



ulm university

universität  
**uulm**

# **Incorporating Knowledge into Statistical Acoustic Models for Spoken Language Dialog Systems**

**Dissertation**

**Sakriani Watiasri Sakti**

**Prof. Dr. Dr.-Ing. W. Minker**  
**Department of Information Technology**  
**University of Ulm, Germany**

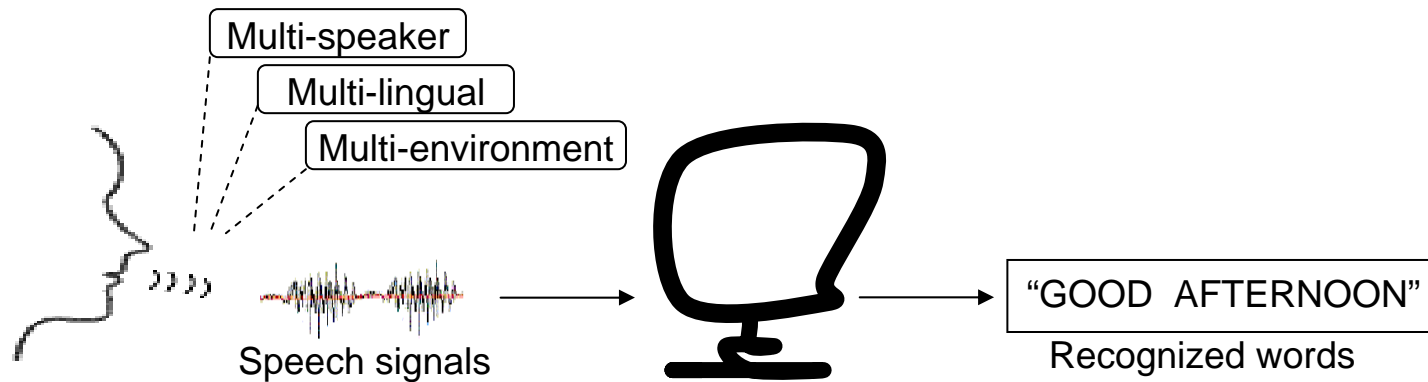
**Prof. Dr. Satoshi Nakamura**  
**ATR Spoken Language Communication**  
**Research Laboratories, Kyoto, Japan**

## Outline

- Introduction
  - Motivation and background
  - Thesis objectives
  
- Graphical framework to incorporate knowledge sources (GFIKS)
  - Apply at different levels of speech recognition
  - Use various kinds of knowledge
  
- Closing
  - Thesis contributions
  - Future directions

## Spoken Language Dialogue System

Human-machine communication



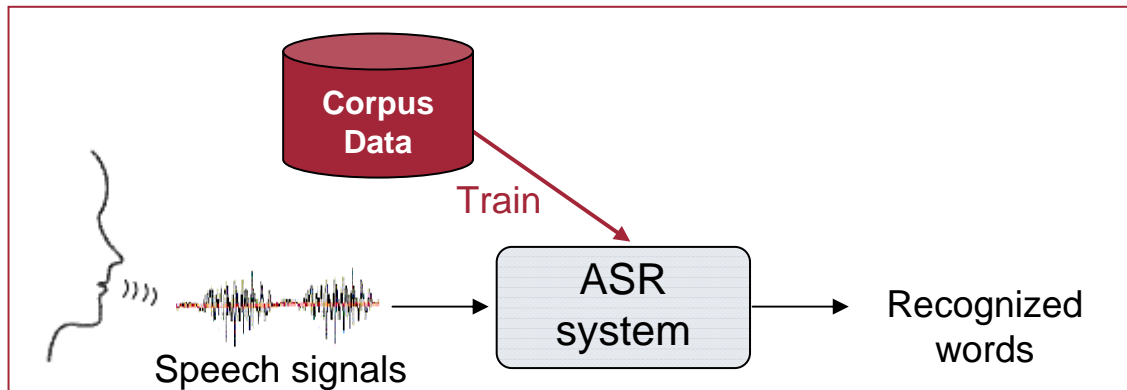
Fundamental technology:

Automatic speech recognition (ASR)

- Commercially used in controlled conditions
- Not ready for use by **any** speaker, **any** language in **any** environment

## Approaches to ASR

Data-driven (statistical) approach:



