



ulm university

universität  
uulm

# Statistical Pronunciation Modeling for Non-native Speech

Dissertation

Rainer Gruhn

Nov. 14<sup>th</sup>, 2008

Institute of Information Technology

University of Ulm, Germany

In cooperation with Advanced Telecommunication Research Labs, Kyoto

## Outline

- Introduction
  - Motivation and background
  - Thesis objectives
  
- Hidden Markov Models as statistical lexicon
  - Initialization and training
  - Application
  
- Experiments
  - ATR non-native speech database
  - Evaluation
  
- Closing
  - Thesis contributions
  - Publications

## Non-native English speech

- Relevant in many applications of speech recognition:
  - Automatic tourist information system
  - Car navigation with user going abroad
  - Speech recognition in the media domain
- Mispronunciations include phoneme insertions, deletions and substitutions  
(e.g. in German English: /tʰ/)
- Different patterns for each language (→ Accent)
- Example: “Certainly. What time do you anticipate checking in?”



Chinese



Indonesian

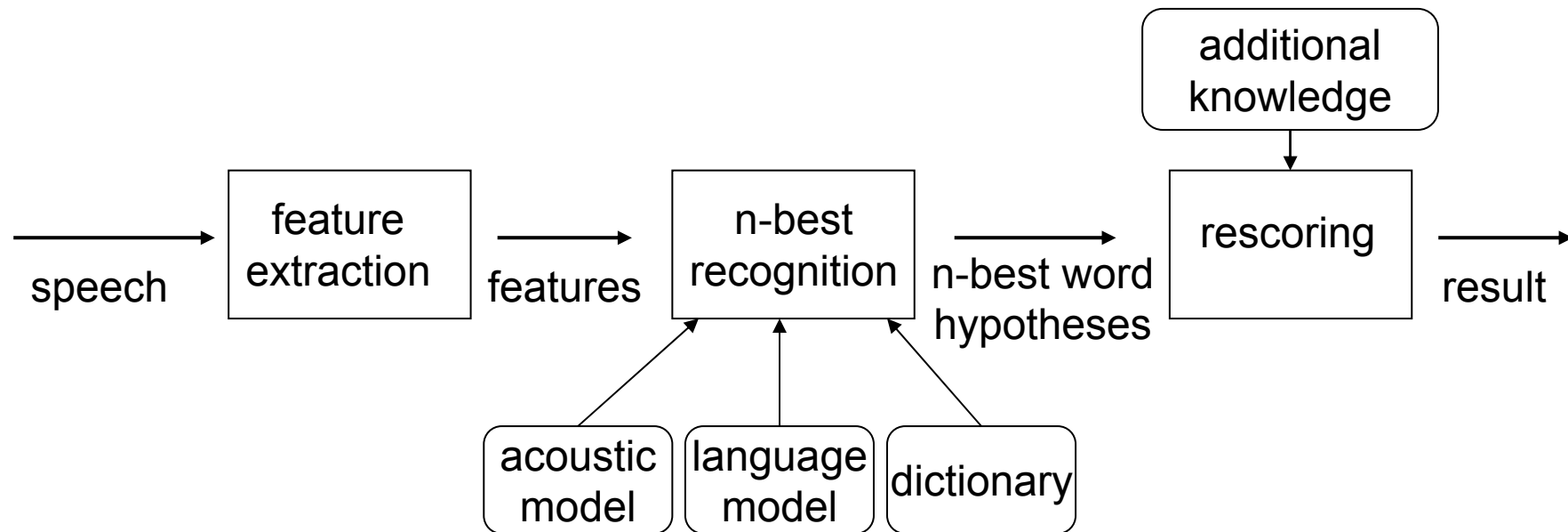


Japanese

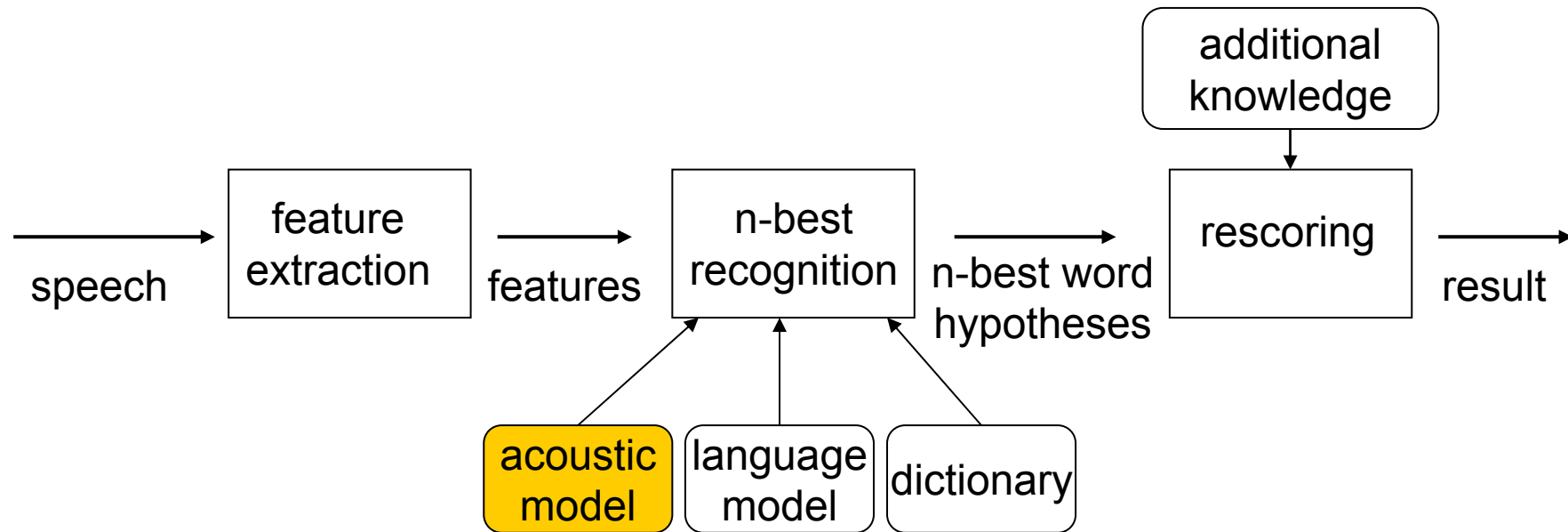


French

## Schematic Outline of a Speech Recognition System



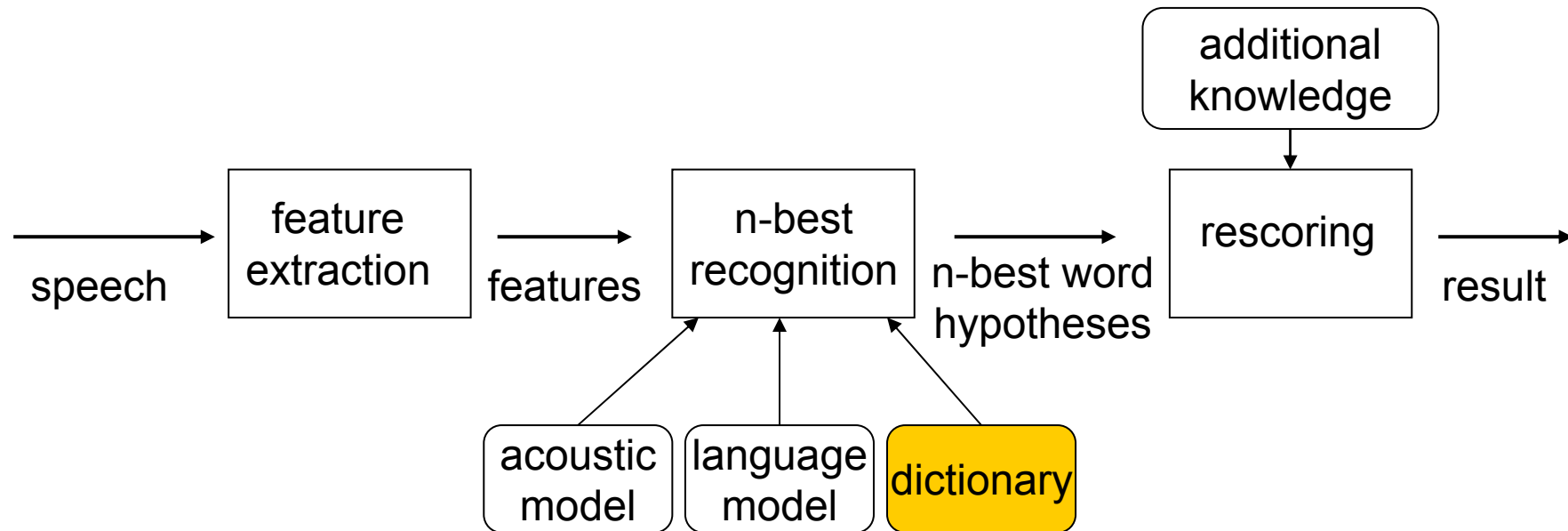
## Schematic Outline of a Speech Recognition System



Improve performance for individual speakers:

→ acoustic model adaptation (e.g. Maximum A Posteriori)

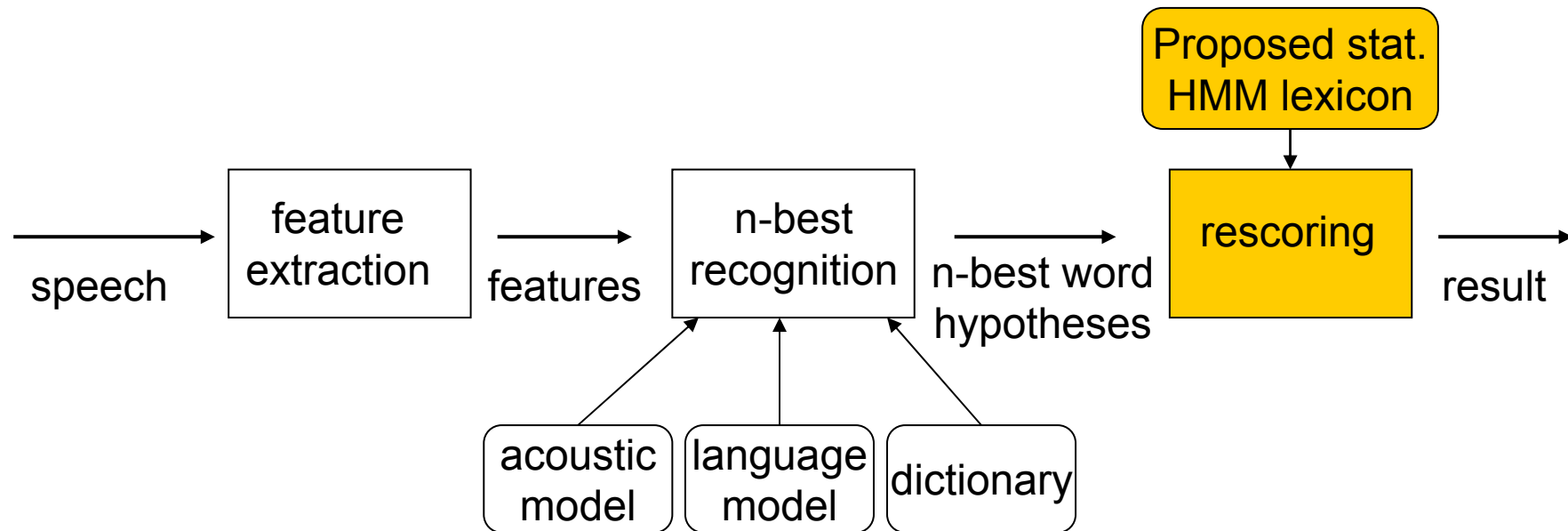
## Schematic Outline of a Speech Recognition System



Common approach for non-native speakers:

→ rule-based dictionary enhancement (Goronzy 2002, Mayfield-Tomokiyo 2001)

## Schematic Outline of a Speech Recognition System



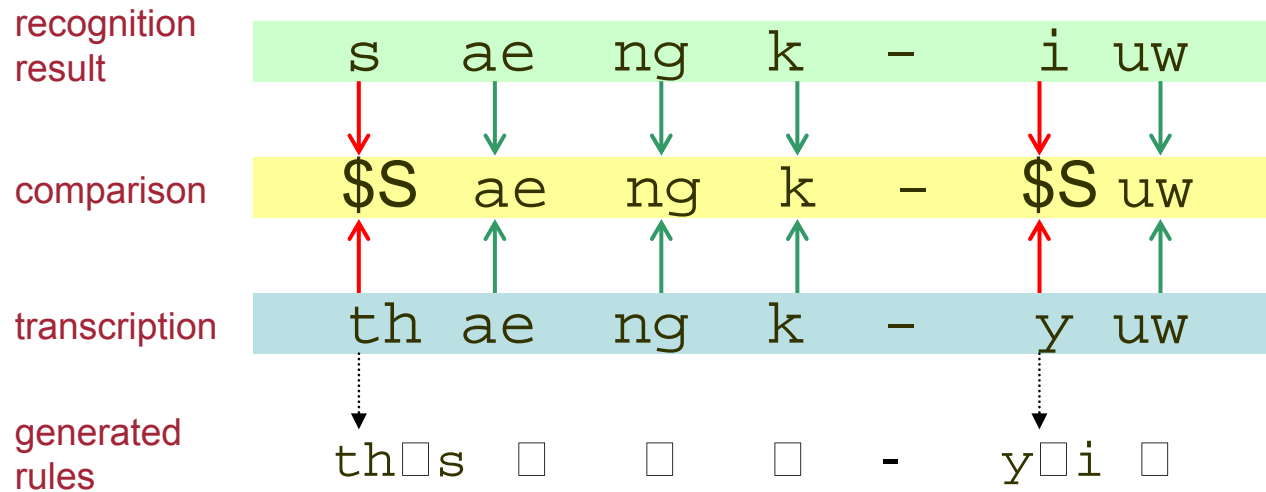
Proposed method:

→ rescoring with HMMs as statistical lexicon

## Common Approach: Rules

Common approach:

- Phoneme confusion rules (data driven / knowledge based)



