What’s the difference? Comparing humans and machines on the Aurora 2 speech recognition task

Bernd T. Meyer
Medical Physics, Carl-von-Ossietzky Universität Oldenburg, Germany
bernd.meyer@uni-oldenburg.de

Abstract
The comparison of human speech recognition (HSR) and machine performance allows to learn from the differences between HSR and automatic speech recognition (ASR) and serves as motivation for using auditory-inspired strategies in ASR. The recognition of noisy digit strings from the Aurora 2 framework is one of the most widely used tasks in the ASR community. This paper establishes a baseline with a close-to-optimal classifier, i.e., our auditory system by comparing results from 10 normal-hearing listeners to the Aurora 2 reference system using identical speech material. The baseline ASR system reaches the human performance level only when the signal-to-noise ratio is increased by 10 or 21 dB depending on the training condition. The recognition of 1-digit recordings was found to be considerably better for HSR, indicating that onset detection is an important feature neglected in standard ASR systems. Results of recent studies are considered in the light of these findings to measure how far we have come on the way to human speech recognition performance.

Index Terms: Human speech recognition, automatic speech recognition, human-machine comparison

1. Introduction
Recognition of spoken language is a task that is easily performed by the healthy human auditory system, yet automatic speech recognition (ASR) often fails to accurately transcribe speech, especially in noisy conditions. It is this performance gap between human speech recognition (HSR) and ASR that motivates the use of auditory features for ASR with the aim of increasing its overall robustness [15].

Earlier studies compared the classification results of humans and machines, which served both as a motivation for using auditory-inspired signal processing and also allows to learn from the differences between HSR and ASR. Lippmann [8] presented a review of numerous human and machine results and reported ASR error rates in some cases to be an order of magnitude higher than those of humans. For a phoneme recognition task that effectively excludes lexical knowledge, the gap was reported to be smaller, but ASR error rates were still about 2.5 times higher compared to HSR [9].

This paper extends this line of research by comparing results from 10 listeners and a standard ASR system for a digit recognition task that has been widely used by the ASR community, i.e., the Aurora 2 speech database [4]. Listening tests with similar data have been performed earlier: Robertson et al. [13] presented a digit-in-noise test based on Aurora 2 speech material; the aim of the study was to model speech intelligibility of digit triplets (a specific subset of the database) in babble noise with one normal-hearing and one hearing-impaired listener. Leonard [6] reported an error rate of 0.1% for clean speech resynthesized from linear prediction coefficients measured with the TIDigits database from which Aurora 2 was generated. In contrast to [13] and [6], the main goal of the present paper is to measure human performance in conditions that closely resemble the original Aurora 2 task to establish an competitive baseline with the best recognition system that we currently have: our auditory system. The human-machine gap should be quantified, both in terms of the absolute recognition difference and the SNR-increase that is required to achieve human recognition rates with machine learning approaches. The human listeners for this comparison were native (L1) and non-native (L2) listeners of American English, which allows to analyze if and to what extend L1-L2-effects result in different recognition performance for this digit-in-noise test. Further important questions are how different noise types affect HSR and ASR, and whether memory effects play a role for the chosen recognition task, which might give us some insight into specific weaknesses of ASR systems. Finally, the HSR scores are compared to current studies using Aurora 2, to answer the question how far we have come to reach human recognition abilities with new approaches in speech recognition.

The remainder of this paper is structured as follows: The listening experiments, speech material and the experimental ASR setup are specified in Sections 2 and 3. The experimental results regarding human and machine word recognition rates and discussions are presented in Section 4. The paper is summarized in Section 5.

2. Listening tests
2.1. Description of the Aurora 2 database
Aurora 2 was developed as a reference system to provide a framework that allows the comparison of ASR results of different front-ends or noise suppression algorithms [4]. It specifies speech data, a standard feature set, and the configuration of a Hidden Markov Model (HMM) classifier. The speech data consists of spoken digit strings (one to seven digits covering the words one to nine, zero, and oh) taken from the TIDigits database [6]. It was produced with 104 speakers for the testing set and 110 different speakers for training, and includes clean speech as well as speech mixed with ten different noise types. All data was resampled to a sampling frequency of 8 kHz.

Aurora 2 defines two ASR training conditions that use either clean digits or a mixture of noisy and clean data for training (multi-condition). Testing is performed with clean and noisy data, using combinations of various noises grouped in different test sets: Subway, babble, car, exhibition (‘Test A’, which is used during multi-condition training), restaurant, street, airport, and station (‘Test B’, noise types that are unknown to the
classifier), and subway m and street m ("Test C", which has been filtered to simulate the transmission characteristics of a phone line).