[+] Automatically learn from data

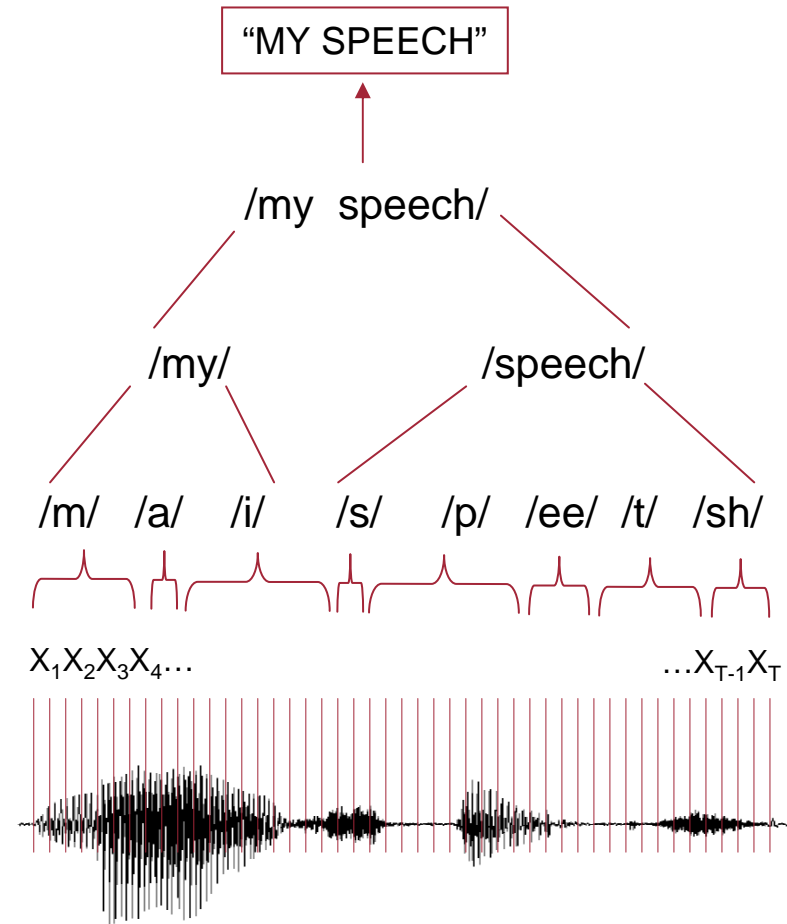
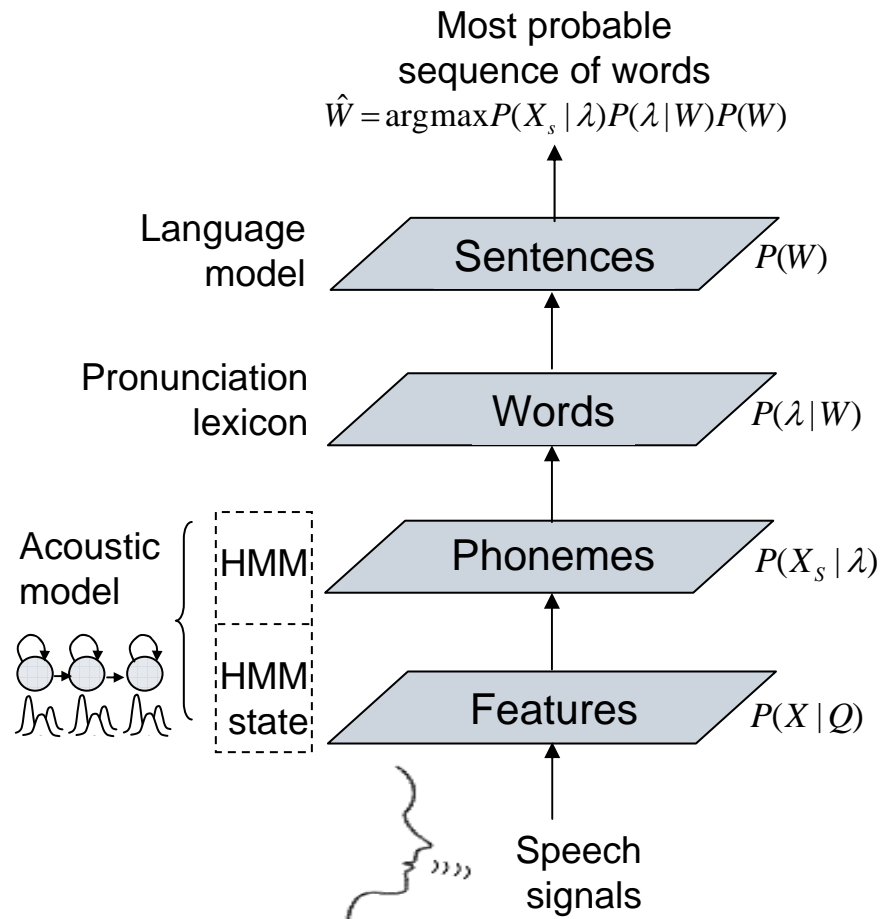
[-] Performs well on observed events

[-] Performance degrades if conditions change

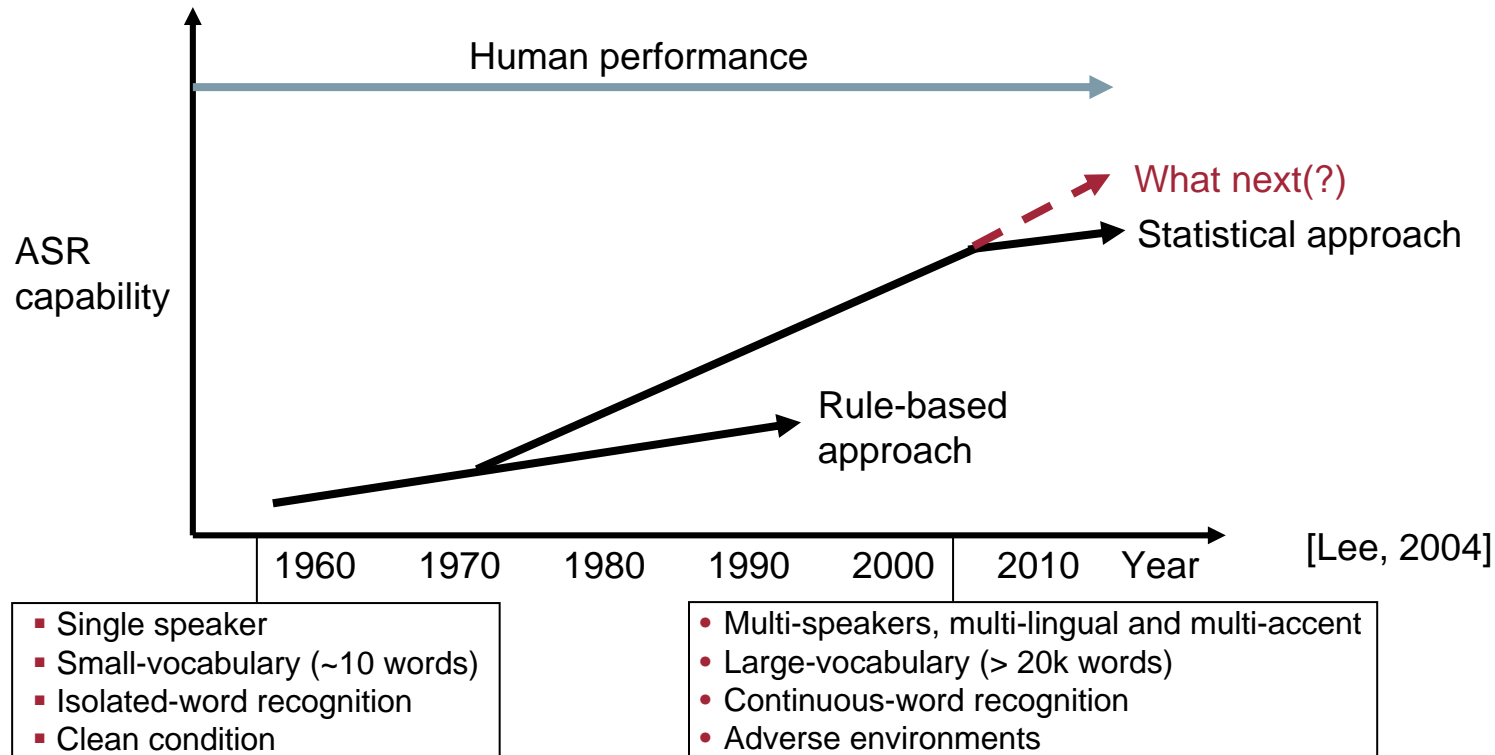
[-] To improve performance, more representative data required

