Automatic Speech Recognition for Wireless Mobile Devices

(무선 모바일 단말에서의 음성인식 시스템)

2005. 04. 22

김홍국

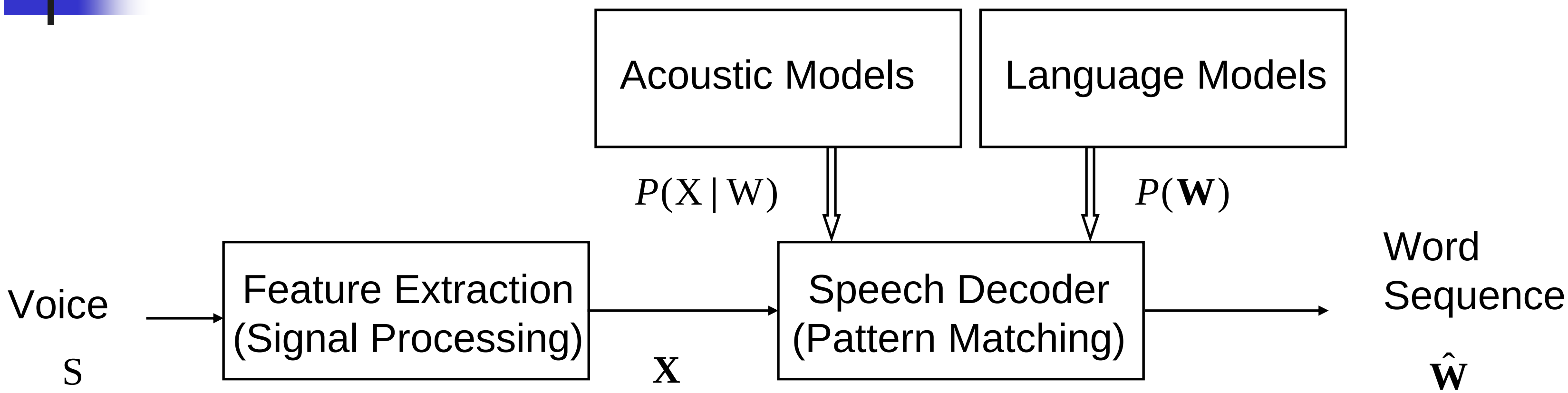
광주과학기술원, 정보통신공학과

http://salc.gist.ac.kr
E-mail: hongkook@gist.ac.kr

Contents

- Overview of ASR Systems
  - Functional description of ASR
- ASR for Mobile Applications
  - Mobile applications
  - ASR Architecture for wireless mobile devices
  - ASR in client-server scenarios
- Robust ASR in Client-Server Scenarios
- Ubiquitous ASR
- Summary, Discussion, and Q&A

Overview of ASR Systems



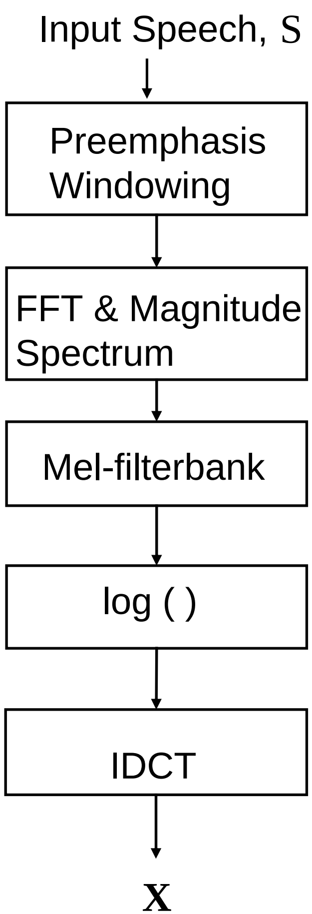
- For a given acoustic observation $\mathrm{X} = X_1, X_2, \cdots, X_T$, to find out the corresponding word sequence $\hat{\mathbf{W}} = w_1, w_2, \cdots, w_m$ such that

$$\hat{\mathbf{W}} = \arg\max_{\mathbf{W}} P(\mathbf{W}\,|\,\mathbf{X}) = \arg\max_{\mathbf{W}} \frac{P(\mathbf{X}\,|\,\mathbf{W})P(\mathbf{W})}{P(\mathbf{X})}$$

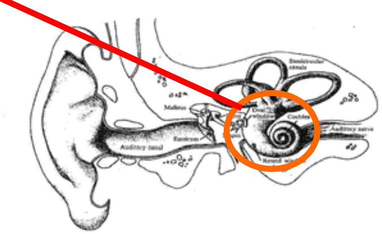
Functional Components of ASR

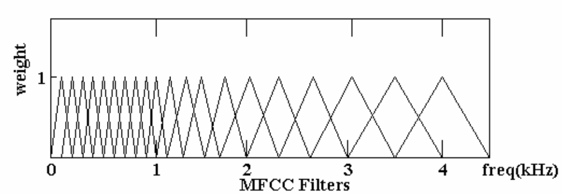
- Feature extraction
- Acoustic model
- Language model
- Search/ Decoding (Pattern Matching)
- Robust ASR: Compensation, Adaptation
- Performance

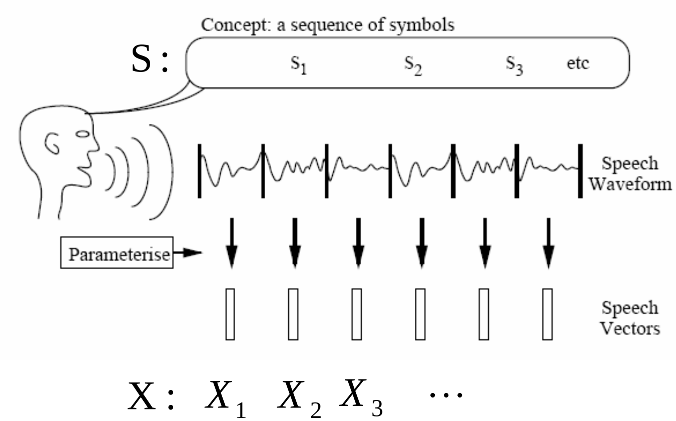
Feature Extraction



- Goal
  - To extract salient features that are useful for acoustic matching
- Typical Setting
  - 20~30ms Hamming window, frame rate = 100Hz
  - MFCCs (Mel-Frequency Cepstral Coefficients): 39-dim
  - CMN, Energy Normalization



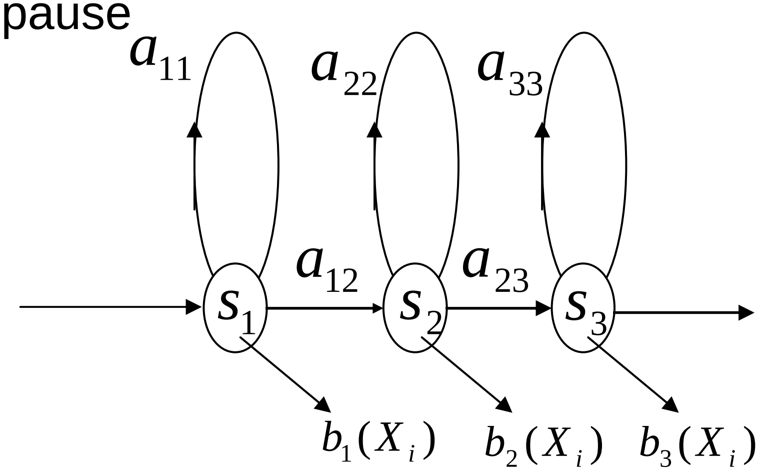




SALC Speech, Audio and Language Comm. Lab.

