Applying Non-Audible Murmur (NAM) microphone in speech recognition systems

Panikos Heracleous
COE-Postdoctoral Fellow
Speech and Acoustics Processing Laboratory
NAIST

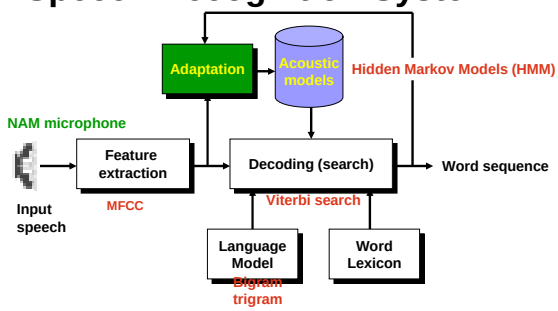
Preface

- Automatic Speech Recognition
- Non-Audible Murmur (NAM) definition
  - Scenario – Applications
  - Similar approaches
- NAM recognition based on adaptation
  - Maximum Likelihood Linear Regression (MLLR)
  - Maximum A Posteriori (MAP) adaptation
  - Combination of MLLR and MAP
  - Experiments and Results

Automatic Speech Recognition

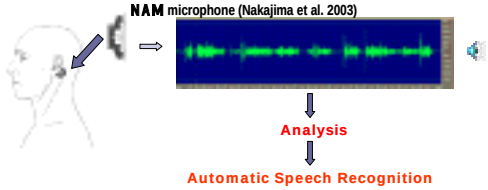
- Goal
  - Convert a speech signal into a text message
  - Device-, environment-, speaker-independent
- Applications
  - Automation of operator-based tasks
    - Dictation
    - Customer care, help
    - Etc.
- Speech receiver
  - Close-talking microphone
  - Head-set
  - Microphone array
  - **NAM microphone**

Speech Recognition System

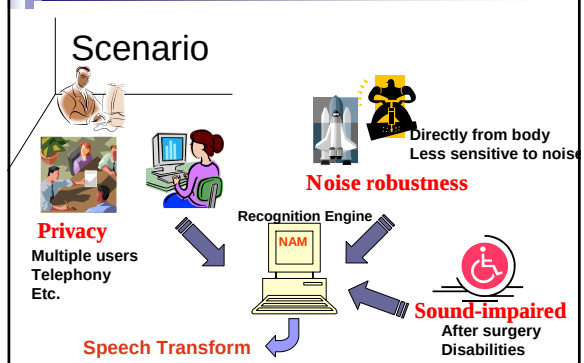


Non-Audible Murmur (NAM)

DEFINITION
Non-Audible Murmur (NAM) is murmur uttered very quietly, which cannot be heard by other listeners near the talker.

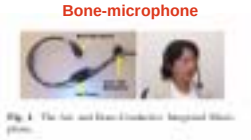


Scenario



Similar approaches

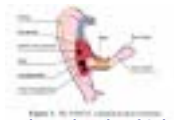
Bone-microphone



Microsoft [ASRU2003]

-Noise-robustness
-Unable to recognize normal speech
-Unable to recognize inaudible murmur
-Complementary usage

PARAT



Norwegian University of Science and Technology [ASRU2003]

-Noise-robustness
-Unable to recognize inaudible murmur
-Practical difficulties (communication)

The NAM microphone can be used for normal speech, and inaudible murmur recognition. Preliminary experiments show noise-robustness. No additional techniques are required.

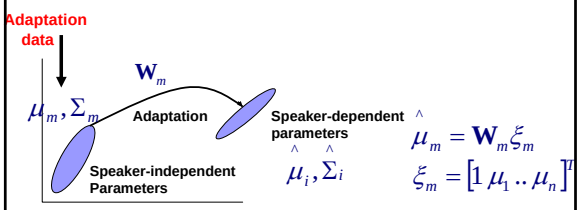
NAM recognition using adaptation

- Adaptation techniques
  - Maximum Likelihood Linear Regression (MLLR) adaptation
  - Maximum A Posteriori (MAP) adaptation
- Advantages
  - Require only a small amount of data
  - Speaker and environment adaptation
  - High performance

MLLR adaptation

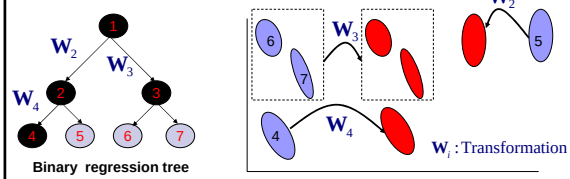
- **Aim**

  Obtain a set of transformations matrices for the model parameters that maximizes the likelihood of the adaptation data. For the solution, the Expectation-Maximization (EM) technique is used.



