SIGNAL PROCESSING FOR CLIENT/SERVER APPLICATIONS IN NEXT GENERATION CELLULAR TELEPHONY

Naveen Parihar and Joe Picone
Institute for Signal and Information Processing
Mississippi State University

- Contact Information:
  Box 9571
  Mississippi State University
  Mississippi State. Mississippi 39762
  Tel: 662-325-3149, Fax: 662-325-2298
  Email: parihar@isip.msstate.edu

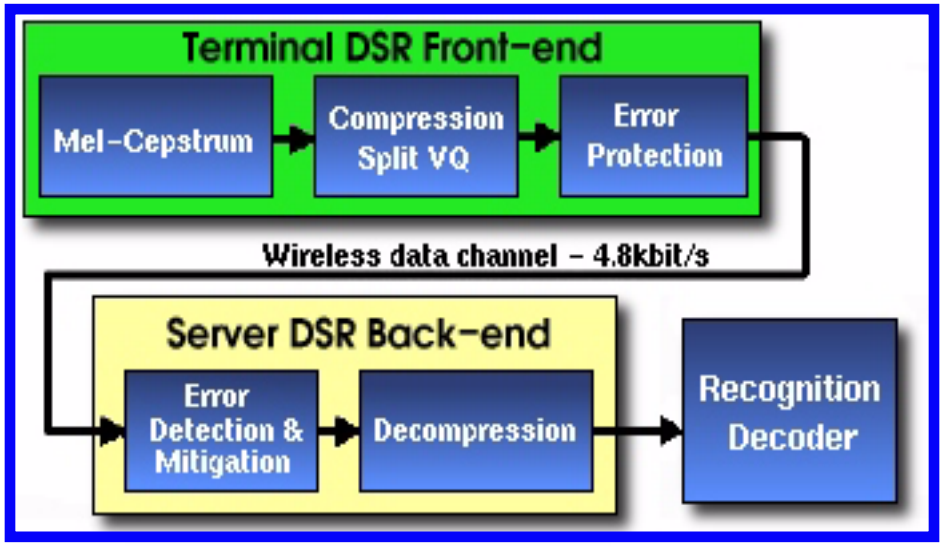
- This material is based upon work supported by the European Telecommunications Standard Institute. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the ETSI.

This talk is available at:
http://www.isip.msstate.edu/publications/seminars/ece_weekly/2003/aurora_evals

ORGANIZATION OF PRESENTATION

- Distributed Speech Recognition — client/server application
- Aurora Evaluations WI007 — objectives and results
- Aurora Evaluations WI008 — objectives
- LVCSR in Aurora Evaluations WI008
- Speech Recognition Review
- ISIP Baseline WSJ0 System — objectives and results
- Baseline results for Aurora Evaluations WI008 — analysis
- Aurora Evaluations WI008 results and state-of-the-art technology
- Conclusions and References

INTRODUCTION TO DSR



- Distributed Speech Recognition (DSR) — client/server application
- Terminal DSR front end with limited compute power
- Common back end speech recognition system
- Speech recognition in adverse environments
- Proposal for an advanced front end (AFE) standard for LVCSR applications

AURORA EVALUATIONS WI007

- Objective — evaluate front ends in adverse environment with an aim towards the development of the future advanced front end (AFE)
- Two MFCC front ends — Aurora, and HTK
- Small vocabulary task — TIDigits (11 words)
- Speech recognition technology — HMM-based word models
- Matched Conditions (Average WER across 4 noise types)

Front end	SNR/db							
	clean	20	15	10	5	0	-5	Average
Aurora	1.5%	2.3%	3.1%	5.1%	12.2%	38.3%	75.4%	12.2%
HTK	1.5%	2.5%	3.1%	5.4%	12.4%	40.2%	76.5%	12.7%

- Mismatched Conditions (Average WER across 4 noise type)

Front end	SNR/db							
	clean	20	15	10	5	0	-5	Average
Aurora	1.5%	2.8%	4.2%	6.9%	15.3%	39.5%	74.1%	13.8%
HTK	1.5%	3.1%	4.6%	7.4%	16.2%	41.4%	76.2%	14.5%

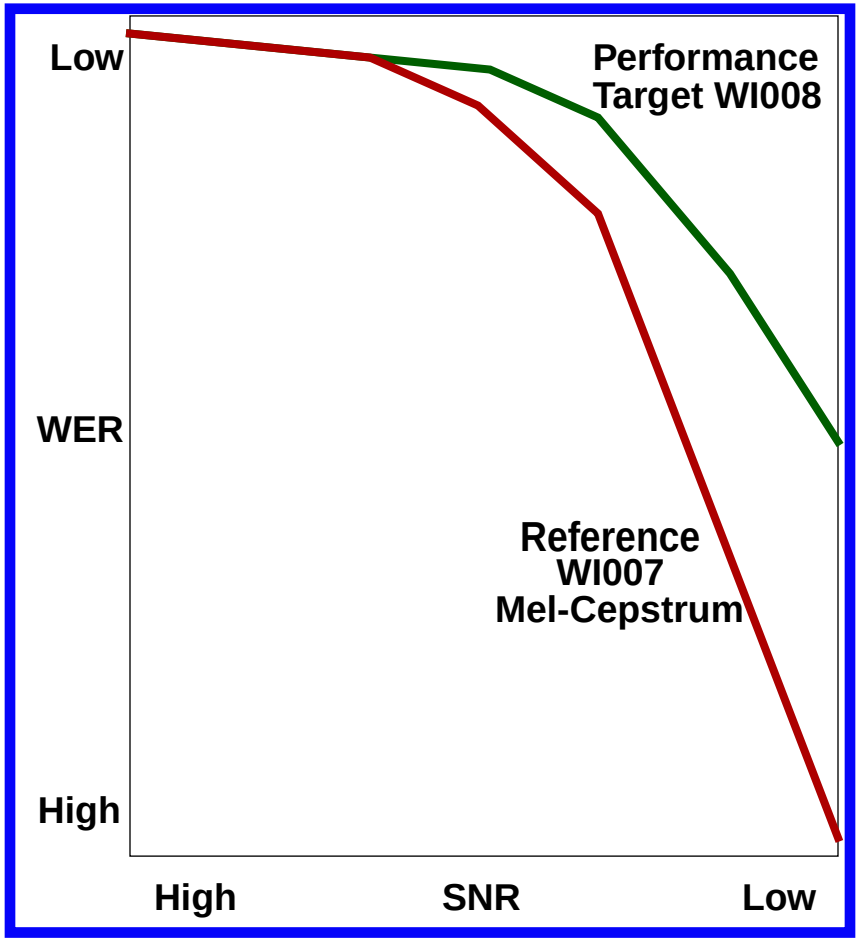
AURORA EVALUATIONS WI008

- Objective — standardize more advanced front end (AFE-WI008) than the MFCC WI007 front end

- Extended to Large Vocabulary Task — WSJ0 (5000 words)

- Extension of the AFE to include a range of European languages — SpeechDat-Car noisy sub-sets in Finnish, Italian, Spanish, German, and Danish

- Performance of WI008 in low noisy background conditions not worst than WI007, and significantly better in demanding environments



- Speech recognition technology — HMM-based word, and sub-word models

- Custom real-time Voice Activity Detection (VAD) algorithm

LVCSR FOR AURORA EVALUATIONS

- Wall Street Journal Task (WSJ0) — closed-loop 5000 words, read sentences out of the Wall Street Journal Magazine

