

Automatic Live Monitoring of Communication Quality for Normal-Hearing and Hearing-Impaired Listeners

Jan Rennies, Eugen Albertin, Stefan Goetze, and Jens-E. Appell

Fraunhofer IDMT, Hearing, Speech and Audio Technology, Oldenburg, Germany
jan.rennies@idmt.fraunhofer.de

http://www.idmt.fraunhofer.de/eng/hearing_speech_audio_technology/index.htm

Abstract. This contribution presents a system, which allows for a continuous monitoring of speech intelligibility from a single microphone signal. The system accounts for the detrimental effects of environmental noise and reverberation by estimating the two relevant parameters signal-to-noise ratio and reverberation time, and feeding them to a speech intelligibility model. Due to its real-time functionality and the fact that no reference signal is required, the system offers a wide range of opportunities to monitor communication channels and control further signal enhancement mechanisms. A priori knowledge of the individual hearing loss can be used to make the system applicable also for hearing-impaired users.

1 Introduction

At modern workplaces as well as in the private sphere, chat or video-conferencing systems and mobile devices have become an integral part for acoustic communication. Under adverse acoustic conditions, however, factors such as background noise, competing speech or reverberation can reduce the quality of the communication or even prevent the user of such systems from understanding at all. Particularly for hearing-impaired people, non-optimal acoustic conditions can considerably influence the ability to communicate. In contrast to visual impairment, which can be cured by purely passive devices such as lenses or glasses in most cases, even the most recent hearing aids cannot completely restore the functionality of the complex hearing system. Further technical support in acoustic communication systems for hearing-impaired people is therefore highly desirable.

In modern teleconferencing systems, one problem of hearing-impaired users and also often of normal-hearing people is that the near-end speaker is not aware of the potentially low intelligibility at the location of the distant listener. Without technical support the distant listener needs to make constant interventions, which obviously disturbs the communication. This contribution describes a novel tool to automatically monitor the quality of acoustic communication, which can be used to create awareness for the particular problems of the far-end listener and to control further means to enhance the communication quality, e.g. by

acoustic signal processing schemes for noise reduction [1,2] or dereverberation [3]. Estimating the relevant physical parameters from the signal recorded by a microphone, the system makes use of a well-evaluated speech intelligibility model [4,5].

The remainder of this contribution is organized as follows. Section 2 provides a detailed description of the implementation of the system and its components as well as an extension for hearing-impaired listeners, before the system is evaluated in Section 3. Section 4 concludes the paper.

2 Extended Speech Intelligibility Model for Real-Time Monitoring of Acoustic Communication

The schematic structure of the system is shown in Fig. 1. Let k denote the discrete time index. The acoustic signal $y[k]$ recorded by the microphone can be described as a mixture of the speech signal $s[k]$ and environmental noise $n[k]$, i.e. $y[k] = s[k] + n[k]$. The speech is affected by the acoustic properties of the room, which can mathematically be described by the so-called room impulse response (RIR) $h[k]$. The reverberant speech signal $s[k]$ is the convolution of the clean speech signal $s'[k]$ and the RIR $h[k]$, i.e. $s[k] = s'[k] * h[k]$. For practical applications, the system needs to work with low delay. The recorded signal is therefore analyzed in blocks of length L . For each block ℓ , signal processing strategies are used to extract the main physical parameters affecting speech intelligibility, namely signal-to-noise ratio (SNR) and reverberation time (T_{60}) (see Section 2.1). These two parameters are averaged for each block and subsequently processed by a speech intelligibility model, which relates the physical parameters to a perceptual quantity. The basic principle of the speech intelligibility model is the same as used in the well-known Speech Transmission Index (STI) [4,5], which is widely applied both in research and practical systems. It is beyond the scope of this contribution to detail on the concept underlying the STI, the interested reader is referred to the literature. Briefly, the effects of reverberation and background noise are characterized by the modulation transfer function (MTF), which is related to perceived intelligibility after accounting for (among other things) auditory masking, the different importance of certain frequency regions for speech recognition and the threshold of hearing. For hearing-impaired listeners, the system can be provided with information from the audiogram (see

