PROCEEDINGS of the
23rd International Congress on Acoustics

9 to 13 September 2019 in Aachen, Germany

Prediction of speech intelligibility based on deep machine listening: Influence of training data and simulation of hearing impairment

Jana ROẞBACH[1]; Birger KOLLMEIER[2]; Bernd T. MEYER[3]

[1,2,3] Medical Physics and Cluster of Excellence Hearing4all, Carl von Ossietzky Universität Oldenburg, Germany

ABSTRACT

An accurate prediction of speech intelligibility (SI) is a useful tool for the development of speech enhancement algorithms; if a model is blind it could also serve as real-time SI monitor in real-world applications. A blind processing strategy implies that speech or noise references are not used for predictions. Previous work [Spille et al. (2018), Comp. Speech & Lang. doi:10.1016/j.csl.2017.10.004] introduced an accurate SI model using automatic speech recognition based on a deep neural network, which however had access to a priori information about the noise signal. This current work investigates model predictions for mismatched noise with the speech reception threshold (SRT) as target measure. Moreover, we simulate hearing impairment to explore if SI can also be predicted for hearing-impaired listeners. For similar but mismatched noises, we obtain an SRT RMSE of 1.5 dB in contrast to 1.9 dB obtained with the previous model. Further, several baseline models (SII, ESII, STOI, and mr-sEPSM) are outperformed. The hearing-impaired model reached an RMS error of 4.2 dB which is less accurate as for normal-hearing listeners. This result has been expected because of the larger variety in the SI of hearing-impaired listeners.

Keywords: speech intelligibility, deep neural networks, automatic speech recognition

1. INTRODUCTION

Communication and speech intelligibility (SI) are essential for social interaction in everyday life. Additive noise can negatively affect SI, and the strength of this influence depends on the type and relative level of the masker as well as on the individual hearing ability.

Speech enhancement algorithms have the potential to increase SI. During their development process, it is often informative to assess their influence on the SI. One option is to perform SI measurements with subjects, but this is very time-consuming and expensive. By using a model with an accurate prediction of SI, this effort can be reduced.

In the last decades, several models have been developed to predict the SI but all of them need a different amount of a priori knowledge for the prediction. An interesting question is whether it is possible to develop completely reference-free models which could be used for SI monitoring in real-world scenarios.

Spille et al. [1] developed a SI prediction model based on automatic speech recognition (ASR) that uses a deep neural network (DNN) as acoustic model. This model does not need the separated signals but requires a different kind of reference since it used identical noise signals for training and testing of the ASR system.

To investigate if the use of the identical noise signals leads to an overfitting of the model, we modified the training procedure. We used only similar noises for training and testing instead of the identical noise. Additionally, we included hearing impairment to the model to investigate if the SI of hearing-impaired listeners can also be predicted.

[1] jana.rossbach@uni-oldenburg.de

2. METHODS

2.1 Stimuli

The speech material for the ASR training as well as for the listening experiments was the Oldenburg sentence test (OLSA) [2]. It is a matrix test with the structure name, verb, number, adjective, object (e.g. “Peter buys eighteen wet shoes”), containing 120 different sentences. For the ASR training the same sentences were used but recorded from 20 different speakers (10 male, 10 female) with a total length of ten hours [3].

To mask the speech signals, eight different noise types were used. The speech shaped noise (SSN), the sinusoidal SSN (SAM-SSN), the SSN multiplied with the envelope of a broadband speech signal (BB-SSN) and the across frequency shifted SSN (AFS-SSN) are all based on an international speech signal (ISS) with a length of 11 hours. Furthermore, a noise vocoded version of the ISS (NV-ISS), a single talker (ST) and a noise vocoded version of the ST (NV-ST) were used for masking. Each noise signal was split into 80 % for the ASR training and 20 % for the testing.

2.2 ASR Training and Testing

The first step of the ASR training is the calculation of the amplitude modulation filter bank (AMFB) features [4] of the mixed audio signals (speech and noise). This is a decomposition of the signal into sub-band amplitude modulation frequency components. For this purpose, a short-time Fourier transform, the application of a mel-filter bank, a discrete cosine transform and an analysis of the amplitude modulations are performed. The AMFB features are the inputs for the DNN that maps feature inputs to phoneme categories (in our case context-dependent triphones). The network is a fully connected feed-forward model with seven hidden layers and 2048 units per layer. The DNN output is processed by a hidden Markov model to find the transcript that is most likely given the triphone estimates. The resulting model is referred to as automatic speech intelligibility prediction model (ASIP) in the following.

The training was carried out with the recorded sentences of the OLSA. The speech files were mixed with the eight noise types at random signal to noise ratios (SNRs) between -10 and 20 dB, resulting in 80 hours of mixed sound files. For the ASR testing, the original recordings of the OLSA sentences were used. These were mixed at 400 random SNRs between -30 and 20 dB with each of the noises types to sample the psychometric functions.

2.3 Inclusion of Hearing Loss in the Model

The individual prediction for hearing-impaired listeners was done by including the hearing loss into the calculation of the AMFB features. After the application of the mel-filter bank, a comparison of the magnitudes and the individual threshold takes place. All magnitudes which were smaller than the threshold were raised up to the threshold.

