

# Overview of the PASCAL CHiME Speech Separation and Recognition Challenge

Jon Barker<sup>1</sup>, Emmanuel Vincent<sup>2</sup>, Ning Ma<sup>1</sup>, Heidi Christensen<sup>1</sup>,  
and Phil Green<sup>1</sup>

<sup>1</sup>Department of Computer Science, University of Sheffield, UK

<sup>2</sup>INRIA Rennes - Bretagne Atlantique, France

1st September, 2011



The  
University  
Of  
Sheffield.



**PASCAL**

Pattern Analysis Statistical Modelling and  
Computational Learning



# Outline

- 1 CHiME Challenge motivation and design
- 2 Human listening test results
- 3 Overview of CHiME Challenge entrants

# Outline

- 1 CHiME Challenge motivation and design
- 2 Human listening test results
- 3 Overview of CHiME Challenge entrants

# Outline

- 1 CHiME Challenge motivation and design
- 2 Human listening test results
- 3 Overview of CHiME Challenge entrants

# Previous speech separation challenges

## PASCAL single-channel separation challenge, Interspeech 2006

- Instantaneous speech + speech mixtures from the Grid corpus.
- Not multisource in the sense that the number of sources is known a priori
- Best solutions built models of each speaker and combined the models to explicitly model the mixture
- 'super human' results. Too artificial?

## PASCAL microphone array separation challenge, MLMI 2007

- Simultaneous live readings of WSJ recorded by microphone array.
- Small number of competitors.
- Very poor results. Too challenging?

# Previous speech separation challenges

## SiSEC evaluation campaign, ICA 2009 and LVA/ICA 2010

- 2- to 5-channel datasets, where the number of sources is generally known a priori.
- One exception: denoising dataset including real multisource outdoor noise (subway, cafeteria, town square).
- Performance evaluated in terms of source separation quality only.

# The PASCAL CHiME challenge

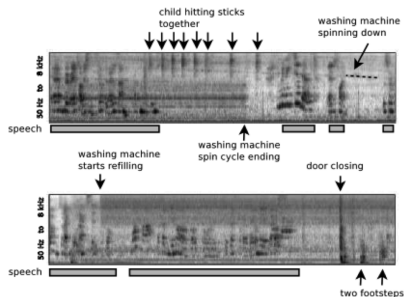
## PASCAL CHiME challenge, 2011

- Using Grid corpus - small vocabulary and fixed grammar; continuity with 1st PASCAL challenge
- Real multisource environment – a domestic living room.
- Convolutional mixtures using impulse responses recorded in the room.
- Binaural recording – to provide link to hearing research and comparisons with human performance

# The CHiME noise background

Noise backgrounds collected from a family home,

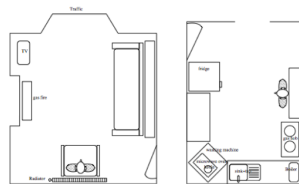
- it's noisy ... plenty of sources and potential for low SNRs
- it's easy to collect,
- potential application interest,
- well defined 'domain' with a learnable noise 'vocabulary' and 'grammar'.





# Recording Details

- Recordings made in the main living room.
- Recorded using a B&K 'head and torso' simulator.
- Total of 50 hours of stereo audio at 96 kHz, 24bit.
- Morning and evening sessions over course of several weeks.
- Set of binaural room impulse responses recorded.



# The target speech data

Target utterances come from the Grid corpus.

VERB	COLOUR	PREP.	LETTER	DIGIT	ADV.
bin	blue	at	a-z	1-9	again
lay	green	by	(no 'w')	+ zero	now
place	red	in			please
set	white	with			soon

- Small vocabulary so easy to build recognisers and computationally cheap.
- Still represents significant challenge for its size – letter set highly confusable.
- Small number of speakers (34) but a lot of data from each (1000 utterances). So can focus on speaker dependent models.
- Provides continuity with 1st PASCAL separation challenge.

# Preparing the mixed data

The aim was to simulate the effect of Grid utterances being spoken from a fixed position within the room.

