ON MUTUAL INFORMATION AS A MEASURE OF SPEECH INTELLIGIBILITY

*Jalal Taghia, Rainer Martin**

Ruhr-Universität Bochum, Germany
{jalal.taghia,rainer.martin}@rub.de

Richard C. Hendriks

Delft University of Technology,
The Netherlands
{R.C.Hendriks}@tudelft.nl

ABSTRACT

Speech intelligibility prediction of noisy and processed noisy speech is important in a number of application domains such as hearing instruments and forensics. Most available objective intelligibility measures employ either a signal-to-noise ratio (SNR)-based or correlation-based comparison between frequency bands of the clean and the processed speech. In this paper, we approach the speech intelligibility prediction from the angle of information theory and show that an information theoretic concept provides a unified viewpoint on both the SNR and the correlation based approaches. Two objective intelligibility measures are introduced based on estimated mutual information between the clean speech and the processed speech in the time and the frequency subband domain. Our proposed measures show high correlation with subjective intelligibility measure (i.e. word correct scores) and comparative results with the short-term objective intelligibility measure (STOI).

Index Terms— Intelligibility prediction, mutual information, speech enhancement

1. INTRODUCTION

Speech enhancement algorithms often aim at improving both quality (e.g. speech pleasantness and naturalness) and intelligibility of noisy speech. Although the improvement of speech intelligibility is a difficult task, recent progress in this field has also triggered new interest in the instrumental evaluation of speech intelligibility. Most available objective intelligibility measures, which have been published (e.g. [1, 2, 3]), employ either a signal-to-noise ratio (SNR)-based or correlation-based comparison between frequency bands of the clean and the processed speech. Promising results for the speech intelligibility prediction were reported for measure based on Dau auditory model (DAU) [4] (examined in [5]), the coherence speech-intelligibility index (CSII) [1], the normalized subband envelope correlation (NSEC) [6], the frequency-weighted segmental SNR (FWS) [7], the normalized covariance based STI (NCSTI) [1], the measures based on computing SNR loss incurred in critical bands [2], and the short-term objective intelligibility measure (STOI) [3, 8]. Among the existing objective intelligibility measures, it has been shown in [3] that STOI is one of the most promising candidates to predict speech intelligibility for data processed by time-frequency (TF) varying gain functions. STOI measure is based on correlation coefficient between the temporal envelopes of the clean and the processed speech per frequency band. To derive the correlation coefficient, temporal envelopes of the processed speech in frequency bands are further normalized and clipped with a clipping factor. Besides the clipping process, the difference between STOI and NCSTI is that in STOI the correlation coefficient is computed for short-time segments and not for the whole signal at once. In the paper, we compare the results derived for our proposed objective intelligibility measures with STOI.

In this paper, speech intelligibility prediction is investigated from the viewpoint of information theory, and novel objective speech intelligibility measures are introduced. The proposed measures are based on mutual information (MI) [9]. We show how the proposed measures can be used to predict speech intelligibility of noisy speech and processed noisy speech delivered by single-channel noise reduction approaches. Experimental results demonstrate that our proposed objective intelligibility measures have high correlation with intelligibility listening tests. Furthermore, for the special case of Gaussian random variables we show that the MI measure depends solely on the correlation which is in turn fully determined by the SNR. Thus MI provides a unified viewpoint on existing measures.

The remainder of the paper is organized as follows: In Section 2 mutual information and the estimation procedure of mutual information is briefly described. Section 3 outlines the proposed intelligibility measures. The experimental framework is described in Section 4 and followed by a discussion of experimental results and conclusions in Section 5.

