Modelling human speech recognition in challenging noise maskers using machine learning

Birger Kollmeier*, Constantin Spille, Angel Mario Castro Martínez, Stephan D. Ewert and Bernd T. Meyer

Medizinische Physik & Cluster of Excellence Hearing4all, Universität Oldenburg, D-26111 Oldenburg, Germany

Abstract: The advantage and limitations of utilizing automatic speech recognition (ASR) techniques for modelling human speech recognition are investigated for a set of “critical” speech maskers for which many standard models of human speech recognition fail. A deep neural net (DNN)-based ASR system utilizing a closed-set sentence recognition test is used to model the speech recognition threshold (SRT) of normal-hearing listeners for a variety of noise types. The benchmark data from Schubotz et al. (2016) include SRTs measured in conditions with an increasing complexity in terms of spectro-temporal modulation (from stationary speech-shaped noise to a single interfering talker). The DNN-based model as proposed in Spille et al. (2018) produces a higher prediction accuracy than baseline models (i.e., SII, ESII, STOI, and mr-sESPM) even though it does not require a clean speech reference signal (as is the case for most auditory model-based SRT predictions). The most accurate predictions are obtained with multi-condition training with known noise types and ASR features that explicitly account for temporal modulations in noisy sentences. Another advantage of the approach is that the DNN can serve as valuable analysis tool to uncover signal recognition strategies: For instance, by identifying the most relevant cues for correct classification in modulated noise, it is shown that the DNN is listening in the dips. Finally, we present preliminary data indicating that the WER of the model can be replaced with an estimate of the WER, which does not require the transcript of utterances during test time and therefore eliminates an important limitation of the previous model that prevents it from being used in real-world scenarios.

Keywords: Human speech recognition, Automatic speech recognition, Auditory model

PACS number: 43.71.An, 43.72.Ne, 43.72.Bs [doi:10.1250/ast.41.94]

1. INTRODUCTION

A quantitative processing model of the auditory system is advantageous for predicting human performance for a variety of psychoacoustic and speech perception tasks with a minimum set of underlying assumptions. It also provides the possibility to test the relevance of certain auditory functions and features — mostly motivated from physiology — for the functioning of the complete auditory system in more or less complex listening experiments.

While the “classical” auditory-processing model-based approach has the advantage of relating to several auditory functions but makes strong assumptions about the underlying detection process (such as, e.g., an “optimum detector” as in [1]), machine learning tools employed in automatic speech recognition (ASR) make much less strict assumptions and better exploit the statistical properties of the presented input signals.

Auditory processing models are — roughly speaking — better suited to describe feature extraction and processing principles in the auditory system with a standard recognition backend (like an optimum detector or a simple speech recognizer like a Dynamic time warping (DTW) algorithm [2]). Machine learning methods, on the other hand, employ a better recognition backend that is ideally trained to classify the data presented even if an imperfect preprocessing and/or feature extraction has to be compensated for by massive statistical learning from large training data sets. A comparison of both approaches should lead to the possibility to combine the best from two worlds, i.e., to utilize the best suited frontend features with a high-performance backend.

The current contribution therefore reviews recent advances in ASR-based prediction of human speech recognition that utilize an auditory-model based frontend (using Amplitude Modulation Filterbank (AMFB) features) and a Deep Neural Net (DNN)-based backend (Sects. 2.1–3.1). An analysis of how this DNN operates and weighs
information within the input signal spectrum at different instants of time is described in Sect. 3.2. Finally, an extension of this approach is presented in Sect. 4 to predict performance without requiring a transcription of the test data, thus enabling the approach to become language independent. A discussion and conclusion section (Sects. 4 and 5) conclude this contribution.

2. METHODS AND DATA EMPLOYED

2.1. Human Benchmark Data

Schubotz et al. [3] published sentence recognition threshold (SRT) data from 8 young normal-hearing listeners obtained with the German Matrix Sentence Test (Oldenburger Satztest, see [4] for a review). Eight noise maskers with increasing complexity in terms of spectro-temporal modulations (stationary noise to single talker) and a male/female version for each masker were employed as listed in Table 1. They pose a challenge to classical speech recognition prediction models (referred to as “baseline models,” see below) since none of these models were able to correctly predict human performance [3].

The SRT, i.e., the SNR with 50% error rate, was obtained with an adaptive procedure according to [5]. Schubotz et al. demonstrated that none of the standard speech intelligibility prediction models employed were able to predict performance in an accurate way [3], thus making this benchmark data set especially suitable to test and compare different model approaches. The baseline models include the speech intelligibility index (SII, [6]), the extended SII [7], the short-time objective intelligibility measure (STOI, [8]), as well as the multi-resolution speech envelope power spectrum model (mr-sEPSM; [9]).

