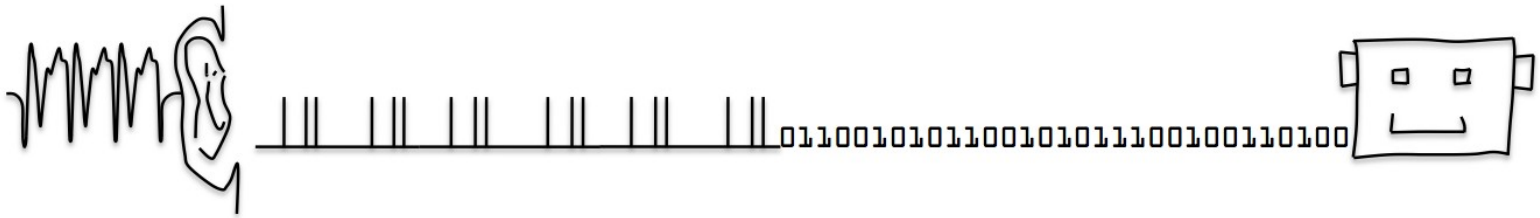
What's the difference? Quantifying errors in human and automatic speech recognition

Bernd T. Meyer
Medical Physics Group
University of Oldenburg

Workshop Errare
Oct 21st 2013

Motivation: The auditory approach

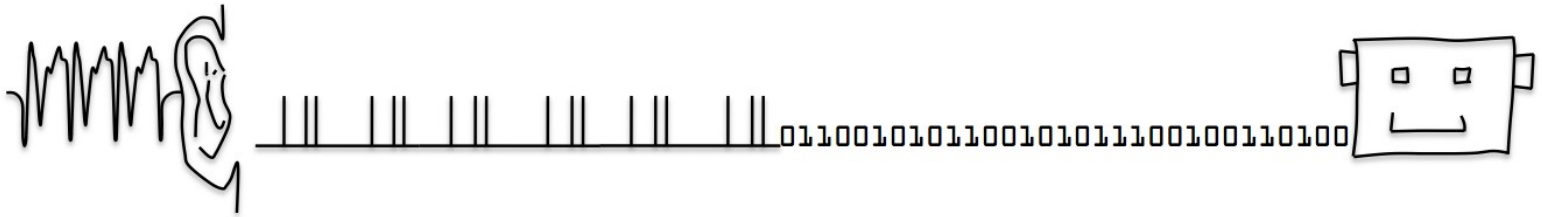


Human speech recognition (HSR)

Automatic speech recognition (ASR)

- Motivation: Our auditory system is very good at extracting the relevant cues for speech recognition
- Aims: Improvement of speech processing by considering principles of the human auditory system
- One approach: Learn about HSR-ASR differences by comparing humans and ASR

Motivation: The auditory approach



Human speech recognition (HSR)

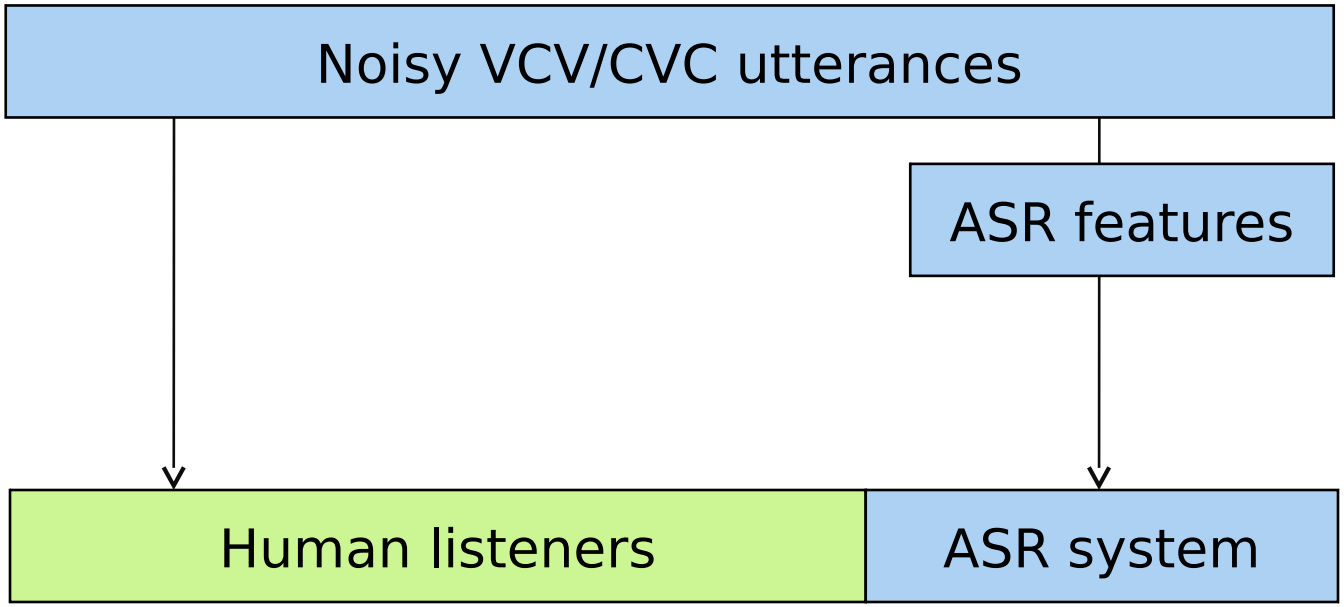
Automatic speech recognition (ASR)

