INTERSPEECH
2009 BRIGHTON

An Evaluation of Objective Quality Measures for Speech Intelligibility Prediction

Cees H. Taal[1], *Richard C. Hendriks*[1], *Richard Heusdens*[1], *Jesper Jensen*[2] *and Ulrik Kjems*[2]

[1]Delft University of Technology, Dept. of Mediamatics, [2]Oticon A/S
{c.h.taal, r.c.hendriks, r.heusdens}@tudelft.nl, {jsj, uk}@oticon.dk

Abstract

In this research various objective quality measures are evaluated in order to predict the intelligibility for a wide range of non-linearly processed speech signals and speech degraded by additive noise. The obtained results are compared with the prediction results of a more advanced perceptual-based model proposed by Dau *et al.* and an objective intelligibility measure, namely the coherence speech intelligibility index (cSII). These tests are performed in order to gain more knowledge between the link of speech-quality and speech-intelligibility and may help us to exploit the extensive research done into the field of speech-quality for speech-intelligibility. It is shown that cSII does not necessarily show better performance compared to conventional objective (speech)-quality measures. In general, the DAU-model is the only method with reasonable results for all processing conditions.

Index Terms: Speech intelligibility prediction, speech quality, objective Measure.

1. Introduction

In speech processing systems, degradations and modifications are often introduced to clean or noisy speech signals. Examples could be quantization noise in a speech-coder, residual noise and speech distortion in a speech enhancement scheme or an intelligibility improvement algorithm in a hearing aid. To determine the perceptual consequences of these artifacts, the algorithm at hand can be evaluated by means of subjective listening tests and/or a particular objective machine-driven quality assessment. Accurate and reliable objective evaluation methods are of interest since they might replace subjective tests, at least in some stages of the algorithm development process, and in this way save time and costs. Although it is not straightforward how the overall quality of a speech system to describe, people tend to divide the evaluation into the attributes speech-quality, (i.e. pleasantness/naturalness of speech) and speech-intelligibility. In this paper the focus will be on speech intelligibility.

One of the first objective intelligibility measurements was developed at AT&T Bell Labs by French and Steinberg in 1947 [1], currently known as the 'Articulation Index' (AI). Later, Steeneken and Houtgast proposed the 'Speech Transmission Index' (STI) [2], which, in contrast to AI, is also able to predict several non-linear degradations. The majority of recent published models are still based on these two objectives measures. Although these measures are suitable for several types of degradation, there still exists a wide range of processing methods for which the existing measures are less appropriate.

The lack of generally applicable intelligibility measures probably explains why new speech-algorithms are often only evaluated in terms of speech quality, despite the importance of intelligibility. In contrast to speech intelligibility, a wide range of objective speech quality measures are available, where important research is done by Quackenbush [3]. Currently, one of the most used objective quality measurements, standardized as ITU-T Recommendation P.862 in 2002, is 'Perceptual Evaluation of Speech Quality' (PESQ). This measure shows high correlation with Mean Opinion Scores (MOS) for various types of distortions [4].

Unfortunately, the relation between speech quality and intelligibility is not well understood. While one intuitively would state that a better quality would imply a better intelligibility, the contrary can also be true. For example, the quality of a noisy speech signal may be improved by applying some speech enhancement algorithm, while the intelligibility may actually decrease. This is confirmed by Hu *et al.* [5] who have shown that none, out of several well known speech enhancement algorithms, is able to improve intelligibility.

In this paper, we evaluate a wide range of objective quality measures in terms of their ability to predict the results of intelligibility listening experiments for speech degraded by additive noise and various non-linear signal modifications. The results may lead to more insight in the relation between speech-quality and intelligibility, and may help us to exploit the extensive research done into the field of speech-quality for speech-intelligibility assessment. In addition, an objective measure which is able to predict intelligibility can be used as a potential candidate to optimize for in an enhancement scheme.

Although a study on intelligibility prediction by some speech-quality measures exists [6], only very little research has been done to compare the performance of these objective measures with the intelligibility prediction of a more advanced objective intelligibility measure. Therefore, in this study we include the objective intelligibility measure from [7] and a more advanced perceptual-based model [8]. In addition, six extra measures are added, including perceptual-based quality measures originally designed for audio coding. Moreover, our research contributes by proposing model adjustments in order to improve correlation with speech intelligibility.