⇒ Outperforms rule-based approach and becomes state-of-art ASR system

## State-of-the-art Statistical ASR



## ASR Technology Progress

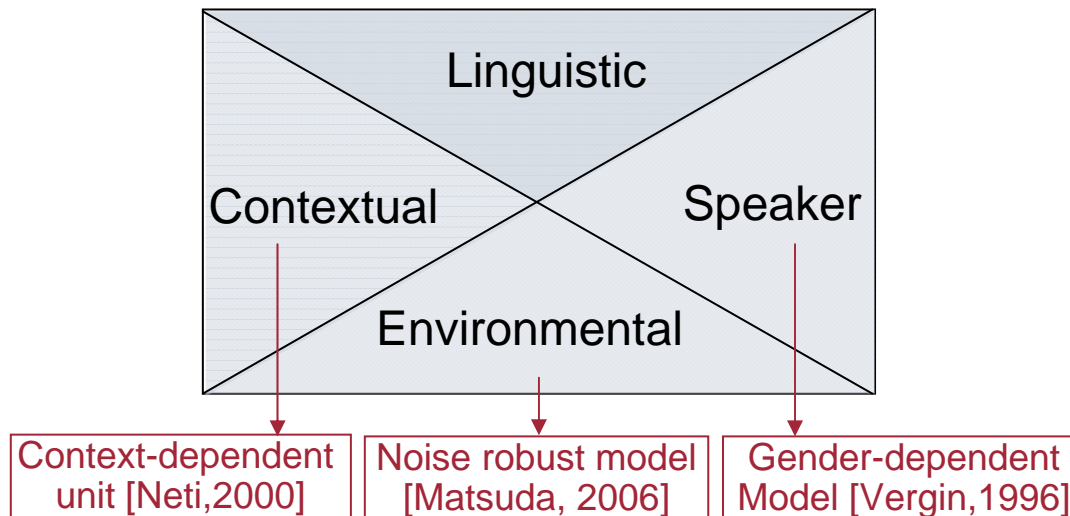


State-of-art ASR:

- **Performance drops** as task constraints are relaxed
- Still far **below human performance**
- Relying on statistical approaches with **more data is not enough**

## Sources of Speech Variability

Major variabilities:



Incorporating variability:

- Impose structure to explain variation
- Reduce uncertainty and increase predictability

State-of-the-art techniques:

- Heuristic
- Not unified framework

⇒ Efficient and unified approach is needed

## Thesis Objectives

Provide a novel ASR framework

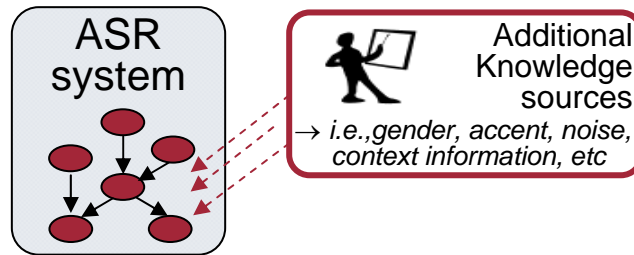
1. Incorporate various knowledge in unified way
2. Keep model complexity low
3. Improve speech recognition performance

## Outline

- Introduction
  - Motivation and background
  - Thesis objectives
- Graphical framework to incorporate knowledge sources (GFIKS)
  - Apply at different levels of speech recognition
  - Use various kinds of knowledge
- Closing
  - Thesis contributions
  - Future directions

## Proposed Approach

- Graphical framework to incorporate knowledge sources (GFIKS)
- Based on Bayesian network (BN)



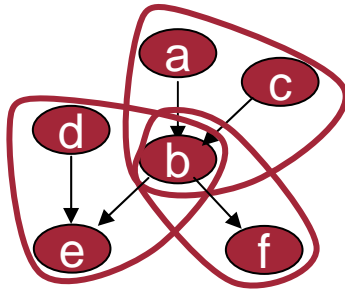
### Offers important advantages:

1. Universal & efficient to encode any structure through its topology
2. Provides a way to simplify network complexity

## The Use of Bayesian Networks

A Bayesian network:

- Consists of variables (nodes) and directed edges (causal relationships)
- Learn conditional probabilities between any variables



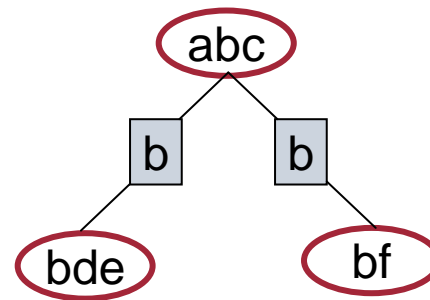
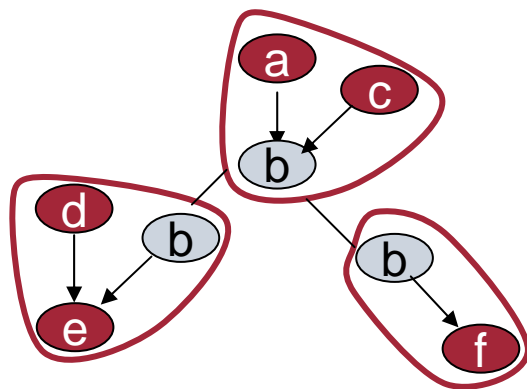
BN joint PDF:

$$P(a, b, c, d, e, f)$$

$$= P(b | a, c)P(a)P(c)P(e | d, b)P(d)P(f | b)$$

Inference:

- Direct on BN:** if PDF can be solved analytically
- Junction tree algorithm:** decompose global probability function