Man-machine comparison

...on sublexical level with focus on speech-intrinsic variability

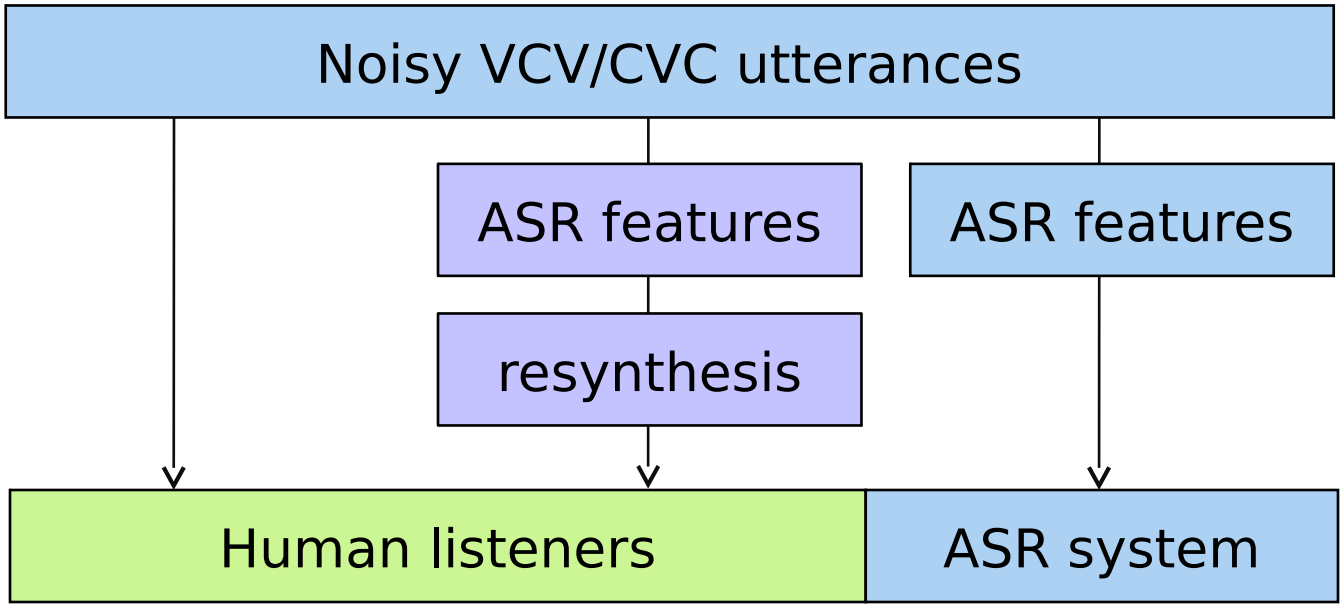
...on word level (noisy digit strings)

Man-machine comparison: Overview



Identical (nonsense) utterances for HSR and ASR tests
 focus is laid on sublexical level

Man-machine comparison: Overview



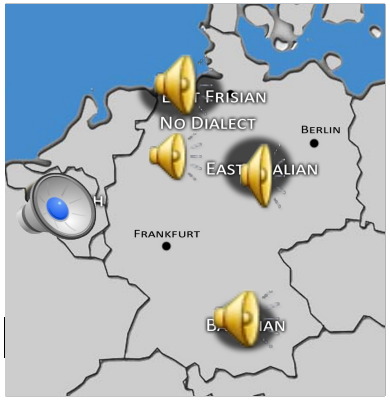
Is the information contained in standard features (MFCCs) sufficient for human listeners to recognize speech?

- Resynthesize MFCCs to audible signals
- Human listeners and ASR systems get similar information

Oldenburg Logatome Corpus

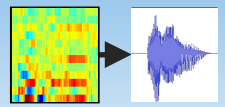
- Logatomes: Simple phoneme combinations(vowel-consonant-vowel, consonant-vowel-consonant)
- Suitable for HSR und ASR experiments
- 150 different logatomes
- 50 speakers
- Sources for intrinsic variability
  - Speaking rate (fast vs. slow)
  - Speaking effort (loudly and softly spoken utts.)
  - Speaking style (rising pitch/question and normal
  - Dialect and accent
- Freely available at http://medi.uni-oldenburg.de/ollo

p a p p
p i p p
p o h p
p u p p
p a h p
p e p p
p i e p
p u h p
p e h p
p o p p



Meyer, Jürgens, Wesker, Brand, Kollmeier (2010). "Human phoneme recognition as a function of speech-intrinsic variabilities." JASA 128

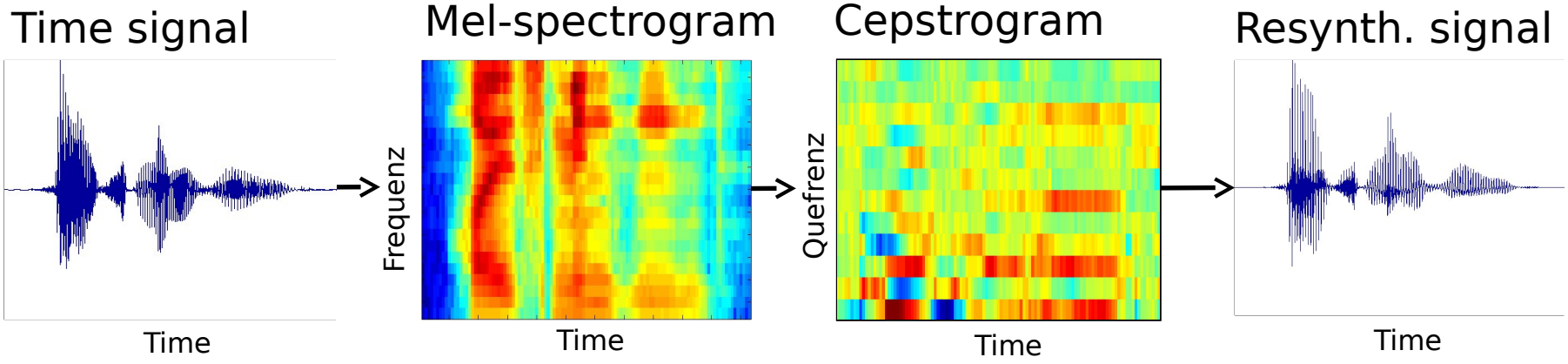
ASR features and resynthesis



- Cepstral coefficients: Encode the spectral envelope of the short-time Fourier transform

$$x_c(n) = \mathcal{F}^{-1}(\log|\mathcal{F}(x(n)|)$$

- Loss of information
  - Phase und fine structure need to be estimated for resynthesis (algorithm supplied by Demuynck et al., 2004)
  - Fine structure: Pulse train with fixed fundamental frequency

