

# Proceedings of Meetings on Acoustics

---

Volume 14, 2011

<http://acousticalsociety.org/>

---

**162nd Meeting  
Acoustical Society of America  
San Diego, California  
31 October - 4 November 2011  
Session 2pSCa: Speech Communication**

---

## **2pSCa2. Improving automatic speech recognition by learning from human errors**

**Bernd T. Meyer\***

**\*Corresponding author's address: Dept. of Physics, Medizinische Physik, Carl von Ossietzky Universität Oldenburg, Oldenburg, 26111, Niedersachsen, Germany, [bernd.meyer@uni-oldenburg.de](mailto:bernd.meyer@uni-oldenburg.de)**

This work presents a series of experiments that compare the performance of human speech recognition (HSR) and automatic speech recognition (ASR). The goal of this line of research is to learn from the differences between HSR and ASR, and to use this knowledge to incorporate new signal processing strategies from the human auditory system in automatic classifiers. A database with noisy nonsense utterances is used both for HSR and ASR experiments with focus on the influence of intrinsic variation (arising from changes in speaking rate, effort, and style). A standard ASR system is found to reach human performance level only when the signal-to-noise ratio is increased by 15 dB, which can be seen as the human-machine gap for speech recognition on a sub-lexical level. The sources of intrinsic variation are found to severely degrade phoneme recognition scores both in HSR and in ASR. A comparison of utterances produced at different speaking rates indicates that temporal cues are not optimally exploited in ASR, which results in a strong increase of vowel confusions. Alternative feature extraction methods that take into account temporal and spectro-temporal modulations of speech signals are discussed.

---

Published by the Acoustical Society of America through the American Institute of Physics

## 1 Introduction

Automatic speech recognition (ASR) has come a long way in the last decades, from the speaker-dependent recognition of isolated digits to large vocabulary, speaker-independent recognizers used in commercial systems. Still, machines that recognize speech as well as the healthy human auditory system have not yet been realized. In contrast to ASR, human speech recognition (HSR) is very robust in the presence of variability in spoken language. This variability can be caused by either extrinsic sources (e.g., additive noise or reverberation) or intrinsic sources (speaker- and speech-related factors such as gender, emotional state, age, or speaking style), and human listeners can adapt very well to both of these (Benzeguiba et al., 2007).

This paper summarizes experiments and results published in (Meyer et al., 2010) and (Meyer et al., 2011). The aim is to measure the gap between HSR and ASR, and to identify the specific differences between our auditory system and standard ASR systems. The outcome of these experiments is potentially useful for incorporating novel signal processing strategies into ASR to increase its robustness, and simultaneously reducing the human-machine gap. An overview of the experiments presented in this study is shown in Fig. 1: Utterances from a database of nonsense words (referred to as logatomes) were presented to human listeners and also used as input to an ASR system. In each case, the task was to identify the central phoneme in vowel-consonant-vowel (VCV) or consonant-vowel-consonant (CVC) combinations, which lays the focus on the sublexical level.

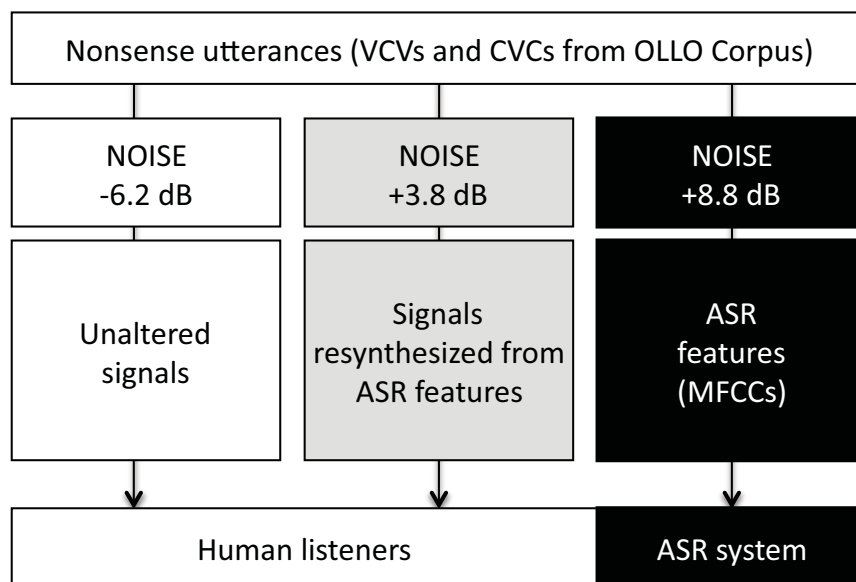


Fig. 1. Original and resynthesized, noisy signals were presented to six listeners. The same utterances were used as input to an ASR system. In both cases, the task was the identification of the central phoneme.