- WI008 AFE must have a performance improvement of at least 25% over the MFCC WI007 baseline

- Two sampling frequencies — 16 kHz and 8 kHz

- Three training conditions to account for model-match and model-mismatch into the final evaluation

- Seven additive noise conditions at various SNR's — clean, street-traffic, train-station, car, babble, restaurant, and airport

- Filtering to simulate the frequency characteristics of the terminal device

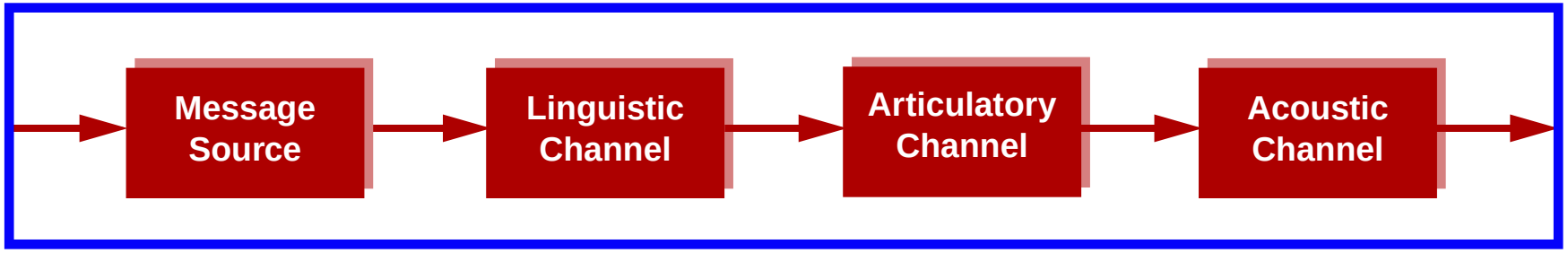
- Two microphone conditions

- Lossy VQ-based compression algorithm

- Utterance detection (VAD)

SPEECH RECOGNITION OVERVIEW

A noisy communication channel model for speech production and perception:



Bayesian formulation for speech recognition:

$$P(W|A) = P(A|W)P(W)/P(A)$$

Objective: minimize the word error rate by maximizing $P(W|A)$

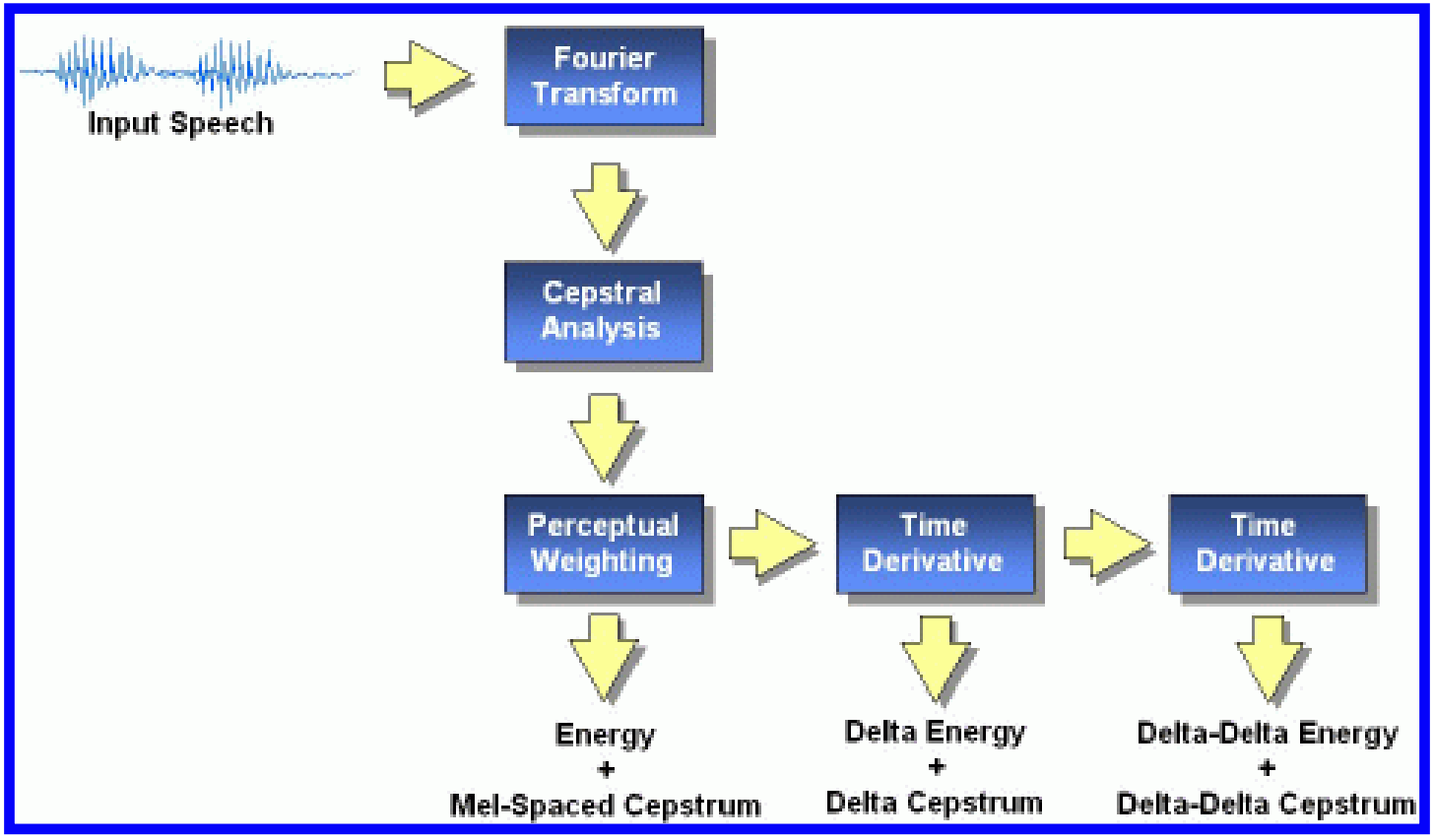
Approach: maximize $P(A|W)$ (training)

Components:

- $P(A|W)$: acoustic model (hidden Markov models, Gaussian mixtures)
- $P(W)$: language model (statistical, N-grams, finite state networks)
- $P(A)$: acoustics (ignore during maximization)