2. Subjective Listening Tests

The subjective data we use is obtained from two different listening experiments adopted from Kjems *et al.* [9]. In the first experiment, speech is degraded by four additive noise sources: speech shaped noise, cafeteria noise, noise from a bottling factory hall and car interior noise. Average-user psychometric curves are estimated with an adaptive procedure for each noise

The research is supported by the Oticon foundation and the Dutch Technology Foundation STW.

Abbr.	Objective Measure
sSNR	Segmental SNR [3]
LSD	Log Spectral Distance
LLR	Log-Likelihood Ratio [3]
IS	Itakura-Saito [3]
CEP	Cepstral Distance Measure [3]
WSS	Weighted-Spectral Slope Metric [11]
FWS	Frequency Weighted Segmental SNR [12]
FWSn	Normalized Frequency Weighted Segmental SNR [13]
PAR	van de Par *et al.* [14]
TAA	Taal *et al.* [15]
HNS	Hansen *et al.* [16]
PESQ	PESQ [4]
DAU	Dau *et al.* [8]
cSII	Coherence SII [7]

Table 1: *The evaluated objective measures with their corresponding abbreviations.*

type. For the second experiment the noisy signals are non-linearly processed by a technique called ideal time frequency segregation (ITFS) [10]. This technique can improve the intelligibility of the noisy speech signal significantly by applying a binary modulation pattern in a time-frequency representation. Both tests are performed with sentences consisting of five words, all spoken by the same Danish female speaker. The sentences are of the grammatical form name-verb-numeral-adjective-noun (e.g. Ingrid owns six old jackets), where each word in the sentence is picked randomly from a list of 10 possible words. For more details of the listening experiments and processing conditions we refer the reader to [9].

From the first experiment, speech signals are regenerated for each noise type and 9 different SNR values, equally spaced between -30dB and 10dB resulting in 36 different conditions. For the non-linear experiment a subset from the original settings of [9] are used, which corresponds to 96 different conditions covering a wide range of noisy speech-manipulated signals. The complete set of 132 conditions covers subjective intelligibility scores ranging from 0% to 100%. The set includes extreme cases where essentially a noise-only signal (-60dB SNR) is ITFS-processed, resulting in almost 100% intelligible speech; a challenging condition for many objective measures, since all fine-structure of the clean speech is lost.

3. Objective Measures

In Table 1 all evaluated objective measures are presented in combination with their corresponding abbreviations. This includes several conventional methods (sSNR, LSD, LLR, IS, CEP, WSS, FWS), where the critical band weighting function of FWS is set to the AI weights, (e.g. [17]). FWSn is based on FWS and is modified such that it correlates well with the speech-quality of enhanced speech signals [13]. In addition, three perceptual-based coding measures are included (PAR, TAA, HNS) followed by the objective speech-quality measure PESQ. For comparison we added the objective intelligibility measure (cSII) and an advanced perceptual model (DAU), where its distance measure is calculated by means of a linear correlation coefficient as proposed in [18].

All objective measures are based on a sample-rate of 20 kHz, except for PESQ where a sample-rate of 16 kHz is used. The frame-length is set to 512 samples (≈ 26 ms) with an overlap of 50%. Implementations of LLR, IS, CEP, WSS, FWS, FWSn and PESQ are based on the software included in [17].

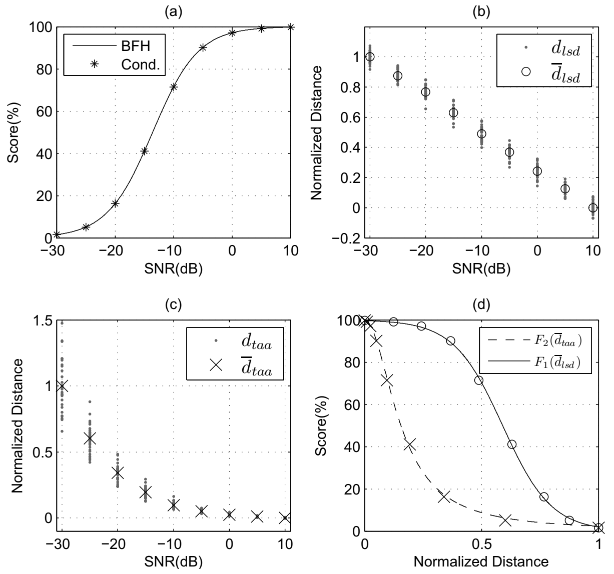

Figure 1: *Examples for the mapping procedures F_1 and F_2 for the objective measures LSD and TAA, respectively. See text for more details.*

3.1. General Procedure

For the measures that work on a frame-per-frame basis with no explicit definition how to combine per-frame information, we define the following procedure,

$$d\,(x,y) = \frac{1}{|\mathcal{M}|}\sum_{k\in\mathcal{M}} d\,(x_k,y_k), \qquad (1)$$

where x and y denote the clean and processed speech, respectively, k the frame index and $\mathcal{M}$ the set of frames with clean speech energy within 40 dB of that of the frame with maximum energy. The cardinality of $\mathcal{M}$ is indicated by $|\mathcal{M}|$. With this procedure, time-frames with no significant speech energy (mainly silence regions) and therefore no contribution to the intelligibility will not be included.

