Synchronisation of Senses
From Text to Speech … to Movie

Rémi Ronfard

CVAM/ICCV

Oct 23, 2017

Introduction

- IMAGINE team at INRIA on natural interfaces for designing shapes, motions and stories
- Build interactive narrative environments where the *user is the director*
  - Requires an explicit representation of story goals : character actions, events and their causal relations
  - Requires a directable film crew of virtual actors, cameramen, lighting technicians, etc.

Scientific challenges

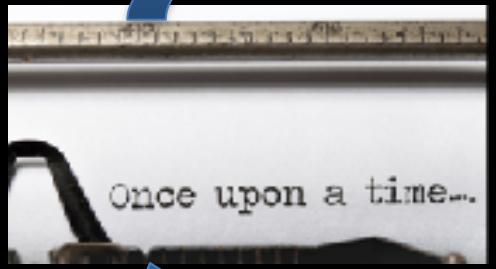
- Natural language and story understanding for script analysis
- Generative audio-visual models
  - Procedural models for 3D scene generation
  - Behavior-based 3D animation for directing virtual actors
  - Virtual cinematography for placing lights and cameras automatically and editing them together to a single string of film

Outline

- Text-to-movie
- Generative audiovisual prosody model for virtual actors
- Eisenstein's theory of vertical montage
- Continuity editing for 3D animation

Motivation: Text-to-Movie

Script



Storyboard

Stage

Editing Room

Video Game / Live Action / 3D Animation

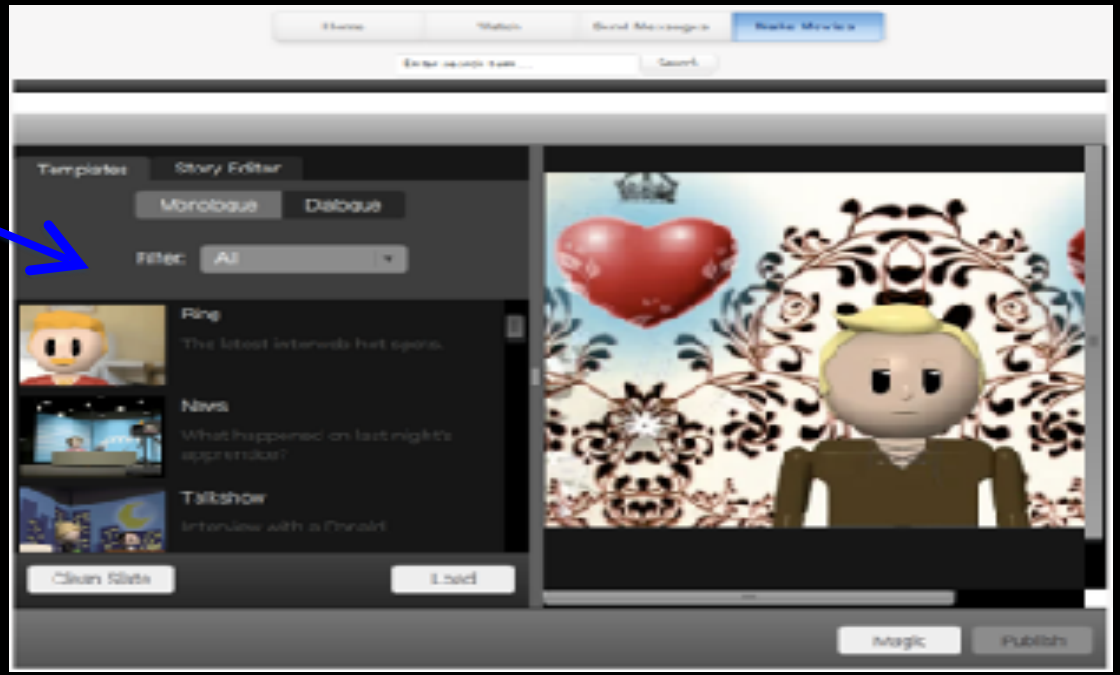
Hitchcock's dream of a machine in which he'd "*insert the screenplay at one end and the film would emerge at the other end*" (Truffaut/Hitchcock, p. 330)

Xtranormal Text-to-Movie ©

- Startup created in 2006 in Montreal
- Mission statement : 3-D animation tools for digital storytelling
- « If you can write, you can make movies »
- Shut down in 2013, re-born in 2015 as « Nawmal »

Create a short movie in four easy steps….