Acoustic Model (1)

- Goal
  - To provide $P(\mathbf{X}\,|\,\mathbf{W})$, which includes the representation of knowledge about acoustics, phonetics, and variations regarding to speaker, microphone, environments, etc.
- Typical Acoustic Model
  - Continuous-density Hidden Markov Model (HMM): $\lambda = (A, B, \pi)$
  - Distribution: Gaussian Mixture
  - HMM Topology: 3-state left-to-right model for each phone
    1-state for silence or pause

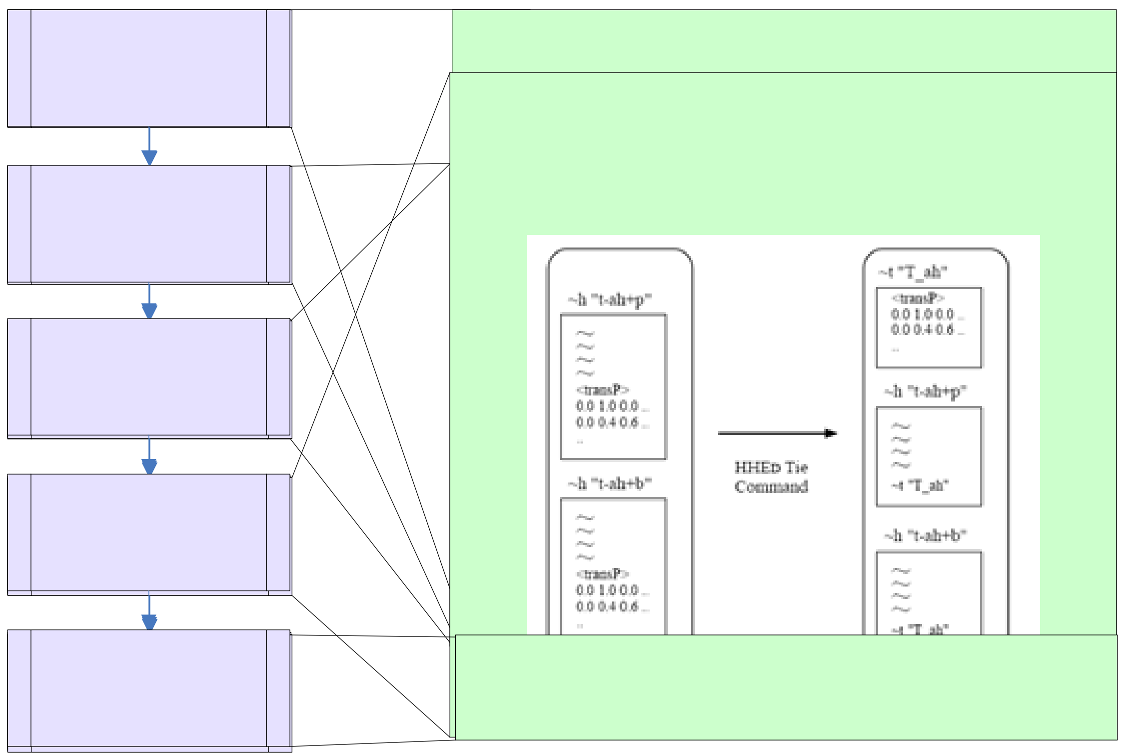


Acoustic Model (2) - example

- Monophones (41)
  - 15 vowels
    - AH,EY,EH,AE,OW,AW,IY,UW,AO,AY,ER,AA,OY,UH,IH
  - 24 consonants
    - R,N,B,D,NG,T,M,S,HH,V,L,SH,Z,JH,W,G,P,Y,K,CH,F,TH,DH,ZH
  - 2 silence models
    - sil (silence), sp (short pause)
- 8360 Cross-Word Triphones, 23085 states
  - Eg. This → sil-th+ih th-ih+s ih-s+sp
  - Clustered by a decision tree
  - Reduced to 5356 states

Acoustic Model (3)

- Training Paradigm (Based on Triphone Models)



Language Model (1)

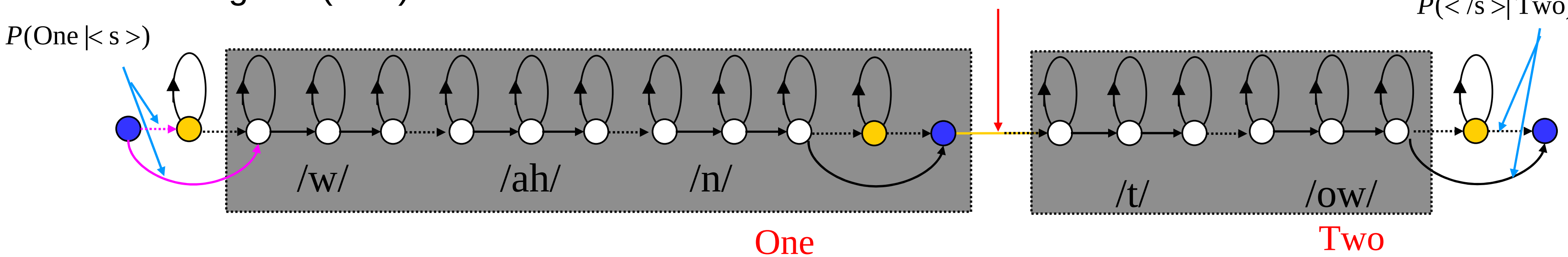
- Goal: Language model reflects how frequently a string **W** occurs as a sentence
  - Built from a training corpus
- N-gram

$$P(\mathrm{W}) = P(w_1, w_2, \cdots, w_n)$$
$$= P(w_1)P(w_2 \mid w_1)P(w_3 \mid w_1, w_2)\cdots P(w_n \mid w_1, w_2, \cdots, w_{n-1})$$
$$= \prod_{i=1}^{n} P(w_i \mid w_1, w_2, \cdots, w_{i-1})$$

  - In practice,

$$P(w_i \mid w_1, w_2, \cdots, w_{i-1}) \cong P(w_i \mid w_{i-N+1}, w_{i-N+2}, \cdots, w_{i-1})$$

Bigram (N=2)



SALC Speech, Audio and Language Comm. Lab.

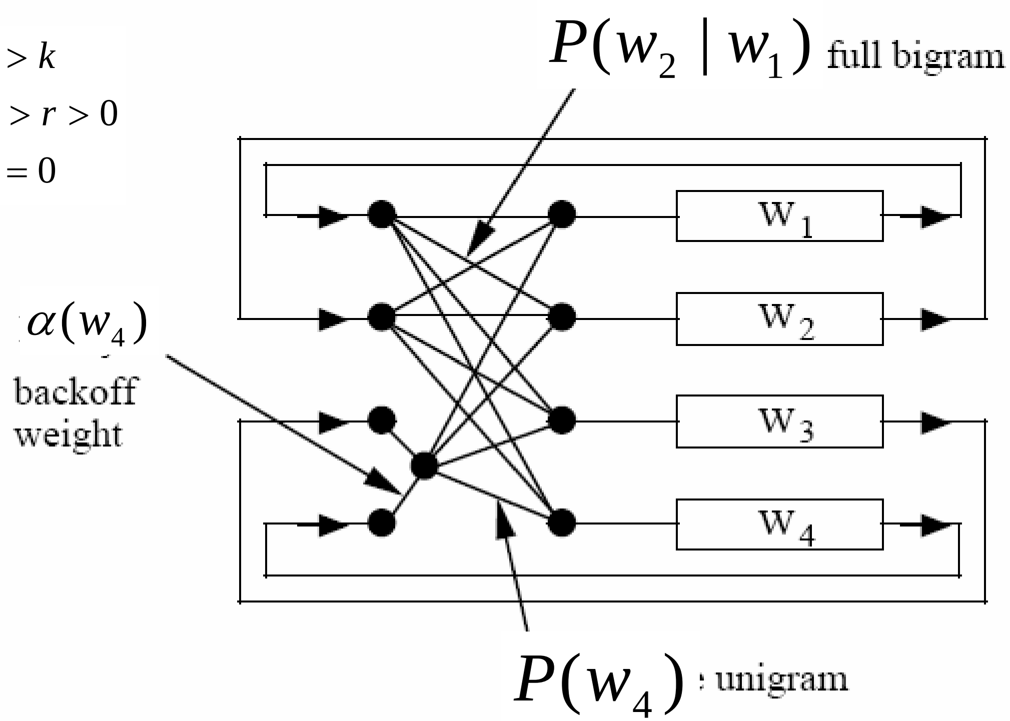
Language Model (2)