2.2. Selection of SNRs and noise types
As Aurora 2 contains more than 70,000 digit strings in the test sets, a subset was compiled from the database that is suitable for listening tests with humans. Five out of ten noises were chosen to cover the three Aurora 2 test sets and a range of ASR recognition scores. Babble (Test set A) and airport (Set B) resulted in a low-performance (<50% word accuracy for clean ASR training and averaged over SNRs from 0 to 20 dB). Medium performance (50-60% accuracy) was achieved with noises car (A) and station (B), while subway m (C) produced high accuracies (>60%) for a clean-trained ASR system. Pilot measurements with two participants (one English and one German listener) were performed to determine the SNRs at which normal-hearing listeners actually produce errors with these noise types to enable a statistical analysis. For clean data, a 100% identification score for 650 word presentations was measured. At 5 dB SNR the error rates were still above 98%; hence, the tests with all listeners were performed at the lower SNRs (0 and -5 dB).

2.3. Subjects and experimental setup
Five native listeners of American English (L1) and five German listeners (L2) participated in the study. The listeners were aged between 25 and 39 and were normal-hearing (i.e., their hearing threshold for pure tones in standard audiology did not exceed +20 dB at more than one data point and +10 dB at more than two data points in the pure tone audiogram). One L1-listener was not available for audiometry tests but reported to have normal hearing. Listeners were asked to read a description of the experiments and were verbally instructed to enter a sequence of digits via a graphical interface using a mouse. A training phase of 5 to 10 minutes with feedback in case of incorrect responses preceded each measurement to familiarize the participants with the interface and the noisy audio material. After the presentation of a few clean digit strings, utterances from the noisy ASR training set were presented at the lowest SNR in that set (5 dB), and listeners could adjust the volume to a comfortable level. Since participants listened to some noise types that reoccur in the testing phase, this procedure resembles ASR multi-condition training. It seems fair to compare HSR and multi-condition trained ASR for a second reason, i.e., the life-long training that enables the listeners to quickly adapt to the noises (especially since the noises were recorded in everyday situations). During training, feedback was provided by displaying the correct transcript in case of a false response. Listeners were paid and encouraged to take regular breaks to avoid errors due to fatigue. Each listener participated in 2-4 measurement runs with each run not exceeding 2.5 hours including breaks. Randomized stimuli were presented in a sound-proof booth and with audiological headphones (Sennheiser HD200) after an online free-field equalization and normalization of the RMS power was applied to the signals.

3. ASR setup
The Aurora2 reference recognizer uses 13-dimensional Mel-frequency cepstral coefficients (MFCCs) [2] along with delta and double-delta features, which are computed from the speech data using the front-end provided with the HTK toolkit [16]. The resulting 39-dimensional features are used to train and test a HMMs. The setup uses whole-word HMMs with 16 states and with a 3-Gaussian mixture with diagonal covariances per state. Skips over states are not permitted in this model. For better comparability (and if not stated otherwise), all ASR results reported here are obtained with the five noise types used for HSR (instead of using the full test set with 10 noises).

4. Results
Word recognition accuracies (WRAs) for HSR are shown in Figure 1. The results are based on a total of 8,773 strings (28,705 words) collected from ten listeners. First it was tested if the two groups of listeners produce significantly different results. While the average L1-accuracies are higher for all SNRs (Table 1), a two sided t-test showed that the differences are not significant (p-value for -5 dB: p = 0.075, 0 dB: p = 0.074). Hence, the results in the next sections are averaged over the two groups. The comparison of human and machine results is shown in Figure 2 for different signal-to-noise ratios. The ASR test set consists of 28,028 digit strings (92,785 words). When comparing this result to L1 vs. L2 experiments on phoneme level [3], the differences obtained with digits are not as large as for sub-word units. This is presumably caused by the fact that complex stimuli enable listeners to employ a higher number of cues, and to identify each speech token by using contextual factors such as the number of syllables [1]. This compensates the degraded phoneme classification in L2 listeners [3].

![Figure 1: HSR results for individual listeners. Circles and crosses represent data from German and English listeners, respectively.](image)

<table>
<thead>
<tr>
<th>SNR / dB</th>
<th>US (L1)</th>
<th>GER (L2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>-5</td>
<td>82.3</td>
<td>77.0</td>
</tr>
<tr>
<td>0</td>
<td>96.4</td>
<td>94.6</td>
</tr>
<tr>
<td>5</td>
<td>99.5</td>
<td>98.7</td>
</tr>
</tbody>
</table>

Table 1: Word recognition scores for native and non-native listeners.

