Microscopic Models of Intelligibility and Bandwidth Reduction

Juan-Pablo Ramirez[1,2], Alexander Raake[2]

[1] *Quality and Usability Lab, Technical University Berlin, E-Mail:juan-pablo.ramirez@telekom.de*
[2] *Assessment of IP-Based Applications, Technical University Berlin, E-Mail: alexander.raake@telekom.de*

Introduction

The present study concerns the modeling of the human speech recognition in noisy environments. The intelligibility of speech is traditionally predicted by the Speech Intelligibility Index [1] (SII) or the Speech Transmission Index [2] (STI), considering the long-term-averaged features respectively of the target speaker and of the masking sources, as well as the recognition task performed. These models were developed for steady state impairments and are not relevant in the case of time varying masking [3].

A method by [3] proposed to predict the Speech Reception Threshold (SRT) for amplitude modulated noise by averaging the SII calculated within 10 ms to 30 ms time windows. The SRT is the signal-to-noise ratio that yields to an intelligibility of 50%. Even though [3] improved traditional models for fluctuating masking, it did not provide intelligibility scores on their full range from 0% to 100%.

In a recent study, [4] proposed an intelligibility assessment method based on the performances of an Automatic Speech Recognizer (ASR) [5] [6] [7]. These so-called *microscopic* models present two major assets over traditional approaches in the evaluation of intelligibility:

- similarly to the human listener, their unique input is the mono-channel mixture of speech and noise,
- they provide almost instantly phone probability estimations which can be mapped in real-time onto human average performances.

The first part of the present study describes the ASR-based model (ASRp) and the robustness of its predictions in noise for wideband speech (50 Hz to 8 kHz band-pass). The second part compares the ASRp to the SII when speech in stationary noise is linearly filtered, pointing out the limits of the model proposed.

ASR-based intelligibility predictions

The ASR algorithm uses 13 Mel Frequency Cepstral Coefficients (MFCC) [8] extracted from the acoustic wave in time-frames of 25 ms with 70% overlap between adjacent windows. For each frame, first and second derivatives of MFCC feature are concatenated to the original MFCC features forming a 39 elements acoustic feature vector. The feature vector extracted is used as input to a Multi-Layer Perceptron (MLP). The MLP is trained on the TIMIT data base [9] with acoustic features at input and phone labels at the output as target classes.
The model estimates posterior probability for each of the 39 phonemes of its library (q^i) as $p(q_t^i|x_t)$, where q_t^i is the phoneme i at time-frame t, and x_t is the acoustic feature vector for frame t. Given q_t^c the correct phone (reference) at a time-frame t, the ASRp is the average probabilities attributed to q_t^c across all N time frames of the sentence, as given in equation 2.

$$ASRp = \frac{1}{N} \sum_{t=1..N} p(q_t^c \mid x_t) \qquad (1)$$

A strong correlation between the ASRp and the SII was verified for various speakers by [4]. Figure 1 illustrates the correlation between the SII and the ASRp for one speaker and 10 different sentences masked by stationary speech-shaped noise at various levels. A 4th order polynomial fitting of the ASRp onto the SII scale is given in [4]. The model matches the predictions of the SII for stationary noise, with the assets mentioned in the introduction.

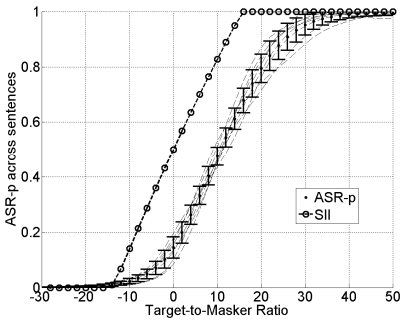

Figure 1: In open circles is the SII. In lighter lines is the ASR-based intelligibility model (ASRp) for 10 sentences of a same speaker; in dots, the average ASRp and its standard deviation.

Model robustness to bandwidth reduction

The SII predicts the intelligibility loss not only due to masking by noise, but also resulting from linear filtering. This part compares the predictions of the ASRp with those of the SII in stationary speech-shaped noise for various bandwidths.

Figure 2 shows SII predictions from the ASRp when a low-cutoff frequency (F_{low}) is applied on speech in stationary noise. The predictions are good for F_{low} below 510 Hz. For higher values of F_{low}, the ASRp may not be used to predict intelligibility. This limit is acceptable if compared to the narrow-band telephony bandwidth from 400 Hz to 3500 Hz.

In figure 3, speech in noise is low-pass filtered. The ASRp is hardly robust to the loss of upper frequencies, as its predictions do not fit the SII for high-cutoff frequencies (F_{up}) below 6400 Hz.

