INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Microscopic multilingual Matrix test predictions using an ASR-based speech recognition model

Marc René Schädler, David Hülsmeier, Anna Warzybok, Sabine Hochmuth, Birger Kollmeier

Medical Physics and Cluster of Excellence Hearing4all
University of Oldenburg D-26111 Oldenburg, Germany

{marc.r.schaedler, david.huelsmeier, a.warzybok, sabine.hochmuth, birger.kollmeier}
@uni-oldenburg.de

Abstract

In an attempt to predict the outcomes of matrix sentence tests in different languages and various noise conditions for native listeners, the simulation framework for auditory discrimination experiments (FADE) and the extended Speech Intelligibility Index (eSII) is employed. FADE uses an automatic speech recognition system to simulate recognition experiments and reports the highest achievable performance as the outcome, which showed good predictions for the German matrix test in noise. The eSII is based on the short-time analysis of weighted signal-to-noise ratios in different frequency bands. In contrast to many other approaches, including the eSII, FADE uses no empirical reference. In this work, the FADE approach is evaluated for predictions of the German, Polish, Russian, and Spanish matrix test in stationary and fluctuating noise conditions. The FADE-based predictions yield a high correlation (Pearsons $R^2 = 0.94$) with the empirical data and a root-mean-square (RMS) prediction error of 1.9 dB outperforming the eSII-based predictions ($R^2 = 0.78$, RMS = 4.2 dB). FADE can also predict the data of subgroups with only stationary or only fluctuating noises, while the eSII cannot. The FADE-based predictions seem to generalize over different languages and noise conditions.

Index Terms: speech intelligibility prediction, matrix sentence test, robust ASR features, modeling approaches

1. Introduction

Traditional macroscopic speech intelligibility models, such as the articulation index (AI; [1]) or speech intelligibility index (SII; [2]), aim to predict the human performance in speech recognition tasks in noise from macroscopic signal properties like the frequency-weighted signal-to-noise ratio (SNR). In more recent developments, microscopic models were employed to perform speech recognition experiments [3, 4, 5], from which the outcome of speech recognition experiments with human listeners, such as the speech reception threshold (SRT), is then statistically derived. Most macroscopic and microscopic models rely on empirical reference data for calibration or make strong assumptions about the a-priori knowledge to the extent that models have knowledge about the exact temporal alignment of the to-be-recognized signals, sometimes referred to as "frozen speech" or "frozen noise". Both calibration, e.g., offset compensation, and "frozen" signal approaches result in the unfortunate situation where the models perform different tasks than the human listeners, or perform the task at a different (usually higher) SNR. This lack of parallelism can be expected to eventually limit the models' ability to generalize to other noise conditions, speech materials or languages.

In the current study, a framework for auditory discrimination experiments [2,8] which uses an automatic speech recognition (ASR) system and that neither requires calibration with empirical data nor the temporal alignment of the to-be recognized signal was used as a microscopic model to simulate and hence predict the outcome of the matrix test across several languages and noise conditions as empirically determined in [6]. FADE was shown to accurately predict speech intelligibility of the German matrix test in different stationary noise conditions [7]. Its scope was then successfully extended to a fluctuating noise conditions and even basic psychoacoustical experiments [8].

The matrix sentence test was developed for several languages in order to make speech recognition measurements as comparable across languages as possible. The first matrix test was proposed in [9] for Swedish and modified in [10, 11, 12] for German. Further matrix tests were developed for—at the time of writing—14 languages including Polish [13], Russian [14], Spanish [15], and other languages (see overview in [16]). Each matrix test consists of a 50-word base matrix which is used to generate semantically unpredictable and grammatically fixed sentences, like "Peter sees eight wet chairs" in the English version. The close-set structure and limited speech material makes this type of test particularly suitable for the use with ASR-based models, such as FADE.

