The characterization of the relative information content by spectral features for the objective intelligibility assessment of nonlinerly processed speech

Anton Schlesinger, Marinus M. Boone

Acoustical Imaging and Sound Control, Delft University of Technology, The Netherlands

a.schlesinger@tudelft.nl, m.m.boone@tudelft.nl

Abstract

The objective intelligibility assessment of nonlinearly enhanced speech is a widely experienced problem. Nonlinear speech enhancement processors operate primarily on the low-level and transient components of speech. As these sections contain important acoustic cues as well as context-constitutive information, they dominate speech intelligibility. For that reason, short-time intelligibility measures at low-level and transient components are weighted with their contribution to the overall intelligibility. In this report, spectral features are calculated from auditory sub-bands and are utilized to label these sections of high information content. A genetic optimization is performed to adapt the spectral feature measures to the linearly and nonlinearly processed speech material of a listening test. No improvement is found over existing methods of objective speech intelligibility assessment, using short-time intelligibility calculation and level-dependent weighting. Therewith, the reported results contribute to pinpoint practicable solutions to the problem.

Index Terms: speech intelligibility, spectral features, entropy

1. Introduction

One of the important questions in nonlinear speech enhancement asks for the balance between the SNR improvement and the introduced distortion that is most beneficial for speech intelligibility [1]. While linear algorithms are often applied in speech processors, such as hearing aids, they generally pose a suboptimal solution in constantly changing acoustics. Nonlinear processors aim to approximate the Wiener filter, i.e., the optimal filter, in changing acoustics and can be realized as time-frequency mask-based approaches. It is yet the nonlinearity that constitutes a challenge for the subjective and objective evaluation of the audiological benefit. In the subjective evaluation of speech processors, speech reception threshold (SRT) tests are preferred over percent correct scores, which show a limited basis for generalization of the audiological benefit [2]. SRT tests on the other hand require an invariant signal-to-interferer improvement in order to derive the algorithm’s effect on speech intelligibility. Nonlinear algorithms do not satisfy this condition [2]. Moreover, subjective listening tests are laborious tasks and are not feasible during the development of complex algorithms, which are equipped with parameter sets that depend on the acoustic scene and, when considering hearing aids, on the particular auditory performance of a hearing impaired person. Objectively, speech intelligibility can be calculated from the sum of the audible contributions in different frequency bands. This forms the empirical basis of the Articulation Index (AI) theory. The Speech Intelligibility Index (SII) and the Speech Transmission Index (STI) are based on this concept and are successful in predicting intelligibility for a large number of linear degradations. In spite of that, the classical SII and the classical STI are not adequate to assess degradations of nonlinear speech enhancement. Nonlinearity violates the principle of superposition and therefore, the requirement of the classical SII approach to calculate the SNR from the isolated spectra of speech and noise cannot be met. The classical STI employs artificially modulated tones and calculates the SNR in bands from the modulation depth of the intensity envelope. The principle of superposition would again be needed to derive an overall result from these test signals. Furthermore, nonlinear speech processors tend to increase the modulation depth abnormally. The classical STI mistakes this deterioration for an increase in speech intelligibility.

Intrusive measures have been developed to deal with nonlinear distortions. These are measures that relate degraded speech to its clean reference. Some evolve with the methods of SII [3] and STI [4], others build on sophisticated perceptual models, as e.g., PESQ adapted for SI [5] or the Dau-Model, as applied in [6]. Most of these methods correctly predict the perceptual trends for gradual changes of additive noise and nonlinear distortions, respectively. However, for the comparison between the input and output of a nonlinear speech enhancement algorithm, one functional relationship for these different kinds of distortion between subjective perception and objective prediction is indispensable. The application of complex perceptual models could not solve this problem as long as, and this was identified as the key issue, no account was made on the time-invariant nonlinear processing. I.e., speech enhancement is usually associated with an envelope thresholding distortion and merely modifies the low-level and transitional components of speech. Yoo et al. [7] found that these portions hold only 2 % of the energy of the original speech but are almost equally intelligible. In order to compensate for this characteristic, Kates and Arehart [3] developed a coherence based SII (CSI) that includes a time-course weighting based on the short-time RMS level. They showed that short-time sequences of 0 to -10 dB with respect to the overall RMS primarily contribute to speech intelligibility. Recently, Taal et al. [8] advanced with a different approach of an intrusive short-time measure for mask weighted noisy speech. Their model is based on equally contributing transitional intelligibility scores, calculated in windows of 400 ms length. As the analysis windows overlap by half, this particular window-length is in the order of the syllabic rate of speech, which has a maximal modulation transfer at 4 Hz.