2. MUTUAL INFORMATION AND THE ESTIMATION PROCEDURE

Mutual information (MI) is a general measure of the dependence between two random variables and shows the quantity of information one has obtained on random variable X by observing random variable Y. Mutual information between two random variables X and Y on discrete spaces can be defined as

$$I(X;\,Y) = \sum_{x,\,y} p_{XY}(x,y)\,\log_2\left(\frac{p_{XY}(x,y)}{p_X(x)\,p_Y(y)}\right), \qquad (1)$$

where $p_X(x)$ and $p_Y(y)$ are the marginal probability density functions of random variables X and Y, and $p_{XY}(x,y)$ is the joint probability density function. Since the base of the logarithm is 2, the MI is measured in bits. MI can be defined by the Shannon entropy of random variables as well. Shannon entropy of a random variable $H(X)$ expresses the degree of information that the observation of random variable X provides, and it is defined as

$$H(X) = -\sum_{x} p_X(x)\,\log_2(p_X(x)). \qquad (2)$$

Then, the definition of mutual information can be rewritten as

$$I(X;\,Y) = H(X) - H(X|Y), \qquad (3)$$

*This work was funded by the European Commission within the Marie Curie ITN AUDIS, grant PITNGA-2008-214699.

where $H(X|Y)$ is the conditional entropy of random variable X given random variable Y which is defined as

$$H(X|Y) = -\sum_{x,y} p_{XY}(x,y)\ \log_2(\,p_{X|Y}(x|y)). \tag{4}$$

The conditional entropy can be written as

$$H(X|Y) = H(X,Y) - H(Y), \tag{5}$$

where $H(X,Y)$ is the joint entropy. The mutual information is always greater than or equal to zero, with equality if X and Y are independent. In other words the higher the mutual information, the stronger the dependency between X and Y. To gain an insight in mutual information and its properties consider, for instance, two normally distributed random variables X and N with realizations x and n. It is assumed that X and N are independent, and may be added to generate random variable $Y = X + N$. In our experiment, we consider random variables X , N, and Y as clean signal, noise and noisy signal respectively. The distribution of these random variables is specified as $X \sim \mathrm{N}(\mu_x,\, \sigma_x^2)$ (i.e. $p_X(x) = \frac{1}{\sqrt{2\pi\sigma_x^2}} e^{-\frac{(x-\mu_x)^2}{2\,\sigma_x^2}}$), $N \sim \mathrm{N}(\mu_n,\, \sigma_n^2)$, and $Y \sim \mathrm{N}(\mu_x + \mu_n,\, \sigma_x^2 + \sigma_n^2)$. By assuming the random variables to be zero mean (i.e. $\mu_x = \mu_n = 0$), it can be proved that [9]

$$H(X) = \frac{1}{2}\ \log_2(2\pi e\sigma_x^2), \tag{6}$$

and

$$H(Y) = \frac{1}{2}\ \log_2\left[2\pi e\sigma_y^2\right]. \tag{7}$$

The joint entropy between the two random variables X and Y is similarly derived as

$$H(X,\,Y) = \frac{1}{2}\ \log_2\ \left[(2\pi e)^2\sigma_x^2\sigma_y^2(1-\rho^2)\right], \tag{8}$$

where $\sigma_y^2 = \sigma_x^2 + \sigma_n^2$, and $\rho = \frac{COV(X,Y)}{\sqrt{\sigma_x^2\,\sigma_y^2}}$ the correlation coefficient between X and Y. In our case, since random variables X and N are independent $COV(X,\,Y) = \sigma_x^2$ and therefore we have,

$$H(X,\,Y) = \frac{1}{2}\ \log_2\ \left[(2\pi e)^2\sigma_x^2\,\sigma_n^2\right]. \tag{9}$$

Finally, mutual information of X and Y is computed as

$$I(X;\,Y) = -\frac{1}{2}\ \log_2\,(1-\rho^2), \tag{10}$$

where $\rho = \sqrt{\frac{\sigma_x^2}{\sigma_y^2}} = \sqrt{\frac{\sigma_x^2/\sigma_n^2}{1+(\sigma_x^2/\sigma_n^2)}}$. Thus, under the Gaussian assumption the mutual information depends on the SNR only. The (true) mutual information, the correlation coefficient, the conditional entropy, and the joint and marginal entropies are derived for a range of low to high SNRs and shown in Fig.1.

