Overview of the 2nd 'CHiME' Speech Separation and Recognition Challenge

Emmanuel Vincent[1], Jon Barker[2], Shinji Watanabe[3], Jonathan Le Roux[3], Francesco Nesta[4] and Marco Matassoni[5]

[1]Inria Nancy – Grand Est, France
[2]Department of Computer Science, University of Sheffield, UK
[3]Mitsubishi Electric Research Labs, Boston, MA, USA
[4]Conexant Systems, Newport Beach, CA, USA
[5]FBK-Irst, Trento, Italy

1. Motivation
2. Datasets and tasks
3. Baselines
4. Results

Earlier evaluations

Aurora noise-robust ASR benchmarks (2000–2005)

- Single-channel speech (TIDigits or WSJ) + noise
- Isolated noise sounds: too artificial?

PASCAL speech separation challenges (2006–2007)

- Single- or multichannel speech + speech (Grid or WSJ)
- Either 'superhuman' or poor results: still too artificial?

SiSEC source separation campaigns (2008–)

- 2- to 5-channel speech + speech and speech + noise
- Real-world noise scenes (unknown number of noise sources)
- Performance evaluated in terms of source separation metrics

The ‘CHiME’ Challenges

1st ‘CHiME’ Challenge (2011)

- Binaural data – link to hearing research and comparison with humans
- Grid speech corpus – small vocabulary and fixed grammar
- Real environment – Impulse responses and noises recorded in a domestic living room at a fixed position
- Performance evaluated in terms of ASR

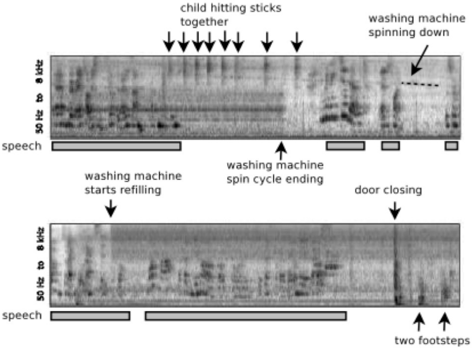
2nd ‘CHiME’ Challenge (2013)

Extends the difficulty along two dimensions:

- small speaker movements (Track 1)
- larger vocabulary (Track 2)

The 'CHiME' noise backgrounds

- Binaural noise backgrounds recorded in a family home (living room).
- Plenty of sources, well-defined application domain with a learnable noise 'vocabulary' and 'grammar'.



- Total of 14 h of audio in 0.5 to 1.5 h sessions over several weeks.

Data generation procedure

Simulate speakers at 2 m distance in front of the listener:

- record binaural room impulse responses (BRIRs) around that position,
- convolve clean speech utterances with the BRIRs and add to the noise backgrounds at specific times so as to match one of 6 possible SNRs: -6, -3, 0, 3, 6 or 9 dB (no rescaling).

Noise characteristics highly SNR dependent:

- 9 dB backgrounds fairly stationary ambient noise,
- -6 dB backgrounds highly non-stationary energetic events.

For each Track, generate:

- 3 training sets (clean, reverberated and noisy),
- 1 noisy development set (isolated or embedded utterances),
- 1 noisy test set (isolated or embedded utterances).

Track 1: small vocabulary, small speaker movements

- Target utterances from the Grid corpus.
- Speaker movements within each utterance on a straight left-right line for a distance of at most 5 cm at a speed of at most 15 cm/s.