- Word network based on a back-off bigram
  - Reduce the network size
  - Speed up the decoding
  - Smooth the n-gram due to data sparseness in a real training set

$$P_{Katz}(w_i \mid w_{i-1}) = \begin{cases} C(w_{i-1}w_i)/C(w_{i-1}) & \text{if } r > k \\ d_r C(w_{i-1}w_i)/C(w_{i-1}) & \text{if } k > r > 0 \\ \alpha(w_{i-1})P(w_i) & \text{if } r = 0 \end{cases}$$

$$d_r = \frac{\frac{r^*}{r} - \frac{(k+1)n_{k+1}}{n_1}}{1 - \frac{(k+1)n_{k+1}}{n_1}}, r^* = (r+1)\frac{n_{r+1}}{n_r},$$

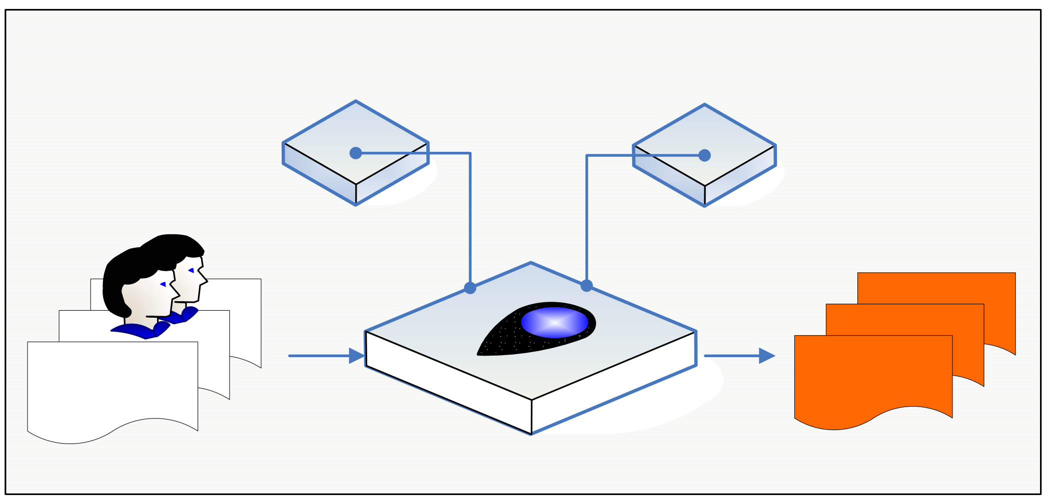
$$\alpha(w_{i-1}) = \frac{1 - \sum_{w_i:r>0} P_{Katz}(w_i \mid w_{i-1})}{1 - \sum_{w_i:r>0} P_{Katz}(w_i)}$$



Speech Decoding (1)

- Goal
  - To find a sequence of words whose corresponding acoustic and language models bet match the input signal

$$\hat{W} = \arg\max_{W} P(W \mid X) = \arg\max_{W} \frac{P(X \mid W)P(W)}{P(X)}$$



SALC Speech, Audio and Language Comm. Lab.

Speech Decoding (2)

- Viterbi Search
  - Find state sequences that maximizes $P(S, \mathrm{X} \mid \mathrm{W})$

Step 1 : Intialization

$$V_1(i) = \pi_i b_i(X_1), \quad 1 \le i \le N$$

$$B_1(i) = 0$$

$$\max_{1 \le i \le N} V_T(i) = P(\mathrm{X}, \mathrm{S}^* \mid W) \cong P(X \mid W)$$

Step 2 : Induction

$$V_t(j) = \max_{1 \le i \le N} \left[V_{t-1}(i) a_{ij} \right] b_j(X_t), \quad 1 \le j \le N, 2 \le t \le T$$

$$B_t(j) = \arg\max_{1 \le i \le N} \left[V_{t-1}(i) a_{ij} \right]$$

Step 3 : Ternimation

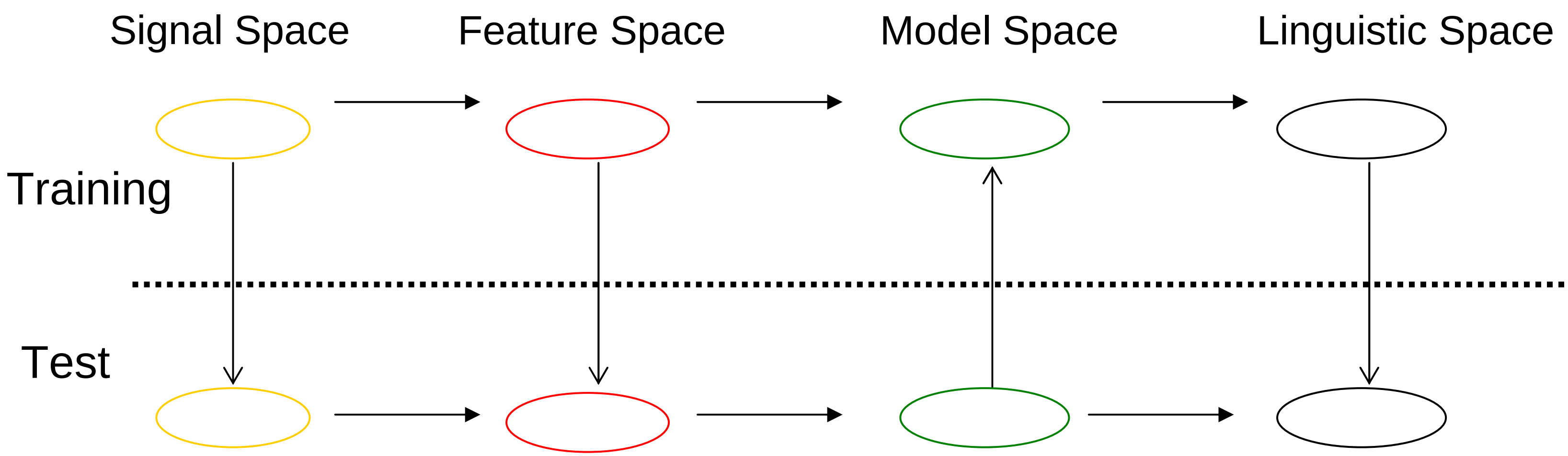
$$\text{Best Score} = \max_{1 \le i \le N} V_T(i)$$

$$\text{Best state at T} : s_T^* = \arg\max_{1 \le i \le N} B_T(i)$$

Step 4 : Backtracking

$$s_t^* = B_{t+1}(s_{t+1}^*) \qquad t = T-1, \cdots, 1$$

Adaptation



Signal Enhancement
- Spectral subtraction
- Wiener, MMSE
- Signal subspace

Feature Compensation
- Spectral subtraction
- Spectral mapping
- CMN
- RASTA

Model Compensation
- Parallel model combination (PMC)
- Stochastic matching
- State-based Wiener filtering

Language Adaptation
- Active learning

Current ASR Performance

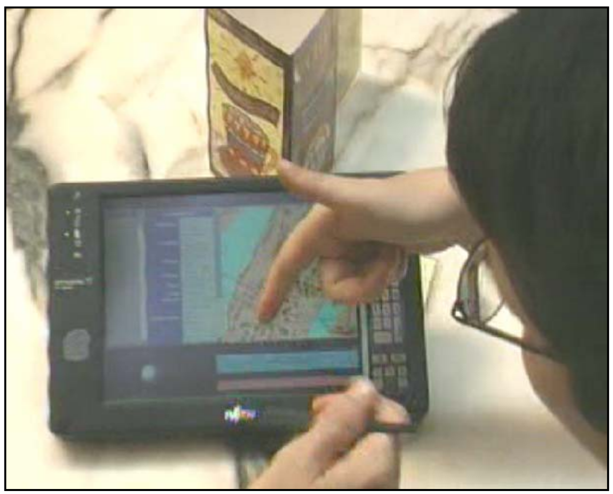
Corpus (DARPA)	Type of Speech	Vocabulary Size	Word Error Rate (%)
Connected Digit String	**Spontaneous**	**10**	**0.3**
Airline Travel Information	**Spontaneous**	**2,500**	**2.5**
Wall Street Journal	**Read Speech**	**64,000**	**6.6**
Switchboard	**Conversational Telephone**	**28,000**	**37**
Call Home	**Conversational Telephone**	**28,000**	**40**
SPINE	**Conversational Speech in Noise**	**2,000**	**20**