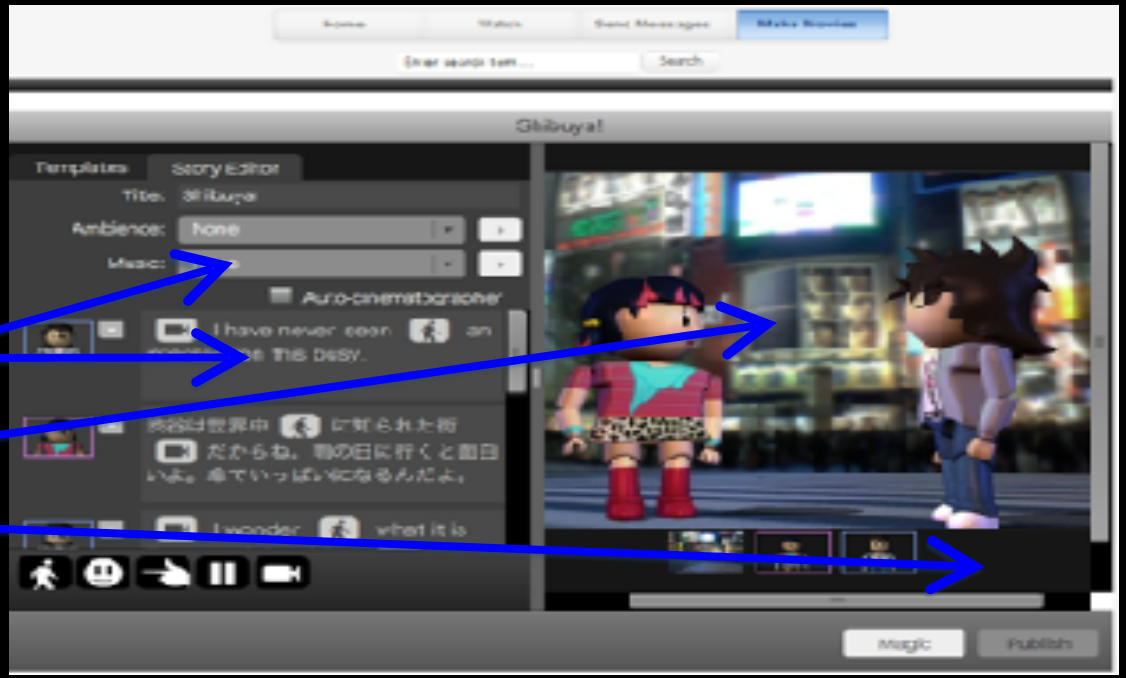
1. Pick template, characters & voices from libraries…