- Apply rules on pronunciation dictionary

Rule set: `th□s, y□i`

`thank : /th ae ng k/ , /s ae ng k/;`  
`you : /y uw/, /i uw/;`

## Problems about Rules

- Pronunciation variations also depend on context
- Variations unseen in training data cannot be modeled
- Knowledge-based: Manual rule generation
- When rules are applied to pronunciation dictionary: tradeoff between:
  - Large dictionary (including all possible variations as entry)
  - Losing information (choosing to apply only some rules)

## Thesis Objective

- Non-native speech: many pronunciation variations, automatic speech recognition difficult
  - Improve automatic speech recognition of non-natives
  - Target: Model those variations automatically and statistically
  - Cover all pronunciation variations
- Approach: Train discrete Hidden Markov Models (HMM) for each word as pronunciation model

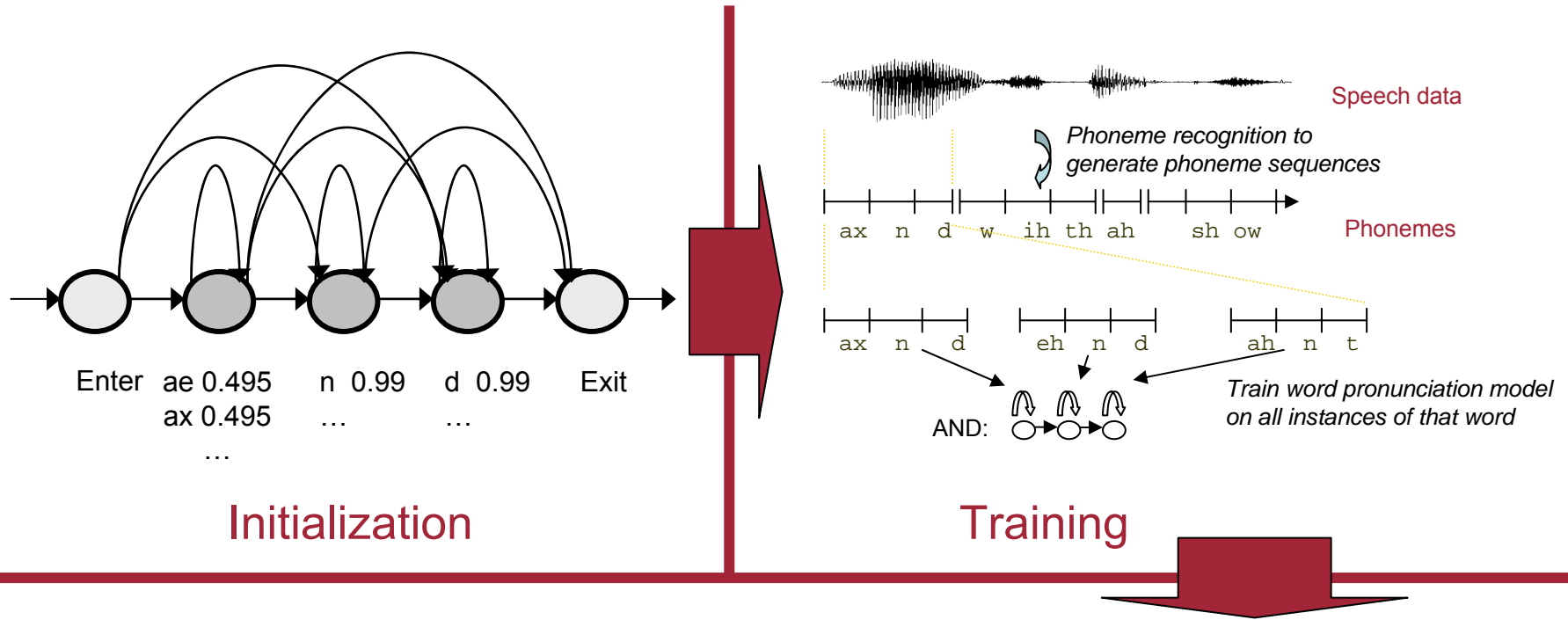
## Outline

- Introduction
  - Motivation and background
  - Thesis objectives
- Hidden Markov Models as statistical lexicon
  - Initialization and training
  - Application
- Experiments
  - ATR non-native speech database
  - Evaluation
- Closing
  - Thesis contributions
  - Publications

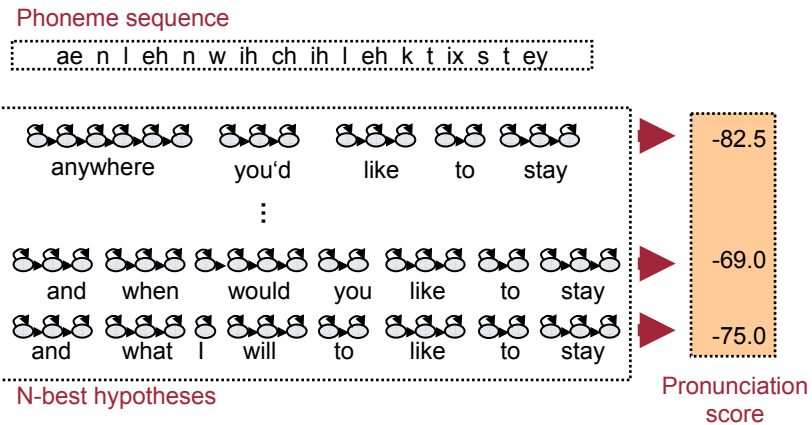
## Statistical Lexicon

- HMMs to represent pronunciations (not explicitly representing the confusions)
- One discrete HMM model for each word
- Initialization on baseline lexicon
- Training on phoneme sequences generated by phoneme recognition