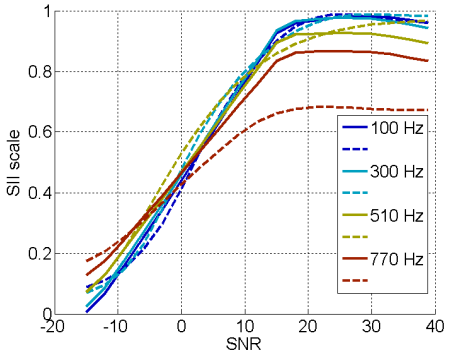

Figure 2: In solid lines is the SII, in dashed lines the predictions from the ASRp. Speech is masked by stationary speech-shaped noise. Each color corresponds to a low-cutoff frequency. Predictions are accurate below 510 Hz.

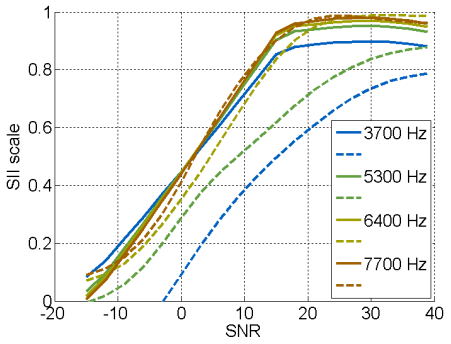

Figure 3: In solid lines is the SII, in dashed lines the predictions from the ASRp. Colors correspond to a high-cutoff frequency. Predictions are accurate above 6400 Hz.

Figure 4 shows the root mean square error (rmse) between the traditional SII and the mapping of the ASRp onto the SII scale for combinations of high- and low-cutoff frequencies. The model is not acceptable for rmse values above 0.1.

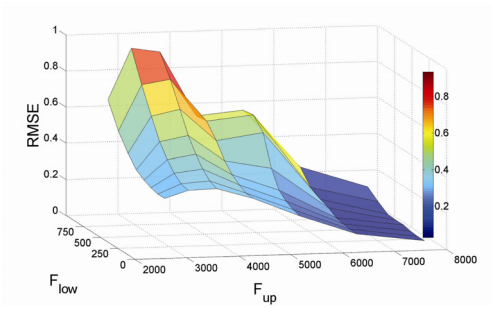

Figure 4: Root mean square error between the SII and the predictions from the ASRp for bandwidth filtering. F_{up} and F_{low} are resp. high- and low-cutoff frequencies in Hz.

Conclusion

The prediction of intelligibility by traditional means is limited to stationary listening conditions whereas in real life, disturbances such as noise are often fluctuating. In order to narrow the gap between the human speech understanding and its models (SII or STI among others), various works [4] [5] [6] [7] considered the performance of automatic speech recognizers (ASR) in noise. Band filtering being an impairment commonly found in speech transmissions, the present study questioned the accuracy of intelligibility predictions based on ASR performances as described in [4] in the presence of noise and band-pass filtering. The model is robust to low-cut filtering up to 510 Hz, which represents a satisfactory limit. It however is unreliable for frequencies cut below 6400 Hz. This drawback should to be considered in the planning of future ASR-based intelligibility models.

Literature

[1] Methods for calculation of the speech intelligibility index. ANSI Report No. S3.5-1997, American Standards Institute, New York (1997).

[2] Steeneken, H. J. M. and Houtgast, T.: A physical method for measuring speech transmission quality. J. Acoust. Soc. Am., 67 (1980), 318–326.

[3] Rhebergen, K. S., Versfeld, N. J., "A Speech Intelligibility Index-based approach to predict the speech reception threshold for sentences in fluctuating noise for normal hearing listeners", J. Acoust. Soc. Am., 117:2181-2192, 2005.

[4] Ramirez, J.-P., Ketabdar, H., and Raake, A., "Intelligibility Predictions for Speech against Fluctuating Masker", Proceedings of INTERSPEECH 2010, International Speech Communication Association, Makuhari Messe International Convention Complex, Chiba, Makuhari, Japan.

[5] Cooke, M., "A glimpsing model of speech perception in noise", J. Acoust. Soc. Am., 119:1562-1573, 2006.

[6] Holube, I., and Kollmeier, B., "Speech intelligibility prediction in hearing-impaired listeners based on a psychoacoustically motivated perception model", J. Acoust. Soc. Am., 100:1703-1716, 1996.

[7] Jürgens, T., and Brand, T., "Microscopic prediction of speech recognition for listeners with normal hearing in noise using an auditory model", J. Acoust. Soc. Am., 126:2635-2648, 2009.

[8] Davis, S., and Mermelstein, P., "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences," IEEE Trans. Acoust., Speech, Signal Processing, 8:357-366, 1980.

[9] Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., and Dahlgren, N. L., "DARPA TIMIT acoustic-phonetic continuous speech corpus CDROM", National Institute of Standards and Technology, NTIS Order No. PB91-505065, 1990.