VERB	COLOUR	PREP.	LETTER	DIGIT	ADVERB
bin lay place set	blue green red white	at by on with	a-z (no ‘w’)	1-9 and zero	again now please soon

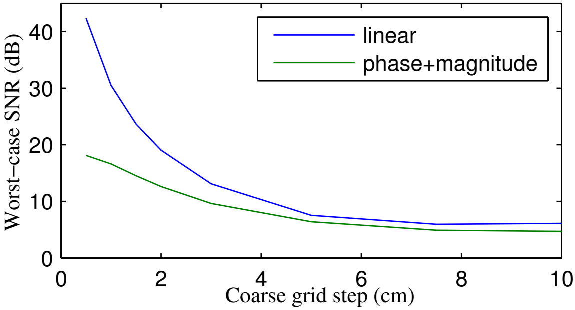
Clean | Reverberated | -6 dB | -3 dB | 0 dB | 3 dB | 6 dB | 9 dB

Only 34 speakers but 500 utterances each: makes it possible to learn speaker-dependent models and exemplar-based models.

No need to master ASR. . . but still has applications (house automation) and represents a significant challenge (letter set highly confusable).

BRIR interpolation details

- BRIRs recorded for 121 positions covering a horizontal square grid of 20 cm side with a grid step of 2 cm.
- Simulation of intermediate positions by linear interpolation.



Track 2: medium vocabulary, no speaker movements

- Target utterances from the Wall Street Journal (WSJ0) read speech corpus (5000-word vocabulary).
- No speaker movements (single BRIR).

Last month overall goods-producing employment fell 68,000 after a 32,000 job rise in February.

Clean Reverberated -6 dB -3 dB 0 dB 3 dB 6 dB 9 dB

More speakers but few sentences each: use speaker-independent models.

More challenging. . . but more difficult for non-experts in ASR.

Task and instructions

Task:

- Track 1: report the 'letter' and 'digit' tokens,
- Track 2: transcribe the whole utterance.

Allowed:

- exploit knowledge of the speaker identity and spatial location,
- exploit knowledge of the temporal location of each utterance,
- exploit the acoustic context of each utterance.

Forbidden:

- exploit knowledge of the SNR,
- tune algorithm parameters on the test set,
- exploit the fact that different datasets involve the same speech and/or noise signals (note that this forbids "stereo data" approaches).

1. Motivation
2. Datasets and tasks
3. Baselines
4. Results

Baseline ASR systems (HTK-based)

- Target signal enhancement: none
- Features: 12 MFCCs+log-energy+Δ+$\Delta\Delta$ with Cepstral Mean Subtraction (CMS)
- Track 1 decoder:
  - word-level HMMs - 2 states per phoneme,
  - states modelled with GMMs - 7 components with diagonal covariance,
  - flat start training on all data then on speaker-dependent data,
  - Viterbi decoding using Grid grammar, no pruning.
- Track 2 decoder:
  - triphone-level HMMs - 3 states, 1860 triphones,
  - states modelled with GMMs - 8 components with diagonal covariance,
  - reestimation of the HMM/GMM parameters from a pretrained speaker-independent clean speech model,
  - Viterbi decoding with pruning using the standard WSJ 5K non-verbalized closed bigram language model.

Difficulty of the task

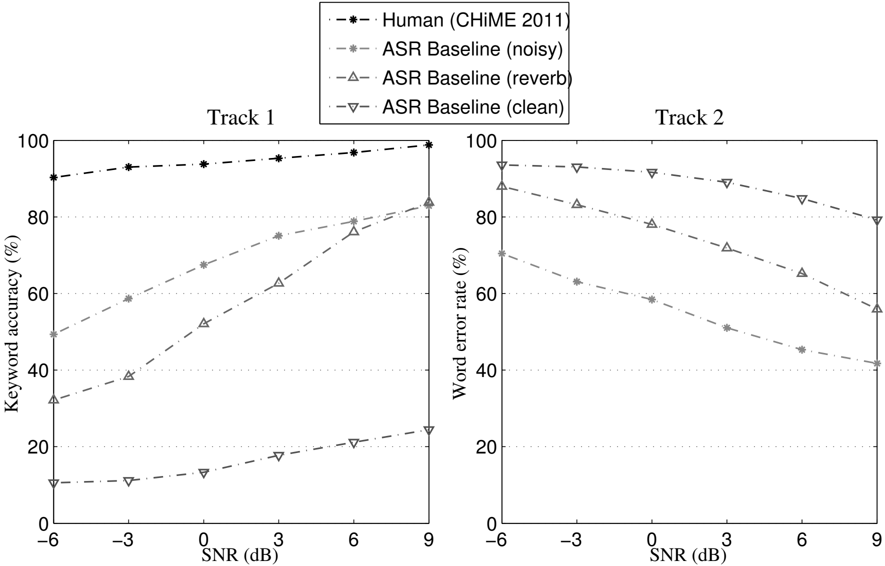
Training/test	Keyword accuracy (Track 1)	Word error rate (Track 2)
Clean/clean	97.25%	7.49%
Reverb/reverb	95.58%	18.40%
Noisy/noisy	68.72%	55.00%