Contents

- Overview of ASR Systems
  - Functional description of ASR
- ASR for Mobile Applications
  - Mobile applications
  - ASR Architecture for wireless mobile devices
  - ASR in client-server scenarios
- Robust ASR in Client-Server Scenarios
- Ubiquitous ASR
- Summary, Discussion, and Q&A

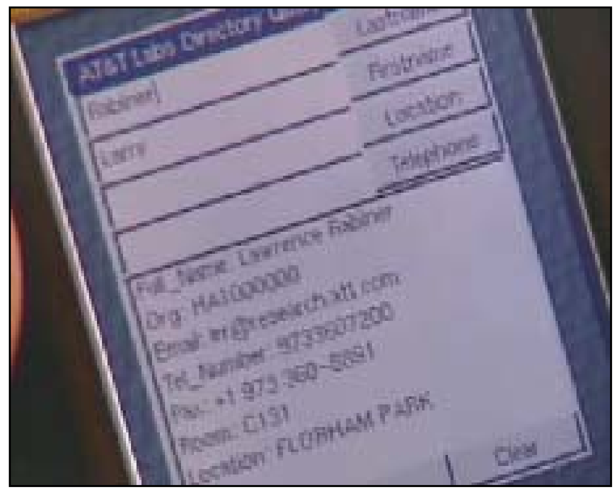
Mobile Applications

Multimodal Dialog
(Database query)



- **MATCH: Multimodel access to city help** [Johnston et al,2002]
  - **Fujitsu Tablet PC**

Voice Form Filling
(Directory retrieval)



- **Multimodal Directory Retrieval** [Rose and Parthasarthy,2001]
  - **Compaq Ipaq PDA**

Command
and Control

- **Name/Digit Dialing**
  - **Motorola M70 Digital Cellular Phone**
- **Speech-to-SMS**
  - **Samsung P207** **[CES'05]**

Practical Issues in Mobile Applications

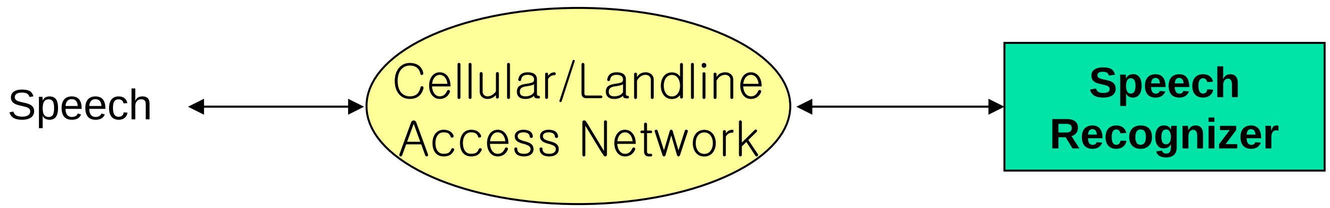
- Memory requirement: acoustic model, language model, lexicon
- Search speed: network size
- ASR accuracy
- Display, Battery & GUI

ASR Architecture

- ASR on Mobile Devices
  - Client-based scenarios, Embedded implementation
- ASR over Current Existing Networks
  - Server-based scenarios
- ASR over Data Networks
  - Client/server-based scenarios
  - Distributed speech recognition (DSR)

ASR over Current Existing Networks (1)

- ## ASR over Voice Networks

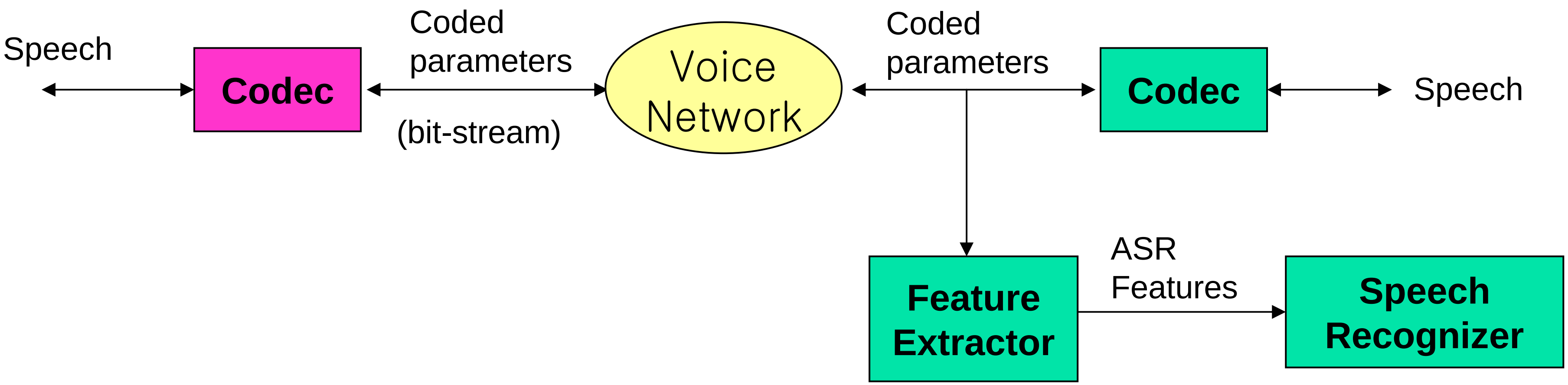


- **Features**
  - ASR service is available from any telephone
  - ASR resides entirely on the server
- **Issues**
  - Speech signal is subject to distortions, degrades ASR performance
    - Channel – especially cellular
    - Background noise at the source and its effect on speech codecs
    - Source coding – data rates higher than 8 kbps does not degrade ASR performance
  - User interface is constrained
    - Speech only input, Speech only output

ASR over Current Existing Networks (2)

[H.K. Kim et al, 2001]

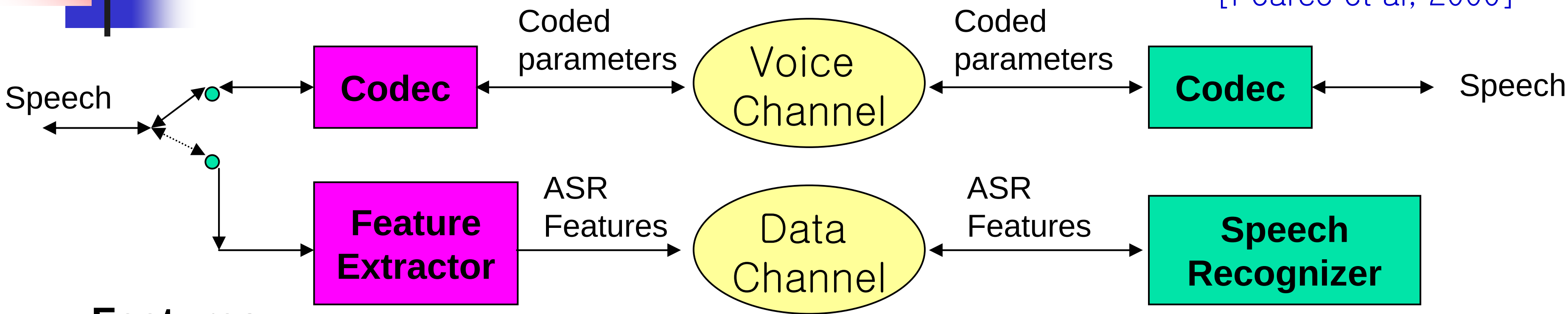
- ASR over Wireless Networks



- **Features**
  - Not susceptible to reconstruction losses
  - Codec parameters can be combined to improve ASR performance
  - No changes required in the handsets or transmission protocols
- **Issues**
  - The bit-stream has to be made available to the recognizer
    - What if there are multiple carriers ?
    - The coded parameters chosen for telephony is not optimized for ASR

ASR over Data Networks

[Pearce et al, 2000]