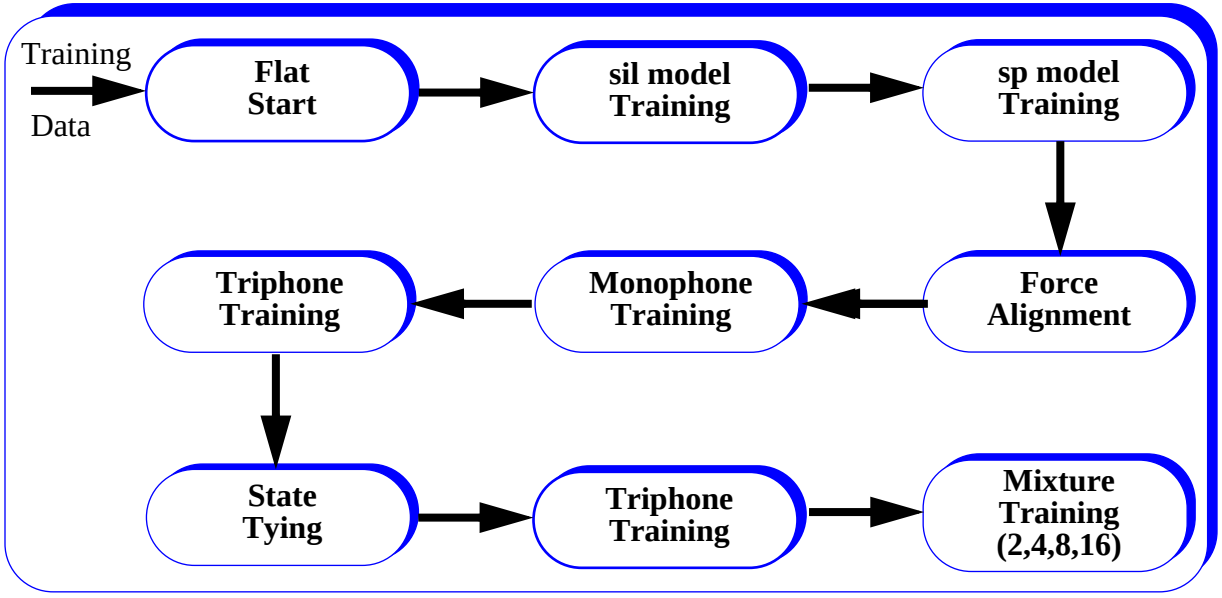
The language model typically predicts a small set of next words based on knowledge of a finite number of previous words (N-grams) — leads to search space reduction.

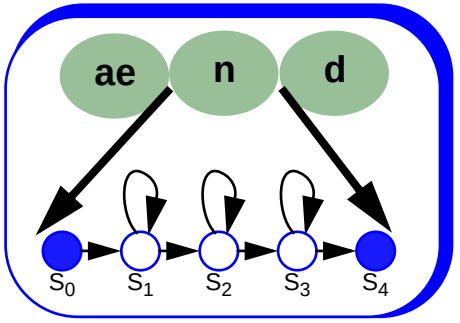
ISIP BASELINE WSJ0 SYSTEM



- 100 frames per second, 25 msec Hamming window
- 12 base FFT-derived mel cepstra with CMS and log-energy
- Energy normalization
- Delta and acceleration coefficients

ISIP BASELINE WSJ0 SYSTEM (cont.)





- ~15 hours of WSJ0 training data including 83 speakers
- phonetically state-tied 16-mixture cross-word triphone models
- ~40 minutes of November'92 WSJ0 evaluation data used for decoding
- single pass decoding using a bigram backoff language model

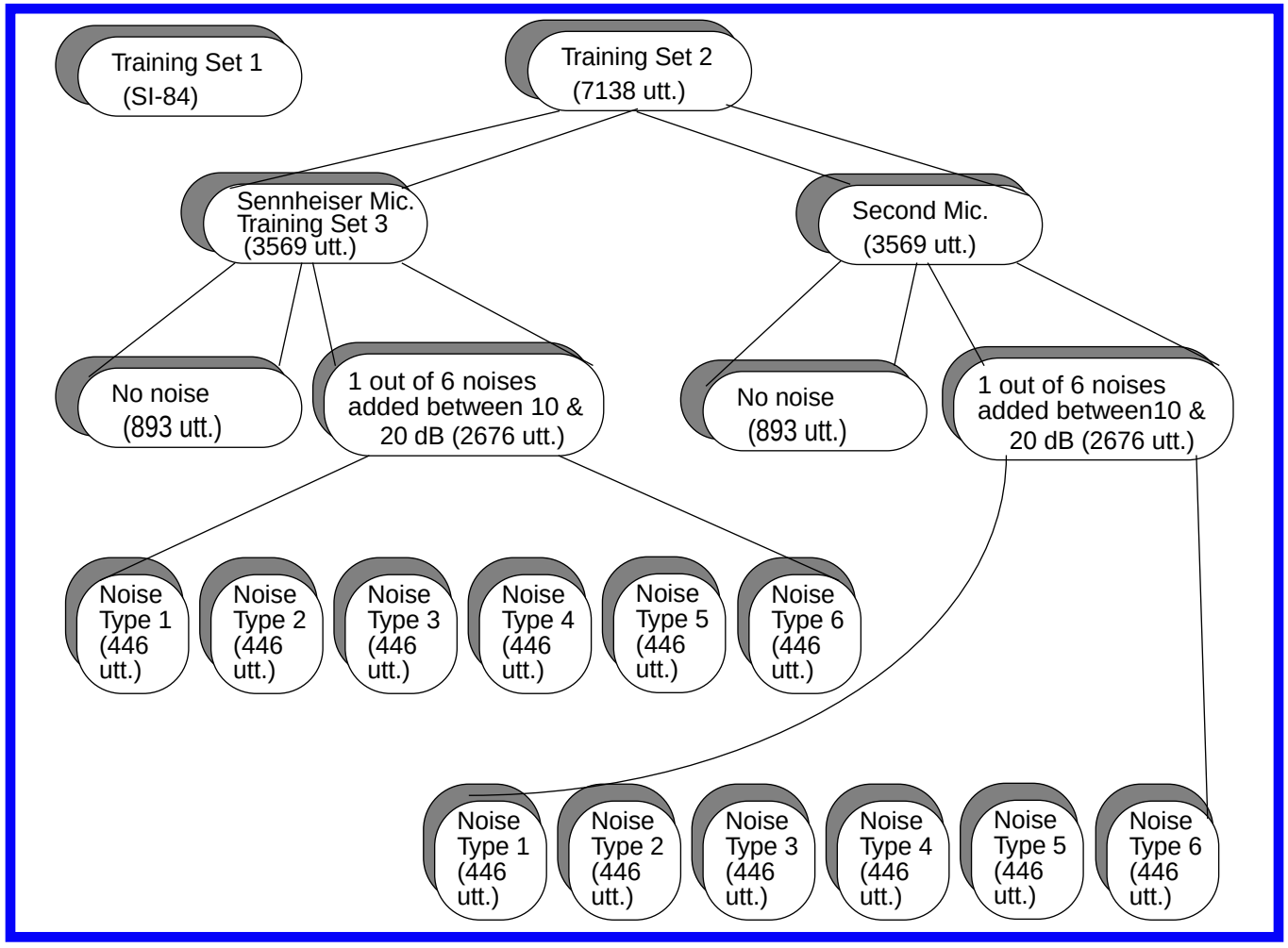
Site	Acoustic Model Type	WER
CU	word-internal / gender-independent	8.1%
UT	word-internal / gender-dependent	7.1%
ISIP	cross-word / gender-independent	8.2%
CU	cross-word / gender-independent	6.9%
LT	cross-word / gender-independent	6.8%

CU: Cambridge University, UK
UT: University of Technology, Germany
LT: Lucent Technologies