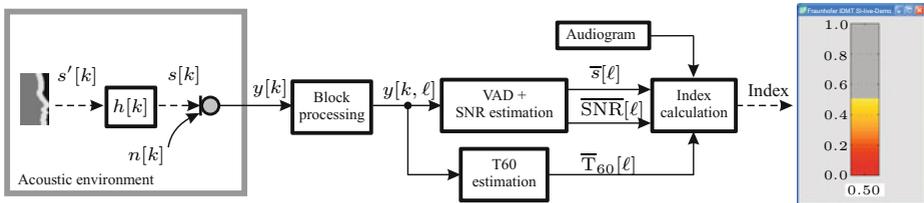


Fig. 1. Schematic structure of the processing stages for the system to estimate speech intelligibility from a single microphone signal

Section 2.2). The calculation of the STI results in an index between 0 and 1, where 0 indicates no intelligibility and 1 indicates perfect intelligibility. Intermediate values of the index can be transformed into other measures for intelligibility like e.g. percentage of correctly understood words or sentences (see for example [6]). At the output of the system, an intelligibility index is available for each block, and can be graphically displayed or used for further processing.

2.1 Estimation of Physical Parameters

As mentioned above, signal-to-noise ratio (SNR) and reverberation time (T_{60}) are the main factors affecting speech intelligibility. For estimation of the SNR the system uses a single-channel noise-reduction technique [1] combined with a voice activity detector (VAD) [2]. This technique provides an estimation of the SNR in several frequency bands n for each short-time frame ℓ . The SNR estimation is used for those frames when voice activity was detected.

The estimation of reverberation time (T_{60}) is based on the cepstral averaging procedure [7], which uses the theory of blind homomorphic deconvolution to convert the convolution $s[k, \ell] = s'[k, \ell] * h[k, \ell]$ into a sum of cepstra. If n denotes the discrete frequency index, \Re denotes the real part and $\text{IDFT}\{\cdot\}$ denotes the discrete inverse Fourier transform, the equivalent description of the convolution in the frequency domain is $S[n, \ell] = S'[n, \ell] \cdot H[n, \ell]$, and the cepstra can be calculated as

$$c_s[k, \ell] = \Re \{ \text{IDFT} \{ \log(S'[n, \ell]) + \log(H[n, \ell]) \} \} = c_{s'}[k, \ell] + c_h[k, \ell] \quad , \quad (1)$$

where $c_{s'}[k, \ell]$ and $c_h[k, \ell]$ are the cepstra of the clean speech and the RIR, respectively. Under the assumption that $s'[k]$ is not strongly correlated for different blocks ℓ , averaging over several blocks N gives an estimate of the cepstrum of the RIR

$$\hat{c}_s[k] = \sum_{i=1}^N c_{s'}[k, i] + \sum_{i=1}^N c_h[k, i] \approx \hat{c}_h[k] \quad . \quad (2)$$

Inverting the cepstrum as $\hat{h}[k] = \text{IDFT} \{ \exp(\text{DFT} \{ \hat{c}_h[k] \}) \}$ yields an estimate of the RIR, from which the reverberation time can be calculated (for details, see [7]).

2.2 Effects of Hearing Impairment

Apart from physical parameters, the calculation of the STI also uses quantities related to the auditory system (see [5] for details). One of these quantities is the threshold of hearing, i.e. the minimum level at which sounds are audible. In typical applications of the STI, normal hearing is assumed. However, since one of the main indications of hearing impairment is an elevated hearing threshold, it is possible to modify the threshold of hearing in the system to account for hearing losses. Obviously, information about the individual hearing loss of the listener cannot be estimated from the acoustic signal, and therefore has to be provided

