Maximum Likelihood Beamforming for Robust Automatic Speech Recognition

Barbara Rauch
barbara@lsv.uni-saarland.de
IGK Colloquium, Saarbrücken, 16 February 2006

Agenda

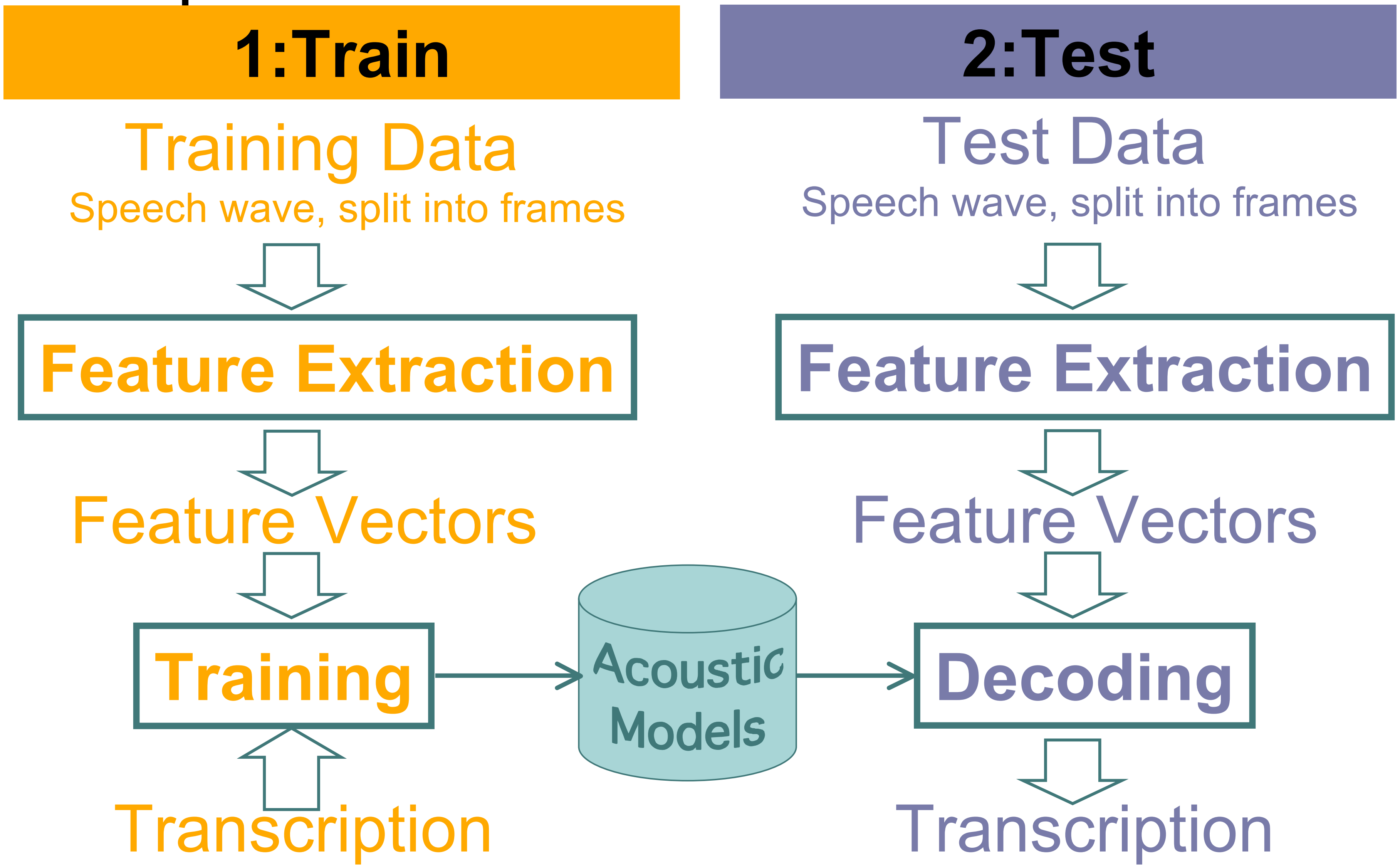
- Background: Standard ASR
- Robust ASR
- Background: Standard beamforming
- Maximum Likelihood Beamforming
  - Michael Seltzer’s Ph.D. work
  - Our version, open issues etc.

Introduction

- Reimplementation project at LSV
- Current participants: Andreas Beschorner, Marcela Charfuelan & Barbara Rauch
- Algorithm developed by Michael Seltzer at CMU, Ph.D. thesis in 2003
- Significant reduction in WER for recognition of noisy/reverberant speech

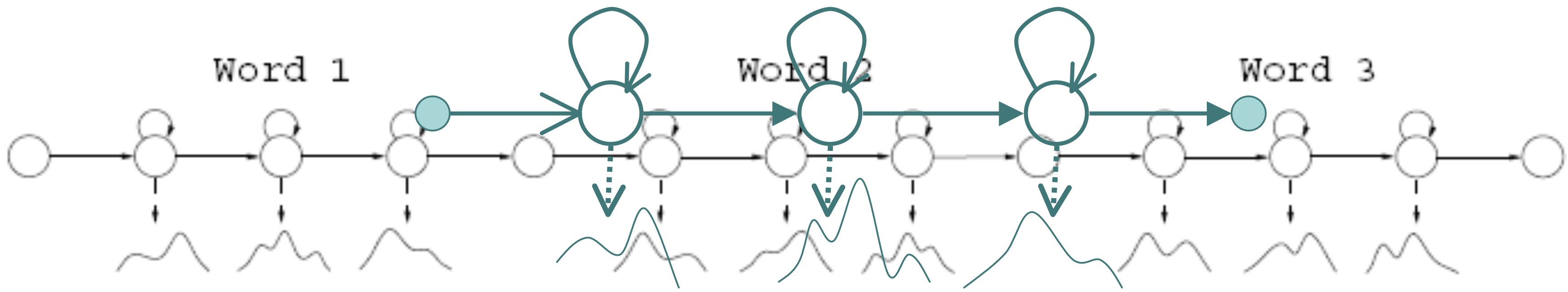
Background: Standard ASR

Standard ASR Process



HMM Acoustic Models

- Standard Acoustic Models: HMMs



- Decoding means search:
  - Alignment of frames with states = path through network of HMM states
  - Find most likely alignment / path
  - HMM parameters tell us likelihood of observation for a particular state sequence
- Transcription can be deduced from alignment

2nd figure taken from [1]

