Available online at www.sciencedirect.com

ScienceDirect

Computer Speech & Language 52 (2018) 123–140

www.elsevier.com/locate/csl

# Comparing human and automatic speech recognition in simple and complex acoustic scenes☆

Constantin Spille*, Birger Kollmeier, Bernd T. Meyer*

*Medizinische Physik and Cluster of Excellence Hearing4all, Carl von Ossietzky Universität, Küpkersweg 74, Oldenburg 26129, Germany*

Received 22 June 2017; received in revised form 1 December 2017; accepted 11 April 2018
Available online 14 April 2018

**Abstract**

Former comparisons of human speech recognition (HSR) and automatic speech recognition (ASR) have shown that humans outperform ASR systems in nearly all speech recognition tasks. However, recent progress in ASR has led to substantial improvements of recognition accuracy, and it is therefore unclear how large the task-dependent human-machine gap still remains. This paper investigates this gap between HSR and ASR based on deep neural networks (DNNs) in different acoustic conditions, with the aim of comparing differences and identifying processing strategies that should be considered in ASR. We find that DNN-based ASR reaches human performance for single-channel, small-vocabulary tasks in the presence of speech-shaped noise and in multi-talker babble noise, which is an important difference to previous human-machine comparisons: The speech reception threshold, i. e., the signal-to-noise ratio with 50% word recognition rate is at about −7 to −8 dB both for HSR and ASR. However, in more complex spatial scenes with diffuse noise and moving talkers, the SRT gap amounts to approximately 12 dB. Based on cross comparisons that use oracle knowledge (e.g., the speakers' true position), incorrect responses are attributed to localization errors or missing pitch information to distinguish between speakers with different gender. In terms of the SRT, localization errors and missing spectral information amount to 2.1 and 3.2 dB, respectively. The comparison hence identifies specific components in ASR that can profit from learning from auditory signal processing.


*Keywords:* Human-machine comparison; Speech recognition threshold; Deep neural networks; Speech intelligibility prediction; Spatial scenes

## 1. Introduction

A strong motivation for research on automatic speech recognition (ASR) is the observation that human speech recognition (HSR) easily outperforms machine listening in most recognition tasks and acoustic scenes. Hence, a quantitative comparison of ASR and human listeners enables us to focus the development of ASR systems on achieving the same performance as HSR and to quantify the progress of ASR towards reaching human performance levels. A second aspect of human-machine comparisons is the possibility to model human speech recognition performance by systematically changing the acoustic conditions and recognition task and tailoring ASR to minimize the human-

☆ This paper has been recommended for acceptance by Prof. R. K. Moore.

* Corresponding authors.
*E-mail address:* constantin.spille@uni-oldenburg.de (C. Spille), bernd.meyer@uni-oldenburg.de (B.T. Meyer).

https://doi.org/10.1016/j.csl.2018.04.003
0885-2308/ 

machine gap in as many conditions as possible. The current paper follows this approach by considering ASR and HSR in simple recordings with one audio channel and spatial acoustic conditions recorded with multiple microphones of behind-the-ear (BTE) hearing aids (multi-channel). While a systematic comparison of ASR and HSR in spatial scenes has not been performed so far, many comparisons in simple single-channel scenes have been conducted in the past. In a classic review, Lippmann (1997) reported error rates of machines to be "often more than an order of magnitude greater than those of humans for quiet, wideband, read speech," referring to results for the Switchboard corpus (conversational speech), for which the HSR error rate was 4%, while the ASR error rates were at 43% (Lippmann, 1997; Peskin et al., 1996). The human error rate was later reexamined by Xiong et al. (2016) with trained professional transcriptionists and was found to be at 5.9%. Since the study by Lippmann (1997), the ASR error rate has been first reduced to 8% (Saon et al., 2015) and more recently to 5.9%, i.e. achieving human parity, demonstrating the achievements in this time period (Xiong et al., 2016). Due to the vanishing gap between human and machine performance, an increasing interest in the differences and similarities between errors of human listeners and ASR system could be recently observed (Stolcke and Droppo, 2017; Saon et al., 2017). These studies were conducted with conversational speech making it difficult to pinpoint the weaknesses of ASR systems that lead to this gap because of the entanglement of different error sources. In an attempt to decouple those error sources, Shen et al. (2008) focused on the influence of the language model by comparing recognition scores for sentences in which one word was replaced by a near-homophone so that discrimination could only be performed based on acoustic input and not by a language model. Recently, Shen et al. (2017) estimated the gap in language modeling between humans and different current language models. Another important potential ASR error source lies on the sublexical level, which is analyzed in the current study by performing experiments with non-sense words or sentences with very simple grammatical structure. Similar research has been performed before:

Sroka and Braida (2005) performed experiments with nonsense syllables in additive noise and found a man-machine gap of 10 dB with an ASR system using cepstral features, 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. On the other hand, high- and low-pass filtering reduced (and for some conditions even eliminated) the gap. Based on an analysis of articulatory features such as voiced/unvoiced speech, Sroka and Braida (2005) concluded that human listeners and ASR systems use different cues for speech recognition. Other studies showed that human listeners outperform automatic recognizers in speech-like and modulated additive noise in experiments with nonsense syllables and for a digit recognition task (Carey and Quang, 2005; Cooke and Scharenborg, 2008). Cooke and Scharenborg (2008) showed that humans achieved 85% lower error rates relative to ASR in clean speech in a consonant recognition task. Meyer and Kollmeier (2011) compared ASR with Mel-frequency cepstral coefficients (MFCC) and HSR in a phoneme recognition task and found that ASR error rates are about 2.5 times higher and that the man-machine gap was 15 dB. By presenting resynthesized speech from MFCCs to listeners, the information loss due to MFCC processing could be quantified to an amount of 10 dB. From these results, it was concluded that ASR systems should incorporate spectral fine structure and temporal as well as spectro-temporal properties of speech in a more appropriate way than it is done by MFCCs. More recently, Mandel et al. (2016) used a data-driven framework to measure the importance of time-frequency points to the intelligibility of syllables for humans and ASR systems based on Gaussian Mixture Models (GMM) and Hidden Markov Models (HMM).

Recent progress in ASR has mainly been driven by developments in deep learning: Saon et al. (2015) showed that the main contributions to the improvements of ASR scores were related to implementations of deep neural networks (DNN). The concept of an artificial neuron or perceptron was first introduced by Rosenblatt (1957) in the late 50s, but training networks with multiple layers have become possible in the 80s by the development of the error back-propagation algorithm (Rumelhart et al., 1986). However, using deep neural networks with many hidden layers and large numbers of neurons per layer has only become possible in the last decades due to the increase in computing power provided by graphical processor units (GPU) and efficient training algorithms, such as complementary priors for deep belief networks (Hinton et al., 2006). The combined work of different research groups has further pushed the development of DNNs and has led to the development of different deep learning toolkits (Hinton et al., 2012) of which the Kaldi speech recognition toolkit by Povey et al. (2011) is the most widely used in ASR. Due to these developments, so called hybrid DNN-HMM approaches where DNNs are used to produce posterior probabilities for each state of a HMM have become omnipresent in ASR. Prior to DNNs, GMMs were used to model these posterior probabilities. The use of DNN-based models closed the speech recognition gap between humans and machines in

some conditions (Xiong et al., 2016). These breakthroughs allow for a human-machine comparison in more complex scenes that more and more differ from a simple speech in additive noise task.

The current paper compares human recognition performance with DNN-based ASR systems for several tasks, ranging from recognition of matrix sentences in different noises (single-channel recordings) to speech in spatial scenes with competing moving speakers and diffuse noise that were recorded with multiple channels of BTE hearing aids. Human speech recognition performance in single-channel recordings is evaluated across different studies by Jürgens and Brand (2009), Hochmuth et al. (2015b) and Schubotz et al. (2016). In spatial scenes, a machine listening system from Spille et al. (2017) is used which combines processing from computational auditory scene analysis (CASA) and ASR, e.g. the position of sound sources is used to steer a beamformer for signal enhancement. All comparisons are performed on a sublexical level with tasks being designed not to be influenced by language model processing (e.g. by analysing sentences with a fixed grammar).

The aim is to compare different architectures of ASR systems and different feature representations to human speech recognition, which allows to identify processing steps in ASR that are responsible for the human-machine gap. Identifying those processing steps allows the development of new processing strategies that further improve ASR performance and close the human-machine gap. In simple acoustic scenes with one audio channel, the focus is laid on the feature representation and the ASR backend, i.e. the system's architecture, to measure the human-machine gap induced by the features or the architecture, which would also be visible in more complex scenes. Afterwards, differences between ASR and HSR in complex spatial scenes can then led back to differences in the processing of the acoustic scenes. The similarity between HSR and ASR is quantified by measuring the overall performance in terms of the speech recognition threshold (SRT), i.e. the SNR at which the recognition accuracy is 50%. Factors between HSR and ASR error rates often strongly depend on the SNR and previous studies have shown the SRT to be more stable over a wide range of SNRs (Meyer, 2013; Spille and Meyer, 2014).

The paper is structured as follows: In Section 2 we describe how the SRT is used to compare HSR and ASR, the comparisons experimental details as well as the ASR features and systems. After presenting the results of the different comparisons in Section 3 we discuss them in Section 4 and give some conclusions in Section 5.

