Speech Intelligibility Prediction with hybrid auditory Model- And ML-based Methods: The Best of two Worlds?

Birger Kollmeier [1], David Hülsmeier& Anna Warzybok

Medizinische Physik & Cluster of Excellence Hearing4All, Universität Oldenburg
[1] birger.kollmeier@uol.de

Summary: This contribution reviews the usage of a hybrid approach to model human speech recognition with an auditory-model-based frontend and a machine learning (ML)-based backend. The Framework for Auditory Discrimination Experiments (FADE [15]), for example, utilizes a physiology-inspired Gabor-Filter feature extraction in combination with an HMM/GMM ASR system as backend. Its performance in comparison to standard procedures is evaluated for predicting the speech recognition threshold (SRT), i.e., the signal-to-noise ratio corresponding to 50% sentence intelligibility for normal and hearing-impaired listeners in various interfering noise conditions. The results highlight the advantage of combining the best of two worlds, i.e., a model-based frontend to allow for an individualizations strategy for the respective perception task given any individual hearing impairment and a ML-based ASR backend as a “generalized optimum detector”.

1 Introduction

While machine learning-based automatic speech recognition systems exhibit convincing performance without providing much explainability, auditory models strive for high interpretability in simulating the human speech or sound recognition process, but achieve a low generalization ability for unknown sound samples. This general statement can also be made for methods predicting human speech recognition (HSR) that have achieved an increased attention recently (for a review see Feng and Chen, 2022).

This contribution reviews the usage of a hybrid approach to model human speech recognition with an auditory-model-motivated frontend and a ML-based backend. The Framework for Auditory Discrimination Experiments (FADE [15]), for example, utilizes a physiology-inspired Gabor-Filter feature extraction in combination with an Gaussian mixture model/hidden Markov model (HMM/GMM) automatic speech recognition (ASR) system as backend. Its performance in comparison to standard procedures is evaluated for predicting the speech recognition threshold (SRT), i.e., the signal-to-noise ratio (SNR) corresponding to 50% sentence intelligibility for normal and hearing-impaired listeners in various interfering noise conditions.

2 Methods

2.1 Human speech recognition

For measuring human speech recognition (HSR), the matrix sentence recognition test is often used (closed-set of 50 words to compose in each list 10 semantically unpredictable 5-word sentences with a fixed syntax) which is available in more than 20 languages [9] and allows for a high measurement accuracy and reproducibility. Note that employing such a closed-set sentence recognition test for HSR appears to be quite limited in its generalization ability to real life due to its limited representation of the whole language. However, the test design warrants that the respective language properties (such as, e.g., phoneme distribution, syllabic structure of test items, all test items included in the typical vocabulary even of school children) is maintained as well as possible [9]. Another disadvantage of a closed-set HSR test structure is the training effect for the first 1-2 sentence test lists employed with the Matrix test which is counterbalanced by the exact comparability of different test lists due to its construction from the same words that appear in different context. The syntactically fixed, but semantically unpredictable structure furthermore enables a repeatability for usage in humans (who can not memorize the individual test sentences over time) which also helps to certify the equivalence of test lists when used with human subjects.

Employing a closed-set HSR test for prediction by ASR further has the advantage, that the respective ASR task is an "easy" job due to its limited reference vocabulary. Hence, when matching between HSR and ASR conditions, no man-machine gap occurs which clearly helps to model human performance. This "recognizing something which is already (nearly) known" is not acceptable as a performance measure for an ASR system. However, this limited-vocabulary approach helps to focus the performance prediction on acoustic properties in HSR rather than on linguistic content or on other cognitive functions for performing the task.

2.2 HSR prediction with FADE

For predicting the SRT, FADE simulates the speech recognition process with an ASR system, i.e., by transforming the incoming signals to a multidimensional pattern in the feature space that roughly resembles the neurosensory transformation of the auditory system. Machine-learning-based ASR techniques (like GMM/HMM speech pattern recognizer utilized in FADE) are then used to recognize the presented speech utterances given the restrictions of the lossy transformation process in an optimum way. Thus, the ASR-based recognition process can be considered as a "generalized optimum detector approach" to model human speech recognition. This provides the chance of simulating the detrimental effect of imprecisions in the signal representation that are intended to emulate the effect of the audiogram and of suprathreshold distortions.

