Proceedings of 20th International Congress on Acoustics, ICA 2010

23–27 August 2010, Sydney, Australia

# Reverberation-Based Post-Processing for Improving Speech Intelligibility

**Magnus Schäfer, Marco Jeub, Bastian Sauert, and Peter Vary**

Institute of Communication Systems and Data Processing (ind), RWTH Aachen University, 52056 Aachen
E-Mail: {schaefer,jeub,sauert,vary}@ind.rwth-aachen.de
Web: www.ind.rwth-aachen.de

**PACS:** 43.55.Lb, 43.60.Dh, 43.71.Gv, 43.72.Gy

## ABSTRACT

When evaluating new algorithms for speech and audio coding or enhancement systems (e.g., noise reduction, echo control, or artificial bandwidth extension), one will usually listen to audio examples on headphones and not use any loudspeaker setup that might be available. The reasoning behind this choice is that using a headphone reproduction system makes it easier to identify even small signal processing artifacts which would be at least partly concealed by room reflections in listening rooms.

Usually, these artifacts due to coding or signal enhancement can not be completely removed but only minimized with respect to the constraints of the application. Examples could be a limited data rate for speech and audio coding or a trade-off decision between noise attenuation and speech distortion in noise reduction algorithms.

Based on the aforementioned superiority of headphones for making these artefacts noticeable, this contribution presents a postfilter that mimics the properties of listening rooms to conceal residual errors and artifacts. This postfilter is a finite impulse response filter that is designed according to measured or simulated room impulse responses.

The main focus of this contribution lies on the evaluation of different types of impulse responses for a reverberation-based postfiltering of speech signals that were transmitted by speech codecs at low data rates. In an exemplary study based on the Adaptive Multi-Rate Wideband (AMR-WB) speech codec, the proposed post-processing leads to an increase in the speech transmission index (STI), which indicates a better intelligibility. Optimized impulse responses for the different data rates of AMR-WB are given in order to maximize the STI.

## INTRODUCTION

Reverberation usually has a detrimental effect on various aspects of speech or audio presentation. Especially speech intelligibility was shown to be severely degraded in reverberant acoustical environments. In [1], it was even shown that the effect of reverberation on speech intelligibility could not be adequately explained by simple masking effects alone but that a combination of overlap- and self-masking has to be considered.

In contrast to that, it is often argued that having some reverberation can have a positive influence on speech intelligibility [2]. Based on this qualitative argument, the effect of short impulse responses (IRs) is quantified here by means of the speech transmission index (STI), which is a well developed measure for the intelligibility of speech in various conditions, especially taking into account the effects of additive noise and reverberation. For different scenarios and test signals, different variants of the STI were proposed and extensive testing of the different approaches has been carried out in the past. The so-called *envelope regression method* [3] was recommended in a recent comparative study [4] for the use with speech input signals and hence will be used for the comparison in this contribution.

The comparison will focus on the impact of the chosen impulse response on the speech intelligibility. There are two different types of impulse responses that have to be considered: measured and simulated IRs. There are some measured IRs available covering some environments (from low to high reverberation times) and source-receiver setups (from single to multiple sources and receivers or binaural setups with dummy heads) [5, 6, 7]. For the simulation of IRs, one has the choice of either simulating the entire room impulse response (e.g., by means of the image method [8]) or focusing on either the early reflections (e.g., in the form of a sparse IR [9]) or the diffuse, late reverberation (e.g., by means a statistical model [10]).

The different IRs will be tested as postfilters in an application scenario where speech intelligibility is an absolute necessity: telephony in a mobile, fixed-line, or voice over IP (VoIP) environment. It can be seen that even codecs that are currently being introduced into the networks fail to reach acceptable STI values especially at lower data rates. One prominent example is the Adaptive Multi-Rate Wideband (AMR-WB) codec [11] whose three lowest data rates of this specific codec are not able to provide a good speech intelligibility according to the STI.

The remainder of this contribution is organised as follows: First, the STI is shortly introduced and the specific method that will be used here is presented. A presentation of the different types of IRs follows. Subsequently, the structure of the reverberation-based post-processing is described. Optimized IRs are derived from STI measurements and explicit recommendations are deduced. The paper concludes with further possible use cases for the post-processing scheme.

## SPEECH TRANSMISSION INDEX

The basis for the STI [12] was laid in the context of measurements of early very-high-frequency-radio systems. There has been a continuous development in this area for more than three decades now, beginning with the early works of Houtgast and Steeneken [13, 14].