- **Features**
  - Improved speech recognition performance
    - Minimizes impact of channel errors: error protected data channel
    - Eliminates mismatch due to different speech codecs in use [Besacier, 2001]
  - Enables integration of speech input and data output on devices with displays
  - Potential for improved feature extraction from wideband speech or from multiple microphones
- **Issues**
  - Handsets have to incorporate feature extraction
  - Protocols need to be established to select between modes: speech codec or ASR features

Comparison of Architectures

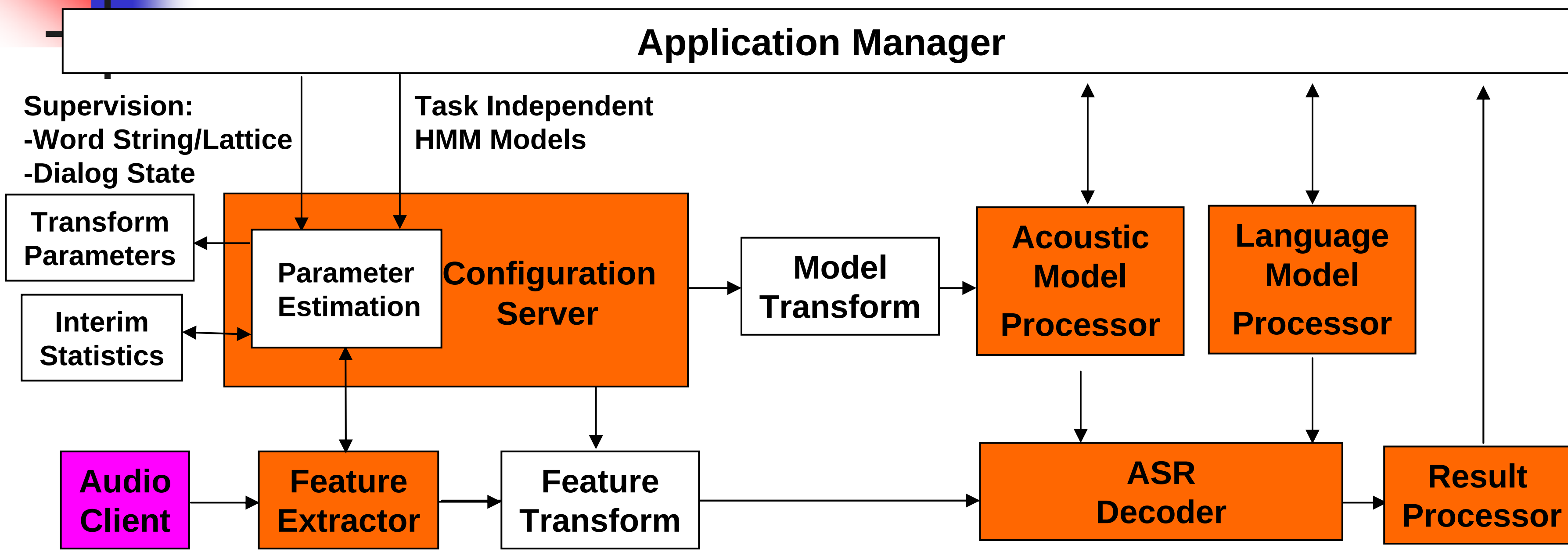
	Client (Embedded)	Server (Bit-stream)	Client-Server (DSR)
Computation Load in Devices	High	None	Low (same to speech coding, ~17WMOPS)
Network Protocol	None[1)]	None	Required
Robustness to Channel	Don't worry	Weak	Strong
Vocabulary Size	Small-medium	No restriction	No restriction
Performance Degradation Caused by	Restricted Computation	Channel & Speech Coding Distortion	Parameter quantization Channel distortion

1) It is required depending on the applications

ASR in Client-Server Scenarios

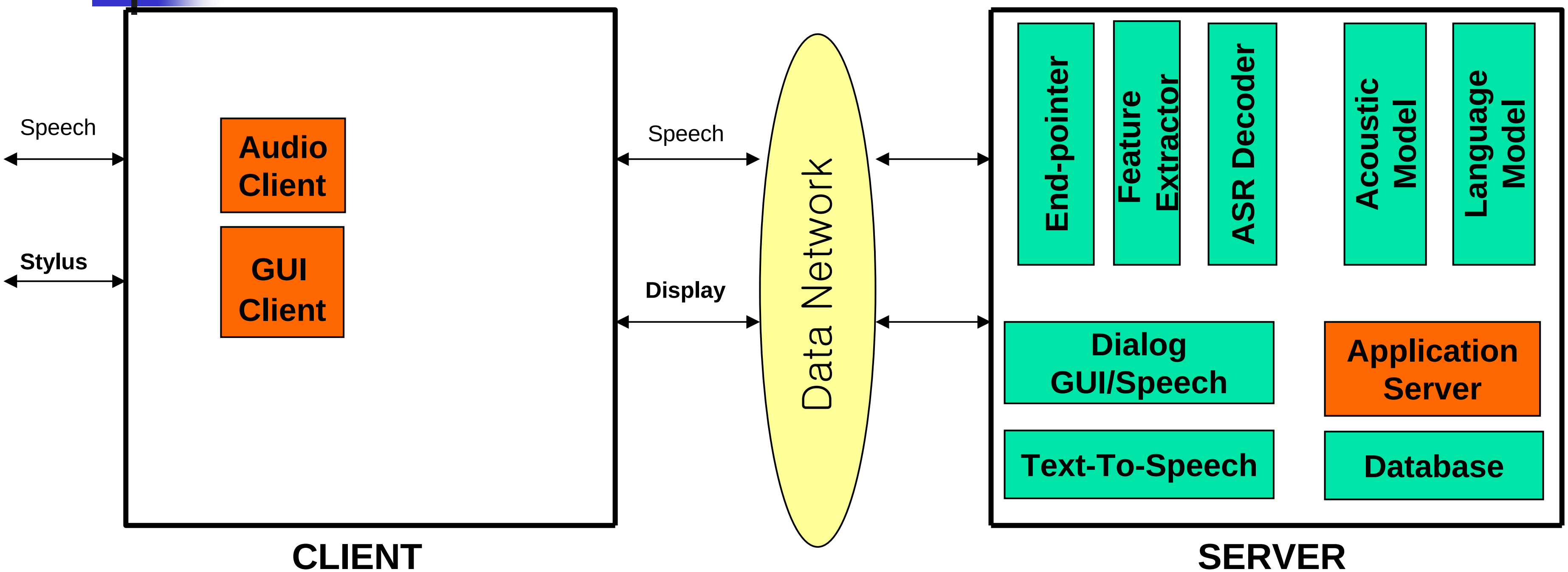
- Thin client
  - Client resources limited
  - Reliable high speed data network
  - All ASR functions on the server
- Moderate client resources, limited bandwidth network
  - Migrate some ASR functionality to the device
- Fat client
  - ASR resides mostly on client

