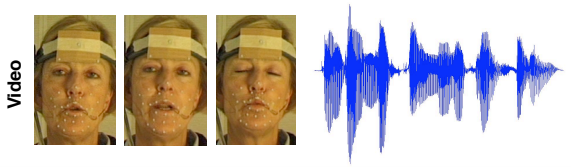
Audiovisual-to-articulatory speech inversion using Active Appearance Models for the face and Hidden Markov Models for the dynamics

Athanassios Katsamanis, George Papandreou, Petros Maragos

School of E.C.E., National Technical University of Athens, Athens 15773, Greece

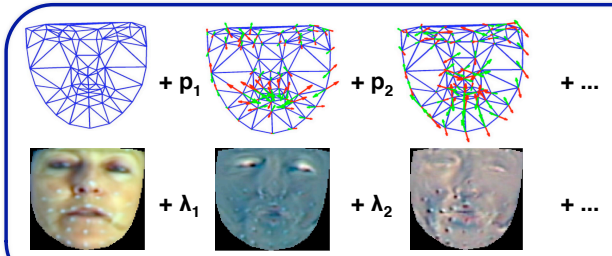
Active Appearance Modeling

❑Automatic visual feature extraction from frontal view, without markers
❑Account for both facial shape and appearance variations



Switching Linear Modeling

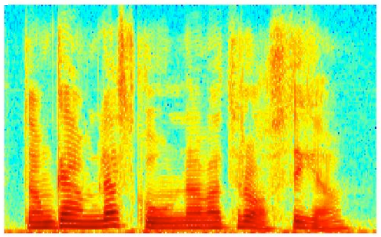
❑Complex audiovisual-articulatory interactions are captured in a piece-wise manner
❑Constituent mappings are built using Canonical Correlation Analysis



Speech inversion ?

❑Recover vocal tract geometry from the speech signal and speaker's face
❑Applications in Speech Therapy, Language Tutoring, Speech Coding,

y_v y_a

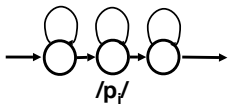


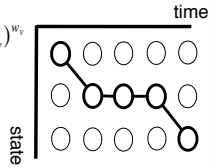
spectral characteristics/MFCC

Switching Process

❑Switching is governed by a hidden Markov process
❑Phoneme/Viseme HMMs are trained using Baum-Welch
❑State sequence determined using the Viterbi algorithm

$$b(c \mid y) \propto N(y_a;\mu_{c,a},\Sigma_{c,a})^{w_a} N(y_v;\mu_{c,v},\Sigma_{c,v})^{w_v}$$



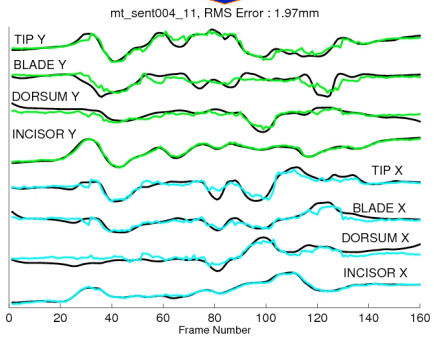


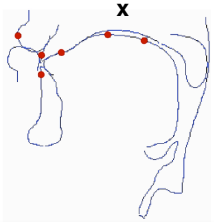
Audiovisual Fusion

❑Audio and visual mapping switching processes can interact at various synchronization levels

Features	Level	Type	States	RMS(mm)	ρ_{xx}
Audio (A)	Phoneme	HMM	2	2.56	0.60
Qualisys (QS)	Phoneme	HMM	2	2.30	0.65
A-QS	Phoneme	HMM	3	2.24	0.66
A-QS	Phon.-Phon.	HMM+LF	2-2	2.02	0.71
A-QS	Phon.-Viseme	HMM+LF	2-2	1.99	0.72
A-QS	**Phoneme**	**MS-HMM**	**2**	**1.95**	**0.74**

LF: Late Fusion, MS: Multistream





Maximum A Posteriori Estimation

Time t, state i:

$$y_t = A_i x_t + \varepsilon_t$$

Maximum A Posteriori articulatory parameter estimate:

$$\hat{x} = (\sigma_x^{-1} + A_i^T Q_i^{-1} A_i)^{-1}(\sigma_x^{-1}\bar{x} + A_i^T Q_i^{-1} y)$$

Q_i is the covariance of the approximation error

The prior of x is considered to be Gaussian determined at the training phase

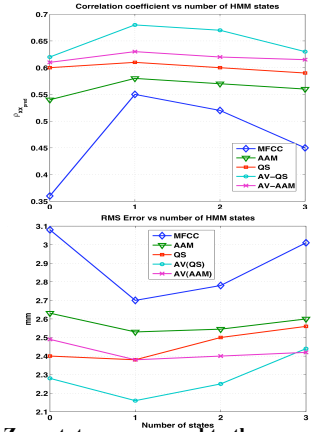
$$x_{prior} \sim N(\bar{x},\sigma_x)$$

Canonical Correlation Analysis

❑Analyze the co-variability of audiovisual and articulatory data
❑Determine the linear mappings by only keeping the first canonical correlation directions

Evaluation

❑Exploiting jointly audio and visual information in the proposed scheme clearly improves performance relative to either audio or visual- only estimation.



Zero states correspond to the case of a global linear model.

References

S. Dupont and J. Luettin, "Audio-visual speech modeling for continuous speech recognition," IEEE Tr. Multimedia, vol. 2, no. 3, pp. 141–151, 2000.

O. Engwall, "Introducing visual cues in acoustic-to-articulatory inversion," in INTERSPEECH, 2005, pp. 3205–3208.

O. Engwall and J. Beskow, "Resynthesis of 3D tongue movements from facial data," in EUROSPEECH, 2003.

S. Hiroya and M. Honda, "Estimation of articulatory movements from speech acoustics using an HMM-based speech production model," IEEE TSAP, vol. 12, no. 2, pp. 175–185, March 2004.

J. Jiang, A. Alwan, P. A. Keating, E. T. Auer Jr., and L. E. Bernstein, "On the relationship between face movements, tongue movements, and speech acoustics," EURASIP Journal on Applied Signal Processing, vol. 11, pp. 1174–1188, 2002.

H. Kjellstrom, O. Engwall, and O. Balter, "Reconstructing tongue movements from audio and video," in Interspeech, 2006, pp. 2238–2241.

H. Yehia, P. Rubin, and E. Vatikiotis-Bateson, "Quantitative association of vocal-tract and facial behavior," Sp. Comm., vol. 26, pp. 23–43, 1998.

K. Richmond, S. King, and P. Taylor, "Modelling the uncertainty in recovering articulation from acoustics," Computer Speech and Language, vol. 17, pp. 153–172, 2003.

A. Katsamanis, G. Papandreou, and P. Maragos, "Audiovisual-to-articulatory speech inversion using HMMs," in Proceedings of IEEE Int'l Workshop on Multimedia Signal Processing (MMSP 2007).

T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," IEEETrans. on Pattern Analysis and Machine Intelligence, vol. 23, no. 6, pp. 681–685, 2001.

P. Viola and M.J. Jones, "Rapid object detection using a boosted cascade of simple features," in Proc. IEEE Conf. on Comp. Vision and Pat. Recog., 2001,vol. I, pp. 511–518.