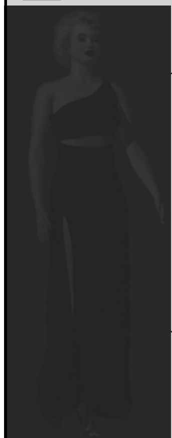
MIRALab
Where Research means Creativity

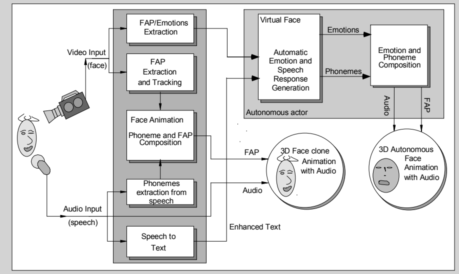


Speech Animation

Nadia Magnenat-Thalmann
MIRALab, University of Geneva
thalmann@miralab.unige.ch

Where do we stand today?

Face to Virtual Face Communication



- What are technologies?
- What is the progress?
- What is still absent?
- What is the future?

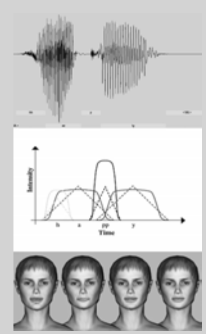
Magnenat Thalmann N., Kalra P., Pandzic I.S., **"Direct Face-to-Face Communication Between Real and Virtual Humans"**, International Journal of Information Technology, Vol.1, No.2, 1995, pp.145-157.

What are the technologies?

Face to Face Communication

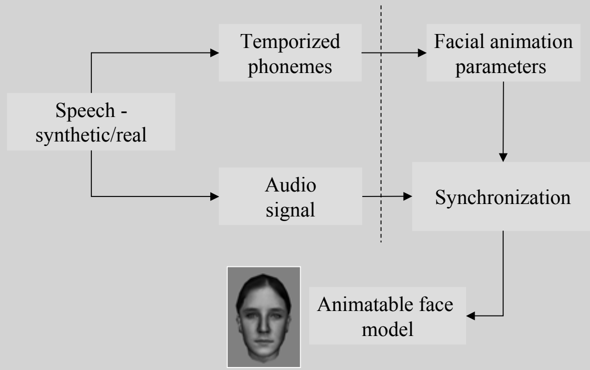
- Facial Animation
- Speech Synthesis and Recognition
- Face Recognition and Tracking
- NLP and AI

Speech Animation : Hierarchy



Step	Technology	Methods
Temporized phonemes from speech (synthetic or real)	Phoneme recognition	Manual, semi-automatic or automatic
Phoneme transition	Co-articulation	Rules based, automatic
Viseme generation and animation	Viseme definition, Synchronization with sound	Automatic

Animated Talking Heads – a typical system

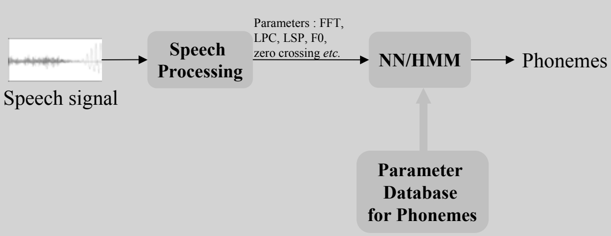


Speech Animation from

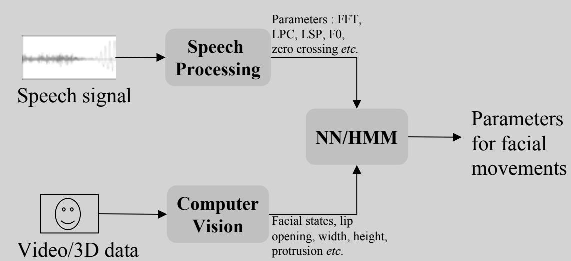
Natural Voice

(for cloned avatars)

(Only) Acoustic Analysis



Acoustic and Visual Analysis



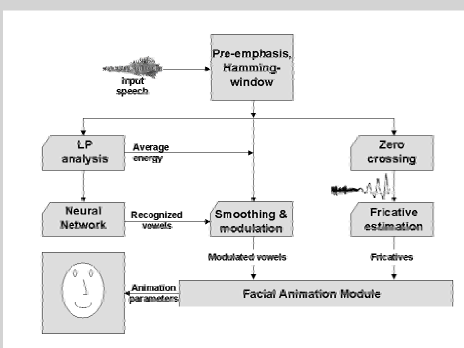
Comparison

Acoustic Analysis	Acoustic-Visual Analysis
The output "phonemes", suitable for any facial animation system	Output parameters (facial states/lip width, height *etc.*) are tied to a particular animation system
Resulting facial animation not affected by training database	Resulting facial animation closely affected by the training database
Ease in training data collection (only speech)	Training data is synchronized speech and video/3D capture
Only lip/mouth movements can be generated	Technique can be used for synthesis of other facial movements (eyebrow, nods)
Co-articulation model needs to be applied to resulting phonemes	Co-articulation effect is inherently taken care of in analysis
Greater language dependence	Less language dependence