Example Implementation – Configuration Server



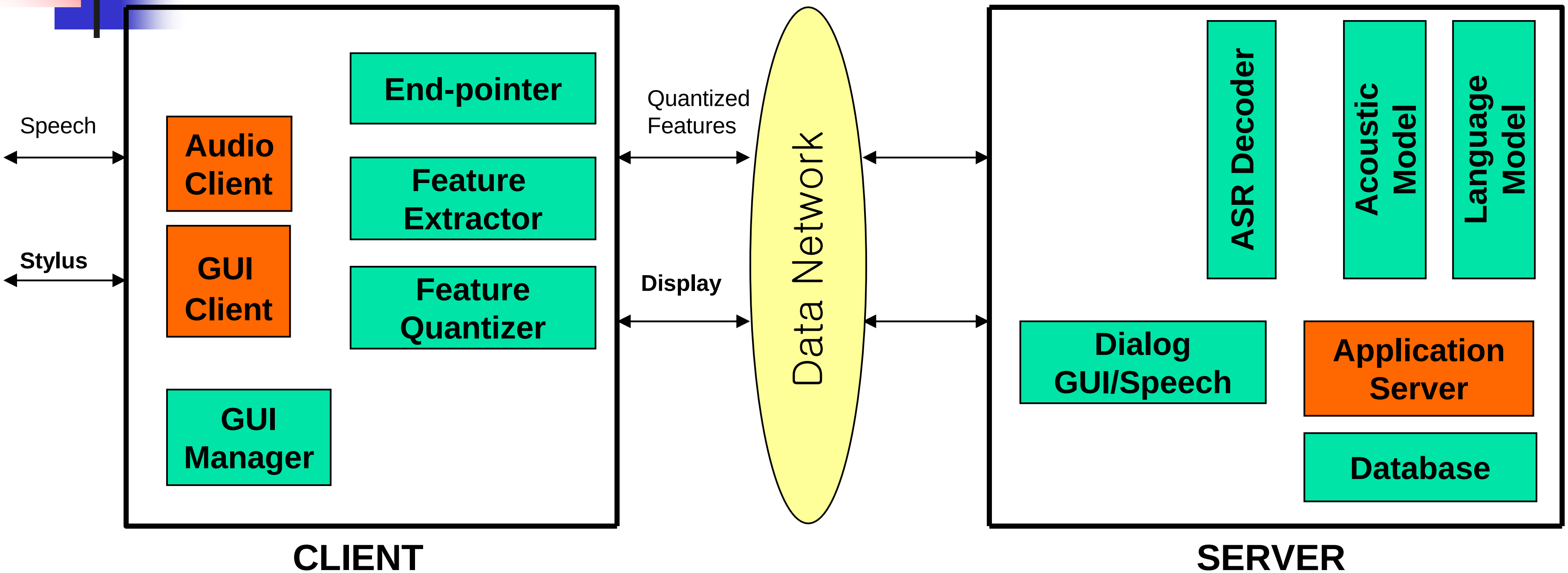
- Configuration Server
  - Continuous unsupervised estimation of model/feature space transformations
- Computational requirements
  - Estimation of parametric transformations
  - Application of transformation during recognition
- User specific storage requirements
  - Parametric model/feature space transformations and interim statistics for estimating transformation parameters
  - Dialog state and task completion status obtained from application provide supervision for parameter estimation algorithms

Thin Client



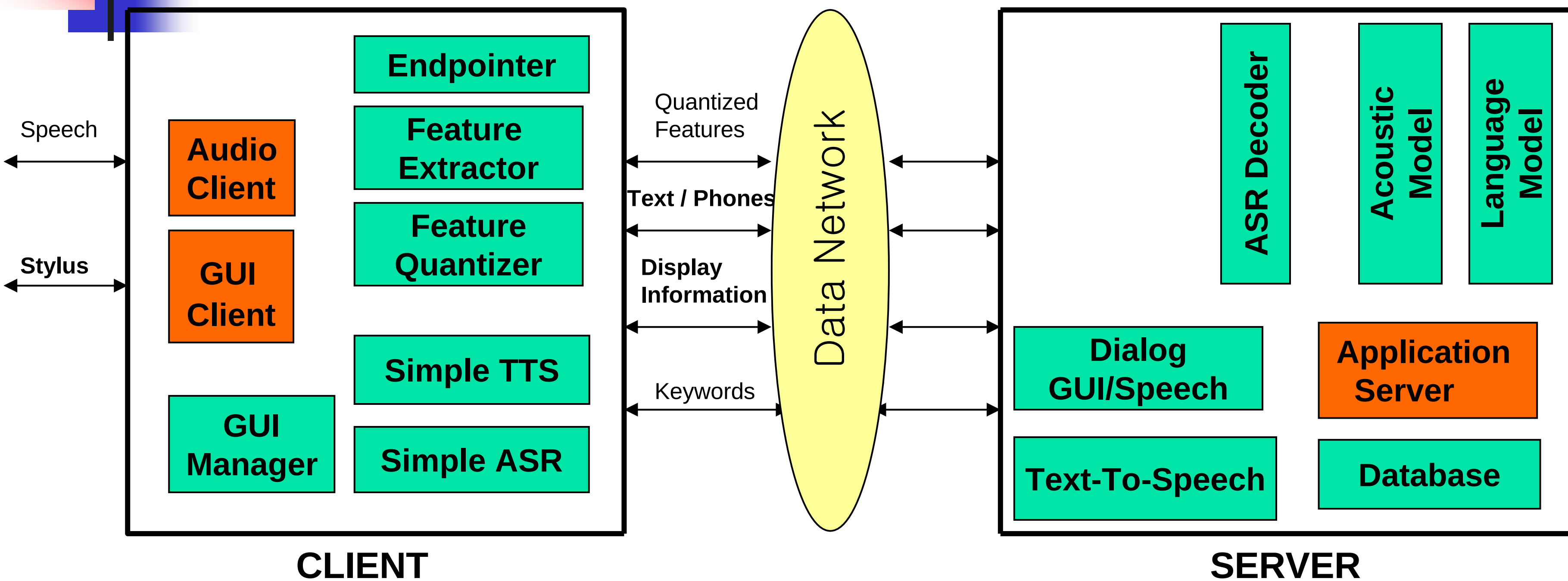
- Client is simply an input-output device
- ASR functions are tightly coupled; changes in functional blocks are transparent to the client
- Network delays will degrade the user interface unless carefully designed

Moderate Client



- Migrating the front-end to the client allows
  - transmission of ASR features at a low bit-rate: 4.8Kbps in the case of Aurora
  - Potential for signal processing: noise-cancellation; multiple microphones; wideband features
- A suitable client-server protocol allows
  - programmable features: parameterized and controlled by server; downloadable code
- Information for display can be transmitted in compressed form over the network and rendered for display by the GUI manager

Fat Client



- Client has sufficient computing resources to run a complete recognizer
- All recognition done on the client: connect to the server only for access to central databases
- ASR for only terminal functions only on the client: command and control, name-dialing, local address book
- Phone recognizer: phone strings or lattices transmitted to the server; domain knowledge not available to the client

Contents

- Overview of ASR Systems
  - Functional description of ASR
- ASR for Mobile Applications
  - Mobile applications
  - ASR Architecture for wireless mobile devices
  - ASR in client-server scenarios
- Robust ASR in Client-Server Scenarios
- Ubiquitous ASR
- Summary, Discussion, and Q&A

Robust ASR in Client-Server Scenarios

[Rose and Kim, 2004]

- Robustness Issues for Mobile Domains
  - Feature extraction scenarios
  - Robustness with respect to channel distortions
  - Acoustic and channel variability: Importance of acoustic environment

- Adaptation / Normalization Algorithms
  - Implementation of adaptation algorithms
  - Parameter adaptation architecture
  - Example

Robustness Issues for Mobile Domains (1)

- Feature extraction scenarios
  - **Client-Server based:** ETSI Aurora DSR Standard transmitted over (protected) data channel [Pearce et al, 2000]
  - **Server-based:** Deriving features from the transmitted voice channel bit stream [H.K. Kim et al, 2001]
  - **Client-based:** Relying on existing Adaptive Multi-rate speech coder to trade-off voice source coding and channel coding bit assignment [Kiss et al, 2003]
- Robustness with respect to channel distortions
  - **Modified ASR Decoder** [Potamianos et al, 2001]
    - Similar to missing feature theory
    - Derive confidence measures from channel decoder
    - Use confidence to weight or censor Gaussian computation in the Viterbi algorithm
  - **Channel Error Concealment Strategies:**
    - Frame interpolation and frame repetition [Milner, 2001]
    - Sub-vector based concealment

Robustness Issues for Mobile Domains (2)

- Acoustic and Channel Variability: Importance of acoustic environment [Sukkar et al, 2002]
  - GSM channel was shown to roughly double word error rate (WER) relative to landline channel
  - Noisy car environment increased WER by nearly an order of magnitude relative to landline channel