In this contribution, it is analyzed whether spectral feature measures can be applied to identify transitional components of speech and in how far these are suitable to weight short-time
intelligibility predictions with respect to the varying, i.e., relative information content. The utilization of spectral features in speech processing is not new. Spectral features are, by way of example, successfully used to improve automatic speech recognition tasks [9, 10]. In the next section four spectral feature measures for the separation of speech into voiced and unvoiced segments are presented. Adapted to the locations of relative high information content, these short-time measures form feature vectors that are applied to the time-course of the short-time CSII method of Kates and Arehart [3]. The successive third section deals with the parameter optimization of the measures to psychoacoustical data and compares the results to state-of-the-art speech intelligibility measures. In the last paragraphs, we discuss the results and draw a conclusion.

2. Algorithm

The algorithmic approach is divided into parts. In the first part, the extraction and adaption of the spectral features is examined. In the second part, the coherence based SI method is presented.

2.1. Extraction of source information with spectral features

The source-filter model is a widely applied method in speech processing to separate the excitation characteristics of the vocal chords from the resonator characteristics of the mouth. In this model, speech is assumed to be short-time stationary in the range of 10 to 30 ms. Accordingly, the model can be formulated as a linear convolution:

\[
x(t) = s(t) * h(t),
\]

where \( s(t) \) is the source and \( h(t) \) is the filter that form the speech signal \( x(t) \) as a function of time \( t \). Linear prediction and cepstrum techniques are often used to separate \( h(t) \) from \( s(t) \). In here, the main interest lies simply on the differentiation whether the glottis produces white noise for unvoiced speech or a periodic stimulation for voiced speech. This feature can be directly calculated from \( x(t) \). In a preparatory study, four spectral measures were identified that correlate with this chance of articulation. These are the Renyi Entropy (RE), the Shannon Entropy (SE), the Spectral Band Energy (SBE) [9] and the Madhu Flatness Measure (MF) [11]. As the transitions between phonemes and formants, i.e., the low-level and transient components of speech, are essential for intelligibility, the first order derivative with respect to the short-time frames was taken to flag these segments of speech. In Fig. 1 the four differentiated spectral feature measures are shown. For comparability, the waveform of the analyzed sentence is plotted in the background (gray shaded). In detail, the measures were calculated from the clean speech waveform that was sampled at 22.05 kHz. An analysis window of 706 samples (32 ms) was Hann weighted and padded with zeros prior to a 1024-point FFT. With a window-overlap of 50 %, a new spectrum was calculated every 16 ms. The STFT spectra were filtered block-wise with center frequencies and bandwidths of the auditory critical bandwidth filters as given in Table I of ANSI S3.5-1997. Prior to the calculation of the entropy measures the sub-band spectra were normalized to obtain the probability mass functions:

\[
\hat{X}_{i,b}(f) = \frac{|X_{i,b}(f)|}{\sum_{f=lb}^{ub}|X_{i,b}(f)|},
\]

where \( X_{i,b}(f) \) is the STFT representation of frame \( i \) that was subdivided in non-overlapping bands \( b \), with an lower frequency bound \( lb \) and an upper frequency bound \( ub \). RE was then calculated with:

\[
RE_{i,b} = \frac{1}{1-\alpha} \log_2 \left( \sum_{f=lb}^{ub} \left| X_{i,b}(f) \right|^\alpha \right).
\]

