

# The SRI-ICSI Spring 2007 Meeting Recognition System

*Andreas Stolcke*

*Xavier Anguera*

*Kofi Boakye*

*Özgür Çetin*

*Adam Janin*

*Mathew Magimai-Doss*

*Chuck Wooters*

*Jing Zheng*

SRI International, Menlo Park, CA, USA

International Computer Science Institute, Berkeley, CA, USA

**Funding:** DARPA/CALO (SRI), AMI & IM2 (ICSI)



# Overview

- What's new this year
- Data and microphone conditions
- System architecture
- Audio preprocessing
- Acoustic modeling
- Language modeling
- Eval system results
- Conclusions



# What's New This Year ... and What Isn't

- MDM, ADM, etc.: cleaned-up beamforming software
- IHM: energy normalization in cross-channel features
  - Already presented as a post-eval improvement last year
- Acoustic modeling:
  - Unified acoustic models for all conditions
  - All models trained with fMPE-MAP and MPE-MAP
  - Additional AMI and NIST training data
  - Retained feature MLP from last year
- Language modeling:
  - Incorporated additional AMI and NIST transcripts for confmtg LM
  - But lectmtg LM remained unchanged
- Same system architecture
  - based on SRI CTS system
- New genre: Coffee breaks
- New task: Speaker-attributed speech-to-text (SASTT)



# Preliminaries



# Conference Meeting Datasets

- **eval07**: NIST RT-07S conference meetings
- **eval05, eval06**: NIST RT-05S and RT-06S meetings
  - Used for development
  - Many parameters not retuned from last year
- Meeting training data
  - AMI (170 meetings, 100 hours) – **more data this year**
  - CMU (17 meetings, 11 hours) – Lapel personal mics, no distant mics
  - ICSI (73 meetings, 74 hours)
  - NIST (27 meetings, 28 hours) – **more data this year**
- Acoustic background training data (same as last year)
  - CTS (Switchboard + Fisher, 2300 hours)
  - BN (Hub-4 + TDT2 + TDT4, 900 hours)

# Lecture Datasets

- **eval06: RT-06S lecture eval set**
  - Used for development
  - No independent test set was available
  - Many parameters (e.g., rescoring weights) were copied from conference meeting system without retuning
- **Training data:**
  - All conference training data
  - Background data as for conference data
  - CHIL training data (close-talking mics only, 38 meetings, ~7 hours)
  - CHIL dev06 distant mic data
  - TED lecture recordings (boom mics only, 39 meetings, ~9 hours)
  - **We didn't have time to process any of the CHIL data released since RT-06S!**
  - **In particular: we “developed” the cbreak system by guessing the best components from the confmtg and lectmtg systems**

# Evaluation Tasks

Conference room meetings (**confmtg**):

- **MDM** Multiple distant microphones
- **IHM** Individual headset microphones
- **SDM** Single distant microphone

Lecture room meetings (**lectmtg**) and coffee breaks (**cbreak**), in addition to the above:

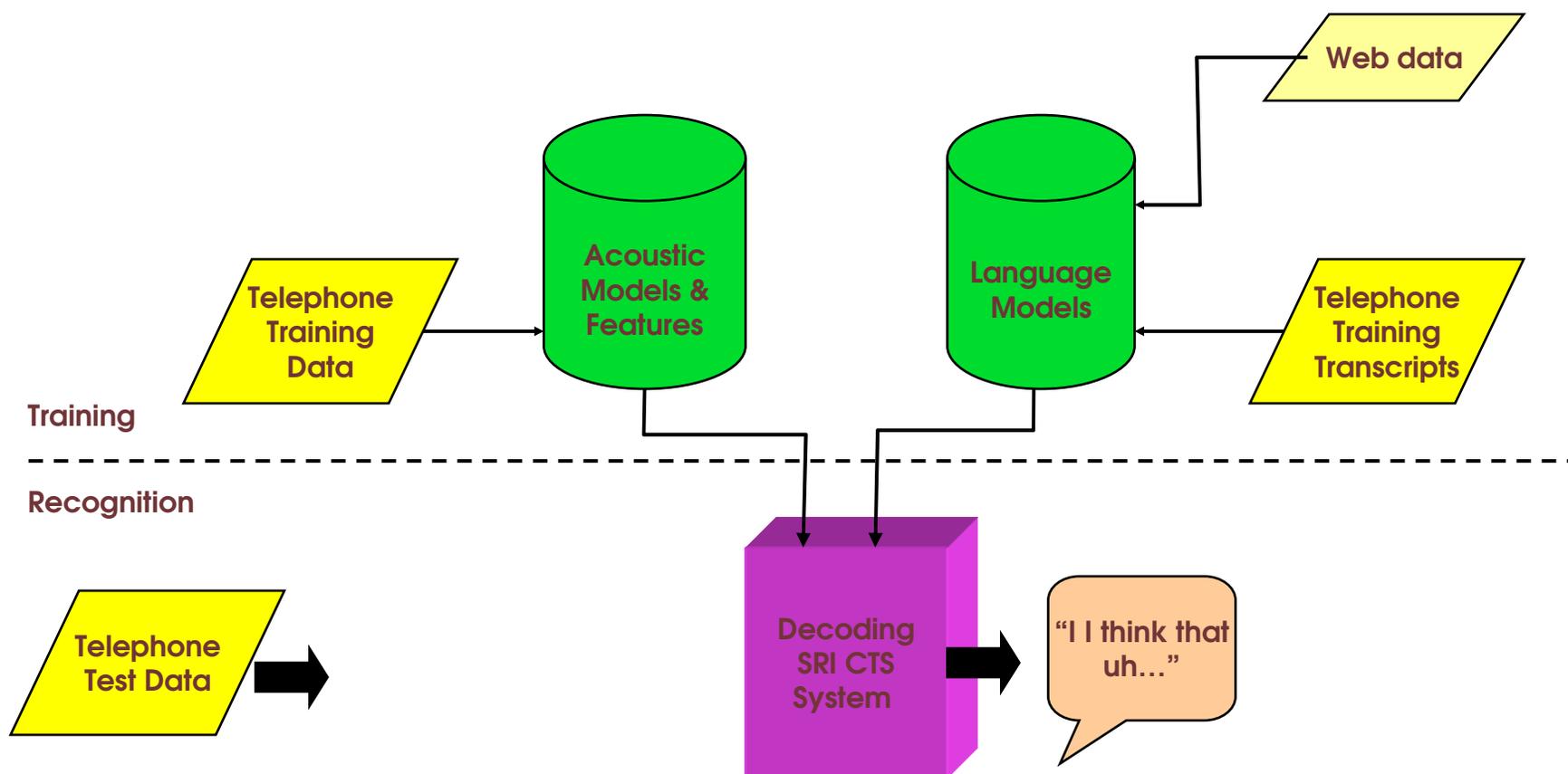
- **ADM** All distant microphones (i.e., table-top and array)
- **MSLA** Multiple source-localization arrays
- **MM3A** Multiple Mark III microphone arrays

