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.

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

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

Audiospectral Features

Kullback-Leibler Divergence Analysis:
• Goal: quantify deviation from neutral speech
• Distribution determined using K-means algorithm (K=40)

\[ KLD(P||Q) = \sum_{i} P(i) \log \left( \frac{P(i)}{Q(i)} \right) \]

• Data partitioned in two: reference and testing (cross-validation)
• Reference partition, \( P_{ref} \), uses only neutral speech condition

\[ \Delta_{KLD} = \frac{KLD(P_{ref}||P_{test}) - KLD(P_{test}||P_{ref})}{KLD(P_{test}||P_{ref})} \times 100 \]

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

Visual features are less affected by changes between neutral and whisper conditions

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:

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