Audiovisual Corpus to Analyze Whisper Speech

Tam Tran, Soroosh Mariooryad and Carlos Busso

Multimodal Signal Processing (MSP) Laboratory
Erik Jonsson School of Engineering & Computer Science
University of Texas at Dallas
Richardson, Texas 75083, U.S.A.

ICASSP 2013
Vancouver Convention & Exhibition Centre
May 26 - 31, 2013 • Vancouver, Canada

Motivation

Improve ASR Robustness

- **Security** – Protect privacy in public
- **Robustness** – ASR systems cannot detect whisper easily
- **Audiovisual** – Add visual modality to improve recognition

Previous Works

- We showed improvement of 37% by including visual modality
- Limited by only using one subject

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%**

Goals:

- Create a corpus to study audiovisual whisper speech
- Identify changes on acoustic/facial features in whisper speech

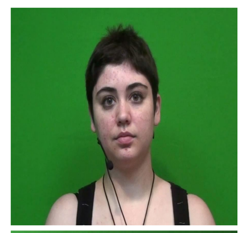
Audiovisual Whisper (AVW) Corpus

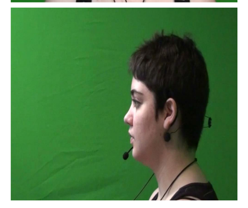
Description

- 25 Speakers (13 male, 12 female)
  - Analysis uses only 11 speakers
- Read speech (~ 20 min per subject)
  - Part 1 – 60 Neutral/ 60 Whisper TIMIT Sentences
  - Part 2 – 11 digits (1-9, zero, oh) 10x per mode/digit
- Spontaneous speech (~ 10 min per subject)
  - Part 3 – 10 questions (5 per mode) ~ 45 sec each

Equipment (sound booth):

- Audio – 48 KHz mono WAV
- Video –1440x1080 pixels, 29.7fps
  - Frontal & side cameras
  - Two LED light panels





Speech features are extracted with openSMILE and Praat

Spectral LLDs	
Rfilt AudSpec [X]	RASTA-style filtered auditory spectrum bands 1-26 (0-8kHz)
MFCC [X]	Mel-frequency cepstral coefficients 1-12
Fband [F1-F2]	Spectral energy 25-650Hz, 1000-4000Hz
Spectral roll-off [X]	Spectral roll-off point 0.25, 0.50, 0.75, 0.90
Spectral [statistic]	Spectral flux, entropy, variance, skewness, kurtosis, slope
Formants [X}	*Spectral Formants* 1-5 (extracted with Praat)
Prosody LLDs	
AudSpec L1	Auditory spectrum L1-norm (loudness)
Rfilt AudSpec L1	RASTA-style filtered auditory spectrum L1-norm
RMS Energy	RMS Energy
ZCR	Zero-crossing rate
F0	Fundamental frequency
Voicing prob	Voicing probability
Voice Quality LLDs	
Jitter	Frame-to-frame F0 deviations
Δ Jitter	Frame-to-frame Jitter deviations
Shimmer	Frame-to-frame amplitude deviations

Extracted facial features using CERT

Action unit	Description	Example Image
AU 10	Lip Raise	
AU 12	Lip Corner Pull	
AU 15	Lip Corner Depressor	
AU 18	Lip Pucker	
AU 20	Lip Stretch	
AU 23	Lip Tightener	
AU 24	Lip Presser	
AU 25	Lips Part	
AU 26	Jaw Drop	
AU 28	Lips Suck	
Lip Features		
Lip spreading	Horizontal Lip Spreading	