2.2. ASR-based Prediction of Human SRT Data

The ASR features employed utilize Mel-spectrograms (40 frequency channels) or Amplitude Modulation Filterbank Features (AMFB) [10]. They were used as input to a deep neural net (DNN) as acoustic model in combination with a hidden Markov model (HMM) as backend, which was implemented in the Kaldi ASR toolkit (as proposed in [11], without modifications except for the training data, see below). The DNN uses 7 hidden layers with 2048 neurons per layer. Its output are context-dependent triphone targets that are fed to a HMM to produce the transcript of the sentence to be recognized.

As the training base set, 10 hours of matrix sentences from 10 male and 10 female speakers were employed for a speaker-independent training [12]. It was artificially extended to an 80h set by adding noise at random SNRs in the range of $[-10, +20]$ dB which is a typical amount for current DNN-based systems. The training procedure was either matched (i.e., the training and testing noise are identical and 16 models are obtained) or multi-condition (2 sets for male/female and a random selection of all noise types, which also implies that one of the noise types is later seen during testing). Note that the model is language-specific and requires a range of SNRs for each condition to be learned, but does not require clean speech test signals for training or testing.

For testing ASR, the Oldenburg Sentence Test (1 speaker, not seen during training) was presented to the system with 400 random SNRs drawn from a range of $[-30, +20]$ dB in order to sample the psychometric function for each noise condition. The ASR-predicted SRT is then given by the SNR corresponding to 50% speech recognition.

3. RESULTS

3.1. Comparison between Measured and Predicted SRT

Figure 1 presents the results from the different benchmark models (left panel) in comparison to the measured data (in green) for the various conditions employed. The predictions using the filterbank feature set is displayed in the middle column whereas the amplitude modulation filterbank feature set was employed for the results displayed in the right column.

Obviously, the ASR-based predictions clearly outperform the baseline models where the best match between human performance and prediction is obtained for the amplitude modulation based filterbank in the multi-condition training. This can also be seen from the prediction error which was calculated as RMS error in Table 2.

3.2. Looking Inside the Black Box: Analysis of DNN-based Models

Deep learning algorithms are often considered as approaches that work well, but merely represent black boxes that cannot be further analyzed. However, recent studies have shown that activations and/or learned net parameters can be useful to provide insight into signal processing strategies of DNNs: Tüske et al. [14] showed that DNNs learn an auditory filterbank mapping when using the raw waveform as input for ASR. Nagamine et al.
[15] have shown that early layers learn to separate easily separable phonemes (/t/, /n/, /s/), while later layers separate difficult classes (/p/, /t/, /k/). To analyze our speech intelligibility model, we applied a relevance algorithm [16] that determines time-frequency patches relevant for correct phoneme recognition. The algorithm traces back the activations from the output (phoneme) layer to the input (spectrogram) layer. The result for modulated maskers adapted from [17] shows that high-relevance patches coincide with modulation valleys of the masker (Fig. 2): Panel C shows the relevance of the respective time-frequency bins that were used for correct classification, which coincide with modulation valleys of the masker. The color represents intensity on a scale from deep blue (minimum) to light yellow (maximum).

### 4. REQUIREMENT OF WORD LABELS

The underlying DNN is constructed to be used with open-set speech materials that clearly go beyond the closed-set Matrix test materials used in this study as common benchmark for humans and machine. One limitation of the model described in [13], however, is that—although it does not require the clean speech token—the prediction is based on the WER of the ASR model, which requires a correct transcript of spoken sentences during test time and hence limits the the model to known, correctly labeled speech materials. This prevents it from being used in real scenarios with arbitrary speech input, e.g., for constant monitoring of speech intelligibility in a hearing aid or cochlear implant. To overcome this limitation, we explore the relation of speech intelligibility with the estimated WER, which is based on an algorithm that quantifies the representation of the DNN output [18]. This output corresponds to phoneme posterior probabilities for each time frame (which is referred to as posteriorgram). For clean speech, distinct phoneme activations are obtained, which are smeared in the presence of noise. This degradation of the DNN output is related to the WER of an ASR system and has therefore been applied in multi-stream ASR for the selection of the optimal stream [19], by

---

**Table 2** Average RMS prediction error (in dB) for baseline models (left) and four different DNN-based models (right) used either with training with a fixed noise or with multicondition (mc) training employing several noise conditions.