The STI characterizes the system-under-test based on the comparison of two signals: the input (or probe) signal $x(k)$ and the output (or response) signal $y(k)$ with the time index $k$. The original proposal of measuring STI with an artificial probe signal was later extended by different approaches to use speech as the probe signal. A good overview on the various speech-based STI approaches and a comparison thereof can be found in [4]. The basic system that is used for the calculation of the STI in all concepts can be found in Figure 1.

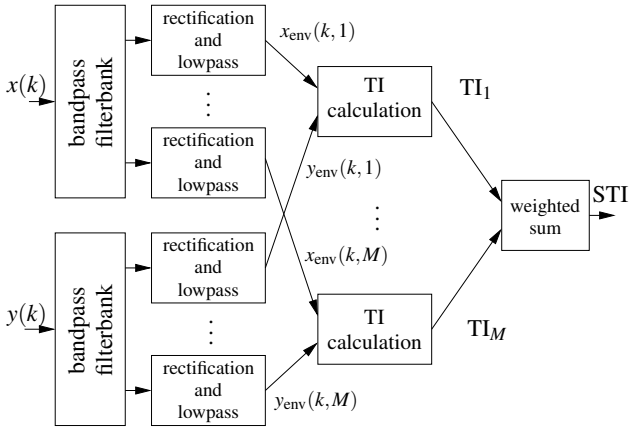


Figure 1: Block diagram of the STI calculation.

The STI is calculated as a weighted summation of the individual band transmission indices $\mathrm{TI}_m$. These are calculated in each frequency band $m \in \{1,2,\ldots M\}$ based on the envelope signals $x_{\mathrm{env}}(k,m)$ and $y_{\mathrm{env}}(k,m)$ of the bandpass-filtered input and output signals $x(k,m)$ and $y(k,m)$.

For the evaluation in this contribution, the so-called *envelope regression method* according to Ludvigsen et al. [3] is used. The extensive comparison of the different speech-based STI procedures by Goldsworthy and Greenstein [4] has shown that this method leads to equivalent results as the common non-speech-based STI method at a reasonable computational complexity.

The specific property of this method in comparison to other known approaches is that it calculates the apparent signal-to-noise ratio in each band $\mathrm{aSNR}_m$ by comparing the input and output envelope signals based on a linear regression analysis. The details can be found in [3] and [4].

## MEASURED AND SIMULATED ROOM IMPULSE RESPONSES

When evaluating or developing signal processing algorithms that are related to acoustical reverberation, one has the choice of using either measured or simulated impulse responses. Both approaches have their advantages and disadvantages:

- *Measured impulse responses* inherently capture all properties of real-world environments and are hence more precise when it comes to replicating the reality. On the other hand, there is no infinite number of properly measured IRs available that are representative for all possible application environments. This might lead to overfitting the algorithms to the available datasets.
- *Simulated impulse responses* can be calculated for practically any environment so that there is no risk of developing an algorithm only for a few rooms that happen to be measured in the past. However, simulated impulse responses do not give a perfect representation of every aspect of real IRs.

### Real Impulse Responses – the AIR Database

For the evaluation in this paper, impulse responses from the Aachen impulse response (AIR) database[1] [5] will be used as measured real-world room impulse responses. The main purpose of this database is the evaluation of speech enhancement algorithms dealing with room reverberation. The measurements with and without a dummy head took place in a low-reverberant studio booth, an office room, a meeting room, and a lecture room. Due to the different dimensions and acoustic properties, it covers a wide range of situations where digital hearing aids or other hands-free devices can be used.

The IRs in the AIR database are measured binaurally at a sampling frequency of 48 kHz. For the application as a reverberation postfilter, only single channel IRs are used. Additionally, the measurements without the dummy head are better suited since the additional shadowing and the reflections of the dummy head could lead to a false spatial impression. Hence, the left channel of each measurement without the presence of a dummy head was used here.

Depending on the measurement room, the AIR database includes different lengths of the direct path between source and receiver. The details for the excerpt that is used for the evaluation in this contribution can be found in Table 1. With this variability, different direct-to-reverberant energy ratios (DRRs) are represented in the excerpt, which allows a first look at the importance of the different parts of the IR for a possible change in speech intelligibility.

| Room | Lengths of the direct paths in m |
|---|---|
| Studio booth | 0.5, 1.0 and 1.5 |
| Office room | 1.0, 2.0 and 3.0 |
| Meeting room | 1.45, 1.7, 1.9, 2.25 and 2.8 |
| Lecture room | 2.25, 4.0, 5.56, 7.1, 8.68 and 10.2 |

Table 1: Room configurations for the AIR database.