4.1. Human-machine gap in terms of the SNR
The relative increase of error rates (as, e.g., reported in [8, 9]) strongly depends on the specific SNR for the Aurora 2 dataset. At the lowest SNR, the ASR error rates (multi-condition training) are increased by a factor of 4 compared to HSR. At 0 and
5 dB, this factor amounts to 13 and 35, respectively. For clean training, these factors are even higher (at 5 dB, the ASR error rate of 82.2% is 133 times the HSR error rate of 0.62%). When the performance differences are as large as reported here, the increase of errors appears to be a rather fragile measure. Hence, the human-machine gap is also reported in terms of the SNR increase required for ASR to reach the human performance level. Therefore, the speech reception threshold (SRT), i.e., the SNR at which listeners achieve a 50% accuracy is estimated. The SRT is usually the steepest point in the SNR-dependent curve of recognition rates, thereby enabling a consistent determination of a reference SNR.

The SRT for HSR is estimated based on data with digit-triplets from Aurora 2 in different babble noises [13]: The WRA for 5 dB, as shown in Figure 2, lies between the performance for 2- and 4-talker babble published in [13]. When a similar degradation with decreased SNR is assumed for our data, the WRA at -10 dB can be estimated to be 51.3% and (assuming a linear degradation near this data point) the SRT corresponds to -10.2 dB. The SRTs for ASR are +0.2 dB and 11 dB for multi-condition and clean training, respectively. Hence, the human-machine gap for digit recognition at 50% speech intelligibility can be estimated as the difference between these SRTs. It amounts to approximately 10 dB for multi-condition training and (10+11) dB for clean-trained ASR systems (Figure 2).

4.2. Effect of noise types

The noise types in Aurora 2 differ with respect to the frequency and temporal modulation characteristics, which leads to large differences in ASR accuracy as reported in Section 2.2. The effect of the Aurora 2 noises on HSR is analyzed and compared to ASR performance in Table 2.

The lowest HSR performance by far is obtained in babble noise: The error rates for this type of noise are at least increased by 50% in comparison to any other masker. Since the Aurora 2 babble noise is a multi-talker noise with unintelligible speech, informational masking presumably does not play a role and the difference to other noise types arises from effective energetic masking. For noise types that are different from speech either spectrally or temporally, HSR accuracies are consistently high. The noise types employed for the training sessions (babble and car) result in high errors, which suggests the additional training is negligible compared to the listeners’ life-long training (which is relevant since we are dealing with noises recorded in everyday environments).

ASR results for clean training are similar to human accuracies, i.e., low accuracies are obtained with babble noise, high accuracies are obtained with subway m. The results for multi-condition trained ASR are quite different: The lowest accuracies are observed for subway m, which simulates signals transmitted via a phone line. Although it is fairly stationary, car noise also results in low recognition results. Subway m and car are examples for noise conditions that human listeners can easily cope with in contrast to the MFCC-based recognizer.

<table>
<thead>
<tr>
<th>Test set</th>
<th>Babble</th>
<th>Car</th>
<th>Airport</th>
<th>Station</th>
<th>Subway M</th>
</tr>
</thead>
<tbody>
<tr>
<td>HSR (-5)</td>
<td>70.2</td>
<td>82.9</td>
<td>83.7</td>
<td>85.0</td>
<td>87.6</td>
</tr>
<tr>
<td>HSR (0)</td>
<td>93.8</td>
<td>96.1</td>
<td>97.4</td>
<td>96.7</td>
<td>96.0</td>
</tr>
<tr>
<td>ASR (-5, clean)</td>
<td>5.2</td>
<td>8.6</td>
<td>5.2</td>
<td>7.6</td>
<td>9.5</td>
</tr>
<tr>
<td>ASR (0, clean)</td>
<td>6.4</td>
<td>11.1</td>
<td>7.9</td>
<td>8.5</td>
<td>14.2</td>
</tr>
<tr>
<td>ASR (10, clean)</td>
<td>37.4</td>
<td>44.4</td>
<td>41.1</td>
<td>42.9</td>
<td>52.9</td>
</tr>
<tr>
<td>ASR (-5, multi)</td>
<td>16.8</td>
<td>14.2</td>
<td>18.4</td>
<td>15.6</td>
<td>15.0</td>
</tr>
<tr>
<td>ASR (0, multi)</td>
<td>45.1</td>
<td>43.3</td>
<td>52.1</td>
<td>46.3</td>
<td>38.3</td>
</tr>
<tr>
<td>ASR (10, multi)</td>
<td>89.2</td>
<td>94.0</td>
<td>90.9</td>
<td>90.6</td>
<td>88.4</td>
</tr>
</tbody>
</table>

Table 2: Word recognition rates depending on the type of noise for HSR and ASR. Test SNR and training condition are denoted in parentheses.