MLLR: Transforms sharing

- Acoustically similar components are transformed together. Splitting algorithm using Euclidean distance.
  - Less data for accurate estimation
  - Unseen parameters can be adapted



MAP adaptation

- Informative priors are used
  - Speaker-independent parameters
- Requires more adaptation data, than MLLR
- Together with MLLR more effective

$$\hat{\mu}_{jm} = \frac{N_{jm}}{N_{jm}+\tau}\bar{\mu}_{jm} + \frac{\tau}{N_{jm}+\tau}\mu_{jm}$$

j : state, m : mixture, τ : weight

$\hat{\mu}_{jm}$: Updated mean

μ_{jm} : Speaker - independent mean

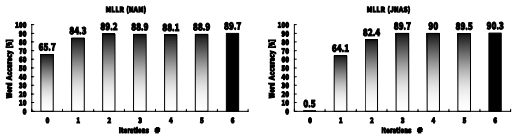
$\bar{\mu}_{jm}$: Obtained from adaptation data

N_{jm} : Occupation likelihood (forward - backward

Experiments

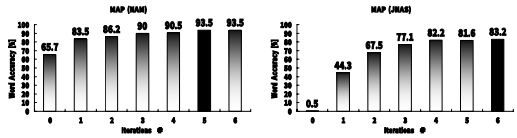
- Experimental conditions
  - Initial models
    - Normal-speech PTM HMMs (JNAS)
    - NAM-speech PTM HMMs (1800 utterances)
  - Adaptation techniques
    - MLLR
    - MAP
    - MLLR+MAP
  - Adaptation data
    - 710 NAM utterances – Male speaker
  - Test set
    - 24 NAM uuterances

Results using MLLR adaptation



-Higher initial performance using SI NAM initial HMMs
-After few iterations almost equal performance is achieved
-The initial models do not have serious effect
-Due to the components grouping, unseen components are also transformed

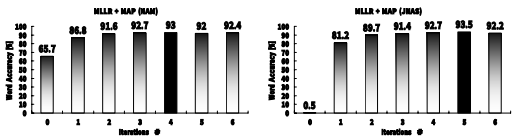
Results using MAP adaptation



-Higher initial performance using SI NAM initial HMMs
-No regression tree is used – Unseen components are not transformed
-Initial models have significant effect
-High performance **93.5%** using NAM-speech initial models

! Iterative MAP performs much better that single-iteration MAP.
MAP formula very similar to EM formula. Only in priors are different

Results using MLLR + MAP



-Simultaneous MLLR and MAP adaptation (one pass)
-Transformed components by MLLR are used as priors
-Initial models do have significant effect
-High performance **93.5%** - Same as MAP adaptation with NAM initial models

Discussion

- MLLR adaptation
  - Requires less amount of data
  - Initial models do not have serious effect
    - Components grouping – Multiple iterations
  - Critical factor: Number of classes (32 or 128)
- MAP adaptation
  - Performs better when more data are available
  - Initial models are very important
  - Critical factor: Scaling factor
- MLLR and MAP adaptation
  - Very useful when there are not NAM initial models
  - Equal performance with MAP

Conclusion

- Brief introduction of Non-Audible Murmur
- Automatic Speech Recognition
  - MLLR adaptation
  - MAP adaptation
  - MLLR + MAP
- **93.5%** word accuracy
  - 20k vocabulary
  - Only 700 utterances
  - Very promising result
  - Comparable to normal-speech recognition
- On-going work
  - Silicon NAM microphone

Thank you!