Challenges

- Independent of language and speaker
- Independent of face model used for animation
- Minimal training requirements
- Simplicity of tools, algorithm and implementation

Speech Driven Talking Head: an example



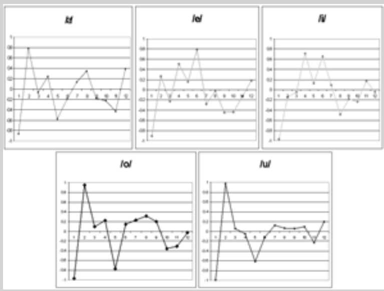
Speech Analysis

Parameter extraction

- Choice of LP Analysis
  - LP derived reflection coefficients are directly related to vocal tract shape [Wakita]
  - Phonemes can be characterized by vocal tract shape
- Limitations
  - Works well for vowels so we choose the most common five vowels /a/, /e/, /i/, /o/, /u/.
  - For the consonants?

Use of Neural Network

Typical plots for reflection coefficients for five chosen vowels



Three layer back propogation

12 input nodes, 10 hidden nodes, 5 output nodes

Five vowels used

/a/, /e/, /i/, /o/, /u/

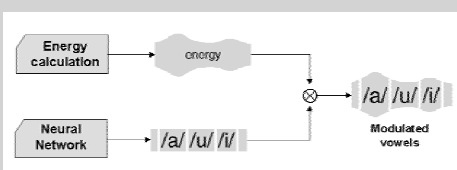
12 male and 5 female speakers

Results of NN training

		Recognized				
		/a/	/e/	/i/	/o/	/u/
Expected	/a/	241	2	15	11	0
	/e/	0	177	89	0	5
	/i/	0	3	301	0	2
	/o/	10	0	0	224	36
	/u/	4	12	0	88	143

Energy Analysis

Vowel-Vowel Transition
Semi-vowels
Consonants

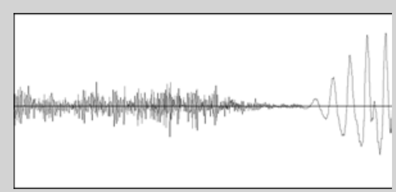


The parameters corresponding to the vowels are modulated

What more?

Zero crossing for affricates and unvoiced fricatives (/sh/, /dzh/) and /h/

Zero crossing rate is 49 per 10 msec for unvoiced, and 14 per 10 msec for voiced speech

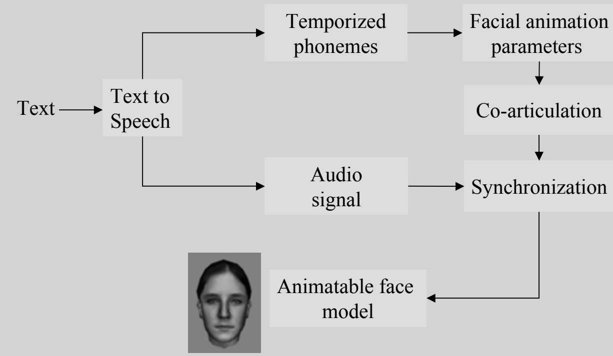


Speech Animation from

Text/Synthetic Speech

(for autonomous virtual humans)

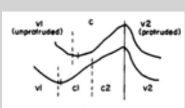
Synthetic Speech Driven Talking Head



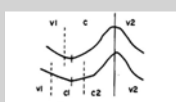
Speech Co-articulation

Co-articulation is a phenomenon observed during fluent speech, in which facial movements corresponding to one phonetic or visemic segments are influenced by those corresponding to the neighboring segments.

