Simulating spatial speech recognition performance with an automatic-speech-recognition-based model

Marc René Schädler[1], Paul Kranzusch[1], Christopher Hauth[1], Anna Warzybok[1]
[1] *Medical Physics and Cluster of Excellence Hearing4All, 26111 Oldenburg, Germany*

Introduction

The speech recognition performance of human listeners is an important factor in the evaluation of acoustics and signal processing in communicative scenarios. Especially for listeners with hearing problems, the realistic assessment of the effect of environmental factors on their speech recognition performance is important to identify accessible communication environments. Moreover, for listeners with impaired hearing, the speech recognition performance in the same acoustic scene can strongly vary, depending on the acoustical parameters as well as on the individual hearing abilities. When non-linear or time-dependent signal processing (e.g. hearing aid processing) is part of the communication channel, it is often difficult to predict the effect on speech recognition performance using objective methods [1]. Recently, the individual speech recognition performance of listeners with aided impaired hearing was predicted with high accuracy (3.4 dB root-mean-square (RMS) prediction error) in static *monaural* noisy listening conditions [2]. Those predictions were performed with the simulation framework for auditory discrimination experiments (FADE, [3]), where a re-purposed automatic speech recognition (ASR) system was used to simulate speech recognition experiments.

However, the majority of communication environments allow for head movements and *binaural* listening. In these scenes, binaural signal processing schemes (e.g. with binaural hearing devices) can interact with effects related to spatial hearing. At least in anechoic listening conditions, the masking release due to a spatial separation of speech and noise signals can be up to 15 dB for the matrix sentence test [4]. Due to the different effect of the head shadow on spatially separated signals, the signal-to-noise ratios (SNRs) at each ear is different, which results in a better ear and a worse ear. However, the binaural performance is usually observed to be better than the monaural performance with the better ear, i.e. the information of the worse-ear signal improves the performance. This requires an interaction between the signals of both ears which is traditionally modeled with an equalization-cancellation (EC) principle [8] that can be interpreted as an adaptive SNR-optimizing beamformer. The speech recognition thresholds (SRTs) in [4] were modeled with the binaural speech intelligibility model (BSIM), which is an extension of the speech intelligibility index (SII) by prepending an EC stage to it.

In a first approach with FADE, binaural speech recognition was simulated by concatenating two feature vectors (left ear and right ear) [1]. This can be interpreted as a model of automatic better-ear listening (ABEL). The approach was used to predict the beneficial effect of binaural noise suppression schemes on SRTs. While the predicted *improvement* in SRT showed good agreement with the measured data in a cafeteria scene, the *unaided* SRTs were predicted to be worse than the measured SRTs. This shows that the better-ear listening approach is not sufficient to model binaural listening with FADE.

Implementations of the EC principle require a comparatively high temporal signal resolution. Contra-lateral inhibition describes an alternative model where the signal on the ipsi-lateral side can suppress portions of the signal on the contra-lateral side in a longer temporal context. This concept was recently implemented to improve speech recognition performance for users of cochlear implants [5]. The main idea is that information which is already encoded on one side can be removed from the other side to unmask possibly masked information. In the current contribution, this concept is simplified and implemented in FADE by taking the difference between the left and the right feature vectors as an additional feature vector into account. The extended version of the FADE simulation approach is evaluated with respect to basic binaural listening experiments, i.e., SRTs in spatial configurations and binaural masking level differences (BMLDs). This is possible, because in FADE the same simulation approach can be used for speech recognition as well as for basic psycho-acoustic experiments.

With this unique feature, model parameters of the speech recognition model can be inferred from measured psycho-acoustic data. In a recent study [2], the loss of information due to the individual hearing impairment was implemented in the feature extraction stage, where the parameter values of the signal degradation model were inferred from tone(-in-noise) detection experiments [2]. For this, a parameter controlling a supra-threshold signal degradation, the level uncertainty u_L was used to model the distortion component of hearing loss [6] which was postulated earlier [7]. The same parameter can be expected to remove information from the difference of the feature vectors of the left and the right channel, and hence influence binaural speech recognition performance. In this contribution the u_L parameter value is inferred from the BMLD experiments and used to predict SRTs in anechoic spatial listening conditions.