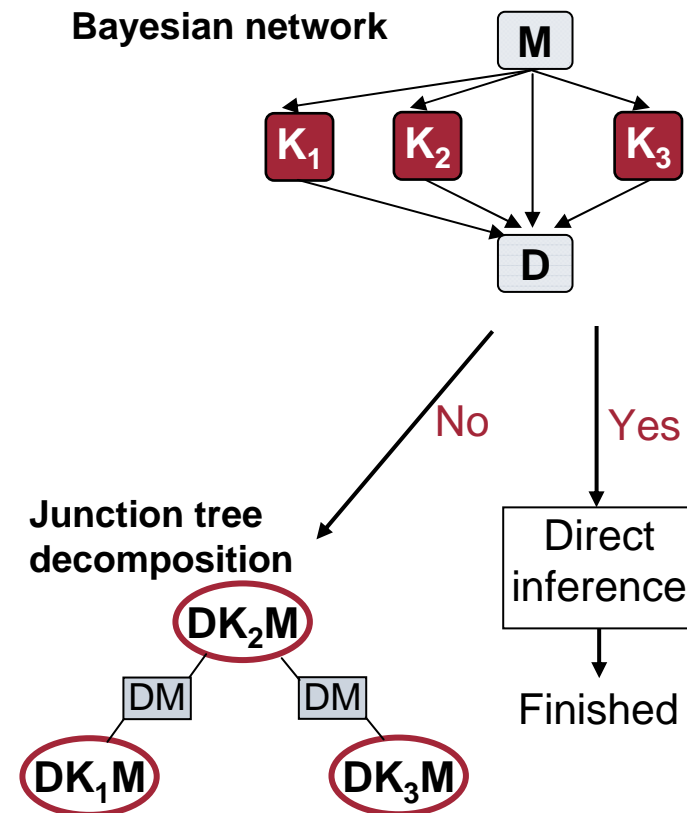
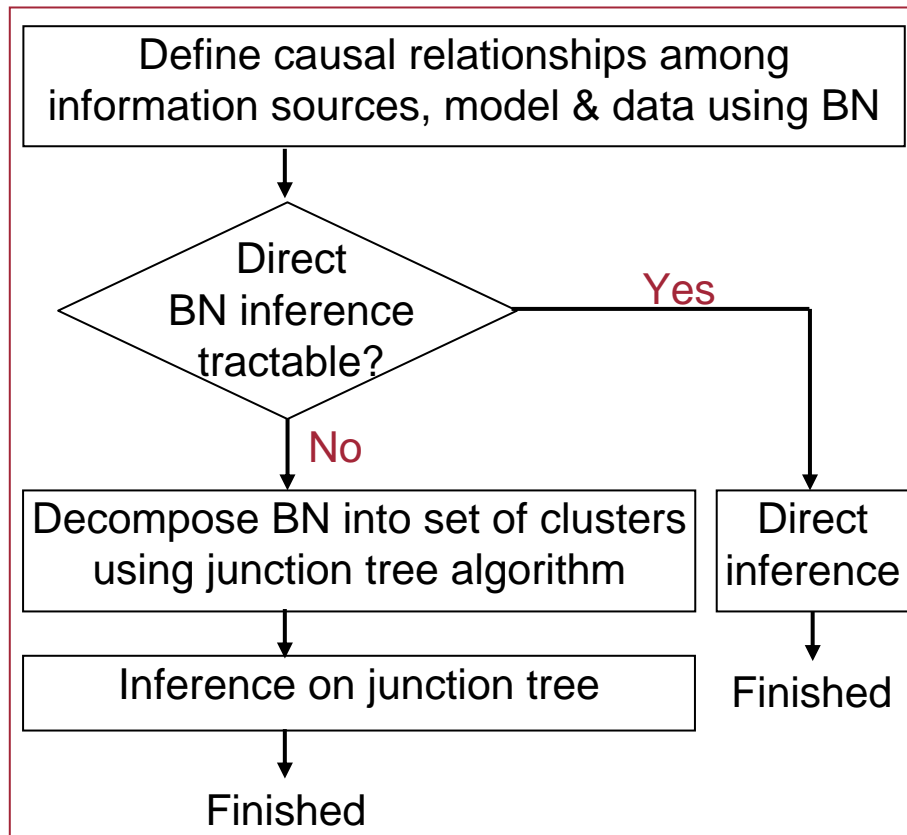


Overall PDF:

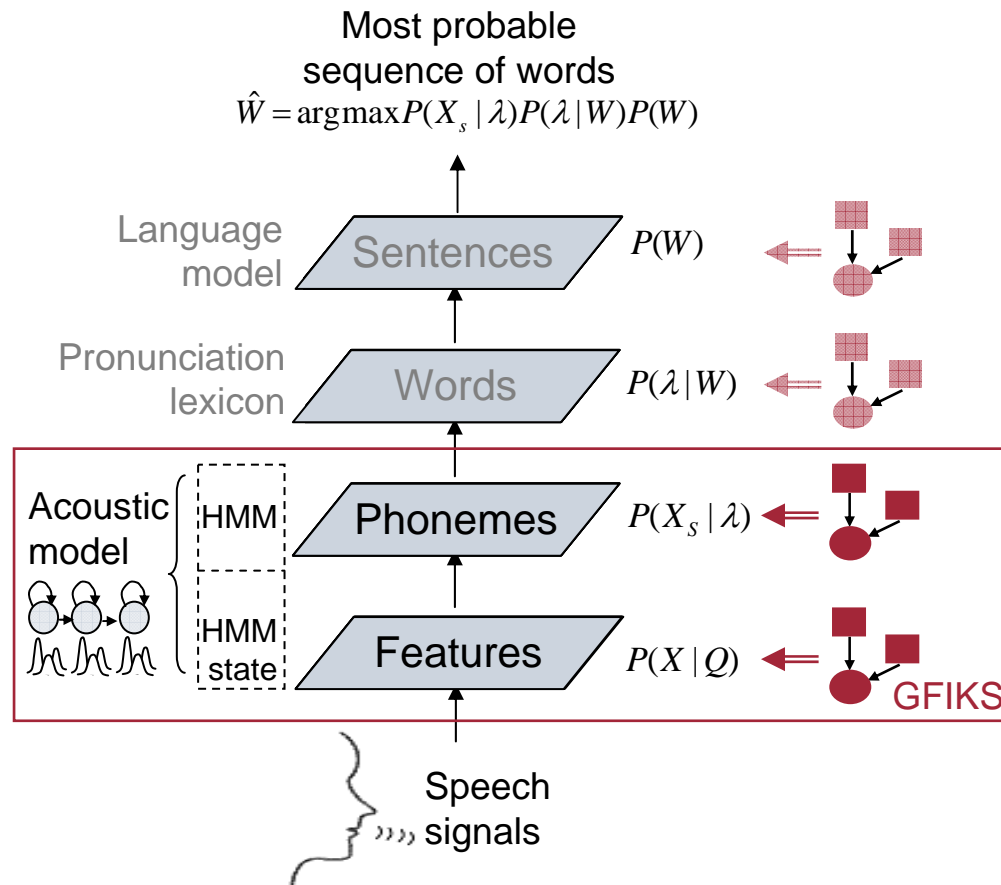
$$P(a, b, c, d, e, f)$$

$$= \frac{P(a, b, c)P(b, d, e)P(b, f)}{P(b)P(b)}$$

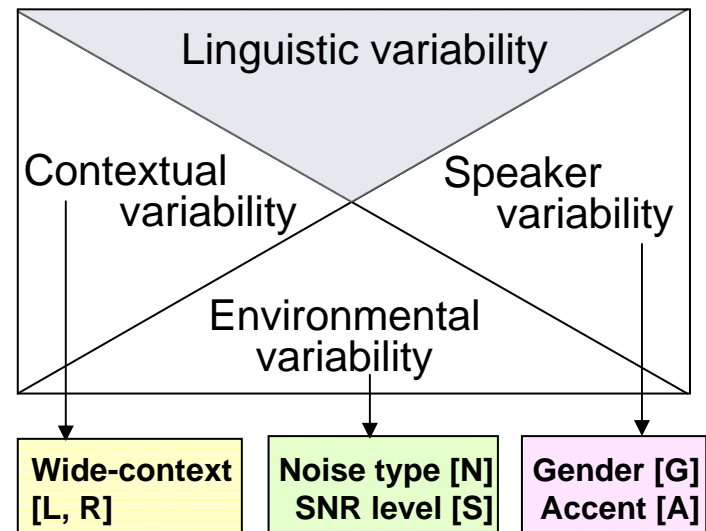
## GFIKS Procedure



## GFIKS in State-of-the-art Statistical ASR

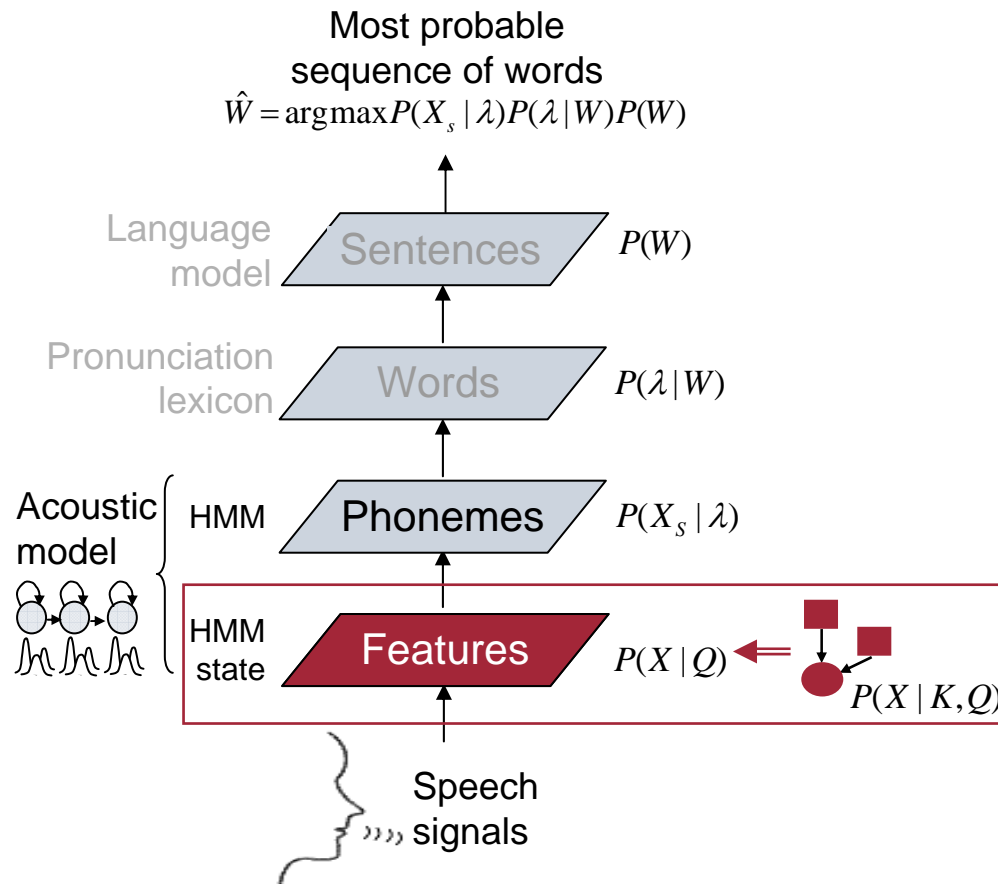


Knowledge sources:

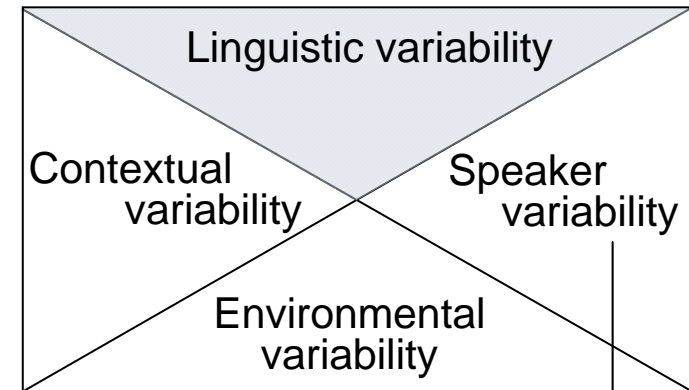


⇒ ATR speech recognition system used as baseline

## GFIKS at Feature Level



Knowledge sources:



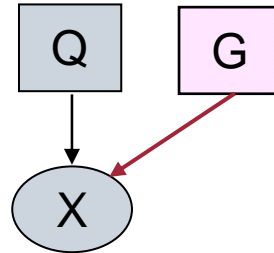
Gender [G]