3 Results

3.1 Performance of FADE for predicting SRT with normal listeners

The papers published so far (i.e., [5], [8],[14]-[18]) revealed an excellent prediction of the empirical SRT including measurements in different noise conditions, for different languages, binaural unmasking or assessment of noise reduction algorithms. An example for SRT predictions in different languages is provided in Fig. 1 (from [14]).

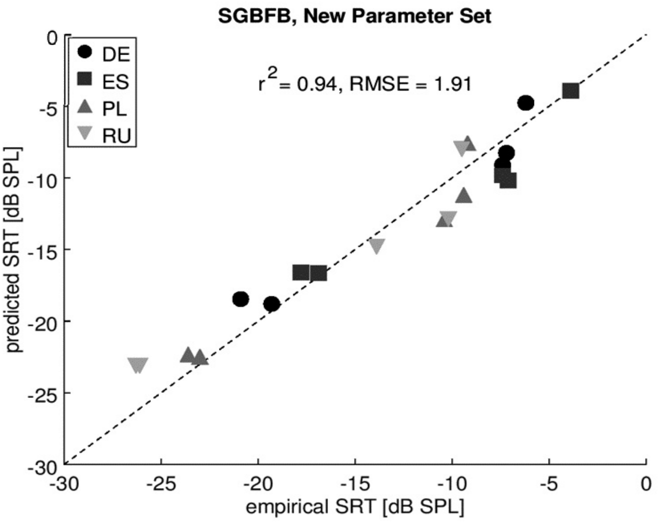

Figure 1 - Measured and predicted SRTs using FADE for normal listeners with different language-specific noise maskers from [14]). The diagonal line indicates a 1-to-1 mapping of the predictions and measured SRTs (i.e., a perfect correlation).

3.2 Prediction for hearing-impaired listeners

An even more challenging modelling task involves the SRT prediction for hearing-impaired listeners that differ in their empirical SRT considerably even if roughly the same pure-tone audiogram is involved, yielding root-mean-square (RMS) errors of about 6.5 dB for a large range of hearing losses. An improved prediction accuracy can be achieved if the model does not only involve the properties of the speech and noise materials employed as well as the individual audiogram, but if some independent psychoacoustic measures of suprathreshold distortions are used: Fig. 2 shows prediction results (from Hülsmeier & Kollmeier, submitted) where the RMSEs were about 3.3 dB which is close to test-retest reliability of the tests used in the empirical measurements (RMSEs of about 1.4 – 2.8 dB dependent of the masking condition used [22]).

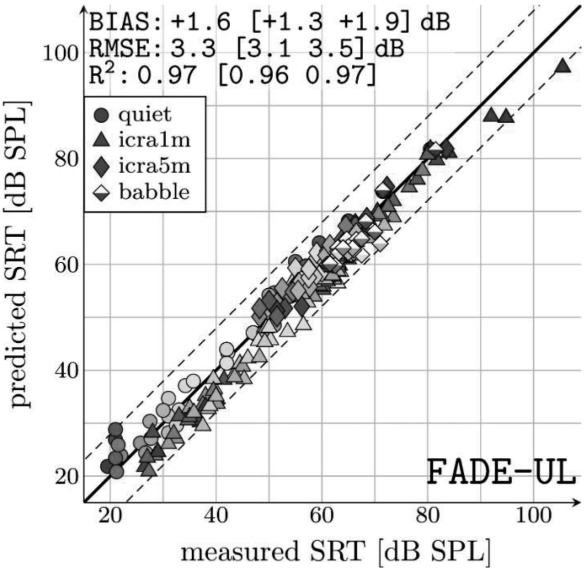

Figure 2 - Measured and predicted SRTs using FADE for listeners with different degree of hearing loss utilizing an exactly measured audiogram and a tone-in-noise detection threshold for parameter individualization. The dashed diagonal lines indicate the 95th percentile of the prediction errors found by Schädler et al. (2020, Tab. 1)