The order \( \alpha \) was set to 3 throughout this study. SE was calculated with:

\[
SE_{i,b} = -\sum_{f=lb}^{ub} \hat{X}_{i,b}(f) \log_2 \hat{X}_{i,b}(f),
\]

and the MF measure with:

\[
\log_2(MF_{i,b} + 1) = -\frac{1}{\log_2(N_0)} \sum_{f=lb}^{ub} \hat{X}_{i,b}(f) \log_2 \hat{X}_{i,b}(f),
\]

where \( N_0 \) is the amount of frequency bins in frame \( b \). Not a information theoretic measure is the SBE, which displays the relative energy distribution and which was calculated with:

\[
SBE_{i,b} = \frac{\sum_{f=lb}^{ub} \left| X_{i,f} \right|^2}{\sum_{f=lb}^{ub} \left| X_{i,f} \right|^2}.
\]

Thereafter, the results in sub-bands were weighted with the band importance function for average speech (ANSI S3.5-1997) and summed. This resulted in short-time feature vectors. To adapt these to the transitions of high information content in an optimization routine, the feature vectors were either compressed or expanded by an exponent \( p \). Subsequently to the differentiation, a first order recursive smoothing was applied to allow the feature vectors to align with the phonemic transitions. Let
\[
\rho_i \equiv \frac{[\Delta RE^p_i / \Delta i, \Delta SE^p_i / \Delta i, \Delta MF^p_i / \Delta i, \Delta SBE^p_i / \Delta i]}{\Delta i},
\]
the smoothed output was calculated with:
\[
\rho_i = (1 - \alpha) \rho_i + \alpha \rho_{i-1},
\]
where \(\alpha = e^{-\Delta T^*/\tau}\), \(\Delta T^*\) the frame shift and \(\tau\) the time constant. This procedure of contrast enhancement was finalized with a threshold value according to:
\[
\rho_i = \begin{cases} 
\rho_o, & \rho_i > \epsilon \\
0, & \rho_i \leq \epsilon 
\end{cases}.
\]

2.2. The coherence SII

The intrusive CSII predicts the SNR between the input \(x(t)\) and the output \(y(t)\) of a system by calculating the signal power fraction that is linear related. This is possible through the magnitude squared coherence function:
\[
|\gamma(f)|^2 = \frac{\sum_{i=1}^{M} X_i(f) Y^*_i(f)^2}{\sum_{i=1}^{M} |X_i(f)|^2 \sum_{i=1}^{M} |Y_i(f)|^2},
\]
where \(X_i(f)\) and \(Y_i(f)\) are the STFT representation of the input and the output signal, respectively, and the asterix denotes the complex conjugate. Accordingly, the SNR can be calculated in auditory critical bandwidth filters:
\[
SNR_{R_0} = \frac{\sum_{i=0}^{M} W_0(f) |\gamma(f)|^2 \sum_{i=1}^{M} |Y_i(f)|^2}{\sum_{f=0}^{N} W_0(f)(1 - |\gamma(f)|^2) \sum_{i=1}^{M} |Y_i(f)|^2},
\]
in here, \(W_0(f)\) is a matrix of rounded-exponential filters that are described in [3]. To include the weighting with the feature vectors \(\rho_i\), the summations over \(M\) frames in the equations (9) and (10) were replaced with \(\sum_{i=1}^{M} \rho_i\), where \(\rho\) is either the cross-spectral density or the auto-spectral density. After the calculation of the SNRs, in equation (10), the CSII index is conform with the standard ANSI S3.5-1997.