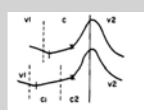
Example: a V1-C-V2 sequence where V1 is un-protruded (eg. 'a') and V2 is protruded (eg. 'u')



Transition towards v2 starts as soon as v1 ends



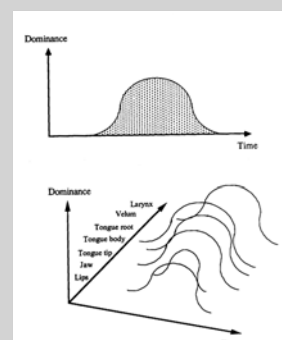
Transition towards v2 starts a fixed time interval before v2 begins



Transition towards v2 takes place in two phases

M. M. Cohen, D.W. Massaro, "Modelling coarticulation in synthetic visual speech", in N. M. Thalmann and D. Thalmann, *Models and techniques in Computer Animation*, Spinger-Verlag, 1993, pp. 139-156.

Articulatory Gesture Model



- Each speech segment (typically a viseme) has dominance that increases and decreases over time
- Adjacent visemes have overlapping dominance functions that will blend over time
- Each viseme may have a different dominance function for each articulator

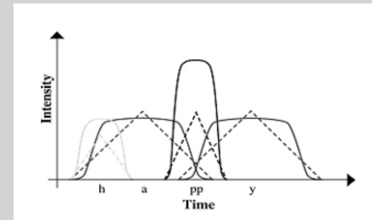
A. Löfqvist, "Speech as audible gestures", in *Speech Production and Speech Modeling, Kluwer Academic Publishers, 289-322*

Co-articulation Models for Talking Head

Pelachaud (1991) :
“Look ahead” model based on deformability of phonemes
Also considered muscle contraction times

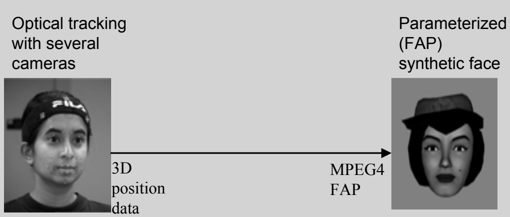
Cohen & Massaro (1992) :
Non-linear dominance and blending functions designed for each phoneme

In Summary



Define weight (dominance), and overlap according to phoneme group.

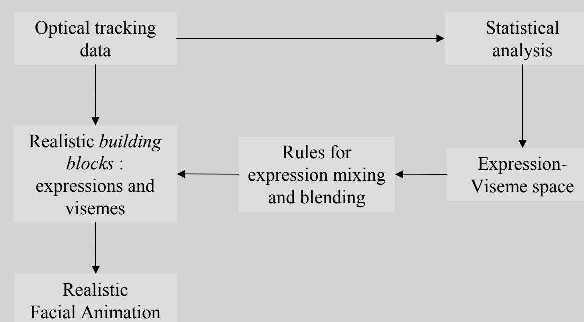
Performance Driven Facial Animation



Enhances realism to a great degree
Enables design of the *building blocks*
Limitations : complex equipment, availability of skilled performer

Realism in Talking Heads

Can we combine *flexibility* of facial animation design and *realism* of performance driven facial animation? How?



What is PCA

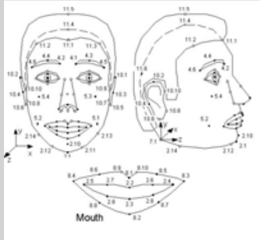
PCA is a well-known multivariate statistical analysis technique aimed at :

- reducing the dimensionality of a dataset, which consists of a large number of interrelated variables
- retaining as much as possible of the variation present in the dataset
- transforming the existing dataset into a new set of variables called the principal components (PC)

The PCs are uncorrelated and are ordered so that the first few PCs retain the most of the variation present in all of the original dataset.

Why PCA

Use of MPEG4 Feature points and FAP



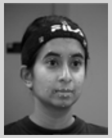
For facial capture

High correlation between facial feature points

Large amount of capture data for speech

Capturing individual as well as collective movement dynamics important during expressive speech

Data **Capture**



Optical Tracking system : Vicon

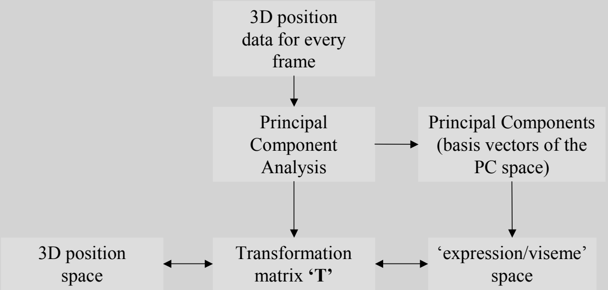
27 optical markers, 6 Cameras

Extraction of 3D positions of markers

100 phoneme rich sentences from TIMIT database

3D position data of 14 markers around lips and cheeks used for PCA

Data Analysis



Analysis results into a transformation between 3D position space and the newly constructed expression/viseme space

