INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Listening in the dips: Comparing relevant features for speech recognition in humans and machines

Constantin Spille and Bernd T. Meyer

Medizinische Physik and Cluster of Excellence Hearing4all,
Carl von Ossietzky Universität, Oldenburg, 26129, Germany

constantin.spille@uni-oldenburg.de, bernd.meyer@uni-oldenburg.de

Abstract

In recent years, automatic speech recognition (ASR) systems gradually decreased (and for some tasks closed) the gap between human and automatic speech recognition. However, it is unclear if similar performance implies humans and ASR systems to rely on similar signal cues. In the current study, ASR and HSR are compared using speech material from a matrix sentence test mixed with either a stationary speech-shaped noise (SSN) or amplitude-modulated SSN. Recognition performance of HSR and ASR is measured in term of the speech recognition threshold (SRT), i.e., the signal-to-noise ratio with 50% recognition rate and by comparing psychometric functions. ASR results are obtained with matched-trained DNN-based systems that use FBank features as input and compared to results obtained from eight normal-hearing listeners and two established models of speech intelligibility. For both maskers, HSR and ASR achieve similar SRTs with an average deviation of only 0.4 dB. A relevance propagation algorithm is applied to identify features relevant for ASR. The analysis shows that relevant features coincide either with spectral peaks of the speech signal or with dips of the noise masker, indicating that similar cues are important in HSR and ASR.

Index Terms: man-machine comparison, deep neural networks, automatic speech recognition, relevance propagation

1. Introduction

The capability of human listeners to recognize speech in extremely adverse conditions motivated comparisons between automatic speech recognition (ASR) and human speech recognition (HSR) with the aim of pinpointing weaknesses in the processing strategies in ASR. Previous studies revealed a significant man-machine gap in terms of recognition performance, which serves another aim of human-machine comparisons, i.e., to answer the question how far we have come in achieving human performance levels with machine listeners. These studies often reported humans and machines to rely on different signal cues.

For instance, experiments with nonsense syllables in additive noise showed a man-machine gap of 10 dB with an ASR system using Mel-frequency cepstral coefficients (MFCCs), i.e. the ASR system achieved the same performance as normal-hearing listeners only if the signal-to-noise ratio (SNR) was increased by 10 dB. High- and low-pass filtering reduced (and for some conditions even eliminated) the gap. The results indicated the voicing feature to be suboptimally exploited in ASR [1]. Similarly, results from another study suggested that ASR should incorporate spectral fine structure and temporal as well as spectro-temporal properties more explicitly than standard spectral features [2].

Different processing strategies in HSR and ASR were observed when comparing speech recognition in modulated vs. stationary noise: While human recognition scores are higher in a modulated masker, ASR performance was lower compared to a stationary speech-shaped masker [3]. Increased intelligibility in modulated maskers is a well-known phenomenon in humans which is often referred to as *listening in the dips* or *glimpsing* [4, 5, 6, 7, 8], where *glimpse* refers to time-frequency regions with a positive local SNR. It was found that it is sufficient for listeners to observe several glimpses of each word to achieve high intelligibility and that intelligibility is highest for modulations around 8-10 Hz [9, 4]. Glimpsing was also investigated in ASR [10, 11] and correlations between the number of glimpses and recognition accuracy was found. All of the previously mentioned studies employed Gaussian mixture models (GMMs) as acoustic models, but did not consider deep neural networks (DNNs), i.e., the current state-of-the-art [12]. The use of DNNs in ASR closed the speech recognition gap in single-channel two-talker scenes with a small vocabulary speech material [13] and for large vocabulary telephone conversations, in which both an elaborate ASR system as well as human transcribers achieved an error rate of 6% [14].

In this study, we build upon this progress to perform a comparison of ASR and HSR with a focus on modulated and unmodulated noise types because these were reported to be problematic for ASR earlier. Since the speech task is relatively simple (with a vocabulary of only 50 words), ASR could potentially reach similar performance as HSR, i.e., it could serve as a model for speech intelligibility. To evaluate our ASR system as such a model, results are compared to predictions from two established models of intelligibility. Further, we investigate if DNN-processing allows to perform listening-in-the-dips or glimpsing, which could explain the advances attributed to deep learning, and compare the result to word recognition scores from eight normal-hearing listeners. For the analysis, a relevance algorithm is chosen that identifies the time-frequency regions that are most relevant to obtain the classification result. This layer-wise relevance propagation was suggested before to analyze results in image recognition [15] and electroencephalography [16], and is applied in a speech-context in our work. This paper therefore presents data and methods for comparing for the first time HSR and DNN-based ASR on a sublexical level.