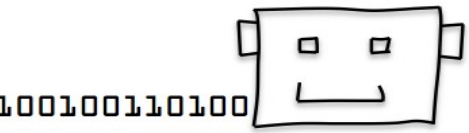


HSR: Speech data and measurements

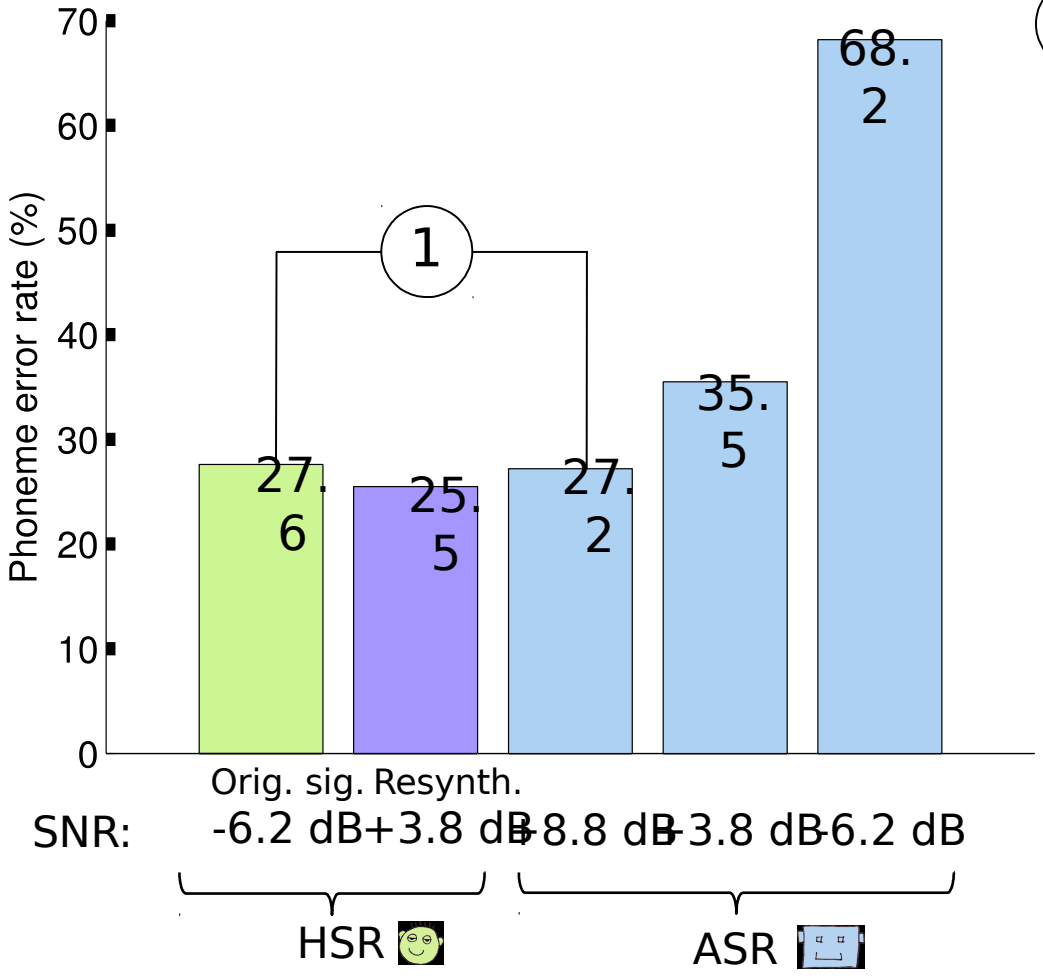
- Aim: Quantification of the influence of speaking rate, style, and effort
- Stimuli: Utterances from 4 talkers
- Pilot experiments: Estimation of SNR for 60-80% phoneme recognition rate
- Masking noise: Stationary noise with speech-shaped frequency characteristics
- Six normal-hearing subjects
- Task: identification of central phoneme
- Each subject listened to 2 (orig. & resynth. signals) ×3600 utterances (4 talkers ×6 variabilities ×150 logatomes)

ASR: experimental setup

- Features: Cepstral coefficients with 13 components + delta and acceleration coefficients
- Classifier: Hidden Markov Model Toolkit (HTK), 3 states, 8 Gaussian mixtures per state
- Phoneme recognizer, closed test
- Test data: Same as in HSR
- Training data: 6 talkers w/o dialect (speaker-independent ASR)



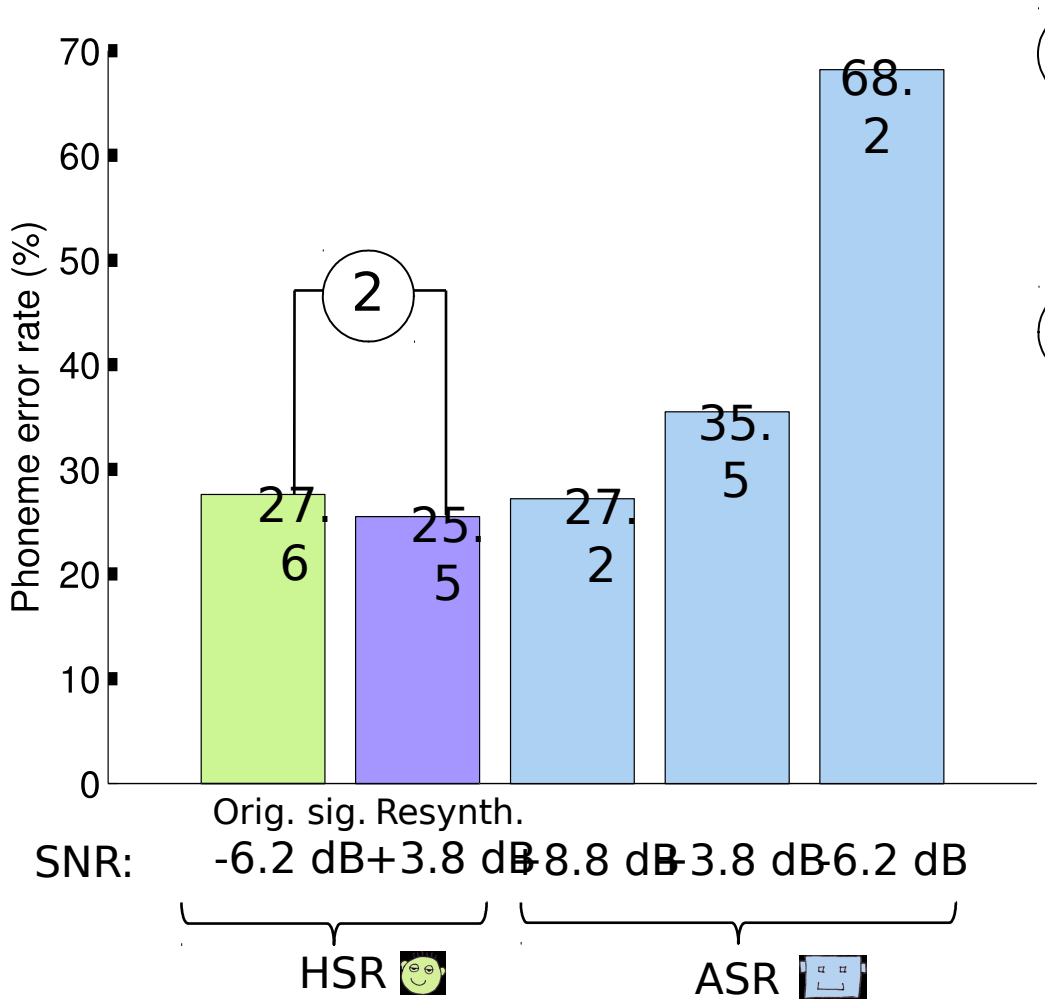
Comparison of error rates of man and machine



1 Human-machine gap: ASR reaches human performance level when SNR is increased by 15dB

Meyer, Brand, Kollmeier (2011), "Effect of speech-intrinsic variations on human and automatic recognition of spoken phonemes", JASA 129.