Further, it was investigated if the most common features in ASR (mel-cepstral coefficients, MFCCs) contain sufficient information so that human listeners are able to identify noisy phonemes when those features are resynthesized (i.e., are converted to audible signals). The idea is that HSR and ASR are provided with identical or similar information. When the information relevant for the recognition of noisy speech is retained during feature calculation, the intelligibility of original and resynthesized speech should be identical in listening tests. On the other hand, an increase of error rates when using resynthesized instead of original signals could be exploited to quantify the loss of information that is relevant for speech recognition. Finally, the effect of intrinsic variation in spoken language was analyzed by using speech stimuli that were produced with different speaking rates, efforts, and styles. The effect of such variations compared to normally produced speech was evaluated both in listening tests and ASR experiments.

The next section gives a short description of the speech database, the resynthesis of ASR features, and the experimental setup for HSR and ASR tests. The results and the discussion are presented in Sections 3 and 4, respectively.

## 2 Methods

### 2.1 *Speech database*

The Oldenburg Logatome Corpus (OLLO) is a database that was designed for speech intelligibility tests with human listeners and for experiments with automatic classifiers (Meyer et al., 2010). It consists of nonsense utterances or logatomes, i.e., words without semantic meaning which comply with phonetic and phonotactic rules. The logatomes are composed of triplets of vowels (V) and consonants (C) with the outer phonemes being identical. The central phonemes in utterances were /b/, /d/, /f/, /g/, /k/, /l/, /m/, /n/, /p/, /s/, /ʃ/, /t/, /v/, /ts/, /a/, /ɛ/, /ɪ/, /ɔ/, /ʊ/, /a:/, /e/, /i/, /o/, /u/. 70 VCVs and 80 CVCs were recorded with different speaking styles, efforts and speaking rates, thus enabling an analysis of the effect of such intrinsic variations of speech. During the recordings, 50 talkers were asked to produce each logatome either normally or in one of five variations (fast, slow, loud, soft, question, the latter referring to utterances with rising pitch). The corpus is freely available for research at '<http://medi.uni-oldenburg.de/ollo>'.

For HSR and ASR testing, a subset was selected from the database that was comprised of data from 4 talkers (2 male, 2 female) and contained logatomes for all six categories or speaking styles (normal + 5 variations). These speaking styles were equally distributed in the test set that contained 3,600 recordings (150 (logatomes)  $\times$  6 (speaking styles)  $\times$  4 (talkers)).

## 2.2 *Resynthesis of ASR features*

Mel-frequency cepstral coefficients encode the spectral envelope of short-time segments of speech. In order to calculate MFCC features from speech, signals with 16 kHz sampling frequency are windowed with 30 ms Hann windows and a frame shift of 10 ms. Each frame undergoes the same processing steps: Calculation of the amplitude spectrum, reduction of the frequency resolution using a mel-scaled filter bank and calculation, log-compression of the amplitude values, and application of the inverse discrete cosine transformation (IDCT). The 12 lowest coefficients plus an additional energy feature are selected for the ASR experiments and HSR tests with resynthesized speech.

This processing scheme results in a loss of spectral fine structure and phase information, which may be detrimental for speech recognition in noisy environments. In order to generate audible signals from MFCCs, an algorithm proposed by Demuynck et al. (2004) has been used, which uses a linear neural net to obtain a mel-spectrogram from the cepstral coefficients. Since information about the original excitation signal is discarded in MFCCs (and therefore not used by standard ASR), an artificial excitation signal needs to be employed because the addition of voicing for human listeners would be an advantage over the ASR system. Pilot experiments with noisy and periodic signals with a fixed fundamental frequency were performed to estimate an excitation signal that results in a high intelligibility, and a periodic pulse train with a frequency of 130 Hz was found to produce good results and hence was chosen for the resynthesis (with signals that sounded artificial, but were perfectly understandable in the absence of noise).