Estimating mutual information is a challenging task since probability distributions are not given in practice. There are many parametric and non-parametric estimation approaches in the literature (a survey on different approaches for the estimation of MI can be found in [10]). Non-parametric estimation approaches are statistical methods which do not need to assume that data is from a known distribution. The k-nearest neighbor (KNN) estimator [11] is such a non-parametric estimation approach which is used here for deriving our proposed measures. It has been concluded in [10] that KNN estimator performs better than the kernel density and histogram based

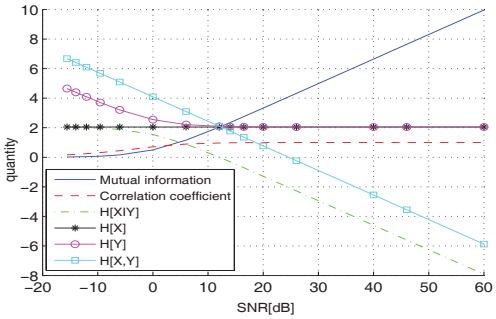

Fig. 1. True mutual information, the correlation coefficient, the conditional entropy, the joint and marginal entropies.

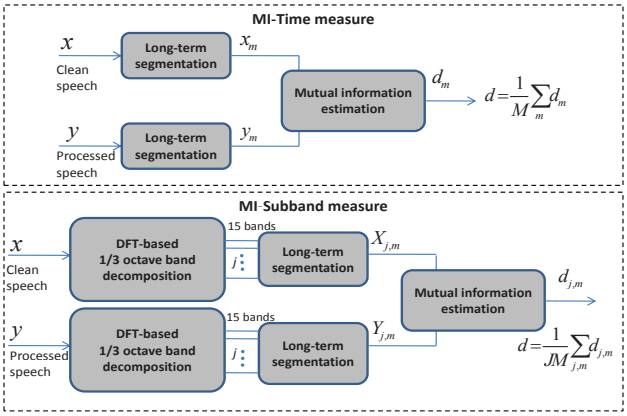

Fig. 2. Proposed procedures for deriving objective intelligibility measures: MI-Time measure (on top) is the proposed procedure for the first objective measure in which the mutual information is estimated in time domain. MI-Subband measure is the proposed procedure for the second objective measure in which the mutual information is estimated in subband domain.

methods for estimating MI. Assuming $z_i = (x_i,\, y_i),\ i = 1, \ldots, L$ are bivariate measurements from realizations of a random variable $Z = (X,\, Y)$ with joint density $p_{XY}(x,y)$, MI is estimated by calculating Shannon entropies based on Kozachenko-Leonenko estimate method in which the probability distribution for the distance between a realization of a random variable and its k-th nearest neighbor is computed [11]. Entropy estimation is achieved for a random sample of L realizations for random variables X and Y. Therefore, parameters involved in the estimation process are L and k. It has been shown in [11] that entropy estimates are data-efficient and systematic error is scaled as a function of k/L. However, the choice for k in KNN estimator is important and has an impact on the final result.

3. PROPOSED INTELLIGIBILITY MEASURES

Here, we introduce two objective measures which both require the clean and the processed speech, respectively denoted by x and y, to provide the speech intelligibility score. Our two proposed measures are completely based on the estimation of mutual information and their difference is in the domain in which MI is estimated. As we have shown in Fig.2, the first proposed objective measure (called MI-Time measure) is derived by segmenting the input signals (the

