



---

---

# Use of Model Transformation for Distributed Speech Recognition

Naveen Srinivasamurthy

Shrikanth Narayanan

Antonio Ortega

[\[snaveen,shri,ortega\]@sipi.usc.edu](mailto:[snaveen,shri,ortega]@sipi.usc.edu)

Integrated Media Systems Center (IMSC)

and Dept. of Electrical engineering

USC

# Distributed Speech Recognition



**Network**



## ● Client

- ◆ Speech acquisition
- ◆ Feature extraction
- ◆ Noise compensation
- ◆ Compression

## ● Server

- ◆ Decompression
- ◆ Channel compensation
- ◆ Speech recognition
- ◆ User interaction

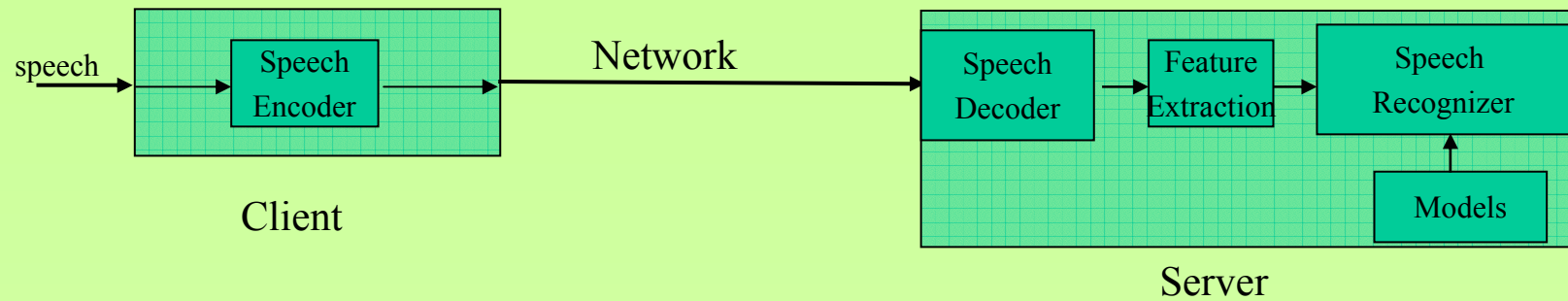


# DSR Possibilities

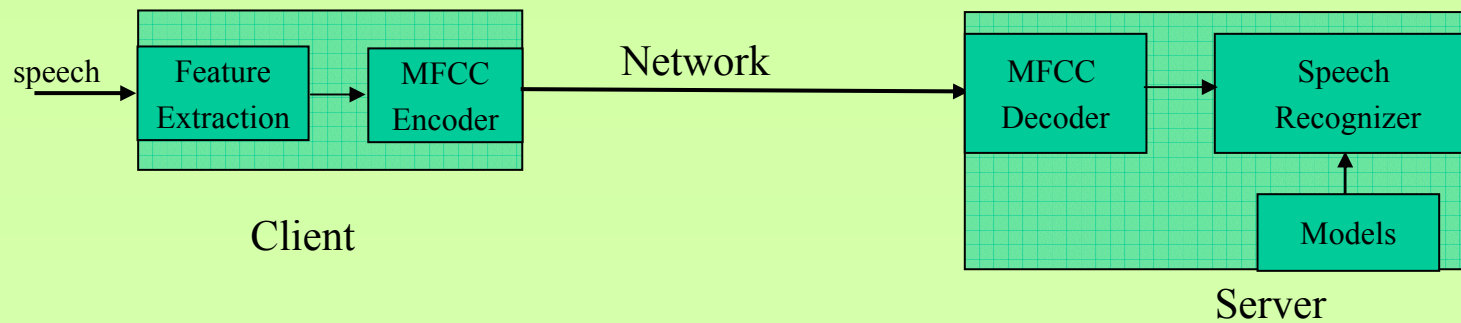
---

- Complex recognition tasks
  - ◆ Computation, Memory not an issue
- Improved robustness
  - ◆ Background noise estimate
  - ◆ Better quality of speech
    - ✗ Higher sampling rate
    - ✗ Better microphones
- Reduced bandwidth
  - ◆ Compression

