Mutual Information Approach for Predicting Speech Intelligibility

Ranjitha A[1] Ramyashree M[2]
[1]M. Tech Student [2]Assistant Professor
[1,2]Department of DECS
[1,2]VTU PG Centre, Mysore-570019

Abstract— The proposed intelligibility prediction model makes use of fundamental principle theoretic tools like entropy and mutual information. It emerges as natural to employ tools developed to characterize data or information transmission. Subsequent to all, the speech communication process be able to view as the method of transmitting a speech signal across a dynamic, transient channel (acoustic channel, auditory periphery and higher stages of the auditory pathway) to arrive at the brain of the receiver. The expression for the AI demonstrates well-built resemblance to the expression for the ability of a memory less Gaussian channel. The fundamental thought of the proposed technique is to compare the critical band amplitude envelopes of the clean and noisy or processed signal for estimating the intelligibility of the noisy or processed signal.

Key words: Intelligibility, prediction entropy and mutual information

I. Introduction

To predict the intelligibility of speech signal in presence of background noise a number of measures have been proposed, thus intelligibility can be calculated effectively. Among these measures, today the most commonly used techniques to predict the speech intelligibility in presence of background noise include Articulation Index (AI) and the Speech Transmission Index (STI). This AI method was advanced to produce Speech intelligibility index (SII). This SII method is based on the idea that intelligibility of speech signal will depend on the quantity of the spectral information that is audible to the listeners and thus the intelligibility is computed by dividing the spectrum into 20 bands thus the intelligibility is equally contributed and the weighted average of the signal-to-noise ratio is estimated in each frequency band. The signal-to-noise ratio in each band is prejudiced by the band importance functions (BIF) that differ across several speech resources. The SII measures successfully calculate the effects of additive noise and linear filtering on intelligibility of speech signal.

Both speech quality and intelligibility are measured through listening tests, these tests are very slow and also expensive, thus there is a great demand for objective measures that correlate well with subjective performance. Traditional method for measuring the intelligibility was Signal-to-noise ratio, which measures the squared differences between the original, pre-corruption speech, and the output of the improvement stages. Thus a high SNR will give a speech that is similar to the original sound speech signal and achieves high quality and thus the intelligibility is high. However, for complex nonlinear speech enhancement schemes, SNR is a very poor predictor of intelligibility.

Intelligibility is a gauge of how understandable the speech is, or the degree to which a speech can be able to understand. Intelligibility is affected by verbal clearness, explicitness, fluency, clarity, self-expression, and accuracy. For appropriate communication, the average speech level must exceed that of an intrusive noise by 6dB; lower sound to noise ratios are seldom acceptable. Manifesting in a broad frequency range, speech is fairly resistant to many kinds of masking frequency cut-off, for instance, that a band of frequencies range from 1000 Hz to 2000 Hz is adequate (sentence articulation score is around 90%).

In the present work, the speech intelligibility prediction method aim to calculate the average intelligibility of noisy and processed signals, judged by group of listeners. The main motivation for studying this speech intelligibility predictors are of two reasons,

1) Replacement for the costly listening test that may occur during the early stages of development, thus there is a great demand for the intelligibility predictors that are reliable.
2) Developing and studying of these intelligibility predictors will guide to a better understanding mechanism behind human intelligibility capability.

The present work predicts the average intelligibility of noisy signals and the processed signals. A model that performs this prediction is based on the hypothesis that intelligibility can be monotonically related to the mutual information between the critical band amplitude envelope of clean speech signal and the corresponding noisy speech signal. Thus the intelligibility predictor will be a simple function of mean square error (mse), arising while estimating the clean critical band amplitude envelopes using a minimum mean square error estimator (mmse) on the basis of noisy amplitude.