In this work, FADE was evaluated for speech intelligibility predictions of matrix tests in several languages including German, Polish, Russian, and Spanish and in stationary as well as modulated noise conditions. The predictions were compared to those of the extended speech intelligibility index (eSII; [17, 18]) which was proposed to extend the standard SII for predictions in modulated noise conditions. One of the main advantages of the FADE approach over many other approaches to speech intelligibility prediction is that it performs the same task as the human listeners, i.e. recognition of speech signals at the same SNR without knowledge about the temporal alignment. The hypothesis is that this parallelism aids the models ability to generalize. Two different front-ends are used with FADE to simulate the speech recognition experiments and hence predict their outcome. The first one, commonly used in ASR, extracts Mel-frequency cepstral coefficients (MFCCs) with their first and second order temporal derivatives, in the implementation from [19]. The second one uses separate spectral and temporal modulation filter banks of Gabor filters to extract robust auditory-inspired features called separable Gabor filter bank (SGBFB) features, which were reported to improve the robustness of ASR system compared to MFCCs [19].

http://dx.doi.org/10.21437/Interspeech.2016-1119

2. Methods

2.1. Multilingual matrix sentence test

For each language, the 50-word base matrix consists of ten alternatives per each of the five word classes (name, verb, numeral, adjective, and object). The sentences are constructed by choosing one of the ten alternatives per word class. The German and Polish tests were recorded with a male speaker and the Russian and Spanish tests with a female speaker. All matrix tests were optimized for speech intelligibility measurements in noise.
The empirical data were taken from [6], which provides a detailed description of the measurement procedure. Speech reception thresholds (SRT) defined as speech-to-noise ratio (SNR) yielding 50%-word-correct performance, were measured with native listeners for each language (i.e., German, Polish, Russian, and Spanish). All listeners were normal-hearing with pure tone thresholds not exceeding 20 dB HL for each of the octave frequencies between 125 and 8000 Hz. The signals were presented monaurally over headphones (Sennheiser HDA200) to the listener's preferred ear. To measure the SRT, the noise level was fixed at 65 dB SPL while the speech level was varied adaptively for converging to the SRT [20]. The task of the subjects was to indicate the words she/he understood on a touch screen which displayed all 50 words of the matrix test.
Of the noise signals used in [6], a subset was examined in this study including a test-specific speech-shaped noise (TSN) generated from the speech material of the respective test, four standardized ICRA noises, and a multitalker noise. Regarding the ICRA noises, a stationary, speech-shaped noise with male (ICRA1m) and female (ICRA1f) characteristics and spectrally female- and male-shaped fluctuating noises (termed ICRA4f-250 and ICRA5m-250, respectively) were considered [21]. The fluctuating ICRA noises mimic the envelope of a single talker with pause durations limited to 250 ms. The multitalker (MT) babble was generated from the recordings of 12 female and 8 male competing talkers reading different English passages (c.f. [6] for details). This noise was considered to be stationary due to the synchrony of all talkers.

2.2. Extended speech intelligibility index predictions

To calculate the extended speech intelligibility index (eSII), the signals were analyzed in short time frames of 1024 samples at 44100 Hz sampling rate and a frame shift of half of the frame length. These parameters were set based on the findings in [18, 22]. The resulting eSII was transformed into an intelligibility value using a nonlinear transform derived from a mapping function for sentence intelligibility (cf. [23, Table III, Fig. 7]). The SRT for a given condition was calculated by selecting a fixed reference eSII value and varying the SNR until the eSII equals this reference value. A value of 0.26 was used as the reference, which corresponds to the SRT of the German matrix test in the test-specific noise condition. The test-specific noises (TSN) were used as speech input for the eSII model as proposed by [22].