3.2. Intelligibility Prediction

To obtain a total objective measure outcome for one particular condition, 30 five-word sentences are randomly generated and processed by the particular condition (e.g. additive speech shaped noise with an SNR of -10dB). We then define the total distance measure for this condition as the average of these 30 values, say $\bar{d}$, which we normalize such that $0 < \bar{d} < 1$.

The distance outcome $\bar{d}$, for each individual objective measure, should be mapped somehow to the corresponding subjective intelligibility score. For additive noise, it is widely accepted to model the psychometric-curve using a logistic curve as a function of the SNR. Therefore, the same logistic curve can be used when an objective measure shows a linear relationship with the SNR, for one particular noise type. Hence,

$$F_1\left(\bar{d}\right) = \frac{1}{1+e^{a\bar{d}+b}}, \qquad (2)$$

where a and b are two free parameters which are adapted in order to fit the subjective data. Some objective measures are not linearly, but exponentially related with the SNR. Motivated by this, the following mapping is also included,

$$F_2\left(\bar{d}\right) = \frac{1}{1 + e^{a\log\left(\bar{d}+c\right)+b}}, \tag{3}$$

where an extra free parameter c is introduced. To fit F_1 and F_2 to the subjective scores for a certain set of conditions, a non-linear least squares procedure is used. Both mappings are included in the fitting procedure, where eventually the best mapping is chosen to evaluate the performance.

Fig. 1 shows an example for both mappings, where the set of conditions equals 9 different SNR values for the bottling factory hall noise (BFH) [1]. Subplot (a) shows the subjective average-user psychometric curve for BFH as a function of the SNR, where the 9 different conditions (Cond.) are illustrated by the stars. The average, and the 30 individual distances for the objective measures LSD and TAA are plotted in (b) and (c), respectively. Clearly, LSD shows a linear relation with the SNR while TAA indicates exponential behavior. Subplot (d) indicates that all individual data points are fitted reasonably by the proposed mappings, and thus, can predict the intelligibility score given $\bar{d}$ for this set of conditions.

3.3. Normalization

For some processing conditions of the ITFS-procedure, the energy of the processed speech can be significantly larger than the energy of the clean speech. In general, we observed that these energy differences will dominate the results for the objective measures which are based on the difference between the processed and the clean speech, say $\varepsilon = y - x$. Since these energy differences will not affect the subjective intelligibility score, the prediction performance for these types of objective measures is decreased drastically. To overcome this problem a general normalization procedure is needed.

We propose to scale the processed signal y with a factor α, such that the squared correlation coefficient between the clean signal x and $(\alpha y - x)$ is minimized, that is

$$\alpha^* = \min_{\alpha}\left(\frac{\langle x, \alpha y - x\rangle^2}{\|x\|^2\,\|\alpha y - x\|^2}\right) = \frac{\|x\|^2}{\langle x, y\rangle}, \tag{4}$$

where $\langle\cdot\,,\cdot\rangle$ denotes the inner product and α^* the optimal solution. In this manner, speech similarities between x and y will have the same energy and therefore will cancel when obtaining ε. Consider the example where speech is degraded by statistically independent and additive noise, i.e. $y = x + n$, and we are working with long signal sequences, then with the proposed normalization procedure we obtain, as expected, $\alpha \approx 1$. Note, that this implies no normalization.

In general, we observed that better correlation scores with speech intelligibility is obtained compared to applying no normalization, or a normalization procedure where the energy of x and y are equalized, which is the specified approach for HNS [16], and essentially also the procedure in PESQ [4]. Note, that for the measures PESQ and FWSn the normalization procedure is a fundamental part of the algorithm and therefore not changed.