What are the Principal Components

(a) Open mouth
(b) Lip protrusion
(c) Lip sucking
(d) Raise cornerlips

The facial movements are controlled by single parameters, as opposed to several MPEG4 parameters needed to control the same facial movement

Eg. 'Open Mouth ' affects not only lips, but jaw and cheek region also

Thus the Principal Components take care of global facial movements using minimum number of parameters and provide higher level parameterization for facial animation design

Expression and Viseme Space

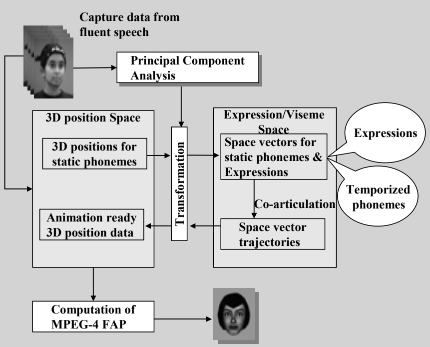
•The 'Principal Components' form the basis or the 'principal axes' of the abstract *Expression and Viseme* space

•Each point in the *Expression and Viseme* space is a facial expression, a viseme, or a combination

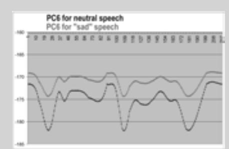
•Transition in this space from one point (expression) to another, results in smooth and realistic transition in the 3D position space giving a new way of achieving keyframe animations.

•A combination of points in this space results in realistic blending and combination of visemes and expressions in 3D position space, and hence a realistic expressive speech animation.

Application to Speech Animation



Expressive Speech



Each expression and viseme is a vector in the *Expression and Viseme* space

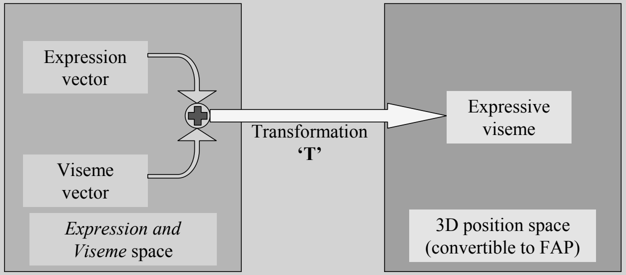
Mixing between Viseme and Expression is a simple vector addition in that space

Transforming back to 3D position space results into 'Expressive Speech'

For *happy* expression, PC2 and PC3 are most effective, as it controls lip protrusion

For *sad* expression, PC4 and PC6 is found to be most effective, that controls corner lip movements

Blending Speech with Expressions



Further Reading…

- E. Yamamoto, S. Nakamura, K. Shikano, "Lip movement synthesis from speech based on Hidden Markov Models", *Speech Communication*, Elsevier Science, (26)1-2 (1998) pp. 105-115.
- Matthew Brand, "Voice puppetry", *Proc. SIGGRAPH 99 Computer Graphics Proceedings*, Annual Conference Series, pp. 21-28.
- Sumedha Kshirsagar, Nadia Magnenat-Thalmann, Lip Synchronization Using Linear Predictive Analysis, *Proceedings of IEEE International Conference on Multimedia and Expo*, New York, August 2000.
- D. R. Hill, A. Pearce, B. Wyvill, "Animating speech: an automated approach using speech synthesized by rule", *The Visual Computer*, 3, pp. 277-289, 1988
- B. Grandstrom, "Multi-modal speech synthesis with applications", in G. Chollet, M. Di Benedetto, A. Esposito, M. Marinaro, *Speech processing, recognition, and artificial neural networks*, Springer, 1999
- M. M. Cohen, D.W. Massaro, "Modelling co-articulation in synthetic visual speech", in N. M. Thalmann and D. Thalmann, *Models and techniques in Computer Animation*, Springer-Verlag, 1993, pp. 139-156
- C. Pelachaud (1991), *Communication and Coarticulation in Facial Animation*, PhD thesis, University of Pennsylvania, 1991
- T. Kuratate, H. Yehia, E. V-Bateson, "Kinematics-based synthesis of realistic talking faces", *Proceedings AVSP'98*, pp. 185-190
- Sumedha Kshirsagar, Tom Molet, Nadia Magnenat-Thalmann, Principal Components of Expressive Speech Animation, *Proceedings Computer Graphics International 2001*, July 2001, IEEE Computer Society, pp 38-44.