## 2.3 *Experiments with human listeners*

The selected OLLO subset was presented to six normal-hearing listeners (three male, three female, aged 18 to 35), who were asked to identify the central phoneme in a 1-out-of-N forced-choice paradigm. In case of clean speech, the phoneme error rates for original and resynthesized signals was approximately 1%. To enable the statistical analysis errors, a stationary masking noise with speech-shaped frequency characteristics (Dreschler et al., 2001) was added to the signals. Pilot experiments with a small test set and one normal-hearing subject showed that an SNR of -6.2 dB for original signals and +3.8 dB for resynthesized signals result in error rates between 20-40%. These SNRs were used for the HSR measurements.

Randomized sequences of logatomes were presented in a soundproof booth via audiological headphones (Sennheiser HDA200) after an online free-field equalization was performed. After a training phase, listeners were presented

a randomized sequence of logatomes at a level of 70 dB SPL, which was the preferred level for most listeners. After each presentation, an item had to be selected from a list of logatomes, which triggered the presentation of the next listening item after a short pause. The number of choice alternatives was either 10 (corresponding to the 10 central vowels used in CVCs) or 14 (since 14 different central phonemes are used for the VCVs).

## 2.4 ASR experiments

For the ASR experiments, mel-frequency cepstral coefficients (MFCCs) and their discrete temporal derivatives (delta and double-delta coefficients) were calculated from clean and noisy speech files. A standard Hidden Markov model (HMM) with three states per phoneme and eight Gaussian mixtures per state implemented in HTK was used as back end. The system was set up to resemble the 1-out-of-N identification task used for HSR, i.e., the recognizer identified the central phoneme in VCV and CVC utterances. Further, the test utterances were identical to the speech data used for HSR experiments, with the additional repetitions that were recorded for the OLLO database. ASR training was performed with data from six talkers not contained in the testing data, which resulted in a speaker and gender-independent ASR setup. Recognition of noisy utterances was tested at several SNRs (ranging from -6.2 to +8.8 dB) with matched conditions for training and testing. The same noise signal as for HSR (Dreschler et al., 2001) was used.

## 3 Results

### 3.1 Overall results

The overall phoneme error rates for ASR and HSR averaged over all factors of intrinsic variation and all phonemes in the database are shown in Fig. 2. When HSR and ASR are tested at the same SNR of -6.2 dB (Labels A and E in Fig. 2), the ASR error rates are more than twice as high as the HSR error rates. The error rates for A, B, and C are in the same range, and can therefore be used to estimate the overall human-machine gap in terms of the SNR: The ASR system reaches human performance levels only when the SNR is increased by 15 dB. The results also indicate that standard ASR features do not carry all information required to decode a speech signal in noise, since the error rates with original and resynthesized signals are similar (A and B), although the SNR for resynthesized speech is 10 dB higher than for original signals.

On the other hand, even when humans are provided with the information that is also available to the MFCC-based recognizer, they still perform better than ASR: The error rates with resynthesized signals (B) are comparable to the ASR result with an SNR of 8.8 dB (C), corresponding to an SNR gap of 5 dB, which is an estimate for the gap introduced by the back end. When identical SNR conditions are compared (B and D), the relative increase of errors due to the backend can be estimated to be 30% (absolute increase: 8.3%) for this medium range of speech intelligibility.

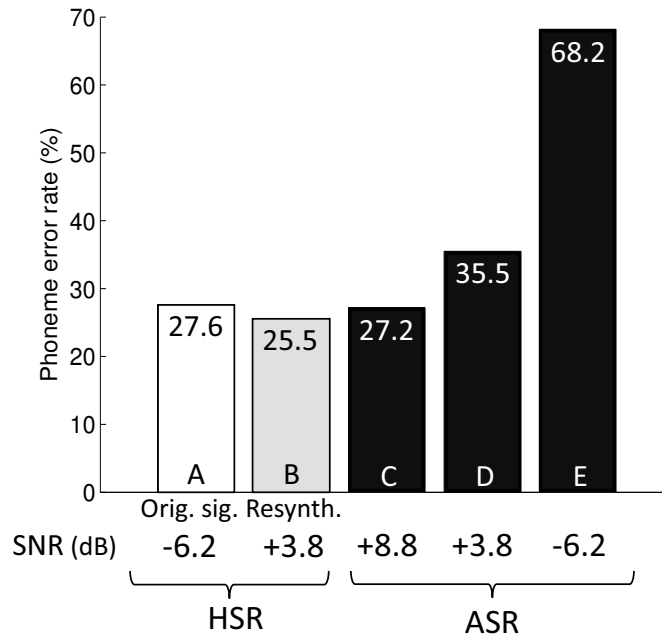


Fig. 2. Phoneme error rates obtained in HSR and ASR experiments at different SNRs (denoted below each bar).