The room parameters that influence the reverberation characteristics of the measurement rooms differ significantly. While the volume of the studio booth is small (3.00 m × 1.80 m × 2.20 m) and it is specifically designed to have a short reverberation time that is approximately constant over frequency, the lecture room is fairly large (10.80 m × 10.90 m × 3.15 m) and has very reflective surfaces (three walls mostly consist of glass windows, one wall is painted concrete and the floor is parquet). The average reverberation times for the four rooms are given in Table 2.

| Room | Average reverberation time |
|---|---|
| Studio booth | 0.12 s |
| Office room | 0.43 s |
| Meeting room | 0.23 s |
| Lecture room | 0.78 s |

Table 2: Average reverberation times for the different rooms.

### Simulation Methods

In addition to the measured impulse responses, two different simulation strategies will also be tested with respect to their

[1] The Aachen Impulse Response (AIR) database can be found at `http://www.ind.rwth-aachen.de/AIR`

applicability for improving speech intelligibility. Two significantly different models were chosen due to the fact that real-world impulse responses can be divided into two parts:

- early reflections (including the direct path) and
- late, diffuse reverberation.

In order to separately examine the influence of both components of the IR, one of the models only simulates the late reverberant tail while the other one only consists of a few strong early reflections.

The representative for the late reverberant tail is the design according to the exponential decay model by Polack [10]. It is trying to mimic the late reverberation properties of real environments (e.g. [2]) by envelope shaping of white noise.

In the first step, this model generates a white Gaussian noise signal $n(k)$ of length $T \cdot F_s$ with the target duration $T$ of the impulse response and the sampling frequency $F_s$. This signal has zero mean and is uncorrelated.

$$E\{n(k)\} = 0 \tag{1}$$

$$E\{n(k) \cdot n(k+\kappa)\} = 0 \quad \text{for} \quad \kappa \neq 0 \tag{2}$$

This noise $n(k)$ is then shaped by an exponential decay $b(k)$ which has the same length $T \cdot F_s$ as the noise and can be parameterized by the reverberation time $T_{60}$:

$$b(k) = \mathrm{e}^{-\frac{3 \cdot \ln(10)}{T_{60}} \cdot k}. \tag{3}$$

The final impulse response $h(k)$ can then be calculated as the multiplication of the two signals:

$$h(k) = n(k) \cdot b(k). \tag{4}$$

The model can be extended to include a delay for representing the length of the direct path. For the application as a signal processing postfilter, this is omitted as it would only cause additional processing delay which is generally undesirable.

This model does not consider early, individual reflections, which for most rooms form the first 50-80 ms of the IR after the arrival of the sound on the direct path. Instead, it focuses on the diffuse reflections that occur later in the IR.

An alternative that emphasizes the strong individual components that are present in real-world acoustic environments are sparse impulse responses. These consist of just very few components $h(k) \neq 0$. In the most simple setup, such an IR only consists of two coefficients: the direct path at $k = 0$ with the amplitude $h_{\text{direct}}$ and a single reflection at $k = k_1$ with the amplitude $h_{\text{reflection}}$.

It can be expressed by a two-tap finite impulse response (FIR) filter with transfer function

$$H(z) = h_{\text{direct}} + h_{\text{reflection}} \cdot z^{-k_1}. \tag{5}$$

Just like in the case of the Polack model, a delay for the length of the direct path is not included.

## POSTFILTER DESIGN

The structure of the system that is necessary for investigating the properties of the measured or simulated room impulse responses is depicted in Figure 2. It consists of a FIR filter which is used for post-processing of the respective system (e.g., speech codec).

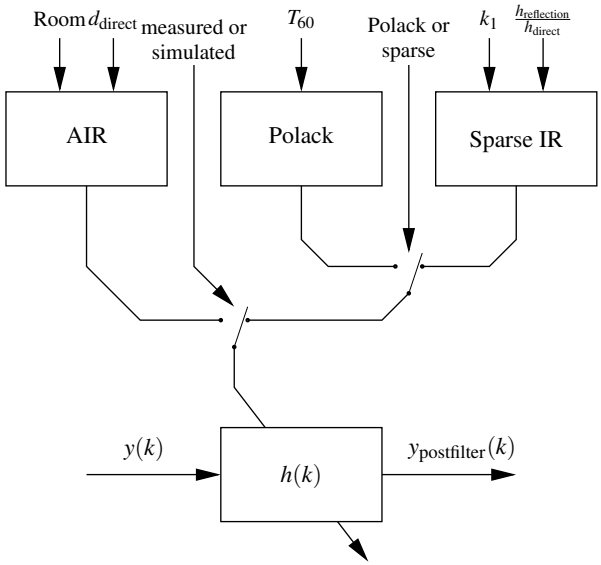


Figure 2: Block diagram of the proposed postfilter.