WI007 BASELINES

Training Database Design

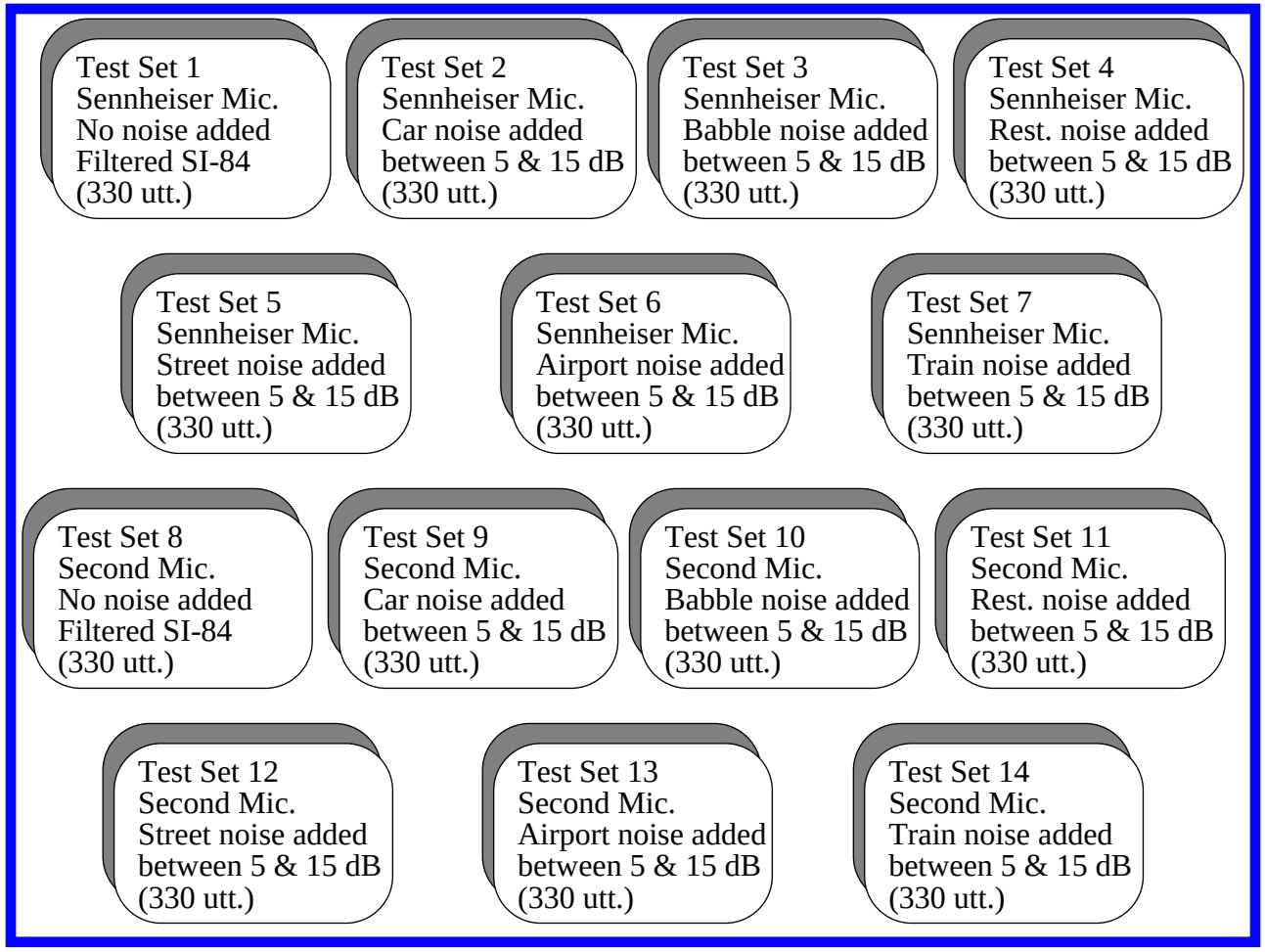
- **Three training sets — Training Set 1, Training Set 2, Training Set 3**
- **6 noise conditions, randomly chosen SNR between 10 and 20 dB**
- **2 microphone conditions — Sennheiser mic. and Secondary mic.**
- **G.712 filtering at 8 kHz and P.341 filtering at 16 kHz**



WI007 BASELINES (cont.1)

Evaluation Database Design

- **14 test sets**
- **7 recorded on Sennheiser mic. and 7 recorded on Secondary mic.**
- **6 noise conditions, randomly chosen SNR between 5 and 15 dB**
- **G.712 filtering at 8 kHz and P.341 filtering at 16 kHz**



WI007 BASELINES (cont.2)

Computational Considerations

- 11 Baseline training conditions and 14 test conditions for each training set means ~1034 days on an 800 MHz x86 CPU!

Training Sets	Training Time (days)	Decoding Sets	Decoding Time (days)	Total Time (days)
1	~10			~10
		1	~6	~ 6
11	~11*10	14	~11*14*6	~1034

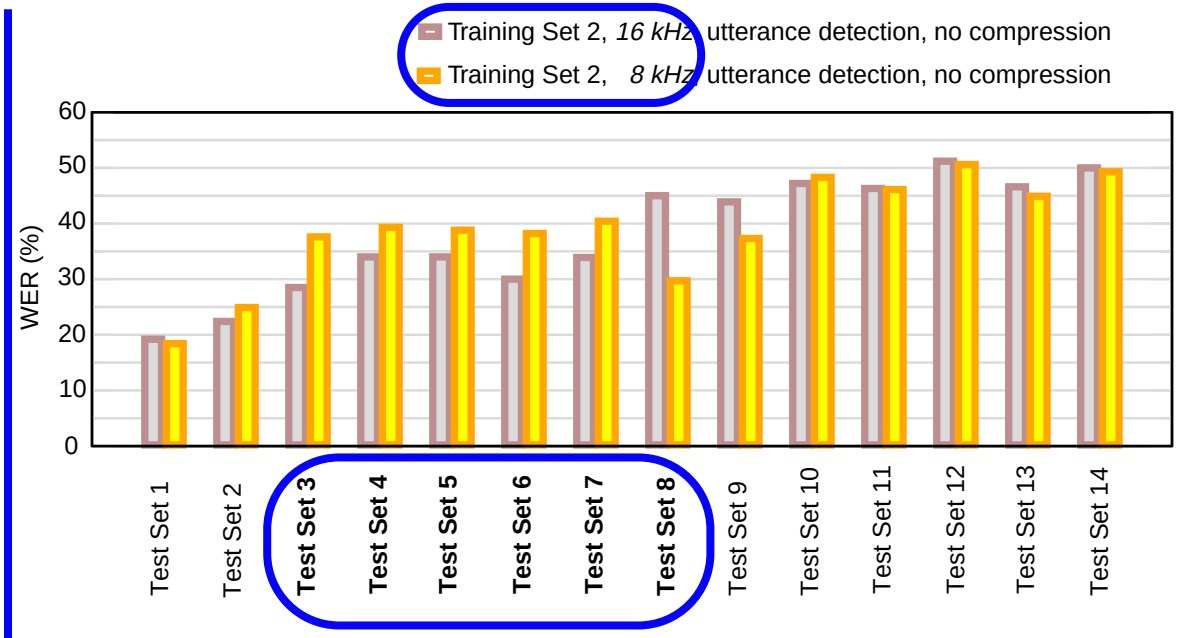
Three steps to reduce the CPU requirements

