Inversion from Audiovisual Speech to Articulatory Information by Exploiting Multimodal Data

N. Katsamanis, T. Roussos, P. Maragos, M. Aron, M.-O. Berger

Computer Vision, Speech Communication and Signal Processing Group
http://cvsp.cs.ntua.gr
Audiovisual Speech Inversion

\[ Y_a \]

\[ Y_v \]

\[ X \]

\[ \ldots \longrightarrow C_n \longrightarrow C_{n+1} \longrightarrow C_{n+2} \longrightarrow C_{n+3} \longrightarrow \ldots \]
Audiovisual Speech Inversion

\( Y_a \)

\( Y_v \)

phoneme

\( \hat{X} \)
Audiovisual Speech Inversion

$Y_a$

$Y_v$

$X$

http://cvsp.cs.ntua.gr

ISSP 2008, Strasbourg
Audiovisual Speech Inversion

\( Y_a \)

\( Y_v \)

\( X \)

\( y_a^n, y_v^n, y_a^{n+1}, y_v^{n+1}, y_a^{n+2}, y_v^{n+2}, y_a^{n+3}, y_v^{n+3} \)

\( X_n, X_{n+1}, X_{n+2}, X_{n+3} \)

\( C_n, C_{n+1}, C_{n+2}, C_{n+3} \)
Audiovisual Speech Inversion

$Y_a$

\[ \cdots \rightarrow C_1 \rightarrow C_2 \rightarrow C_3 \rightarrow \cdots \]

phoneme

$\hat{X}_a$

\[ \cdots \rightarrow \]

Fusion

$\hat{X}$

\[ \cdots \rightarrow D_2 \rightarrow D_1 \rightarrow \cdots \]

viseme

$\hat{X}_v$

$Y_v$

http://cvsp.cs.ntua.gr
Vocal Tract Representation

Points on articulators
(EMA, X-ray Microbeam)

Mermelstein, 1972
(3D) Birkholz et al., 2006

Maeda, 1990
(3D) Badin et al., 1998

http://cvsp.cs.ntua.gr
Articulatory Model Building (Xrays and Grid)

- 3-part semipolar grid (30 gridlines in total)
- approx. 700 VT contours (IPS data)
- Vocal tract semi-automatically annotated (IPS & LORIA)
Articulatory Model Building (Principal Component Analysis)

96% of the variance is explained by the 6-parameter model
Audiovisual Speech Inversion

Articulatory Parameter Extraction

Articulatory Model Training
Articulatory Parameter Extraction
Registration of Multimodal Data

Head Coordinate System

Face markers
Forehead
Outer VT wall
EM sensors
US img area

CAM 1
CAM 2
Articulatory Parameter Extraction
Registration of Multimodal Data

Head Coordinate System

- Face markers
- Forehead
- Outer VT wall
- EM sensors
- US img area

CAM 1

CAM 2
Articulatory Parameter Extraction
Ultrasound Tongue Tracking

Denoising
Aron et al.
2008

Detection of maximum image intensity on each gridline

http://cvsp.cs.ntua.gr
Articulatory Parameter Extraction
Model Fitting

6 parameters are specified for each US frame
Audiovisual Speech Representation

optical flow
head movement compensation

CAM 1
CAM 2

MFCCs
Inversion Experiments and Results

- Reference shapes are in dashed lines
- Corpus: VCVs, VVs and a number of phonetically balanced French sentences (38 phonemes)
- 6 minutes of recordings (US+EM+SV+A)
- 10% testing data
- US @ 25Hz, SV @ 120Hz, EM @ 40Hz, A @ 44kHz
Conclusions

Audiovisual Speech Inversion

Articulatory Parameter Extraction

Articulatory Model Training
Articulatory Speech Synthesis

Audiovisual Speech Inversion

Articulatory Parameter Extraction

Articulatory Model Training
Articulatory Speech Synthesis

Audiovisual Speech Inversion

Articulatory Parameter Extraction

Articulatory Model Training
Acknowledgements

- We would like to thank
  - Yves Laprie and Erwan Kerrien at LORIA for their help in the acquisition and distribution of the articulatory data
  - Shinji Maeda from ENST, Jean Schoentgen from ULB and all other ASPI participants for many fruitful discussions
  - Fabrice Hirsch and Rudolph Sock for enduring the data acquisition process
Thank you!