There are various parameters that can be set depending on the type of impulse response that is used. For the measured IRs, one has the choice between four different rooms with three to six different lengths of the direct path between source and receiver.

For the simulated IRs, the first choice has to be between the two models: either the statistical model from Polack or the sparse impulse response. The statistical model can then be parameterized by the reverberation time $T_{60}$. The sparse IR needs two input parameters: the position $k_1$ of the second filter tap in relation to the first tap and the amplitude relation $\frac{h_{\text{reflection}}}{h_{\text{direct}}}$ between the two filter taps.

To allow for a fair comparison between the different IRs, a normalization of the impulse response is carried out. This ensures that the STI is unaffected by possibly different energy levels of the signals. This is also the reason why the amplitude relation is a sufficient description of the sparse IR.

As described in the last section, the two simulation models do not incorporate an additional delay for the length of the direct path so that they inherently do not lead to an additional algorithmic delay. The measured impulse responses do have an algorithmic delay $t_{\text{direct}}$ that is related to the length of the direct path $d_{\text{direct}}$ by

$$t_{\text{direct}} = c \cdot d_{\text{direct}} \tag{6}$$

with $c$ as the speed of sound. Removing the first $t_{\text{direct}} \cdot F_s$ samples from the impulse response is a simple yet effective countermeasure and leads to an identical algorithmic delay of zero samples for all IRs. This however does not render the different measured IRs from one room identical as they still exhibit, e.g., different DRRs.

The complexity of the postfilter is directly proportional to the number of non-zero filter taps. Each non-zero filter tap requires one multiply and one add operation per sample. Since this can be computationally expensive for long filters if the processing is carried out in the time domain. Frequency domain processing could be used in those cases to increase the efficiency. Postfiltering with the sparse IR on the other hand can easily be executed in the time domain due to the very low number of non-zero taps.