3.3 Comparison with other predictions measures

In the speech processing literature, index-based measures like STI [6], SII [1] or the STOI measure [19] are often treated as the "gold standard" for speech intelligibility prediction (SIP) even though in the HSR literature it is well known that other measures perform better (e.g., [11],[2]), especially in fluctuating noise and reverberation. It should also be mentioned (as described in [3]) that the concept of what later was called STOI has been described before independently by Ludvigsen et al (1990, [10]) and Kollmeier (1990, [7]) and was already used and described, e.g. by Holube and Kollmeier (1996, [4]). An in-depth-review of STOI and related issues has been recently given by Feng and Chen (2022, [2]). Unfortunately, a direct comparison of FADE with STOI and other intrusive instrumental speech intelligibility measures is difficult because the input quantities differ considerably: While FADE requires the mixed speech and noise signal for a number of predetermined signal-to-noise ratios, STOI (as well as other index-based measures like AI, SII or STI) requires clean speech as reference signal. Nevertheless, in those instances where a direct comparison was employed (e.g., [15], [20], [5]), the FADE approach or a comparable approach by [12] appear advantageous for the following reasons:

- Few assumptions about SNR estimation in each frequency band: While index-based measures either derive their SNR estimate directly from previous knowledge of the clean speech and noise signal, the ASR-based methods are trained to optimally cope with the "effective" speech feature deterioration at a given SNR. Note that the a priori assumptions about the SNR determination may not apply in certain conditions (e.g., in fluctuating noise where the modulation transmitted in the output signal are not necessarily due to speech envelope-based modulations which is, e.g., presumed for STOI.)

- Optimized weighting of frequency bands: While index-based measures employ a fixed, rule-based weighting of the "effective" SNR across frequency bands, ASR-based measures utilize the available information in each speech band in an optimum way which is trained with the help of the reference material with predefined SNR values.

As a consequence, ASR-based predictors of HSR have certain advantages. This should at least lead to a balanced discussion about the sometimes misleading and uncritical usage of STOI and other index-based measure in the speech processing literature.

4 Discussion

4.1 Performance of FADE

The experimental SRT outcomes for various noise masker conditions for each subject can be predicted with a high accuracy using FADE with an appropriate individualization strategy: The highest prediction accuracy for the individual SRT (coefficient of determination R2 = 0.97, RMSE = 3.3 dB) is achieved if a precise measure of the individual audiogram is used and if the individual suprathreshold processing deficit is determined, e.g. from individual tone-in-noise detection thresholds. These results highlight the advantage of combining the best of two worlds, i.e., a model-based frontend to allow for an individualizations strategy for hearing impairment and a ML-based ASR backend as a “generalized optimum detector”.

4.2 Comparison HSR with ASR

How to compare human speech recognition with machine-learning-based automatic speech recognition (ASR)? The test and tasks employed for HSR are limited in their generalization ability to real life. This also applies to any test corpus which an ASR system is used for. Using the same speech material for HSR as for ASR (for example using the matrix test or logatome databases like the OLLO, Wesker et al., 2005) has a clear advantage over non-structured comparisons.

Mapping function (of HSR versus ASR) versus matched training: which approach is better? FADE as well as related approaches employ matched training (using, e.g., the matrix test) which has the advantage, that no HSR-ASR-gap occurs. Other open-set prediction systems most often encounter the HSR-ASR gap and hence need a mapping function relating ASR and HSR to each other. It may be expected that an approach like FADE that needs no mapping function has a higher face validity and generalization ability to other comparisons as well than approaches that require a mapping function which might markedly differ for yet unseen tasks. However, this is still open and needs to be explored.