2.4 Listening Experiments

To verify the predictions of the model, we compared our data with the listening experiments of Schubotz et al. [5] who measured the OLSA with eight normal-hearing listeners for the same eight noise types as described in 2.1. Moreover, listening experiments with hearing-impaired listeners were performed to check the prediction accuracy of the model for this group of listeners. The subjects either had a very mild, mild or moderate hearing loss and were between 53 and 80 years old (median 71 years).

2.5 Comparative Models

To evaluate the model predictions, five comparison models were used. Four of them are established baseline models and the other is the original model of Spille et al. [1]. All of them need a different amount of a priori knowledge. The speech intelligibility index (SII) [6] and the extended speech intelligibility index (ESII) [7] require separate speech and noise signals. The short-time objective intelligibility measure (STOI) [8] needs the clean speech and the multi-resolution speech envelope power spectrum model (mr-sEPSM) [9] requires information about the type of speech material (e.g. single words, meaningful or meaningless sentences). The ASR based model of Spille et al. [1] knows the noise signal from the training and needs the transcripts for the SI prediction.

3. PRELIMINARY RESULTS AND DISCUSSION

To quantify the differences between the subjects and the model predictions, the root mean squared error (RMSE) was calculated for the ASR models and also for the baseline models SII, ESII, STOI and mr-sEPSM. The results are listed in Table 1. The RMSE of ASIP for normal-hearing listeners is close to the results of the original model, and both outperform the baseline models. The RMSE of ASIP for hearing-impaired subjects is larger, but since individual predictions were being made (based on the listener's audiogram) this prediction is a more difficult task than the prediction of averaged data from NH listeners. Nevertheless, the RMSE of ASIP HI is smaller than the RMSE of three of the baseline models with normal-hearing predictions. Perhaps the model predictions for hearing-impaired listeners can be improved by inserting additionally a distortion component instead of using only the pure tone audiogram.

Table 1 – Root mean square error (RMSE) of the SRT 50 prediction for different baseline models [5], the DNN based model of Spille et al. [1] and the current study. The current study is split in the predictions of ASIP for normal-hearing listeners and hearing-impaired listeners.

	Baseline models				ASR models		
	SII	ESII	STOI	mr-sEPSM	Spille	ASIP NH	ASIP HI
RMSE in dB	7.9	5.6	9.2	3.5	1.9	1.6	4.2

4. CONCLUSION

The results of the modified model are similar to the results of the original model [1], which means that overfitting is not an issue with the DNN-based SI model. The modification of the training procedure is one step in the direction of creating SI models that do not require a speech or noise reference. The speech intelligibility can also be predicted for hearing-impaired listeners, but the model could be extended by a distortion component to increase the accuracy for this group of listeners. In future work, different schemes for including hearing impairment into the model will be explored.

ACKNOWLEDGEMENTS

This research was supported by the DFG (Cluster of Excellence 1077/1 Hearing4all). The authors want to thank Jasper Ooster for his support during the measurements and Wiebke Schubotz and Stephan Ewert for sharing their data.

REFERENCES

[1] Spille, C., Ewert, S. D., Kollmeier, B., und Meyer, B. T. (2018). Predicting speech intelligibility with deep neural networks. Computer Speech and Language, 48:51–66.

[2] Wagener, K. C., Brand, T., and Kollmeier, B. (1999a). Development and evaluation of a German sentence test Part III: Evaluation of the Oldenburg sentence test. Zeitschrift für Audiologie, 38(3):86 95.

[3] Meyer, B. T., Kollmeier, B., and Ooster, J. (2015). Autonomous measurement of speech intelligibility utilizing automatic speech recognition. Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, 2015 January:2982 2986.

[4] Moritz, N., Adiloglu , K., Anemüller, J., Goetze , S., and Kollmeier, B. (2017). Multi-Channel Speech Enhancement and Amplitude Modulation Analysis for Noise Robust Automatic Speech Recognition. Computer Speech and Language, 46:558-573.

[5] Schubotz, W., Brand, T., Kollmeier, B., and Ewert, S. D. (2016). Monaural speech intelligibility and

detection in maskers with varying amounts of spectro temporal speech features. The Journal of the Acoustical Society of America, 140(1):524-540.

[6] ANSI (1997). ANSI S3.5-1997, American national standard methods for calculation of the speech intelligibility index. American National Standards Institute, New York.

[7] Rhebergen, K. S. und Versfeld, N. J. (2005). A Speech Intelligibility Index based approach to predict the speech reception threshold for sentences in fluctuating noise for normal-hearing listeners. The Journal of the Acoustical Society of America, 117(4):2181–2192.

[8] Taal, C. H., Hendriks, R. C., Heusdens, R., und Jensen, J. (2011). An algorithm for intelligibility prediction of time-frequency weighted noisy speech. IEEE Transactions on Audio, Speech and Language Processing, 19(7):2125–2136.

[9] Jørgensen, S. und Dau, T. (2011). Predicting speech intelligibility based on the signal-to-noise envelope power ratio after modulation-frequency selective processing. The Journal of the Acoustical Society of America, 130(3):1475–1487.