Distant Speech Recognition for Home Automation: Preliminary Experimental Results in a Smart Home

Benjamin Lecouteux, Michel Vacher, François Portet

Laboratory of Informatics of Grenoble - UMR CNRS/UJF/INP 5217
Summary

- ANR project between 2 academics: LIG, Esigetel, and 3 SMEs: Theoris, Camera Contact, Technosens.

- Goals:
  - Assistance via natural man-machine interaction (voice and tactile command)
  - Social inclusion
  - Detection of distress situations
  - Less intrusive: no camera use

⇒ The person will be able to pilot their environment at any time in the most natural way possible
Automatic Speech Recognition system in Sweet-home

**Issues**

- Presence of echo and reverberation
- Signal to Noise Ratio (SNR)
- Distance between speaker and microphones
- Potential audio pollution (TV, radio, vacuum...)

**Record configuration**

- Several microphones set in each room
- Distance between speaker and microphones is at min 2/3 meters
**Smart Home and Corpora**

- **DOMUS Smart Home**
  - CTL - Carnot LSI
  - MULTICOM Team - LIG

- **21 recorded speakers** in constrained conditions
  - (7 womens, 14 mens)

- **2 record phases**:
  - **Phase 1**: 40 sentences, activity scenario including phone talk simulation
  - **Phase 2**: 44 read sentences in 3 rooms and 3 status (396 sentences) : vacuum, radio, everything off
First experiments

**Inventory**
- Data without external noise are selected
- Test of two ASR systems
- Classical adaptation of acoustic and language models

**Training and development data**
- Phase 1 sentences (manually annotated)
- Acoustic models are adapted for each speaker
- About 2 minutes of speech per speaker

**Test data**
- Phase 2 sentences (manually annotated)
ASR systems

Sphinx
- CMU
- Beam search, Sphinx 3.3
- Contextual model, 3 or 5 states HMM (120 hours)
- MFCC

Speeral
- LIA
- A* on a phone lattice
- Contextual model, 3 states HMM (100 hours)
- PLP
### Task

#### Configuration

- Working in about 1xReal Time
- Same language model for the two ASR systems

#### Keyword detection

- Domotic orders:
  - "Allumer la lumière", "Monter la température"
- Distress situation:
  - "Appeler le docteur", "Aidez moi"
Reducing linguistic variability

Interpolation between generic model and a targeted model

- **Generic Model (GM):** Trained on journal papers Le Monde + Gigaword
- **Specialist model (SM):** Trained on read sentences
- GM alone: without knowledge about environment
- SM alone: some degradation when speaker deviates
- Best configuration: GM 10%, SM 90%
Baseline WER on the 2 best SNR channels

**Sphinx** (57% average), **Speeral** (35% average)
Speaker adaptation

- Maximum A Posteriori:
  - **Sphinx** (62% WER), **Speeral** (29% WER)

- Maximum Linear Likelihood Regression:
  - **Sphinx** (28% WER), **Speeral** (13% WER)
Combination : ROVER

Baseline

- Combination of 4 streams (2 ASR systems, 2 channels)
- High computing cost: 4 decoding pass
- Show the complementarity between streams
- Final result: 10% WER
Combination : DDA

Principle

- The second stream is driven by the first one
- The second pass is very fast
- Three streams are requested for ROVER, only two for DDA
Combination : DDA, results

- Mean average : 12.35% WER
- About the mean observed on the two streams
- Good approach to combine informations
Principle

- The first pass selects sentences in order to drive the second pass
Combination: DDA-2, results

Average result: 7.9% WER

The best results
Detection of predefined sentences:

GDR (good detection results)

- **Baseline**: 83.1%
- **ROVER**: 88.2%
- **DDA1**: 87.4%
- **DDA2**: 92.5%
Conclusion

- Record of a **Distant Speech Corpus**
- Experiments with two ASR systems
- Best “baseline“ : MLLR + Specialized language model
- **Multichannel decoding** : DDA
  - Allows to use only 2 streams (best SNR)
  - Possibility to introduce *a priori* sentences in the decoding pass
  - Best results than the ROVER

Perspectives

- Use source separation (ICA)
- Factor analysis (FA)
Thank you for your attention.
For Further Reading

M. Vacher, A. Fleury, F. Portet, J.-F. Serignat, N. Noury
Complete Sound and Speech Recognition System for Health Smart Homes: Application to the Recognition of Activities of Daily Living,

M. Vacher, F. Portet, A. Fleury and N. Noury
Development of Audio Sensing Technology for Ambient Assisted Living: Applications and Challenges,

A. Fleury, M. Vacher and N. Noury
SVM-Based Multi-Modal Classification of Activities of Daily Living in Health Smart Homes: Sensors, Algorithms and First Experimental Results,