What to learn from comparisons between HSR and ASR? Possible issues might be the improvement of auditory modeling (learning from ASR towards better understanding of HSR) as opposed to improving ASR systems based on better understanding of HSR, e.g., by including auditory-model-based speech features in ASR. It is quite open, however, how the still existing gap between HSR and ASR does influence this field.

4.3 Interpretability vs. complexity

Auditory-model-motivated front ends have the advantage of incorporating knowledge which helps to improve the interpretability of the results, especially if different front-end features produce different ASR results. End-to-end ASR solutions – employing quite the opposite approach - do not provide any interpretable intermediate steps and require a larger amount of training data to train properties of speech processing and feature extraction that is already known from expert knowledge. Hence, it remains to be seen if the promising approach of end-to-end HSR systems provide any advantage for the purpose of speech intelligibility prediction for humans.

Knowledge-based versus data-driven approach: Modeling/predicting HSR through ASR techniques have the advantage over expert-knowledge-driven approaches that they can automatically select the "best features" to be used for the respective recognition task. This is perhaps the strongest advantage of ASR-based SIP-methods over "conventional" SIP methods (like STI or STOI). Hence, this point has to be taken into consideration when combining two worlds, i.e., take knowledge-driven front ends and machine-learning driven backends to build interpretable ASR systems that at least have the advantage of modelling human performance with great accuracy.

Supported by Deutsche Forschungsgemeinschaft, Cluster of Excellence Grant EXC2177 "Hearing4all" (Project ID 390895286)

References

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

[2] FENG, Y. and CHEN, F. (2022) "*Nonintrusive objective measurement of speech intelligibility: A review of methodology". Biomedical Signal Processing and Control,* Vol 71, Part B, https://doi.org/10.1016/j.bspc.2021.103204

[3] GOLDSWORTHY and GREENBERG (2004) *Analysis of speech-based speech transmission index methods with implications for nonlinear operations. The Journal of the Acoustical Society of America* 116, 3679. https://doi.org/10.1121/1.1804628

[4] HOLUBE, I and KOLLMEIER, B. (1996) *"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. https://doi.org/10.1121/1.417354

[5] HÜLSMEIER, D., M. BUHL, N. WARDENGA, A. WARZYBOK, M. R. SCHÄDLER, and B. KOLLMEIER (2021). *"Inference of the distortion component of hearing impairment from speech recognition by predicting the effect of the attenuation component".* In: *International Journal of Audiology* 0.0. PMID: 34081564, pp. 1–15. doi: 10.1080/14992027.2021.1929515.

[6] IEC (1998) *Sound system equipment - Part 16: Objective rating of speech intelligibility by speech transmission index*, Geneva, Switzerland: IEC

[7] KOLLMEIER, B. (1990). *Meßmethodik, Modellierung und Verbesserung der Verständlichkeit von Sprache* (Habilitation dissertation, Universität Göttingen).

[8] KOLLMEIER, B., M. R. SCHÄDLER, A. WARZYBOK, B. T. MEYER, and T. BRAND (Sept. 2016). *"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".* In: *Trends in Hearing* 20.0, p. 2331216516655795. doi: 10.1177/2331216516655795

[9] KOLLMEIER, B., A. WARZYBOK, S. HOCHMUTH, M. A. ZOKOLL, V. USLAR, T. BRAND, and K. C. WAGENER (2015). *"The multilingual matrix test: Principles, applications, and comparison across languages: A review".* In: *International Journal of Audiology* 54.sup2, pp. 3–16. doi: 10.3109/14992027.2015.1020971.

[10] LUDVIGSEN, C., ELBERLING, C., KEIDSER, G., and POULSEN, T. (1990). *Prediction of intelligibility of non-linearly processed speech. Acta oto-laryngologica,* 109(sup469), 190-195.

[11] RELAÑO-IBORRA, H., MAY, T., ZAAR, J., SCHEIDIGER, C., and DAU, T. (2016). *Predicting speech intelligibility based on a correlation metric in the envelope power spectrum domain. The Journal of the Acoustical Society of America,* 140(4), 2670-2679.