2.3. FADE predictions

The simulation framework for auditory discrimination experiments (FADE; [8]) was used to simulate the outcome of the matrix test in different languages and various noise conditions (c.f. [7] for details). For each language and noise condition separately, simple ASR systems using Gaussian Mixture Models (GMMs) and Hidden Markov Models (HMMs) were trained and tested on the noisy speech signals over a wide range of SNRs (-24 dB to 6 dB in 3-dB steps). During the training phase, 50 (5 word classes * 10 alternatives) whole-word models with six emitting states and one mixture component were learned for each considered SNR. With each trained system, recognition of the test data at the same SNRs was then performed which resulted in a quadratic recognition result map. This map shows the recognition performance depending on the training an testing SNR (c.f. [7]) and was used to interpolate the lowest achievable SRT which is reported as the predicted (or simulated) SRT. The training and the testing data were generated in the same way, by mixing the speech signals with random portions of the noise signal, but not identical. The details of this modeling approach were explained in [8]. The actual performance and hence the prediction depends, above all, on the speech signal, the noise signal, and the signal representation, i.e. the features.
For the front-end of the ASR system, Mel-frequency cepstral coefficients (MFCCs) and separable Gabor filter bank (SGBFB) features were considered. Here, only a brief summary of the feature extraction steps is provided, a detailed explanation can be found in [19].
MFCCs were calculated by performing a discrete cosine transform (DCT) of the spectral dimension of a logarithmically scaled Mel-spectrogram (LogMS). Only the first 18 DCT-coefficients were used and concatenated with their first and second order discrete temporal derivative in order to form a feature vector, which is referred to as MFCC features. Separable Gabor filter bank (SGBFB) features were extracted from a LogMS by first filtering the spectral dimension with the spectral modulation filters of the SGBFB, and then filtering the output of each spectral filter with the temporal modulation filters of the SGBFB. The spectral modulation filters had center frequencies of 0.000, 0.029, 0.060, 0.122, and 0.250 cycles/Mel-Band, the temporal modulation filters of 0.0, 6.2, 9.9, 15.7, and 25.0 Hz. The LogMS was calculated as follows. The linear frequency axis of an amplitude spectrogram with a window length of 25ms and a windows shift of 10ms was transformed into a Mel-frequency axis by integrating the frequency bins from 64 to 8000 Hz into 31 equally-spaced Mel-bands. Subsequently, the amplitude values were compressed with the decade logarithm. MFCC as well as SGBFB feature vectors were normalized using mean-and-variance normalization on a per-utterance basis.

3. Results

The empirical SRTs measured in [6] and the predicted SRTs using the eSII and FADE with both front-ends are reported in Table 1.

3.1. Empirical data

The empirical SRTs from [6] in Table 1 range from about -26 dB SNR for the Russian matrix test in the fluctuating noise conditions to about -4 dB for the Spanish matrix test in the multitalker babbel noise condition. Independently of the language, the multitalker babble noise resulted in the highest SRTs, and one of the fluctuating noise conditions (ICRA4/5) in the lowest SRTs, while the SRTs in stationary noise conditions lie in between. In general, the SRTs of the Polish test were lower than the corresponding SRTs of the Spanish and the German test, and higher than for the Russian test. Despite the different sex of the speakers (male for the Polish and German test, female for the Spanish and Russian), the SRTs for the male and female version of the ICRA noises resulted in very similar SRTs (absolute differences below 2 dB).

	Noise	DE	PL	RU	ES
Empirical	TSN	-7.2	-9.4	-10.2	-7.4
	ICRA1f	-8.0	-10.5	-13.4	-6.7
	ICRA1m	-7.4	-10.4	-13.9	-7.1
	ICRA4f-250	-20.9	-23.6	-26.1	-16.9
	ICRA5m-250	-19.3	-23.0	-26.3	-17.8
	multitalker	-6.2	-9.2	-9.5	-3.9
eSII	TSN	-7.2	-7.2	-7.2	-7.2
	ICRA1f	-5.8	-6.9	-6.4	-4.6
	ICRA1m	-5.4	-6.5	-6.5	-4.7
	ICRA4f-250	-22.4	-20.3	-18.8	-20.1
	ICRA5m-250	-15.7	-15.2	-16.5	-16.2
	multitalker	-7.6	-8.6	-8.2	-6.3
$\text{FADE}_{\text{MFCC}}$	TSN	-7.8	-10.1	-8.3	-8.4
	ICRA1f	-8.3	-10.7	-11.3	-7.7
	ICRA1m	-7.7	-10.6	-12.4	-8.2
	ICRA4f-250	-15.6	-16.1	-19.1	-15.1
	ICRA5m-250	-15.5	-15.8	-18.6	-12.8
	multitalker	-5.2	-7.4	-7.5	-3.9
$\text{FADE}_{\text{SGBFB}}$	TSN	-7.9	-10.8	-12.7	-9.4
	ICRA1f	-9.4	-12.4	-14.0	-9.7
	ICRA1m	-9.4	-12.7	-14.5	-10.1
	ICRA4f-250	-19.4	-21.0	-23.7	-17.3
	ICRA5m-250	-18.2	-21.1	-22.7	-16.5
	multitalker	-5.1	-7.3	-7.7	-3.7