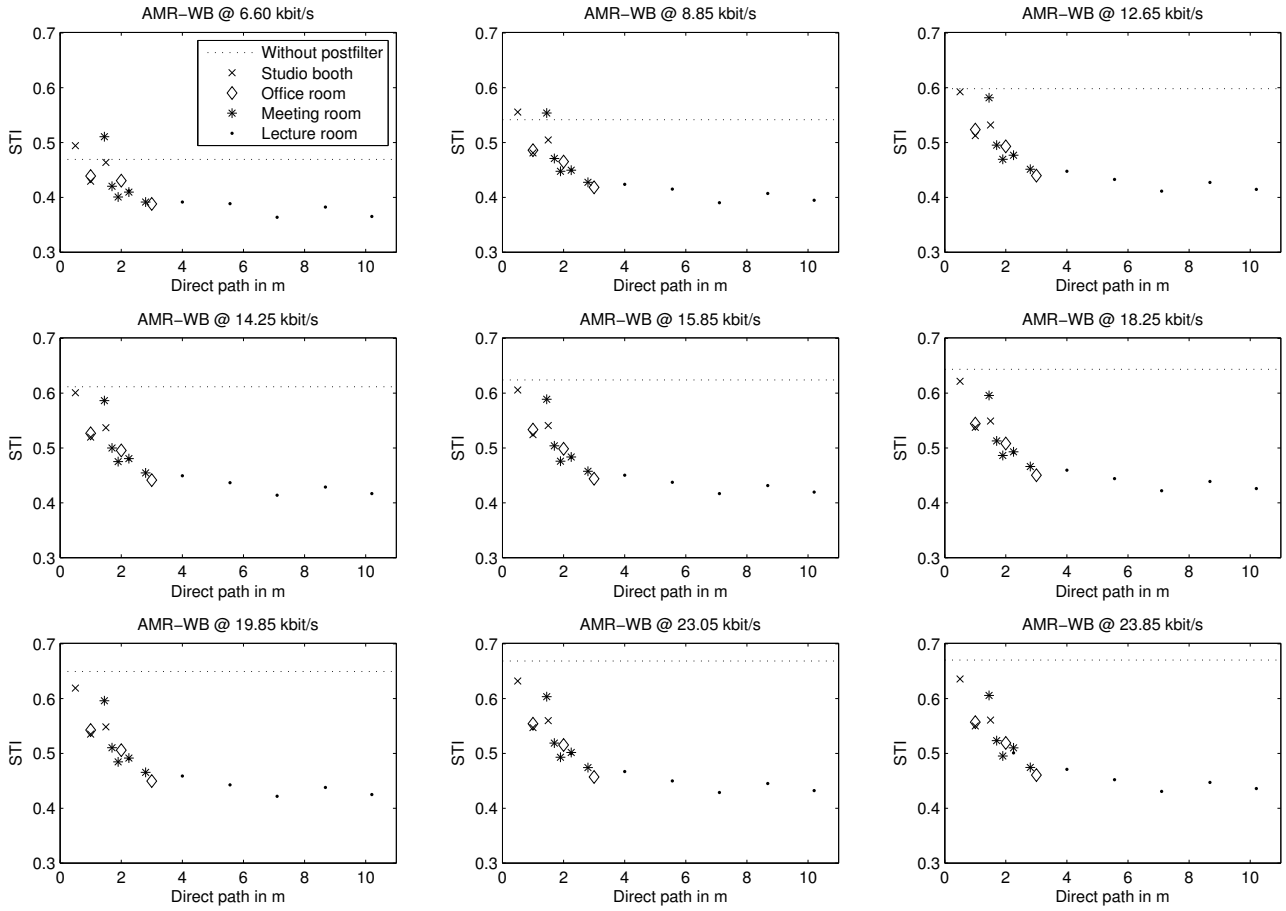


Figure 3: STI for AMR-WB after postfiltering with impulse responses from the AIR database.

## MEASUREMENT RESULTS

The proposed post-processing was evaluated as a postfilter for the AMR-WB speech codec [11]. The NTT speech corpus [15] was used as the dataset for the evaluation.

As a reference, the STI was calculated between the clear speech signal as the probe signal and the output of the AMR-WB speech codec (encoding and decoding without transmission errors) as the response signal. Each file in the speech corpus was processed individually and the STI values were averaged, the resulting mean values are given in Table 3. Usually, systems with an STI of 0.6 or greater are considered good [16] while a value of 0.5 should at least be reached for an acceptable intelligibility.

| Data rate in kbit/s | Average STI |
|---|---|
| 6.60 | 0.4693 |
| 8.85 | 0.5416 |
| 12.65 | 0.5983 |
| 14.25 | 0.6118 |
| 15.85 | 0.6242 |
| 18.25 | 0.6436 |
| 19.85 | 0.6494 |
| 23.05 | 0.6686 |
| 23.85 | 0.6703 |

Table 3: Average STI values for the different possible data rates of AMR-WB.

The first measurement results are those for a post-processing with measured impulse responses from the AIR database in four different rooms, they can be found in Figure 3. Since the AMR-WB speech codec operates at a sampling frequency of $F_s = 16\,\text{kHz}$, a downsampled version of the AIR database was used. The dotted line marks the average STI for the particular data rate of the AMR-WB speech codec without post-processing. It can be seen that most impulse responses lead to a decrease in STI with the notable exception of very short lengths of the direct path in the less reverberant rooms (studio booth and meeting room), where an increase in STI for the lower data rates is present.

The resulting STI values for the nine different operation modes of AMR-WB in combination with the proposed postfilter for the model of Polack are depicted in Figure 4. Again, the dotted line marks the average STI without post-processing. It can be seen that even for low data rates and very short reverberation times $T_{60}$, there is no increase in STI and especially for higher data rates, a significant drop in STI is obvious.

The last results are those for a postfiltering with the sparse IRs with just two non-zero coefficients in $h(k)$, which can be found in Figure 5. For all data rates, the largest STI values can be observed for the case that the second non-zero coefficient directly follows the direct path (i.e., $k_1 = 1$). The behaviour with respect to the amplitude relation $\frac{h_{\text{reflection}}}{h_{\text{direct}}}$ is less explicit, the changes between the values are significantly smaller. For the two lowest data rates, the maximum STI can be found for $\frac{h_{\text{reflection}}}{h_{\text{direct}}} = 1$ while for all the other data rates, a quotient of $\frac{h_{\text{reflection}}}{h_{\text{direct}}} = 0.3$ leads to the largest STI. An overview on the achievable STI in comparison to the STI without post-processing can be found in Table 4.