### 3.2 Effect of intrinsic variability

The effect of intrinsic variation in speech was analyzed by breaking down the average scores with respect to the speaking styles represented in the OLLO database. Table 1 shows the relative increase of error rates for varied speaking rate, effort, and pitch for SNR conditions that yield similar average error rates. The relative increase was calculated based on the corresponding scores for the reference condition ('normal speaking style'). In almost all cases, changes compared to the reference condition increase the error rates, i.e., the presence of intrinsic variability covered in this study increases the phoneme error rates, which is observed both for HSR and ASR. A condition that consistently results in strong increases of error rates is a high speaking rate. Hence, the relationship of phoneme duration and error rates was analyzed more closely.

	SNR	Normal	Fast	Slow	Loud	Soft	Question
HSR (Orig. signals)	-6.2	0.0	35.0	-5.9	31.6	32.9	4.2
HSR (Resynth. sig.)	3.8	0.0	29.4	6.5	-3.3	71.5	10.7
ASR	8.8	0.0	63.8	10.2	68.4	55.6	33.7

Table 1

Relative increase of phoneme error rates for different speaking styles. The increase is shown for noise conditions that produced similar average (absolute) error rates (HSR, orig: 25.5%; HSR, resynth.: 27.6%; ASR: 27.2%).

A forced-alignment procedure that takes into account pronunciation variations of utterances was used to obtain estimates for the phoneme durations (Kipp et al., 1996). The durations of central phonemes were compared to the corresponding error rates in HSR and ASR. Fig. 3 shows the general trend that was observed for central vowel phonemes for HSR and ASR at the same SNR. In HSR, two groups of vowels emerge for which an increased phoneme duration either increases the error rate ( $/a/$ ,  $/\varepsilon/$ ,  $/i/$ ,  $/o/$ ,  $/u/$ ) or decreases the error rate ( $/a:/$ ,  $/e/$ ,  $/i/$ ,  $/o/$ ,  $/u/$ ). The observation that durational cues affect HSR error rates (although most of the vowel phonemes also differ with respect to their spectral properties) is in accordance with earlier studies (Hillenbrand et al., 1995). When the data obtained at the same SNR with ASR is plotted, the same trend as for HSR is observed for two phonemes ( $/a/$  and  $/a:/$ ). However, in all other cases, either no clear trend is observed (for instance for  $/i/$ ) or the opposite trend as in HSR is found (an example being  $/u/$ ). These examples are highlighted by the blue and green lines in Fig. 3. This result shows that temporal cues are not optimally exploited in standard ASR systems, and suggests to pay more attention to temporal processing. One approach to do this is to incorporate more information about the temporal context on feature level, e.g., by using spectro-temporal Gabor filters (Meyer and Kollmeier, 2011) that can be parametrized to either perform a purely spectral processing (and thereby mimicking the functionality of MFCC features), spectro-temporal processing (to detect formant transients that are often represented by diagonal structures in the spectrogram), or temporal processing (which might help to distinguish between phonemes with different durations).

## 4 Summary

The comparison of human and automatic speech recognition on a sub-lexical task showed that a large gap between HSR and ASR still exists. In the presence of stationary noise, ASR error rates were more than twice as high as for HSR. The comparison of noise conditions that resulted in comparable error rates enabled an estimation of the human-machine gap in terms of the SNR, since a standard ASR system reached human performance only when the SNR