→ Female

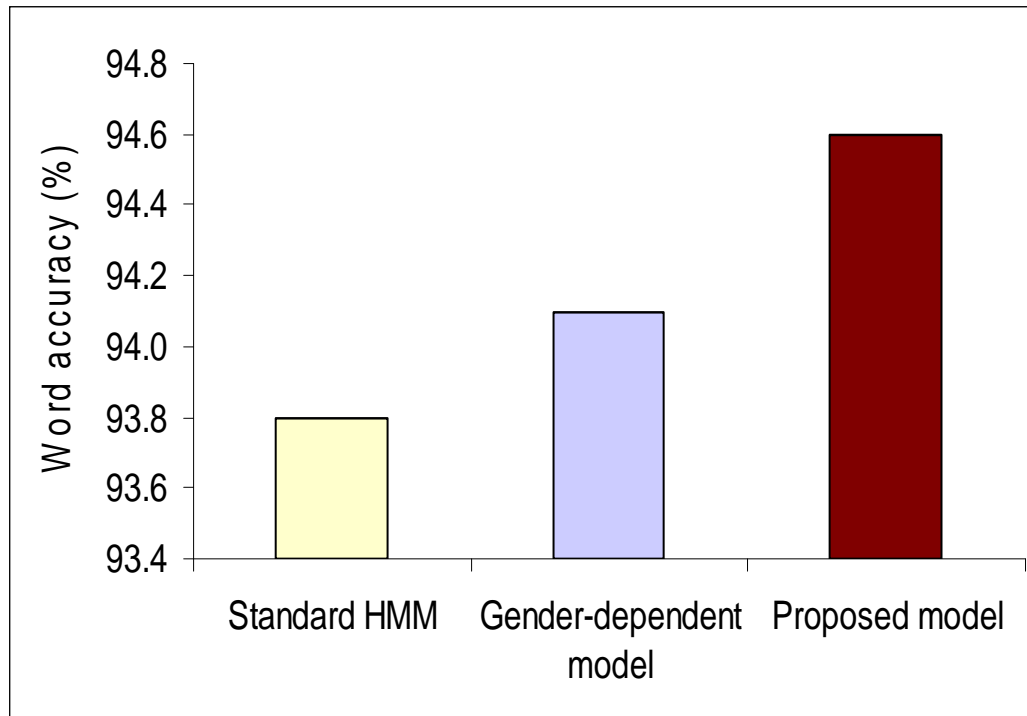
→ Male

## Incorporating Gender Information

BN topology:



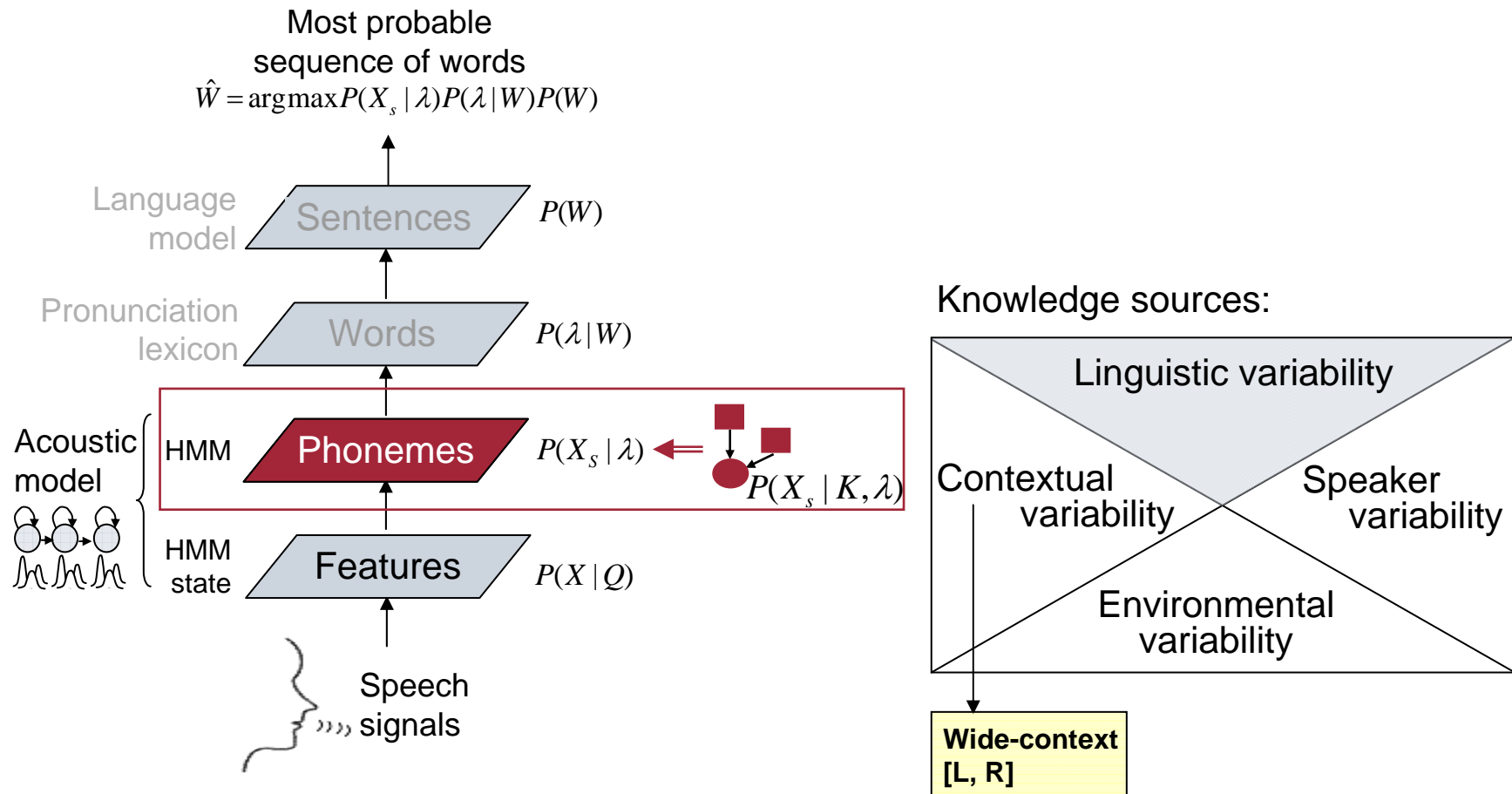
Results:



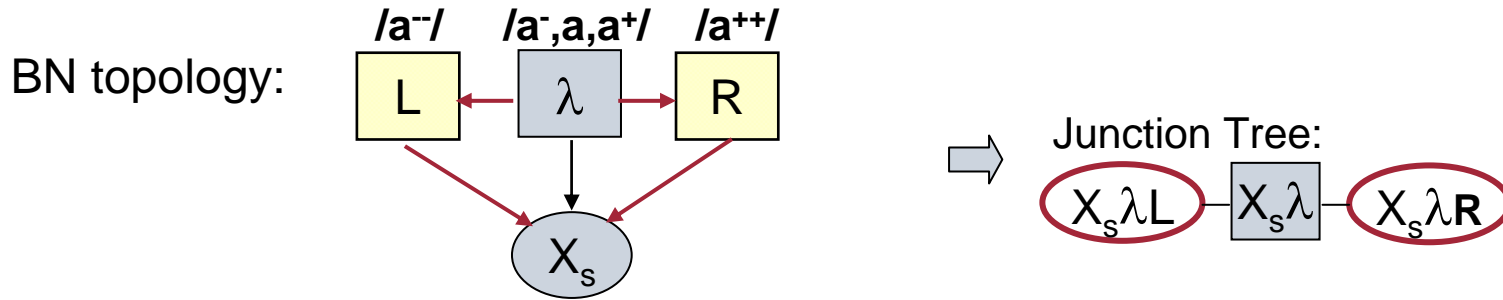
### Experimental evaluation:

- LVCSR task on news domain (WSJ)
- Follow official benchmark test set-up
- 60 hrs training data
- Direct inference on BN
- Performance : **12.9% relative WER reduction (significant)**

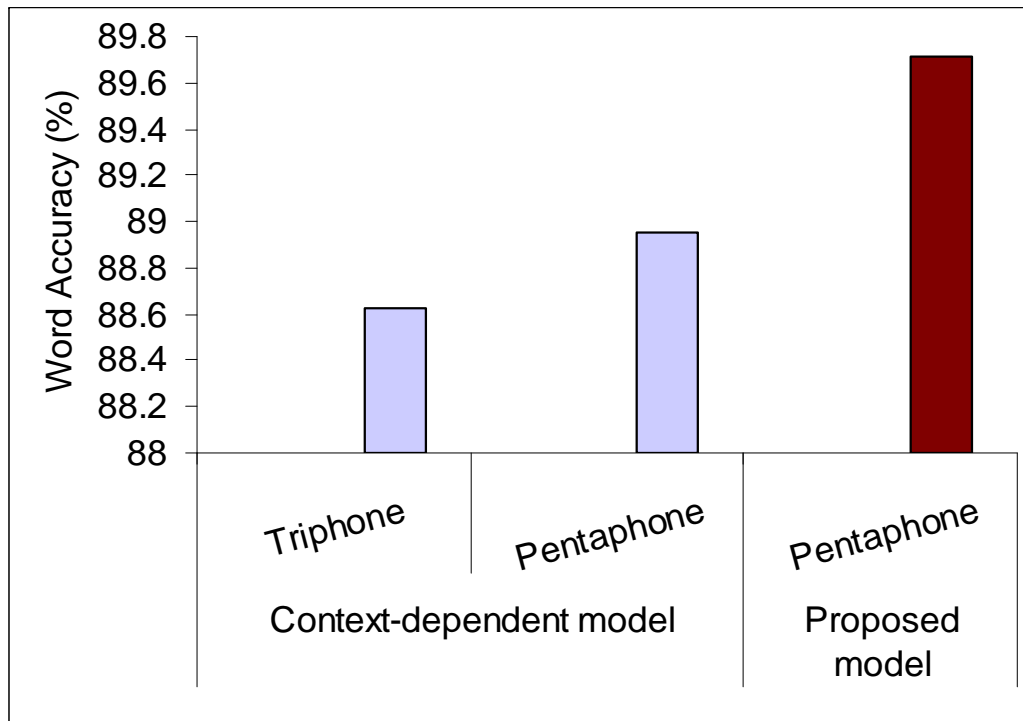
## GFIKS at Phonemes Level



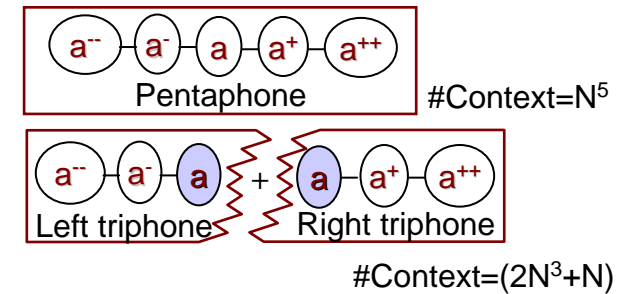
## Incorporating Context Information



### Results:



### Pentaphone composition:



### Experimental evaluation:

- LVCSR task on travel domain (ATR-BTEC)
- 60 hrs training data
- Inference on junction tree
- Performance : **9.5% relative WER reduction (significant)**