# Initialization, Training and Application

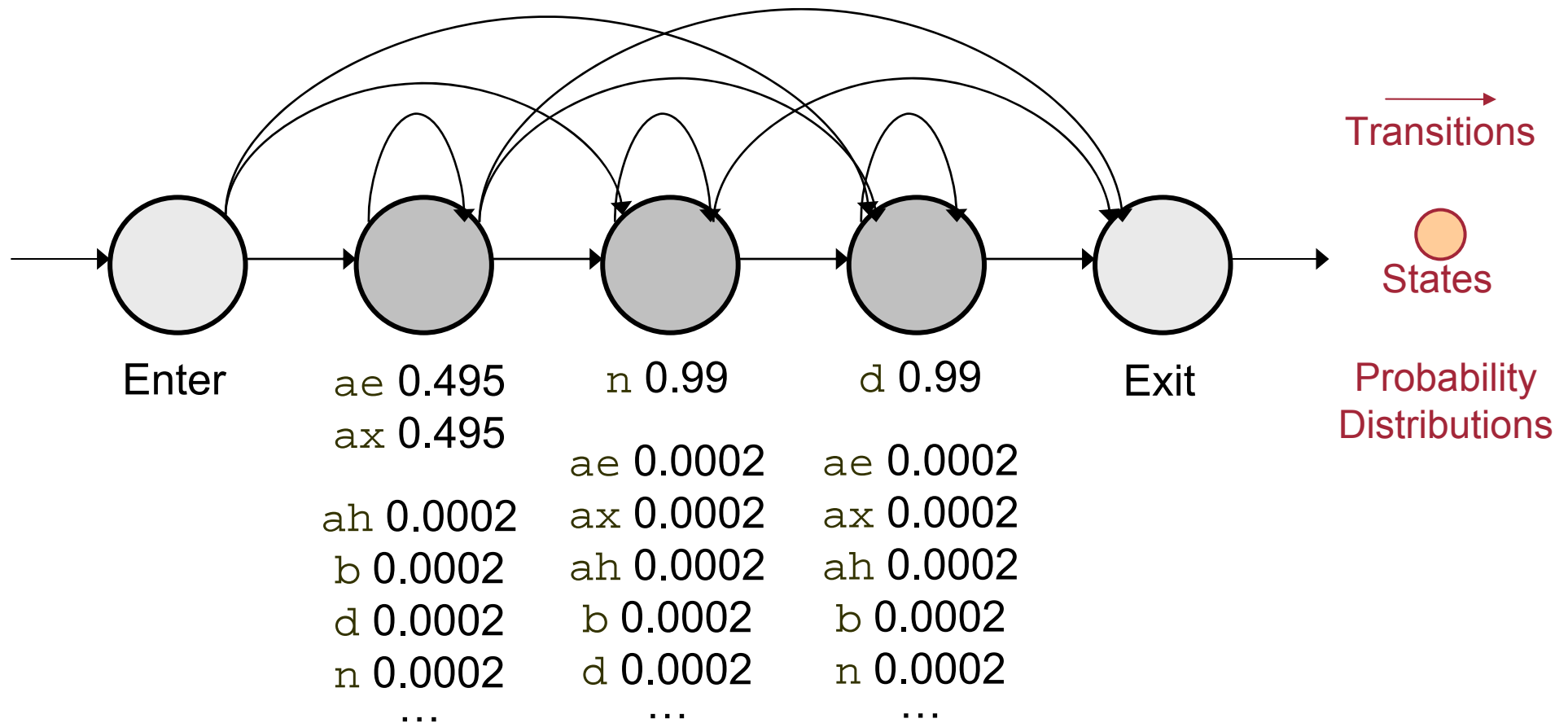


## Application of Models



## Word Model Example: AND

AND: /ae n d/  
 /ax n d/



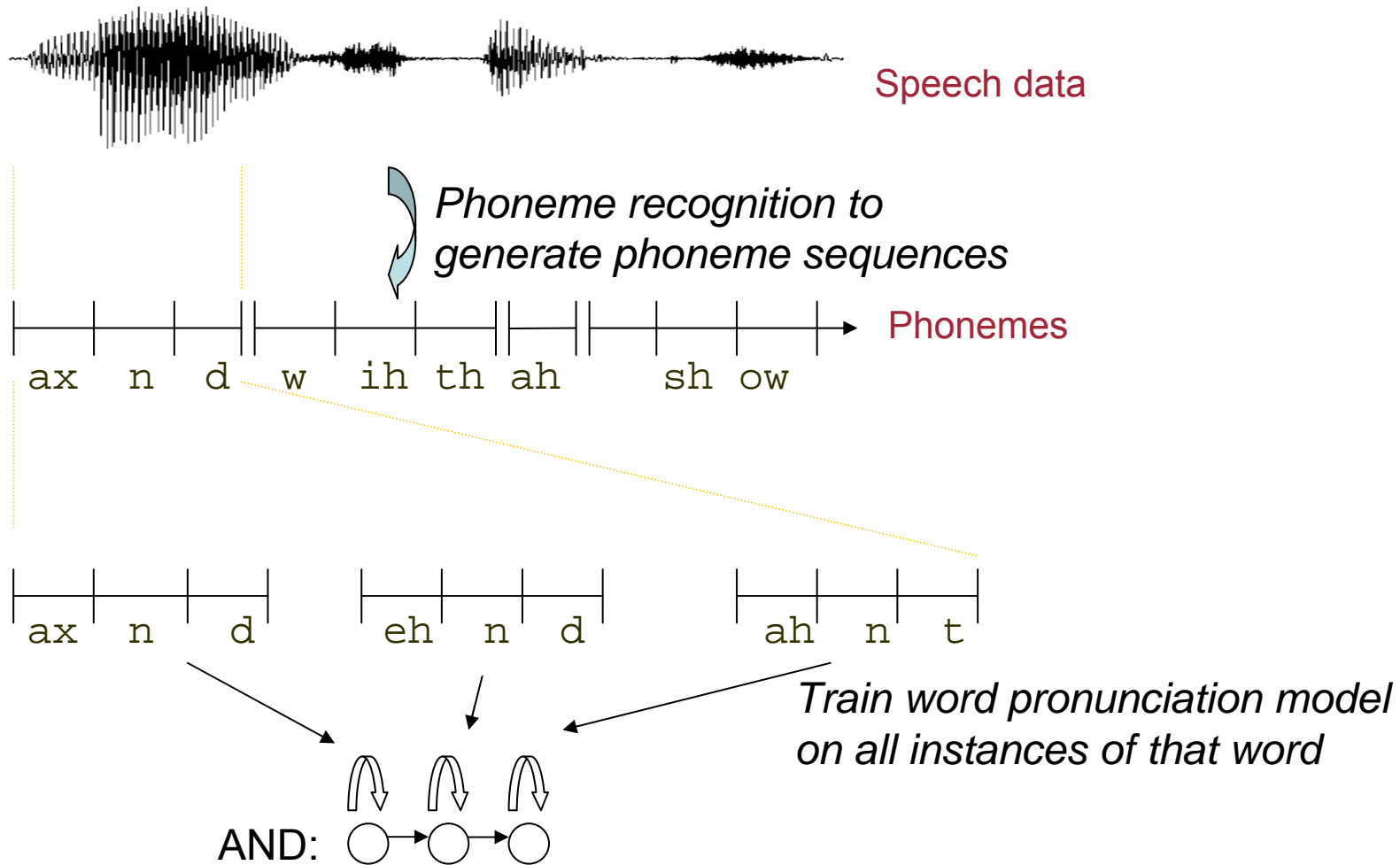
## Model Initialization

- Given: standard pronunciation dictionary
- One discrete HMM for each word
- Number of states equals number of baseline phonemes (+ enter, exit states)
- Several pronunciation variants in dictionary are integrated into word model