Comparison of error rates of man and machine

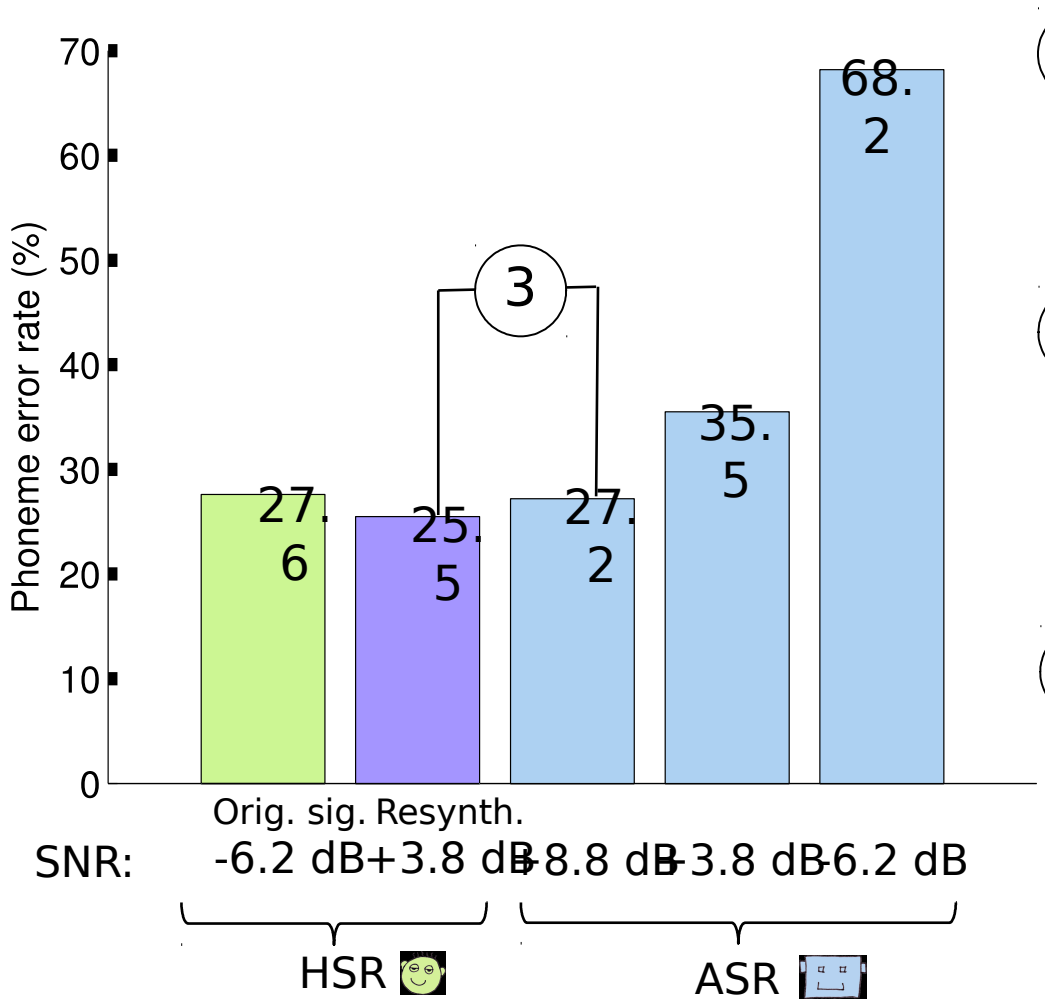


1 Human-machine gap: ASR reaches human performance level when SNR is increased by 15dB

2 Information loss due to feature extraction amounts to 10 dB MFCCs to not contain all information relevant for SR

Meyer, Brand, Kollmeier (2011), "Effect of speech-intrinsic variations on human and automatic recognition of spoken phonemes", JASA 129.