II. Simulation tool

The simulation tool used in present work is the Matlab 7.10 which was released on March 2010 for Intel 32 bit. MATLAB known as matrix laboratory is a multi paradigm numerically computing environment and an fourth-generation programming language. Matlab was developed by Math Works, Matlab allows matrix operations, scheming of functions and data, achievement of algorithms, formation of user interfaces, and interfacing with programs written in other languages, including C, C++, Java,Fortran and Python. Even though MATLAB is primarily intended for numerical computing, an optional toolbox uses the MuPAD symbolic engine, allowing access to symbolic computing capabilities. An additional package called the Simulink is present, that adds graphical multi-domain simulation and Model-Based Design for embedded systems and dynamic systems.

MATLAB is extensively used in academic and research institutions as well as industrial enterprises.An easy way to comply with the conference paper formatting requirements is to use this document as a template and simply type your text into it.

III. METHODOLOGY

The basic auditory model is an auditory model i.e. the auditory system, which is a sensory system for the sense of hearing, this includes auditory part of sensory system (Brain) and sensory organ (ear).

The mutual information lower bound is derived in the second section. This is necessary because in certain simple situations the conditional differential entropy h(P/Q) can be analytically derived from the given joint probability density function $f_{P,Q}$(p,q). However in general case, where the processing leading to Q is complicated or even unknown, it is difficult to compute conditional differential entropy h(P/Q) from the limited data of joint probability density function $f_{P,Q}$(p,q). as a substitute, to evade this difficulty, lower bound on mutual information is proposed. In the next section implementation details is discussed, where signals are sampled to a sampling frequency of 10KHz, thus the frequency region relevant for speech intelligibility is covered. The signals are divided into the frames of length N = 256 samples. and hann analysis window is applied. DFT order of N = 256 samples is used and the resultant DFT coefficient are grouped to perform one third octave band analysis. In the successive section results are obtained by simulating the codes in MATLAB 7.10. The plots of clean speech and noisy speech (both additive noise and low pass filter) are compared to get the mutual information between clean and noisy speech followed by conclusion.

The time domain input signal are divide into successive signal, thus obtaining the band pass filtered signals. These signals are overlapped using analysis frames and an analysis window is applied to these frames thus the resulting frames are transformed from time domain to frequency domain using an discrete Fourier transform (DFT).

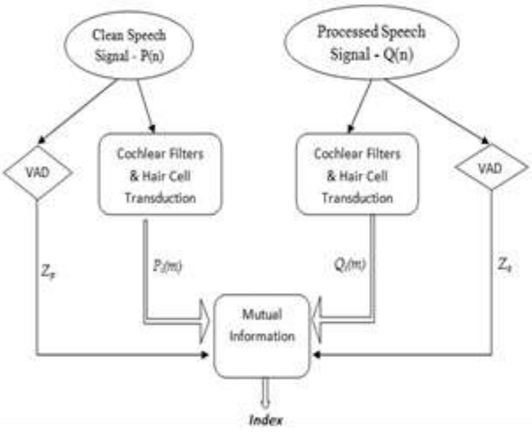
Thus the resulting DFT coefficients arte given by,

$$\tilde{P}(k,m) = \sum_{n=0}^{N-1} P(mD+n)\,\omega(n)\,e^{-j2\pi kn/N}$$

$$\tilde{Q}(k,m) = \sum_{n=0}^{N-1} Q(mD+n)\,\omega(n)\,e^{-j2\pi kn/N}$$

Where,

m = Frame index
k = frequency bin index
D = Frame shifts in samples
N = Frame length in samples
ω(n) = Analysis window


Clean Speech Signal - P(n)
Processed Speech Signal - Q(n)
VAD
Cochlear Filters & Hair Cell Transduction
Cochlear Filters & Hair Cell Transduction
VAD
Z_s
$P_i(m)$
$Q_i(m)$
Z_x
Mutual Information
Index