…without worrying about cinematography and editing

2. Type dialog and insert gestures and effects…
3. View & edit your work …

4. Publish



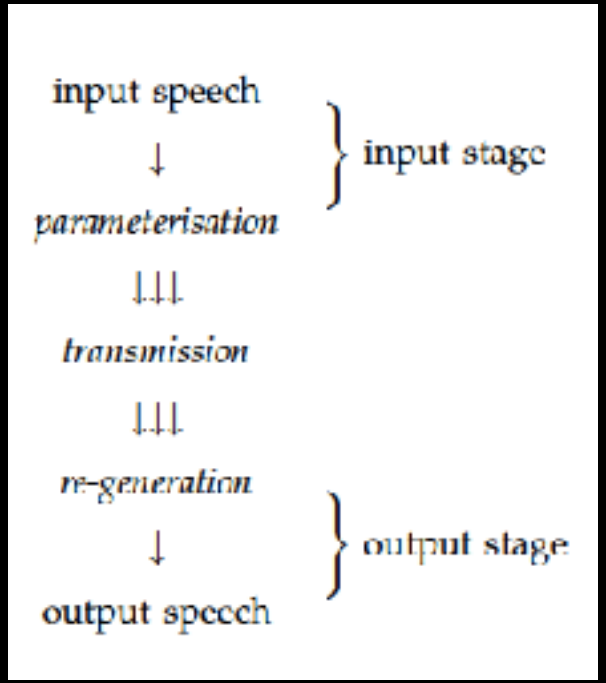
Text-to-movie: Nawmal Make

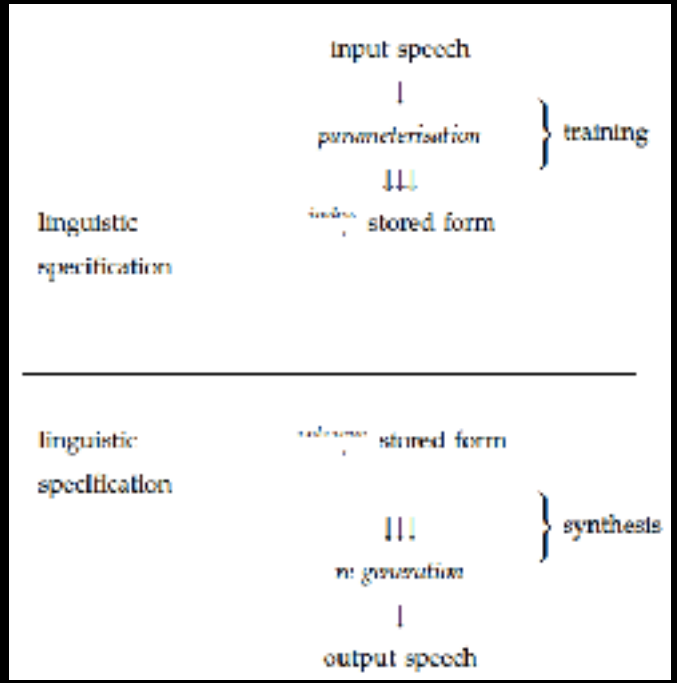
Text-to-movie: Nawmal smart cameras

Text-to-speech (TTS)

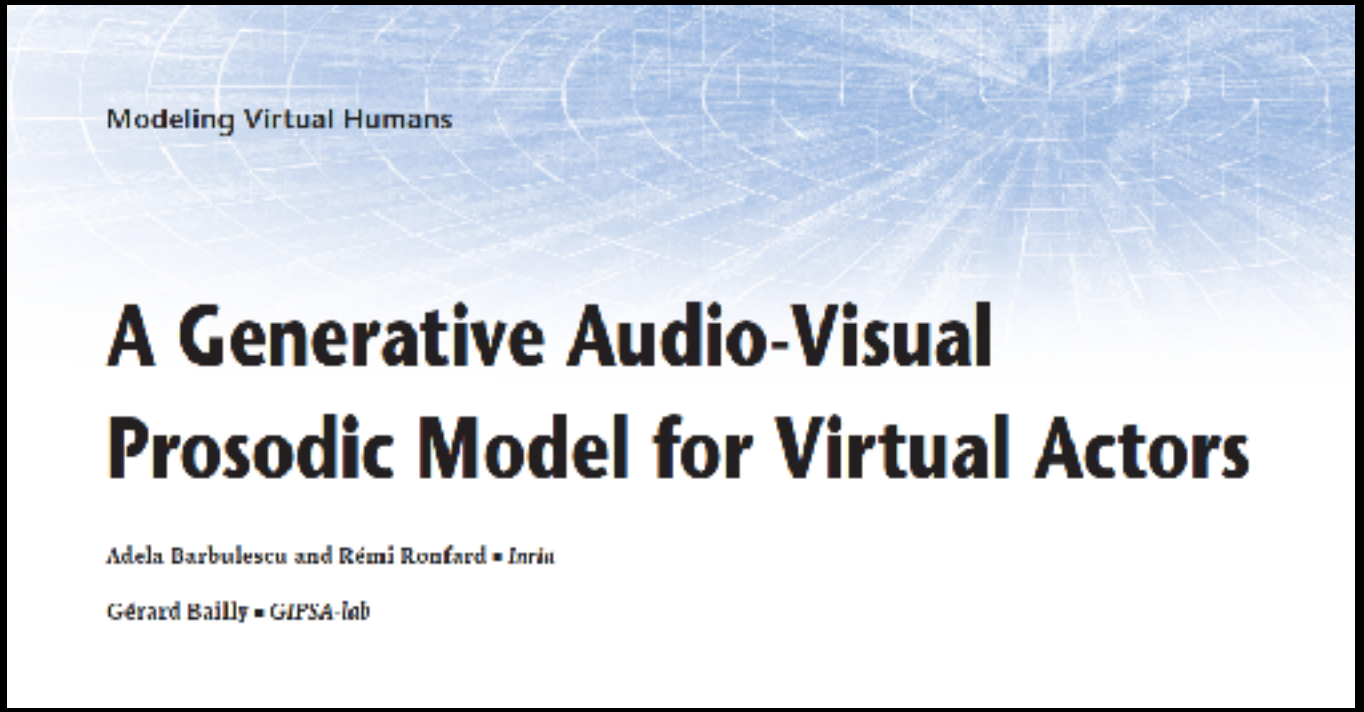
- A text-to-speech (TTS) system converts normal language text into speech; other systems render symbolic linguistic representations like phonetic transcriptions into speech.
- Allen, Jonathan; Hunnicutt, M. Sharon; Klatt, Dennis (1987). From Text to Speech: The MITalk system. Cambridge University Press.