## 2. Methods

The experimental procedure of this study is as follows: First, speech material is used to create different acoustic scenes (single or multi-channel, see Section 2.2). The signals are then presented to normal-hearing listeners whose responses (word sequences) are logged. Second, the same signals are transcribed by different ASR models (Section 2.3). ASR is compared to human speech recognition in terms of the speech recognition threshold (Section 2.1).

### 2.1. Speech recognition threshold

As a global measure for performance differences, the speech recognition threshold (SRT) is employed. When the recognition accuracy is plotted against the signal-to-noise ratio (SNR), the resulting psychometric function has a sigmoidal shape. At very low SNRs, the recognition accuracy is close to zero while at very high SNRs it is usually at 100%. The SRT is the SNR at which 50% of the presented words are recognized.[1] It is usually located at the steepest point of the psychometric function and thus can be reliably determined. Note that *lower* SRTs represent a *better* performance. In screening tests, an adaptive method is used to measure the subjects SRT and the slope of the psychometric function at this point (Brand et al., 2002). In ASR, the recognition accuracy is measured at different SNRs and the SRT is interpolated from this data.

As shown in Fig. 1, the SRT is a more stable measure of the difference in overall performance than word error rate (WER) at distinct SNRs: SRTs for different accuracies (horizontal arrows) have approximately the same length, although they are related to very different accuracies. The increase of WER however, reaches from extremely low to high values when comparing low and high SNRs (vertical arrows in Fig. 1).

[1] In speech audiometry, the SRT at which 80% accuracy is reached is often measured for hearing-impaired listeners, and it is referred to as $SRT_{80}$. However, *SRT* without supplement usually means $SRT_{50}$.

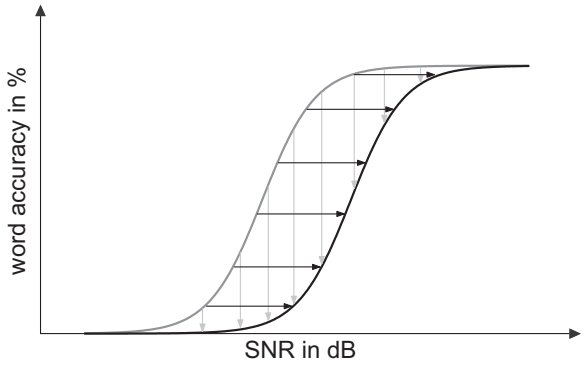


Fig. 1. Schematic of two psychometric functions. The shift in SNR (black arrows) is similar over a wide range of SNRs while the shift in word error rates (grey arrows) strongly depends on the specific SNR.

### 2.2. Acoustic scenes and speech databases

As summarized in Table 1, a total of three different acoustic conditions are considered here: two single-channel and one spatial condition. All experiments were conducted with the a German matrix sentence test, i.e. the Oldenburg sentence test (OLSA), as a basis for the speech data which will be briefly described in the following.

#### 2.2.1. Matrix sentence test

The Oldenburg sentence test (OLSA) was developed by Wagener et al. (1999) and is becoming a standard test for speech recognition testing in Europe since highly comparable closed-set tests for more than 18 major languages already (see Kollmeier et al., 2015, for a review). The test is used to determine the speech recognition threshold of human listeners, e.g. in order to assess the performance with and without a hearing device. Hence, data of many SRT measurements is available for normal-hearing and hearing-impaired listeners. 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. The original OLSA used in audiology was produced by just one speaker. Because sufficient data to train an ASR system was needed, a speech corpus of 10 h of speech from 20 different speakers was recorded, using the syntactical structure and the vocabulary of the OLSA (Meyer et al., 2015). For multi-condition ASR training, clean and noisy files were used, the latter being obtained by mixing signals with random parts of a stationary speech-shaped noise at random SNRs for each file ranging from −10 to 20 dB.

#### 2.2.2. Single-channel recordings

The first two experiments were conducted in comparatively simple acoustical scenes with a stationary speech-shaped noise and a multi-talker babble noise presented as a single audio channel. During listening experiments, the SRT values and occasionally the slope of the psychometric function at the point of the SRT are determined. The

Table 1
Overview of the experimental details of the different acoustic scenes used in this study.

| Acoustic condition | Type of masker | Signal presentation | No. of speakers | No. of listeners |
|---|---|---|---|---|
| Single-channel recordings | Speech-shaped, stationary noise | Monaural | Training: 20 Test: 9 | 58 normal-hearing<br>34 hearing-impaired |
| | Multi-talker babble noise | | Training: 20 Test: 9 | 10 normal-hearing |
| Multi-channel recorded spatial scenes | Competing talker + speech-shaped stationary noise | Binaural | Training: 20 Test: 10 | 9 normal-hearing |

psychometric function is fully described by the SRT and its slope via the function

$$f(x) = \frac{1}{1 + \exp\left(-\frac{x - L_{50}}{s}\right)} \tag{1}$$

where the slope is given as $m = 1/4s$ and $L_{50}$ corresponds to the SRT (Wagener et al., 1999). SRTs are measured with an adaptive procedure proposed by Brand and Kollmeier (2002). 20 OLSA sentences were presented with an initial SNR of 0 dB; subsequently, the SNR is varied depending on the intelligibility of the previous sentence and the SRT is estimated via an maximum-likelihood estimator. SRTs depend both on the listener and the speaker, which is why different listener–speaker combinations are tested.

The data from listening experiments in stationary speech-shaped noise (SSN) was collected from different studies. The following studies used the original OLSA speaker: Wagener et al. (1999) performed listening tests with 20 normal-hearing listeners and obtained a model function that was used to describe the recognition performance of normal-hearing listeners. Results from 20 normal-hearing listeners and 34 hearing-impaired ears (with a high-frequency hearing loss) were taken from Juergens et al. (2010). Hochmuth et al. (2015b) and Schubotz et al. (2016) measured SRTs of 10 and 8 normal-hearing listeners, respectively. Note that different types of speech-shaped noises were used in these studies, i.e. Wagener et al. (1999) used a test-specific noise that matches the long-term spectrum of the target speaker, Juergens et al. (2010) and Hochmuth et al. (2015b) used the ICRA1 noise (Dreschler et al., 2001) and Schubotz et al. (2016) created a SSN with a long-term spectrum of the International Speech Test Signal by Holube et al. (2010). Nevertheless, results are not significantly different between the groups of normal-hearing listeners as will be seen in Section 3.1. In a study by Hochmuth et al. (2015a), OLSA sentences uttered by eight different speakers were recorded and SRTs of 10 normal-hearing listeners were measured for each of the speakers. Sentences of the speakers were mixed with the ICRA1 noise as well as a multi-talker babble noise which was composed of 20 different speakers. Also sentences of the original OLSA speaker were mixed with this multi-talker babble noise to measure the corresponding SRT (Hochmuth et al., 2015b).

For ASR experiments, speech data from the original OLSA speaker was mixed at random SNRs between −40 and 20 dB to create a testset that samples the whole psychometric function. The SRT of an ASR system is obtained by fitting the function in Eq. (1) to the word accuracies of the ASR system.

#### 2.2.3. *Complex acoustical scenes recorded with multiple channels*

After comparing state-of-the-art ASR systems and human listeners in single-channel tasks, now, the comparison is extended to complex multi-channel scenes. This comparison includes factors that are not covered by previous comparisons, such as localization of speakers.

Experiments in simulated complex multi-channel scenes were conducted with speech material that exhibits the OLSA structure. Because the 10 h speech corpus used for the experiments with single-channel recordings was not available when these experiments were performed, a smaller database was used for testing. The speech data used for testing consists of sentences produced by 10 speakers (4 male, 6 female) that were recorded using close-talk microphones in our lab. The original recordings with a sampling rate of 44.1 kHz were downsampled to 16 kHz. To simulate movements that cover a wide range of azimuth angles, utterances with a duration of 5 to 10 s are required. These were obtained by concatenating three OLSA sentences produced by one speaker. These audio signals are convolved with previously measured head-related impluse responses (HRIR) which allows simulation of spatial scenes with many different speaker constellations. The HRIRs used in this study are a subset of the database described by Kayser et al. (2009): Anechoic HRIRs from the frontal horizontal half-plane measured at a distance of three meters between microphones and loudspeaker and a 5° resolution for the azimuth angles, which was interpolated to obtain a 1° resolution were selected. Each HRIR was measured with eight channels (6 channels from 2 behind-the-ear (BTE) hearing aids, and 2 channels from in-ear microphones).

The speakers initial and final positions, speed and direction of movement on a semi-circle with a radius of 3 m were randomly chosen but constrained to typical walking speeds. The speakers moved linearly from the start to the end point for the duration of the respective stimulus and crossed their tracks with a 50% probability. To avoid speaker tracks that overlap too much, initial and final positions were chosen to ensure an average angle difference of at least 10°. Additionally, the minimal distance between the start and end points was set to 10° and 20° for non-crossing and crossing speakers, respectively. Moving speakers were simulated by employing a frame-wise processing

scheme: 64 ms Hann windows with 50% overlap were applied and each time frame was convolved with the respective HRIR. For a detailed description of the experimental parameters, the reader is referred to Spille et al. (2013b).

The in-ear signals were presented to the listeners via headphones and serve as input to the binaural model in ASR experiments. The beamformer was working on the signals from the BTE hearing aids (6 channels).

*Listening experiments.* Nine normal-hearing subjects participated in the experiment. Their hearing thresholds did not exceed +20 dB at any data point in the audiogram, and not more than +10 dB at more than two data points. Signals were presented in a soundproof booth via audiological headphones (Sennheiser HDA200 with free-field correction). Before the experiment, listeners were verbally instructed and were also asked to read the written instructions to prepare for the task. Additionally, each listener completed a short training phase to get familiar with the speech material, the acoustic scenes as well as the graphical user interface. This training procedure took about 5–10 min.