Methods

Binaural masking level differences

Binaural tone-in-noise detection experiments were simulated and compared to literature data in [8]. Tone-in-

noise detection thresholds were simulated with pure tones of 600 ms duration flanked by 200 ms cosine ramps. The noise maskers were bandpass-filtered white noise signals with a bandwidth of two octaves centered around the target frequency on a logarithmic frequency scale. Two conditions were considered: S_0N_0, in which the left and right signals were identical, and $S_\pi N_0$, in which the tone had an inter-aural phase shift of π. The difference in detection threshold between S_0N_0 and $S_\pi N_0$ is reported as the BMLD.

Spatial speech recognition thresholds

The German matrix sentence test was simulated for the anechoic conditions described in [4]. The matrix sentence test, which exists in more than 20 languages [9], consists of a set of 50 common word, from which syntactically fixed sentences like "Peter kauft fünf schöne Messer" ("Peter buys five nice spoons") are generated. For measurements with human listeners, phonetically balanced lists of 20 sentences are presented in noise, and after each presented sentence the speech level is adapted with a target of 50% correct word recognition score. In [4], the target speaker was located in front of the listener and the SRTs were measured for different positions of a stationary noise masker located around the listener. For the simulations in this work, the same head-related impulse responses (HRIRs) were used. The simulated SRTs are compared to the empirical data and predictions with the speech intelligibility index (SII)-based BSIM presented in [4].

Binaural speech intelligibility model

For the BSIM, the SII was extended with an EC stage [4]. The EC principle works on the time series of amplitude values by finding the relative gain and time delay between the left and right signal channel which minimizes the energy of the difference signal. In this way, the energy of a dominant noise masker, i.e. if the target-to-masker ratio is negative, can be completely removed. It was shown that this approach is suitable to model BMLDs [8]. To model the BMLDs of human listeners, i.e. to avoid the complete removal of the masker energy, an internal noise has to be assumed and its value to be inferred from empirical data. The combination of the EC principle with the SII was shown to be suitable to predict the SRTs in binaural listening conditions [4]. However, for predictions of SRTs, the SII is a latent variable with values between 0 and 1 (0 - unintelligible, 1 - perfectly intelligible) which needs to be mapped to an SRT in a measured reference conditions. Hence, SII-based approaches can only predict differences in SRT relative to a reference condition in which the outcome of the speech test needs to be known. For comparison, the BSIM predictions from the literature [4] are reported.

Simulations of experiments with FADE

FADE version 2.3.1 [11] was used to simulate the described experiments analogous to the standard procedure proposed in [3]. In this approach, a simple ASR system using Gaussian Mixture Models (GMM) and Hidden Markov Models (HMM) is trained on noisy matrix test speech material in matched training conditions to discriminate between the 50 words. The trained system is than used to determine the psychometric function of the system by evaluating recognition rates at various signal-to-noise ratios. From the psychometric function, the requested SRT is interpolated and reported as the predicted outcome of the test. Due to the matched training, i.e. using the same procedure for generating training and testing data, the predicted value can be interpreted as an estimate of the maximum achievable performance, i.e. the minimum achievable SNR. Tone detection thresholds can be predicted with the same approach by interpreting the classes "tone" and "no tone" as two words that need to be discriminated. Predictions with this approach for SRTs and tone detection thresholds were found to be in line with the performance of human listeners [3]. Like in human listeners, the recognition rate of the ASR system is limited due to the masker and the representation of the audio signals, i.e. the feature extraction. In contrast to index-based methods like the SII, no empirical SRT data is needed to predict the outcome of a speech test, e.g. an SRT.