Robust ASR

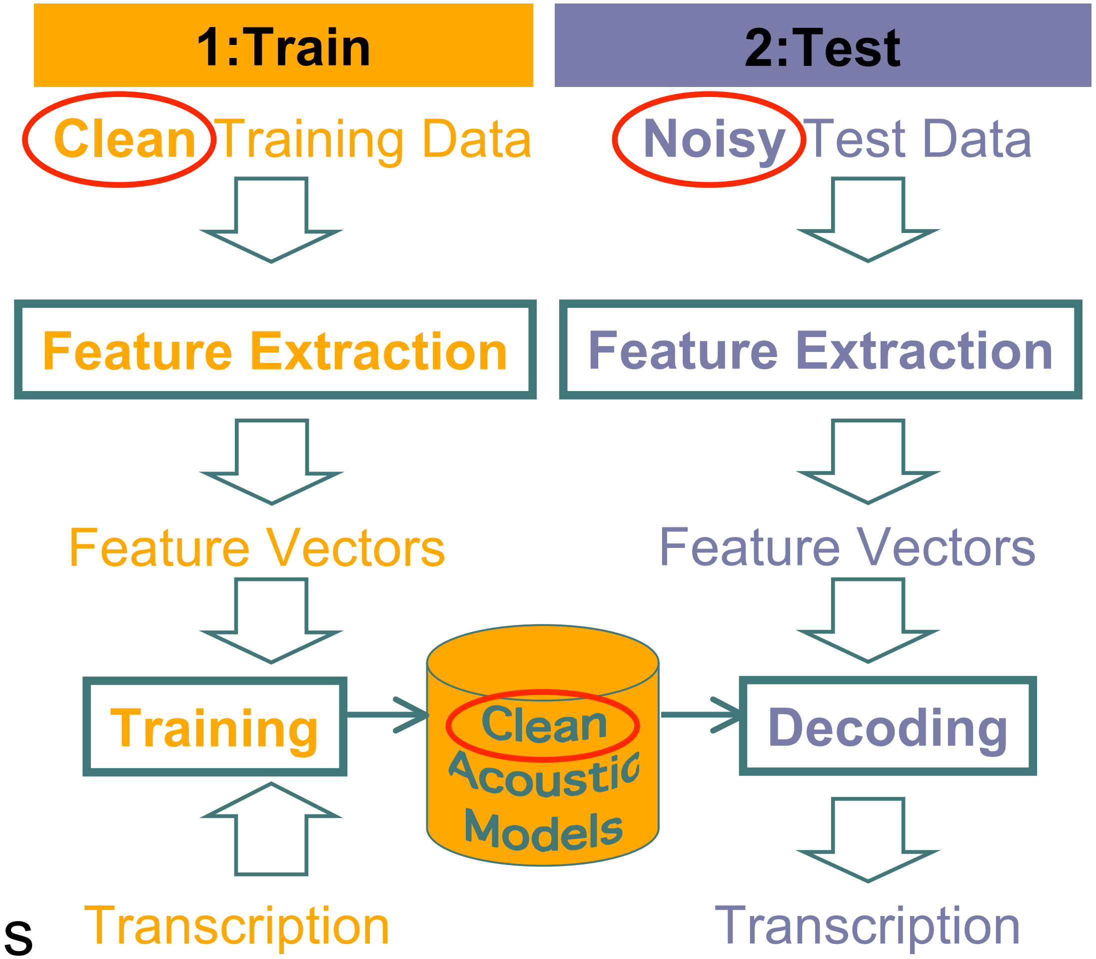
What is Robustness?

- “A smooth degradation in the performance of a system when faced with unexpected input.”
  (ROMAND workshop)

- “The degree to which a system or component can function correctly in the presence of invalid inputs or stressful environmental conditions.”
  (IEEE Standard Glossary of Software Engineering Terminology)

Robustness in ASR

- Invalid/unexpected input: things we didn't train on
- Focus here: noise and reverberation
- Problem: mismatch between test and training conditions
- Assumption: can't anticipate all the different noise conditions



Typical Degradation: Additive Noise

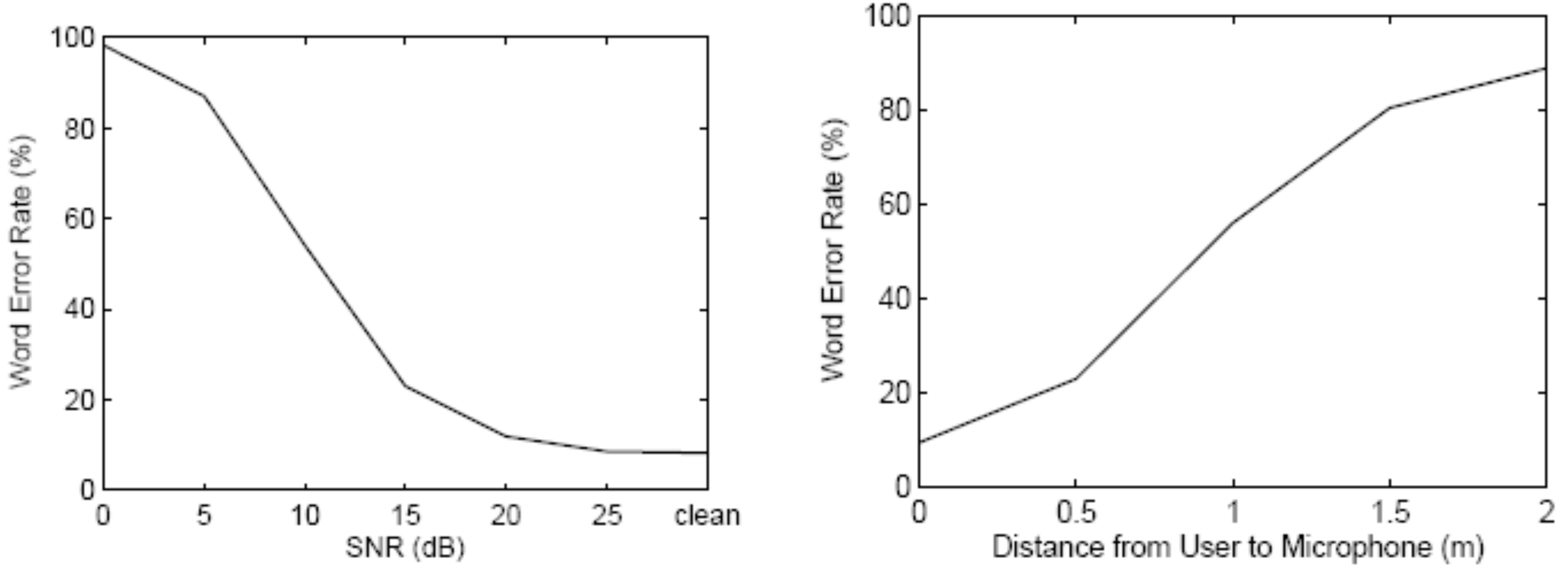
- Aurora (2000): connected digits with additive noise. Baseline WERs:

SNR	Clean	20dB	15dB	10dB	5dB	0dB
Ratio S:N	n/a	10 : 1	5.6 : 1	3.2 : 1	1.8 : 1	1 : 1
WER	1.5%	2.7%	3.8%	7.3%	16.8%	41.6%