clean and the processed speech) into segments of sufficient length (i.e. being long enough for providing estimated mutual information with low systematic error), computing mutual information between the segments of the clean and the processed speech and averaging estimated values over the whole number of segments. For the second objective measure (called MI-Subband measure), mutual information is estimated in the subband domain. The procedure of transforming the input signals into the subband domain is similar to what the STOI measure uses [3]. Here also the model is designed for 10 kHz sampling rate so that input signals in any other sampling rate should be resampled. By employing the discrete Fourier-transform (DFT) based one-third octave decomposition, both the clean and the processed speech signals are decomposed into 15 subband signals. The lowest and the highest center frequencies of subbands are 150 Hz and 4.3 kHz. Long term segmentation is performed on the subband signals followed by applying the MI estimator to the clean and the processed speech in corresponding subbands and segments. The final intelligibility score is the average of estimated mutual information values over all segments and all subbands. The long term segmentation in the procedures of the proposed measures is motivated by the need for decreasing the cost of computational complexity of MI estimation. Moreover, notice that no normalization and clipping processes are employed in the two proposed procedures.

4. EXPERIMENTAL RESULTS

In this section we evaluate the performance of the proposed measures and compare the result with STOI. The experimental settings and the used database are introduced, followed by a description of the evaluation procedure and the results.

4.1. Experimental settings

The database which is used in the intelligibility experiments is the same as in [3]. The clean speech originates from the Dantale II database [12], containing 30 randomly chosen Danish spoken sentences each consisting of five words originating from the same Danish female speaker. The clean signal has a total duration of 65 seconds and is sampled at 20 kHz. The clean speech is corrupted with speech shaped noise at 5 different SNR levels -8.9, -7.7, -6.5, -5.2 and -3.1 dB. In addition to unprocessed speech (denoted by UN), there are two sets of processed signals both derived by applying single-channel noise reduction approaches; one is the Ephraim Malah algorithm (denoted by EM) [13] and the other is an algorithm based on a super-Gaussian speech model (denoted by SG) [14]. Therefore, in total, 15 conditions are provided. The reference subjective evaluation measure in our experiments is the word correct score (WC). To derive it, 15 normal hearing listeners were employed in the listening experiments and the average of their results is considered as the word correct score. For the comparison, the STOI measure is used and the settings of this measure is exactly as in [3]. The optimal choice for the k-nearest neighbor parameter of the KNN estimator can not be found theoretically. In general, it is dependent on the data set and needs to be determined in experiments. Based on initial experiments, we choose $k = 300$ in the estimation process of MI for both the proposed objective measures. All the data is processed in one segment.

4.2. Evaluation procedure and the results

In order to examine the objective measures using performance measures, mapping the values of objective measures to the word correct scores is necessary. By using such a mapping process, first we can explain the nonlinear relation between the values of objective measures and the subjective intelligibility scores. Subsequently, we can linearize this relation by applying the mapping functions. Linearizing the relation is important since some performance measures like correlation coefficient can be used properly. The function which we use for the mapping process is as follows,

$$S_i = \frac{1}{1 + e^{a \log(d_i) + b}}, \tag{11}$$

where S_i denotes the value of word correct score in the i_{th} condition, d_i denotes objective measure value which has to be mapped, and a, b are parameters which have to be found. For the implementation in Matlab, the built in function "lsqcurvefit" can be used which solves nonlinear curve-fitting problems in the least-squares sense. This function is used for all objective measures. The parameters of mapping function are derived separately for each objective measure. We use only the data from unprocessed speech conditions (i.e. UN data) to obtain the required parameters of the mapping function, and then the mapping function is applied to the rest of data available in two other conditions (i.e. EM and SG data). In the evaluation part, data points related to unprocessed speech conditions are excluded since they were already used to derive the mapping functions. In Fig. 3 we show the mapping functions derived for objective intelligibility measures MI-Subband, MI-Time, and STOI. The scatter plots between objective intelligibility measures and word correct scores before and after applying mapping function are presented in Fig. 4.