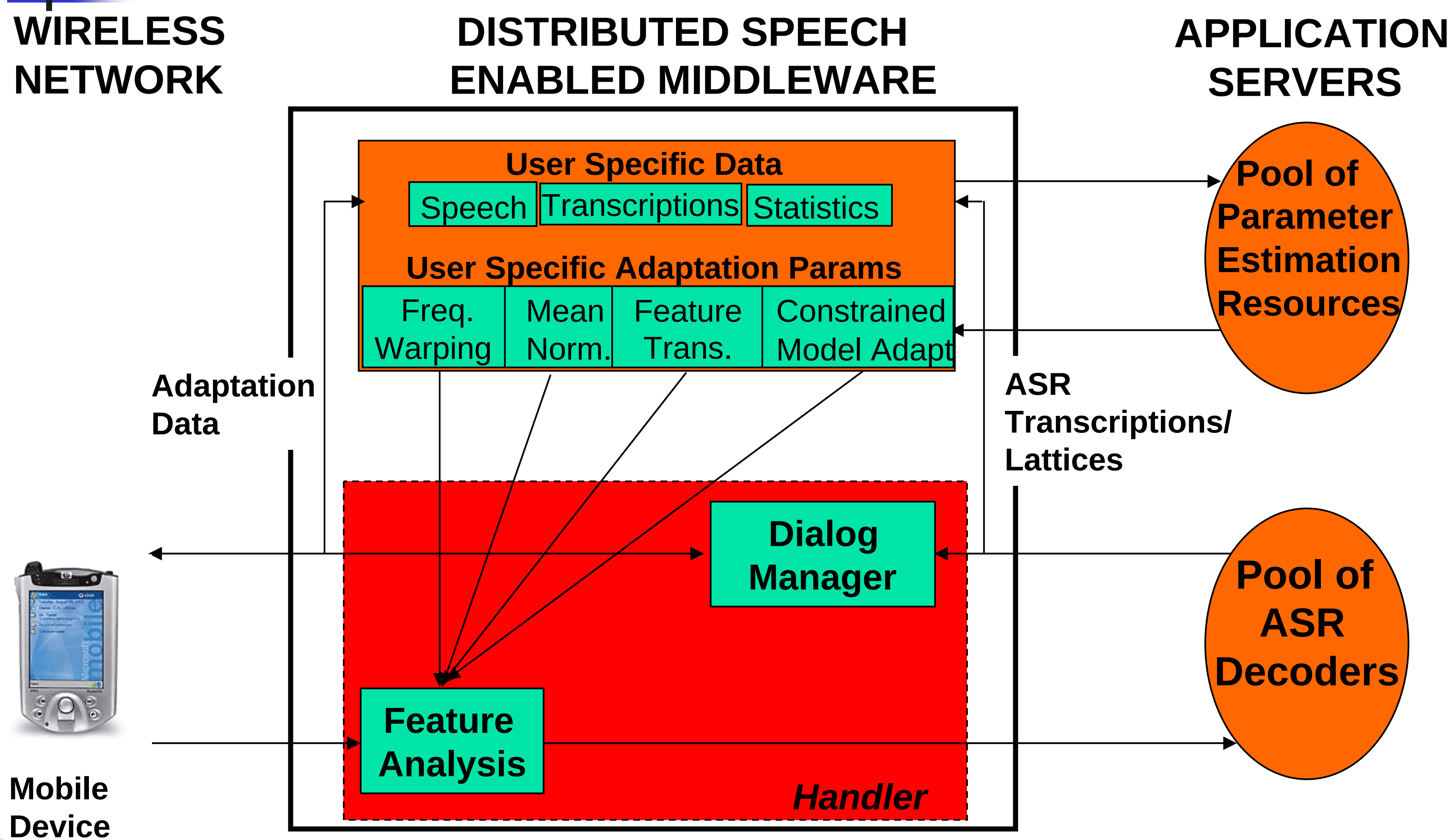
Wireless Standard	Environment	Landline HMM Models (WAC)
Landline	Home/Office	96%
GSM	Home/Office	91%
GSM	Traffic	70%

* Word Accuracy (WAC) measured on SpeechDat Car connected digit corpus (Sukkar et al, 2002)

Adaptation/ Normalization Algorithms

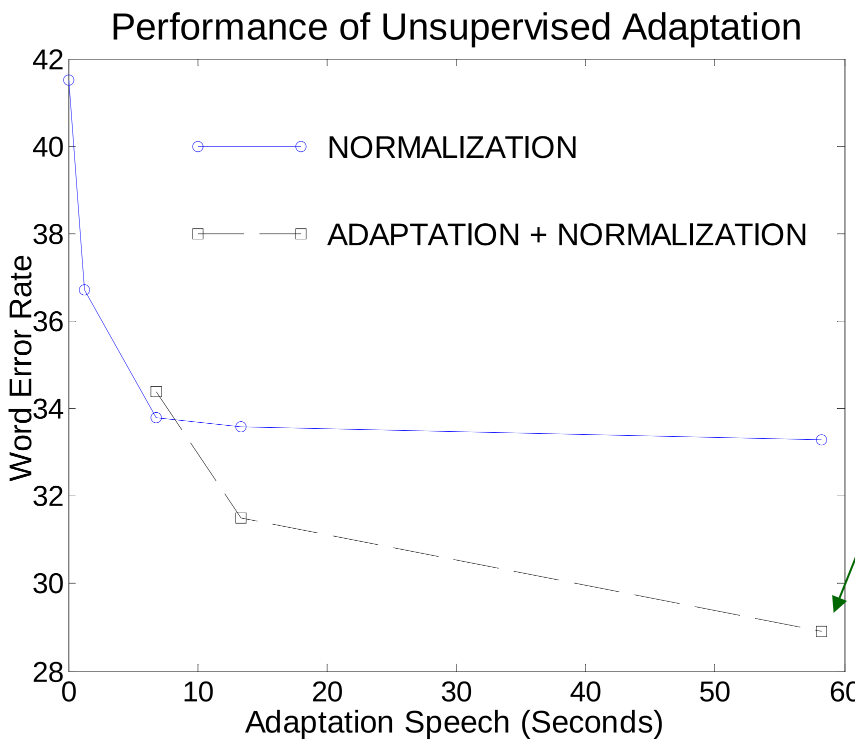
- **Mobile Applications**: Imply wider variety of acoustic environments than wire-line telephone or desk-top applications
- **Personalized Devices**: Implies that representations of speaker, environment, and transducer variability can be acquired through normal use of the device
- **Issues for Robust Modeling in Mobile Domains:**
  - Continual, incremental update of transformation and normalization parameters
  - Need for efficient storage of transformation parameters and interim statistics

Adaptation / Normalization Algorithms – Parameter Adaptation Architecture



Adaptation/Normalization Implementation

- 3000 word name recognition task where users interact with a handheld device with a device mounted far-field microphone
- Normalization: CMN and frequency warping based spkr. norm.
- Adaptation: Unsupervised Constrained Model Adapt.
- WER reduction is highly dependent on adaptation data



[Rose et al, 2003]

31% reduction in word error rate

Ubiquitous ASR

[Tan et. al, 2005]

- Ubiquitous networking and context-aware computing
  - Computing Context
    - Network and terminal capabilities
    - Different scenarios for mobile devices
  - User Context
    - User's profile and location
    - Personalization of acoustic model and language model
    - Personalization/context-aware for spoken language processing
  - Physical Context
    - Environmental awareness