Table 1: *Empirical SRTs and predictions with the eSII and FADE (with traditional MFCC and robust SGBFB front-end) for different noise conditions and languages. The empirical reference for the eSII was the SRT achieved for TSN in German, which resulted in a value of 0.26.*

3.2. Model predictions

The eSII predicts SRTs in the range from about -22 dB for the German test in the female version of the fluctuating noise and about -5 dB for the Spanish text in the female version of the stationary noise (ICRA1f). These conditions do not coincide with the respective conditions of the lowest and highest empirical SRTs. The FADE-based predictions depend on the front-end and range from about -19 dB and -23 dB for the Russian test in the fluctuating noise condition with MFCCs and SGBFB, respectively, to about -4 dB for the Spanish matrix test in the multitalker babbel noise condition for both front-ends. These conditions coincide with the respective conditions of the lowest and highest empirical SRTs for both front-ends.

In line with the empirical data, all models predict the lowest SRTs in the fluctuating noise conditions, independently of the language. While FADE-based predictions show the highest SRTs in the multitalker condition in line with the empirical data, the eSII predicts higher SRTs for the stationary ICRA noises than for the multitalker noise.

The FADE-based models correctly predict the SRTs of the Polish test to be lower than the corresponding SRTs of the German and the Spanish test, and also that the lowest SRTs in each noise condition is achieved with the Russian test. The predictions with the eSII show different language-dependent pattern. It predicts only a minor language-dependence (below 2 dB) for the stationary noise conditions, whereas the empirical data reach a difference of 6.5 dB and it doesn't follow the empirically found order of the languages for the fluctuating noise conditions (ES>DE>PL>RU).

Furthermore, the eSII predictions show large differences (2 dB to 7 dB) between the male and female version of the fluctuating noises. In contrast to the eSII and in line with the empirical data, FADE-based predictions show only small differences between the male and female versions of the four ICRA noises.

Figure 1 depicts the predicted versus empirical SRTs, where data points on the diagonal indicate perfect predictions. The

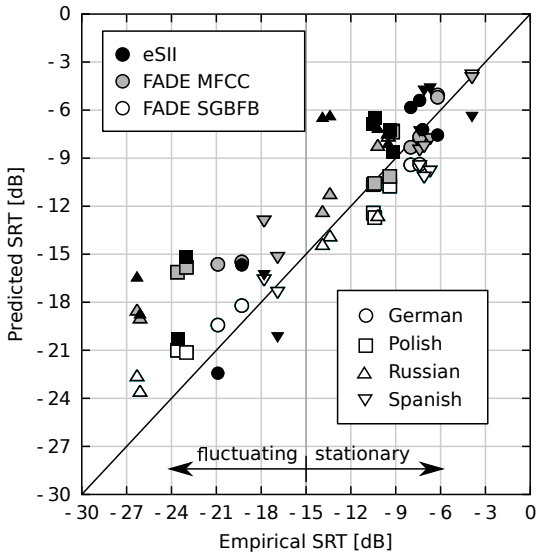

Figure 1: *Scatter plot of predicted SRTs against measured SRTs for the eSII (filled black symbols) and FADE with traditional MFCC (filled gray symbols) or robust SGBFB (open symbols) front-end. The solid line is the bisecting line. The predictions for different languages are indicated by the symbol: German (circle), Polish (square), Russian (upward triangle), Spanish (downward triangle). The conditions with empirical SRTs below -15 dB SNR correspond to fluctuating noise conditions, and above -15 dB SNR to stationary noise conditions.*

data points of the eSII predictions for the stationary noise conditions on the right hand side are rather horizontally aligned, while the FADE-based predictions are mostly aligned to the diagonal. For the fluctuating noise conditions on the left hand side, the data points of the eSII predictions seem to be randomly scattered without any orientation while data points of the FADE-based predictions show a diagonal orientation. However, the predictions with the SGBFB front-end were closer to the diagonal.