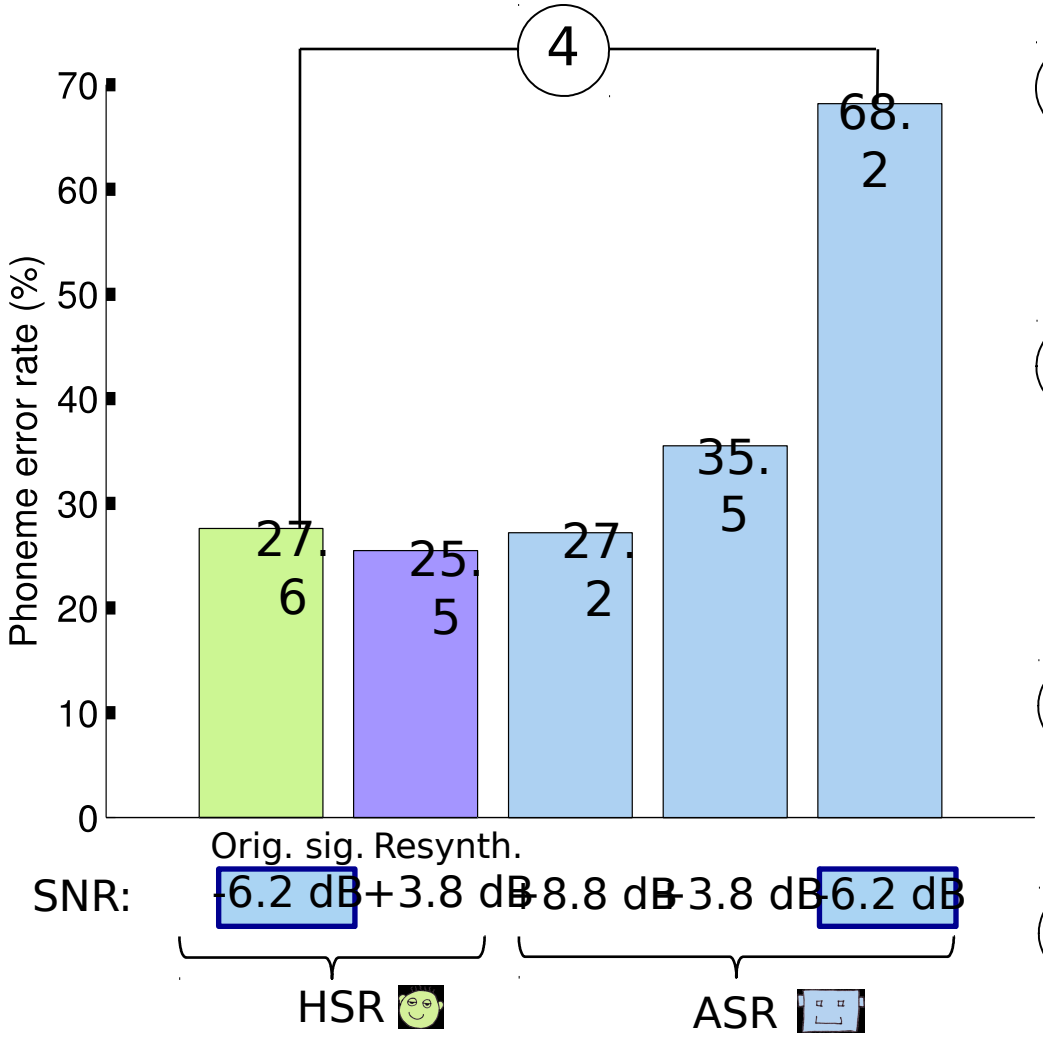
Comparison of error rates of man and machine



1. Human-machine gap: ASR reaches human performance level when SNR is increased by 15dB
2. Information loss due to feature extraction amounts to 10 dB MFCCs to not contain all information relevant for SR
3. Using the same information for HSR and HSR: Gap of 5 dB (can be attributed to HMM)

Meyer, Brand, Kollmeier (2011), "Effect of speech-intrinsic variations on human and automatic recognition of spoken phonemes", JASA 129.

Comparison of error rates of man and machine



1. Human-machine gap: ASR reaches human performance level when SNR is increased by 15dB

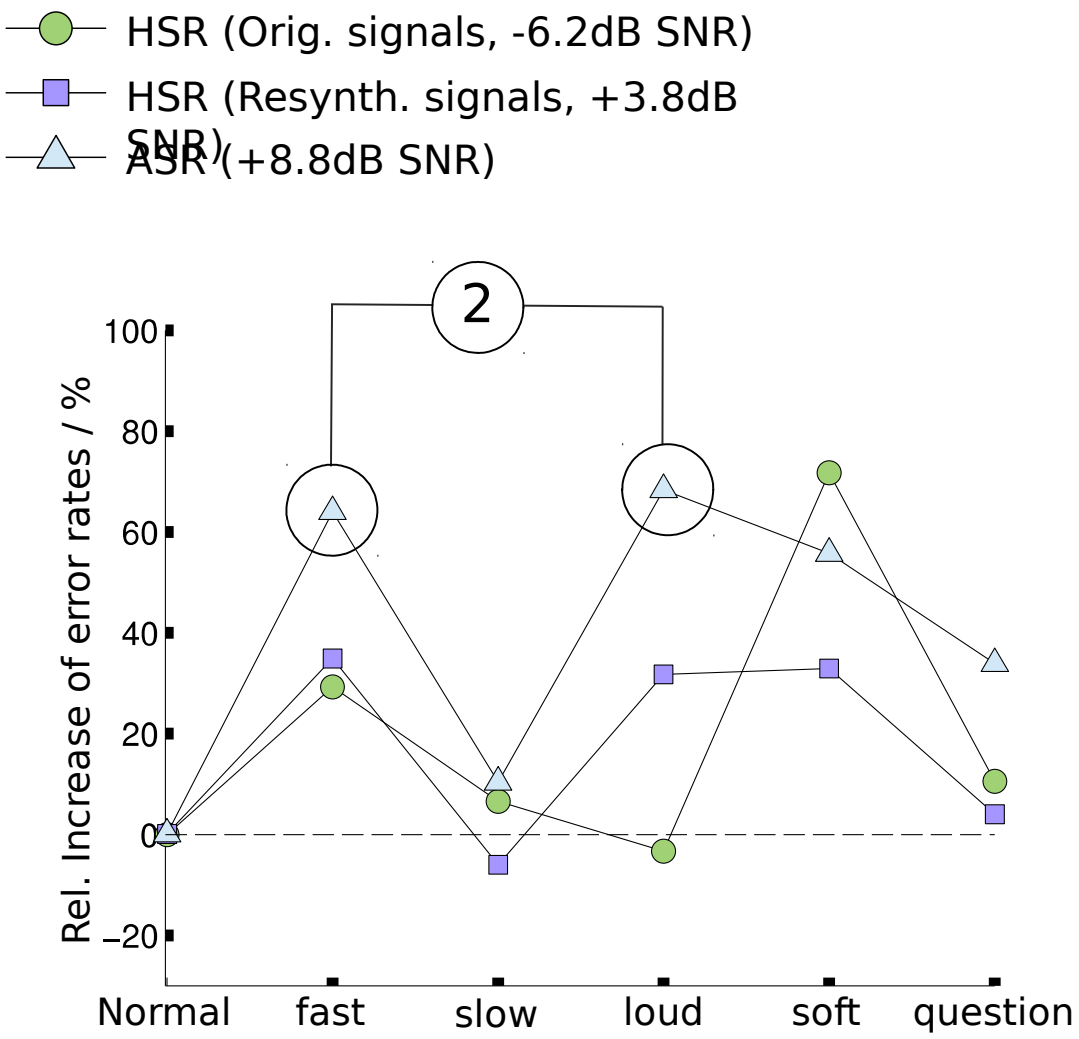
2. Information loss due to feature extraction amounts to 10 dB MFCCs to not contain all information relevant for SR

3. Using the same information for HSR and HSR: Gap of 5 dB (can be attributed to HMM)

4. Identical conditions for ASR and HSR: Error rates are more than doubled.

Meyer, Brand, Kollmeier (2011), "Effect of speech-intrinsic variations on human and automatic recognition of spoken phonemes", JASA 129.