For HSR experiments, a visual marker indicating the initial azimuth of one speaker was presented to the participants before the playback started so that they know to which speaker to attend. Participants were instructed to focus on the source initially located at that angle (ranging from −90 to +90°). All subjects listened to the first of the ten created test sets containing about 40 min of speech data with 1065 sentences, and entered the recognized words produced by the target speaker via the graphical user interface. To mimic the ASR setup with the known vocabulary and a fixed grammar, subjects were presented all word alternatives as a grid of 5 (word groups) × 10 (words per group) on the screen, resulting in a closed-test setup. Subjects could enter the responses for each sentence of the three-sentence presentations in arbitrary order. To avoid incorrect responses due to memory effects (which would be relevant with 15-word sentences of an average duration of 6.8 s), subjects could re-listen to the presentation as often as desired. The total measurement time per subject was approximately six hours, which was divided in multiple sessions that did not exceed 2 h of measurement per session and listener. Word error rates were averaged over all signals and all subjects for each SNR. For ASR experiments, the initial position of one speaker is supplied to the system and used to define the desired target speaker.

### 2.3. ASR system architectures and features

ASR experiments have been performed using the Kaldi speech recognition toolkit (Povey et al., 2011) and DNNs have been trained using the nnet1 recipe.

In all acoustic multi-channel scenes, an additional pre-processing is necessary to deal with the moving speakers and the binaural information available in these scenes. Since some of the features have a physiologic or auditory motivation while others do not, it is possible to investigate if auditory inspired features also produce more human-like results.

#### 2.3.1. Feature extraction

Baseline features are calculated by converting time signals to Mel-Frequency Cepstral Coefficients (MFCCs) (Davis and Mermelstein, 1980) with additional cepstral mean and variance normalization. By adding delta and double-delta features, 39-dimensional feature vectors are obtained per 10 ms step. When deep neural networks (DNNs) are used instead of Gaussian mixture models (GMMs), it has been shown that it is beneficial to provide the output of the mel-filterbank as input to the DNN (Mohamed et al., 2012). Here, a mel-filterbank with 40 channels is used for 16 kHz data. These Filterbank features are mean and variance normalized resulting in a 40 dimensional feature vector per 10 ms step.

The third feature type used in this study is inspired by psycho-acoustic and physiologic findings showing that neurons in the primary auditory cortex of mammals are sensitive to specific spectro-temporal modulations (Mesgarani et al., 2007). Two-dimensional Gabor filters have been applied to model the selectivity of such neurons (Qiu et al., 2003), which motivated their use for ASR. Gabor filters are defined as the product of a complex sinusoidal function $s(n, k)$ (with $n$ and $k$ denoting the time and frequency index, respectively) and an envelope function $h(n, k)$. In this notation, the complex sinusoid is defined as

$$s(n,k) = \exp[i\omega_n(n-n_0) + i\omega_k(k-k_0)] \tag{2}$$

and the Hann function that we chose as envelope (with the parameters $W_n$ and $W_k$ for the window length) is given by

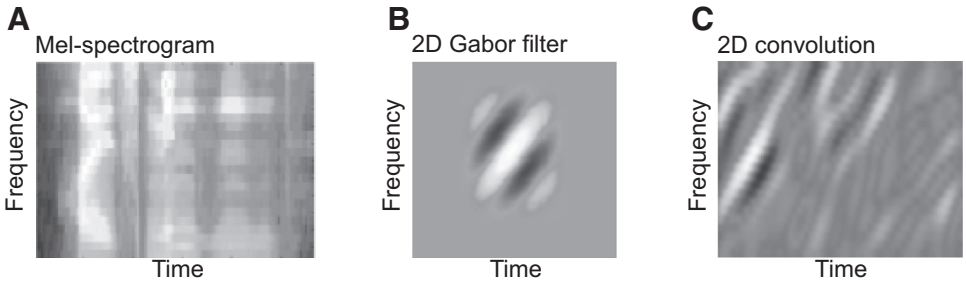


Fig. 2. Convolution of a spectrogram (A) with a 2-D-Gabor filter (real part of complex filter schown in panel (B) results in an enhancement of spectro-temporal patterns (C).

$$h(n,k) = 0.5 \; - \; 0.5 \cdot \cos\left(\frac{2\pi(n-n_0)}{W_n+1}\right) \cdot \cos\left(\frac{2\pi(k-k_0)}{W_k+1}\right).$$

Gabor features are calculated by processing a spectro-temporal representation of the input signal by a number of modulation filters. Filtering is performed by calculating the 2D convolution of the filter and the log-mel-spectrogram (see Fig. 2).

Experiments presented in this paper are based on a spectro-temporal filter bank used in Meyer et al. (2012). The filter bank contains a set of temporal, spectral and spectro-temporal filters that were chosen to cover a wide range of modulation frequencies relevant in speech recognition. The result of the time-aligned convolution for all filters is subsampled based on the spectral width of the filters and used as feature vector. Mean and variance normalization is performed on the final features.

*2.3.2. Single-channel ASR*

The comparison of different ASR system architectures and feature representations with HSR performed here also allows to quantify improvements in ASR due to auditory features and DNNs. It potentially explains specific differences between ASR systems and how they relate to the human-machine gap. Two different ASR architectures commonly used are explored in this study:

1. Gaussian mixture model (GMM) in combination with hidden Markov models (HMMs)
2. Deep neural networks (DNNs) with subsequent HMMs

Both architectures are trained on triphone targets and every phone is modeled with 3 HMM states except the silence phone which is modeled with five states. The number of triphone targets and the number of Gaussians per HMM state differ for the different experimental conditions. DNNs had seven hidden layers with 2048 units per layer, similar to experiments by Mohamed et al. (2012).

The OLSA dataset contains 37 different phones from which 149 context-dependent phones were derived. The number of triphone targets was set during training to a commonly used value of 2000 from which after pruning about 1000−1300 triphones were left. Every HMM state was modeled with 13 Gaussians on average. MFCCs and Filterbank features without any delta coefficients are spliced with a temporal context of $\pm 5$ frames and used as input features. Gabor features are spliced with a temporal context of $\pm 1$ frame since they inherently capture a larger temporal context compared to the other two features. The dimension of the DNN output (which is subject to a softmax transformation) equals the number of triphone targets for the specific feature.

*2.3.3. Multi-channel ASR*

For multi-channel signal processing, the system described by Spille et al. (2017) is used. The binaural signals of the in-ear microphones are fed into the binaural model developed by Dietz et al. (2011) that estimates the direction of arrival of spatially distributed speakers. A particle filter (Särkkä et al., 2007; Hartikainen and Särkkä, 2008) is then used to track the positions of the moving sources over time. Its output steers a minimum variance distortionless

response (MVDR) beamformer (Cox et al., 1987; Bitzer and Simmer, 2001), enhancing the 6-channel speech signal that is to be transcribed by the ASR system. The ASR system's architectures are then the same as for the monaural scenes.

## 3. Results

### 3.1. Single-channel recordings

At first, results from the listening experiments with the original OLSA speaker in a stationary speech-shaped noise from the studies by Wagener et al. (1999), Juergens et al. (2010), Hochmuth et al. (2015b), and Schubotz et al. (2016) are shown in Table 2.

Although these studies used different stationary speech-shaped noises, results are very similar. Welch's t-test shows no significant difference between the different groups (cf. Table 3). Hence, the mean of −7.3 dB over all these studies is used as the mean performance of normal-hearing listeners in a stationary speech-shaped noise.

Additionally to SRTs for the original speaker in stationary speech-shaped noise, Hochmuth et al. (2015b) also measured SRT for the original speaker in a multi-talker babble noise and SRTs for eight different speakers in both stationary and babble noise using data from 10 normal-hearing subjects. All SRTs from normal-hearing subjects and ASR experiments are shown in Fig. 3 and in the appendix in Tables A.5 and A.6.

Fig. 3 shows SRTs of GMM-based systems to be substantially higher than human SRTs in both maskers. The multi-talker babble noise generally yields higher SRTs for humans and ASR. However, all curves have a similar shape, i.e. speakers that are more intelligible for humans also produce lower SRTs in ASR, independently of the system's architecture or input feature. The different curves are merely shifted along the y-axis. Filterbank features and MFCCs result in highest SRTs while Gabor features can improve ASR performance and thus decrease SRTs considerably. DNN-based systems produce SRTs very close to human results for all feature types. The overall deviation from human SRTs is quantified by the root mean squared error (RMSE) between ASR and human SRTs (cf. Table 4).