The STI is known to be well-correlated to the intelligibility of reverberant speech [4, 16]. Informal listening tests support the increase in intelligibility that is indicated by the STI. The

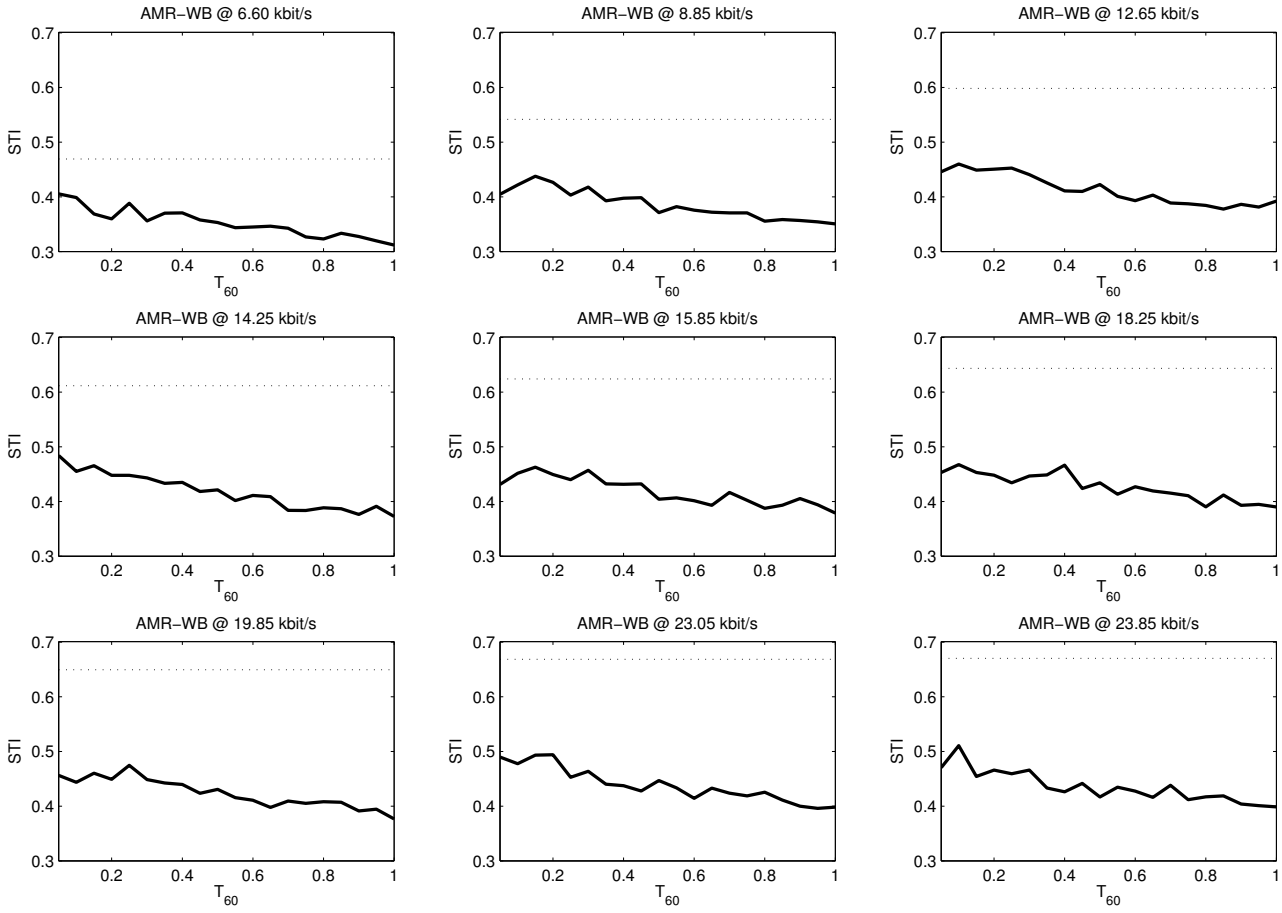


Figure 4: STI for AMR-WB after postfiltering with impulse responses according to the model of Polack.

| Data rate in kbit/s | STI without postfiltering | Achievable STI |
|---|---|---|
| 6.60 | 0.4693 | 0.6445 |
| 8.85 | 0.5416 | 0.7004 |
| 12.65 | 0.5983 | 0.7376 |
| 14.25 | 0.6118 | 0.7456 |
| 15.85 | 0.6242 | 0.7533 |
| 18.25 | 0.6436 | 0.7660 |
| 19.85 | 0.6494 | 0.7691 |
| 23.05 | 0.6686 | 0.7815 |
| 23.85 | 0.6703 | 0.7831 |

Table 4: Average STI values for the different possible data rates of AMR-WB and the maximum STI values for postfiltering with sparse IRs.

magnitude of the increase is currently under investigation by means of specific listening tests for the application scenario that was presented in this contribution.