# Distributed Speech Recognition



Standard speech encoder used at client



Feature extraction/compression at client



# Encoder Variability

---

- Computational resources/load
- Quality of Service
  - ◆ Bandwidth
  - ◆ Complexity
  - ◆ Delay
- Standard encoder
  - ◆ MELP
  - ◆ GSM
  - ◆ G.72x (LPC-based)
- Encoder optimized for recognition
  - ◆ Not (completely) standardized
    - ✗ ETSI : Aurora
- Scalable encoders
  - ◆ Scalable recognition



# Speech Recognition

---

- Diverse encoding schemes
  - ◆ Clients have freedom over compression scheme
- Matched training-testing
  - ◆ Train models for all different encoders
  - ◆ Number of combination large: not an efficient solution
  - ◆ Data collection complicated
- Train  $\Rightarrow$  clean speech, Test  $\Rightarrow$  encoded speech
  - ◆ Better models
  - ◆ Won't performance be affected ?
    - ✗ YES: mismatch

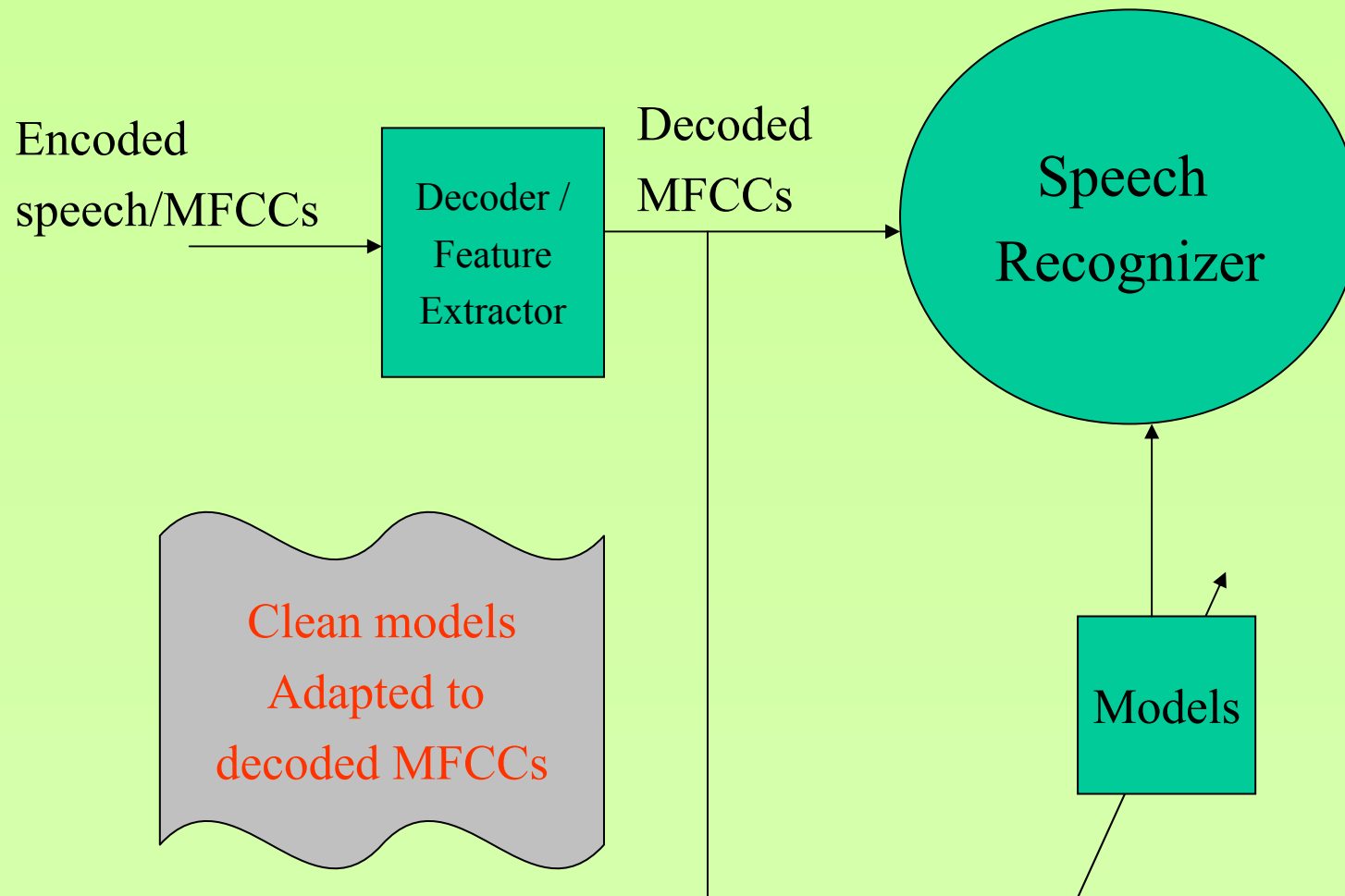


# Model Adaptation

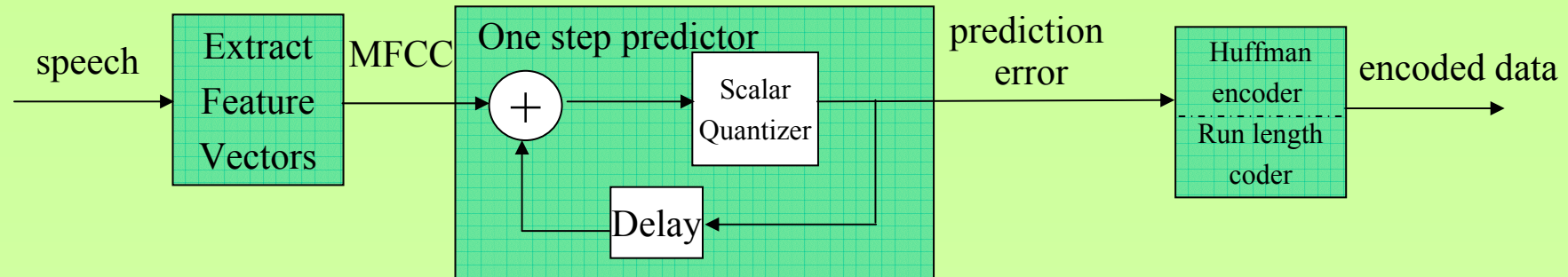
---

- Speaker/Channel/Speaking Environment adaptation
- Compensate for encoder variability
  - ◆ Adapt models: Clean to Encoded
    - ✗ Distortion is non-linear
      - Quantization distortion
      - Source-filter approximation
    - ✗ “Less” random
      - Encoding scheme is known
  - ✗ MLLR
  - ✗ MAP

# Model Adaptation



# MFCC Compression Algorithm



## Encoder: simple

- One step prediction used to predict every MFCC frame
- Quantizer
  - ✗ uniform scalar quantization (USQ)
  - ✗ entropy constrained scalar quantizer (ECSQ)
- Quantization indices corresponding to the non-zero coefficients encoded using Huffman entropy coder
- Bitmap used to indicate position of non-zero coefficients
  - ✗ Bitmap encoded using run length coding



# Experiments

---

## ● Speech Coders

- ◆ MELP
- ◆ GSM
- ◆ MFCC-Encoder

## ● TI Corpus Connected Digits (TIDIGITS)

- ◆ Strings of connected digits
- ◆ 1 to 7 digits in utterances

## ● Speech recognizer

- ◆ HMM: HTK 3.0
- ◆ Number of states per model : 10
- ◆ Mixtures
  - ✗ 4 per state
  - ✗ Diagonal covariance
- ◆ Features: MFCC 1-12