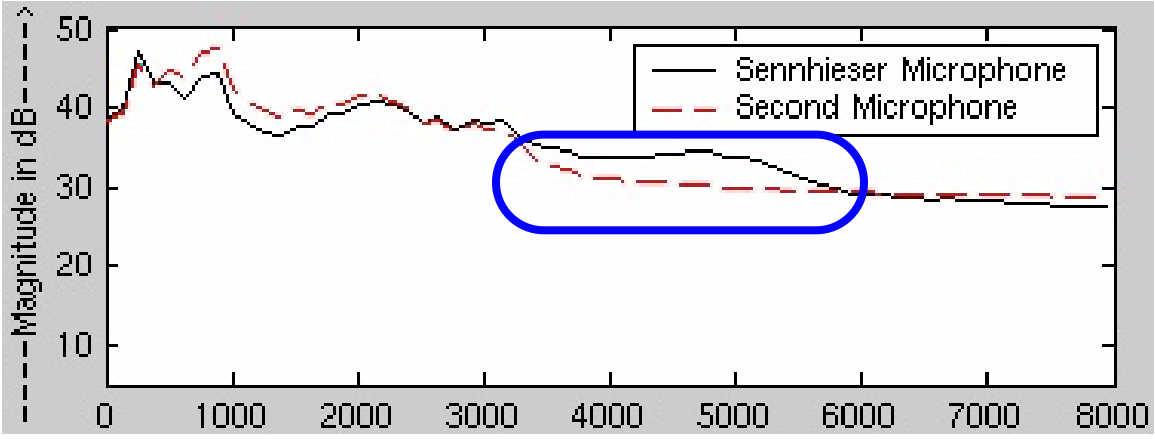
- Reduction in the evaluation set from 330 utterances to 166 utterance — expected reduction in the total decoding time by 50%
- Reduction in the number of gaussian mixtures from 16 to 4 — expected reduction in total training time by a factor of 7/9
- Prune the beam widths — expected reduction in the total decoding time by a factor of 6 with a modest degradation in the performance
- Total expected experimental time dropped to ~163 days on an 800 MHz CPU

WI007 BASELINE RESULTS

Sampling frequency reduction — 16 kHz to 8 kHz

- No trend on mismatched conditions (Training Set 1)
- Significant degradation on Sennheiser mic. on matched conditions (Training Set 2) — attributed to additional information provided by high sampling frequency
- No degradation on perfectly matched conditions (Training Set 1 and Test Set 1) — additional information provided by high frequencies does not influence performance

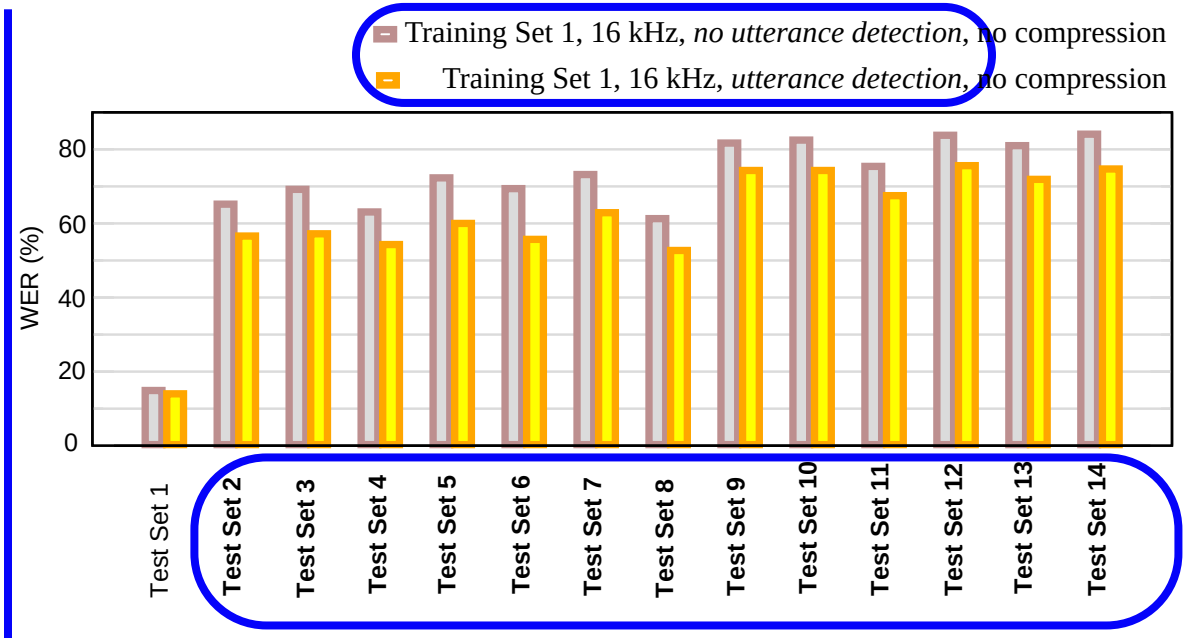




WI007 BASELINE RESULTS (cont.1)

Utterance Detection — Ideal endpoints (200 ms of silence at the start and the end of each utterance)

- Significant improvement in mismatched conditions (Training Set 1) — "silence" model did not model noisy silence and hence, increase in "insertion" type errors
- No significant improvement in matched conditions (Training Set 2) — "silence" model learnt the noisy non-speech segments