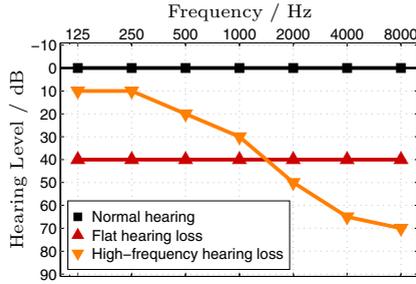


Fig. 2. Audiogram for normal-hearing (squares) and a person with flat (upward-pointing triangles) and high-frequency hearing loss (downward-pointing triangles)

as a priori knowledge. The most common measure for hearing impairment is the audiogram, which indicates the individual hearing threshold relative to the average normal-hearing person (see Fig. 2). The audiogram shows the threshold of pure tones (expressed in dB Hearing Level) as a function of the frequency of the pure tone. A value of zero indicates normal hearing. For clinical applications, a hearing level of up to 15 to 20 dB is considered normal, higher levels indicate a hearing loss. Exemplary audiograms of a normal-hearing (squares) and two hearing-impaired persons (triangles) are shown in Fig. 2. The hearing loss indicated by upward-pointing triangles is a flat hearing loss, i.e. the hearing level is similar for all frequencies. Such a hearing loss is usually observed e.g. when the middle ear is damaged. The hearing loss indicated by downward-pointing triangles is a high-frequency hearing loss, i.e. the hearing levels increase with frequency. Such a shape is typical for age-related hearing losses. The system presented here can include the effects of hearing impairment in terms of an elevated hearing threshold as indicated by the audiogram (see Fig. 1). Such an extension of the STI was already presented in [8].

In the following section, the performance of the proposed system is evaluated with respect to physical changes of the acoustic signal as well as to the effects of hearing impairment.

3 Evaluation of the Influence of SNR, Reverberation and Hearing Impairment

3.1 Methods

To evaluate our system, we tested it using continuous speech from a well-established speech test (Oldenburg sentence test, [9]). The speech was generated by concatenating sentences of the speech test in random order without pauses between the sentences. The speech was subsequently superimposed by a stationary noise, which had the same long-term spectrum as the speech. The signals were reverberated by convolution with an RIR to mimic the room acoustics of a typical office room ($T_{60} = 300$ ms). The intelligibility index was estimated every

2 s. Initially, the speech was superimposed by the stationary noise of relatively low level at an SNR of 15 dB, resulting in good speech intelligibility. After 20 s, the SNR was lowered to 10 dB. In practice, such a reduction in SNR could occur when an additional noise source is switched on. After 40 s, the SNR was further decreased to 0 dB. After 60 and 80 s, the SNR was increased to 5 and 15 dB, respectively, corresponding to a step-wise reduction of the noise source. The intelligibility index calculated with our system for this scenario of time-varying acoustic conditions is shown in Fig. 3 for a period of 100 s. The time instances when the changes in SNR occurred are indicated by vertical dashed lines. To investigate the sensitivity of the system to SNR, reverberation and hearing impairment, the calculations were made based on (i) SNR estimation alone (squares), (ii) based on SNR and T_{60} (circles) and (iii) based on SNR and hearing impairment (triangles) using the same audiograms (and symbols) as in Fig. 2. In condition (i) and (iii), the system assumed no reverberation, i.e. a totally anechoic room.

For all estimation methods, the effect of SNR can be clearly observed: the predicted intelligibility index is highest in the beginning and in the end, when the SNR is 15 dB. Accordingly, it is lowest for an SNR of 0 dB and at intermediate values for SNRs of 5 and 10 dB. The data also indicate slight variations in the predicted index, even when the SNR is constant. This can be explained by two effects. On the one hand, the energy distribution of running speech varies over time, which leads to a time-varying SNR. On the other hand, the estimations of both SNR and reverberation time are based on statistical methods [1,2], which introduce fluctuations in the prediction. In particular at higher SNR, the estimated index is lower, when the reverberation time is considered for calculation of the intelligibility index. This reflects the stronger detrimental effect of reverberation on speech intelligibility at higher SNRs, whereas intelligibility is primarily decreased by background noise at a low SNR. The two exemplary hearing losses included in the analysis result in a decreased intelligibility index for all SNR, which is expected since part of the speech energy is lower than the threshold of the assumed hearing losses.