- Aurora-4 (2002) large vocabulary task (ALV): baseline overall 50.3%

Typical Degradation: Additive Noise (2)

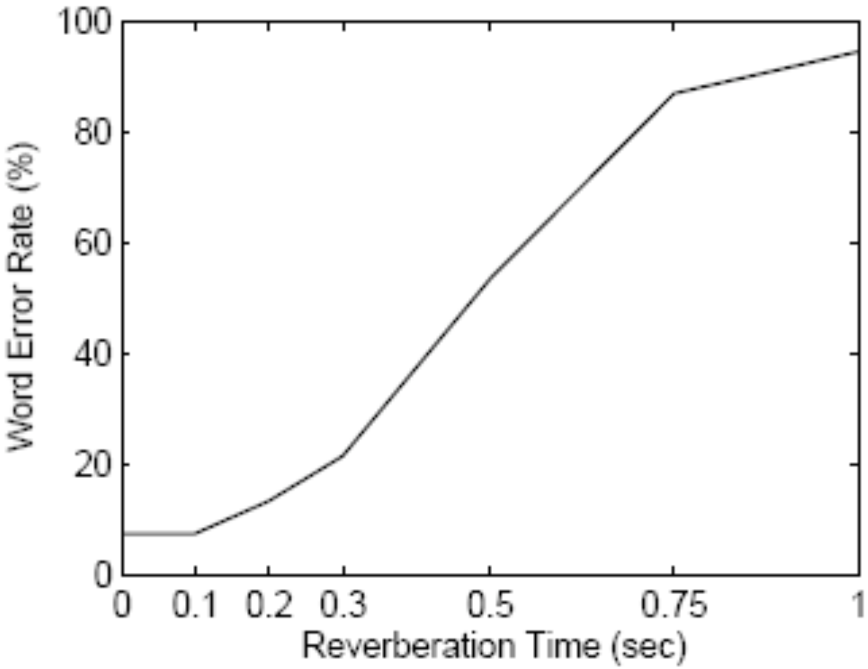
- System trained on clean speech recorded from close-talking microphone. Data: CMU microphone array database, described later

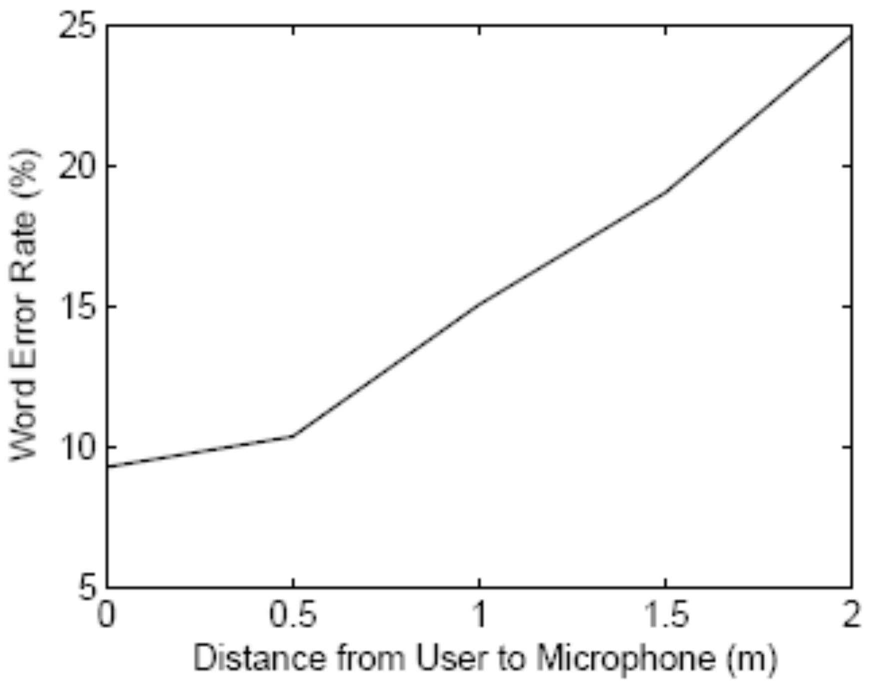


Figures taken from [1]

Typical Degradation: Reverberation

- System trained on clean speech recorded from close-talking microphone. Data: CMU microphone array database, described later