Now these DFT bins are grouped in order to perform one third octave band analysis. This results in critical band amplitude, which is given by

$$P_i(m) = \sqrt{\Sigma_{K\epsilon CB_i} |\hat{P}(K,m)|^2} \quad \text{---------} \ (1)$$

$$Q_i(m) = \sqrt{\sum_{K\epsilon CB_i} |\hat{Q}(K,m)|^2}$$

Where,
CB_i = Frequency index set of ith one third octave band,
i = 1, ..., L

The voice activity detection (VAD) blocks are used to identify and exclude the frames that have low energy from the computation. As signal P(n) contains low energy frames that is the silence region which do not contribute to the speech intelligibility, it can be excluded from computing the mutual information. Thus by applying an simple energy based per frame VAD to P(n) a frame index set Z_s of the active frame is resulted. Similarly the VAD which is present in the lower branch identifies the high and low energy frames in the noisy speech signal X(n).

The low energy frames are the noise or the silence frames only which may occur due to certain types of aggressive processing thus suppressing the speech signal frames that carry the information. These high energy frames are represented by an frame index set Z_x.

To calculate the mutual information, denote the number of frames in a given speech sentence as M and

$$Q = [Q_1(1)\, Q_2(1) \ldots\ldots\ldots Q_L(1)\, Q_1(2) \ldots\ldots\ldots Q_L(M)]^T$$

$$P = [S_1(1)\, S_2(1) \ldots\ldots\ldots S_L(1)\, S_1(2) \ldots\ldots\ldots S_L(M)]^T$$

Be the super random vectors respectively, which are formed by stacking the critical band amplitude spectra of successive frames.

The mutual information between the clean and noisy critical band amplitude is given by

$$\frac{1}{L|Z_p|}\, I\,(P;Q)$$

Where,
Z_p = Set cardinality.
$L|Z_p|$ = Estimates number of speech active critical band amplitude in a clean speech signal.

By assuming that the entries in each super vector are statistically independent, the mutual information

$$I(P;Q) = \frac{1}{L|Z_p|} \sum_{m} \sum_{i=1}^{L} I\{P_i(m), Q_i(m)\}$$

$$I(P;Q) = \frac{1}{L|Z_p|} \sum_{m\epsilon Z_P \cap Z_Q} \sum_{i=1}^{L} I\{P_i(m), Q_i(m)\}$$

As both the signals P(n) and Q(n) are active speech signals, the second equation follows over the frame index set $m\epsilon Z_P \Omega Z_Q$ and excludes I (P_i (m) ; Q_i (m)) terms as it is zero. For the convenience let us, simply replace P_i(m) and Q_i(m) by P and Q. Therefore, the mutual information I (P;Q) between the clean and the noisy critical band amplitudes is given by,

$$I(P;Q) = h(P) - h(P/Q)$$

Where,
h(P) = Differential entropy.
h(P/Q) = Conditional Differential entropy.

The differential entropy and the conditional differential entropy is given by

$$h(P) = \int_{P} f_P(p) \ln f_P(p) dp$$

$$h(P/Q) = \int_Q \int_p f_{P,Q}(P,Q) \ln f_{P/Q}(p/q) dpdq$$

---------- (2)

In certain simple situations it is easy to analytically derive the conditional differential entropy h(P/Q) from the given joint probability density function (pdf) $f_{P,Q}$(p,q). However in some general situation, since the exact processing may be unknown or leads to X can be complicated, estimating from limited statistics the joint pdf $f_{P,Q}$(p,q) required to calculate h(P/Q) is difficult. Thus, to avoid this complexity, the present work propose to lower bound the mutual information I(P;Q) and this requires only 2nd order statistics of $f_{P,Q}$(p,q).

The lower bound on mutual information I(P,Q) is resulting by upper bounding the conditional entropy h(P/Q). The conditional $\mu_{P/q}$ is equal to the mean square error (mse) estimator $\hat{p}_{mse}(q)$ of the clean arbitrary variable P on observing the noisy processed comprehension q.

$$I_{LB,lmmse}(P;Q) \leq I_{LB,mmse}(P;Q) \leq I(P;Q)$$

These lower bound is a function of entropy h(P) of a critical band amplitude of a clean speech signal.