Table 2
Results of normal-hearing subjects from different listening experiments. The sample size *N*, the speech recognition threshold (SRT), the slope of the psychometric function at the SRT and the standard deviations (sd) across listeners of SRTs and slopes are shown for all studies.

| Study | *N* | SRT in dB | sd in dB | Slope in %/dB | sd in %/dB |
|---|---|---|---|---|---|
| Wagener et al. (1999) | 20 | −7.1 | 0.2 | 17.1 | 1.7 |
| Juergens et al. (2010) | 20 | −7.1 | 1.8 | 18.6 | 4.8 |
| Hochmuth et al. (2015b) | 10 | −7.4 | 1.2 | N/A | |
| Schubotz et al. (2016) | 8 | −7.5 | 1.5 | 15.8 | 5.5 |
| **Mean** | | **−7.3** | **1.1** | **17.1** | **4.0** |

Table 3
*P*-values of Welch's t-test between the different groups of normal-hearing listeners (a value $>$ 0.05 indicating no significant difference between the results of the respective studies listed in the row and column).

| | Juergens et al. | Hochmuth et al. | Schubotz et al. |
|---|---|---|---|
| Wagener et al. | 0.39 | 0.27 | 0.28 |
| Juergens et al. | | 0.33 | 0.32 |
| Hochmuth et al. | | | 0.39 |

Table 4
RMS-errors in dB between SRTs obtained by ASR and normal-hearing listeners.

| | Filterbank features | | MFCCs | | Gabor features | |
|---|---|---|---|---|---|---|
| | GMM | DNN | GMM | DNN | GMM | DNN |
| Stationary | 11.7 | 1.5 | 7.9 | 1.3 | 3.0 | 2.0 |
| Babble | 14.2 | 1.8 | 15.6 | 2.4 | 8.3 | 2.9 |
| Average | 12.95 | 1.65 | 11.75 | 1.85 | 5.65 | 2.45 |

When averaged over all speakers and both maskers, GMM-based systems with Filterbank features and MFCCs result in very high deviations of at least 11.8 dB, Gabor features could halve the RMSE to 5.7 dB. DNNs further reduce the RMS error for all features to an average of 2.5–1.7 dB. Note that although Gabor features with DNNs result in lower SRTs than Filterbank features and MFCCs in stationary noise condition, their deviation to human

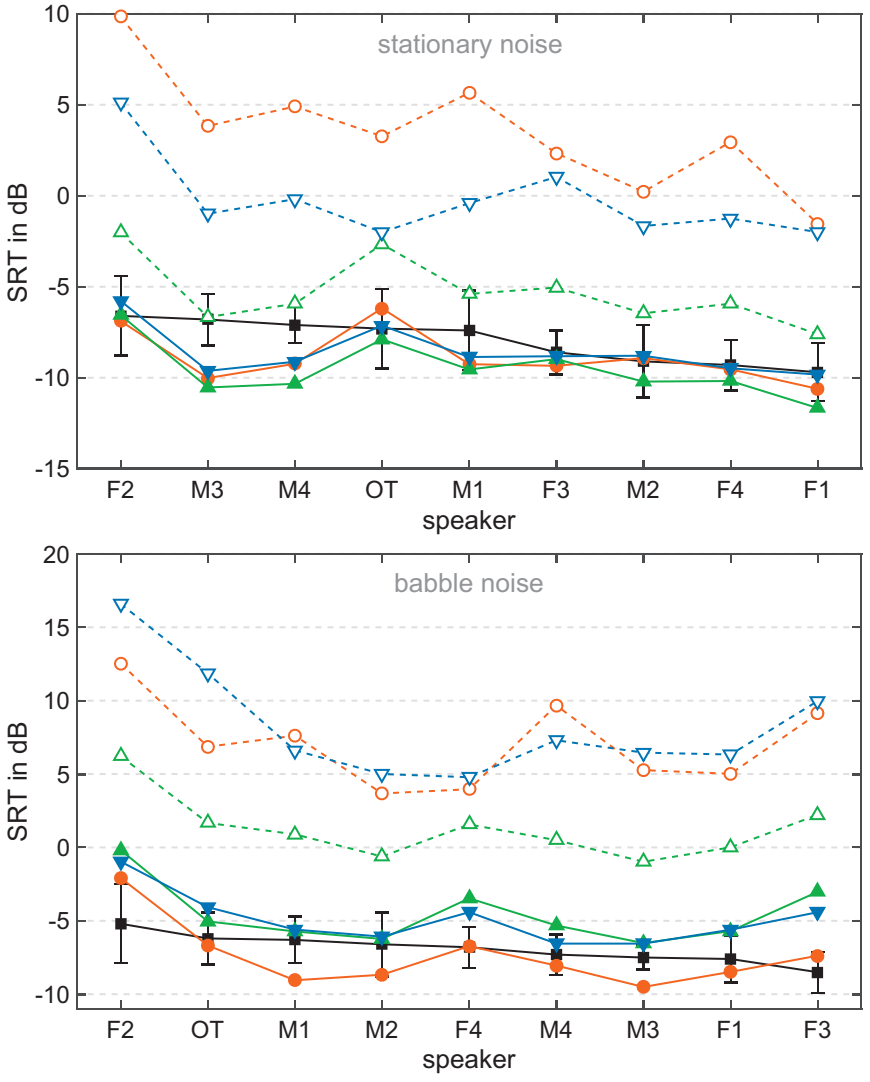


Fig. 3. SRTs in dB of normal-hearing subjects and ASR systems for all speakers in both stationary speech-shaped (top) and multi-talker babble noise (bottom). Human SRTs are shown by black squares. Dashed lines with open symbols and solid lines with filled symbols denote GMM/HMM and DNN/HMM architectures, respectively. The original OLSA speaker is indicated as OT, male speakers as M1–M4 and female speakers as F1–F4. Circles refer to Filterbank features, downward triangles to MFCCs and upright triangle to Gabor features.

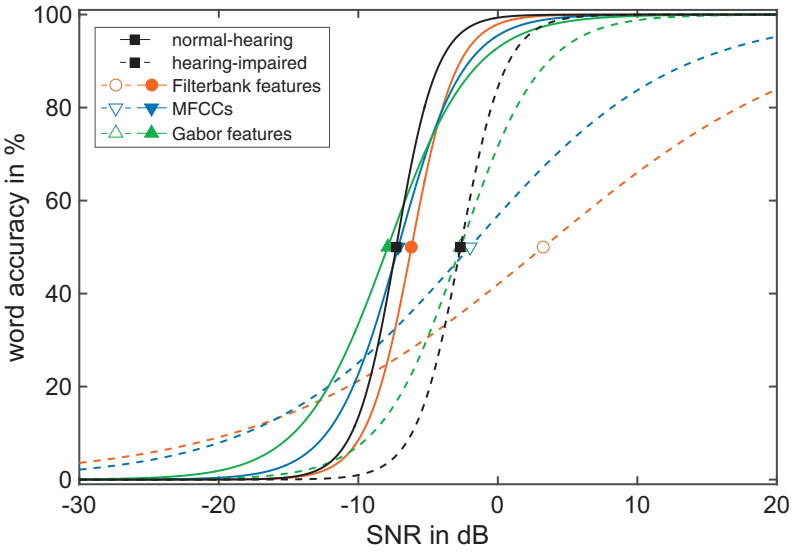


Fig. 4. Word accuracy vs. signal-to-noise ratio for different features and different ASR system architectures in single-channel scenes with German matrix sentences. Solid and dashed black lines denote the performance of normal-hearing and hearing-impaired subjects, respectively. Dashed and solid lines with triangles denote GMM/HMM and DNN/HMM architecture, respectively. The symbols denote the SRT of the corresponding system/listener group.

SRTs is higher. Overall, Filterbank features with DNNs show the best match with human SRTs with an average RMSE of 1.65 dB. For most speakers, SRTs obtained with DNNs lie within the 95% confidence interval of measured human SRTs (cf. Fig. 3).

In stationary speech-shaped noise, the slope of the psychometric function was measured with normal-hearing and hearing-impaired listeners (as mentioned above). This allows to not only compare the SRTs of human listeners and ASR, but also to look at the whole psychometric function to compare the recognition performance over a wide range of SNRs. Jürgens and Brand (2009) determined both the SRT and its slope of hearing-impaired listeners. The SRT was found to be −2.7 dB (sd: 2.3 dB), which is about 5 dB higher than the SRT of normal-hearing listeners and has a standard deviation that is twice as high as for normal-hearing listeners. The slope of 15.8%/dB (sd: 4.6%/dB), however, is only slightly shallower compared to normal-hearing listeners. Fig. 4 shows psychometric function for listeners and ASR systems in stationary speech-shaped noise. GMM-based systems with MFCCs and Gabor features could reach the performance of hearing-impaired listeners, while MFCCs exhibit a much shallower slope of the psychometric function than Gabor features do. As already shown in Fig. 3, DNN-based systems result in similar SRTs as normal-hearing listeners. Nevertheless, the slopes of the psychometric functions differ between the different input features. In particular, they are 8.1%/dB for Gabor features, 10.7%/dB for MFCCs and 15.6%/dB for Filterbank features. The similarity between ASR and HSR can be measured in terms of the RMSE between the psychometric functions. In this case, MFCCs result in the most similar system with an RMSE of 3.8%, followed by Filterbank features with 4.7% and Gabor features with 6.7%.

### *3.2. Complex acoustical scenes recorded with multiple channels*

In complex spatial scenes, spatial cues are available, which potentially help to separate target and masker. In the acoustic scenes considered here, the SRT of normal-hearing listeners was −20.6 dB which is 13.3 dB below the SRT in the monaural condition. Since listeners consistently produced less than 50% errors even at −20 dB SNR, the SRT was estimated from a fit to the observed data points (see Fig. 5).