## Model Training

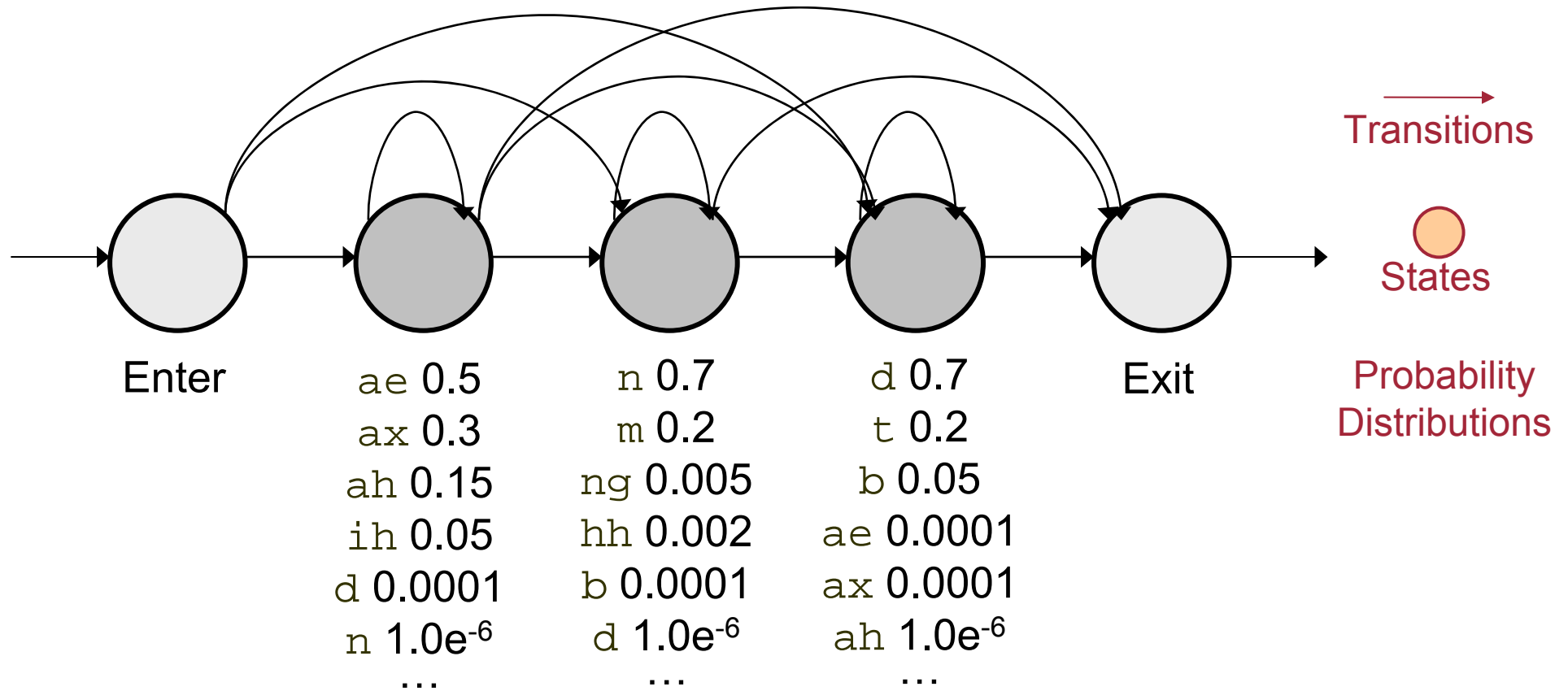
- Segmentation of training data into words
- Phoneme recognition
- Train discrete HMM for each word on phoneme sequence
- Default unseen words to baseline lexicon phoneme sequence(s)