Source: http://www.cs.cmu.edu/~face/facs.htm

Feature Analysis: Neutral Versus Whisper

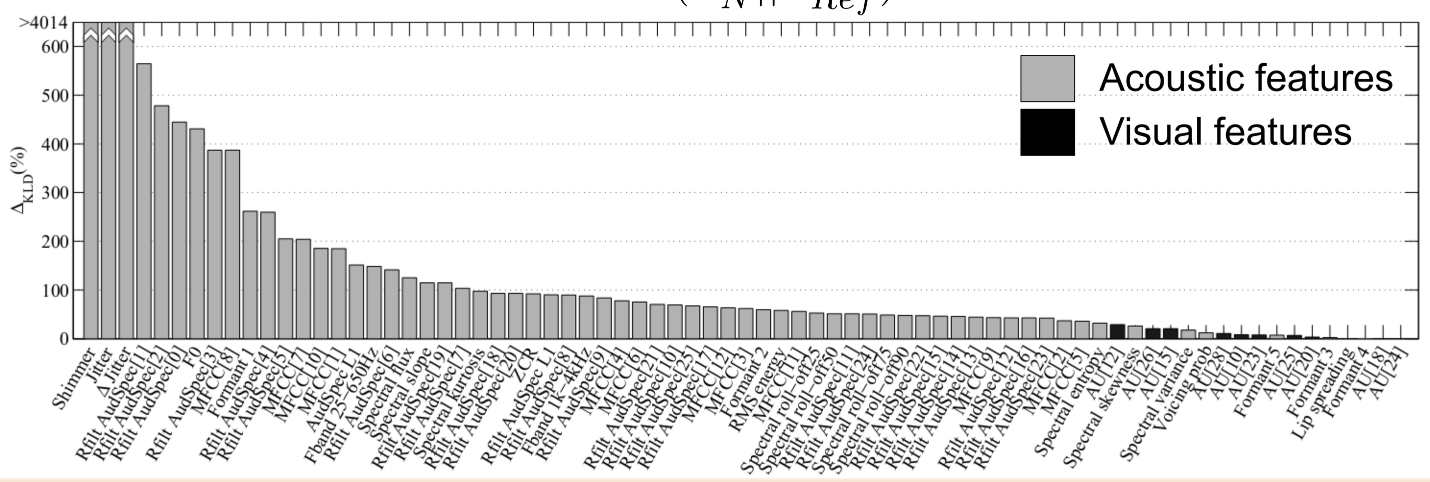
Kullback-Leibler Divergence Analysis:

- Goal: quantify deviation from neutral speech
- Distribution determined using K-means algorithm (K=40)

$$KLD(P||Q) = \sum_i \ln(\frac{P(i)}{Q(i)})P(i)$$

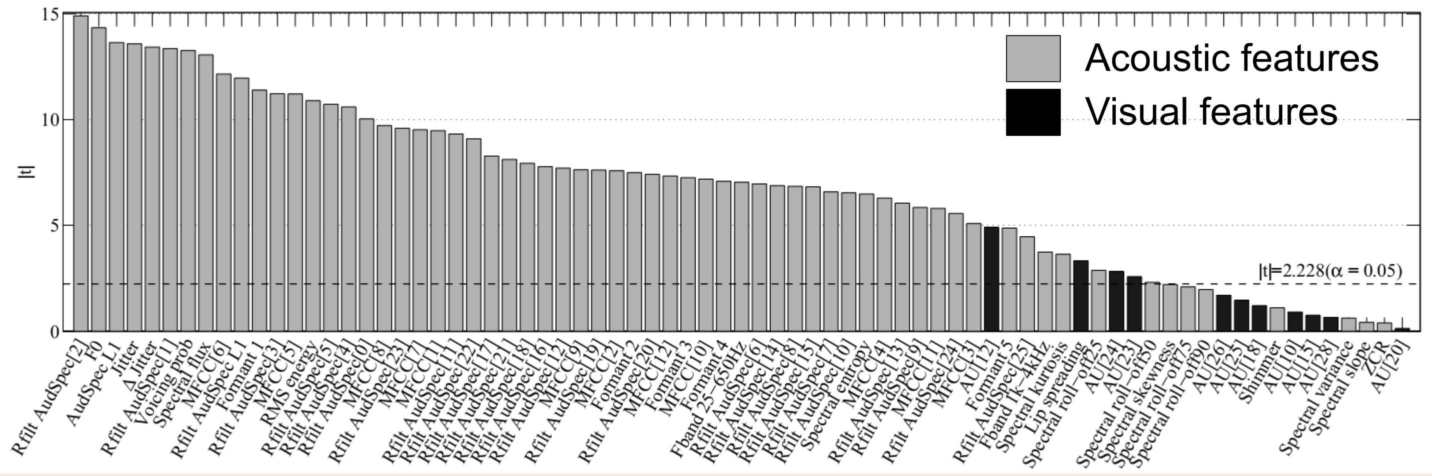
- Data partitioned in two: reference and testing (cross-validation)
- Reference partition, P^f_{Ref}, uses only neutral speech condition

$$\Delta^f_{KLD} = \frac{KLD(P^f_W||P^f_{Ref}) - KLD(P^f_N||P^f_{Ref})}{KLD(P^f_N||P^f_{Ref})} \times 100$$



Statistical Analysis

- Goal: analyze whether the differences are statistically significant
  - Only digits data (11 digits), which is used as the matched condition
- Matched pair two-tailed *t*-test
- Values above threshold have statistically significant differences
  - Acoustic features present differences between speech mode
  - Four visual features present differences between speech mode



Discussion

Conclusions:

- Visual features are less affected by changes between neutral and whisper conditions
- Orofacial area provide whisper-invariant features that can improve ASR performance

Future Directions

- Increase size of corpus (40 speakers)
  - We expect to make the corpus available to the community
- Identify other facial features (DCT, Gabor filter, HOG)
- Identify suitable graphical models to train audiovisual ASR

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.

This work was funded by NSF (IIS-1217183) and Samsung