[12] ROßBACH, J., RÖTTGES, S., HAUTH, C. F., BRAND, T., and MEYER, B. T. (2021, June). *Non-Intrusive Binaural Prediction of Speech Intelligibility Based on Phoneme Classification.*

In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 396-400). IEEE.

[13] SCHÄDLER, M. R. and B. KOLLMEIER (Apr. 2015). "*Separable spectro-temporal Gabor filter bank features: Reducing the complexity of robust features for automatic speech recognition". In*: *The Journal of the Acoustical Society of America* 137.4, pp. 2047–2059. doi: 10.1121/1.4916618.

[14] SCHÄDLER, M. R., WARZYBOK, A., EWERT, S. D., and KOLLMEIER, B. (2016). *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), 2708-2722.

[15] SCHÄDLER, M. R., A. WARZYBOK, and B. KOLLMEIER (2018). *"Objective Prediction of Hearing Aid Benefit Across Listener Groups Using Machine Learning: Speech Recognition Performance With Binaural Noise-Reduction Algorithms".* In: *Trends in Hearing* 22, p. 2331216518768954. doi: 10.1177/2331216518768954

[16] SCHÄDLER, M. R., D. HÜLSMEIER, A. WARZYBOK, and B. KOLLMEIER (2020). "*Individual Aided Speech-Recognition Performance and Predictions of Benefit for Listeners With Impaired Hearing Employing FADE".* In: *Trends in Hearing* 24, p. 2331216520938929. doi: 10.1177/2331216520938929.

[17] SCHÄDLER, M.R., HÜLSMEIER, D., WARZYBOK, A., HOCHMUTH, S., and KOLLMEIER, B. (2016) *Microscopic Multilingual Matrix Test Predictions Using an ASR-Based Speech Recognition Model. Proc. Interspeech* 2016, 610-614, doi: 10.21437/Interspeech.2016-1119

[18] SCHÄDLER, M.R., KRANZUSCH, P., HAUTH, C., and WARZYBOK, A. (2020). *Simulating spatial speech recognition performance with an automatic-speech-recognition-based model. Proc. DAGA* 2020, 908-911.

[19] TAAL, C. H., HENDRIKS, R. C., HEUSDENS, R., and JENSEN, J. (2011) *An evaluation of objective measures for intelligibility prediction of time-frequency weighted noisy speech. The Journal of the Acoustical Society of America* 130(5): 3013–3027. doi:10.1121/1.3641373.

[20] VÖLKER C, WARZYBOK A, and ERNST SM. (2015) *Comparing Binaural Pre-processing Strategies III: Speech Intelligibility of Normal-Hearing and Hearing-Impaired Listeners. Trends Hear.* 2015 Dec 30;19:2331216515618609. doi: 10.1177/2331216515618609. PMID: 26721922; PMCID: PMC4771033.

[21] HÜLSMEIER, D., HAUTH, CH., RÖTTGES, S., KRANZUSCH, P., ROßBACH, J., SCHÄDLER, M.R., MEYER, B.T., WARZYBOK, A., and BRAND, T. (2021) *Towards non-intrusive prediction of speech recognition thresholds in binaural conditions.* In the *proceedings of ITG Tagung*. Kiel, 29 September- 1 October 2021.

[22] WAGENER, K. C., BRAND, T., and KOLLMEIER, B. (2006). *The role of silent intervals for sentence intelligibility in fluctuating noise in hearing-impaired listeners. International Journal of Audiology,* 45(1), 26–33. https://doi.org/10.1080/14992020500243851

[23] WESKER, T., MEYER, B., WAGENER, K., ANEMÜLLER, J., MERTINS, A., and KOLLMEIER, B. (2005). *Oldenburg logatome speech corpus (OLLO) for speech recognition experiments with humans and machines.* In *Ninth European Conference on Speech Communication and Technology.*