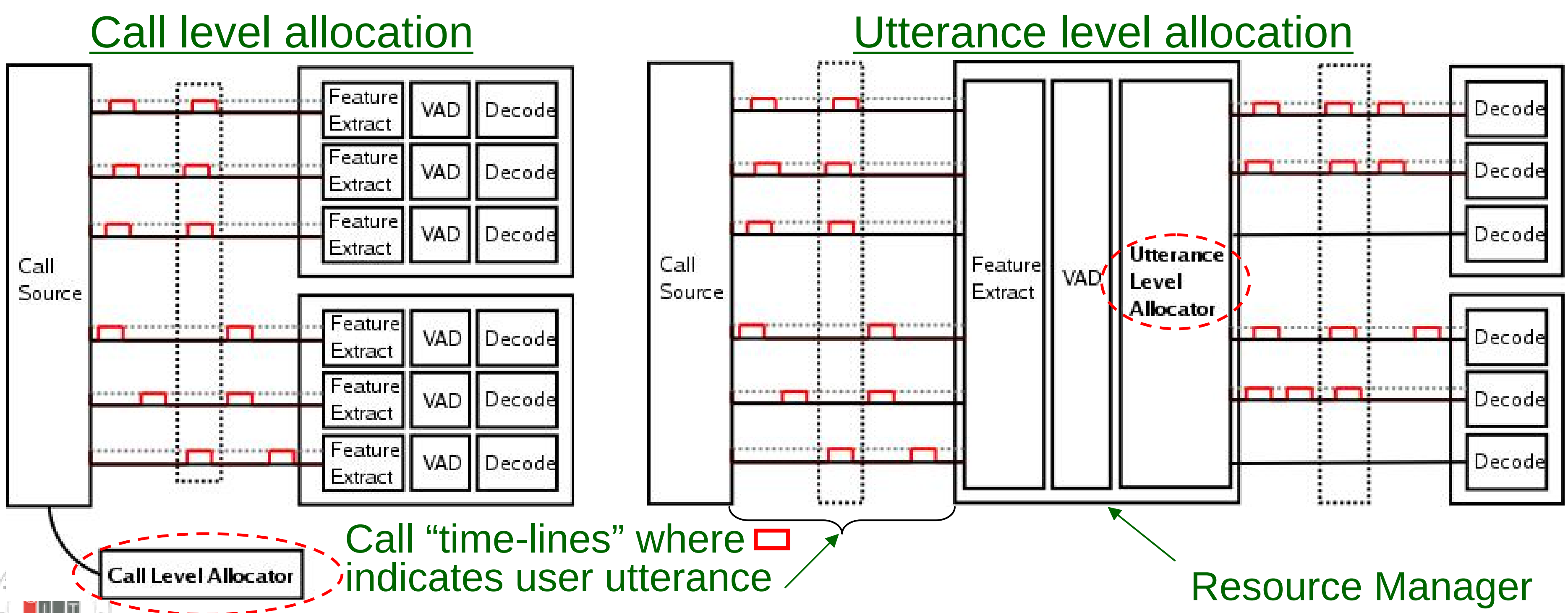
Efficient Use of ASR Resources in Multi-User Scenarios

[Rose and Kim, 2004]

- **Allocation of ASR Servers:** High degree of variability in processing effort for ASR in human-machine dialog scenarios
  - Infrequent occurrence of user utterances in dialog
  - Variability in processing load across different ASR tasks
- **Greater Efficiency:** ASR Server Allocation strategies
  - Call Level Allocation (CLA): Allocate ASR decoder to a computation server for an entire Dialog (Call)
  - Utterance Level Allocation (ULA): Allocate ASR decoder to a server for each utterance

ASR Resource Allocation Strategies

- Strategies for allocating ASR decoders to computation servers
  - Goal: To minimize probability of server overload at peak load times
  - ULA: Rely on a Resource Manager to assign utterances to servers
- Example: Six incoming calls to multi-user configuration with two servers with capacity of two utterances per server

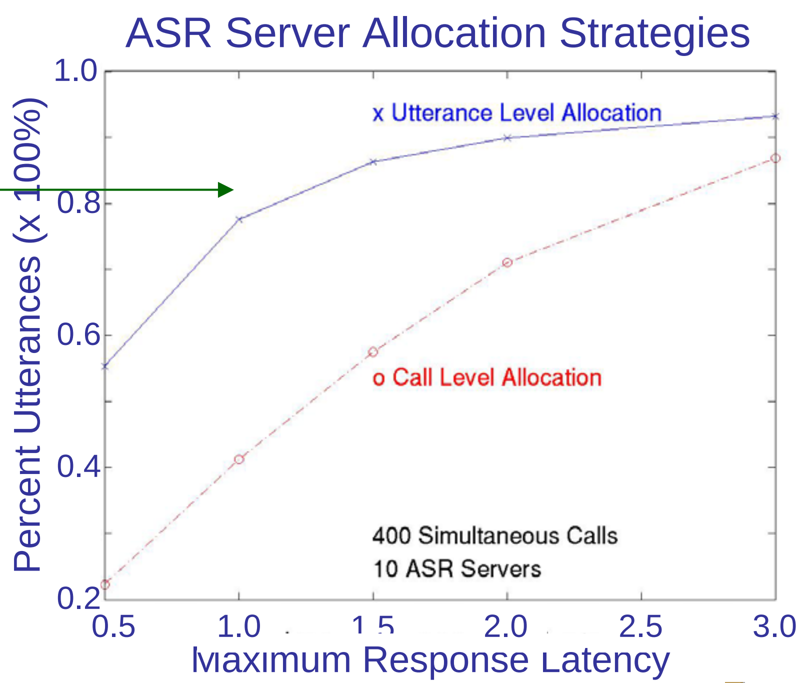


SALC Speech, Audio and Language Comm. Lab.

Performance of Efficient Allocation Strategies

- Compare ASR response latencies observed by users for CLA and ULA strategies:
  - 400 calls of natural language queries presented to a multi-user system
  - User utterances active for approximately 35% of call duration
  - Multi-user system: Ten Linux servers – Running at 1GHz and able to process two simultaneous utterances without overload

ULA strategy is able to handle twice as many calls with sub-second response latencies



SALC Speech, Audio and Language Comm. Lab.

Summary

- ASR Scenarios over Network Environments for Mobile Devices
  - Different configuration of ASR functions according to the resources in devices
  - Client-Server scenarios are more flexible
- Research problems in client-server ASR scenarios
  - The effect of communications channels on ASR performance
  - Implementation of speaker / environment normalization/adaptation algorithms
  - Efficient ASR implementations in multi-user scenarios
- Evolving to Ubiquitous ASR
  - Distributed speech recognition
  - Personalized/ context-aware approach for functional components of ASR
  - Standardization of architecture and scenarios

References

- **R. C. Rose and S. Parthasarthy, a tutorial on "ASR for Wireless Mobile Devices," ICSLP, 2002.**
- **X. Huang, A. Acero, and H.-W. Hon, Spoken language processing: a guide to theory, algorithm, and system development, Prentice Hall, 2001.**
- **S. Young, et al, "The HTK Book (for HTK Version 3.2)," MS, Cambridge Univ., Dec. 2002.**

- R. C. Rose and H. K. Kim, "Robust speech recognition in client-server scenarios," in *Proc. ICSLP*, Jeju, Korea, pp. 2321-2324, Oct. 2004.
- H. K. Kim and R. V. Cox, "A bitstream-based front-end for wireless speech recognition on IS-136 communications system," *IEEE Trans. SAP*, vol. 9, no. 5, pp. 558-568, July 2001.
- D. Pearce and H. Hirsch, "The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Proc. ICSLP*, Beijing, China, Oct. 2000.
- R. C. Rose, S. Parthasarathy, B. Gajic, B, A. E. Rosenberg, and S. Narayanan, "On the Implementation of ASR Algorithms for Hand-Held Wireless Mobile Devices," in *Proc. ICASSP*, May 2001.
- R.C. Rose, I. Arizmendi, and S. Parthasarathy, "An efficient framework for robust mobile speech recognition," in *Proc. ICASSP*, Apr. 2003.
- Z.-H. Tan, P. Dalsgaard, and B. Lindberg, Speech recognition in ubiquitous networking and context-aware computing, Special Session in Eurospeech 2005.

- R. A. Sukkar, R. Chengalvarayan, and J. J. Jacob, "Unified speech recognition for landline and wireless environments." in *Proc .ICASSP*, May, 293-296, 2002.
- L. Besacier, C. Bergamini, D. Vaufreydaz, and E. Castelli, "The effect of speech and audio compression on speech recognition performance," in *Proc. IEEE Multimedia Signal Processing Workshop*. 2001.
- M. Johnston, S. Bangalore, G. Vasireddy, A. Stent, P. Ehlen, M. Walker, S. Whittaker, and P. Maloor, "MATCH: An architecture for multimodal dialog systems," in *Proc. 40th Anniv. Mtg. of Assoc. for Computational Linguistics*, June 2002.
- A. Potamianos and V. Weeackody, "Soft-feature decoding for speech recognition over wireless environments," in *Proc. ICASSP*, May 2002.