Noise is the main difficulty:

- Reverberation increases the error rate by a factor of 1.6 to 2.5.
- Larger vocabulary size further increases it by a factor of 4.2.
- Noise further increases it by a factor of 2.3 to 11 depending on the SNR.

Small speaker movements have little effect on the baseline... but they may start to have one when attempting to enhance the target.

Baseline results



Multicondition training alone greatly improves performance.

Overview of the submitted systems

16 entries, among which 13 adhering to the instructions.

	Enhanced target	Modified features	Modified decoder
FBK-Irst & INESC-ID	X	X	X
Fraunhofer IAIS	X		X
Fraunhofer IDMT & U Oldenburg	X	X	X
Inria & Hörtech	X	X	X
KU Leuven		X	X
KU Leuven & TU Tampere	X		
Mitsubishi Electric	X	X	X
MRC IHR & U Sheffield	X		X
RU Bochum & GAMPT	X	X	X
TU Graz	X	X	
TUM, TUT, KUL & BMW	X	X	X
TU Tampere & KU Leuven	X		X
U Maryland & SRI		X	X

Target enhancement strategies

- Spectral enhancement based on pitch and/or timbre (5 entries)
  - multiple pitch tracking,
  - codebook-based separation,
  - exemplar-based Nonnegative Matrix Factorization (NMF).
- Spatial enhancement based on spatial location (4 entries)
  - fixed or adaptive beamforming,
  - Wiener filtering,
  - clustering of Interaural Time/Level Differences (ITD/ILD).
- Combined spatial and spectral enhancement (2 entries)
  - multichannel NMF.