In ASR, performance levels in terms of the SRT for the different systems are comparable to the monaural experiments. Highest error rates were obtained with an GMM/HMM using Filterbank features as input features for which the SRT amounts to 2.4 dB (i.e., 0.9 dB lower than in monaural scenes with stationary noise). With MFCCs as input

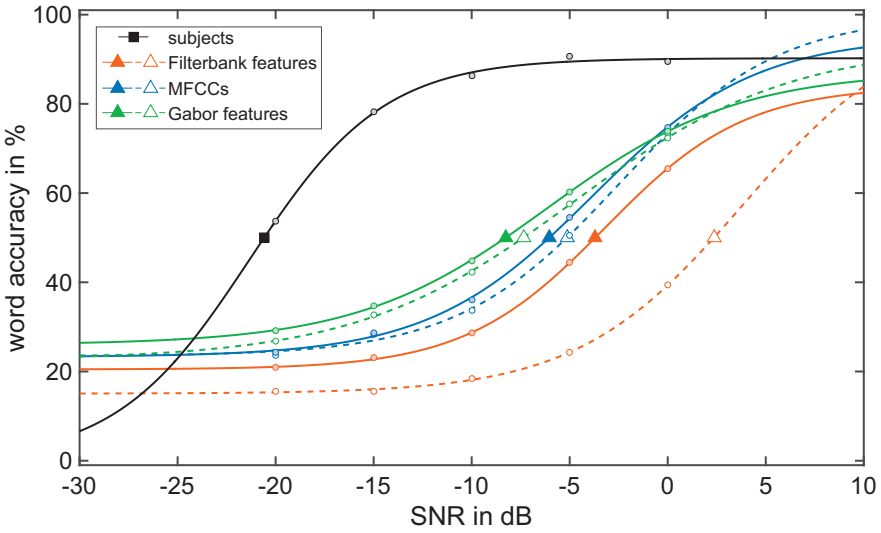


Fig. 5. Word accuracy vs. signal-to-noise ratio for different features and different ASR system architectures in multi-channel scenes with German matrix sentences. Solid black line denotes the performance of normal-hearing subjects. Dashed and solid lines denote GMM/HMM and DNN/HMM architecture, respectively. Little circles and big symbols denote the measured recognition accuracies and SRTs obtained from fitting, respectively.

features the SRT is −5.5 dB (i.e., 3.4 dB lower than in monaural scenes). This shows that the ASR system employed here can also benefit from spatial information. If a DNN is used instead of a GMM, the SRTs are reduced to −3.7 and −6.1 dB for Filterbank features and MFCCs, respectively, which corresponds to an improvement by DNNs of 6.1 dB (filterbank) and 0.6 dB (MFCCs). Spectro-temporal Gabor features reach an SRT of −7.4 dB with GMMs and −8.3 dB with a DNN (0.9 dB improvement by DNN). The total man-machine gap in binaural scenes therefore amounts to 12.3 dB.

### *3.2.1. Analysis of the performance gap*

In the following, a further analysis of the 12.3 dB gap is presented with the aim of identifying the error sources that cause the differences between man and machine. Previous studies have shown that recognition performance considerably decrease with degrading localization performance (Spille et al., 2013a; 2013b). If the binaural model is not sufficient to accurately localize the speaker, the beamformer cannot enhance the signal and hence, WERs increase. To quantify this effect, the true speaker positions (instead of estimated angles as required for a working application) were used for beamforming. Removing these localization errors reduces the SRT of the DNN system with Gabor features by 2.1 dB from −8.3 to −10.4 dB.

For multi-speaker scenes such as the ones investigated, human listeners perform significantly better when voice characteristics of the target and the concurrent speaker are different (Brungart, 2001). In particular, it was shown that word error rates for different-gender speakers was substantially lower than for same-gender speakers. This is also reflected by the data collected for this study: The same-gender recognition accuracies are consistently lower than the different-gender WRAs (by 8.4% on average – see also Fig. 6). An additional factor that influences the complexity of binaural scenes is the spatial distance of target and masker. Since we are interested in how complexity influences HSR and ASR, an analysis of situations with speakers being temporarily at the same position (i.e., the constantly moving speakers cross their ways) was performed. For HSR, the WER for non-crossing speakers is 5.3% lower than for crossing speakers, which was expected since a non-crossing implies a spatial distance that should increase source separability. Fig. 6 shows the improvement in HSR performance in scenes where both speakers did not cross, i.e. both speakers have a larger spatial distance compared to scenes where speakers did cross.

To bridge the man-machine gap in spatial scenes, ASR algorithms should be applied that take into account the same cues that are employed by the auditory system. However, since these algorithms are currently unavailable in our ASR system, we analyzed the man-machine gap in situations in which the mentioned cues are unavailable to the human listener, i.e., scenes with crossing speakers with the same gender. In these situations, the SRT is raised from

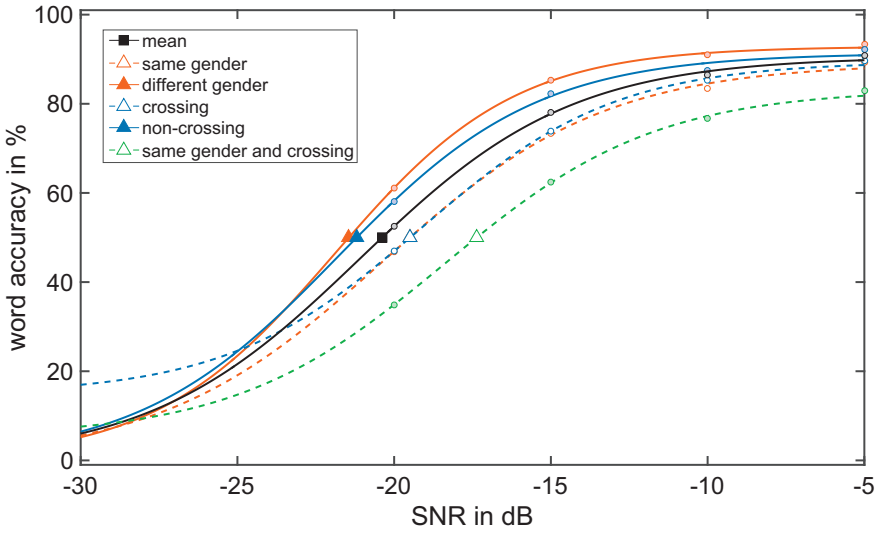


Fig. 6. Recognition accuracies for normal hearing listeners in situations where both speakers have the same or a different gender and where both speakers cross or do not cross each other. Little circles and big symbols denote the measured recognition accuracies and SRTs obtained from fitting, respectively.

the original −20.6 dB by 3.2 dB to −17.4 dB. The gender and position cues hence explain for an SRT shift of 3.2 dB. Assuming that ASR algorithms are developed that optimally exploit these cues, the gap of 10 dB (which was already based on optimal localization) could potentially be further reduced to 7 dB.

### 3.3. Correlation analysis of errors

To further look into the similarities and differences of ASR and human listeners, recognition performances are compared on a more detailed level. First, word error rates for different target speakers are compared. Fig. 7(A) shows word error rates for human listeners and ASR systems for each target speaker for DNN-based ASR systems trained

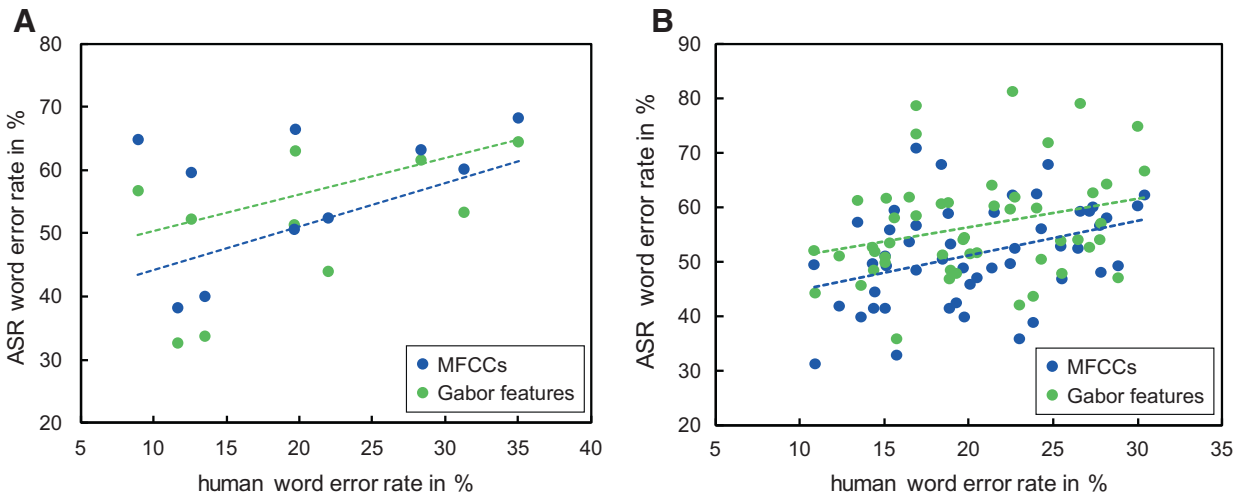


Fig. 7. Correlation between human and ASR word error rates (DNN-based systems) per speaker (A) and per word (B). Each dot represents one of the ten speakers or one of the 50 words, respectively. Blue dots represent results obtained with DNNs and MFCCs and green dots represent results obtained with DNNs and Gabor features. Dashed lines denote the corresponding linear regression lines. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article).

with MFCCs and Gabor features. Pearson's correlation coefficients are 0.48 and and 0.54 for MFCCs and Gabor features, respectively. On a word level, correlation coefficients are much smaller, namely 0.29 for MFCCs and 0.38 for Gabor features.