Deriving an expression for differential entropy by considering the frame size N is large when compare to the correlation time of the clean speech signal p(n), then the imaginary and real parts of DFT coefficient $\tilde{P}(k,m)$ can be considered independent and thus can be modeled as zero mean Gaussian variables. Further assuming that the DFT coefficients within the same critical band $\tilde{P}(k,m)$; $k \epsilon CB_i$ are identically distributed such that $P_i(m)$ given in equation 1 is a scale chi distributed random variable with $k' = 2|CB_i|$ degrees of freedom.

When the real and imaginary parts of $\tilde{P}(k,m)$, $k \epsilon CB_i$ are zero mean, unit variance Gaussians, then the corresponding critical band amplitude, Z has an expected value of

$$E(Z) = \sqrt{2}\frac{\Gamma((k'+1)/2\}}{\Gamma(k'/2)}$$

Variance is given by,

$$\sigma_Z^2 = k' - E(Z)^2$$

and a differential entropy given by,

$$h(Z) = ln\Gamma(k'/2) + \frac{1}{2}(k' - ln2 - (k'-1)\psi(k'/2))$$

Where,

$\Gamma(k'/2)$= gamma function.

$\psi(k'/2)$=Digamma function.

If the real and imaginary parts of $\tilde{P}(k,m)$, $k \epsilon CB_i$ are not unit variance then the differential entropy of the corresponding critical band amplitude is given by,

$$h(P) = h(z) - \frac{1}{2} ln\sigma_Z^2 + \frac{1}{2} ln\sigma_P^2$$

---------- (3)

we know that, for any random variable *Y* and a constant *c*, the differential entropy is given by,

$$h(cY) = h(Y) + ln|c|$$

Since the first two terms in equation (3) are the functions only of number of degrees of freedom k', the differential entropy h(P) is a plain function of the variance σ_P^2 of the critical band amplitudes, and thus it can be computed offline.

Thus the mutual information is given by,

$$I(P;Q) = max\left\{h(Z) - \frac{1}{2}ln\sigma_Z^2 - \frac{1}{2}ln2\pi e + \frac{1}{2}ln\frac{\sigma_P^2}{D_{lmmse}}, 0\right\}$$

Where,

I(P;Q) = Mutual information between clean and noisy speech,

h(Z) = Differential entropy,

σ_Z^2 = Variance of critical band amplitude (Z),

σ_P^2 = Unity,

D_{lmmse} = Linear minimum mean square error.

A. Simulation Details

Here the signals are sampled to 10KHz sampling frequency, thus ensuring that the region of frequency relevant for speech intelligibility is covered. These signals are divided into frames with a length of N=256 samples and finally an hann analysis window ω(n) is applied. A frame shift of D= N/2 = 256/2 =128 samples is used.

A DFT of order N=256 is used and the resultant DFT coefficients are grouped into a total of L=15 third order octave band which has the centre frequency of highest band set of 4.3KHz and a centre frequency of lowest band set of 150 Hz.

As there in block diagram voice activity detection is used to recognize and exclude the energy frames with a condition of energy not less than $\Delta_E = 30$ dB of the signal frame with maximum energy are identified.

The proposed model uses basic three parameters such as, α, Δ_E and I_{max} are as follows,

α = 0.95, Δ_E = 30 dB and I_{max}= 0.2 nats*.

*1 nat = $\frac{1}{ln2} = 1.44\ bits$

Corresponding to a time constant of 250 milliseconds (ms), the value of α is taken as 0.95. Often used time span in speech processing applications are between 20-40 ms. According to one of the latest literature, across the time span of 400ms the statistics was computed for the STOI model and it was recommended that this time span can be connected to sequential integration processes taking place in the auditory system. The value of Δ_E taken as 30 dB should not be a controversial, as most of the speech frames present in the clean speech signal have an energy content larger than the selected threshold. The value of I_{max} taken approximately as 0.2 nats in the present model.