Overlapping speech

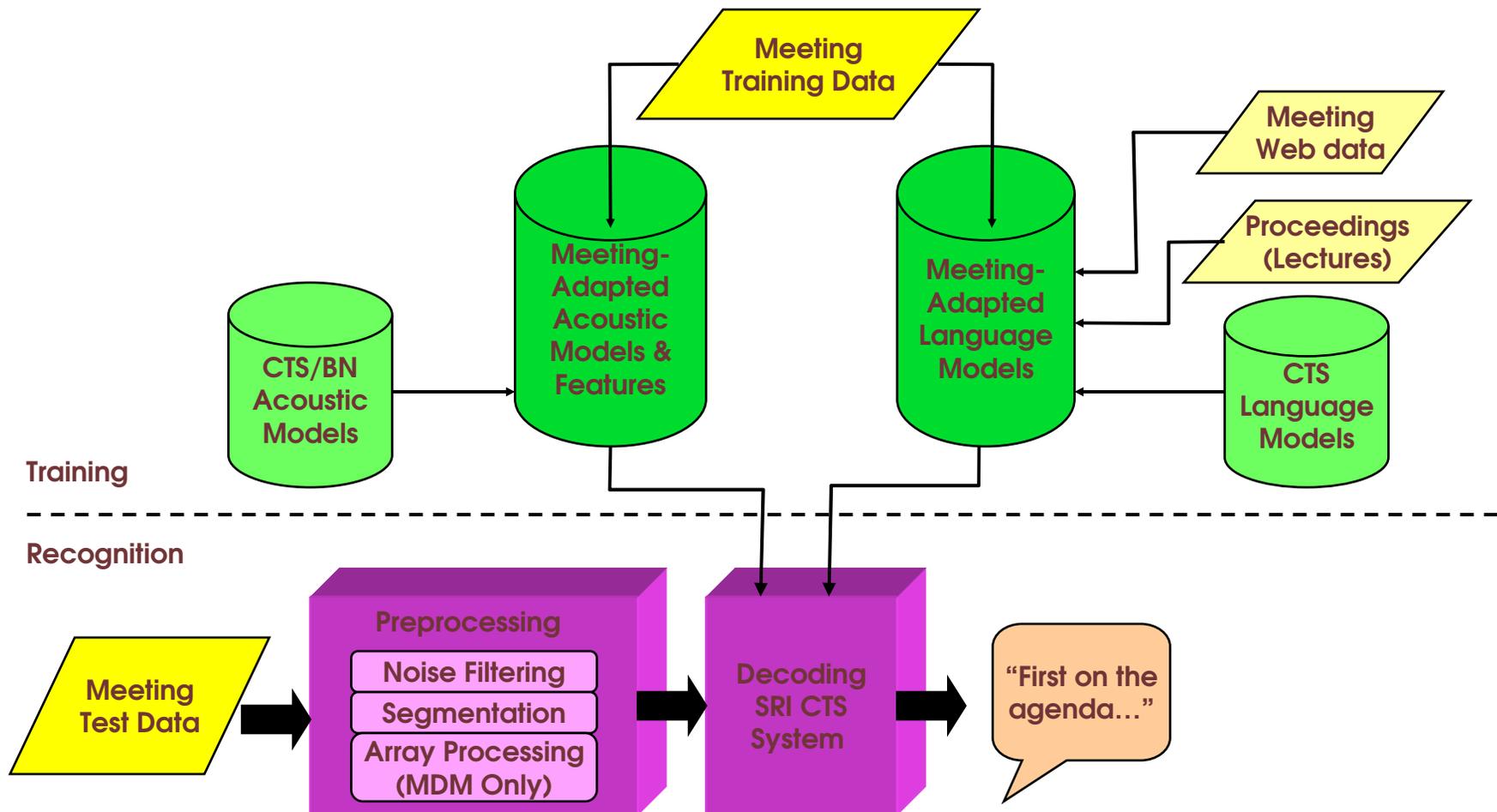
- Although the primary evaluation condition was for overlapped speech, **all results here are for one speaker only** (unless otherwise noted)
- No special processing was done to handle overlap