- A single room location was chosen: 2 metres in front of the binaural manikin.
- Some Grid utterances were recorded from this position to establish a reference speaking level.
- Grid corpus utterances convolved with room impulse responses, inverse filter applied to remove recording coloration, and a testset-wide gain set to match reference level.
- Utterances added to CHiME background recordings at positions chosen so as to match a set of target SNRs.
- Possible to generate SNRs down to -6 dB.

[Original](#)[Convolved](#)[Mixed](#)[Comparison](#)

# Preparing the mixed data

Some points worth noting,

- SNR calculation a little unconventional
  - Two channels, so channels were averaged before SNR computation.
  - Rumble in some CHiME recordings was leading to very low SNRs for perceptually low-noise mixtures...
  - ... so SNR calculation performed after applying a high pass filter with a 80 Hz cut off.
  - SNR was measured over the duration of the entire Grid utterance.
- After mixing the Grid utterances are not evenly spread through the CHiME data
  - The average interval between utterances is about 10 seconds,
  - but asymmetric distribution: 23% < 1 second, 50% < 5 seconds and 70% < 10 seconds.
- Characteristic of noise background highly SNR dependent,
  - 9 dB backgrounds tend to be fairly stationary ambient noise,
  - -6 dB backgrounds highly non-stationary energetic events.

# The recognition task

## Test data

- 600 test utterances at each of 6 SNRs: -6, -3, 0, 3, 6, 9 dB
- All utterances embedded in 20 hours of CHiME audio.

## Task

- Task is to report the 'letter' and the 'digit' spoken by the Grid talker.
- Competition assumes the speaker identity and the temporal location of each utterance are known, but not the SNR.

# Human listening tests

- Listening tests have been performed to allow human machine comparison.
- The 1st PASCAL challenge saw 'super human' performance ...
  - ... but the comparison was arguably unfair in favour of the machines.

## Unfairness in previous comparison

- Task: recognising two simultaneous speakers over a single channel is not a natural task.
- Training: the machines had been trained on Grid corpus, humans were given no specific training.

# Human listening tests

This time around we hope that the comparison is a little fairer...

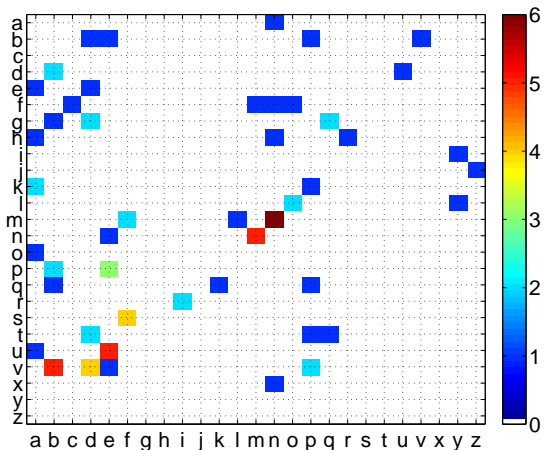
## Reasons that the current comparison is fairer

- The task is more natural - binaural listening in an everyday environment.
- Tests have used one highly motivated listener who is very familiar with the specific CHiME domestic audio environment
- Grid talkers were played in order (i.e. not randomised)
- Reverberant noise free training examples played prior to the test
- Two second of audio context played leading in to each utterance.

Example 6 dB

Example -3 dB

# Listening test confusions: Letters



## Confusions...

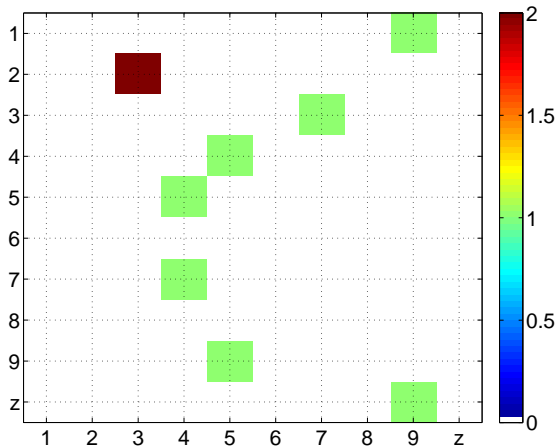
- $m \rightarrow n, n \rightarrow m$
- $v \rightarrow b, v \rightarrow d, p \rightarrow e$
- $s \rightarrow f$
- $u \rightarrow e$

also,

- $d \rightarrow b, g \rightarrow d, v \rightarrow p,$   
 $p \rightarrow b, t \rightarrow d$
- $k \rightarrow a$
- $m \rightarrow f$
- $r \rightarrow i$
- $l \rightarrow o, g \rightarrow q ??$