3.4. Evaluation of the Prediction Performance

The mappings F_1 and F_2, for each objective measure, are fitted on three different data-sets: (1) the speech degraded by additive noise, (2) the non-linear ITFS-processed speech and (3) both data-sets jointly. For fitting both conditions jointly, extra weight is given to the additive-noise data such that it has equal importance compared to the non-linear processing condition, which has more data points. The performance of the objective measures, in terms of intelligibility prediction, is evaluated with three figures of merit. We denote S as the subjective score, D as the mapped objective distance measure, i as the index of the processing condition and N as the total amount of used conditions.

The performance measures consist of the root mean squared error,

$$\sigma = \frac{1}{100}\sqrt{\frac{1}{N}\sum_i \left(S_i - D_i\right)^2}, \tag{5}$$

which is normalized in the range between 0 and 1 and the normalized correlation coefficient

$$\rho = \frac{\sum\limits_i \left(S_i - \bar{S}\right)\left(D_i - \bar{D}\right)}{\sqrt{\sum\limits_i \left(S_i - \bar{S}\right)^2 \sum\limits_i \left(D_i - \bar{D}\right)^2}}, \tag{6}$$

where $\bar{D}$ and $\bar{S}$ denote the average values of the objective and subjective data, respectively. Finally, the Kendall's tau is included

$$\tau = \frac{N_c - N_d}{\frac{1}{2}N\left(N-1\right)}. \tag{7}$$

Here N_c and N_d denote the concordant and discordant pairs in the evaluated set of processing conditions. This performance measure is independent of the applied mapping and tests whether there is a monotonic relation between the predicted and subjective intelligibility score.

4. Results and Discussion

All results are plotted in Fig. 2 and ranked from left to right where the measures positioned on the right indicate better correlation with subjective test results.

In general, most measures show reasonable performance with respect to the conditions where speech is only degraded by additive noise. However, cSII, which is specifically designed for intelligibility prediction, and the more advanced perceptual-based DAU-model, show considerable better results than the remaining objective measures. For the same data most measures perform better than the advanced speech-quality measure PESQ. This is caused by the fact that the objective scores of PESQ remain constant for low SNR-values, in contrast to the subjective intelligibility scores which are still decreasing; a property which is line with the observations in [6]. The same behavior for low SNR's is even stronger present with the FWSn, which is reflected by its lowest ranking. We believe that this behavior of FWSn and PESQ is caused by their similar normalization approach where clean and degraded signal (frame) energies are equalized. Hence, re-investigating this procedure may be worthwhile in order to predict the intelligibility of speech degraded by additive noise.

By comparing the results between the signals degraded by additive noise and the non-linearly processed data, it is clear that the overall performance for most models decreases for the non-linear data. Nevertheless, the DAU-model still turns out to be the best performing model, in contrast to the cSII which reduced

[1] Note, that in the eventual results we fit the mappings on large sets of conditions. The BFH-set is solely used for illustrative purposes.

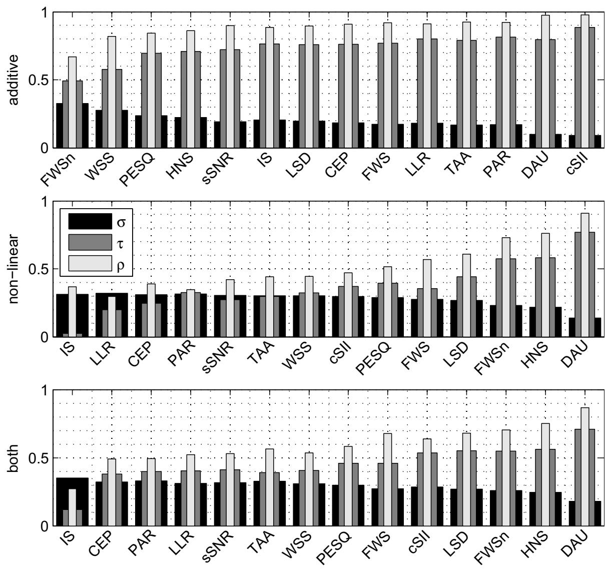

Figure 2: *Intelligibility prediction performance for all objective measures for the additive-noise (top), non-linear (middle) and both additive-noise plus non-linear (bottom) data. Objective measures positioned on the right indicate better correlation with speech intelligibility.*

in performance significantly. Investigating its results, we concludes that cSII is too sensitive for the time-domain waveform difference between the clean and degraded signal; a property which was already reported in [18]. Therefore, the relatively simple measures FWSn and LSD show better results. The high ranking of HNS can be motivated by the fact that it is based on the DAU-model. Somehow surprisingly is the difference in performance between FWS and FWSn. It turns out that for the non-linear data the normalization procedure in FWSn actually *does* improve the intelligibility prediction, which was the other way around for the speech degraded by additive noise. Probably, this can be explained by the fact that the used non-linear processing conditions in this research are closer to the enhanced signals, originally used for evaluating FWSn in [13], where the suggested normalization procedure was found to be critically important.