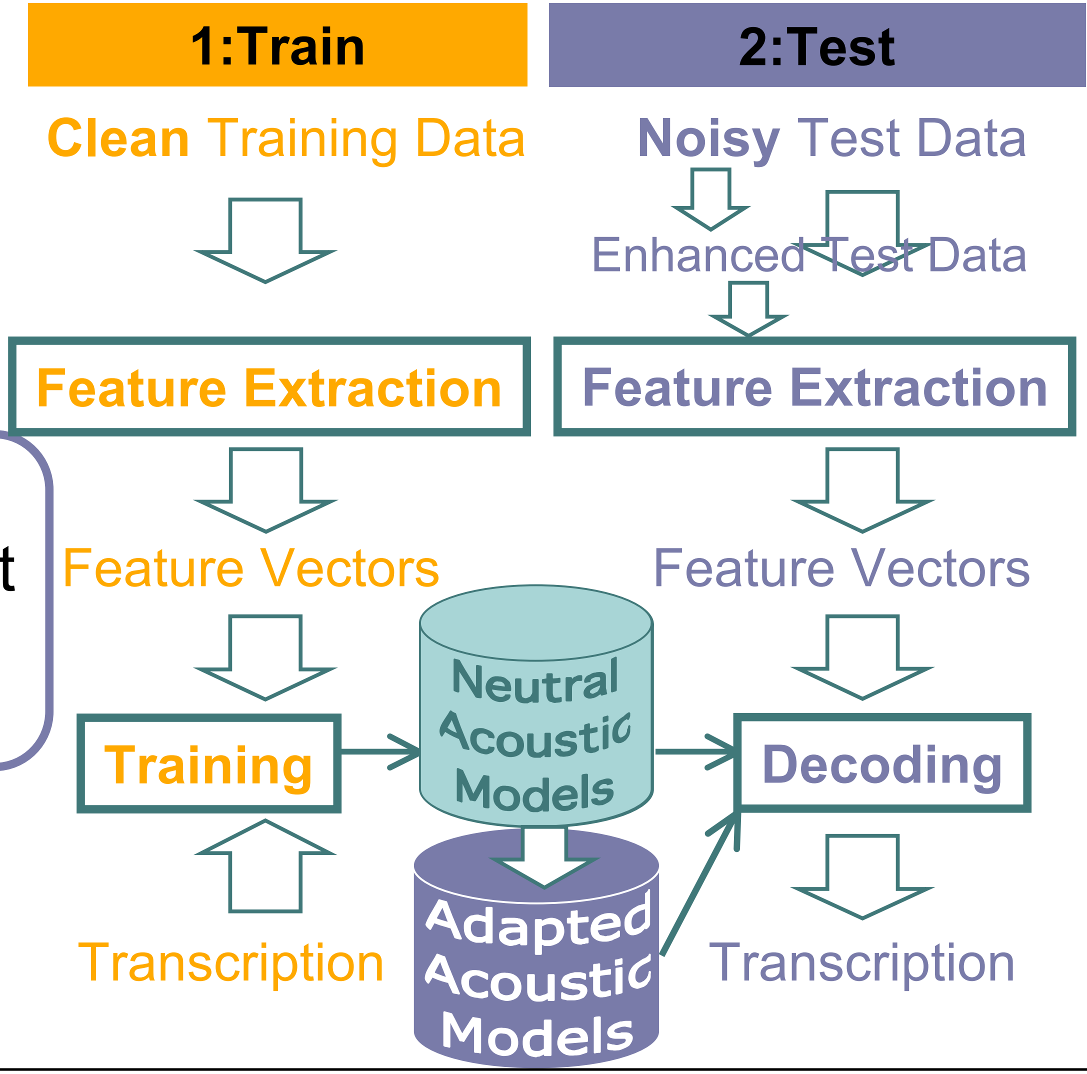




Figures taken from [1]

Possible Approaches

1. Noise resistance
   - Use noise-robust features
   - Cf. last talk, use visual information
2. Feature compensation / Speech enhancement
   - Change test data (signal/parameters)
3. Model compensation
   - Change test models



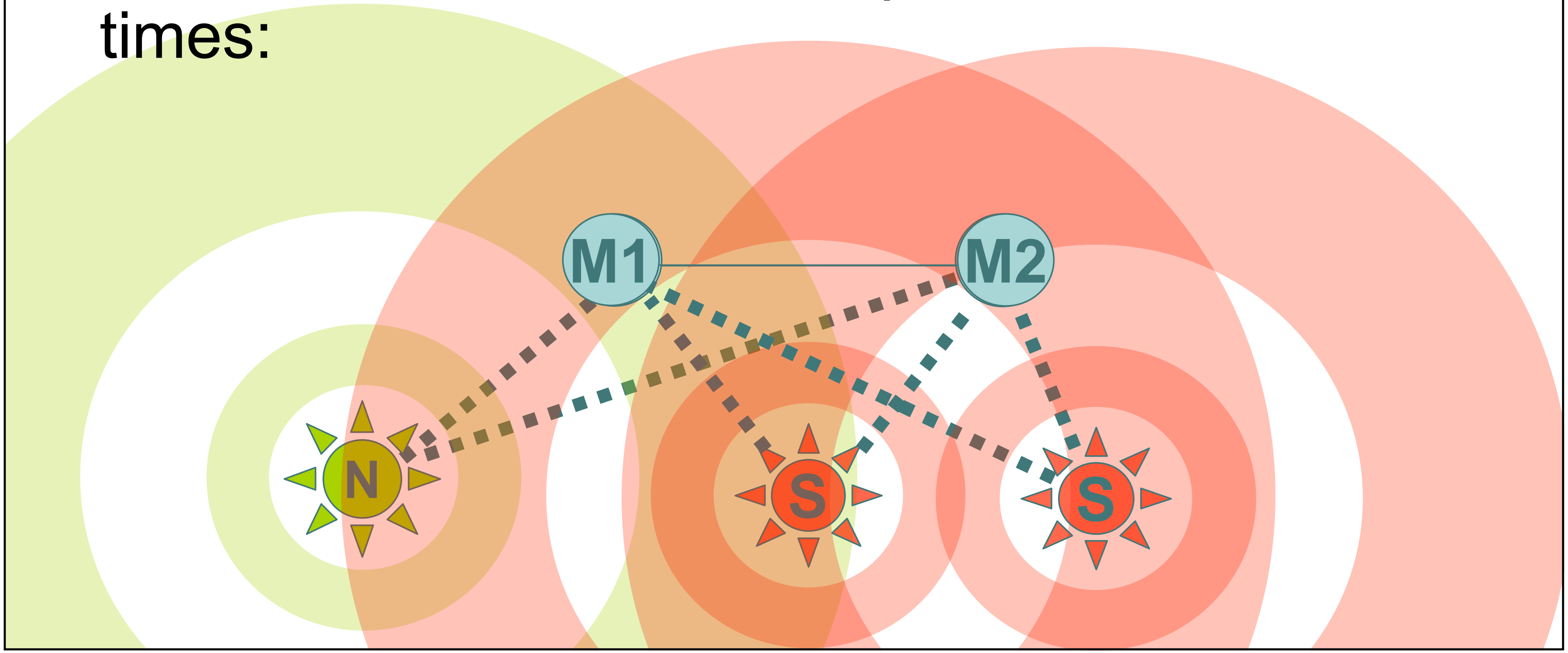
Recent Research Projects & Some Results

- Noise: e.g.
  - AURORA: various types of additive noise. Results: e.g. Large Vocab Task, WER 50%→30-35% (2000)
  - SPINE (‘01/02), ROAR: military noise. Results: e.g. on SPINE-2 data 42%→32% [4]
- ‘Hot’ application: meeting transcription.
  - NIST Evaluations 2002-06
  - CHIL, AMI, …
  - State-of-the-art, Jan ‘06: „[WERs] of 30-40% and large differences to results with close-talking data“ [5]