IV. RESULTS AND DISCUSSIONS

The results are obtained by simulating the codes in MATLAB 7.10. The mutual information is successfully calculated between clean speech signal with additive noise and low pass filter and also the mean square error is taken into account as a measure of purity. The plots of clean speech and noisy speech (both additive noise and low pass filter) are compared to get the mutual information between clean and noisy speech.

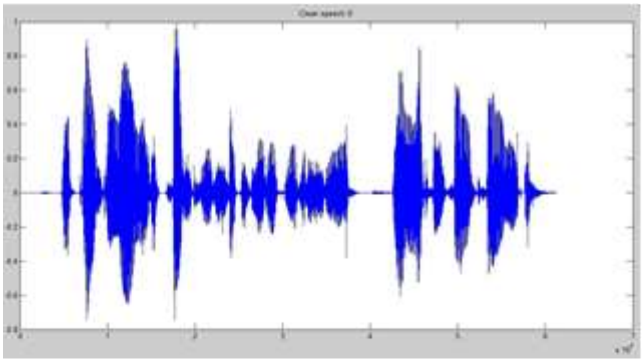


Fig. 1: Depicts the plot of clean speech signal.

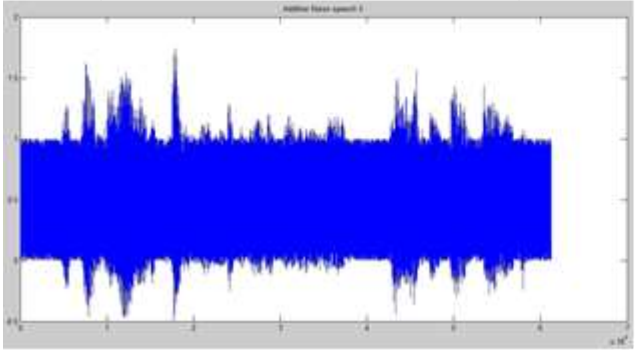


Fig. 2: Depicts the plot of speech signal processed with additive noise.

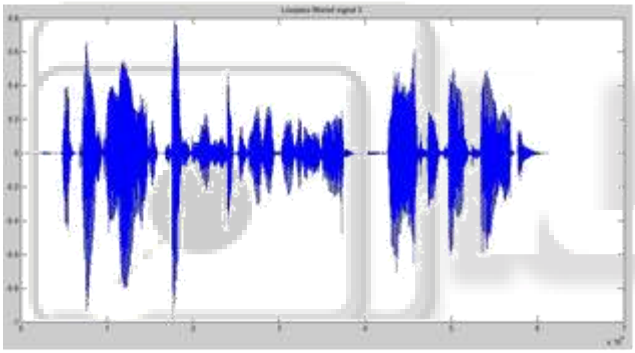


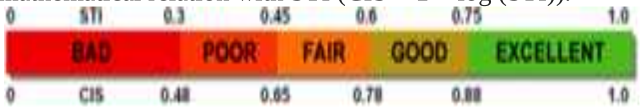
Fig. 3: Depicts the plot of speech signal processed with additive noise.

The simulation results for the mutual information and mean square error between clean speech signal with additive noise and low pass filter are tabulated in table 1.

Sl. No.	Clean speech signal processed with	Mutual Information	Mean Square Error
1	Additive noise	0.08374055	0
2	Low pass filter	0.19411051	0

Table 1: Depicts Mutual Information And Mean Square Error Between Clean Speech Signal With Additive Noise And Low Pass Filter

The Common Intelligibility Scale (CIS), based on a mathematical relation with STI (CIS = 1 + log (STI)).