Parametric text-to-speech (TTS)





- A beginners' guide to statistical parametric speech synthesis, Simon King, 2010.



In press: IEEE Computer Graphics and Applications, Nov/Dec 2017.

Exercises in style

Emotions and attitudes

- Actors express dramatic attitudes using the coordinated prosody of voice, rhythm, facial expressions and head and gaze motion.
- We propose a method for generating natural speech and animation in various attitudes using neutral speech and animation as input.

Audio prosody

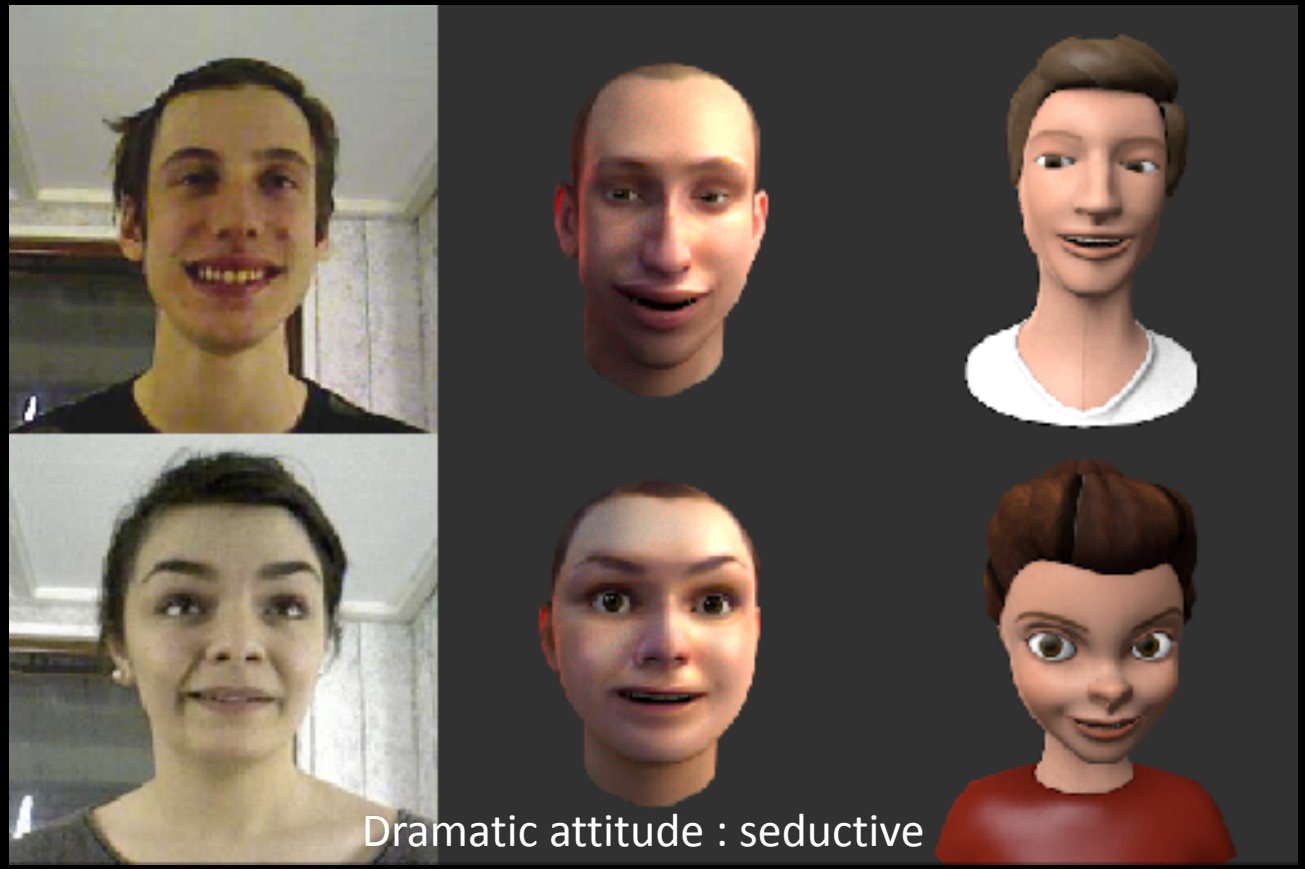
- High-level features: pitch, duration and intensity per syllable
- Low-level features: voice qualities