# Training of Discrete HMMs



# Word Model After Training

AND: /ae n d/  
 /ax n d/

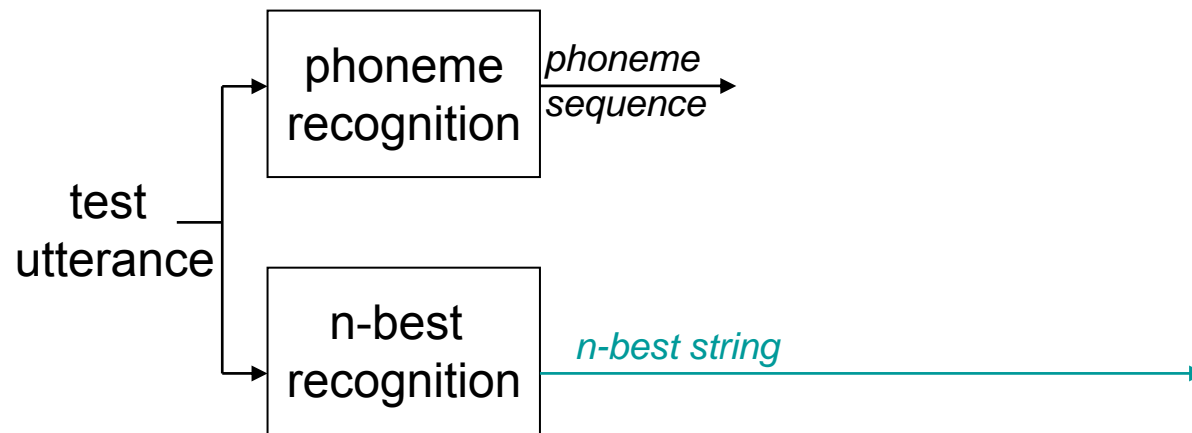


## Model Application



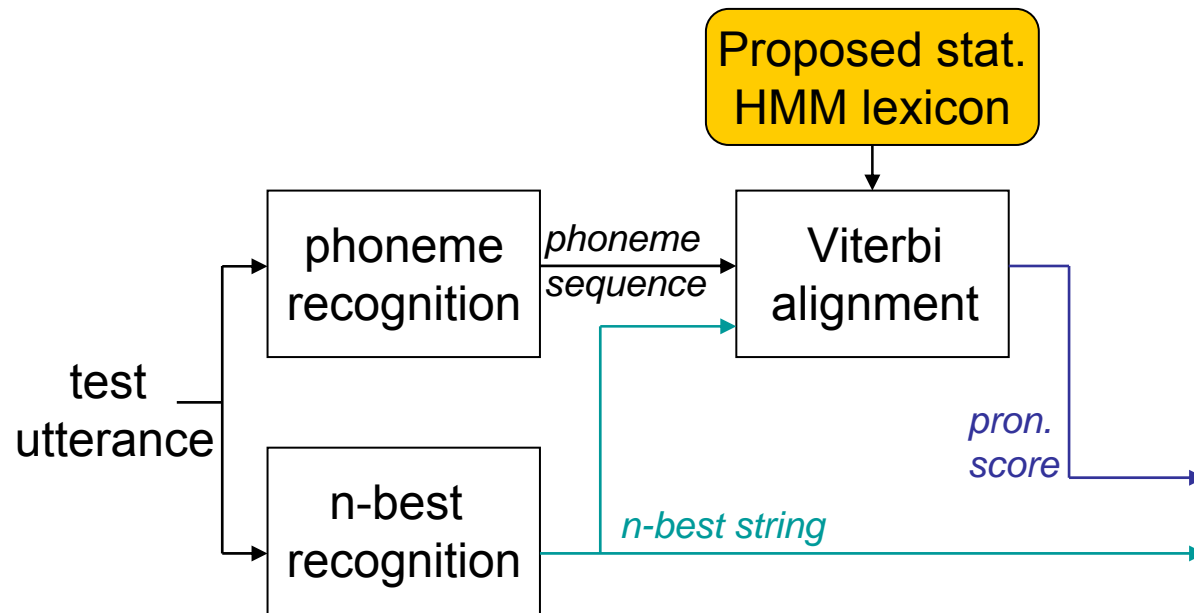
- Standard n-best decoding of test set

## Model Application



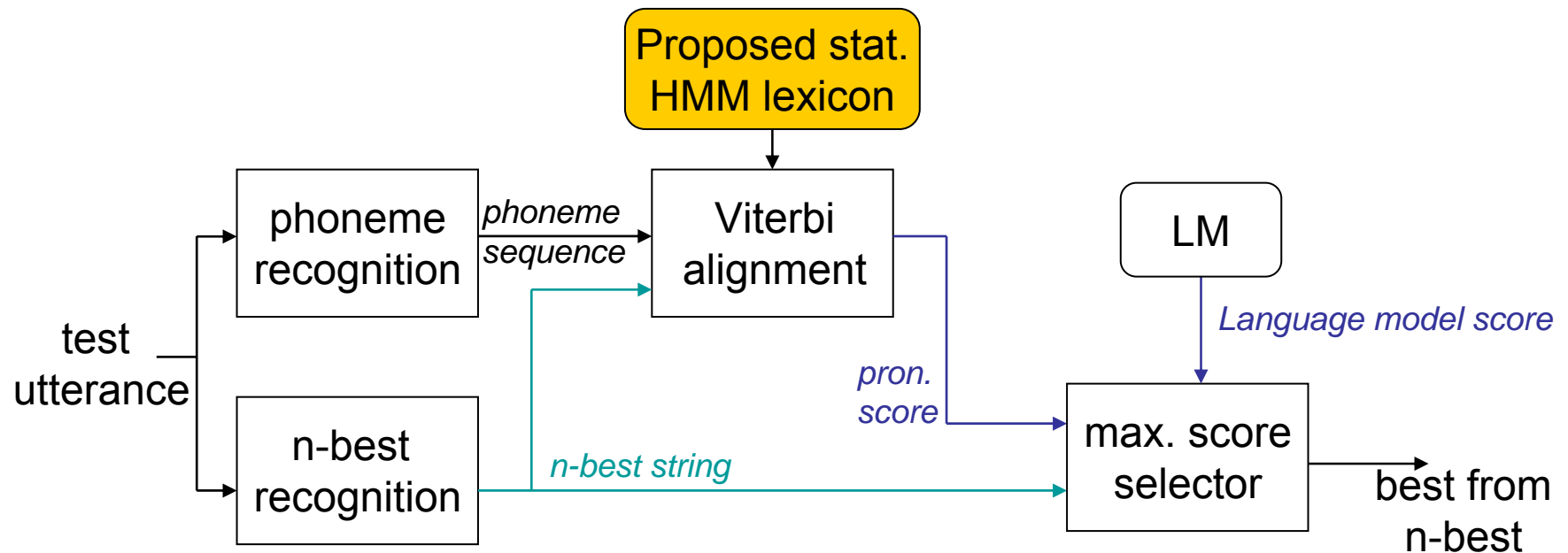
- Standard n-best decoding of test set
- 1-best phoneme recognition of whole utterance

## Model Application



- Standard n-best decoding of test set
- 1-best phoneme recognition of whole utterance
- Calculate pronunciation score of each n-best hypothesis

## Model Application

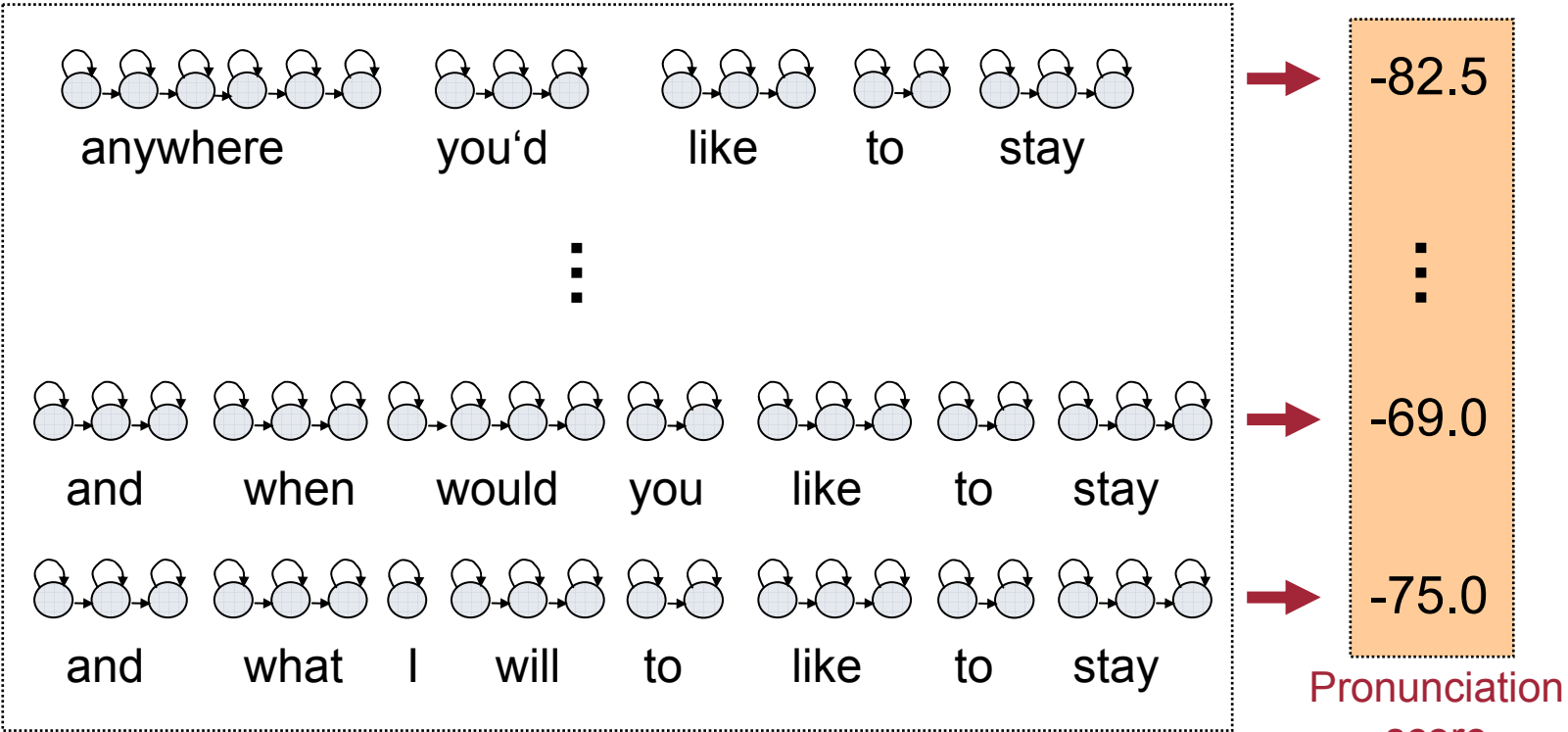


- Standard n-best decoding of test set
- 1-best phoneme recognition of whole utterance
- Calculate pronunciation score of each n-best hypothesis
- Select best hypothesis based on pronunciation score with weighted language model score