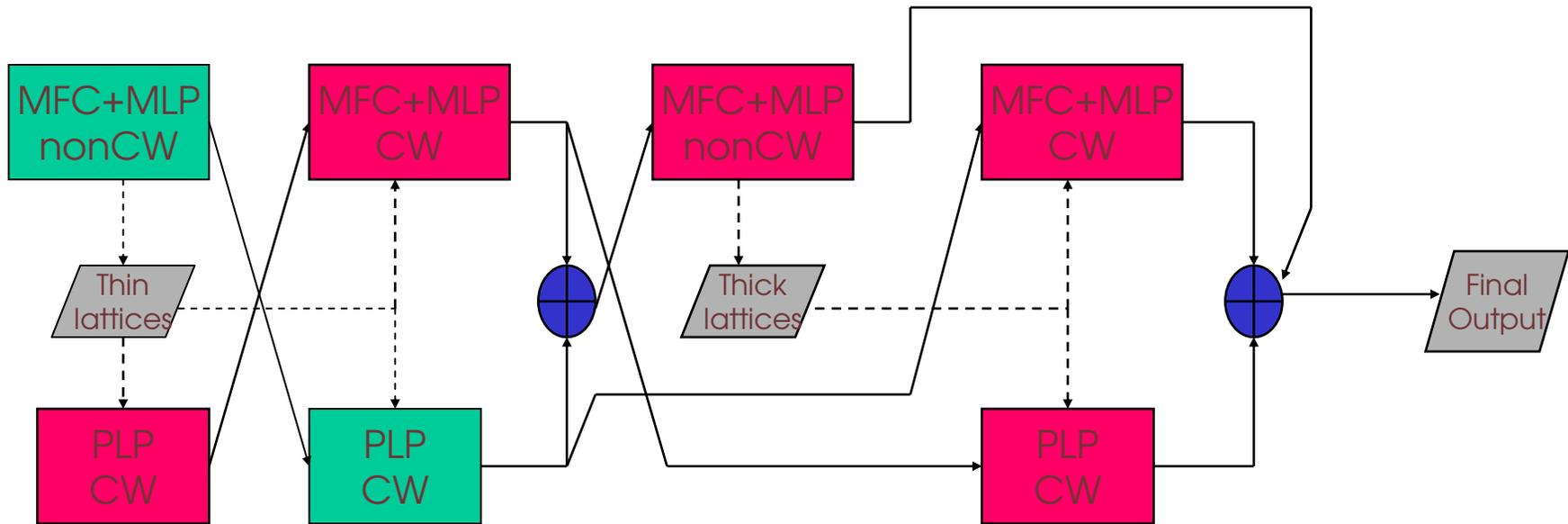
# Development Strategy: Base System



# Development Strategy: Meeting System



# SRI System Architecture



## Legend

-  Decoding/rescoring step
-  Hyps for MLLR or output
-  Lattice generation/use
-  Lattice or 1-best output
-  Conf. Network combination

Runtime: 12xRT (for CTS, Gaussian shortlists)  
25xRT (on meetings, no Gaussian shortlists)

# Preprocessing



# Acoustic Preprocessing

## Recognition

- Distant microphones (same as last two years)
  - Noise reduction using Wiener filtering on all input channels
  - Delay-sum beamforming of all channels, into single enhanced channel (MDM)
    - **Used cleaned-up code, released as BeamformIt-2.0 by Xavi Anguera**
  - Waveform segmentation (speech-nonspeech HMM decoding)
  - Segment clustering (for cepstral normalization, unsupervised adaptation)
- Close-talking (personal) microphones
  - No noise reduction
  - **Energy normalization in cross-channel feature computation**
  - Post-eval improvement by adjusting speech-nonspeech priors

## Training

- Distant microphones (same as last two years)
  - Eliminate overlapping speech (based on personal mic word alignment times)
  - Noise filtering
  - No delay-sum processing
  - Models trained on a selection of distant channel signals

# New vs. Old Beamforming

- Reprocessed eval06 data with BeamformIt-2.0 (Xavier Anguera)
- Results didn't change much for MDM
- Big gain in lecture ADM recognition
  - New code seems to be more robust to large numbers of and/or heterogeneous recording channels
- Also see talk on ICSI speaker diarization system

Testset	eval06 confmtg	eval06 lectmtg	
Condition	MDM	MDM	ADM
System	2006 system		
Old Beamform	34.2	55.5	51.0
New Beamform	33.9	55.8	46.6

# IHM Segmentation (1)

- RT-06 system: cross-channel energy features as an effective means to model cross-talk (Kofi Boakye)
  - Min and max log energy difference between target and all non-target channels
- Post-2006 eval: improved by normalizing channel energies subtracting noise floor (tuned on eval05)
- Most effective in conference meetings
- Not effective on eval07 meetings!
  - More investigation is needed

Testset	eval06		eval07		
	confmtg	lectmtg	confmtg	lectmtg	cbreak
Genre					
System	2006 system		2007 system		
Old segmenter	24.0	30.8	25.6	29.5	n/a
New segmenter	22.8	31.7	25.7	30.5	31.2
Reference seg.	20.2	29.3	22.8	28.1	29.5

## IHM Segmentation (2)

- Noticed large gap between automatic and reference segmentation
- AMI system segmentation was much better!
- Post-eval: tuned prior probabilities for speech/nonspeech (on eval06 confmtg data)
- Significant gains on all testsets, due to lower deletion rate

Testset	eval06	eval07		
		confmtg	lectmtg	cbreak
Genre	confmtg	confmtg	lectmtg	cbreak
System	2007 system			
Old priors	21.9	25.7	30.5	31.2
New priors	20.2	24.0	29.5	30.6
AMI segments		24.0		
Reference seg.	19.1	22.8	28.1	29.5