Visual prosody

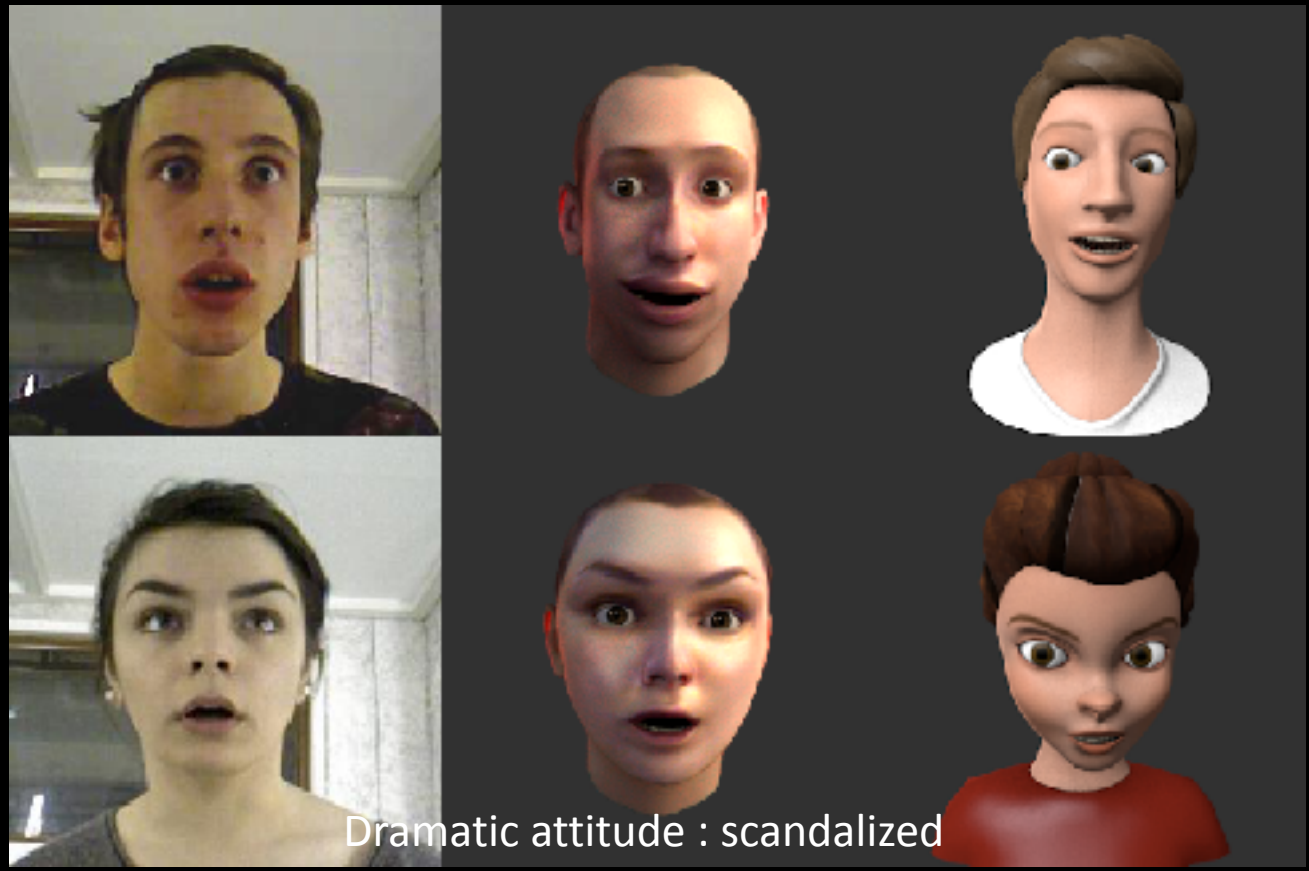
- High-level features: shoulder, head and eye movements
- Low-level features: facial expressions
- Visual Prosody: Facial Movements Accompanying Speech, Hans Peter Graf, Eric Cosatto, Volker Strom, Fu Jie Huang, Face and Gesture, 2002.