Two performance measures are used for evaluating and comparing the performance of objective intelligibility measures i.e. MI-Subband, MI-Time, and STOI. One is root mean squared error (RMSE) and the other is normalized correlation coefficient (NCC) which are defined as,

$$RMSE = \sqrt{\frac{1}{I} \sum_i (S_i - D_i)^2}, \tag{12}$$

$$NCC = \frac{\sum_i (S_i - \bar{S})(D_i - \bar{D})}{\sqrt{\sum_i (S_i - \bar{S})^2 \sum_i (D_i - \bar{D})^2}}, \tag{13}$$

where I is the number of data points, S_i the word correct score value for the i_{th} condition, D_i the mapped objective measure for the i_{th} condition, $\bar{S}$ the averaged value for word correct scores over all conditions, and $\bar{D}$ is the averaged value for objective measure values over all conditions. The performance of the proposed objective measures and STOI in terms of RMSE and NCC is shown in Fig.5.

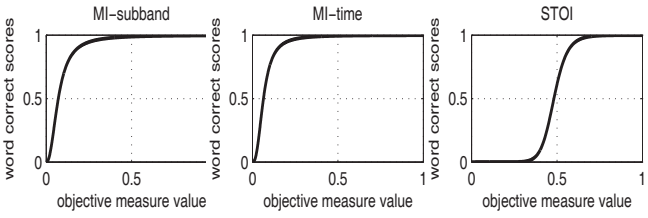

Fig. 3. Mapping functions for objective intelligibility measures MI-Subband, MI-Time, and STOI respectively from left to right.

5. DISCUSSION AND CONCLUSIONS

In this paper, we introduced MI as an objective measure of speech intelligibility and show that this information theoretic concept provides a unified viewpoint. Two objective intelligibility measures are

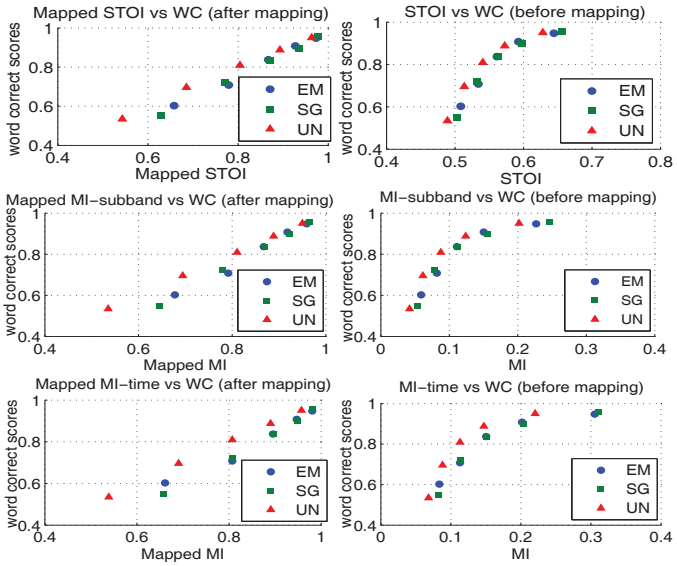

Fig. 4. The scatter plots between objective intelligibility measures and word correct scores before and after applying mapping functions. From top to bottom the results for STOI, MI-Subband, MI-Time are presented respectively. Data points labeled by UN, SG, and EM denote values relevant to unprocessed speech, noisy speech processed by super-Gaussian based algorithm, and noisy speech processed by Ephraim Malah algorithm.

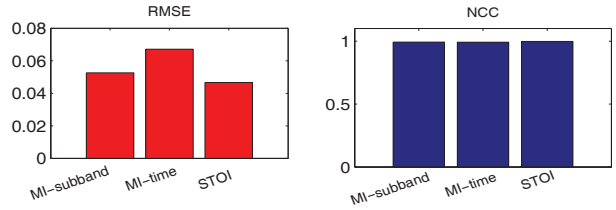

Fig. 5. The performance of proposed objective measures and STOI in terms of root mean squared error (RMSE) and normalized correlation coefficient (NCC).

introduced based on estimated mutual information between the clean speech and the processed speech in time and subband frequency domain. Our proposed measures show nice performance in terms of high correlation with subjective intelligibility measure (i.e. word correct scores). In terms of NCC, the proposed measures show very close performance to STOI. With respect to RMSE, one can observe that the proposed measure in subband domain (i.e. MI-Subband) performs better than the proposed one in time domain (i.e. MI-Time). It should be noted that the proposed measures are not specially tuned for speech intelligibility prediction. Although the adjustment of parameters for MI estimation in the proposed procedures has some impact, it is not overly critical. Nevertheless, the performance depends to some extent on the length of signal segments and k- nearest neighbor parameters. Longer segments, for instance, provide less systematic error in MI estimation but also increase the computational complexity. These tradeoffs will be investigated in more detail in future works.