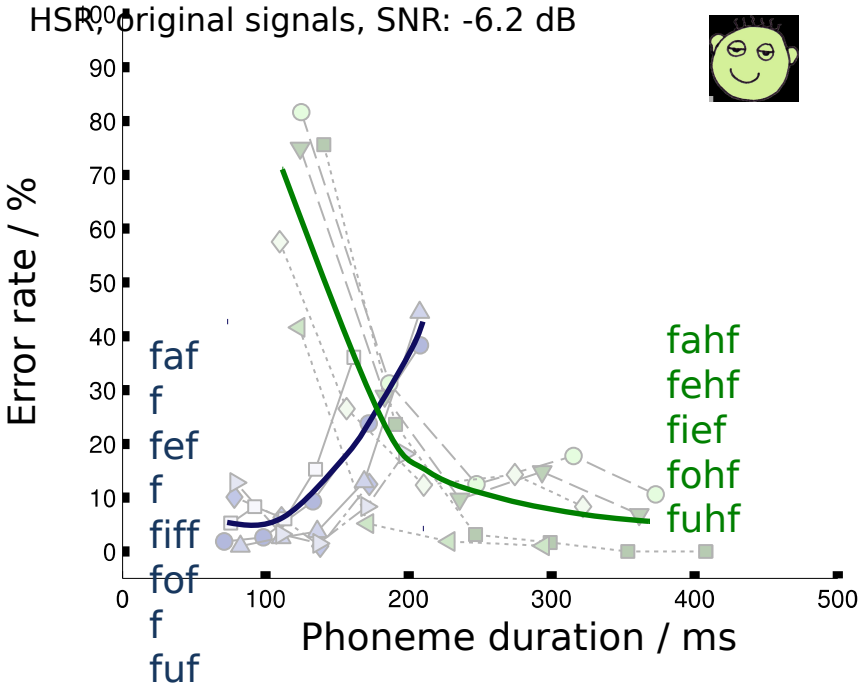
Influence of intrinsic variation

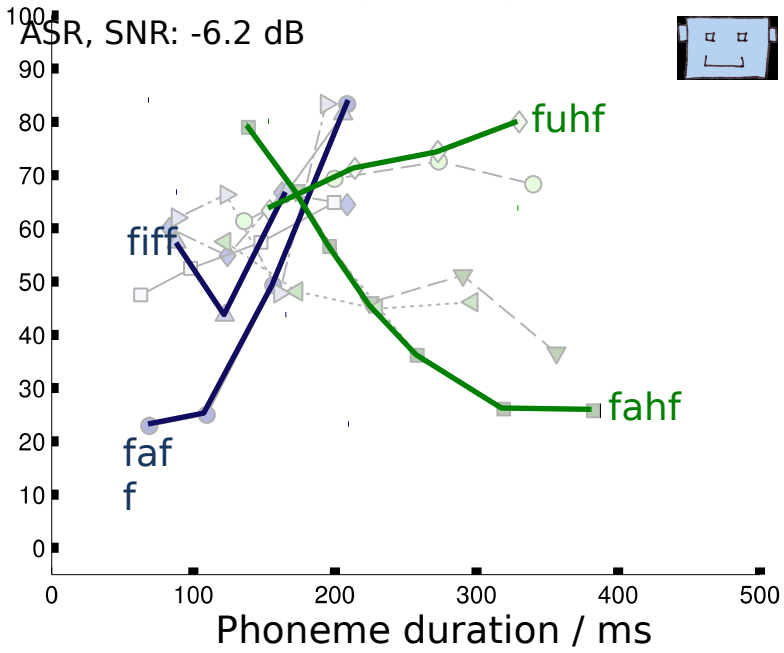


1 Variability increases error rates (HSR and ASR): Rel. increase of up to 70 %

2 ASR: Very high degradation for fast speaking rate and high speaking effort analysis of effect of speaking rate

Influence of phoneme duration on recognition rates

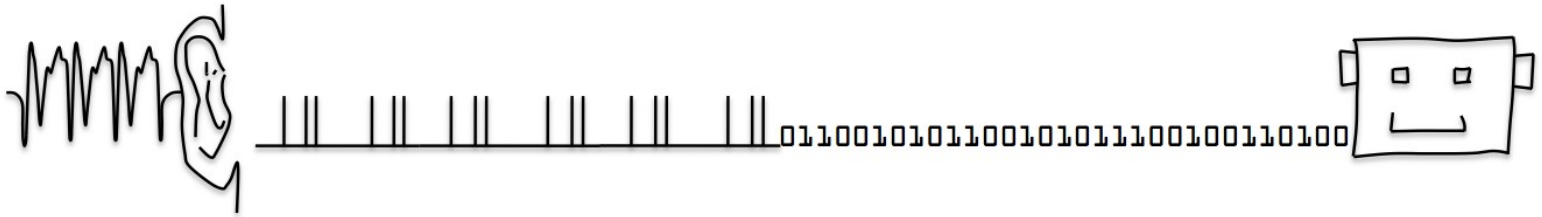




- HSR: Two groups for vowels are observed durational cues are important in HSR (in accordance with Hillenbrand, 1995)
- ASR: Relation between phonene duration and phoneme duration is less pronounced
  Temporal cues are not optimally exploited in ASR

Possible solution:Integration of those cues on feature level

Motivation: The auditory approach



Human speech recognition (HSR)

Automatic speech recognition (ASR)

Man-machine comparison

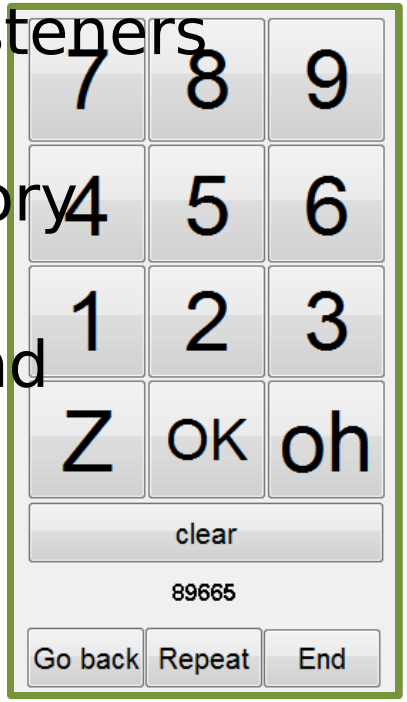
...on sublexical level with focus on speech-intrinsic variability

