Measurement and Prediction of Speech Intelligibility in a Virtual Chat Room

Alexander Raake[1,2], Brian F.G. Katz[2]
[1] Deutsche Telekom Laboratories, Berlin, Germany
[2] LIMSI-CNRS, Orsay, France

Normal hearing human listeners have the remarkable ability to understand a target speech source even under quite unfavourable background noise conditions. The performance is best when the interfering sources are speech or speech-like. This ability of the auditory system is typically referred to as the Cocktail Party effect. Cocktail party processing is based both on monaural as well as binaural auditory cues. Especially the contribution from binaural cues can fruitfully be exploited for multi-party telecommunication scenarios such as teleconferencing or multi-party speech chat rooms. Here, a spatial rendering of the different speech sources transmitted to a given interlocutor enables a considerable increase in speech intelligibility (e.g. Bronkhorst, 2000).

With the advent of flexible transmission technologies such as Voice over IP (VoIP), multi-party speech communication applications have recently had their renaissance. Especially for applications where each interlocutor is located in a different physical environment, binaural technologies can be used for spatial rendering. Here, the sound is presented via headphones to replicate the sound field that is to occur at the listeners' ears in the virtual communication environment (e.g. Blauert, 1997). This can e.g. be achieved by using head-related impulse responses (HRIRs) that characterize the transmission path between a given source position and the listener's ears. For example, our application is a virtual chat room that is based on such binaural rendering technology.

In order to identify the system configuration for a given pair of interlocutors A and B that yields highest intelligibility, it is necessary to estimate the speech intelligibility related to the corresponding listener-target-distracter scenario. Only if the quantitative relation between the transmission and rendering characteristics and speech intelligibility is known, the system can be optimized for speech intelligibility.

To quantify speech intelligibility in noisy environments, intelligibility threshold measurements are often used. They allow the performance differences between a large number of acoustical conditions to be expressed in a compact manner. The sensitive measurement is achieved based on the steep psychometric function of speech identification in noise. For the 50% intelligibility threshold, the so-called speech reception threshold (SRT), slopes between 10 and 20% per dB signal-to-noise ratio (SNR) have been reported in the literature.

The Speech Reception Threshold is typically determined using an adaptive procedure that employs lists containing a certain number of sentences: Each list corresponds to one acoustical test condition. For each sentence of a given list, the speech reproduction level is chosen as a function of the number of (key-) word identification errors made on the previous sentence, targeting 50% intelligibility. The SRT is defined as the SNR at the 50% intelligibility threshold, i.e. the speech level vs. the level of the distracting signal(s).

In a recent paper, we have presented a new SRT test method for French that employs semantically unpredictable sentences (Raake and Katz, 2006). In the present paper, we summarize the test results we have obtained with this method for different conditions of a virtual chat room. The three test series (17 test conditions each) complement the available literature on the topic. Spatial rendering was obtained using HRIR measurements of a KEMAR dummy head. In particular, the tests were focused on:

- The number of sources
- The type of sources, i.e. speech vs. noise distracters
- The source characteristics
- Speech coding
- The spatial source configuration

Based on the test results, we have developed a model for predicting the SRT from a parametric description of the sources, as it applies to the listener under consideration. The model is based on the approach suggested by Bronkhorst (2000), which quantifies the intelligibility advantage of a given spatial configuration over the situation that all sources are presented from front. We extend this approach to distracters other than stationary noise sources based on own tests and newer test results from the literature. Further, we included the impact of speech coding. The initial model has been developed for distracter sources that are presented at equal level, as it is assumed by most tests described in the literature. We extend this approach to distracters of differing source levels, and include considerations on mixed speech and noise distracters.

The paper provides an outline of the model structure and details on its partly additive components. We show that the model performance is quite high, with a linear correlation of ρ^2=0.982 with own test results, and a root mean squared error of about 2 dB SRT. In an outlook, considerations are provided on how the paradigm of intelligibility can be extended towards perceived quality of virtual chat rooms.