Acoustic Model Construction for Speech Recognition Using Unsupervised Selective Training

Tobias Cincarek, D1
Acoustics and Speech Processing Lab
Nara Institute of Science and Technology

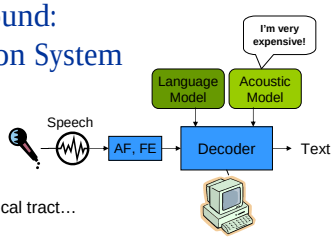
Ultimate Research Goal: Ubiquitous Automatic Speech Recognition

- Realize High-Performance Speech Recognition
  - for everybody
    - Children, Adults, Elderly people, …
  - for any kind of speaking style
    - Read speech, natural speech, spontaneous speech, …
  - under any acoustic conditions
    - Background noise, reverberation, …
  - for any kind of speech quality (transmission channel)
    - High-bandwidth speech, telephone speech, NAM, …

  i.e. there are many sources of acoustic variability …

Research Background: Speech Recognition System

- Acoustic Frontend (AF)
  - Noise reduction, …
- Feature Extraction (FE)
  - Extract information related to the human vocal tract…
- Decoder
  - Calculate most likely word sequence given the input speech
- Language Model
  - Defines, what sentences can be recognized
- **Acoustic Model**
  - Defines, which kind of speech can be recognized



The Acoustic Model (AM)…

- Consists of Hidden Markov models with Gaussian mixture densities, one for each phonetic unit (state-of-the-art)
- is a statistical model, which consists of hundreds of thousands of parameters
- **requires large amounts of training data** to reliably estimate of the model parameters
- However: Collection (recording) and preparation (labeling) of speech data is **very costly and time-consuming**
- Research Objective: **Reduce the Costs of Acoustic Modeling, e.g. save costs for labeling the speech data**

Acoustic Model Construction

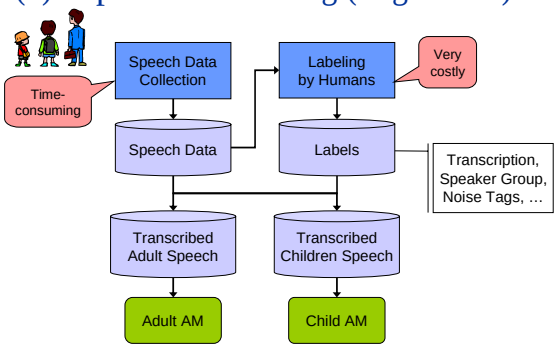
- **Speech and model have to match each other**
  - Children Speech → Children AM
  - Adult Speech → Adult AM
  - Noisy Speech → Noise-superimposed AM

  …
- **Consequence: build one model for each condition**
- **However: high costs for collecting and preparing speech data**

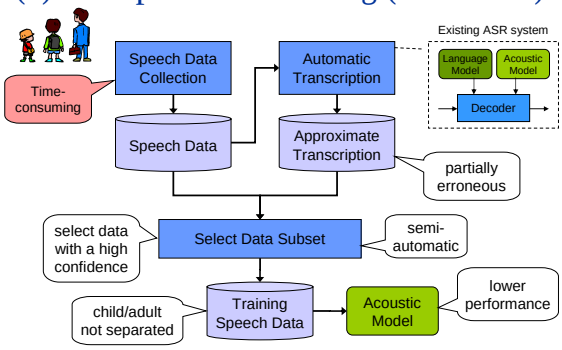
- **Several approaches to AM construction:**

Method	Manner of Learning	Selection criterion	Labeling	Costs & Performance
(1)	Supervised	None	All	High & High
(2)	Unsupervised	"Confidence"	None	Low & Medium
(3)	Active/Superv.	"Confidence"	Partial	Medium & High
Proposed	**Selective/Uns.**	**Likelihood**	**Minimum**	**Low & High?**

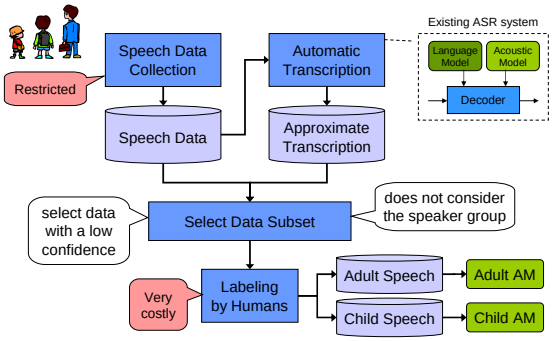
(1) Supervised Training (High-Cost)



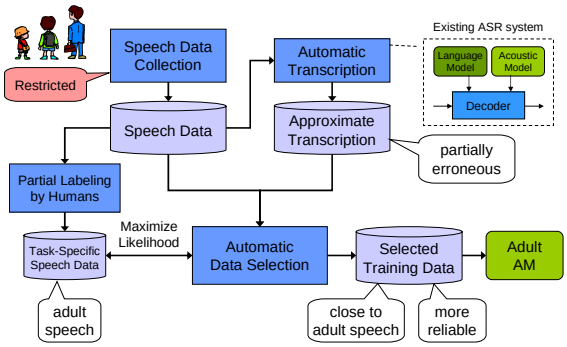
(2) Unsupervised Training (Low-Cost)