# Rescoring of N-best

Phoneme sequence

ae n l eh n w ih ch ih l eh k t ix s t ey



N-best hypotheses






Pronunciation score

## Outline

- Introduction
  - Motivation and background
  - Thesis objectives
- Hidden Markov Models as statistical lexicon
  - Initialization and training
  - Application
- Experiments
  - ATR non-native speech database
  - Evaluation
- Closing
  - Thesis contributions
  - Publications

## ATR Non-native Speech Database

- Existing comparable databases (large, multi-accent):
  - M-ATC, Hiwire: noisy, special military vocabulary
  - Crosstowns: unavailable to public
- Collected in this work
- One of the largest non-native English speech databases
- Data available at ATR
- Total 22h of speech

country	China	France	Germany	Indonesia	Japan	all
						
#speakers	17	15	15	15	28	96

## ATR Non-native Speech Database

- Per speaker: 12 minutes training, 2 minutes test data (2 hotel reservation dialogs)
- Read speech
- Content: Uniform set of
  - hotel reservation dialogs
  - phonetically balanced sentences
  - digit sequences
- Speaker skill: various, rated

## Database Collection

- Non-nativeness vs. anxiousness:
  - Instructor in same room, nodding
  - Non-intimidating environment
  - Words where speaker was not sure how to pronounce: speaker had to try
  - Speakers could repeat sentence until satisfied

## Experimental Setup

- Baseline dictionary: 7311 words, 8875 entries  
→ 7311 pronunciation HMMs
- 10-best word recognition
- Generate pronunciation HMMs separately for each accent group
- Acoustic model: trained on Wall Street Journal database
- Word bigram LM, trained on travel arrangement task text data

Phoneme/Word error rate

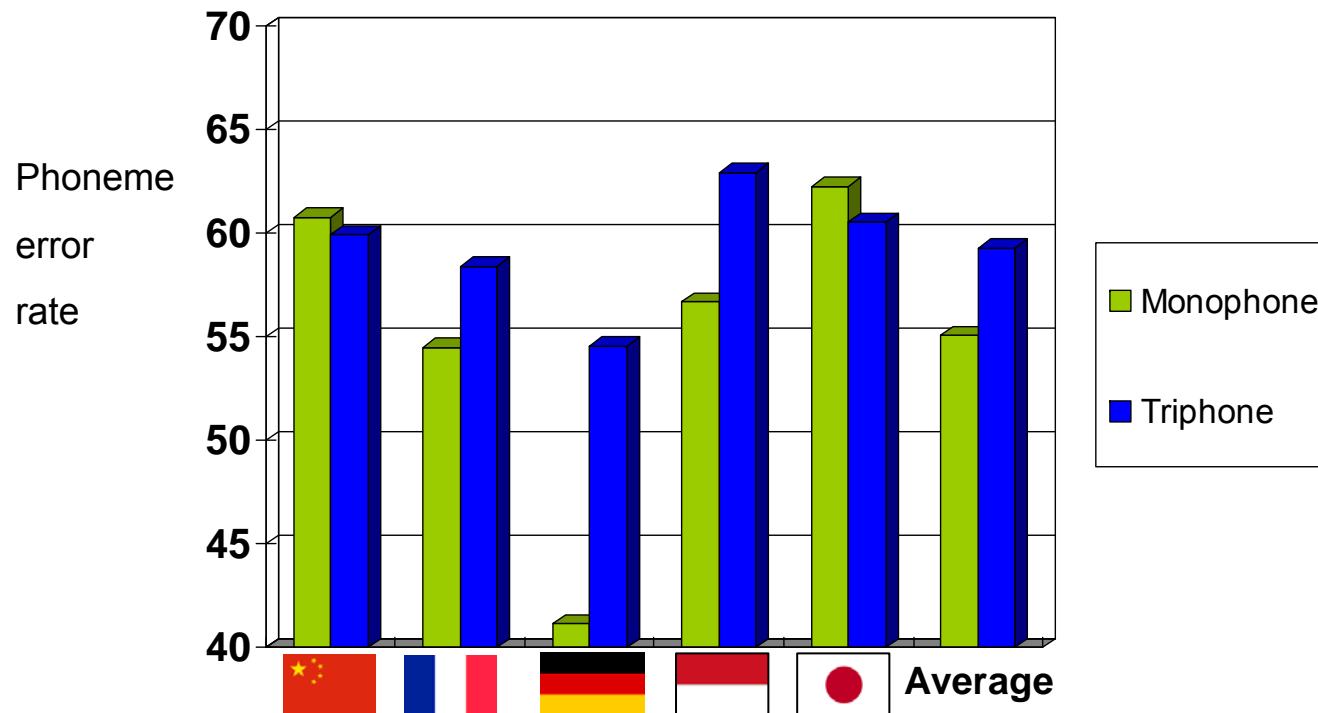
$$\frac{INS + DEL + SUB}{N_{total}}$$

Relative error rate improvement

$$\frac{ERR_{before} - ERR_{after}}{ERR_{before}}$$

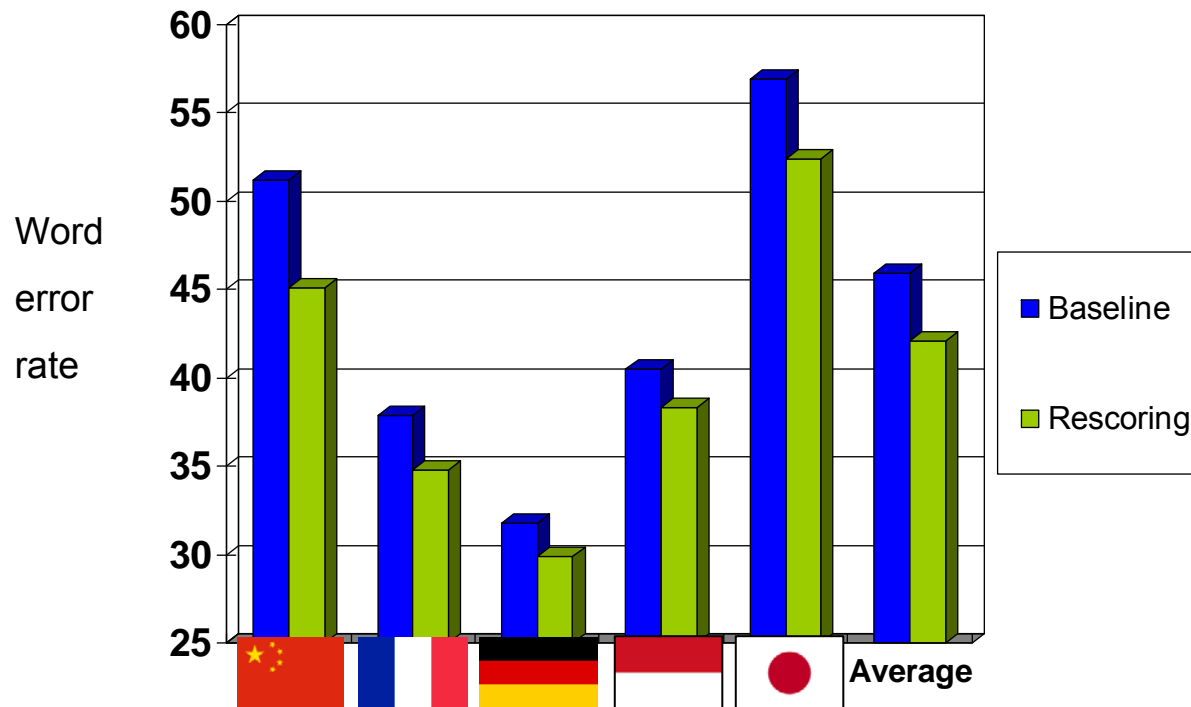
## Phoneme Recognition

- Both pronunciation model training and application steps require phoneme recognition
- Error rate calculated relative to canonical transcription
- Recognition of whole utterance
- Phoneme bigram as phonotactical constraint