# Listening test confusions: Digits



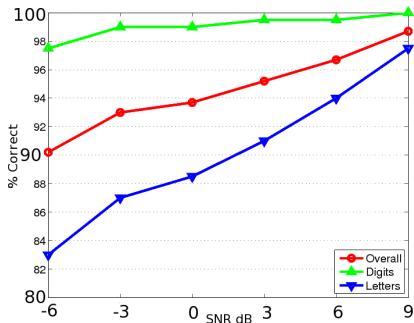
## Confusions ...

Very few.

- one → nine
- four → five, five → four
- nine → five
- zero → nine ?
- seven → four ?
- three → seven ?
- two → three ?

# Listening test results

Percentage digits and letters recognised correctly versus SNR.



- Digit recognition highly reliable: 99% correct down to -3 dB.
- Letter recognition falls steadily with increasing noise level at about 1% per dB: 97% at 9 dB down to 83 % at -6 dB.

# CHiME Challenge Systems

## Training data

- Reverberated noise-free Grid utterances provided for training speaker-dependent speech models. 500 utterances per speaker.
- Access to 6 hours of speech-free background also provided for training noise models.

## Development data

- 600 Grid utterances @ 6 SNRs provided for adapting the speech models to noisy speech.

## Test data

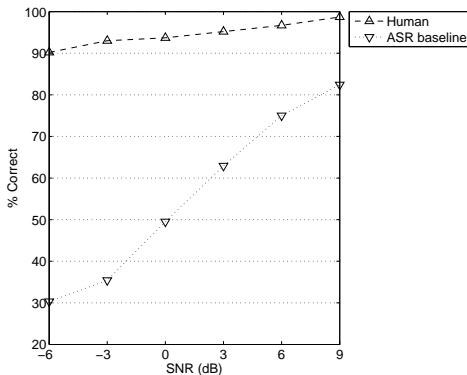
- 600 Grid utterances @ 6 SNRs released shortly before submission deadline.

# Baseline system

## Baseline system configuration

- **Target signal enhancement:** none
- **Features:** MFCC with deltas and delta-deltas computed from magnitude spectra with Cepstral Mean Subtraction (CMS)
- **Decoder:**
  - Word level HMMs - 2 states per phoneme
  - States modelled with GMMs, 7 components with diagonal covariance.
  - Viterbi decoding using Grid grammar, no pruning.
- **Training:**
  - Flat start training.
  - Initial models trained using 34x500 utterance training set.
  - 34 sets of Spkr. Dep. model reestimated using 500 utterances.

# Baseline system



As expected, non-robust baseline system performs fairly well on matched clean data (94%) but it is not robust to additive noise.

# Overview of the 13 accepted entries

	Enhanced target signal	Modified features	Modified decoder	Trained noise model
U. Aalto	X	X	X	
U. Bochum	X	X	X	
U. Erlangen	X		X	
ETRI	X	X	X	
EURECOM	X		X	X
FBK-IRST	X	X	X	
INRIA	X		X	X
K.U. Leuven	X		X	X
T.U. Liberec	X			
T.U. München	X	X	X	X
NTT	X		X	X
U. Sheffield	X	X	X	
T.U. Tampere		X	X	X

# Target signal enhancement strategies

## A wide variety of filters. . .

- Different domains:
  - STFT
  - mel spectrum
  - gammatone spectrum
- Different families of filters:
  - highpass/lowpass
  - beamforming
  - single-/multichannel Wiener filtering
  - binary/soft TF masking
- Tuned implementations:
  - oversubtraction
  - spectral floor/offset
  - temporal smoothing
  - exponentiation
- More fundamental issue: which cues are exploited to discriminate the target speaker from the background?

