Selective Training for Cost-effective Construction of Task-adapted Acoustic Models

Tobias Cincarek, D1
Acoustics and Speech Processing Lab
COE Technical Presentation
Oct 27th, 2005

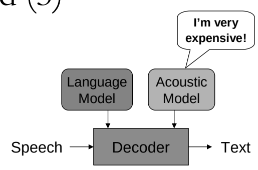
Research Background (1)

- Large number of applications for automatic speech recognition (ASR)
  - Dictation Systems
  - Speech-controlled Dialogue Systems
  - Speech-to-Speech Translation Systems
  - Human-Machine Interfaces
  - Robots
- However, there are only few commercial products which make use of ASR …

Research Background (2)

- Current state of ASR technology
  - High performance under certain conditions (clean read speech, restricted task ~95%)
  - In general performance depends on
    - acoustic conditions (noise) → signal processing
    - speaker characteristics (gender, age, accent) → ?
    - speaking style (read, spontaneous) → ?
    - recognition task (digits, news, dialogue) → ?
  - Impossible to use one ASR system for any application

Research Background (3)



- Design of an ASR system
  - Language Model
    - Grammar-based
    - Corpus-based
  - **Acoustic Model**
    - Robust model training requires a huge amount (> 50,000 utterances) of transcribed speech data
    - Collection and transcription of speech data is **very costly and time consuming!**

→ **Necessity to reduce the costs of acoustic modeling**

Research Goal and Proposed Solution

- Automatic construction of low-cost, task-adapted acoustic models for ubiquitous ASR applications
  - It impractical to collect and transcribe enough speech data for every new ASR application
- Proposed solution
  - Employ existing spoken language resources
  - Reduce effort of data collection to a small amount of task-specific speech data (< 1,000 utterances)
  - Augmentation of the task-specific data by employing **utterance-based** selective training [Cincarek et al, 2005]

Related Research

- **Active Learning** [Hakkani-Tür et al. 2002]
  Only transcribe utterances, which are difficult to recognize based on confidence measures
- **Unsupervised Learning** [Wessel et al. 2001]
  Train the acoustic model with automatically generated transcriptions (error-prone)
- **Active + Unsupervised Learning** [Riccardi et al. 2003]
- **Speaker-based Selective Training** [Yoshizawa et al. 2001]
  Train model with speech from certain speakers
- **Task-independent Acoustic Modeling** [Lefevre et al. 2005]
  Train model with speech data from multiple sources

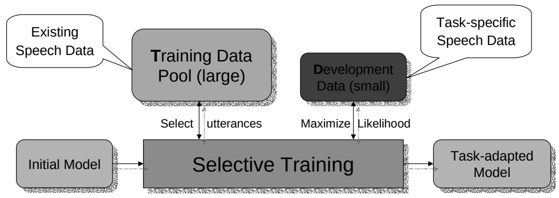
Advantages ⊕ and Shortcomings ⊖

Active and Unsupervised Learning
- ⊕ Relatively few or no costs for transcriptions
- ⊖ Still requires the collection of many speech data

Task-independent Model by Multiple Source Training
- ⊖ Requires huge amounts of transcribed speech data
- ⊕ Good performance for many recognition tasks

Proposed approach
- ⊕ Little effort for data collection and transcription
- ⊕ Model optimization by training data selection
- ⊕ Economical reuse of existing speech data

Proposed Selective Training Framework



- Conventional method
  - Use all training data T
  - Maximize likelihood given the training data T

$$P(T \mid \hat{\Theta}_{train}) > P(T \mid \hat{\Theta}_{init})$$

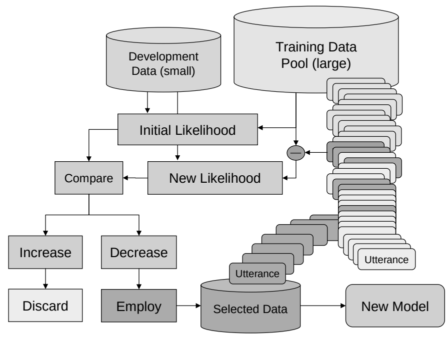
- Proposed method
  - Use a subset of T
  - Maximize likelihood given development data D

$$P(D \mid \hat{\Theta}_{select}) > P(D \mid \hat{\Theta}_{train})$$

Selective Training Algorithm (1)

- There are too many possibilities to select a subset of utterances from the data pool
- Employment of a greedy search technique
  - Start with a model trained on the whole data pool
  - Examine each utterance once for deletion
  - Discard the utterance, if likelihood increases
  - Otherwise, use the utterance for training

Selective Training Algorithm (2)