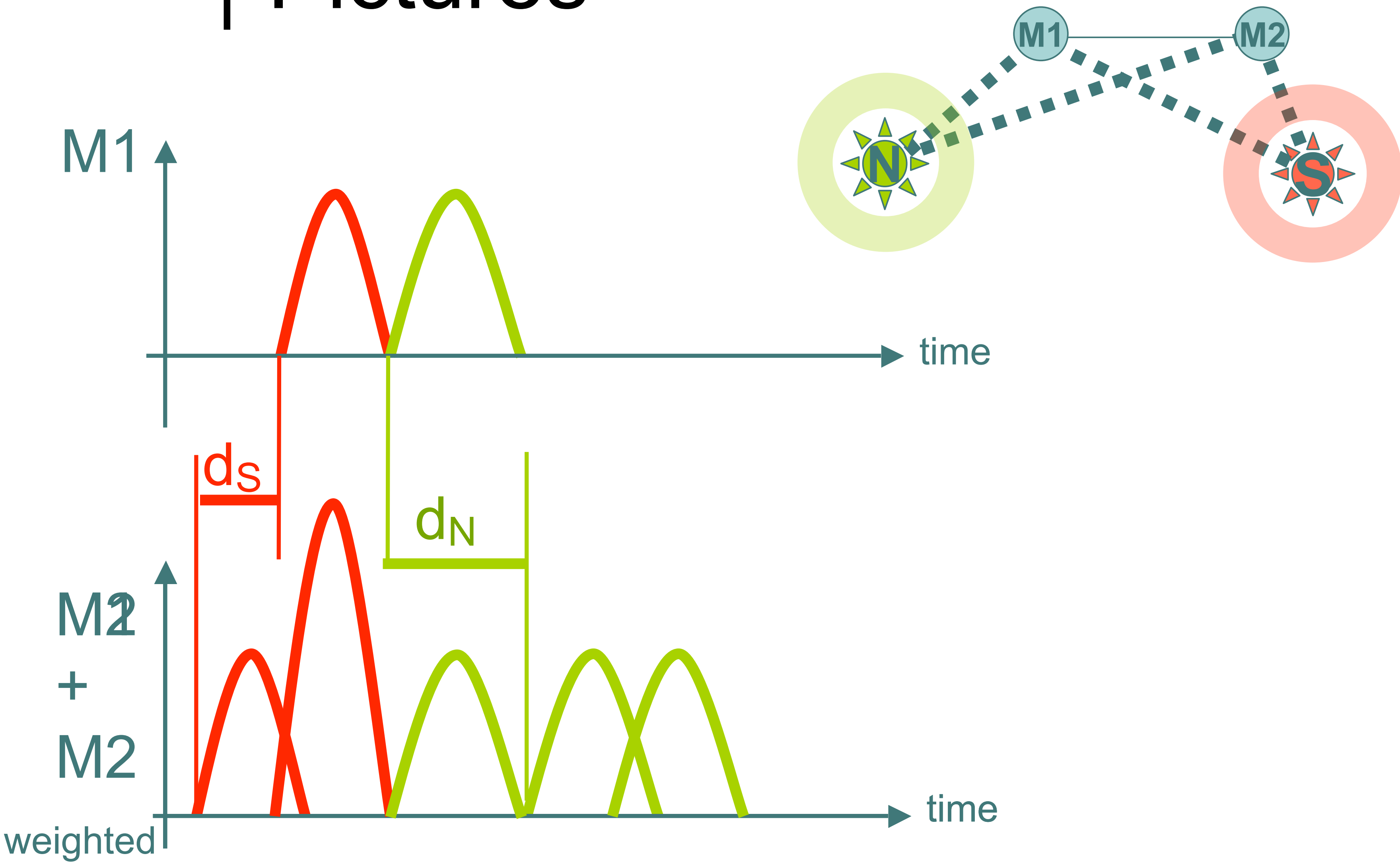
Background: Standard Beamforming

Beamforming in Simple Words

If we have a microphone array, a signal (sound wave) emitted by an off-axis source arrives at the various microphones at different times:



Beamforming in Simple Pictures



Beamforming Summary

- If we know the delay for each microphone (the look direction), we can align the signals and sum them
- Result: signal from desired direction is reinforced, signals from other directions are attenuated
- Simple: Delay-and-sum

$$y[n] = \sum_{m=0}^{M-1} \frac{1}{M} x_m[n - d_m]$$

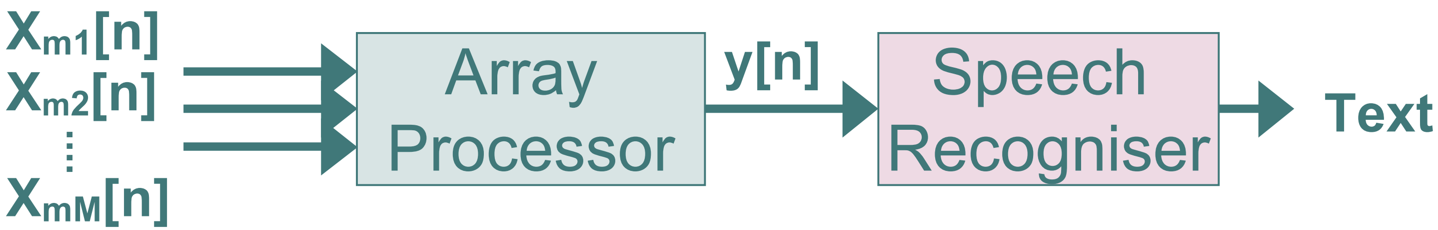
- Extension: Filter-and-sum

$$y[n] = \sum_{m=0}^{M-1} \sum_{p=0}^{P-1} h_m[p] x_m[n - p - d_m]$$

- Choices: Number of microphones *M*, filter length *P*. Parameters to set: delays *d*, filter taps $h_m[p]$

Traditional Beamforming + ASR

- Pipeline: first enhance speech with beamformer, then feed into recogniser



- Speech sounds much better, but WER for more complex beamformers does not improve much on delay-and-sum baseline

Delay&Sum WER vs. other techniques

- System trained on clean speech recorded from close-talking microphone. Data: CMU microphone array database, described later