There are some points which are important in conjunction with the proposed measures: 1) In general, computing actual mutual information is challenging, and in our work mutual information is estimated by using an estimator which is not ideal and includes systematic errors. The purpose of the work, presented here, is not the introduction of a novel mutual information estimator but to show how estimating mutual information between the clean speech and the processed speech can be useful in deriving an objective intelligibility measure. 2) The proposed measures can be employed to rank the performance of single-channel noise reduction algorithms concerning intelligibility of the processed speech. Although in this work noisy speech processed by single-channel noise reduction algorithms is considered, the proposed measures may be employed for the other types of algorithms which apply varying gains to the time-frequency points of noisy speech. Furthermore, linear degradations of speech e.g. additive noise, and more complicated scenarios like speech processed by source separation algorithms are other sorts of conditions which can be considered in an extended work. 3) By taking into account higher order statistics, mutual information is more general than correlation, and independence implies uncorrelatedness. Therefore, the estimation of mutual information in our proposed measures is potentially more general than the correlation coefficient (as employed in STOI).

6. REFERENCES

[1] 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.

[2] J. Ma and P.C. Loizou, "SNR loss: A new objective measure for predicting the intelligibility of noise-suppressed speech," *Speech Comm. (2010)*, doi:10.1016/j.specom.2010.10.005.

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

[4] T. Dau, D. Pschel, and A. Kohlrausch, "A quantitative model of the effective signal processing in the auditory system. i. model structure," *J. Acoust. Soc. Amer.*, vol. 99, no. 6, pp. 3615–3622, 1996.

[5] C. H. Taal, R. C. Hendriks, R. Heusdens, J. Jensen, and U. Kjems, "An evaluation of objective quality measures for speech intelligibility prediction," *in Proc. Interspeech*, pp. 1947–1950, 2009.

[6] J. B. Boldt and D. P. W. Ellis, "A simple correlation-based model of intelligibility for nonlinear speech enhancement and separation," *in Proc. EUSIPCO*, pp. 1849–1853, 2009.

[7] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Audio Speech Lang. Process.*, vol. 16, no. 1, pp. 229–238, 2008.

[8] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short time objective intelligibility measure for time-frequency weighted noisy speech," *in Proc. ICASSP*, pp. 4214–4217.

[9] T. M. Cover and J. A. Thomas, *Elements of information theory*, Wiley, 2006.

[10] J. Walters-Williams and Y. Li, "Estimation of mutual information: A survey," *Lecture Notes in Computer Science*, vol. 5589, pp. 389–396, 2009.

[11] A. Kraskov, H. Stögbauer, and P. Grassberger, "Estimating mutual information," *Phys. Rev. E 69 (6) 066138*, pp. 1–16, 2004.

[12] K.Wagener, J. L. Josvassen, and R. Ardenkjaer, "Design, optimization and evaluation of a danish sentence test in noise," *Int. J. Audiol.*, vol. 42, no. 1, pp. 10–17, 2003.

[13] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. on Acoust., Speech, Signal Process.*, vol. 32, no. 6, pp. 1109–1121, 1984.

[14] J. S. Erkelens, R. C. Hendriks, R. Heusdens, and J. Jensen, "Minimum mean-square error estimation of discrete fourier coefficients with generalized gamma priors," *IEEE Trans. Audio Speech Lang. Process.*, vol. 15, no. 6, pp. 1741–1752, 2007.