In the following, we first describe the two noise types and how they were generated, the speech material used for listening experiments and ASR, as well as the ASR system and the models for predicting speech intelligibility. After a description of the relevance algorithm, we present the average results in terms of the speech reception threshold (SRT), which is the noise level at 50% recognition rate and accuracy-over-SNR curves (*psychometric functions*, Section 3). We then continue to discuss and summarize our findings.

http://dx.doi.org/10.21437/Interspeech.2017-1168

2. Methods

2.1. Maskers

In total, four maskers with different spectro-temporal modulations were adopted from a study by Schubotz et al. [17]. Two maskers are based on a speech-shaped noise (SSN) derived from the International Speech Test Signal (ISTS, [18]), which have the same long-term spectrum as the ISTS signal. The first masker is a stationary speech-shaped noise (SSN), the second is a sinusoidally modulated version of the SSN with a modulation frequency of 8 Hz (SAM-SNN). Since the original ISTS was produced by female speakers, its spectrum (and thus also the spectrum of the SSN and SAM-SSN) matches the long-term spectrum of female speech. Additionally, a male version of the ISTS was generated in [17], which serves as a basis for a male version of the SSN and SAM-SSN.

2.2. Speech material

The speech data used in these experiments originates from a German matrix sentence test, the Oldenburg sentence test (OLSA), which consists of 120 sentences produced by one speaker [19]. The speech material has a fixed syntactical structure: Each sentence contains five words with 10 possible response choices for each word category and a syntax that follows the pattern <name><verb ><number><adjective><object>, which results in a vocabulary size of 50 words. By using this fixed grammatical structure, the focus of the comparison is laid on the sub-lexical level: Both in HSR and ASR experiments, the fixed structure is known to the ASR system/the listener, so that recognition performances does not depend on language model effects. A review of this matrix test and related tests in different languages is presented in [20].

For training the ASR system, a speech corpus of 10 hours of speech from 20 different speakers (10 male, 10 female) was used that exhibits the syntactical structure of the OLSA [21]. A training set of 80 h was created for both maskers by mixing the clean speech signals with random parts of the respective noise at random uniformly distributed SNRs ranging from -10 dB to 20 dB. The 80 h match the amount of training data in commonly used ASR corpora, such as the Wall Street Journal corpus ([22]).

For each of the four maskers, a test set was created by selecting 400 random SNRs between -40 and 20 dB and mixing eight sentences at each of the SNRs, i.e. in total there are 3200 test sentences per masker.

2.3. Speech intelligibility models

Two different established models of speech intelligibility are used as a baseline of intelligibility prediction in the maskers under consideration. The first model is the extended speech intelligibility index (ESII, [23]). Signals are split into 21 frequency bands and grouped into time bins ranging from 35 ms duration in the lowest frequency band to 9.4 ms in the highest frequency band. The SNR is then calculated in all frequency bands and time frames; the average SNR is the resulting ESII score.

The second model is the the multi-resolution speech envelope power spectrum model (mr-EPSM, [24]), which incorporates a modulation filter bank that extracts modulations from 1 hz to 256 Hz in 22 frequency bands independently. Modulation filterbank outputs are split in time windows with a modulation filter-dependent duration, i.e., shorter temporal windows are used for higher modulation frequencies. The SNR of the Hilbert envelopes is then calculated in all time windows for all modulation filters in all frequency channels.

For both models, the output of the models have to be calibrated to an already measured human intelligibility. In this case the models were calibrated with the SRT of the male and female SSN for predictions of the male and female SAM-SSN, respectively.

2.4. ASR system

The ASR system was implemented using the Kaldi speech recognition toolkit [25]. The speech recognition system is a deep neural network (DNN) with five hidden layers, 2048 units per layer and an additional softmax output layer. The DNN was trained as a stack of restricted Boltzmann machines with an unsupervised pre-training and a supervised fine-tuning of the parameters with triphone targets. Every phone was modeled with three Hidden-Markov-Model (HMM) states except for the silence phone which was modeled with five states.