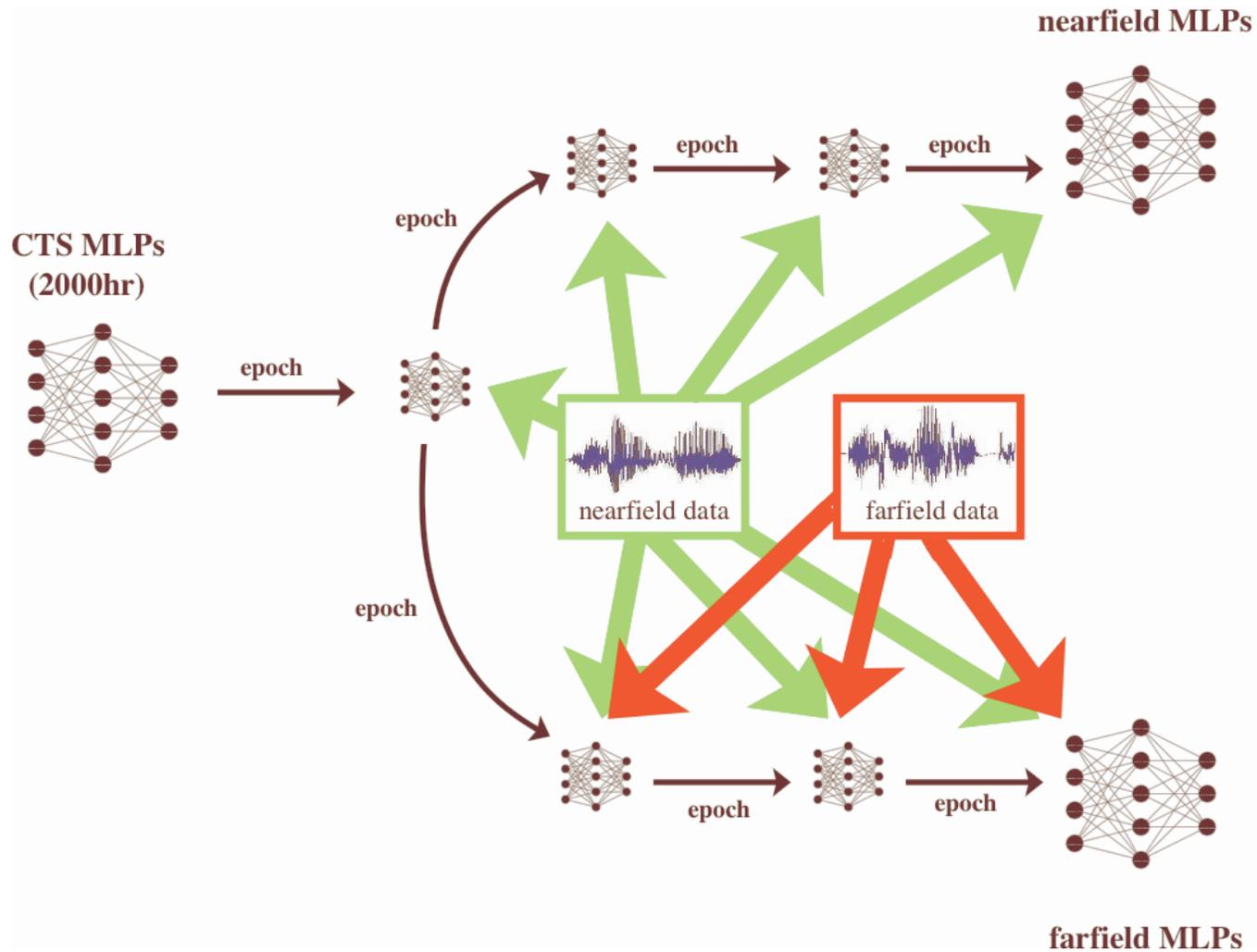
# Acoustic Modeling



# Acoustic Features and Models

- MFCC within- and crossword triphone models
  - Augmented with 2 x 5 voicing features (5 frames around current frame)
  - Augmented with 25-dim Tandem/HATS phone posterior features estimated by multilayer perceptron (MLP features)
  - Gender dependent
  - Base model trained on 1400h of conversational telephone data
  - **fMPE-MAP using meeting data**
  - **MPE-MAP** adapted to meeting data
- PLP crossword triphone models
  - Gender independent
  - Base models trained on 900h of Hub4 and TDT broadcast data
  - **fMPE-MAP using meeting data**
  - **MPE-MAP** adapted to meeting data
- Normalization and adaptation:
  - CMN + CVN, VTLN, HLDA
  - CMLLR (SAT) in training and test (except in first decoding)
  - MLLR with phone-loop in first MFCC and PLP decoding
  - MLLR cross-adaptation in subsequent steps