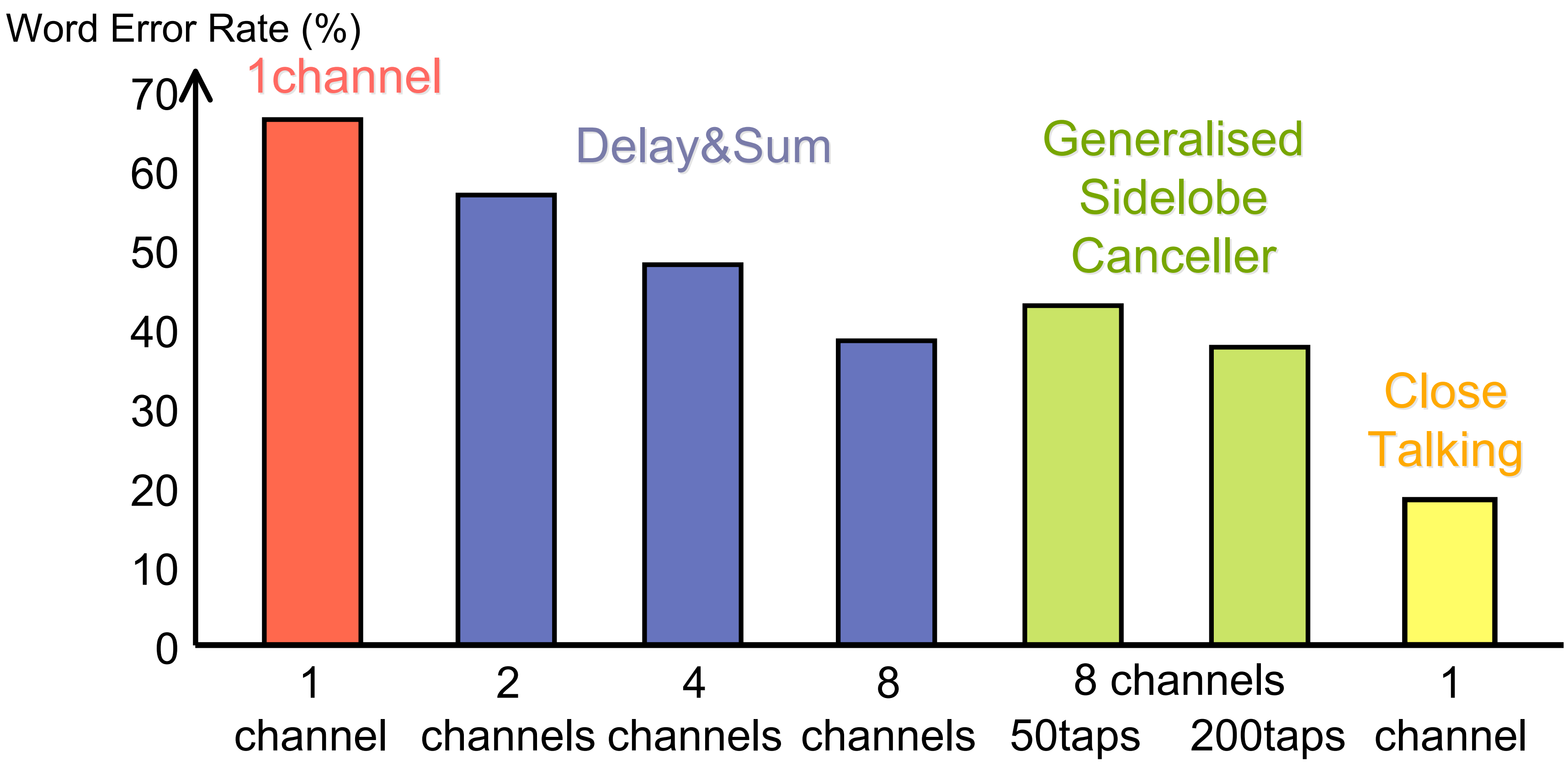


Figure based on [1].

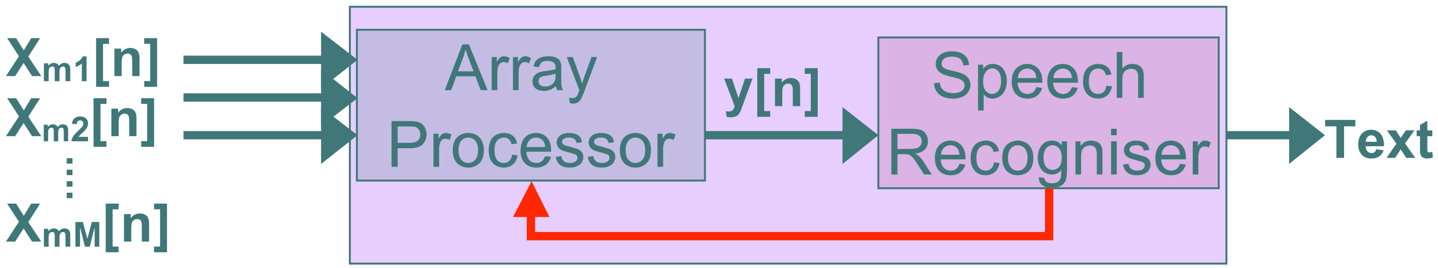
Maximum Likelihood Beamforming (MLB / LiMaBeam)

Seltzer's Data

- Recogniser trained on Wall Street Journal speech corpus (WJS0). 7000 training utterances, 84 speakers.
- Two test sets:
  - **CMU Microphone Array Database**
    Relatively noisy (6.5 dB avg. SNR).
    140 utterances, 10 speakers, vocab. size 138. Flat LM.
  - **Reverberant WSJ0 data**
    Not noisy, but reverberant (artificial); several test sets with different degree of reverberation.
    330 utterances, 8 speakers, vocab. size 5000. Trigram LM.
- We now have the same data at LSV for replication

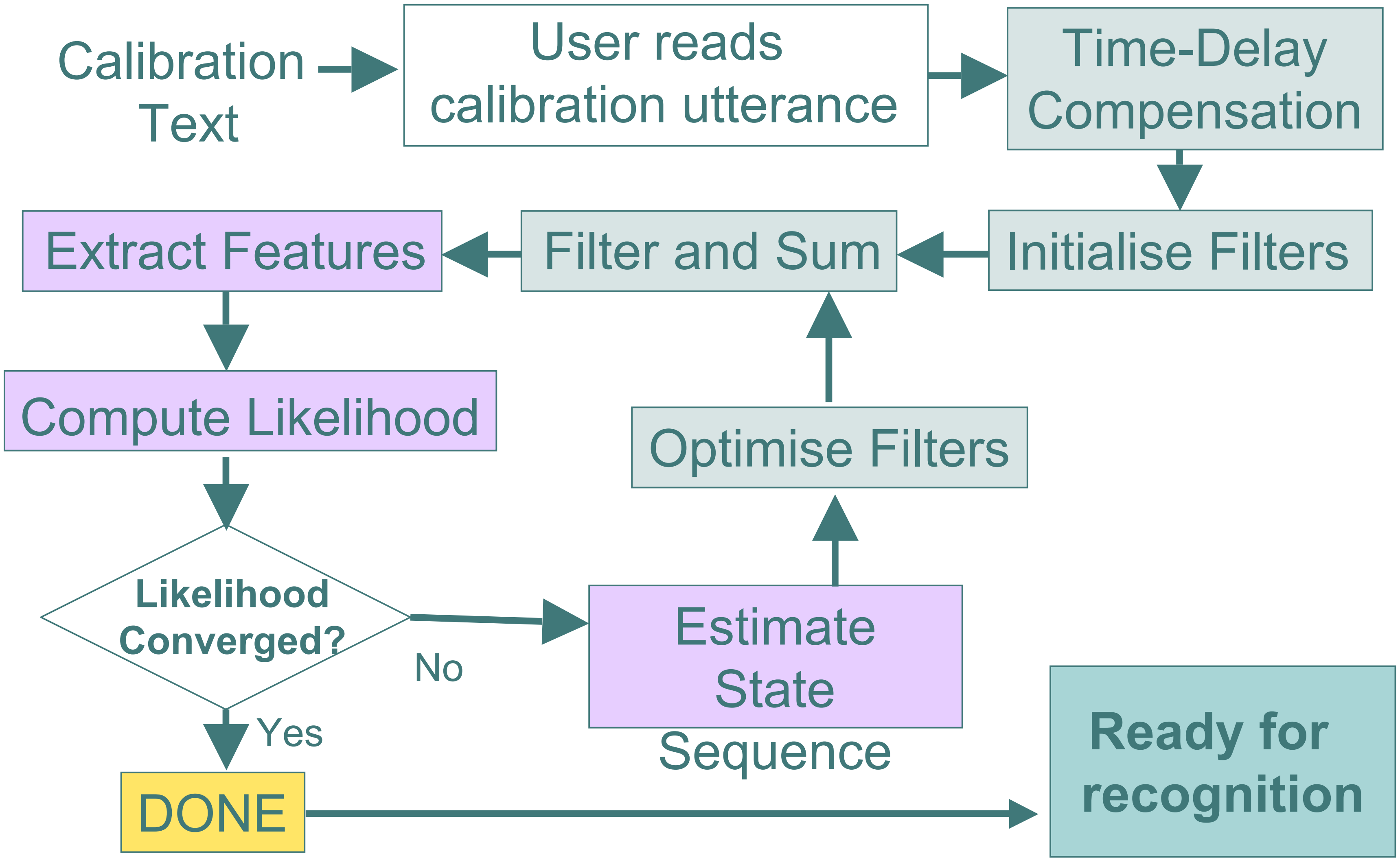
Basic Idea of LiMaBeam

- Break the pipeline; use WER-related criterion to optimise parameters of beamformer



