelligibility (STOI) results that the SNR can be lowered 3-5 dBs without losing intelligibility. By using the proposed method, we are achieving the high speech intelligibility in noise environment. The proposed algorithm is applicable to both processed and unprocessed speech signals

REFERENCES

- [1] P. C. Loizou, *Speech enhancement: theory and practice*, CRC, Boca Raton, FL,
- [2] W. Strange, J.J. Jenkins, and T.L. Johnson, "Dynamic specification Of articulated 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.onAcoust., 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 audiopower 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, NewYork), 1997.
- [9] C. H. Taal and R. Heusdens, "A low-complexity spectrotemporal 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.