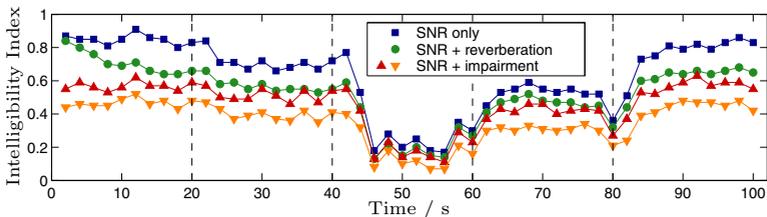


Fig. 3. Estimated intelligibility index as a function of time calculated by the monitoring system. The calculations are based on estimations of both SNR and T_{60} (circles) or SNR alone (squares). Dashed lines indicate points of time at which the SNR was changed (see text). Triangles represent estimations including hearing impairment as indicated in Fig. 2.

3.2 Results

The dependence of the intelligibility index estimated by the proposed system on SNR, reverberation time and hearing impairment shown in Fig. 3 indicates that the system can qualitatively predict the influence of these parameters on speech intelligibility. To verify whether it is also possible to derive a quantitative prediction, the estimated intelligibility was compared to an index that was computed based on a priori knowledge of SNR and T_{60} , i.e. the same index was also computed without estimating these parameters. The correlation between the calculated intelligibility index which is based on perfect knowledge of SNR and T_{60} and the estimated index is shown in Fig. 4 for the same 100 s of speech in noise and the same four conditions as in Fig. 3. Squares indicate index pairs based on an estimation (or a priori knowledge) of the SNR only, circles represent data derived from both SNR and T_{60} , and triangles indicate data including hearing impairment. Ideally, the values of estimated and calculated intelligibility index should be the same. In this case, the system would deliver the same predictions as the well-evaluated STI. Figure 4 shows that this is not the case. In general, the calculated index is higher than the estimated index. However, the deviation is systematic, and the linear correlation between estimated and calculated index is highly significant (correlation coefficient $R=0.86$, p -value <0.001). Further comparison revealed that the index pairs that correspond to time instances when the SNR changed are of particular interest. These data points are highlighted as gray symbols with black edges in Fig. 4. While most of the index pairs fit into the linear relation between calculation and estimation, some data points deviate from this trend, and it can be seen in Fig. 4 that these data points correspond to the index values at the time instances when the SNR was changed. The reason is that the system averages the estimated values over the duration of one block, and in consequence the output of the system is delayed by one block compared to the calculated index. At sudden changes of the SNR, the real intelligibility changes faster than the system can follow, resulting in one sample of over- or underestimated intelligibility. In practical applications, this delay may be a

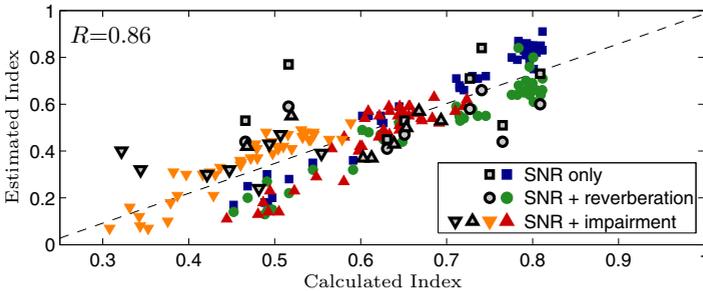


Fig. 4. Correlation between estimated and calculated intelligibility index. The symbol shapes are the same as in Fig. 3 (see text in Section 3). The dashed line indicates the linear regression $y = 1.276 \cdot x - 0.261$ obtained using a least-squares fit.

minor problem, since such sudden changes in SNR do not usually occur. Neglecting those points in the correlation analysis leads to $R=0.90$.

4 Summary and Conclusions

The described intelligibility monitoring system model is sensitive to changes in the acoustic environment such as ambient noise and reverberation as well as to different degrees of hearing impairment. Using the acoustic signal of a single microphone, the perturbing influence of room acoustic properties on speech intelligibility is accounted for by estimating reverberation time and SNR. These estimations are inherently subject to uncertainties. In general, the longer the block over which the parameters are averaged, the more accurate the estimations become. For practical use, a reasonable trade-off between estimation accuracy and the ability to react to changes in the acoustic environment requiring shorter averaging times has to be adjusted. This trade-off is application specific. Here we used an update interval of 2 s. This may be appropriate for telecommunication scenarios in which the near-end talker shall be provided with information on the intelligibility of the far-end listener. However, in case our system is used to control algorithms for acoustic signal processing, other settings may be required.