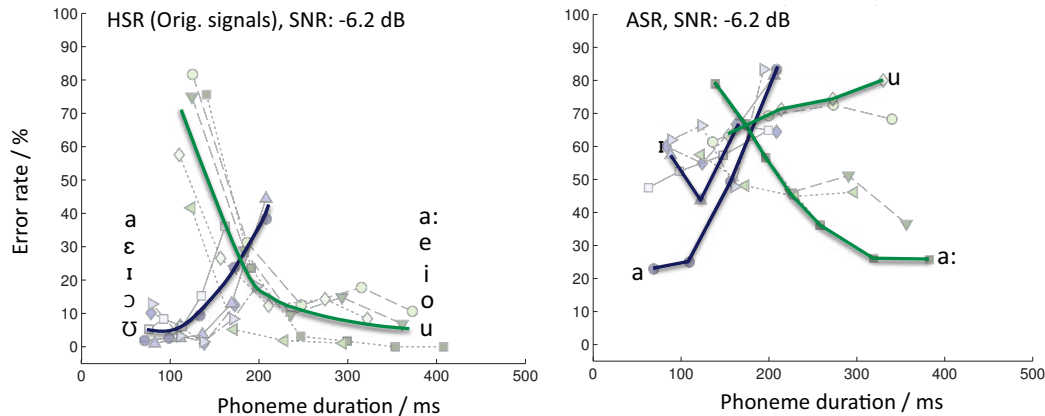


Fig. 3. Phoneme duration of central vowels in CVC utterances vs. the corresponding error rate. In HSR (left panel), two groups of vowels can be identified for which the error rates either increase or decrease with longer phoneme durations. With the exception of the confusion pair ( $/a/$ ,  $/a:/$ ), this trend cannot be observed in ASR (right panel).

was increased by 15 dB. When human listeners were supplied with the information that is available to an ASR backend, the SNR needed to be increased by 10 dB to achieve similar results as with original signals. This shows that even the best back end that we know (the human auditory system) cannot extract the acoustic information from a feature-based signal as well as from original signals, and information relevant for speech recognition seems to be neglected in standard feature extraction.

A closer look at the differences of error rates in the presence of specific factors of intrinsic variation showed that changes in speaking rate severely affect speech recognition. This was the case both in HSR and ASR; however, consistent patterns for specific groups of vowels were observed only in HSR. This result suggest that temporal information is not optimally exploited in current recognizers, and more attention should be paid to temporal processing.

## 5 Acknowledgements

Significant contributions to the research summarized in this study were made by Birger Kollmeier, Thomas Brand, Tim Jürgens, and Thorsten Wesker. It was supported by the DFG (SFB/TRR 31 'The active auditory system'; URL: <http://www.uni-oldenburg.de/sfbtr31>). Bernd T. Meyer has been supported by a post-doctoral fellowship of the German Academic Exchange Service (DAAD).

## References

- Benzeguiba, M., De Mori, R., Deroo, O., Dupont, S., Erbes, T., Jouvét, D., Fissore, L., Laface, P., Mertins, A., and Ris, C. (2007). "Automatic speech recognition and speech variability: A review, *Speech Commun.* 49," 763-786.
- Demuynck, K., Garcia, O., and van Compernelle, D. (2004). "Synthesizing Speech from Speech Recognition Parameters," In *Proc. Interspeech*, pp. 945-948.
- Dreschler, W. A., H, V., Ludvigson, C., and Westermann, S. (2001). "ICRA Noises: Artificial Noise Signals with Speech-like Spectral and Temporal Properties for Hearing Instrument Assessment," *Audiology*, 40(3), 148-157.
- Kipp, A., Wesenick, M.-B., and Schiel, F. (1996). "Automatic detection and segmentation of pronunciation variants in German speech corpora," in *Proc. of the International Conference on Spoken Language Processing (ICSLP)*, pp. 106-109.
- Hillenbrand, J., Getty, L., Clark, M., and Wheeler, K. (1995). "Acoustic characteristics of American English vowels," *J. Acoust. Soc. Am.* 97, pp. 3099-3111.
- Meyer, B. and Kollmeier, B. (2011). "Robustness of spectro-temporal features against intrinsic and extrinsic variations in automatic speech recognition," *Speech Comm.* 53 (5), pp. 753-767.
- Meyer, B. T., Jürgens, T., Wesker, T., Brand, T., and Kollmeier, B. (2010). "Human speech recognition as a function of speech-intrinsic variabilities," *J. Acoust. Soc. Am.* 128 (5), pp. 3126-3141
- Meyer, B.T., Brand, T., and Kollmeier, B. (2011). "Effect of speech-intrinsic variations on human and automatic recognition of spoken phonemes," *J. Acoust. Soc. Am.* 129, pp. 388-403.