The accuracy of the model predictions is assessed by a correlation analyses. Pearson's correlation coefficients (R^2) between predicted and empirical data, including their 95% confidence intervals according to [24], are reported in Table 2 for the stationary, the fluctuating, and all noise conditions along with the probability (p) of the null-hypothesis (no correlation), the root-mean-square (RMS) prediction error, and the bias (B). Over all languages and noise conditions, the eSII predictions show a significantly lower correlation with the empirical data (R^2 = 0.76) than the FADE-based predictions (R^2 = 0.94) and resulted in a higher bias and RMS prediction errors. Analyzing the stationary and fluctuating noise conditions separately, the eSII predictions show no significant correlation with the empirical data, while the FADE-based predictions show significant correlation coefficients exceeding R^2 = 0.70.

Comparing the two front-ends that were used with the FADE model, robust SGBFB features resulted in better predictions for fluctuating noise conditions (R^2 = 0.96 vs. R^2 = 0.77) with a lower bias (1.7 dB vs 5.7 dB) and RMS prediction error (2.1 dB vs. 5.6 dB). Across all noise conditions and languages, the

Model	Cond.	R^2	Interval	p	B[dB]	RMS[dB]
eSII	fluc.	0.00	[0.00 0.53]	0.90	3.6	5.6
$FADE_{MFCC}$	fluc.	0.77	[0.20 0.96]	< 0.01	5.7	6.0
$FADE_{SGBFB}$	fluc.	0.96	[0.80 0.99]	< 0.01	1.7	2.1
eSII	stat.	0.04	[0.00 0.41]	0.43	2.1	3.3
$FADE_{MFCC}$	stat.	0.79	[0.50 0.92]	< 0.01	0.3	1.2
$FADE_{SGBFB}$	stat.	0.73	[0.38 0.90]	< 0.01	-1.0	1.8
eSII	all	0.76	[0.53 0.89]	< 0.01	2.6	4.2
$FADE_{MFCC}$	all	0.94	[0.86 0.97]	< 0.01	2.1	3.6
$FADE_{SGBFB}$	all	0.94	[0.86 0.97]	< 0.01	-0.1	1.9

Table 2: *Statistical analysis of the predicted Speech Recognition Thresholds (SRTs) for the group of stationary, fluctuating and all noise conditions. Pearson's correlation coefficients (R^2) are reported (including their 95% confidence intervals according to [24]) along with the probability (p) of the null-hypothesis (no correlation), the root-mean-square (RMS) prediction error, and the bias (B) for the eSII-based and the FADE-based predictions. Superscript t indicates the used of the traditional MFCC front-end and superscript r the use of the robust SGBFB front-end.*

FADE model using SGBFB features yields the lowest bias (-0.1 dB) and RMS prediction error (1.9 dB).
To summarize, the eSII predicts well only a very general trend showing lower SRTs for fluctuating noise conditions than for the stationary noise conditions, but it fails to predict most of the empirically observed effects between languages and maskers. The FADE-based predictions, in particular when using the robust SGBFB front-end, are mostly in line with the empirical data.