Input features are calculated by converting the time signals into a log Mel spectrogram with 23 Mel filters and a window length of 25 ms and a 10 ms shift. These FBank features were spliced with a temporal context of ±5 frames, resulting in 253 features per input frame.

SRTs and psychometric functions are obtained by decoding all 3200 (400 SNRs × 8) sentences and averaging word error rates (WERs) over all eight sentences at the same SNR. A psychometric function from [19] is fitted to these 400 (SNR,WER)-pairs to obtain the SRT and its slope.

2.5. Layer-wise relevance propagation

To identify input features that are relevant for a specific classification result, an algorithm for layer-wise relevance backpropagation is used [15]. The goal is to assign a relevance value R for the classification result to each input feature component, e.g., to each time-frequency point for FBank input. This is done by starting at the DNN output and giving the output neuron j that encodes the current HMM state a relevance $R_j^{(l+1)} = 1$, where l is the layer index. This relevance is then propagated to all neurons i in the layer below connected to the output neuron. The sum over the relevance of all neurons i is equal to the relevance of this specific neuron j, i.e. the conservation property $\sum_i R_{i,j}^{(l,l+1)} = R_j^{(l+1)}$ holds. In particular, relevance is propagated using the weights and biases obtained during the training stage by the following formulas

$$R_{i,j}^{(l,l+1)} = \begin{cases} \frac{z_{ij}}{z_j+\epsilon} \cdot R_j^{(l+1)} & , \ z_j \geq 0 \\ \frac{z_{ij}}{z_j-\epsilon} \cdot R_j^{(l+1)} & , \ z_j \leq 0 \end{cases} \tag{1}$$

with

$$z_{ij} = x_i w_{ij} \ , \tag{2}$$

$$z_j = \sum_i z_{ij} + b_j \ , \tag{3}$$

$$x_j = g(z_j) \ , \tag{4}$$

where w_{ij} is the weight connecting neurons i and j, b_j is the bias term of neuron j, g is the activation function and $\epsilon \geq 0$ is a predefined stabilizer which was set to 100 (see [15] for more details).

For each time frame, Equation 1 can be used to calculate the relevance of all input features, i.e., spliced FBank features in our case.

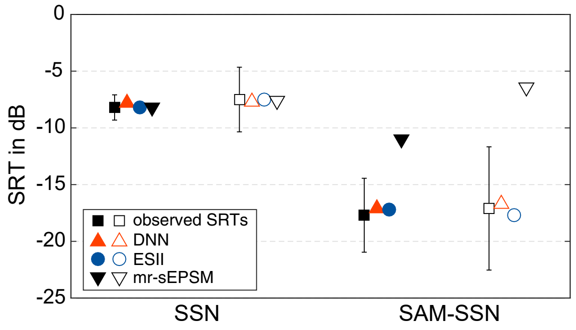

Figure 1: *SRTs of human listeners (black squares) with 95 % confidence intervals (CI), ESII (circles), mr-sEPSM (downward triangles) and DNN-based ASR systems (upward triangles) in both masker conditions. Filled and open symbols refer to the male and female version of the masker, respectively. For each masker, SRTs were obtained by an ASR system trained with speech and the same masker type added (matched training).*

3. Results

3.1. Speech recognition thresholds and psychometric functions

Figure 1 shows SRTs of listeners, ESII, mr-sEPSM and the DNN-based ASR system in all four maskers. Due to the calibration of the ESII and mr-sEPSM to SRTs in both SSN maskers, model data and listener SRTs are identical. ESII shows a good match with SRTs in the SAM-SSN noise whereas predictions by the mr-sEPSM deviate by up to 6.7 and 10.5 dB in male and female SAM-SSN condition, respectively. The DNN-based models that were not calibrated to any of the conditions show only slightly worse predictions compared to the ESII. When averaging over all four maskers, ESII shows the best predictions with a root mean squared error (RMSE) of 0.4 dB, followed by the proposed DNN-based model (0.6 dB RMSE) and the mr-sEPSM (6.2 dB RMSE).