...on word level (noisy digit strings)

Aurora 2: Baseline vs auditory features vs HSR

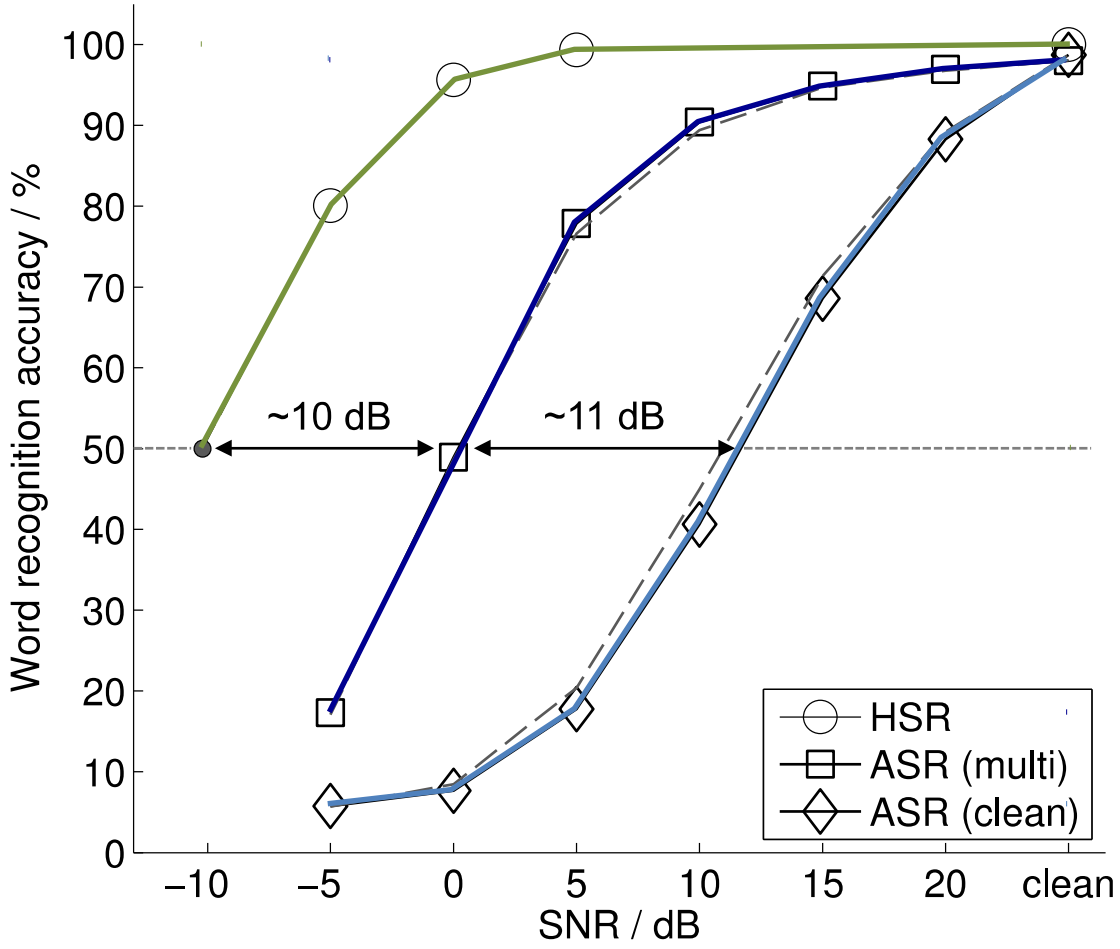
- Aurora 2: Framework for ASR experiments, noisy connected digits, two training conditions (clean and multi-condition)
- is (still) among the most often used corpora for ASR
- Compare HSR and ASR to find out: How far have we come on closing the gap between man and machine?

- Experiments with 10 normal-hearing listeners (L1 and L2) on a subset of Aurora 2
- ASR experiments with MFCC and auditory Gabor features
- ...combined with standard HMM backend



Aurora 2: Baseline vs auditory features vs HSR

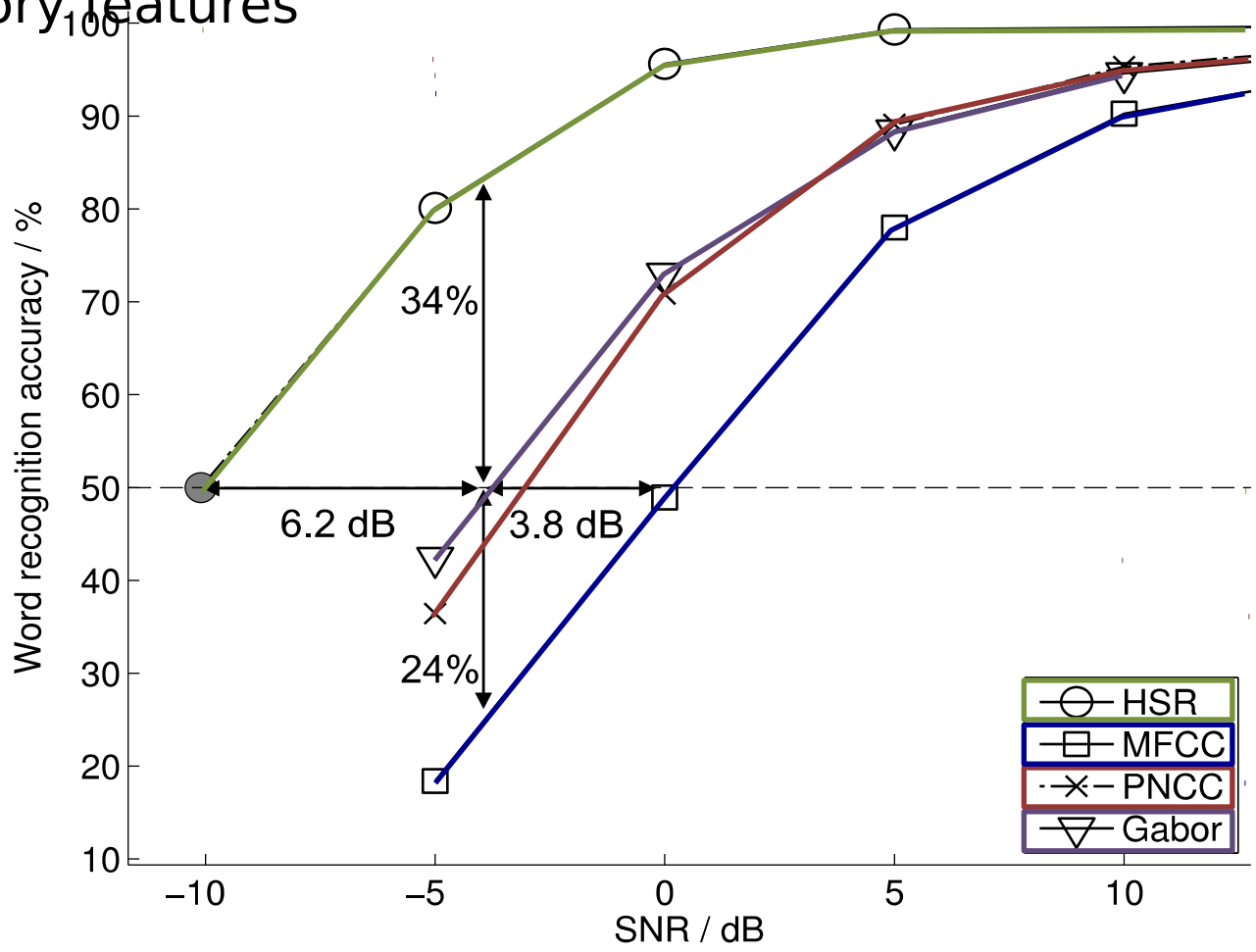
Results for humans and the baseline ASR system