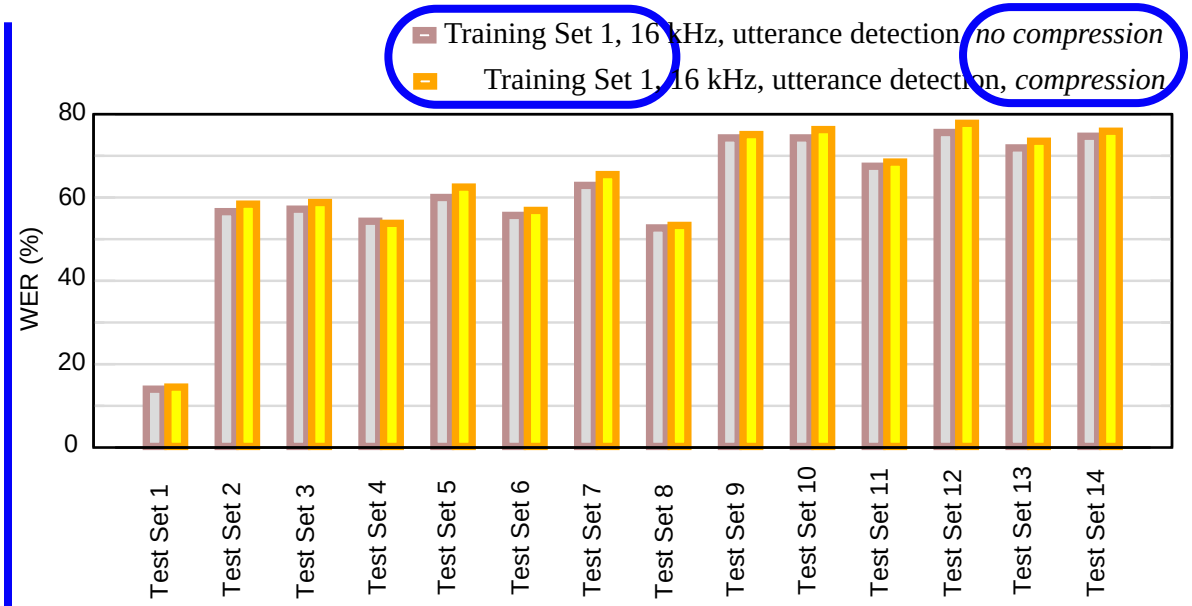


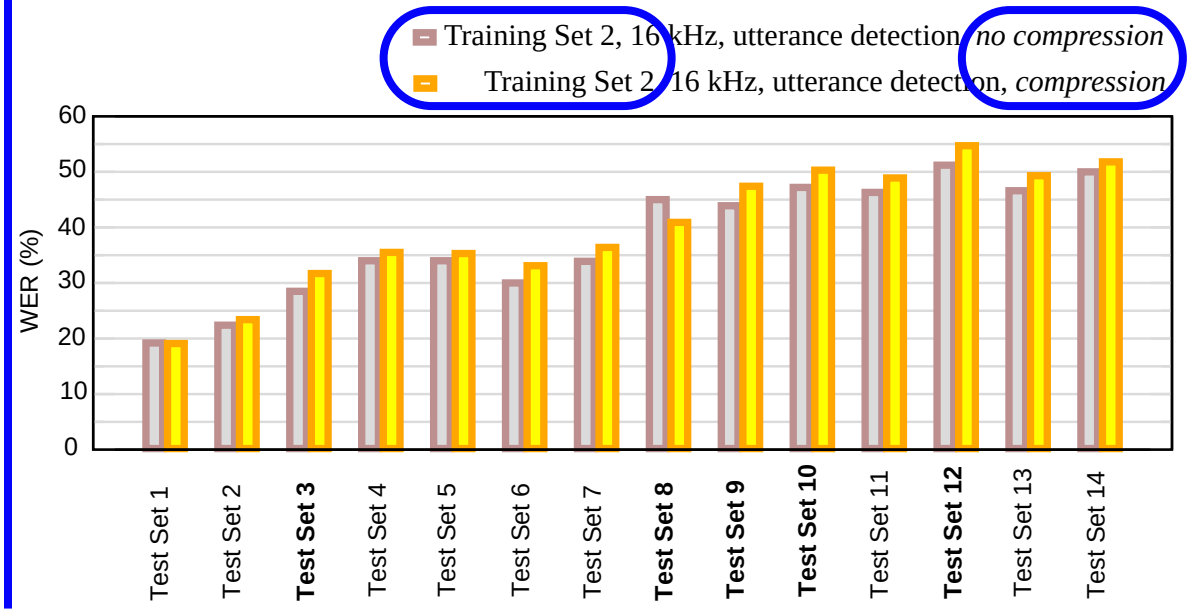
Test Set	Training Set 1							
	Without Utterance Detection				With Utterance Detection			
	WER	Sub.	Del.	Ins.	WER	Sub.	Del.	Ins
1	14.9%	8.8%	1.0%	5.1%	14.0%	9.0%	0.8%	4.1%
2	65.2%	41.4%	3.6%	20.1%	56.6%	40.0%	3.6%	13.0%
3	69.2%	46.0%	6.5%	16.7%	57.2%	40.7%	6.2%	10.2%
4	63.1%	40.5%	12.0%	10.6%	54.3%	36.7%	10.8%	6.9%
5	72.3%	47.0%	11.2%	14.1%	60.0%	39.2%	13.8%	7.1%
6	69.4%	44.6%	7.8%	17.0%	55.7%	37.9%	8.2%	9.6%
7	73.2%	46.6%	14.1%	12.5%	62.9%	42.1%	13.7%	7.1%

WI007 BASELINE RESULTS (cont.2)

Vector quantization based lossy compression of feature-vectors

- No significant degradation in mismatched conditions (Training Set 1)
- Significant degradation on five noisy conditions (3, 8, 9, 10, 12) in matched conditions (Training Set 2) — no consistency
- Two sets of VQ code books — one for speech sampled at 8 kHz or 11 kHz, and one for speech sampled at 16 kHz