Psychometric functions are used to compare the performance of listeners and the proposed DNN system over a wide range of SNRs. Figure 2 shows psychometric functions of individual listeners, the mean over all listeners and the DNN system in all maskers. Psychometric functions of DNNs show an excellent match with average psychometric functions of the listeners for all noise types. The RMSE between the mean function of HSR and ASR is 4.1 % on average, i.e., predicted intelligibility differs by 4.1 % from mean human intelligibility averaged over all four maskers and all SNRs ranging from -40 to 20 dB.

3.2. Relevance analysis

Relevant features are analyzed by applying relevance propagation to the same utterance mixed with noise at -5 dB SNR. For the SSN noise, a large proportion of the signal energy is masked but isolated speech energy peaks are still visible in the noisy spectrogram (cf. Fig 3 (A)). A high relevance is assigned only to these peaks whereas the rest of the spectrogram has very low relevance. A comparison with local SNR shows that regions with high relevance coincide well with time-frequency bins that exhibit a local SNR greater than 1 dB, which indicates that time-frequency glimpses are of high importance in ASR. For the SAM-SSN masker, the same general trend is observed: A high relevance is usually associated with low noise segments. Due to the nature of the modulated masker, the proportion of segments above the 1 dB threshold is considerably higher than for the SSN masker (red overlay in Fig. 3 (B)). The coincidence of relevance and low-noise segments suggests that the DNN-based ASR system is able to listen in the dips, something that GMM-based ASR are failing at [3].

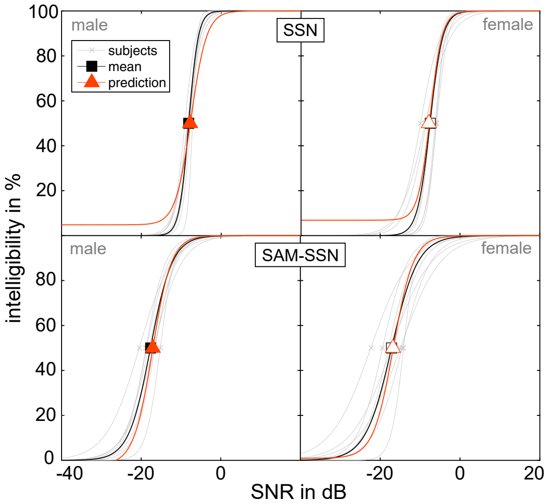

Figure 2: *Psychometric functions of individual listeners (light gray solid lines), the mean over all listeners (black solid), matched-trained ASR system (solid with triangle) in SSN (top panel) and SAM-SSN (bottom panel) noise conditions. SRTs are indicated by crosses, squares and triangles for individual subjects, mean values and ASR, respectively. Filled and open symbols refer to the male and female version of the masker, respectively.*

For the modulated noise, we also observe several segments with a high relevance for noisy (*below* the 1 dB threshold) time-frequency patches. An example is shown in the bottom panel of Fig. 3: The segment at 1.5 seconds is relatively bright (relevant) but does not have a red overlay (is below the threshold). This segment corresponds to the phoneme /ʃ/ with relatively low energy at low frequencies. We interpret this as the DNN having learned the spectral energy distribution of /ʃ/, which is then used to distinguish the phoneme from others. Consequently, a high relevance is assigned to this time-frequency region. The missing energy seems to be an important factor for discriminating phonemes, since it has an effect on the importance of frequency range: For the stationary masker, we consistently observed low relevance at low frequencies. For the modulated masker, several instances as the /ʃ/-example were found, all of which exhibit a combination of low energy, low frequency, and high relevance. The combination (high energy, low frequency, high relevance) is not observed. This suggests for classification with DNN-based system that energy in low frequencies is a poor discriminative property, but the absence of energy can be a relevant factor.