Meyer B. INTERSPEECH 2013

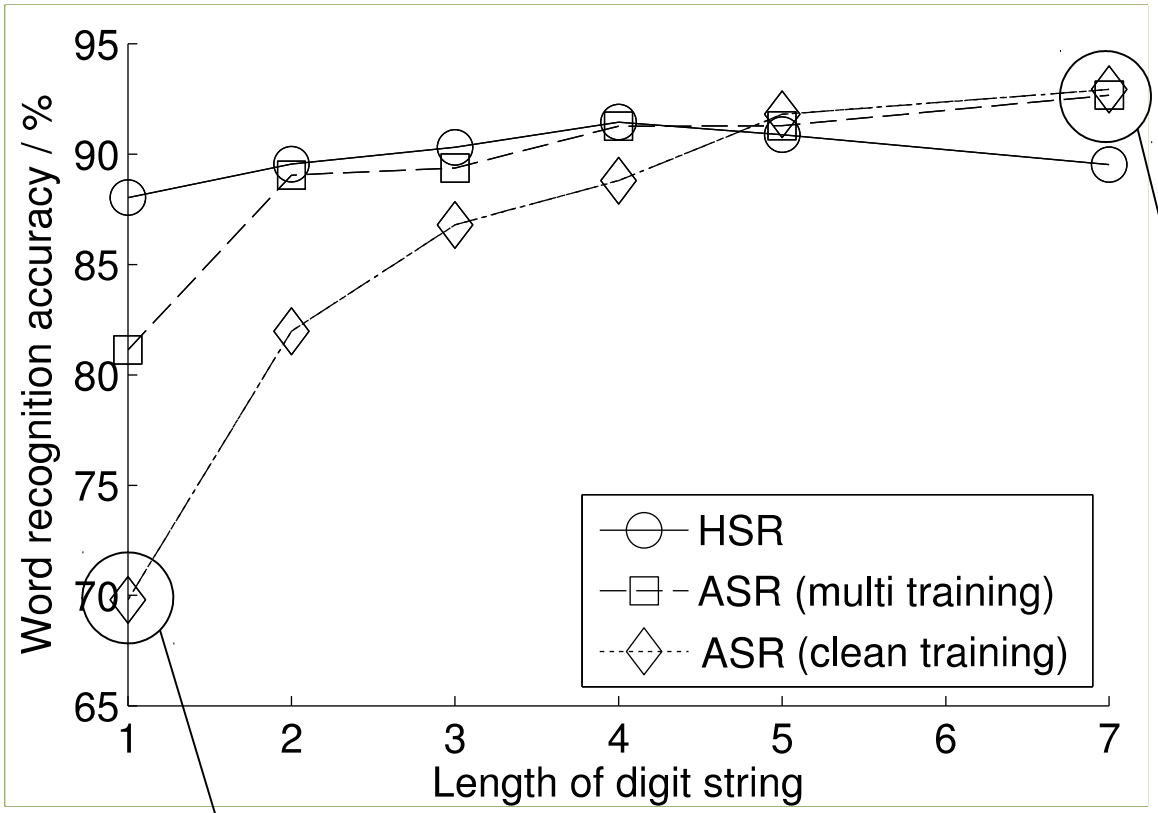
Aurora 2: Baseline vs auditory features vs HSR

Results for humans, baseline (multicondition training) and auditory features

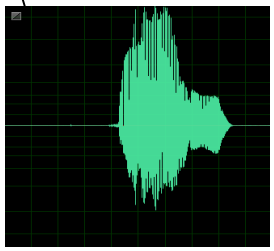


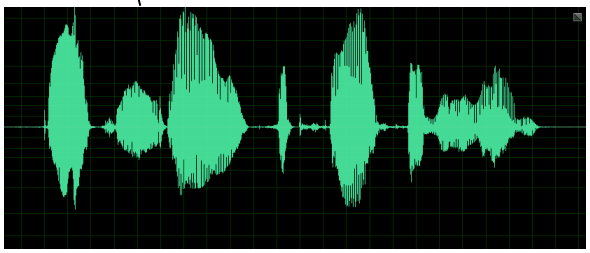
Meyer B. INTERSPEECH 2013

Aurora 2: Baseline vs HSR



- Higher pause-to-signal ratio increases more insertion errors
- Why aren't human listeners affected?
- Possibly because of better onset detectors in auditory system





Meyer B. INTERSPEECH 2013

Summary: Man-machine comparison

- Resynthesized vs. original signals: standard features to not contain all information relevant for speech recognition
- Quantification of the (task-dependent) human-machine gap
  - Phoneme level: In terms of SNR: 15 dB
    - 10 dB caused by feature extraction
    - 5 dB caused by imperfect back end
  - Word level (small vocabulary): In terms of SNR: 10 dB
    - Using auditory Gabor features reduces the gap from 10 dB to 6 dB (without optimizing the backend)
- Intrinsic variability:
  - ASR especially sensitive against high speaking effort and high speaking rate suggests to consider temporal cues in ASR feature extraction

Thanks to...

Birger Kollmeier
Thomas Brand
Tim Jürgens
Thorsten Wesker

...and thank you
for your attention!