4.3. String length and recognition rate

In HSR, the recognition of words in long digit strings might be affected by memory effects. Listeners were given the option to repeat each utterance to avoid this effect, which resulted in word accuracies that are consistent for different string lengths (Figure 3). HSR results were obtained with the -5 and 0 dB data, while ASR SNRs are shifted by 10 (multi-condition training) and 20 dB (clean training) to achieve comparable recognition rates (and following the estimation for the human-machine gap presented before). The effect of string length is stronger for ASR (with increasing WRA for longer strings) than for HSR. The percentage of insertion errors for 1-digit recordings is far higher for ASR than for HSR (85% vs. 50%, multi-condition training). The number of insertions is presumably increased since the proportion of silence is higher for these signals. Further, for longer strings, the probability of detecting at least one word with high confidence increases and - since the temporal dynamics are modeled by the state transitions of the HMM - this also improves estimates of the positions of other digits and increases the overall WRA. To improve these estimates for short digit sequences, onset cues could be used on ASR feature level, which is also reasonable when comparing ASR to HSR, since the auditory system exhibits a relatively robust onset detector [7] and shows a WRA largely independent of the digit length.
4.4. Human-machine gap for novel feature types

The listeners’ data and baseline results are compared to recently developed features to evaluate how far we have come on the way to closing the gap between HSR and ASR.

Power-normalized cepstral coefficients (PNCCs) are features that have been shown to be very robust against a large variety of noise types [5]. They are calculated from the squared magnitude Short-Time Fourier Transform, which is integrated via a Gammatone filter bank since this has been shown to better approximate the place-frequency mapping of the basilar membrane compared to the triangular filters used for MFCCs [12]. A power function nonlinearity that mimics the dependency of the input sound level and the perceived loudness is used to compress the output of the Gammatone filter bank. The compression exponent of 0.1 was derived from the relation of auditory-nerve firings and the level of a presented tone. The medium-time power bias is removed motivated by the fact that the auditory system is sensitive to changes of the signal (in contrast to low-modulated background noises, which are largely ignored).

Secondly, spectro-temporal Gabor features are considered, which are obtained by filtering spectrograms with two-dimensional filters. Spectral and temporal modulation frequencies are free parameters of the filters. In [14, 10], a Gabor filter bank is proposed that arranges filters by modulation frequencies which results in features that cover the modulation space homogeneously, and which have recently been shown to improve a state-of-the-art system to perform ASR in complex acoustic scenes [11]. For the comparison, the features from [10] are chosen, which are based on power-normalized spectrograms borrowed from PNCC processing.

The comparison shown in Figure 4 was obtained with multi-condition training. The curve for MFCCs is therefore identical to ”ASR (multi)” in Figure 2. In terms of the SRT, novel feature extraction approaches reduce the human-machine gap in speech recognition from 10 dB to about 6 dB. It seems that, although large progress has been made in the last decades, there is still quite a way to go to achieve ASR with human performance. This is also reflected in the accuracies: The best ASR system is far better than the baseline (50% vs. 26%, 24% difference), but the same is true for humans compared to the best ASR system (84% vs. 50%, 34% difference).

While this comparison is restricted to the effects of feature extraction alone, recent studies combine feature extraction techniques with signal enhancement and optimized backends or training conditions to reach state-of-the-art performance (e.g., [11]). Such systems could easily be compared to the data presented here using the Aurora 2 speech corpus to measure the man-machine gap.

5. Summary

In this paper, the speech recognition performance of humans and machines was compared based on a digit-in-noise test. The largest differences between HSR and ASR were found at -5 dB SNR, in which case the ASR results are close to chance performance, while human results exceed 98%. This gap was also quantified in terms of absolute performance and the SNR-gap. The gap between the listeners and the reference recognizer amounts to approximately 10 dB for clean and 21 dB for multi-condition trained ASR systems. The widely-used Aurora 2 database was employed for the experiments, which makes the results comparable to many ASR studies. The ASR baseline and human results were compared with novel approaches in feature extraction, which revealed that the gap was reduced from 10 to approximately 6 dB. This shows that ASR has made large progress in the last decades, but further improvements are required before ASR reaches the human performance level.

Analyzing different types of noise showed that human performance is severely degraded by multi-talker babble noise, while the listeners could cope well with other noises that have either different spectral or temporal characteristics. Results for an ASR system that was trained with several noise types were different from HSR results with high error rates for mismatched frequency characteristics, a condition that humans easily dealt with. This highlights the importance of ASR systems that are invariant to channel distortions. The length of the digit string only had a small effect on the word recognition accuracy of listeners when given the option to replay the signal, while ASR results were heavily degraded for 1-digit utterances. This degradation of performance might be compensated by including on-set detectors into the ASR feature extraction stage.

6. Acknowledgements

Supported by the DFG (SFB/TRR 31 'The active auditory system'; URL: http://www.sfb-trr31.uni-oldenburg.de/). The authors would like to Mathias Dietz, Volker Hohmann, and Daniel Marquardt for valuable contributions to this work. thank Birger Kollmeier and Hendrik Kayser for valuable discussions.
7. References