# Posterior Based Features



# Details on Posterior Features

- CTS MLPs trained on 8kHz data
- Tandem
  - 3-layer 9-frame PLP input
- HATs
  - 15 critical-band MLPs with 51 frame input
  - Merger net using hidden activations
  - Only merger net was adapted
- 4 epochs of adaptation
- Learning rate equal to the final learning rate of the CTS nets.
- Farfield adaptation only of non-overlap regions (alignments generated from near-field signal)
- Only one farfield channel (chosen at random) was used.
- Used MLPs from last year
  - Not yet retrained on additional acoustic confmtg training data
  - Not trained on any lecture data

# Use of CTS Training Data

- Is it still worth using background models trained on conversational telephone speech?
- CTS background model requires downsampling all meeting data
- Experiments using first pass of complete eval system, without MLP features
- Conclusion: still worth using CTS background data, especially since we have > 1000h of it.

	<b>eval05 confmtg</b>
Training data	IHM
Fisher 400h	34.0
Mtg 100h 8kHz	33.4
Mtg 100h 16kHz	31.7
Fisher + Mtg 8khz (pooled)	31.9
Fisher + Mtg 8khz (MAP-adapted)	31.5

# Model Adaptation for Nonnative English

- Observation: eval06 lectmtg data seemed dominated by non-American and nonnative speakers of English
- Idea: leverage non-American speakers in Fisher data
  - 1324 conversations, about 220h of speech
- Experiment: adapt Fisher CTS models to meeting and nonnative Fisher data (separately and jointly)
- ML-MAP models with MLP features

	eval06 lectmtg
MAP adaptation data	IHM
confmtg	41.9
Fisher Nonnat.	40.5
confmtg + Fisher Nonnat.	40.0

# fMPE-MAP

- MPE: minimum phone error training of Gaussians (Povey '02)
- fMPE: feature transform based on a sparse, high-dimensional Gaussian posterior vector, trained with MPE objective function (Povey '04)
- Using phone-frame error criterion (MPFE) in all training (Zheng, Eurospeech '05)
- Work on Mandarin (ICASSP'07) shows that MLP features, fMPE feature transforms, and MPE Gaussian training are partly additive
  - Best system combines all three discriminative modeling approaches
- Developed a MAP variant of fMPE that allows transforms to be trained on in-domain data only (Jing Zheng, submitted to Eurospeech'07)
  - Training bootstrapped off of same non-fMPE models as last year
- Our 2006 system used MMI-MAP and ML-MAP for adaptation

# fMPE-MAP Results

Adaptation method	eval06 IHM	
	confmtg	lectmtg
ML-MAP	22.8	34.1
MMI-MAP	n/a	29.8
fMPE-MAP	22.3	28.7
fMPE-MAP+MPE-MAP	22.2	26.3

	eval06 MDM	
	confmtg	lectmtg
ML-MAP	33.7	58.3
fMPE-MAP+MPE-MAP	30.9	48.6

# Adaptation for Lecture Distant Mics

- Goal: have same models for all genres
- Pool meeting and lecture training data
- But: this leads to suboptimal results for distant-mic lecture recognition
  - Maximal mismatch between training and test data
- Solution: perform extra ML-MAP on CHIL dev06 distant-mic lectures

Models	eval06 MDM
	lectmtg
Confmtg models	48.6
Confmtg models + MAP(lectmtg-dev06)	47.8

# Multiple Speaker Clusterings

- In 2005, found that speaker clustering in lectures hurts
- Is it still best to assume a single speaker per lecture?
- Yes, but ...
  - Both systems make significantly different errors, therefore
  - Significant gains from system combination
- For conference meetings, obtain similar gains by combining
  - One clustering with a fixed number (4) of pseudo-speaker clusters
  - Alternate, unlimited clustering with minimum amount of data per cluster

Clustering	eval06 MDM		eval07 MDM		
	confmtg	lectmtg	confmtg	lectmtg	cbreak
1 cluster		47.8		<b>44.6</b>	44.0
4 clusters	30.3		<b>26.2</b>		<b>44.7</b>
Unlimited	30.2	48.1	26.5	44.7	
Combined system	29.4	46.9	25.8	43.6	43.5

# Language Modeling



# Conference Meeting LMs

- Linearly interpolated mixture N-gram LMs
  - Different N-gram orders for different decoding stages
  - Perplexity optimized on held-out data (AMI, CMU, ICSI, NIST)
  - Final LMs entropy-pruned
  - Vocabulary: 54k words, OOV rate < 0.5%
- Conference meeting LM components
  - Switchboard + Fisher CTS (30M words)
  - Hub4 and TDT4 BN transcripts (140M)
  - AMI, CMU, ICSI, and NIST meeting transcripts **(2M) – about 2x data this year**
  - Web data selected to match Fisher (530M) and meeting (382M) transcripts