4. Discussion

The continuous increase in ASR performance that is attended by recent progress in deep learning has led to a point at which ASR systems reach human performance in speech recognition. Previous man-machine comparisons consistently reported a man-machine gap of 10 dB between normal-hearing listeners and GMM-based ASR systems trained on cepstral features

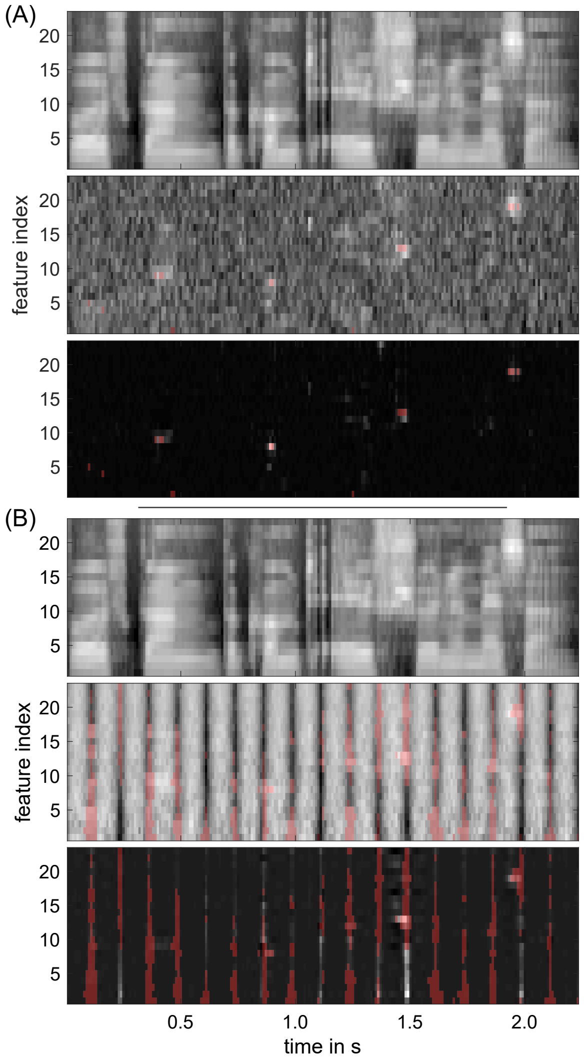

Figure 3: *(A) Clean (top) and noisy (middle) Mel spectrogram and corresponding relevance (bottom) of a test sentence in the SSN masker. The red overlay indicates time-frequency points with a local SNR > 1 dB. (B) The same as (A), but for the SAM-SSN masker. Color coding is in arbitrary units, because features and relevances were normalized for each utterance.*

[1, 2]. Although this gap could be reduced to 6 dB by using spectro-temporal features [26], human performance could only be reached after employing DNNs. The current study confirmed recent results where DNN-based ASR systems and human-listeners achieve the same performance [13, 14] and extended the comparison to an analysis in different noise maskers with and without modulations. Even if there are modulations present in the masker, which has been shown to be problematic for ASR [27], the employed DNN-based system is able to reach human performance, showing that the DNN is able to learn the specific modulations of the masker. While previous comparisons tried to close the gap by using more complex spectro-temporal features such as Gabor features to supply a modulation analysis to the GMM-based system, DNNs seem to be able to extract the relevant features needed in the considered maskers from a simple temporally spliced FBank input. It remains to be seen if this rather simple input is sufficient to reach human performance in more complex maskers such as interfering speakers. An advantage of simple mel-spectrogram input is the potential for identifying time-frequency patches relevant for recognition.

Relevance propagation has been shown to be a useful tool to gain insight into the processing of DNNs in ASR. Analysis shows that DNNs extract relevant information in a similar manner as it is observed in glimpsing and listening in the dips. In the SSN masker where most of the speech is masked, the DNN has learned that spectral peaks are likely to originate from the speech signal and, thus, have a high relevance for recognizing speech. High relevance occurs at time-frequency bins exhibiting a local SNR above 1 dB. In the modulated SAM-SSN masker, a masking release in human listeners is achieved due to a focus on modulation valleys, and a similar trend is observed in DNN-based ASR: Relevance is high at time frames with masker valleys which resembles listening in the dips. However, the relevance analysis gives us more insight than a simple SNR criterion: Missing energy in low frequencies is of high relevance for classifying certain phones such as /ʃ/. Further studies are planned on investigating the behavior of relevance depending on frequency, time, SNR, and phoneme, which will help to gain insight into the high performance levels obtained with DNNs. The similar SRT of DNN-based ASR and normal-hearing listeners motivated a comparison with established measures for predicting speech intelligibility. Our ASR system produced good predictions for noise types although it has never seen the test items before (in contrast to the models, which require separate noise and speech). Since these observations are based on two noise types only, future research will be devoted to test DNN-based ASR as a model for speech intelligibility prediction.