<table>
<thead>
<tr>
<th></th>
<th>SII</th>
<th>ESII</th>
<th>STOI</th>
<th>FBank-mc</th>
<th>FBank</th>
<th>AMFB</th>
<th>AMFB-mc</th>
</tr>
</thead>
<tbody>
<tr>
<td>male</td>
<td>7.9</td>
<td>5.6</td>
<td>9.2</td>
<td>3.5</td>
<td>4.6</td>
<td>2.1</td>
<td>2.1</td>
</tr>
<tr>
<td>female</td>
<td>9.4</td>
<td>7.6</td>
<td>9.7</td>
<td>7.8</td>
<td>4.2</td>
<td>2.3</td>
<td>2.1</td>
</tr>
<tr>
<td>avg.</td>
<td>8.6</td>
<td>6.6</td>
<td>9.5</td>
<td>5.6</td>
<td>4.4</td>
<td>2.2</td>
<td>2.1</td>
</tr>
</tbody>
</table>

---

**Fig. 1** SRTs of normal-hearing listeners (shown in each panel in green), of baseline models (left panel from Schubotz et al. [3], DNN-based model with Mel-spectrogram features (middle, from Spille et al., 2018), and with amplitude modulation features (right, from Spille et al. [13]).

**Fig. 2** (A) Spectrogram representation of a clean speech segment, (B) the same segment in sinusoidally modulated noise, and (C, zoomed panel) the relevance of time-frequency bins that were used for correct classification, which coincide with modulation valleys of the masker. The color represents intensity on a scale from deep blue (minimum) to light yellow (maximum).
calculating the mean temporal distance (MTD) of posteriorgram frames. The main idea of MTD is that temporal smearing of phoneme activations increases the similarity of distant phoneme vectors. A detailed description of MTD is given by [18].

In preliminary experiments, we investigated the relation between the true WER (obtained by decoding the DNN output using the HMM and comparing the transcript to the groundtruth label) and the estimated WER. For the eight maskers derived from the male speaker (“male maskers”), we obtain a good correlation for the diverse noise types employed in this study as shown in Fig. 3(A). Note that error rates can exceed 100% when the number of errors per utterance exceeds the number of words. This clear relation motivated the use of estimated WER in the model. Preliminary data for a subset of the experiments outlined above (AMFB features, male maskers only) are shown in Fig. 3(B). While the match between model and listeners is not as good as for the best DNN-based model using word labels (with an RMSE of 1.9, cf. Table 2), it still achieves a good match (RMSE of 2.35) and outperforms all baseline models on the average. Within these error margins, the application of the model is no more limited to predicting human performance with the Matrix test sentence test but opens the possibility for an open-set speech recognition prediction.

5. DISCUSSION

The DNN-based processing model described here clearly outperforms the baseline models (SII, ESII, STOI and mr-sEPSM) in predicting human speech recognition performance for a number of masker conditions with increasing modulation complexity in the temporal and spectral domain (cf. Fig. 1 and Table 1). Remarkably, a comparably simple, generic front end is used that captures only basic auditory properties such as critical-band scaling and amplitude modulation features. The successive processing steps in the backend (HMM and DNN) perform a transformation of these features which is optimized in a statistical sense by the ASR task trained. This frontend/backend combination seems to yield a certain independence from the exact type of features employed in the frontend and extracts the necessary information for performing the respective recognition task in an unsupervised fashion by appropriate training of the backend. This may be the decisive advantage above the baseline models that typically extract more specific speech features and use an explicit, fixed detection mechanism. If the features employed are not optimal for the speech recognition task in the respective noise masker, the benchmark models may miss important features (like listening in the gap) that are learned by the DNN/HMM-model on the fly.

In comparison with the recently published FADE model [20] which uses a GMM/HMM recognizer specifically trained on the same Matrix test material as used for the prediction, the current DNN/HMM model employs a more sophisticated recognizer that has not seen the exact speech token under test before and could potentially operate on an open speech recognition set as well. However, this comes at the cost of more training data needed for this approach and further restrictions like much more computational resources required of the model parameters for the respective experiment to be modelled. In addition, the current model approach has not been tested yet for predicting psychoacoustic experiments and the effect of hearing impairment in a similar way as successfully implemented with FADE [20].

In addition, the current approach shows a high generalization potential: The estimation method of the word error rate (WER) from posteriorgrams described here yields an independence from the exact phonetic transcription of the test speech material employed. This might enable the model to predict intelligibility for arbitrary input speech — as long as it is similar to the large speech training corpus employed. Thus, the current approach becomes applicable to open speech tests and for unsupervised estimation of ongoing speech intelligibility (e.g., in hearing devices or...
This potential has yet to be explored in the future. Another test of the approach to be explored in the future is its performance in arbitrary noise maskers that are not seen during training. It is still unclear if a selective weighting of performance in arbitrary noise maskers that are not seen in the dips’ for such modulated maskers.

The estimation method using posteriorgrams without employing a transcript of the test data is encouraging since it might enable the model to predict intelligibility for arbitrary input languages.

ACKNOWLEDGEMENTS

Supported by the DFG (Cluster of Excellence 1077/1 Hearing4all; URL: http://hearing4all.eu).

REFERENCES