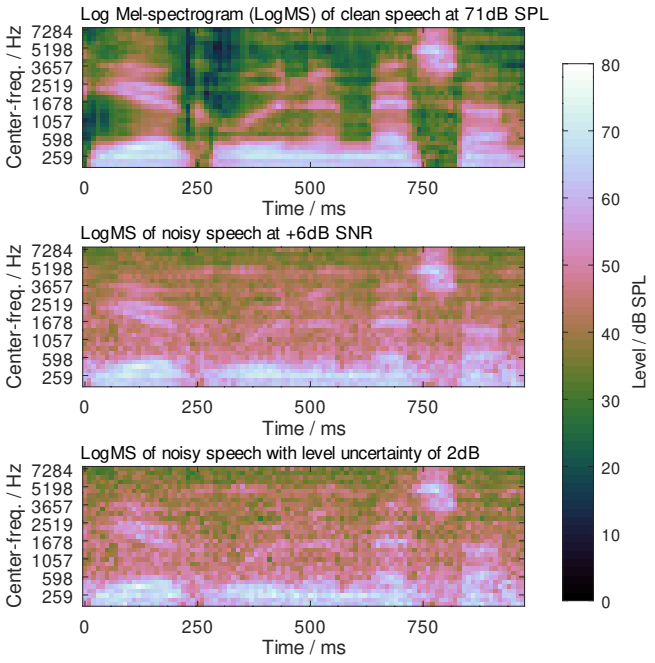

Figure 1: Log Mel-spectrogram (LogMS) of a clean German speech sample (upper panel), the LogMS of the same speech signal in noise at 6 dB SNR (center panel), and the LogMS of the same noisy speech signal with a level uncertainty u_L of 2 dB.

The basis for the feature extraction used in FADE is a spectro-temporal representation, similar to a spectrogram, which is widely used in ASR solutions; the logarithmically scaled Mel-spectrogram (LogMS). An example of a LogMS of a clean speech signal is depicted in the upper panel of Figure 1. It has a temporal resolution of 10 ms and a spectral resolution of about 1 ERB, mimicking the spectral resolution of human auditory filters, where the amplitude values are compressed with the logarithm. The compression of the amplitudes with the logarithm results in masking of information when two sig-

nals are added. This can be observed in the center panel in Figure 1, where the test-specific stationary noise was added to the clean speech signal, before the LogMS was calculated. The time-frequency bins which represent high levels are not altered, while the regions with low levels are masked by the noise signal.

The level uncertainty is implemented in the domain of the LogMS as an additive noise, which may be interpreted as a convolutional noise in the corresponding linear amplitude domain. For this, values are drawn from a standard normal distribution and multiplied with the factor u_L before they are added to the time-frequency bins of the LogMS. In the lower panel in Figure 1 the effect of a level uncertainty of 2 dB is illustrated, where it can be observed that it likewise alters time-frequency bins with high and low amplitudes. This property is the reason why it was proposed to model a limited level resolution, i.e. a supra-threshold or distortion component of hearing loss [6]. In the simulations presented in this study, values of 0.5, 1.0, and 2.0 dB were considered for u_L. As described in [3], the LogMS is only the basis of the feature vectors employed in FADE. From the LogMS, Separable Gabor Filter Bank (SGBFB) features [10] were extracted and used as features for the ASR system.

Binaural extensions of FADE

For the ABEL approach, the feature vectors for the left and the right channel were calculated separately and concatenated [1]. The corresponding code listing describes the concatenation in GNU/Octave syntax: `feat = [feat_left; feat_right];`. This can be interpreted as a model of better ear listening because the ASR system can learn from both channels and, in theory, just use the information in the channels of the respective better ear. The principle of contra-lateral inhibition (KAIN) was implemented by replacing the cited code portion of the SGBFB-ABEL feature extraction with the following code: `feat_diff = feat_left - feat_right; feat = [feat_left; feat_diff; feat_right];`. Hence, an additional feature vector was generated by subtracting the left and right SGBFB feature vectors element-wise. Because the SGBFB features can be described as a linear combination of the time-frequency bins of the LogMS, the difference could as well be calculated on the LogMS; which, however, would be computationally less efficient. For the KAIN approach, the three feature vectors (left, right, and difference) were concatenated.

Results

Binaural masking level differences

The simulation results of the BMLD experiment are depicted in Figure 2 on top of the empirical data from the literature, which is indicated by the black symbols. The ASR system with the ABEL binaural extension approach showed no masking release due to the phase shift of the target signal; i.e., the simulated BMLDs were less than ±1 dB for all frequencies. In contrast, with the KAIN binaural extensions, frequency-dependent binaural masking release was observed, which was less pronounced at high frequencies. As expected, the simulated BMLDs strongly depended on the level uncertainty and were