## CONCLUSIONS

The strong individual reflections that are present in the first part of natural room impulse responses are said to have a positive effect on speech intelligibility. In this contribution, the applicability of this effect for reverberation-based postfiltering of the output signals of signal processing systems was evaluated. A quantitative study of the effect was carried out based on the speech transmission index (STI), a well-developed measure for speech intelligibility in various adverse scenarios.

Different types of impulse responses were evaluated as postfilters for the AMR-WB speech codec in order to explicitly determine which part of the impulse response leads to a reproducible and significant increase in STI.

Measured room impulse responses were shown to increase the STI for short lengths of the direct path in smaller rooms and only at very low data rates. In contrast to that, a clear decrease in STI could be observed for bigger rooms and bigger lengths of the direct path (i.e, smaller DRRs).

Postfiltering with simulated IRs leads to ambiguous results. Impulse responses that were designed according to the model of Polack and thus mimic the late reverberant properties do not offer any gain in STI. The sparse IRs on the other hand can be parameterized to significantly increase the STI even for the highest data rates of the AMR-WB speech codec. Optimum amplitude relations between the two taps of the impulse response could be derived that depend on the on the operation mode of AMR-WB.

Reverberation-based post-processing could also be applied for speech enhancement techniques. A small amount of artificial reverberation could help to conceal signal processing artifacts. Additionally, the positive effect of a certain amount of reverberation on the perceived audio quality is well known from the recording of music performances. This so-called comfort reverb leads to small temporal smearing of the speech or audio material which also overshadows, e.g., small intonation errors. Due to this and in view of the ongoing convergence of speech and audio coding, the proposed reverberation postfiltering might also be used to facilitate a better transmission of music with state-of-the-art speech codecs. Possible use-cases for this, could include improving the perceived quality for music during regular phone calls as well as streaming applications.