LM	eval06 confmtg			
	IHM		MDM	
	AMI	NonAMI	AMI	NonAMI
2006	20.1	23.2	28.9	32.9
2007	19.6	23.1	26.9	33.4

# Lecture Meeting LMs

- Similar to conference meeting LM, but
  - Added CHIL transcripts (70K words)
  - Speech conference proceedings (32M)
  - Removed Fisher web data
  - Collected web data based on CHIL transcripts (512M)
- Vocabulary: added 3781 words from conference proc.
  - OOV rate on CHIL devtest: 0.18%
  - Most common OOV word in CHIL:
- Perplexity optimized on a portion of the CHIL training data
- No 2006/2007 released lecture data used
- Addition of new AMI and NIST transcripts did not affect performance
- Reused 2006 lecture LM unchanged

# Evaluation Results



# Conference Meetings: Overall Results

- Relative WER reduction on eval06 data:
  - 11.4% for MDM
  - 8.8% for IHM
- Additional post-eval gain for IHM > 1%
  - Better tuning of speech/nonspeech priors
- Additional post-eval gain for MDM
  - Combine multiple speaker clusterings
- eval07 **less** difficult than eval06 for **MDM**
- eval07 **more** difficult for **IHM**

System	MDM	SDM	IHM
	eval06 confmtg		
RT-06S	34.2	41.2	24.0
RT-07S	30.3	40.6	21.9
Post-eval	29.4		20.2
	eval07 confmtg		
RT-07S	26.2	33.1	25.7
Post-eval	25.8		24.0

# Lecture Recognition: Overall Results

- Substantial improvements compared to 2006 system
  - 14% relative improvement on 2006 MDM eval data
  - 23% relative improvement on 2006 ADM data
  - 15% relative improvement on 2006 IHM data
- Post-eval improvements by combining two different speaker clusterings (MDM) and adjusting speech/nonspeech prior (IHM)

System	MDM	ADM	MM3A	SDM	IHM
	eval06 lectmtg				
RT-06S	55.5	51.0	56.5	57.3	31.0
RT-07S	47.8	39.3	-	49.6	26.3
Post-eval	46.9				25.7
	eval07 lectmtg				
RT-07S	44.6	42.1	54.0	50.6	30.5
Post-eval	43.6				29.5

# Coffee Break Recognition

- Due to lack of time, developed based on guesses about the domain
- Acoustic models: same as lectmtg system
  - Same recording setup as lectures
- Language model: same as lectmtg system
  - Based on cursory inspection of some 2007 dev data
- Speaker clustering: same as conference meeting system
  - Unlike lecture recognition, which assumes a single speaker
- Post-eval improvements by combining two different speaker clusterings (MDM) and adjusting speech/nonspeech prior (IHM)
- Word error results are comparable to lectmtg

	MDM	ADM	MM3A	SDM	IHM
	eval07 cbreak				
RT-07S	44.7	41.1	51.0	50.0	31.2
Post-eval	43.5				30.6

# Post-eval SASTT Submission

- After first results were published, we decided to submit a baseline SASTT system based on ICSI's diarization output
- Script merges STT CTM and SPKR RTTM output by assigning speaker label to each word
  - Chose longest overlapping speaker if speaker change falls within a word
- We had no diarization output for lectmtg
  - Hypothesized a single speaker
- Resulting system did respectably!

Task (overlap3)	eval07 confmtg		eval07 lectmtg		
	MDM	SDM	MDM	ADM	SDM
STT	37.4	43.6	49.3	47.5	54.8
SASTT	40.3	51.7	60.0	57.6	63.9

# Conclusions

- Post-eval improved IHM segmentation
  - Tuned speech/nonspeech priors
- Improved beamforming
  - Large gains for lectmtg ADM
- Acoustic model improvements
  - fMPE-MAP + MPE-MAP training
  - Addition of nonnative Fisher training data
  - Extra adaptation step for lectmtg distant mic recognition
  - Large gains on “difficult” data (lectures and distant mic)
- Language modeling
  - Additional AMI data helped, but only on AMI meetings
- Speaker clustering for lectures
  - Still doesn’t beat single speaker cluster
  - But gives substantial gains in combination

**Thank You!**