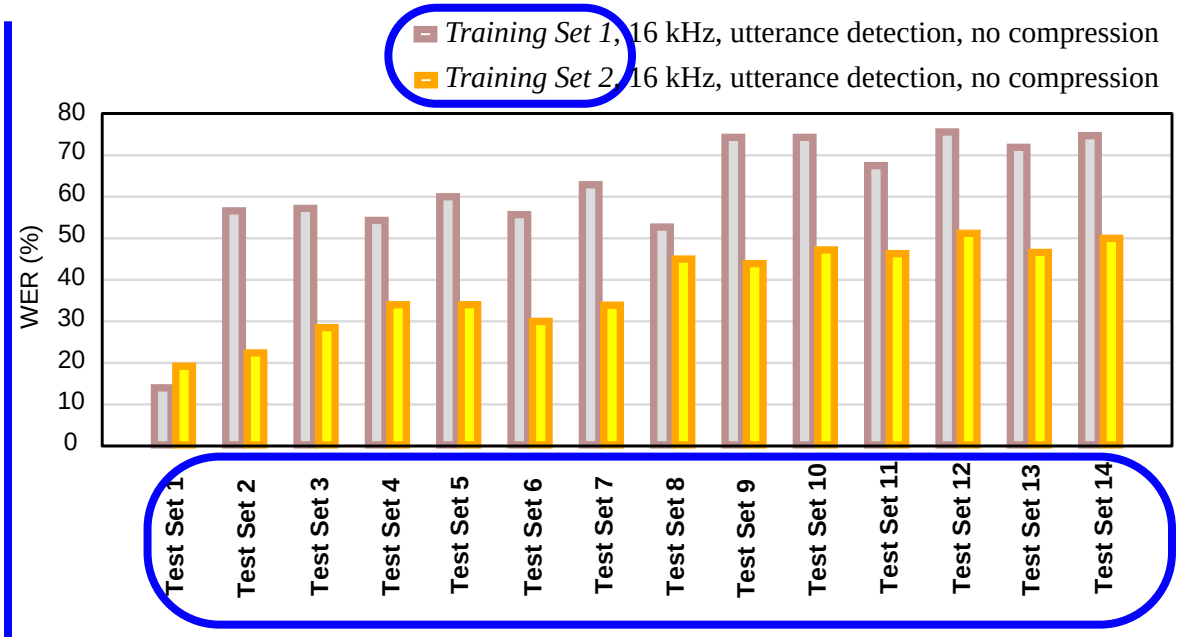


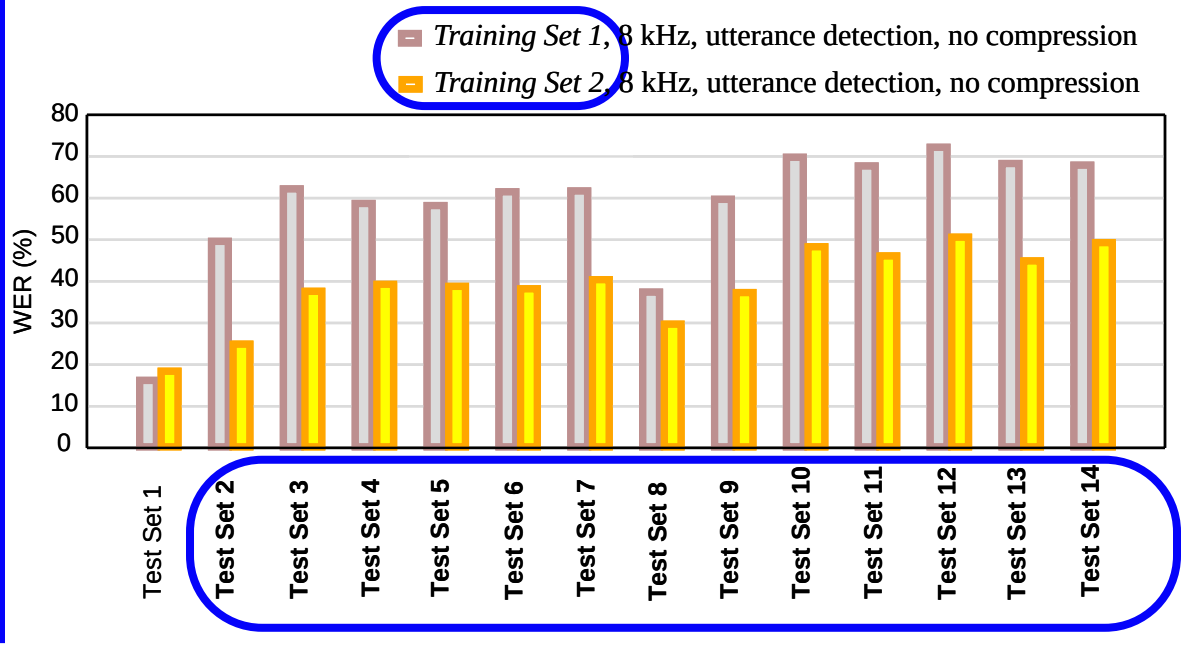


WI007 BASELINE RESULTS (cont.3)

Model Mismatch

- Best performance on perfectly matched conditions (Training Set 1 and Test Set 1)
- Significant degradation in all the mismatched conditions (Training Set 1)
- Matched conditions (Training Set 2) significantly better than the mismatched conditions
- Under maximum likelihood framework, high performance can only be achieved when the test and the training conditions are similar





WI007 BASELINE RESULTS (cont.4)

Microphone Variation — Sennheiser Microphone vs. Secondary Microphone

- Sennheiser HMD-414 is a high quality close-talking microphone whereas the second microphone is one of the 18 common microphone types
- In general, the Sennheiser microphone performed significantly better than the second microphone condition
- Largest degradation observed on clean test condition on Training Set 1
- Less severe but still significant degradation on Training Set 2 — value in exposing the models during training to second microphone condition

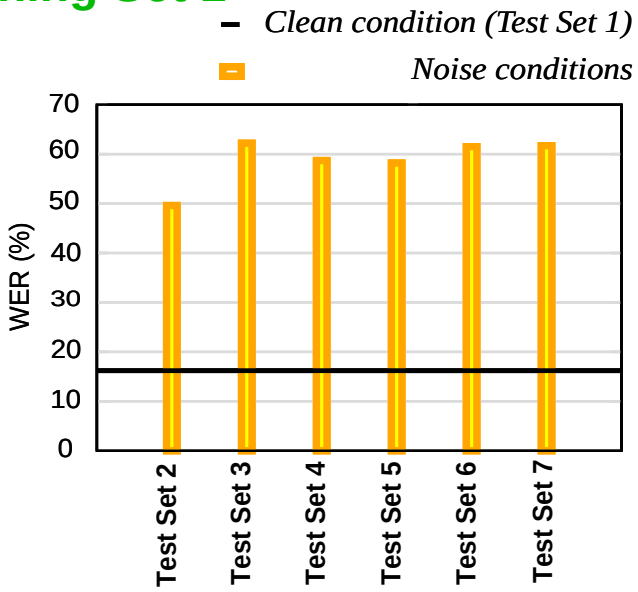
Performance (Without Compression)						
Training Set			Test Set			
Set	Sampling Frequency	Utterance Detection	1 (Sennheiser, Clean)	8 (Second, Clean)	2 (Sennheiser, Car)	9 (Second, Car)
1	8 kHz	Yes	16.2%	37.4%	49.6%	59.7%
2	8 kHz	Yes	18.4%	29.7%	24.9%	37.3%

WI007 BASELINE RESULTS (cont.5)

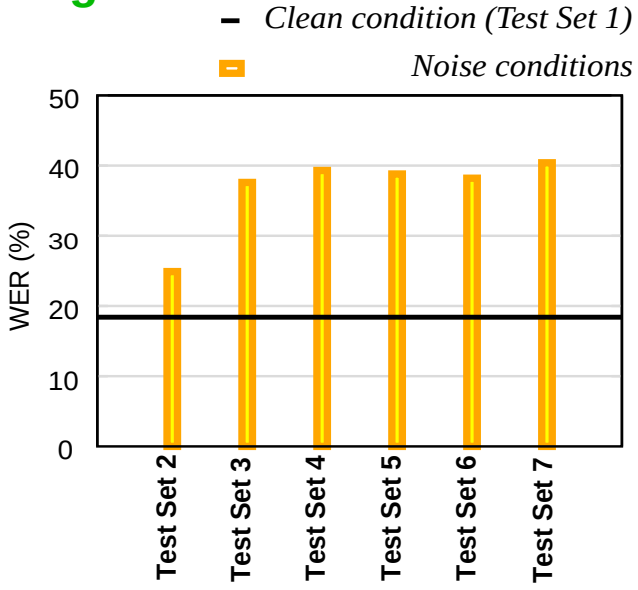
Additive noise

- 7 test conditions — clean, street-traffic, train-station, car, babble, restaurant, and airport
- Training sets SNR randomly chosen between 10 - 20 db
- Test sets SNR randomly chosen between 5 -15 db
- Severe degradation on all noisy conditions
- Severity of the degradation can be limited, though still significant, by exposing the models to noise conditions during the training process (Training Set 2)

Training Set 1



Training Set 2



AURORA EVALUATION RESULTS

- Two participating sites
  * Motorola-FranceTelecom-Alcatel (MFA)
  * Qualcomn-ICSI-OGI (QIO)

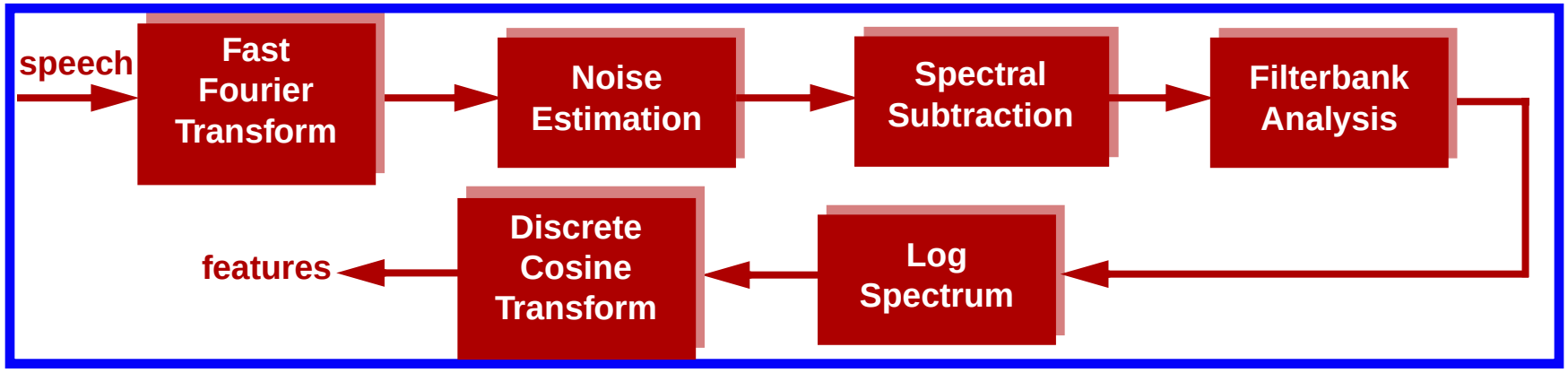
- MFA advanced front end chosen as WI008 standard