- Iterative procedure, utterance-based:
  - Do beamforming
  - Decode (recognise) the utterance
  - Given most likely HMM state sequence, optimise the beamformer parameters for this sequence
  - Stop when likelihood has converged
- Recogniser parameters don't change, only filters

Calibrated Version



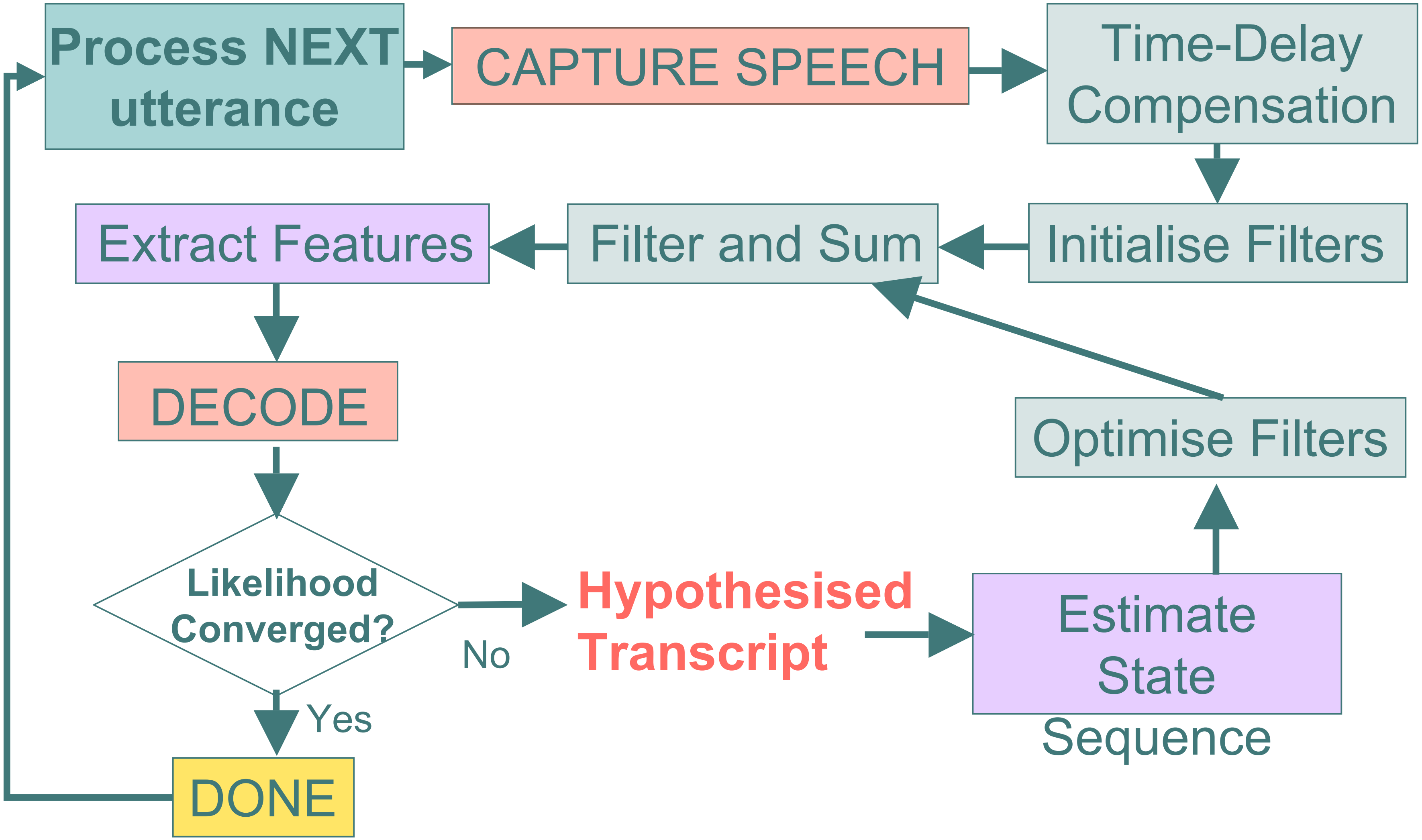
Results: Calibrated LiMaBeam

Method	Approx. WER
1 channel	65 %
Delay and Sum baseline	39 %
Calibrated Limabeam, 50 taps, **3.3 sec** calibration	36 %
CL, 50 taps, **8.3 sec** calibr.	33 %

- Duration of calibration utterance matters

Figure taken from [1]

Unsupervised Version



Results: Unsupervised LiMaBeam

- Average utterance duration influences accuracy:

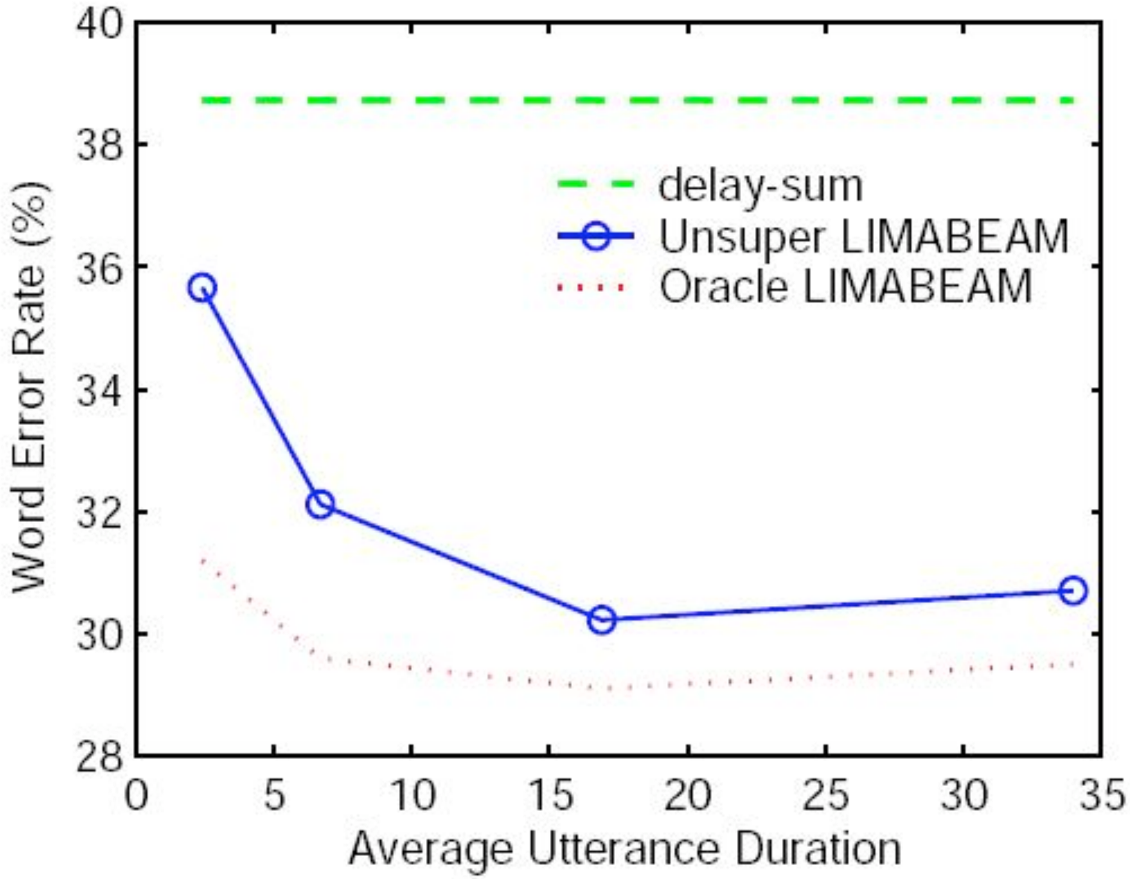


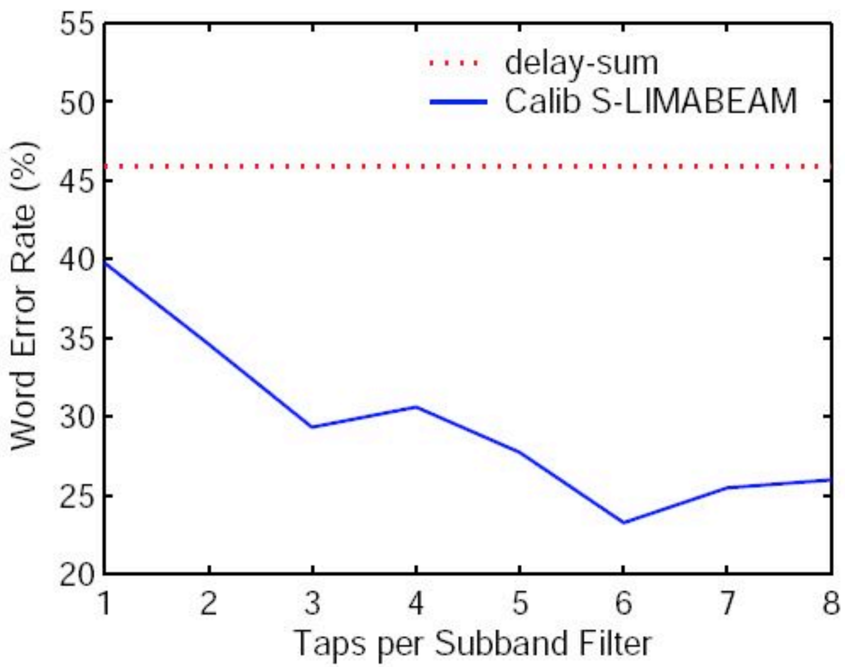
Figure taken from [1]

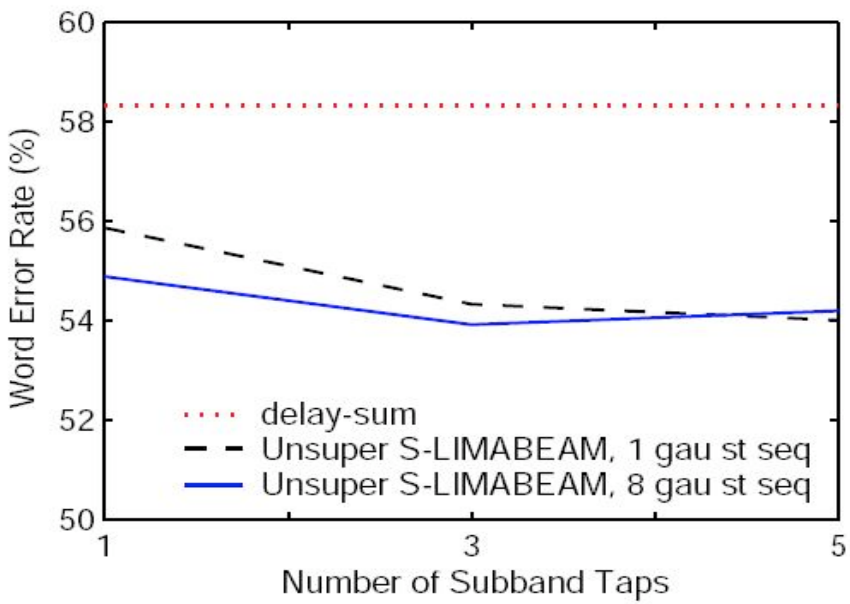
Subband Version

- Calibrated and unsupervised versions worked well on noisy data
- Not so well on reverberant data, same problem with conventional adaptive filtering techniques
- **Subband filtering** improves convergence of conventional adaptive filtering when filter is long and input signals highly correlated
- Input signal is decomposed into independent subbands, which are processed independently, then combined
- Seltzer developed a subband version of LIMABEAM which works also well on reverberant data

Results: S-Limabeam

- Both calibrated and unsupervised subband versions show significant improvement:





Figures taken from [1]

Our Project

- Start by replicating results (initially unsupervised version), exactly same data but different recogniser (HTK)
- Continue with different data
  - ‘Real data’, meeting room test
  - Edinburgh Multi-channel WSJ Audiovisual Corpus[3]?
- Consider related work (Karlsruhe [2], ITC-IRST)

Open Issues and Potential PhD Projects

Inter-related open issues:

- Continuous tracking of a moving speaker
- Multiple speakers
- Incorporate e.g. visual information about speaker

Summary

- Robustness problem in ASR
- Traditional approaches, specifically beamforming
- Maximum Likelihood Beamforming: promising but a number of open issues

References

[1] Michael L. Seltzer, "Microphone Array Processing for Robust Speech Recognition". Carnegie Mellon University, July 2003.

[2] Dominik Raub et al., "A Cepstral Domain Maximum Likelihood Beamformer for Speech Recognition", Proceedings of Interspeech 2004.

[3] Mike Lincoln et al., "The Multi-Channel Wall Street Journal Audio-Visual Corpus (MC-WSJ-AV): Specification and initial experiments", IEEE Automatic Speech Recognition and Understanding Workshop, Nov/Dec 2005.

[4] Bryan Pellom and Kadri Hacioglu, "Recent Improvements in the CU Sonic ASR System for Noisy Speech: The SPINE Task", Proceedings of ICASSP 2003.

[5] AMI "State-of-the-art Overview Conversational Multi-party Speech Recognition using remote microphones", January 2006, http://www.amiproject.org/state_overviews.php