4. Discussion

It was shown that the microscopic, ASR-based FADE modeling approach outperforms the macroscopic eSII modeling approach in predicting SRTs of matrix tests in noise across languages.
The eSII did not predict language-specific effects. In the test-specific noise (TSN) condition, where the average speech and noise spectrum are matched, the eSII predicted the same SRT for each language under test (cf. Table 1). Further, across languages and the stationary or fluctuating noise conditions no significant correlation with the empirical data was found, which suggests that for the considered class of noises differences between languages can not be explained macroscopically by the average portion of the speech signal that is audible for a listener.
In contrast, FADE predicted the language-specific effects very well. It probably learned that certain portions of the speech signal are more important than others, in contrast to the eSII which uses fixed frequency-dependent weights, which was then correctly reflected in the outcome if these portions of the signal were not available due to a masker. Since FADE performs the same task, i.e., word recognition *and* works on noisy data at the same "working point", i.e. SNR, that human listeners encounter in listening tests, it seems to face similar difficulties and hence performs similarly to human listeners on the considered matrix tests. It is probably this parallelism that enables FADE to perform predictions that are unencumbered by any empirical reference and constitutes its predictive power, i.e., its ability to generalize.
Models that require calibration to empirical data can in the best case predict the differences between the reference and other tested conditions but are not able to explain the speech recognition process in human listeners by itself. Also, generalization of the outcomes and better understanding of human performance is limited when predictions are made with models requiring perfect a-priori knowledge about the signals to be recognized since such information is not available to human listeners.
Comparing the FADE approach with two different ASR front-ends, it was shown that the robust SGBFB front-end is better suited for accurate predictions in fluctuating noise conditions. The threshold predicted with a MFCC front-end can model human performance very well in stationary noise conditions but was not able to benefit from fluctuations in noise to the same extent as human listeners did. Therefore, independent of language, the SGBFB features seem to be a more reasonable model of human auditory processing than MFCC features. However, the predicted SRTs in the fluctuating noise condition using the SGBFB front-end are still slightly above the average performance of human listeners and a more robust front-end could possibly further improve the FADE-based predictions.

4.1. Future work

The FADE approach works on processed signal and therefore might also predict the effect of signal processing algorithms on speech intelligibility, which could be evaluated in future studies. Further, acknowledging the fact that hearing loss might be explained by suboptimal signal processing, the effect of hearing loss on speech intelligibility might be modeled by incorporating typically assumed signal processing deficiencies into the feature extraction of the FADE model.

5. Conclusions

The most important findings of this work can be summarized as follows:

1. The microscopic FADE model with robust SGBFB front-end accurately predicts empirically found language-specific and noise-specific influences on the outcome of the matrix test, which results in a correlation coefficient of R^2 = 0.94, a bias of -0.1 dB and a root-mean-square prediction error below 2 dB.
2. In contrast, the macroscopic eSII model was found to only predict that speech reception thresholds in fluctuating noise conditions were lower than in stationary noise conditions, but neither language-specific differences, nor masker-specific differences within the subgroups of stationary and fluctuating noises.
3. Particularly for the fluctuating noise condition, the use of the physiologically-inspired robust ASR-features (SGBFB) resulted in higher correlations with the empirical data (R^2 = 0.96) and a lower RMS prediction error (2.1 dB) compared to using MFCC features (R^2 = 0.77, RMS = 6.0 dB).
4. FADE, unlike many other speech intelligibility models, does not use an empirical reference which avoids a potential language-bias and allowed to accurately predict the effect of specific speakers/languages on speech recognition.

6. Acknowledgements

This work was supported by the Cluster of Excellence Grant "Hearing4all".

7. References

[1] N. R. French and J. C. Steinberg, "Factors governing the intelligibility of speech sounds," *The Journal of the Acoustical Society of America*, vol. 19, no. 1, pp. 90–119, 1947.

[2] A. ANSI, "S3. 5-1997, methods for the calculation of the speech intelligibility index," *New York: American National Standards Institute*, vol. 19, pp. 90–119, 1997.

[3] I. Holube and B. Kollmeier, "Speech intelligibility prediction in hearing-impaired listeners based on a psychoacoustically motivated perception model," *The Journal of the Acoustical Society of America*, vol. 100, no. 3, pp. 1703–1716, 1996.

[4] J. Barker and M. Cooke, "Modelling speaker intelligibility in noise," *Speech Communication*, vol. 49, no. 5, pp. 402–417, 2007.

[5] T. Jürgens and T. Brand, "Microscopic prediction of speech recognition for listeners with normal hearing in noise using an auditory model)," *The Journal of the Acoustical Society of America*, vol. 126, no. 5, pp. 2635–2648, 2009.

[6] S. Hochmuth, B. Kollmeier, T. Brand, and T. Jürgens, "Influence of noise type on speech reception thresholds across four languages measured with matrix sentence tests," *International Journal of Audiology*, vol. 54, no. sup2, pp. 62–70, 2015.

[7] M. R. Schädler, A. Warzybok, S. Hochmuth, and B. Kollmeier, "Matrix sentence intelligibility prediction using an automatic speech recognition system," *International Journal of Audiology*, vol. 54, no. sup2, pp. 100–107, 2015.