Despite the fact that the system has a slightly delayed response to sudden changes of the acoustic conditions, the linear correlation between the estimated intelligibility index and the same index calculated using a priori knowledge of the relevant parameters is highly significant. This makes the systems in principle applicable for the online monitoring of intelligibility in real applications. This constitutes a major advantage of the system compared to well-established speech intelligibility models (e.g. [5,10]), which are typically based on long-term spectra and require speech and noise signals to be available separately (i.e. they require a priori knowledge of the SNR).

The extension of the system to include information of the audiogram represents a first step towards a monitoring system also applicable for hearing-impaired listeners. Current research investigates the effects of other aspects of hearing impairment on speech intelligibility such as for example reduced frequency selectivity, reduced temporal resolution or modified loudness perception. In principle, models describing these effects could be integrated in the presented system to achieve more accurate predictions for hearing-impaired listeners.

Together with research for more elaborate models for hearing impairment, future research will aim at an increased estimation accuracy of the system's components by comparing several methods established in the literature, e.g. blind T_{60} estimation by means of autocorrelation analysis [7] or SNR estimators for fluctuating noise sources [2,11]. In order to optimize the system, further evaluations comprising different speech and noise signals, room acoustic conditions and degrees of hearing impairment as well as evaluations involving normal-hearing and hearing-impaired listeners will be made.

Due to its real-time functionality and the fact that no reference signal is required, the system offers a wide range of opportunities to monitor communication channels, e.g. direct feedback to the speaker about the particular problems

of the listener, the control of further signal enhancement mechanisms like noise reduction or dereverberation, or the control of output modalities in modern user interfaces.

References

1. Ephraim, Y., Malah, D.: Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. *IEEE Transactions on Acoustics, Speech and Signal Processing* 33(2), 443–445 (1985)
2. Marzinzik, M., Kollmeier, B.: Speech pause detection for noise spectrum estimation by tracking power envelope dynamics. *IEEE Transactions on Speech and Audio Processing* 10(2), 109–118 (2002)
3. Habets, E.: Single and Multi-Microphone Speech Dereverberation using Spectral Enhancement. PhD thesis, University of Eindhoven, Eindhoven, The Netherlands (June 2007)
4. Houtgast, T., Steeneken, H.: A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria. *J. Acoust. Soc. Am.* 77, 1069–1077 (1985)
5. International Electrotechnical Commission: IEC 60268-16 Sound System Equipment - Part 16: Objective rating of speech intelligibility by speech transmission index (1998)
6. Fletcher, H., Galt, R.: The perception of speech and its relation to telephony. *J. Acoust. Soc. Am.* 22, 89–151 (1950)
7. Schröder, J., Rohdenburg, T., Hohmann, V., Ewert, S.D.: Classification of reverberant acoustic situations. In: Boone, M. (ed.) *Proceedings of the International Conference on Acoustics NAG/DAGA 2009*, pp. 606–609. DEGA e.V, Berlin (2009)
8. Holube, I., Kollmeier, B.: Speech intelligibility prediction in hearing-impaired listeners based on a psychoacoustically motivated perception model. *J. Acoust. Soc. Am.* 100, 1703–1716 (1996)
9. Wagener, K., Brand, T., Kollmeier, B.: Development and Evaluation of a German Sentence Test. In: *Contributions to Psychological Acoustics - 8th Oldenburg Symposium on Psychological Acoustics*, pp. 439–466. bis-Verlag, Oldenburg (2000)
10. American National Standards Institute: ANSI S3.5-1997 Methods for calculation of the speech intelligibility index (1997)
11. Tchorz, J., Kollmeier, B.: SNR estimation based on amplitude modulation analysis with applications to noise suppression. *IEEE Transactions on Speech and Audio Processing* 11(3), 184–192 (2003)