Regarding the condition where the objective measures predict both the additive-noise and non-linearly processed signals jointly, similar performance is observed compared to the non-linear prediction results. This is caused by the relatively broad range of distance outcomes for the non-linear data compared to the signals degraded by additive noise. Therefore, the same four best objective measures are observed as with the non-linear data, for fitting both conditions jointly. Again, the DAU-model shows indeed the best performance.

5. Conclusions

Various objective quality measures are evaluated in order to predict the intelligibility of additive noise degraded speech and non-linear processed speech. Reasonable performance is obtained for most measures with respect to the additive-noise data, where the advanced perceptual DAU-model and the objective intelligibility measure cSII shows the best results. Most measures give poor results for the non-linear data with the exception of DAU, HNS and FWSn. Remarkable was the fact that the relatively simple measures LSD and FWSn show better performance than the objective intelligibility measure cSII for the non-linear data. In general, the DAU-model is the only method with good results for all processing conditions.

6. References

[1] N. R. French and J. C. Steinberg, “Factors governing the intelligibility of speech sounds,” *J. Acoust. Soc. Am.*, vol. 19, no. 1, pp. 90–119, 1947.

[2] H. J. M. Steeneken and T. Houtgast, “A physical method for measuring speech-transmission quality,” *J. Acoust. Soc. Am.*, vol. 67, no. 1, pp. 318–326, 1980.

[3] S. R. Quackenbush, T. P. Barnwell, and M. A. Clements, *Objective Measures of Speech Quality*. Prentice Hall, 1988.

[4] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, “Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs,” *Proc. ICASSP*, vol. 2, pp. 749–752, 2001.

[5] Y. Hu and P. Loizou, “A comparative intelligibility study of speech enhancement algorithms,” *Proc. ICASSP*, vol. 4, pp. 561–564, 2007.

[6] W. M. Liu, K. A. Jellyman, N. W. D. Evans, and J. S. D. Mason, “Assessment of objective quality measures for speech intelligibility,” *Interspeech*, 2008.

[7] 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.

[8] T. Dau, D. Puschel, and A. Kohlrausch, “A quantitative model of the ”effective” signal processing in the auditory system. i. model structure,” *J. Acoust. Soc. Am.*, vol. 99, no. 6, pp. 3615–22, 1996.

[9] U. Kjems, J. Boldt, S. Pedersen, T. Lunner, and D. Wang, “Role of mask pattern in intelligibility of ideal binary-masked noisy speech,” *J. Acoust. Soc. Am. (In Review)*, 2009.

[10] D. S. Brungart, P. S. Chang, B. D. Simpson, and D. L. Wang, “Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation,” *J. Acoust. Soc. Am.*, vol. 120, no. 6, pp. 4007–4018, 2006.

[11] D. Klatt, “Prediction of perceived phonetic distance from critical-band spectra: A first step,” *Proc. ICASSP*, vol. 7, pp. 1278–1281, 1982.

[12] J. Tribolet, P. Noll, B. McDermott, and R. Crochiere, “A study of complexity and quality of speech waveform coders,” *Proc. ICASSP*, vol. 3, pp. 586–590, 1978.

[13] Y. Hu and P. C. Loizou, “Evaluation of objective quality measures for speech enhancement,” *IEEE Trans. Acoust. Speech Signal Processing*, vol. 16, no. 1, pp. 229–238, 2008.

[14] 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. Appl. Signal Process.*, vol. 2005, no. 9, pp. 1292–1304, 2005.

[15] C. H. Taal and R. Heusdens, “A low-complexity spectro-temporal based perceptual model,” *Proc. ICASSP*, pp. 153–156, 2009.

[16] M. Hansen and B. Kollmeier, “Using a quantitative psychoacoustical signal representation for objective speech quality measurement,” *Proc. ICASSP*, vol. 2, pp. 1387–1390, 1997.

[17] P. C. Loizou, *Speech Enhancement: Theory and Practice*. CRC Press, 2007.

[18] C. Christiansen, “Speech intelligibility prediction of linear and nonlinear processed speech in noise,” *M.Sc. Thesis, Technical University of Denmark, Copenhagen*, 2008.