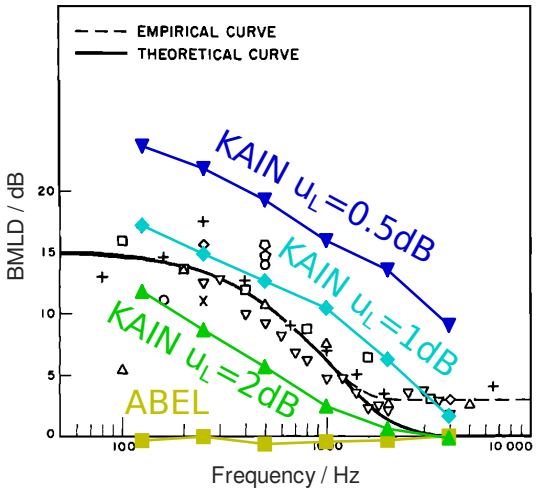

Figure 2: Modified from [8]. Black symbols indicate the measured BMLDs from the literature. The colored symbols represent the simulated BMLDs with the ABEL and KAIN approaches with different values for u_L.

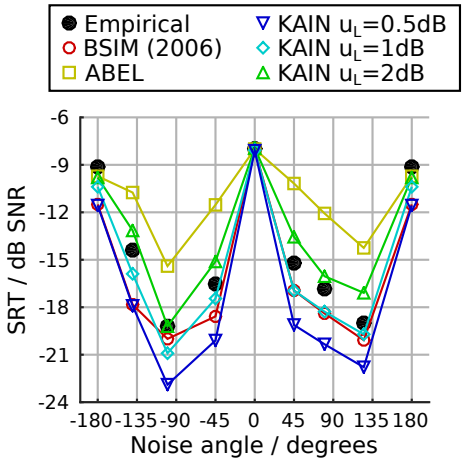

Figure 3: Empirical SRTs and predictions with BSIM from [4]. Black symbols indicate the measured SRTs and red symbols the BSIM predictions from the literature. The other colored symbols represent the simulated BMLDs with the ABEL and KAIN approaches with different values for u_L.

found to be lower for increased values of u_L. For tone frequencies up to 2 kHz, the simulations seem to describe the literature data from [8] best with a value between 1 and 2 dB for u_L. The empirical data from the literature shows small BMLDs of about 3 dB at frequencies above 2 kHz. In that frequency range, the data would be better described with a lower value for u_L of about 1 dB. The solid line in Figure 2 indicates the best description of the empirical data with the EC-based model according to [8], which describes well the empirical data below 2 kHz but misses the lower BMLDs at frequencies above 2 kHz.

Spatial speech recognition thresholds

Figure 3 shows the simulation results of the spatial speech recognition test along with the empirical SRT data and the corresponding BSIM predictions from the literature. The empirical results show a huge benefit, i.e. decrease in SRT, when the noise signal was spatially separated from speech signal, which was most pronounced at −100°

and +125° with an SRT of about -19 dB. All models accurately predicted the SRT in the co-located condition, where speech and noise signal were located in front of the listener at 0°. For the BSIM, this was the reference condition, i.e. the value was not predicted but set to the empirical value. When the noise was located behind the listener at ±180°, all models slightly overestimated the release from masking by 1 to 3 dB. Moreover, all models predicted a decrease in SRT when the noise signal was moved to the sides and correctly predicted that the lowest SRTs were at −100° and +125°.

However, the predictions deviated differently from the measured data when the noise was lateralized. BSIM over-estimated the human performance by 1 to 4 dB. The FADE-based predictions showed the same ranking than in the predictions of the BMLD, where ABEL predicted the lowest, and KAIN with u_L=0.5 dB the highest performance. With ABEL, human performance was underestimated by 3 to 5 dB. But, even if the system did not show a BMLD for tone detection, it predicted a decrease by about 8 and 6 dB when the noise was moved form the front (0°) to −100° and +125°, respectively. This is because the SNR improves on the contra-lateral ear due to the head shadow effect and the modeling approach was able to exploit that information. With KAIN, human performance was slightly under-estimated by 0 to 2 dB with u_L=2 dB and slightly over-estimated by 1 to 2 dB with u_L=1 dB. The predicted SRTs with KAIN and u_L=0.5 dB were found to be 3 to 4 dB lower than the observed SRTs in the lateralized conditions. The most accurate FADE-based predictions were observed with the values for u_L which were inferred from the simulations of the BMLD experiment, i.e. a value between 1 and 2 dB.