3. Fitting of feature vectors and results

In order to adapt the weighting of the feature vectors to subjective intelligibility scores, a nonlinear optimization was performed. The listening test set comprised five conditions of additive speech shaped noise and five conditions of envelope thresholding, which is considered a strong nonlinear distortion in speech enhancement. Details on the implementation of the envelope thresholding conditions are found in [3]. Eight people with a normal hearing (15 dB HL) participated in a percent correct score test that is based on a semantically unpredictable sentence (SUS) corpus in German [12]. The subjects had to respond to three versions of each condition, which were presented through a small filter coefficient in Table 1. The absence of a clear feature pattern however leads to a high threshold and weighted CSII [3] and an optimized one RMS level CSII were calculated. The optimization of the one RMS level CSII was also subject to a genetic algorithm. In this procedure, two level ranges and their weighting in a logistic function were optimized (in total six parameters). As the combination of two level ranges and their logistic weighting can be expressed as a linear combination, the algorithm converged to one optimal level with an upper dB bound of -4.9 dB and a lower dB bound equal to -7.7 dB (with \(a = 1\) and \(b = 8\)). The results of the optimization are shown in Fig. 2. As figures of merit, the \(r^2\) measure and Kendall’s \(\tau\), a rank statistic, are given. As can be seen, none of the spectral feature measures yields a frame-weighting that improves the correlation with high information content as compared to the existing measures. Only RE shows some improvement over the unweighted CSII at the expense of a higher standard deviation. This result was confirmed by others, who found RE to be beneficial for enhancing the performance in automatic speaker identification tasks [9]. Good results were achieved with the STOI, the RMS three level CSII and the one RMS level optimized CSII. All three measures performed close to the significance level (based on the correlation coefficient, here the square root of the \(r^2\)). To generalize these results, the performance of the measures, especially the one RMS level CSII, has to be confirmed in broader listening tests (the STOI measure already experienced exhaustive testing in [8] and showed a correlation coefficient of 0.95 on comparable data).

4. Discussion

The application of the analyzed spectral measures did not yield an advantage over existing methods in labeling speech sections of high information content. In spite of that, the presented spectral feature measures showed from their first inspection to their adaption in the short-time frame weighting of speech intelligibility a fair degree of conformity. In particular the differentiated SBE, MF and SE measures show a comparable pattern, to strongly respond to changes between voiced and unvoiced speech (cf. Fig. 1). Their sensitivity to these changes is confirmed by low threshold values (\(e < 0.33\)) and small expansion values (\(p < 1.6\)) as found in the optimization. Their first order low pass filter time constants \(\tau\) are in the range of 40 to 70 ms. Apparently, the smoothing corrected consistently for the misalignment of the feature vectors with the transitional parts in speech. Even though no improvement over the unweighted CSII was achieved, the spectral measures proofed to label speech information to a fair extent. This is obvious when considering that envelope thresholding acts gradually on low level regions. If a spectral feature measure would consistently miss these speech sections, speech intelligibility would be strongly underestimated. RE is less sensitive to the changes of voiced and unvoiced speech. Nevertheless, the differentiated RE reveals steep slopes in speech sections of high information content, which is expressed through a small filter coefficient in Table 1.
expansion exponent, which results into a low $r^2$ value due to the increased standard deviation.

In view of the good performance of the level-weighted CSII measure, it can be stated that the observed feature measures are less adequate predictors of the relative speech information content. Since the feature measures were simply extracted from auditory critical filters, it can be questioned whether better results can be achieved by applying these to the source signal or the filter signal alone in (1). If, e.g., entropy measures are calculated from the slow changing transfer function of the vocal tract, more local stability and a better identification of transitional speech parts might be yielded [10].

In here entropy was only used to discern between voiced and unvoiced passages. In [13] Leijon refers to the observation that the acoustic speech information rate and the performance of speech perception are coupled. He further showed that there is no direct relation between information rate in frequency bands and the empirical additivity-concept of audible contributions in different frequency bands of the AI theory. Taken together, these are only a few examples to relate to and to integrate the concepts of information theory in speech processing.

5. Conclusion

This contribution deals with the objective assessment of linearly and nonlinearly processed speech. It was analyzed whether spectral feature measures can be utilized to label the relative information content in short-time frames of speech and to establish a weighting based on their contribution to an overall speech intelligibility measure. Although the analyzed spectral feature measures are capable of labeling transitional sections, which are important for speech intelligibility, they do not represent an alternative to the existing level-based weighting methods for short-time intelligibility measures.

6. Acknowledgements

The authors like to thank Juan-Pablo Ramirez, who conducted the listening test at the Telekom Quality and Usability Lab. in Berlin.

7. References