- ~30% relative improvement over the baseline WI007

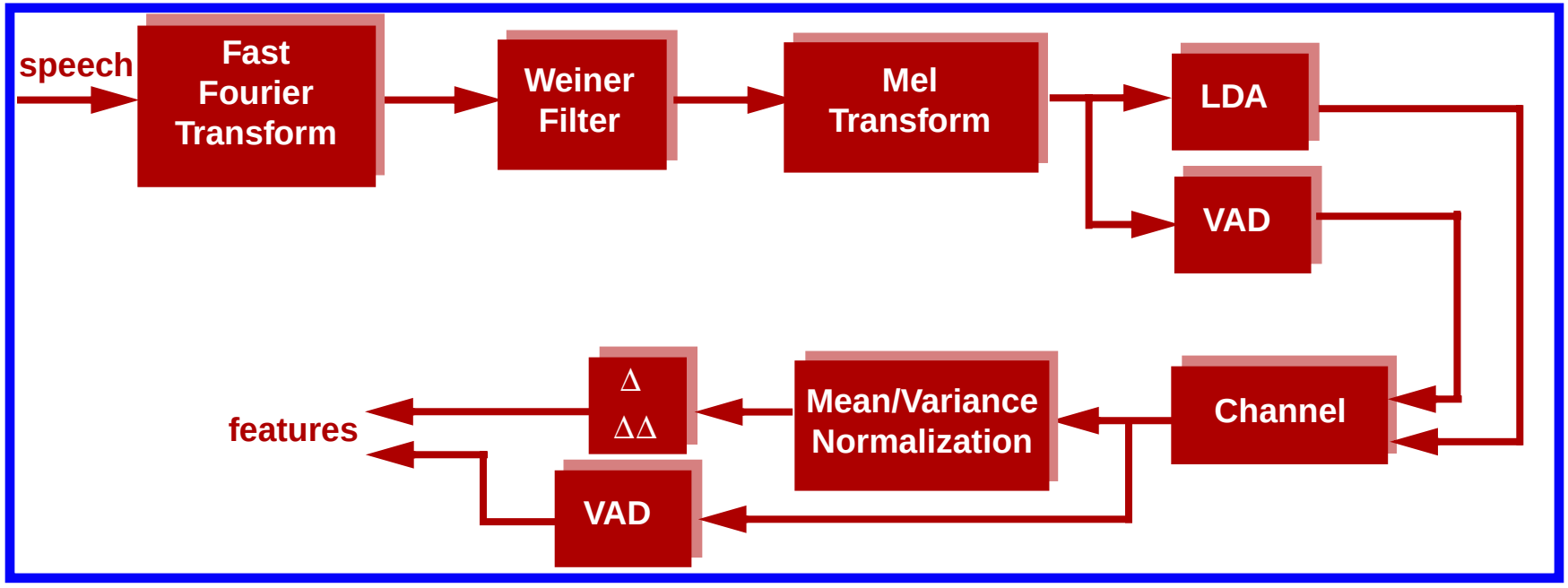
	Performance Summary						
Site	Training Set		Test Set				
			Sennheiser Microphone		Secondary Microphone		Average WER
	Set	Sampling Frequency	Clean	Noise	Clean	Noise	
Base	1	8 kHz	15.4%	49.4%	36.6%	59.9%	58.1%
MFA	1	8 kHz	15.0%	36.4%	23.1%	44.8%	37.5%
QIO	1	8 kHz	16.5%	42.5%	28.7%	50.7%	43.2%
Base	2	8 kHz	20.7%	26.4%	30.9%	38.7%	41.0%
MFA	2	8 kHz	17.2%	30.6%	22.7%	36.1%	31.4%
QIO	2	8 kHz	20.8%	32.7%	23.6%	38.3%	33.6%

STATE OF THE ART

Commercial front ends use adaptive noise compensation



Advanced front ends use a variety of techniques including subspace methods, normalization, and multiple time scales



CONCLUSIONS

- Advanced front end standardization is an elaborate effort!
- Computational resources is a big issue with LVCSR
- Exposing models to different noisy conditions and microphone conditions improves the speech recognition performance in adverse conditions
- Vector Quantization based compression is robust in DSR environment
- 16 kHz sampling frequency results in significant improvement only in noisy test conditions
- Future directions:
  Signal Processing algorithms vs. adaptation in maximum likelihood framework

REFERENCES

[1] "ETSI ES 201 108 v1.1.2 Distributed Speech Recognition; Front-end Feature Extraction Algorithm; Compression Algorithm," ETSI, April 2000.

[2] D. Pearce, "Enabling New Speech Driven Services for Mobile Devices: An overview of the ETSI standards activities for Distributed Speech Recognition Front-ends," presented at the Applied Voice Input/Output Society Conference (AVIOS2000), San Jose, California, USA, May 2000.

[3] D. Pearce, "Overview of Evaluation Criteria for Advanced Distributed Speech Recognition," ETSI STQ-Aurora DSR Working Group, October 16, 2001.

[4] D. Pearce, "Advanced DSR Front-end: Definition of Required Performance Characteristics," ETSI STQ-Aurora DSR Working Group, October 15, 2001.

[5] D. Paul and J. Baker, "The Design of Wall Street Journal-based CSR Corpus," *Proceedings of the International Conference on Spoken Language Systems (ICSLP)*, pp. 899-902, Banff, Alberta, Canada, October 1992.

[6] G. Hirsch, "Experimental Framework for the Performance Evaluation of Speech Recognition Front-ends on a Large Vocabulary Task," ETSI STQ Aurora DSR Working Group, June 2001.

[7] N. Parihar and J. Picone, "DSR Front End LVCSR Evaluation AU/384/02," *http://www.isip.msstate.edu/ publications/ reports/aurora_frontend/2002/report_012202_v21.pdf*, Aurora Working Group, December 2002.

[8] "Recommendation G.712 — Transmission performance characteristics of pulse code modulation channels," International Telecommunication Union (ITU), Geneva, Switzerland, November 1996.

[9] "Recommendation P.341 — Transmission characteristics for wideband (150-7000 Hz) digital hands-free telephony terminals, International Telecommunication Union (ITU), Geneva, Switzerland, February 1998.

[10] "Benchmark Tests, Matched Pairs Sentence-Segment Word Error (MAPSSWE)," *http://www.nist.gov/speech/tests/ sigtests/mapsswe.htm*, Speech Group, NIST, USA, January 2001.

[11] P. C. Woodland, J. J. Odell, V. Valtchev, and S. J. Young, "Large Vocabulary Continuous Speech Recognition using HTK," *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing,* Adelaide, Australia, pp. II/125-II/128, April 1994.

[12] K. Beulen, and H. Ney, "Automatic Question Generation for Decision Tree Based State Tying," *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, pp. 805-808, Seattle, Washington, USA, April 1998.

[13] W. Reichl, and W. Chou, "Decision tree state tying based on segmental clustering for acoustic modeling," *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, pp 801-804, Seattle, WA, USA, April 1998.