Discussion and conclusions

Spatial SRTs in noise in an anechoic setup were predicted by an ASR-based model that directly exploits the difference of robust ASR features between the left and right channel: FADE with the KAIN binaural extension approach with a level uncertainty u_L between 1 and 2 dB. The level uncertainty, originally introduced to simulate a monaural supra-threshold component of hearing loss [6], was shown to strongly modulate the binaural performance, and suitable parameter values for u_L were inferred by comparing simulated to measured BMLDs. The prediction performance was close to or better than with the BSIM, which implements an EC principle. An important difference to the approach with BSIM is that the left and right signal are compared on a stage where the temporal resolution is about 10 ms which is not sufficient to resolve inter-aural time differences due to the HRIRs. The implementation of the binaural interaction can be interpreted as the difference of the LogMS of the left and the right channel, which hence encodes level differences between the two ears.

While the results do not allow to draw conclusions about the human auditory system, they show that the outcome of a simple experiment on spatial speech recognition performance in anechoic listening conditions can be predicted without explicitly using the inter-aural time or phase differences present in the input signals. It is also notable that the same model could predict BMLDs, which was not possible when the difference feature vector was missing, i.e. FADE with ABEL. This can be interpreted as a hint that robust ASR features already contain sufficient binaural information to explain some basic binaural phenomena which the GMM/HMM back-end fails to decode if this information is not explicitly encoded. In this regard, a deep-neural-network-based approach might show to be better suited in the future. The (originally monaural) level uncertainty was found to be well suited to explain, i.e. limit, binaural listening performance; this immediately raises the question if the binaural hearing performance of listener with impaired hearing could be predicted based on monaural parameters.

Of course, the validity of the presented approach should be tested in more realistic spatial communication configurations and in combination with binaural signal processing algorithms.

Acknowledgements

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - Projektnummer 352015383 - SFB 1330 A 3.

References

[1] Schädler, M. R., Warzybok, A., Kollmeier, B.: Objective prediction of hearing aid benefit across listener groups using machine learning: Speech recognition performance with binaural noise-reduction algorithms. Trends in Hearing 22 (2018)

[2] Schädler, M. R., Hülsmeier, D., Warzybok, A., Kollmeier, B.: Individual aided speech recognition performance and predictions of benefit for listeners with impaired hearing employing FADE. Trends in Hearing (2020)

[3] Schädler, M. R., Warzybok, A., Ewert, S. D., Kollmeier, B.: A simulation framework for auditory discrimination experiments: Revealing the importance of across-frequency processing in speech perception. The Journal of the Acoustical Society of America 139(5) (2016), 2708–2722

[4] Beutelmann, R., Brand, T.: Prediction of speech intelligibility in spatial noise and reverberation for normal-hearing and hearing-impaired listeners. The Journal of the Acoustical Society of America 120(1) (2006), 331–342

[5] Lopez-Poveda, E. A., Eustaquio-Martín, A.: Objective speech transmission improvements with a binaural cochlear implant sound-coding strategy inspired by the contralateral medial olivocochlear reflex. The Journal of the Acoustical Society of America 143(4) (2018), 2217–2231

[6] Kollmeier, B., Schädler, M. R., Warzybok, A., Meyer, B. T., Brand, T.: Sentence recognition prediction for hearing-impaired listeners in stationary and fluctuation noise with FADE: Empowering the attenuation and distortion concept by Plomp with a quantitative processing model. Trends in hearing 20 (2016)

[7] Plomp, R.: Auditory handicap of hearing impairment and the limited benefit of hearing aids. The Journal of the Acoustical Society of America 63(2) (1978), 533–549

[8] Durlach, N. I.: Equalization and cancellation theory of binaural masking-level differences. The Journal of the Acoustical Society of America 35(8) (1963), 1206–1218

[9] Kollmeier, B., Warzybok, A., Hochmuth, S., Zokoll, M. A., Uslar, V., Brand, T., Wagener, K. C.: The multilingual matrix test: Principles, applications, and comparison across languages: A review. International Journal of Audiology 54(sup2) (2015), 3–16

[10] Schädler, M. R., Kollmeier, B.: Separable spectro-temporal Gabor filter bank features: Reducing the complexity of robust features for automatic speech recognition. The Journal of the Acoustical Society of America 137(4) (2015), 2047–2059

[11] FADE - Simulation framework for auditory discrimination experiments version 2.3.1, https://doi.org/10.5281/zenodo.3734203