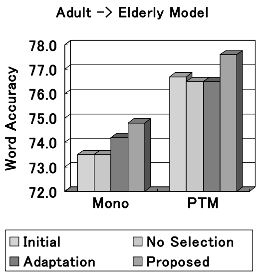
Experimental Evaluation

- Application of selective training to build
  - Elderly-adapted Model
  - Infant-adapted Model
- Analysis of the proposed algorithm's behavior
  - Influence by the development data set size
  - Comparison to standard adaptation methods
    - Maximum A Posteriori (MAP)
    - Maximum Likelihood Linear Regression (MLLR)
  - Computational complexity

Speech Data collected with the Takemaru Dialogue System

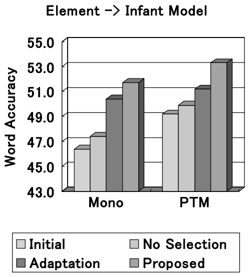
(Subjective*) Classification	Age	Number of Inputs
Total (3 years)	-	> 300,000
Transcribed (2 years)	-	> 200,000
Infants (Preschool Children)*	~6	few → 15,899
Elementary School Children*	6~12	65,767
Junior-high School Children*	12~15	21,074
Adults*	15~70	21,299
Elderly people*	70~	very few → 533

Experiment (1) Build Elderly-adapted Model with Adult Speech



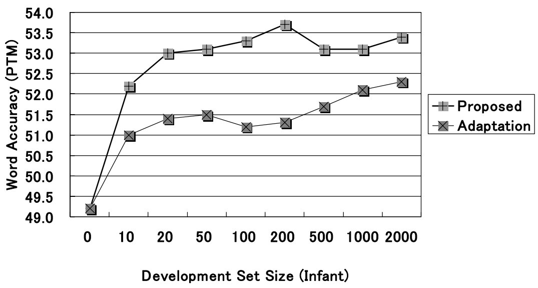
- **Initial Model**
  - Mono: 100k parameters
  - PTM: 180k parameters
- **Development Data**
  - 53 elderly utterances
- **Training Data Pool**
  - 17,874 adult utterances
  - Selection rate: 43%
- **Evaluation**
  - 400 utterances (1,609 words)
  - 20k Language Model

Experiment (2): Build Infant-adapted Model with Speech from Elementary School Children



- **Initial Model**
  - Mono: 100k parameters
  - PTM: 250k parameters
- **Development Data**
  - 100 infant utterances
- **Training Data Pool**
  - 29,776 element. utterances
  - Selection rate: 35%
- **Evaluation**
  - 1,554 utterances (5,742 words)
  - Infant Language Model

Influence of the development set size



Complexity in run time and disk space

- Fast likelihood computation with sufficient statistics (SS)
- Requires to store the SS of all training utterances
- Almost same computational requirements of model training and SS calculation
- Selection of utterances is possible within a short time
- Multiple times of speedup is possible by parallelization

One CPU

Model # Par.	# Utter.	Run time	Disk space
Mono 100k	29,776	20 m	2.5 GB
PTM 250k	29,776	3 h	4.5 GB

Conventional model training can take days or even weeks!

Conclusion

- Introduction of a practical algorithm for utterance-based selective training
- Already effective with only 10 utterances
- Enables fast selection of training utterances
- Addresses the issue of cost reduction
- Successful application of the algorithm to build an infant- and elderly-adapted model

Future Work

- Examine different selection strategies
- Apply algorithm to different databases and task adaptation problems
- Combination of selective training with unsupervised learning
  - Training or development data is untranscribed
  - Obtain utterance transcriptions automatically
  - Automatic selection of “good” training utterances
  - Comparison to active and unsupervised learning

References

[1] T. M. Kamm et al, "Robustness Aspects of Active Learning for Acoustic Modeling", Proc. of ICSLP, 2004

[2] D. Hakkani-Tür et al, "Active Learning for Automatic Speech Recognition", Proc. of ICASSP, 2002

[3] C. Huang et al, "Transformation and Combination of Hidden Markov Models for Speaker Selection Training", Proc. of ICSLP, 2004

[4] S. Yoshizawa et al, "Evaluation of Unsupervised Speaker Adaptation based on Sufficient HMM Statistics of Selected Speakers", EUROSPEECH, 2001

[5] F. Wessel et al, "Unsupervised Training of Acoustic Models for Large Vocabulary Continuous Speech Recognition", ASRU, 2001

[6] G. Riccardi et al, "Active and Unsupervised Learning for Automatic Speech Recognition", EUROSPEECH, 2003

[7] F. Lefevre, et al, "Genericity and Portability for Task-dependent Speech Recognition", Computer, Speech and Language, 2005

[8] T. Cincarek, et al, "Selective EM Training of Acoustic Models based on Sufficient Statistics of Single Utterances", ASRU, 2005 (accepted)