[8] M. R. Schädler, A. Warzybok, S. D. Ewert, and K. Birger, "A simulation framework for auditory discrimination experiments: Revealing the importance of across-frequency processing in speech perception," *Submitted to the Journal of the Acoustical Society of America*, 2015, under review.

[9] B. Hagerman, "Sentences for testing speech intelligibility in noise," *Scandinavian audiology*, vol. 11, no. 2, pp. 79–87, 1982.

[10] K. C. Wagener, V. Kühnel, and B. Kollmeier, "Entwicklung und Evaluation eines Satztests für die deutsche Sprache I: Design des Oldenburger Satztests," *Zeitschrift für Audiologie*, vol. 38(1), pp. 4–15, 1999.

[11] K. C. Wagener, T. Brand, and B. Kollmeier, "Entwicklung und Evaluation eines Satztests für die deutsche Sprache Teil II: Optimierung des Oldenburger Satztests," *Zeitschrift für Audiologie*, vol. 38(2), pp. 44–56, 1999.

[12] K. C. Wagener, T. Brand, and B. Kollmeier, "Entwicklung und Evaluation eines Satztests für die deutsche Sprache Teil III: Evaluation des Oldenburger Satztests," *Zeitschrift für Audiologie*, vol. 38(3), pp. 86–95, 1999.

[13] E. Ozimek, A. Warzybok, and D. Kutzner, "Polish sentence matrix test for speech intelligibility measurement in noise," *International Journal of Audiology*, vol. 49, no. 6, pp. 444–454, 2010.

[14] A. Warzybok, M. Zokoll, N. Wardenga, E. Ozimek, M. Boboshko, and B. Kollmeier, "Development of the Russian matrix sentence test," *International Journal of Audiology*, vol. 54, no. sup2, pp. 35–43, 2015.

[15] S. Hochmuth, T. Brand, M. A. Zokoll, F. Z. Castro, N. Wardenga, and B. Kollmeier, "A Spanish matrix sentence test for assessing speech reception thresholds in noise," *International Journal of Audiology*, vol. 51, no. 7, pp. 536–544, 2012.

[16] B. Kollmeier, A. Warzybok, S. Hochmuth, M. A. Zokoll, V. Uslar, T. Brand, and K. C. Wagener, "The multilingual matrix test: Principles, applications, and comparison across languages: A review," *International Journal of Audiology*, vol. 54, no. sup2, pp. 3–16, 2015.

[17] K. S. Rhebergen and N. J. Versfeld, "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*, vol. 117, no. 4, pp. 2181–2192, 2005.

[18] R. Beutelmann, T. Brand, and B. Kollmeier, "Revision, extension, and evaluation of a binaural speech intelligibility model," *The Journal of the Acoustical Society of America*, vol. 127, no. 4, pp. 2479–2497, 2010.

[19] M. R. Schädler and B. Kollmeier, "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*, vol. 137, no. 4, pp. 2047–2059, 2015.

[20] T. Brand and B. Kollmeier, "Efficient adaptive procedures for threshold and concurrent slope estimates for psychophysics and speech intelligibility tests," *The Journal of the Acoustical Society of America*, vol. 111, no. 6, pp. 2801–2810, 2002.

[21] W. A. Dreschler, H. Verschuure, C. Ludvigsen, and S. Westermann, "Icra noises: Artifical noise signals with speech-like spectral and temporal properties for hearing instrument assessment," *Audiology*, vol. 40, no. 3, p. 148, 2001.

[22] K. S. Rhebergen, N. J. Versfeld, and W. A. Dreschler, "Extended speech intelligibility index for the prediction of the speech reception threshold in fluctuating noise," *The Journal of the Acoustical Society of America*, vol. 120, no. 6, pp. 3988–3997, 2006.

[23] H. Fletcher and R. H. Galt, "The perception of speech and its relation to telephony," *The Journal of the Acoustical Society of America*, vol. 22, no. 2, pp. 89–151, 1950.

[24] W. D. Fisher, "On grouping for maximum homogeneity," *Journal of the American statistical Association*, vol. 53, no. 284, pp. 789–798, 1958.