## Pronunciation Scoring: Results

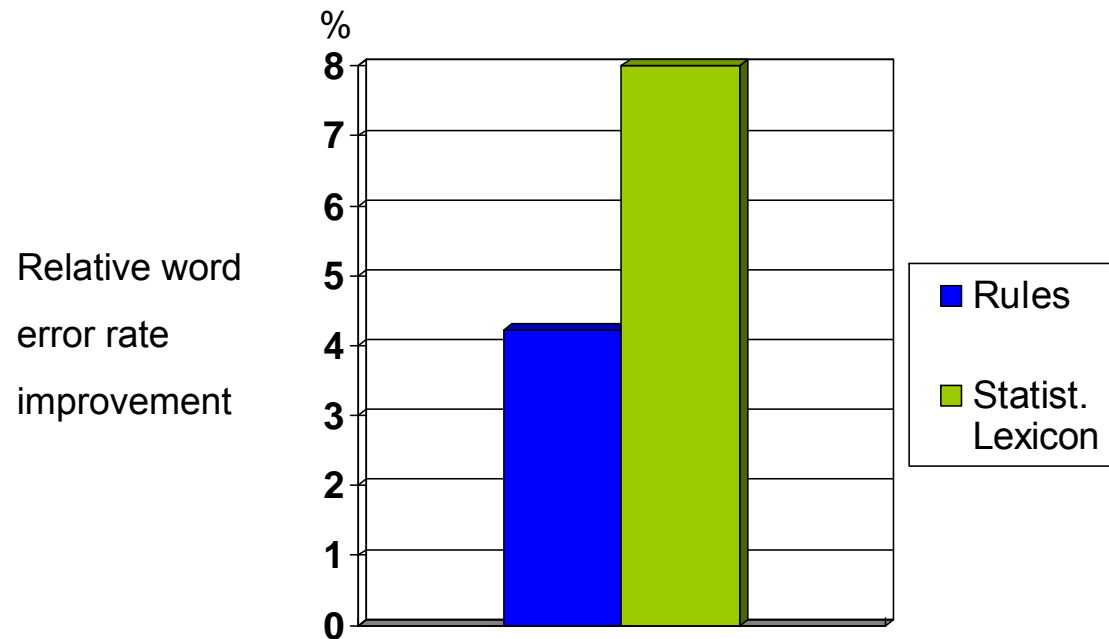
Word error rates for non-native speech recognition, with and without pronunciation rescoring



Accent type	CH	FR	GER	IN	JP	Avg
rel. WER impr.	11.9	8.3	5.9	5.4	8.0	8.2

## Comparing to Standard Technology

- Standard approach to adjust for non-native speech: Rule-based Dictionary modification
- Comparison of relative improvements



Improvement vs. pronunciation alternatives added to dictionary



Evaluated for the Japanese speaker set

## Outline

- Introduction
  - Motivation and background
  - Thesis objectives
  
- Hidden Markov Models as statistical lexicon
  - Initialization and training
  - Application
  
- Experiments
  - ATR non-native speech database
  - Evaluation
  
- Closing
  - Thesis contributions
  - Publications

## Thesis Contributions

- **Theoretical**
  - Integrated framework for statistical pronunciation modeling
  - Both learned and unseen variations are considered
  - Data-driven: No expert knowledge about accent is required
  
- **Practical**
  - Collected a large non-native English speech database
    - 22h of speech uttered by 96 speakers
    - among the largest such databases existing
  
- **Experimental**
  - Consistently improved performance for any type of accent
  - Largest improvement achieved: **11.9%** relative WER reduction

## Publications (Excerpt)

1. *A Statistical Lexicon for Non-Native Speech Recognition*  
Rainer Gruhn, Konstantin Markov, Satoshi Nakamura, **ICSLP** 2004
2. *Discrete HMMs for statistical pronunciation modeling*  
Rainer Gruhn, Konstantin Markov, Satoshi Nakamura, SLP 2004
3. *A multi-accent non-native English database*  
Rainer Gruhn, Tobias Cincarek, Satoshi Nakamura, ASJ 2004
4. *A Statistical Lexicon Based on HMMs*  
Rainer Gruhn, Satoshi Nakamura, IPSJ 2004
5. *Probability Sustaining Phoneme Substitution for Non-Native Speech Recognition*  
Rainer Gruhn, Konstantin Markov, Satoshi Nakamura, ASJ 2002
6. *CORBA-based Speech-to-Speech Translation System*  
Rainer Gruhn, Koji Takashima, Atsushi Nishino, Satoshi Nakamura, **ASRU** 2001
7. *A CORBA based Speech-to-Speech Translation System*  
Rainer Gruhn, Koji Takashima, Atsushi Nishino, Satoshi Nakamura, ASJ 2001
8. *Multilingual Speech Recognition with the CALLHOME Corpus*  
Rainer Gruhn, Satoshi Nakamura, ASJ 2001
9. *Cellular Phone Based Speech-To-Speech Translation System ATR-MATRIX*  
Rainer Gruhn, Harald Singer, Hajime Tsukada, Atsushi Nakamura, Masaki Naito, Atsushi Nishino, Yoshinori Sagisaka, Satoshi Nakamura, **ICSLP** 2000
10. *Towards a Cellular Phone Based Speech-To-Speech Translation Service*  
Rainer Gruhn, Satoshi Nakamura, Yoshinori Sagisaka, MSC 2000
11. *Scalar Quantization of Cepstral Parameters for Low Bandwidth Client-Server Speech Recognition Systems*  
Rainer Gruhn, Harald Singer, Yoshinori Sagisaka, ASJ 1999

Total: 46 Publications

## Patents

2001-222292 A computer with a speech processing system and program in memory

2001-222531 A computer with a program in memory that provides speech translation and feedback

2002-135642 A speech to speech translation system

2002-304392 A speech to speech translation system

2002-311983 A speech to speech translation system

2002-320037 A speech to speech translation system

2005-234504 A method for training HMM pronunciation models for speech recognition

2005-292770 A method for acoustic model generation and speech recognition

2006- 84965 A system and program for speech data collection

2006- 84966 A method and program for automatic rating of spoken speech

Total: 10 Patents, all granted by Japanese Patent Office

## Future Directions

- Applicability on native speech
  - Baseline dictionary with no pronunciation variants
  - Speech controlled services on mobile devices
- Experiments on word level → smaller units?
  - Syllables
  - N-phones
- Special states to model insertion errors
- Accent recognition

**! THANK YOU !**