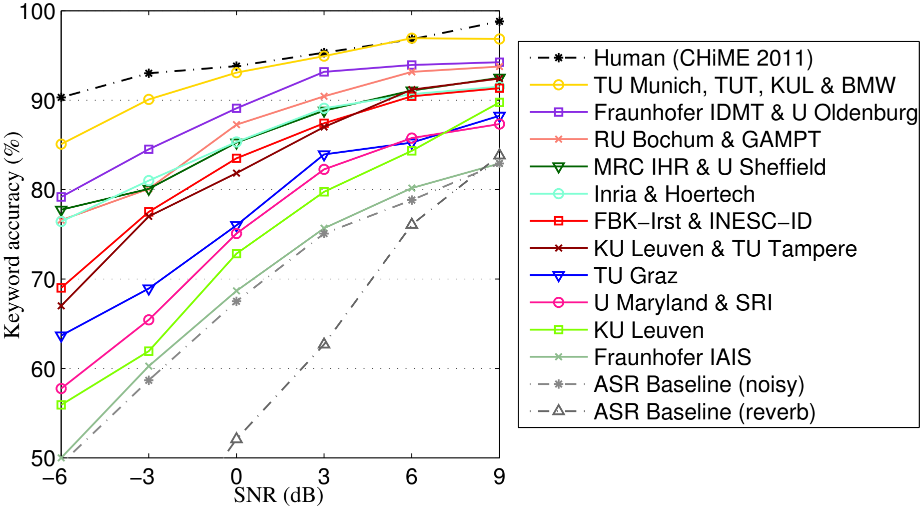
Feature extraction strategies

- Robust features (9 entries)
  - MFCC with spectral floor,
  - Gammatone Frequency Cepstral Coefficients (GFCC),
  - Normalized Modulation Cepstral Coefficient (NMCC),
  - Mel spectra,
  - Gabor Filterbank features (GBFB),
  - GMM posterior features,
  - recurrent neural network (BLSTM) features,
  - nonnegative sparse coding (NSC) features,
  - vocal Tract Variable (TV) trajectories.
- Feature transforms (2 entries)
  - Principal Component Analysis (PCA),
  - Maximum Likelihood Linear Transformation (MLLT),
  - Speaker Adaptive Training (SAT),
  - Linear Discriminant Analysis (LDA),
  - feature-space Maximum Mutual Information (f-MMI),
  - variance normalization.

Decoding strategies

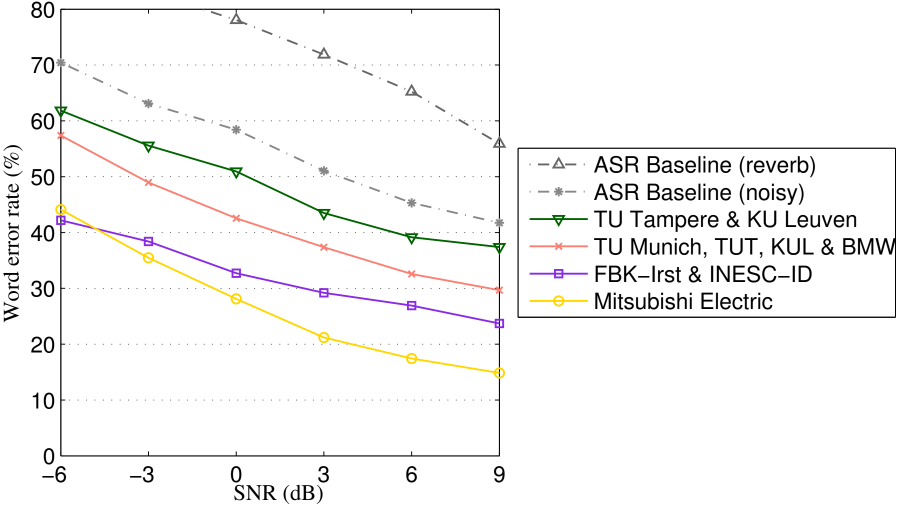
- Noise adaptive training (6 entries)
- Modified training/decoding objectives (3 entries)
  - MLLR/MAP speaker adaptation,
  - MMI discriminative training
  - Discriminative Language Modeling (DLM),
  - Minimum Bayes Risk (MBR) decoding.
- Noise-aware decoding (3 entries)
  - missing-data fragment decoding,
  - uncertainty decoding.
- Optimized HMM level/topology/size (2 entries)
- System combination (2 entries)
  - multistream decoding,
  - Recogniser Output Voting Error Reduction (ROVER).
- Exemplar-based decoding (1 entry)

Track 1 results



Best system: exemplar-based enhancement, MFCC, BLSTM and NSC features, multi-stream decoder with MAP speaker adaptation

Track 2 results



Best system: spatial enhancement, MLLT, SAT, LDA, f-bMMI, feature augmentation, bMMI noise-adaptive training, DLM and MBR decoding

Some outcomes

- For small vocabulary, best entry 30% worse than a trained human.
- Multicondition training and spatial enhancement are the most effective single strategies...
- ...but (even for these small and medium vocabulary tasks) the best systems are highly complicated and tuned setups resulting from collaborative efforts.
- Small source movements do not increase difficulty...
- ...but vocabulary size does: for medium vocabulary, careful design of the ASR back-end plays a major role in performance.
- Further outcomes to be obtained from the analysis of the transcripts and from the refinement of the instructions in future challenges.

Best Task 2 entry now available as a Kaldi baseline.

Please try it!