## 4. Discussion

Since this work is focused on comparing HSR and ASR to identify processing steps currently missing in ASR, differences between the ASR architectures will rarely be discussed; instead we compare our results and the human-machine gap with previous findings.

### *4.1. Single-channel recordings*

For single-channel recordings utilizing matrix sentences, human listeners outperform all ASR systems with GMMs in this task which is in line with literature: Sroka and Braida (2005) reported ASR scores to be 15−30 percentage points lower than human accuracy for consonant-vowel-consonant and consonant-vowel syllables in additive speech-shaped noise. Based on this data, the SRT gap was estimated to be around 10 dB for MFCCs. This is in the same range as results reported by Meyer et al. (2011), where HSR and ASR performance was measured with vowel-consonant-vowel logatomes in speech-shaped noise for which a man-machine gap of 11.8 dB was reported. Furthermore, Meyer (2013) determined the human-machine gap in a later study to be around 10 dB between multi-condition trained ASR system with MFCCs and human listeners in the Aurora2 task (English digits in noise). Note that Sroka and Braida (2005) and Meyer et al. (2011) used matched training conditions in their experiments which generally performs better than multi-condition training used in Meyer (2013). But at the same time the speech material used by the former two studies is more complex which partially compensates the effect of matched training leading to similar results. Meyer (2013) could further reduce the human-machine gap to 6.2 dB by using spectro-temporal Gabor features instead of MFCCs.

In experiments with a stationary speech-shaped noise performed in this study, the human-machine gap amounts to 5.2 dB and 4.8 dB with MFCCs and Gabor features, respectively, which are smaller than the previously reported results. Differences mainly arise from an increase in ASR performance which can be traced back to the use of context-dependent triphone targets instead of monophone targets, which has now become the standard since current training databases provide sufficient data for a more fine-grained estimation on triphone level. The use of DNNs improved performance with all features reducing the human-machine gap to 1.3, 0.3 and −0.5 dB for Filterbank features, MFCCs and Gabor features, respectively, i.e. the human-machine gap could ultimately be eliminated for this task. Experiments with different speakers and different noises show consistent results across speakers and maskers which highlights the applicability of DNN-based ASR systems as models for predicting SRTs of normal-hearing listeners. However, this study tested only two different noises and it has to be thoroughly investigated if this can be generalized to different noises. Nevertheless, DNN-based ASR systems offer various advantages over traditional models of speech intelligibility prediction, e.g., the exact time signal of speech and/or noise is not needed to predict speech intelligibility which makes it possible to predict intelligibility of unknown signals. Further, the system does not need to be calibrated to some previously measured human intelligibility, i.e. the system is reference-free.

### *4.2. Complex acoustical scenes recorded with multiple channels*

In complex spatial scenes with moving speakers and stationary diffuse noise the man-machine gap is heavily increased compared to the single-channel case. Human-listeners benefit from binaural cues, such as interaural phase and level differences (Beutelmann and Brand, 2006), and also from spectral difference between different speakers (Brungart, 2001; Brungart et al., 2001),which lowers the SRT to about −20 dB. At the same time, ASR has to be explicitly tailored to be able to make use of such cues. The ASR system used here exploits spatial information to estimate the azimuth angle of active sound sources and has been shown to improve performance in scenes with an interfering speaker compared to single-channel processing, which corresponds to better-ear-listening (Spille and Meyer, 2014; Spille et al., 2017).

At 0 and −5 dB, human listeners recognize around 90% of the presented words and at −10 dB HSR accuracy still is at 86%. These values can be compared to data from literature: Roman et al. (2003) performed listening experiments in scenes with a target speaker and a localized babble noise at 0 and 5° azimuth angle. Keyword scores were reported to be around 94, 72 and 20% at 0, −5 and −10 dB, respectively. Payton and Uchanski (1994) reported keyword scores to be at 79% on average for anechoic signals mixed with speech-shaped noise at 0 dB SNR and Kollmeier and Koch (1994) reported an average sentence score of around 53% for speech coming from 0° mixed with a speech-shaped masker at −90° at an SNR of −10 dB. These recognition rates are far below the ones obtained in this study which we attribute to several reasons: First, listeners were able to listen to the signals multiple times which is why the results in this work can be regarded as an upper bound for HSR. Second, through the fixed structure of sentences (combined with words that exhibit the same number of syllables) the relative temporal position of each word group is relatively well-known, which is an important difference to keyword spotting. Third, human speech intelligibility also depends on the spatial configuration of sound sources (Payton and Uchanski, 1994; Kollmeier and Koch, 1994; Bronkhorst, 2000; Roman et al., 2003; Beutelmann and Brand, 2006). The spatial separation of speech sources was comparatively large in our study, with an average separation angle of 60°. Roman et al. (2003) also tested a scenario with three sources at −30, 0 and 30° at −10 dB and reported human scores of 36% which is still 50% below this study's scores, but indicates that spatial separation improves recognition performance of human listeners which will be discussed later.

A detailed analysis of human performance in complex scenes with respect to the speakers gender and their location is in line with previous studies showing that humans can better segregate competing speakers if they differ in gender (Brungart, 2001; Chang, 2004). Darwin et al. (2003) showed that HSR in two-speaker scenes was significantly improved if the difference in fundamental frequency (F0) was greater than 2 semitones and if the ratio of vocal-tract lengths was greater than 1.08. When altering both F0 and vocal-tract length, which simulates a shift in gender, substantially larger improvements than differences in one of the two (F0 and vocal-tract length) alone were produced. These improvements were found to be similar to the improvements obtained by different-gender speakers. Differences in fundamental frequency and vocal-tract length are not implicitly exploited in the ASR system here. Brungart (2001) showed that listeners recognition scores are increased by 20–30% at target-to-masker ratios of -10 to 0 dB if the gender of masking and target speaker are different. Consistently, Chang (2004) reported an increase of recognition accuracy by around 20% in two-talker scenes when masker and target differ in gender. Note that these studies were performed with single-channel signals without any binaural information.

In the current study, benefit through different gender is not as prominent: human listeners recognized about 10% more words compared to the same-gender scenario. However, our acoustic scenes exhibit a higher complexity due to the movement of speakers and the presence of a stationary diffuse speech-shaped noise. Thus, a smaller gain than in single-channel signals was expected. The SRT in scenes where speakers had the same gender was found to be −19.4 dB which is 2.4 dB higher than the average SRT of human listeners. The same increase in SRT could be observed in scenes where both speakers crossed each other and in scenes where both speakers had the same gender and crossed each other. This increase in SRT was found to be 3.2 dB in total. These 3.2 dB represent the benefit of human listeners by exploiting differences in fundamental frequency and spatial differences of speakers. The already mentioned study of Roman et al. (2003) showed that at −10 dB, listeners recognize 20% of the presented words in a scenario with two sources where both are only separated by 5°. If the sources are further apart, i.e. 30° separation, scores increase by 16% to 36%. Here, at −10 dB SNR, differences between crossing, i.e. spatially close, speakers and non-crossing speakers are 3.4% but increase to up to 12.4% at −20 dB SNR. The ASR system employed in this study was trained with speech material from female and male talkers, which is the usual procedure for speaker-independent recognition, following the intuition that the ASR system should later generalize well both to new female and male speakers. Our results show that this seems not to be the case, since ASR is outperformed by far by HSR. A possible remedy might be the use of gender-specific models, or the inclusion of fundamental frequency differences of speakers on feature level such as periodicity features which can help to separate speakers of different gender (Josupeit et al., 2016).

A mismatch between training and test data can severely affect ASR and usually performance is much better when training and test data are very similar, e.g., if they contain the same noise type, which is referred to as

matched training. Experiments in spatial scenes in this study correspond to mismatched training, since ASR was trained on signals containing only one speaker in a diffuse stationary noise and tested on signals containing two speakers in diffuse noise. Thus, the ASR system has never seen any interfering speech before. This setup was chosen since it is unrealistic for most applications to have prior knowledge about the interfering speaker, hence matched training was not performed. It is still possible that ASR performance increases when a large number of different interfering talkers are considered for training, which will be investigated in future work. On the other hand, human listeners can easily cope with interfering speech independent of the specific interfering speaker. The auditory system seems to exploit several speaker specific cues, such as fundamental frequency and speaking rate, and can rapidly adapt to the specific combination of target and interferer at hand. Speaker adaptation techniques for ASR have not been investigated in the current study, but are also a possibility to increase recognition of the target speaker.

The test data used for HSR and ASR experiments is based on read matrix sentences, which were checked for undesired signal properties during postprocessing and therefore do not contain disfluencies. However, in realistic scenarios, disfluencies can be encountered which could increase ASR error rates (Goldwater et al., 2010) but presumably have a limited effect on HSR, thereby increasing the human-machine gap. As a solution to partially compensate this effect, disfluency in speech could be automatically detected (Mahesha and Vinod, 2015) to select an acoustic model optimized for disfluencies (Stolcke and Shriberg, 1996).

## 5. Summary and conclusions

This study compared DNN-based ASR systems to normal-hearing and hearing-impaired human listeners in simple acoustical scenes recorded with one audio channel and complex spatial scenarios recorded with multiple channels. The human-machine gap was measured in terms of the speech recognition threshold, i.e. the SNR at which 50% of the presented words were understood correctly.