Speech Intelligibility may be expressed by a single number value. Two scales are most commonly used: STI and CIS Basically, an SII of 0 implies that none of the speech information, in a given setting, is available (audible and/or usable) to improve speech understanding. An SII of 1.0 implies that all the speech information in a given setting is both audible and usable for a listener. there is a monotonic relationship between the SII and speech understanding. That is, as the SII increases, speech understanding generally increases.

V. Conclusion

The mutual information is successfully calculated between clean speech signal with additive noise and low pass filter. In the present work the clean speech is processed by considering the additive noise and low pass filter but not necessarily stationary noise sources, and a non linear processing is considered, thus it can be generally referred to as Time Frequency weighting. Algorithm needed for estimating the outcome of intelligibility listening test are of great importance in order to reduce the number of costly listening test that may occur during the early stages of algorithm development and also have the potential to lead into new auditory system. The processing considered is quite broader and hence it can be used in single channel noise reduction algorithm. The present work follows the theory that, it can be monotonically related to Shannon information about a clean critical band amplitude envelopes and upon observing, the noisy processed counterparts can be learnt. Then the lower bound for this mutual information is derived that can be traced analytically. Thus the information lower bound can be computed as a function of MMSE that arise by estimating the critical band amplitude of a clean speech signal from its noisy counterparts. Using an MSE estimator will lead to highest speech quality.

References

[1] J.Jensen and R.Hendriks, Jan. “Spectral magnitude minimum mean square error estimation using binary and continuous gain functions,” IEEE Trans. Audio, Speech, Lang. Process., vol. 20, no. 1, pp. 92–102, 2012.

[2] J.Koop man, R.Houben, W.A.Dreschler and J.Verschuure, “Development of a speech in noise test (matrix),” in Proc. 8th EFAS Congr., 10th DGA Congr., Heidelberg, Germany, Jun. 2013

[3] C. H. Taal, R. C. Hendriks, R. Heusdens, J. Jensen, and U. Kjems, “An evaluation of objective quality measures for speech itelligibility prediction, ”in Proc. Inter speech., Brighton, U.K., Sep.6-10, 2013

[4] J. B. Boldt and D. P. W. Ellis, “A simple correlation based model of intelligibility for non linear speech enhancement and separation,” in Proc. 17th Eur. Signal Process. Conf.(EUSIPCO). 2014

[5] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, “An evalu- ation of objective measures for intelligibility prediction of time-fre- quency weighted noisy speech,” J. Acoust. Soc. Amer., vol. 130, pp. 30.

[6] Jesper B. Boldt and Daniel P. W. Ellis,"A Simple Correlation-Based Model Of Intelligibility For Nonlinear Speech Enhancement And Separation", submitted to EUSIPCO-2009.

[7] Nasser Mohammadiha, Paris Smaragdis, Arne Leijon, “Supervised and Unsupervised Speech Enhancement Using Nonnegative Matrix Factorization”. IEEE transactions on audio, speech, and language processing, 2013.

[8] Jesper Jensen, Cees H. Taal, “Speech Intelligibility Prediction Based on Mutual Information”, IEEE/acm transactions on audio, speech, and language processing, vol. 22, no. 2, February 2014.

[9] J. Ma, Y. Hu, and P. Loizou, “Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions,” J. Acoust. Soc. Amer., vol. 125, no. 5, pp. 3387–3405, 2009.

[10] Cees H. Taal, Richard C. Hendriks, Richard Heusdens, and Jesper Jensen, “An Algorithm for Intelligibility Prediction of Time–Frequency Weighted Noisy Speech”, IEEE transactions on audio, speech, and language processing, vol. 19, no. 7, September 2011

[11] Jalal Taghia, Rainer Martin Richard C. Hendriks, “On Mutual Information As A Measure Of Speech Intelligibility”, ICASSP 2012 IEEE 978-1-4673- 2012.

[12] http://en.wikipedia.org/wiki/Intelligibility_(communication).

[13] https://www.meyersound.com/support/papers/speech/section2.htm.