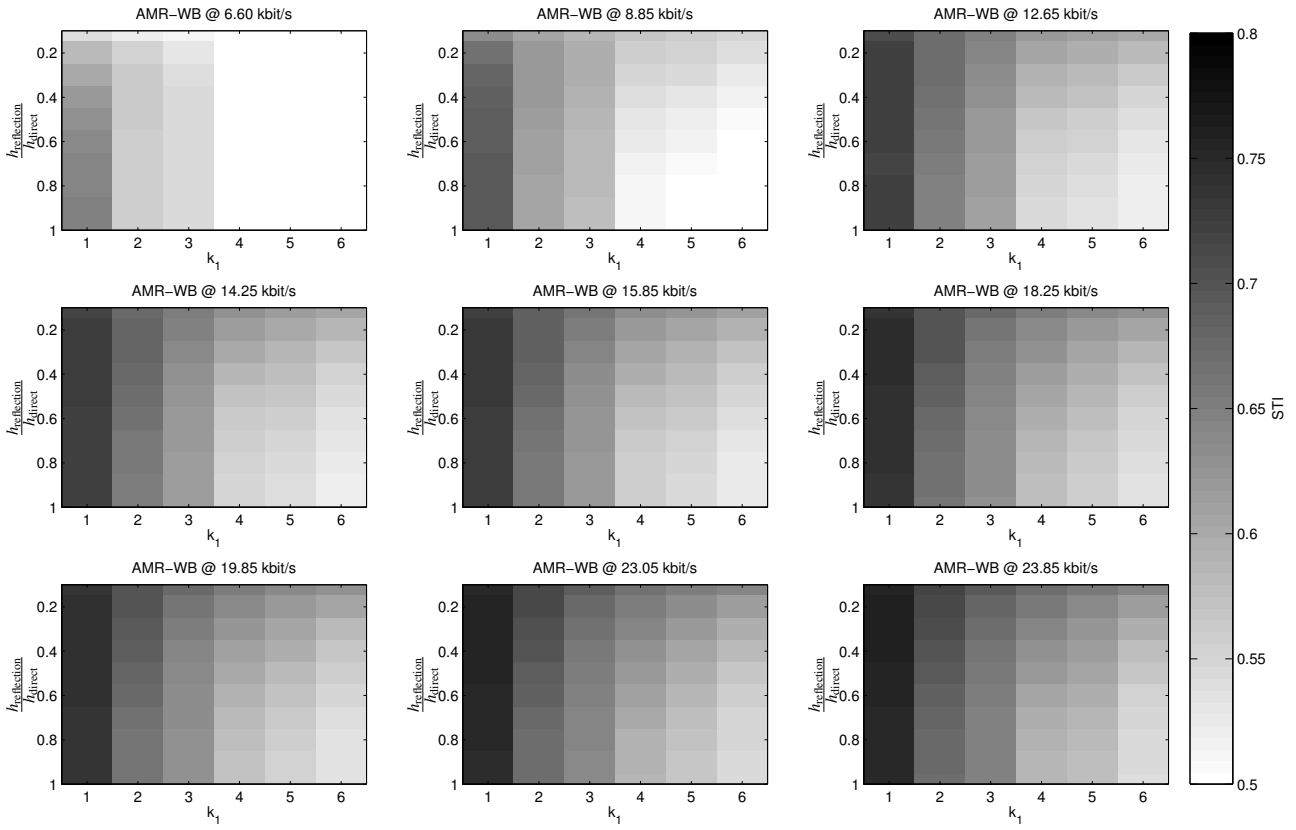


Figure 5: STI for AMR-WB after postfiltering with sparse IRs.

## REFERENCES


[1] Anna K. Nábělek, Tomasz R. Letowski, and Frances M. Tucker. Reverberant overlap- and self-masking in consonant identification. *The Journal of the Acoustical Society of America*, 86(4):1259–1265, October 1989.

[2] H. Kuttruff. *Room Acoustics*. Spon Press, Oxon, 2009.

[3] Carl Ludvigsen, Claus Elberling, Gitte Keidser, and Torben Poulsen. Prediction of intelligibility of non-linearly processed speech. *Acta oto-laryngologica. Supplementum*, 469:190–195, 1990.

[4] Ray L. Goldsworthy and Julie E. Greenberg. Analysis of speech-based speech transmission index methods with implications for nonlinear operations. *The Journal of the Acoustical Society of America*, 116(6):3679–3689, 2004.

[5] Marco Jeub, Magnus Schäfer, and Peter Vary. A binaural room impulse response database for the evaluation of dereverberation algorithms. In *Proc. 16th International Conference on Digital Signal Processing*, 2009.

[6] J.Y.C. Wen, N.D. Gaubitch, E.A.P. Habets, T. Myatt, and P.A. Naylor. Evaluation of speech dereverberation algorithms using the MARDY database. In *Proc. Int. Workshop on Acoustic Echo and Noise Control (IWAENC)*, Paris, France, 2006.

[7] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano. The cipic hrtf database. In *Proc. IEEE Workshop the Applications of Signal Processing to Audio and Acoustics*, pages 99–102, 2001.

[8] Jont B. Allen and David A. Berkley. Image method for efficiently simulating small-room acoustics. *The Journal of the Acoustical Society of America*, 65(4):943–950, 1979.

[9] Jacob Benesty, Yiteng Huang, Jingdong Chen, and Patrick A. Naylor. Adaptive algorithms for the identification of sparse impulse responses. In *Topics in Acoustic Echo and Noise Control*. 2006.

[10] J.-D. Polack. *La transmission de l'energie sonore dans les salles*. PhD thesis, Universite du Maine, Le Mans, France, 1988.

[11] ITU-T Rec. G.722.2. Wideband coding of speech at around 16 kbit/s using Adaptive Multi-Rate Wideband (AMR-WB), 2003.

[12] International Electrotechnical Commission. Sound system equipment – part 16: Objective rating of speech intelligibility by speech transmission index. IEC 60268-16:2003, May 2003.

[13] Tammo Houtgast and Herman J. M. Steeneken. Evaluation of speech transmission channels by using artificial signals. *Acustica*, 25:355–367, 1971.

[14] Herman J. M Steeneken and Tammo Houtgast. A physical method for measuring speech-transmission quality. *The Journal of the Acoustical Society of America*, 67(1):318–326, January 1980.

[15] NTT-AT. Multi-lingual speech database for telephonometry, 1994.

[16] Tammo Houtgast, Herman Steeneken, Wolfgang Ahnert, Louis Braida, Rob Drullman, Joost Festen, Kenneth Jacob, Peter Mapp, Steve McManus, Karen Payton, Reinier Plomp, Jan Verhave, and Sander van Wijngaarden. *Past, Present and Future of the Speech Transmission Index*. TNO Human Factors, Soesterberg, The Netherlands, 2002.