Results show that DNN-based ASR systems can close the man-machine gap in single-channel recordings with stationary speech shaped noise and multi-talker babble noise. Auditory Gabor features in combination with a DNN resulted in the best performing ASR system in stationary noise, while Filterbank features were best in multi-talker babble noise. For different speakers, ASR results reflect differences in measured SRTs, i.e. speakers that more intelligible also produce lower SRTs in ASR.

In complex binaural scenes, listeners make excellent use of spatial and spectral cues, which is in stark contrast to ASR, which is strongly degraded in these scenes. This is reflected in a substantial man-machine gap of 12.3 dB in terms of SRT. When the true positions of speakers are supplied to the ASR system, this gap is reduced by about 2 dB which highlights the importance of robust estimation of direction of arrival. A further important difference between ASR and HSR was found when analyzing scenes with two speakers that are either of equal or different gender: When speech was presented in a situation with two speakers of different gender, the SRT in HSR was decreased by 3 dB, while this effect was not observed in ASR. This indicates that differences in fundamental frequency between the target and the interfering speaker are exploited by humans to separate both speakers but are neglected in ASR. Candidates to perform an efficient separation between speakers are features that make use of joint spectro-temporal features sensitive to periodicity cues, whereas the spectral smoothing that is applied in most current ASR systems appears to be unsuitable for this kind of speaker separation.

## Acknowledgments

Supported by the DFG (Cluster of Excellence EXC 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 Mathias Dietz, Volker Hohmann and especially Hendrik Kayser for valuable contributions to this work and Marc René Schädler for sharing the code of the Gabor feature calculation. Special thanks go to Daniel Marquardt for providing and helping with the code of the beamformer.

## Appendix

Table A.5
SRTs in dB of normal-hearing listeners and ASR systems for all 9 speakers in monaural scenes with stationary speech-shaped noise.

| Speaker | Listeners | Filterbank features | | MFCCs | | Gabor features | |
|---|---|---|---|---|---|---|---|
| | | GMM | DNN | GMM | DNN | GMM | DNN |
| F2 | −6.6 | 9.8 | −6.9 | 5.1 | −5.8 | −2.0 | −6.6 |
| M3 | −6.8 | 3.8 | −10.0 | −1.0 | −9.6 | −6.7 | −10.5 |
| M4 | −7.1 | 4.9 | −9.2 | −0.2 | −9.1 | −5.9 | −10.3 |
| OT | −7.3 | 3.3 | −6.2 | −2.0 | −7.1 | −2.7 | −7.9 |
| M1 | −7.4 | 5.7 | −9.3 | −0.4 | −8.9 | −5.4 | −9.5 |
| F3 | −8.6 | 2.3 | −9.3 | 1.0 | −8.8 | −5.0 | −9.0 |
| M2 | −9.1 | 0.2 | −8.9 | −1.7 | −8.8 | −6.5 | −10.2 |
| F4 | −9.3 | 3.0 | −9.5 | −1.3 | −9.5 | −5.9 | −10.2 |
| F1 | −9.7 | −1.5 | −10.6 | −2.0 | −9.8 | −7.6 | −11.7 |

Table A.6
SRTs in dB of normal-hearing listeners and ASR systems for all 9 speakers in monaural scenes with multi-talker babble noise.

| Speaker | Listeners | Filterbank features | | MFCCs | | Gabor features | |
|---|---|---|---|---|---|---|---|
| | | GMM | DNN | GMM | DNN | GMM | DNN |
| F2 | −5.2 | 12.5 | −2.1 | 16.6 | −1.0 | 6.2 | −0.2 |
| OT | −6.2 | 6.9 | −6.7 | 11.9 | −4.1 | 1.7 | −5.0 |
| M1 | −6.3 | 7.6 | −9.0 | 6.6 | −5.6 | 0.9 | −5.7 |
| M2 | −6.6 | 3.7 | −8.7 | 5.0 | −6.1 | −0.6 | −6.2 |
| F4 | −6.8 | 4.0 | −6.7 | 4.8 | −4.4 | 1.6 | −3.5 |
| M4 | −7.3 | 9.7 | −8.0 | 7.3 | −6.6 | 0.5 | −5.3 |
| M3 | −7.5 | 5.3 | −9.5 | 6.5 | −6.6 | −1.0 | −6.5 |
| F1 | −7.6 | 5.0 | −8.5 | 6.3 | −5.6 | 0.0 | −5.7 |
| F3 | −8.5 | 9.1 | −7.4 | 10.0 | −4.4 | 2.2 | −3.0 |

## References


Beutelmann, R., Brand, T., 2006. Prediction of speech intelligibility in spatial noise and reverberation for normal-hearing and hearing-impaired listeners. J. Acoust. Soc. Am. 120, 331–342.

Bitzer, J., Simmer, K.U., 2001. Superdirective microphone arrays. In: Brandstein, M., Ward, D. (Eds.), Microphone Arrays. Springer, pp. 1021–1042. doi: 10.1007/978-3-540-49127-9.

Brand, A., Behrend, O., Marquardt, T., McAlpine, D., Grothe, B., 2002. Precise inhibition is essential for microsecond interaural time difference coding. Nature 417 (6888), 543–547.

Brand, T., Kollmeier, B., 2002. Efficient adaptive procedures for threshold and concurrent slope estimates for psychophysics and speech intelligibility tests. J. Acoust. Soc. Am. 111 (6), 2801–2810. doi: 10.1121/1.1479152.

Bronkhorst, A.W., 2000. The cocktail party phenomenon: a review of research on speech intelligibility in multiple-talker conditions. Acta Acust. United Acust. 86 (1), 117–128. URL: https://dx.doi.org/10.3758/s13414-015-0882-9.

Brungart, D.S., 2001. Informational and energetic masking effects in the perception of two simultaneous talkers. J. Acoust. Soc. Am. 109 (3), 1101–1109. URL: http://www.ncbi.nlm.nih.gov/pubmed/11303924.

Brungart, D.S., Simpson, B.D., Ericson, M.a., Scott, K.R., 2001. Informational and energetic masking effects in the perception of multiple simultaneous talkers. J. Acoust. Soc. Am. 110 (5), 2527. doi: 10.1121/1.1408946.

Carey, M.J., Quang, T.P., 2005. A speech similarity distance weighting for robust recognition. In: Proceedings of the Ninth European Conference on Speech Communication and Technology, pp. 1257–1260. URL: papers2://publication/uuid/CF272B6B-78D6-4003-B384-47649CCA0186.

Chang, P., 2004. Exploration of behavioral, physiological, and computational approaches to auditory scene analysis. http://www.cse.ohio-state.edu/~dwang/pnl/theses/Chang_MSThesis04.pdf.

Cooke, M., Scharenborg, O., 2008. The interspeech 2008 consonant challenge. In: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, pp. 1765–1768.

Cox, H., Zeskind, R., Owen, M., 1987. Robust adaptive beamforming. IEEE Trans. Acoust. Speech Signal Process. 35 (10), 1365–1376. doi: 10.1109/TASSP.1987.1165054.

Darwin, C.J., Brungart, D.S., Simpson, B.D., 2003. Effects of fundamental frequency and vocal-tract length changes on attention to one of two simultaneous talkers. J. Acoust. Soc. Am. 114 (5), 2913. doi: 10.1121/1.1616924.

Davis, S.B., Mermelstein, P., 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans. Acoust. Speech Signal Process. 28, 357–366. URL: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1163420.

Dietz, M., Ewert, S.D., Hohmann, V., 2011. Auditory model based direction estimation of concurrent speakers from binaural signals. Speech Commun. 53 (5), 592–605. doi: 10.1016/j.specom.2010.05.006.

Dreschler, W.A., Verschuure, H., Ludvigsen, C., Westermann, S., 2001. ICRA noises: artificial noise signals with speech-like spectral and temporal properties for hearing instrument assessment. International Collegium for Rehabilitative Audiology. Audiol. Off. Organ Int. Soc. Audiol. 40 (3), 148–157. doi: 10.3109/00206090109073110.

Goldwater, S., Jurafsky, D., Manning, C.D., 2010. Which words are hard to recognize? prosodic, lexical, and disfluency factors that increase speech recognition error rates. Speech Commun. 52 (3), 181–200. doi: 10.1016/j.specom.2009.10.001.

Hartikainen, J., Särkkä, S., 2008. RBMCDAbox-matlab toolbox of rao−blackwellized data association particle filters. Technical Report. Department of Biomedical Engineering and Computational Science. URL: http://www.lce.hut.fi/research/mm/rbmcda/documentation.pdf.

Hinton, G., Deng, L., Yu, D., Dahl, G.E., Mohamed, A.-r., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T.N., Kingsbury, B., 2012. Deep neural networks for acoustic modeling in speech recognition. IEEE Signal Process. Mag. 82–97. doi: 10.1109/MSP.2012.2205597.

Hinton, G., Osindero, S., Teh, Y., 2006. A fast learning algorithm for deep belief nets. Neural computation 18 (7), 1527–1554. URL: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6796673.

Hochmuth, S., Jürgens, T., Brand, T., Kollmeier, B., 2015a. Talker- and language-specific effects on speech intelligibility in noise assessed with bilingual talkers: which language is more robust against noise and reverberation? Int. J. Audiol. 2027, 1–12. doi: 10.3109/14992027.2015.1088174.

Hochmuth, S., Kollmeier, B., Brand, T., Jürgens, T., 2015b. Influence of noise type on speech reception thresholds across four languages measured with matrix sentence tests. Int. J. Audiol. 54, 1499–2027. doi: 10.3109/14992027.2015.1046502. ISSNOnline

Holube, I., Fredelake, S., Vlaming, M., Kollmeier, B., 2010. Development and analysis of an International Speech Test Signal (ISTS). Int. J. Audiol. 49 (12), 891–903. doi: 10.3109/14992027.2010.506889.