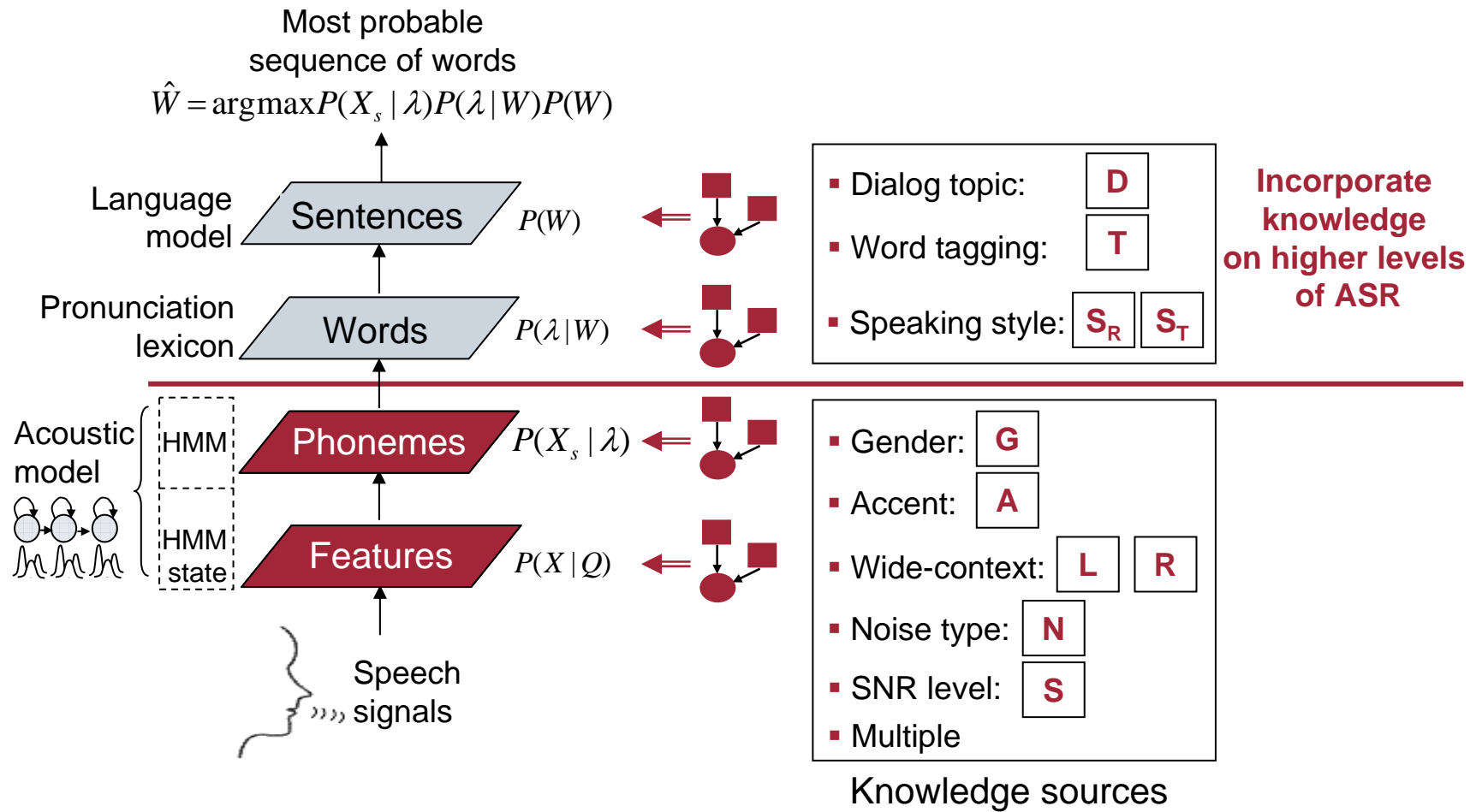
## Outline

- Introduction
  - Motivation and background
  - Thesis objectives
  
- Graphical framework to incorporate knowledge sources (GFIKS)
  - Apply at different levels of speech recognition
  - Use various kinds of knowledge
  
- Closing
  - Thesis contributions
  - Future directions

## Thesis Contributions

- **Theoretical**
  - Developed framework incorporating any knowledge at any ASR level
  - Probabilistic relationships among sources learned through BN
  - Junction tree based complexity reduction
  
- **Application-related**
  - Different levels: features and phonemes
  - Knowledge sources: gender, noise, accent, wide-phonetic context
  
- **Experimental**
  - Compared with state-of-the-art HMM-based speech recognition systems
  - Consistently improved and gave significantly better performance
  - Largest improvement achieved **25.3%** relative WER reduction
  
- **Practical**
  - Implemented in the latest version of ATR speech recognition system

## Future Directions



## Publications and Patents (1)

### Scientific Journals:

1. S. Sakti, et. al, "Incorporating Knowledge Sources into a Statistical Acoustic Model for Spoken Language Communication Systems", IEEE Transactions on Computers 2007.
2. S. Sakti, et. al, "Improving Acoustic Model Precision by Incorporating a Wide Phonetic Context Based on a Bayesian Frame-work", IEICE Transactions on Information and Systems 2006.
3. S. Sakti, et. al, "A Hybrid HMM/BN Acoustic Model Utilizing Pentaphone-Context Dependency", IEICE Transactions on In- formation and Systems 2006.

### Peer-reviewed Papers at International Conferences:

1. S. Sakti, et. al, "Development of Indonesian Large Vocabulary Continuous Speech Recognition System within A-STAR Project", in Proc. TCAST, India, 2008.
2. S. Sakti, et. al, "A method to Integrate Additional Knowledge Sources Into HMM Based on Junction Tree Decomposition", in Proc. EUSIPCO, Poland, 2007.
3. S. Sakti, et. al, "An HMM Acoustic Model Incorporating Various Additional Knowledge Sources", in Proc. EUROSPEECH, Belgium, 2007.
4. S. Sakti, et. al, "The Use of Bayesian Network for Incorporating Accent, Gender and Wide-Context Dependency Information", in Proc. ICSLP, USA, 2006.
5. S. Sakti, et. al, "Incorporation of Pentaphone-Context Dependency Based on Hybrid HMM/BN Acoustic Modeling Framework", in Proc. ICASSP, France, 2006.
6. S. Sakti, et. al, "Rapid Development of Initial Indonesia Phoneme-Based Speech Recognition Using Cross-Language Approach", in Proc. O-COCOSDA, Indonesia, 2005.
7. S. Sakti, et. al, "Modeling Quasi-Pentaphone Units with the Hybrid HMM/BN Acoustic Model", in Proc. SPECOM, Greece, 2005.
8. S. Sakti, et. al, "Incorporating a Bayesian Wide Phonetic Context Model for Acoustic Rescoring", in Proc. EUROSPEECH, Portugal, 2005.

## Publications and Patents (2)

### Other Paper Presentations:

1. S. Sakti, et. al, "Large Vocabulary ASR for Indonesian Language in the A-STAR Project", in Proc. ASJ Autumn Meeting, Japan, 2007.
2. S. Sakti, et. al, "Utilizing Junction Tree Decomposition for Incorporating Accent, Gender, and Wide-Context Dependency Information", in Proc. ASJ Spring Meeting, Japan, 2007.
3. S. Sakti, et. al, "Utilizing Bayesian Network and Junction Tree Decomposition for Incorporating Additional Knowledge Sources into a Statistical Acoustic Model", in Proc. IEICE, Japan, 2006.
4. S. Sakti, et. al, "A Hybrid Pentaphone HMM/BN Acoustic Model", in Proc. ASJ Spring Meeting, Japan, 2006.
5. S. Sakti, et. al, "Composing a Wide Phonetic Context Unit based on Bayesian Framework", in Proc. ASJ Autumn Meeting, Japan, 2005.

### Invention Patents:

1. S. Sakti, et. al, "Probability Calculation Apparatus for Incorporating Knowledge Sources and Computer Program" (ongoing Process).
2. S. Sakti, et. al, "Acoustic Modeling Developing Apparatus and Computer Program", Japan Patent No. P2007-155833A , Issue on June 21st, 2007.
3. S. Sakti, et. al, "An Apparatus of Rescoring a Hypothesis in a Speech Recognizing System", Japan Patent No. P2007-52165A , Issue on March 1st, 2007.
4. S. Sakti, et. al, "A Method of Preparing a Wide-Context Acoustic Model and an Automatic Speech Recognition", Japan Patent No. P2007-52166A, Issue on March 1st, 2007.

THANK YOU