# Lipreading Approach for Isolated Digits Recognition under Whisper and Neutral Speech


**Fei Tao and Carlos Busso**

**Multimodal Signal Processing (MSP) Laboratory**
Erik Jonsson School of Engineering & Computer Science
University of Texas at Dallas
Richardson, Texas 75080, U.S.A.



Interspeech 2014 Singapore
Celebrating the Diversity of Spoken Languages
15th Annual Conference of the International Speech Communication Association, 14 - 18 Sept 2014, MAX Atria @ Singapore EXPO


## Abstract


**Background:**

- Whisper is a speech production mode with low energy and lack of vocal cord vibrations
- The acoustic differences degrade the performance of ASR

**Proposed Solution:**

- Visual features are applied to improve whisper recognition
- Facial features are less affected by whisper speech [Tran et al, 2013]
- Previous work with one subject proved the concept [Fan et al.,2011]


**Word accuracy using HMM (Fan et al., 2011)**

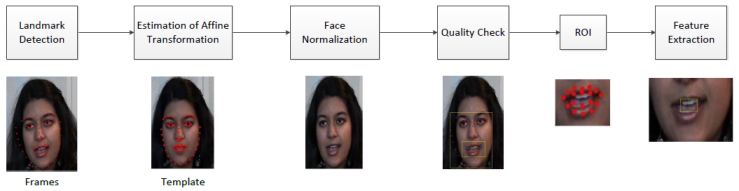
| stream | training | test | Word Accuracy |
|---|---|---|---|
| audio data | neutral | neutral | 98.7% |
| audio data | whisper | whisper | 83.3% |
| audio data | neutral | whisper | 42.7% |
| video data | neutral | neutral | 70.7% |
| video data | whisper | whisper | 68.0% |
| video data | neutral | whisper | 54.7% |
| **combined (best)** | **neutral** | **whisper** | **79.7%** |

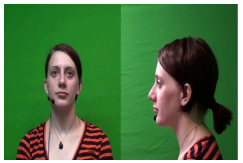
## Data Preparation

**Audiovisual Whisper (AVW) Corpus**

- 40 Speakers (20 male, 20 female)
- Isolated digits, read sentence and spontaneous
  - Recorded with whisper and neutral speech
  - Include data from audio and video channels
- Study relies on isolated digits
  - 1-9, "zero" and "oh"

**Video Processing and Feature Extraction**



- We identify 66 facial landmarks using CSIRO face analysis SDK
- We normalize head pose using an affine transformation (from facial landmarks)
- Quality check: second facial landmark detector (mouth corners and nose tip)
- From ROI, we estimate 25 DCT plus 5 geometric features
  - 90D feature vector [25-DCT + 5D-distance], plus Δ, and ΔΔ

## Experimental Evaluation

**Recognition Task Setting:**

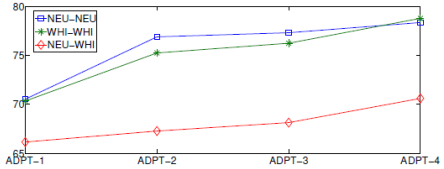
- HMM is used for the recognition task (left-to right, 10 states)
- Conditions (leave-one-out cross validation)
  - Speaker independent (SI)
  - Speaker dependent (SD)
- We explore adaptation schemes (MAP +MLLR)

**Results:**

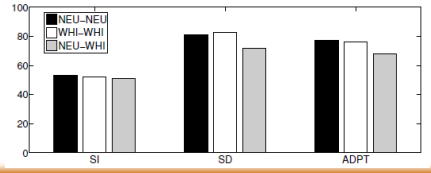
- SD: accuracy in matched conditions is above 80%
- ADPT: Adaptation helps in reducing gap between SD and SI

| Train | Test | SI(%) | SD(%) | ADPT(%) |
|---|---|---|---|---|
| Neutral | Neutral | 52.93 | 80.78 | 77.31 |
| Whisper | Whisper | 52.34 | 82.64 | 76.24 |
| Neutral | Whisper | 50.87 | 71.85 | 68.14 |

- ADPT: Accuracy versus number of samples per digit




- ADPT: Results with three samples per digits




## Discussion

**Conclusions:**

- HMM approach with geometric and appearance based features
- Lipreading approach is a feasible alternative to improve the performance of whisper speech recognition.

**Future Directions**

- Fuse the proposed system with acoustic features (accuracy ~83%)
- Explore the use of phoneme/viseme models to extend to large vocabulary continuous speech recognition

**References:**


X. Fan, C. Busso, and J.H.L. Hansen, "Audio-visual isolated digit recognition for whispered speech," in European Signal Processing Conference (EUSIPCO-2011), Barcelona, Spain, August-September 2011, pp. 1500–1503.

T. Tran, S. Mariooryad, and C. Busso, "Audiovisual corpus to analyze whisper speech," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2013), Van- couver, BC, Canada, May 2013, pp. 8101–8105.



**This work was funded by NSF (IIS-1217104) and Samsung**