Josupeit, A., Kopčo, N., Hohmann, V., 2016. Modeling of speech localization in a multi-talker mixture using periodicity and energy-based auditory features. J. Acoust. Soc. Am. 139 (5), 2911–2923. doi: 10.1121/1.4950699.

Juergens, T., Fredelake, S., Meyer, R.M., Kollmeier, B., Brand, T., 2010. Challenging the speech intelligibility index : macroscopic vs . microscopic prediction of sentence recognition in normal and hearing-impaired listeners. In: Proceedings of the Interspeech, pp. 2478–2481.

Jürgens, T., Brand, T., 2009. Microscopic prediction of speech recognition for listeners with normal hearing in noise using an auditory model. J. Acoust. Soc. Am. 126 (5), 2635–2648. doi: 10.1121/1.3224721.

Kayser, H., Ewert, S.D., Anemüller, J., Rohdenburg, T., Hohmann, V., Kollmeier, B., 2009. Database of multichannel in-ear and behind-the-ear head-related and binaural room impulse responses. EURASIP J. Adv. Signal Process.(1). doi: 10.1155/2009/298605.

Kollmeier, B., Koch, R., 1994. Speech enhancement based on physiological and psychoacoustical models of modulation perception and binaural interaction.. J. Acoust. Soc. Am. 95 (3), 1593–1602. URL: http://www.ncbi.nlm.nih.gov/pubmed/8176062.

Kollmeier, B., Warzybok, A., Hochmuth, S., Zokoll, M.A., Uslar, V., Brand, T., Wagener, K.C., 2015. The multilingual matrix test: principles, applications, and comparison across languages: A review. Int. J. Audiol. 54 (Supl 2), 3–16. doi: 10.3109/14992027.2015.1020971.

Lippmann, R., 1997. Speech recognition by machines and humans. Speech Commun. 22 (1), 1–15. URL: sciencedirect.com/science/article/pii/S0167639397000216.

Mahesha, P., Vinod, D.S., 2015. Support vector machine-based stuttering dysfluency classification using gmm supervectors. Int. J. Grid Util. Comput. 6 (3/4), 143–149. doi: 10.1504/IJGUC.2015.070680.

Mandel, M.I., Yoho, S.E., Healy, E.W., 2016. Measuring time-frequency importance functions of speech with bubble noisea. J. Acoust. Soc. Am. 140 (4), 2542–2553. doi: 10.1121/1.4964102.

Mesgarani, N., Stephen, D., Shamma, S., 2007. Representation of phonemes in primary auditory cortex: how the brain analyzes speech. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP).

Meyer, B.T., 2013. What's the difference? Comparing humans and machines on the Aurora 2 speech recognition task. In: Proceedings of the Interspeech, pp. 2634–2638.

Meyer, B.T., Brand, T., Kollmeier, B., 2011. Effect of speech-intrinsic variations on human and automatic recognition of spoken phonemes. J. Acoust. Soc. Am. 129 (1), 388–403. doi: 10.1121/1.3514525.

Meyer, B.T., Kollmeier, B., 2011. Robustness of spectro-temporal features against intrinsic and extrinsic variations in automatic speech recognition. Speech Commun. 53 (5), 753–767. doi: 10.1016/j.specom.2010.07.002.

Meyer, B.T., Kollmeier, B., Ooster, J., 2015. Autonomous measurement of speech intelligibility utilizing automatic speech recognition. In: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, pp. 2982–2986.

Meyer, B.T., Spille, C., Kollmeier, B., Morgan, N., 2012. Hooking up spectro-temporal filters with auditory-inspired representations for robust automatic speech recognition. In: Proceedings of the Interspeech.

Mohamed, A.-r., Dahl, G.E., Hinton, G., 2012. Acoustic modeling using deep belief networks. IEEE Trans. Audio Speech Lang. Process. 20 (1), 14–22. doi: 10.1109/TASL.2011.2109382.

Payton, K.L., Uchanski, R.M., Bradia, L.D., 1994. Intelligibility of conversational and clear speech in noise and reverberation for listeners with normal and impaired hearing. The Journal of the Acoustical Society of America. 95 (3), 1581–1592.

Peskin, B., Connolly, S., Gillick, L., Lowe, S., McAllaster, D., Nagesha, V., 1996. Improvements in switchboard recognition and topic identification. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). IEEE, pp. 303–306. doi: 10.1109/ICASSP.1996.540418.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., Vesely, K., 2011. The Kaldi speech recognition toolkit. In: Proceedings of the IEEE 2011 Workshop on Automatic Speech Recognition and Understanding. URL: https://infoscience.epfl.ch/record/192584.

Qiu, A., Schreiner, C.E., Escabí, M.A., 2003. Gabor analysis of auditory midbrain receptive fields: spectro-temporal and binaural composition. J. Neurophysiol. 90 (1), 456–476. doi: 10.1152/jn.00851.2002.

Roman, N., Wang, D., Brown, G.J., 2003. Speech segregation based on sound localization. J. Acoust. Soc. Am. 114 (4), 2236–2252. doi: 10.1121/1.1610463.

Rosenblatt, F., 1957. The perceptron, a perceiving and recognizing automaton project para. Cornell Aeronautical Laboratory. URL: https://books.google.de/books?id=P_XGPgAACAAJ.

Rumelhart, D.E., Hinton, G.E., Williams, R.J., 1986. Learning representations by back-propagating errors. Nature 323 (6088), 533–536. doi: 10.1038/323533a0.

Saon, G., Kuo, H.-K. J., Rennie, S., Picheny, M., 2015. The IBM 2015 english conversational telephone speech recognition system. In: Proceedings of the Interspeech, pp. 3–7. doi: 10.21437/Interspeech.2017-405.

Saon, G., Kurata, G., Sercu, T., Audhkhasi, K., Thomas, S., Dimitriadis, D., Cui, X., Ramabhadran, B., Picheny, M., Lim, L.-L., Roomi, B., Hall, P., 2017. English conversational telephone speech recognition by humans and machines. In: Proceedings of the Interspeech 2017, pp. 132–136. doi: 10.21437/Interspeech.2017-405.

Särkkä, S., Vehtari, A., Lampinen, J., 2007. Rao-Blackwellized particle filter for multiple target tracking. Inf. Fusion 8 (1), 2–15. doi: 10.1016/j.inffus.2005.09.009.

Schubotz, W., Brand, T., Kollmeier, B., Ewert, S.D., 2016. Monaural speech intelligibility and detection in maskers with varying amounts of spectro-temporal speech features. J. Acoust. Soc. Am. 140 (1), 524–540. doi: 10.1121/1.4955079.

Shen, W., Olive, J., Jones, D., 2008. Two protocols comparing human and machine phonetic recognition performance in conversational speech. In: Proceedings of the Interspeech, pp. 1630–1633. URL: http://20.210-193-52.unknown.qala.com.sg/archive/archive_papers/interspeech_2008/i08_1630.pdf.

Shen, X., Oualil, Y., Greenberg, C., Singh, M., Klakow, D., 2017. Estimation of gap between current language models and human performance. In: Proceedings of the Interspeech 2017, pp. 553–557. doi: 10.21437/Interspeech.2017-729.

Spille, C., Dietz, M., Hohmann, V., Meyer, B.T., 2013a. Using binarual processing for automatic speech recognition in multi-talker scenes. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 7805–7809. URL: http://www.uni-oldenburg.de/fileadmin/user_upload/mediphysik/ag/mediphysik/download/paper/spille/CSP_ICASSP_final.pdf; http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6639183.

Spille, C., Kollmeier, B., Meyer, B.T., 2017. Combining binaural and cortical features for robust speech recognition. IEEE/ACM Trans. Audio Speech Lang. Process. 25 (4), 756–767. doi: 10.1109/TASLP.2017.2661712.

Spille, C., Meyer, B.T., 2014. Identifying the human-machine differences in complex binaural scenes: what can be learned from our auditory system. In: Proceedings of the Interspeech, pp. 626–630.

Spille, C., Meyer, B.T., Dietz, M., Hohmann, V., 2013b. Binaural scene analysis with multi-dimensional statistical filters. In: Blauert, J. (Ed.), The Technology of Binaural Listening. Springer, Berlin-Heidelberg-New York, NY.

Sroka, J.J., Braida, L.D., 2005. Human and machine consonant recognition. Speech Commun. 45 (4), 401–423. doi: 10.1016/j.specom.2004.11.009.

Stolcke, A., Droppo, J., 2017. Comparing human and machine errors in conversational speech transcription. In: Proceedings of the Interspeech 2017, pp. 137–141. doi: 10.21437/Interspeech.2017-1544.

Stolcke, A., Shriberg, E., 1996. Statistical language modeling for speech disfluencies. In: Proceedings of the 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP-96, 1. IEEE, pp. 405–408.

Wagener, K., Brand, T., Kollmeier, B., 1999. Development and evaluation of a German sentence test Part III: evaluation of the Oldenburg sentence test. Z. Audiol. 38 (3), 5–15.

Xiong, W., Droppo, J., Huang, X., Seide, F., Seltzer, M., Stolcke, A., Yu, D., Zweig, G., 2016. Achieving human parity in conversational speech recognition. pp. 86–95. arXiv: 1610.05256v1.