# MLLR Adaptation

Compression	Clean Models	Clean Models + MLLR	Matched Models	Matched Models + MLLR	MLLR gain
Clean speech	1.88 (7.56)	1.57 (6.57)	-	-	16.5%
MELP	3.14 (12.07)	2.32 (8.70)	2.70 (10.47)	1.87 (8.53)	26.1%
GSM	2.50 (8.76)	1.73 (7.33)	2.29 (8.61)	1.55 (6.91)	30.8%
MFCC-LR	4.81 (14.78)	2.24 (8.49)	2.70 (10.25)	1.85 (8.08)	53.4%
MFCC-HR	2.10 (8.06)	1.60 (6.82)	2.05 (7.87)	1.58 (6.87)	23.8%

WER and String error rates  
for the different coding schemes

Adaptation  $\approx$  clean  $\approx$  matched

MFCC-HR



# MLLR Adaptation

Compression	Clean Models	Clean Models + MLLR	Matched Models	Matched Models + MLLR
Clean speech	1.88 (7.56)	1.57 (6.57)	-	-
MELP	3.14 (12.07)	2.32 (8.70)	2.70 (10.47)	1.87 (8.53)
GSM	2.50 (8.76)	1.73 (7.33)	2.29 (8.61)	1.55 (6.91)
MFCC-LR	4.81 (14.78)	2.24 (8.49)	2.70 (10.25)	1.85 (8.08)
MFCC-HR	2.10 (8.06)	1.60 (6.82)	2.05 (7.87)	1.58 (6.87)
AURORA Clean	1.99 (8.68)	1.32 (6.39)	-	-
AURORA Encoded	2.30 (9.26)	1.52 (7.09)	2.38 (9.48)	1.51 (7.03)



# MAP Adaptation

Compression	Clean Models	MAP	MAP gain
Clean speech	1.86 (7.54)	0.67 (3.85)	64.0%
MELP	3.12 (12.05)	1.19 (6.06)	61.9%
GSM	2.48 (8.72)	0.91 (4.09)	63.3%
MFCC-LR	4.78 (14.73)	3.34 (10.89)	30.1%
MFCC-HR	2.08 (8.01)	0.91 (4.36)	56.3%

WER and String error rates  
for the different coding schemes



# Adaptation Gain

Compression	Degradation before Adaptation	Degradation after Adaptation	Rate (kbps)
MELP	67.02	47.77	2.4
GSM	32.98	10.19	13
MFCC-LR	155.85	42.68	1.22
MFCC-HR	11.70	1.91	2.07
AURORA	15.58	15.15	3.6

Degradation due to coding before and after adaptation  
Baseline: train/test with clean speech

Adaptation results closer to baseline

Aurora bitrate when only MFCC 1-12 are used with no channel coding



# Adaptation Gain

Compression	Degradation before Adaptation	Degradation after Adaptation
MELP	16.30	24.06
GSM	9.17	11.61
MFCC-LR	78.15	21.08
MFCC-HR	2.38	1.27

Degradation from matched conditions before and after adaptation

Adaptation helps MFCC encoders



# Related Work

---

## ● Scalable encoder and scalable recognition

(Srinivasamurthy, Ortega, Narayanan Eurospeech 2001)

- ◆ First stage recognizer uses coarse layer to find likely models
- ◆ Second stage uses fine layer to refine recognition
- ◆ Rate-Recognition performance trade-off
- ◆ Complexity-Recognition performance trade-off
- ◆ Rate-Delay trade-off



# Conclusions/Future Work

---

- Model adaptation compensates for encoder variability
- Re-synthesis
  - ◆ Send more information (pitch)
  - ◆ Transcription required for tuning
  - ◆ Joint compression-classification-playback framework
- Compression noise is not random
  - ◆ Can we modify adaptation techniques?
    - ✗ MFCC encoders: output is discrete
    - ✗ Speech encoders: voiced/unvoiced regions