# Target signal enhancement strategies

## ... but few discrimination cues

- **Spatial diversity** = spatial location (5 entries)
  - beamforming,
  - geometrically constrained Independent Component Analysis (ICA),
  - clustering of Interaural Time/Level Differences (ITD/ILD).
- **Spectral diversity** = pitch and/or timbre (4 entries)
  - multiple pitch tracking,
  - Gaussian Mixture Model (GMM),
  - Nonnegative Matrix Factorization (NMF),
  - exemplar-based enhancement.
- **Combined spatial and spectral diversity** (3 entries)
  - chained design, *e.g.* ITD clustering followed by exemplar-based enhancement,
  - joint design: joint probabilistic frameworks for ITD and GMM/NMF.



# Feature extraction strategies

## Robust features and robustifying transformations

- **Robust features** (5 entries)
  - Gammatone Frequency Cepstral Coefficients (GFCC): improve robustness to spectrum underestimation thanks to wider filters.
  - Mel spectra: concentrate noise in fewer coefficients.
  - Parallel stream of phoneme predictions generated by a recurrent neural net: model the long-range context.
- **Robustifying feature transformations** (2 entries)
  - Maximum Likelihood Linear Transformation (MLLT).
  - Linear Discriminant Analysis (LDA).

# Decoding strategies

## Four complementary decoding strategies. . .

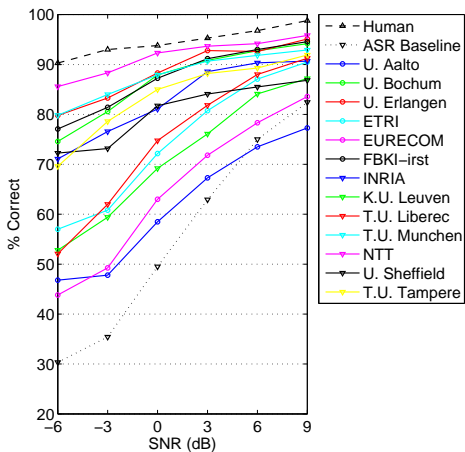
- **Multi-condition training/adaptation** (8 entries)
  - train/adapt the decoder over unprocessed noisy speech,
  - train/adapt the decoder over noisy speech processed by the target enhancement front-end.
- **Robust training** (6 entries)
  - manual setting of the number of Gaussians per mixture,
  - MLLR/MAP/mean-only speaker adaptation,
  - discriminative training.
- **Noise-aware decoding** (5 entries)
  - missing data: fragment decoding, channel-attentive decoding,
  - uncertain data: modified imputation, uncertainty decoding, Dynamic Variance Adaptation (DVA), location-informed decoding.
- **System combination** (4 entries)
  - Recogniser Output Voting Error Reduction (ROVER),
  - multistream decoding.

# Decoding strategies

## ... and one singular strategy

- **Model combination** (1 entry)
  - no target enhancement front-end,
  - jointly decode speech and noise via an exemplar-based model,
  - train the mapping between exemplar activations and likelihoods.

# Overview of ASR results



# Overview of ASR results

## What we can tell...

- Human performance is roughly twice that of the best entry.
- Strategies often present in the top-performing entries include:
  - multi-condition training,
  - robust training,
  - spatial diversity-based enhancement.
- More complex strategies (including trained noise models) seem to bring smaller additional improvement.

## ... and what we cannot tell

- The exact impact of each strategy is unknown, since they have not always been separately evaluated nor combined together.
- This impact may depend a lot on the data and the task.

# Editorial choice

- Five entries chosen for oral presentation at the workshop.
- Not necessarily highest performing: selection bias towards novelty.

# Main questions to think about

- Was this challenge sufficiently **realistic**? If not, in which direction should it evolve?
- How could the **scientific insight** gained from the challenge be increased?
- Is there a way to **facilitate combination** of the best strategies?
- What would be the best **business model** for a regular challenge?

These (and other) issues will be debated during the panel session. Please fill the questionnaire and return it to us before 4pm!