5. Conclusions

This study compared human speech recognition with DNN-based ASR in stationary and modulated noise maskers. Earlier studies based on GMM-HMM architectures reported a human-machine gap of 6-10 dB; this gap was found to increase for modulated noise, since ASR performance degraded, while human listeners could profit from noise modulation valleys and improved performance. Our experiments with a small vocabulary matrix test showed that HSR and ASR achieve very similar performance in both masking conditions with an SRT difference of only 0.4 dB. An algorithm for measuring the relevance of time-frequency patches for the classification result was applied to shed light on the important cues in DNN-based ASR. Our results indicate that current ASR systems are able to rely on glimpsing and listening-in-the-dips, and that the absence of energy in noise modulation valleys can help to discriminate between phoneme classes. Man-machine differences were so small that ASR might be used as a model to predict speech intelligibility (SI) in normal-hearing listeners. When compared to established SI models, the multi-resolution speech envelope power spectrum model was clearly outperformed, while ASR was on par with the extended speech intelligibility index.

6. Acknowledgements

This research was supported by the DFG (Cluster of Excellence 1077/1 “Hearing4all” and SFB/TRR 31 ’The active auditory system’; URLs: http://hearing4all.eu and http://www.sfb-trr31.uni-oldenburg.de). The authors would like to thank Tobias de Tailléz for valuable discussion on the relevance propagation algorithm and Stephan Ewert and Wiebke Schubotz for providing data from listening experiments.

7. References

[1] J. J. Sroka and L. D. Braida, "Human and machine consonant recognition," *Speech Communication*, vol. 45, no. 4, pp. 401–423, Apr. 2005. [Online]. Available: http://dx.doi.org/10.1016/j.specom.2004.11.009

[2] B. T. Meyer and B. Kollmeier, "Robustness of spectro-temporal features against intrinsic and extrinsic variations in automatic speech recognition," *Speech Communication*, vol. 53, no. 5, pp. 753–767, 2011.

[3] M. J. Carey and T. P. Quang, "A speech similarity distance weighting for robust recognition," *Ninth European Conference on Speech Communication and Technology*, pp. 1257–1260, 2005. [Online]. Available: papers2://publication/uuid/CF272B6B-78D6-4003-B384-47649CCA0186

[4] G. A. Miller and J. C. R. Licklider, "The Intelligibility of Interrupted Speech," *The Journal of the Acoustical Society of America*, vol. 22, no. 2, pp. 167–173, 1950. [Online]. Available: http://dx.doi.org/10.1121/1.1906584

[5] R. W. Peters, B. C. Moore, and T. Baer, "Speech reception thresholds in noise with and without spectral and temporal dips for hearing-impaired and normally hearing people." *The Journal of the Acoustical Society of America*, vol. 103, no. 1, pp. 577–87, 1998. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/9440343

[6] A. W. Bronkhorst, "The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions," *Acta Acustica united with Acustica*, vol. 86, no. 1, pp. 117–128, Jan. 2000. [Online]. Available: https://dx.doi.org/10.3758/s13414-015-0882-9

[7] C. Füllgrabe, F. Berthommier, and C. Lorenzi, "Masking release for consonant features in temporally fluctuating background noise," *Hearing Research*, vol. 211, no. 1-2, pp. 74–84, 2006.

[8] M. Cooke, "Glimpsing speech," *Journal of Phonetics*, vol. 31, no. 3-4, pp. 579–584, Jul. 2003.

[9] M. I. Mandel, S. E. Yoho, and E. W. Healy, "Measuring time-frequency importance functions of speech with bubble noisea)," *The Journal of the Acoustical Society of America*, vol. 140, no. 4, pp. 2542–2553, 2016. [Online]. Available: http://dx.doi.org/10.1121/1.4964102

[10] M. Cooke, "A glimpsing model of speech perception in noise." *The Journal of the Acoustical Society of America*, vol. 119, no. 3, pp. 1562–1573, 2006.