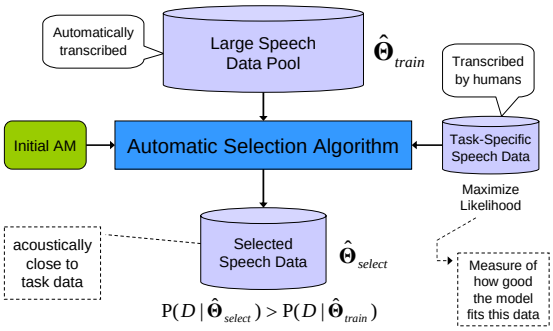
(3) Active Learning (Medium Costs)



Unsupervised Selective Training (Low-Cost)

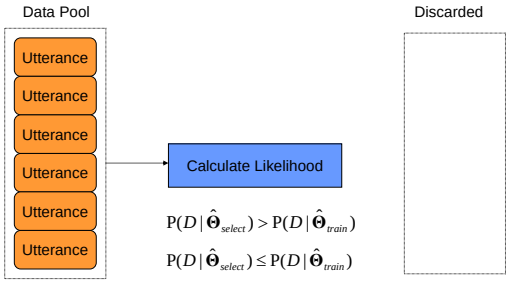


Automatic Speech Data Selection Method



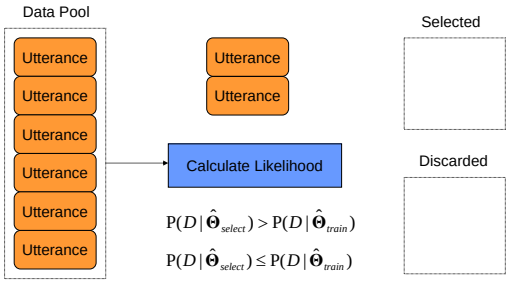
Selection Algorithm (A)

- Successive Deletion and Addition



Selection Algorithm (B)

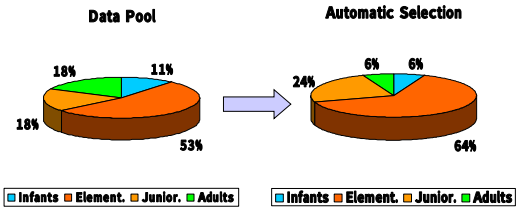
- Independent Deletion of Single Utterances



Experimental Evaluation: AM construction for adults and for children

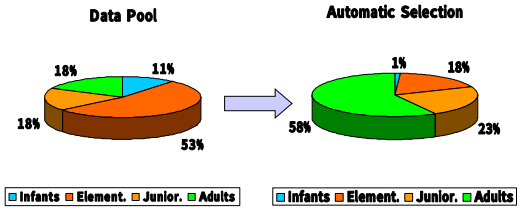
Initial Model	**Adult speech, read speech** Newspaper texts (JNAS database)
Unlabeled Data Pool	**Spontaneous speech from various speakers** Collected with the **Takemaru dialogue system** within the period: 2002/11/08 – 2004/08/18 **89,217 utterances** (only valid speech inputs)
Labeled Task Data	Set 1: adult speech, **1000** utterances (male:female=1:1) Set 2: children speech, **1000** utterances (age balanced)
Evaluation Data Sets	Set 1: adult speech, 476 utterances (2,025 words) Set 2: children speech, 797 utterances (2,795 words)
Language Model	Separate model for adults and children Dictionary contains more than 40,000 words (morphemes)

Selection Result: Children AM



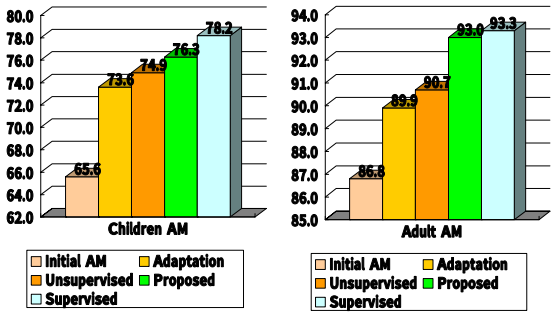
- 42% from the initial data pool are selected
- 88% of the selected data are from children

Selection Result: Adult AM



- 23% from the initial data pool are selected
- 58% of the selected data are from adults

Result of Recognition Experiments



Summary and Future Work

- Framework for acoustic model construction
  - "Unsupervised Selective Training"
  - Less costs for data labeling, but high performance
- Experimental Evaluation
  - Selection of the desired training data is effective
  - Almost maximum performance can be reached
  - Better than conventional unsupervised training
- Future Work
  - Evaluation including non-speech inputs
  - Combining active learning and selective training

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] F. Wessel et al, "Unsupervised Training of Acoustic Models for Large Vocabulary Continuous Speech Recognition", ASRU, 2001.
[4] T. Kemp et al, "Unsupervised Training of a Speech Recognizer: Recent Experiments", EUROSPEECH, 1999.
[5] G. Riccardi et al, "Active and Unsupervised Learning for Automatic Speech Recognition", EUROSPEECH, 2003.
[6] C. Leggetter et al, "Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models", Computer Speech and Language, 1995.
[7] ティンツァレク・トビアス 他, "タスク依存音響モデルのための発話レベでの選択学習法", 信学技法, NLC2005-102, SP2005-135(2005-12).
[8] T. Cincarek, et al, "Selective EM Training of Acoustic Models based on Sufficient Statistics of Single Utterances", ASRU, 2005.