[11] A. Narayanan and D. Wang, "The role of binary mask patterns in automatic speech recognition in background noise." *The Journal of the Acoustical Society of America*, vol. 133, no. 5, pp. 3083–93, May 2013. [Online]. Available: http://link.aip.org/link/?JASMAN/133/3083/1

[12] G. Saon, H.-K. J. Kuo, S. Rennie, and M. Picheny, "The IBM 2015 English Conversational Telephone Speech Recognition System," in *Proc. Interspeech*, 2015, pp. 3–7.

[13] J. R. Hershey, S. J. Rennie, P. A. Olsen, and T. T. Kristjansson, "Super-human multi-talker speech recognition: A graphical modeling approach," *Computer Speech and Language*, vol. 24, no. 1, pp. 45–66, 2010. [Online]. Available: http://dx.doi.org/10.1016/j.csl.2008.11.001

[14] W. Xiong, J. Droppo, X. Huang, F. Seide, M. Seltzer, A. Stolcke, D. Yu, and G. Zweig, "Achieving Human Parity in Conversational Speech Recognition," *arXiv:1610.05256v1*, 2016. [Online]. Available: http://arxiv.org/abs/1610.05256

[15] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On Pixel-wise Explanations for Non-Linear Classifier Decisions by Layer-wise Relevance Propagation," *Plos One*, vol. 10, no. 7, p. e0130140, 2015. [Online]. Available: http://dx.doi.org/10.1371/journal.pone.0130140

[16] I. Sturm, S. Lapuschkin, W. Samek, and K. R. Müller, "Interpretable deep neural networks for single-trial EEG classification," *Journal of Neuroscience Methods*, vol. 274, pp. 141–145, 2016.

[17] W. Schubotz, T. Brand, B. Kollmeier, and S. D. Ewert, "Monaural speech intelligibility and detection in maskers with varying amounts of spectro-temporal speech features," *The Journal of the Acoustical Society of America*, vol. 140, no. 1, pp. 524–540, 2016. [Online]. Available: http://dx.doi.org/10.1121/1.4955079

[18] I. Holube, S. Fredelake, M. Vlaming, and B. Kollmeier, "Development and analysis of an International Speech Test Signal (ISTS)." *International journal of audiology*, vol. 49, no. 12, pp. 891–903, 2010. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/21070124

[19] K. Wagener, T. Brand, and B. Kollmeier, "Development and evaluation of a German sentence test Part III: Evaluation of the Oldenburg sentence test," *Z Audiol*, vol. 38, no. 3, pp. 5–15, 1999.

[20] 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. [Online]. Available: http://dx.doi.org/10.3109/14992027.2015.1020971

[21] B. T. Meyer, B. Kollmeier, and J. Ooster, "Autonomous Measurement of Speech Intelligibility Utilizing Automatic Speech Recognition," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pp. 2982–2986, 2015.

[22] D. B. Paul and J. M. Baker, "The design for the wall street journal-based CSR corpus," in *Proceedings of the workshop on Speech and Natural Language - HLT '91*. Morristown, NJ, USA: Association for Computational Linguistics, Feb. 1992, p. 357. [Online]. Available: http://dl.acm.org/citation.cfm?id=1075527.1075614

[23] ANSI, "Methods for calculation of the speech intelligibility index," in *American National Standards Institute, New York*, 1997, vol. 1969, no. R, pp. 1–35.

[24] S. Jørgensen, S. D. Ewert, and T. Dau, "A multi-resolution envelope-power based model for speech intelligibility." *The Journal of the Acoustical Society of America*, vol. 134, no. 1, pp. 436–46, 2013. [Online]. Available: http://dx.doi.org/10.1121/1.4807563

[25] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, 2011. [Online]. Available: https://infoscience.epfl.ch/record/192584

[26] B. T. Meyer, "What ' s the difference ? Comparing humans and machines on the Aurora 2 speech recognition task," in *Proc. Interspeech*, 2013, pp. 2634–2638.

[27] N. Parihar and J. Picone, "Aurora working group: DSR front end LVCSR evaluation AU/384/02," in *Inst. for Signal and Information Process, Mississippi State University*, 2002.