Exercises in style

Generative Audiovisual Prosodic Model

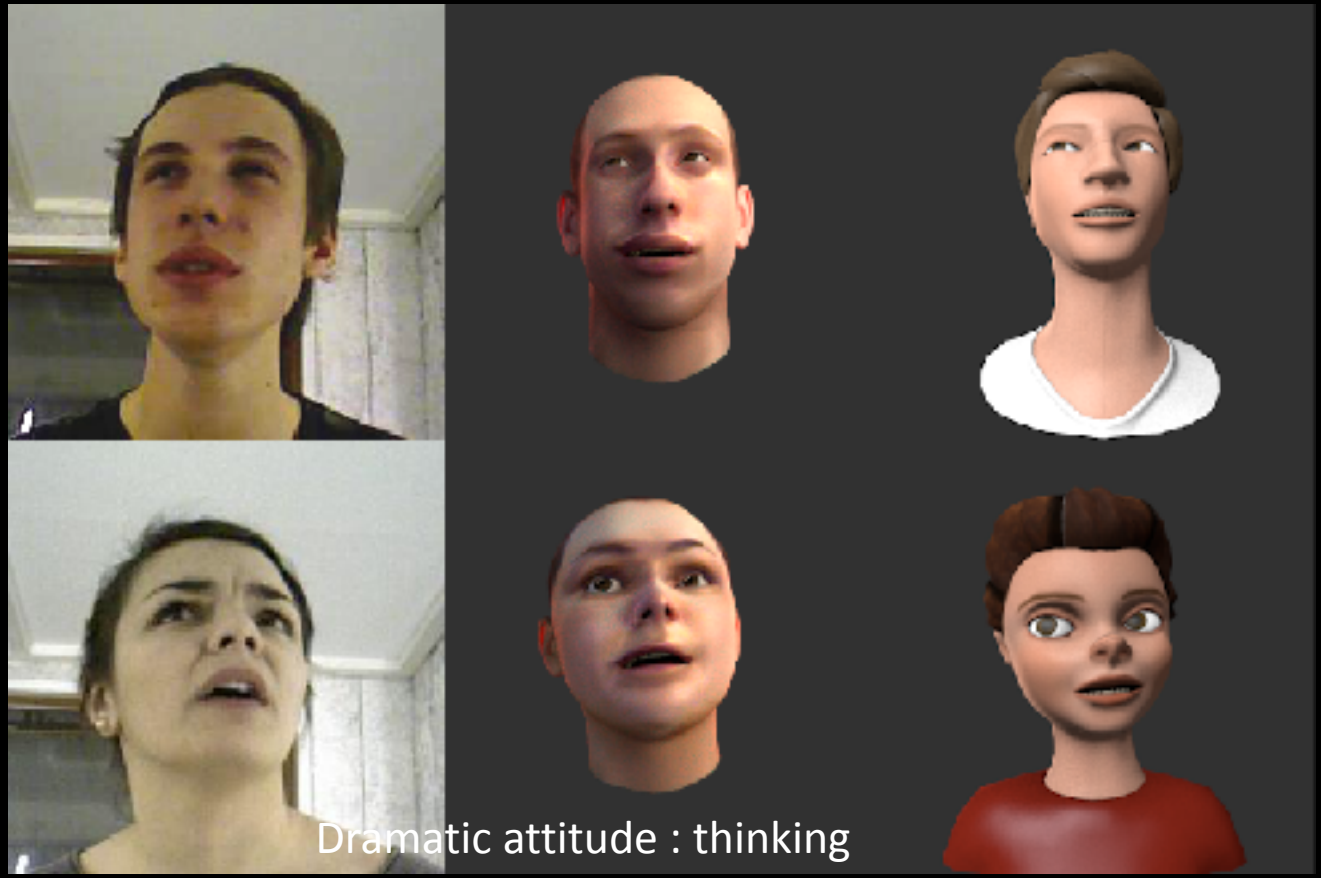


Generative Audiovisual Prosodic Model



Dramatic attitude : scandalized

Generative Audiovisual Prosodic Model



Dramatic attitude : thinking

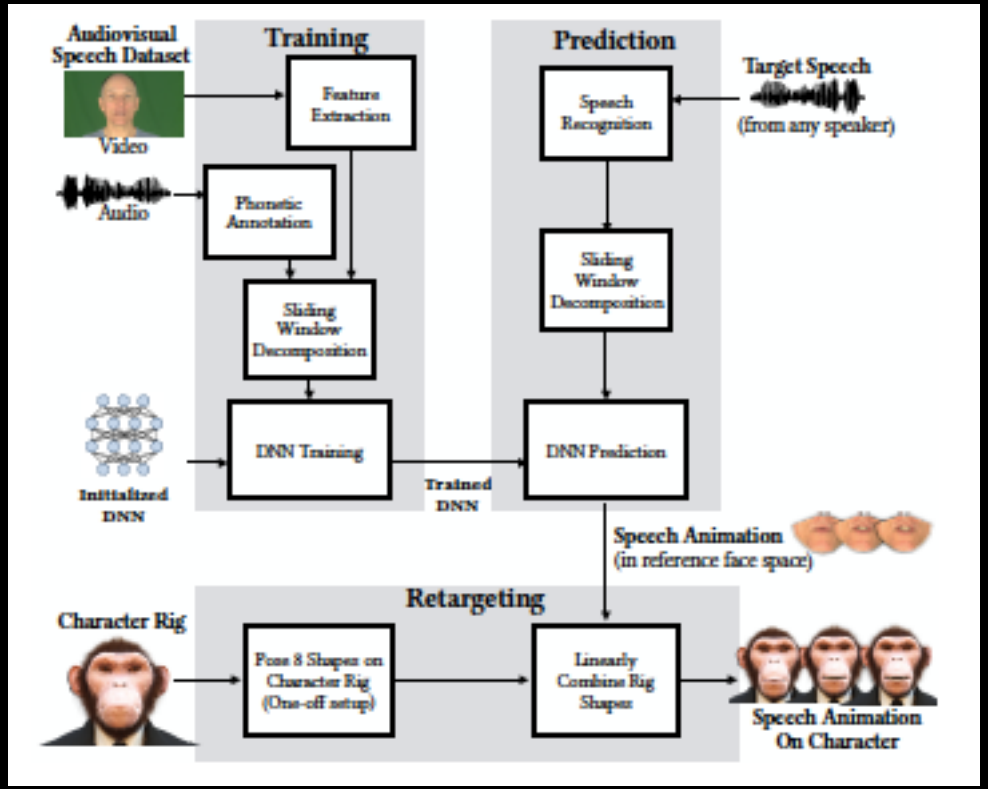
Speech-driven animation

- Erika Chuang and Christoph Bregler. 2005. Mood swings: expressive speech animation. *ACM Trans. Graph.* 2005.
- Stacy Marsella, Yuyu Xu, Margaux Lhommet, Andrew Feng, Stefan Scherer, and Ari Shapiro. 2013. Virtual character performance from speech. *Symposium on Computer Animation* (SCA '13).
- Tero Karras, Timo Aila, Samuli Laine, Antti Herva, and Jaakko Lehtinen. 2017. Audio-driven facial animation by joint end-to-end learning of pose and emotion. *ACM Trans. Graph.* 36, 4, July 2017.

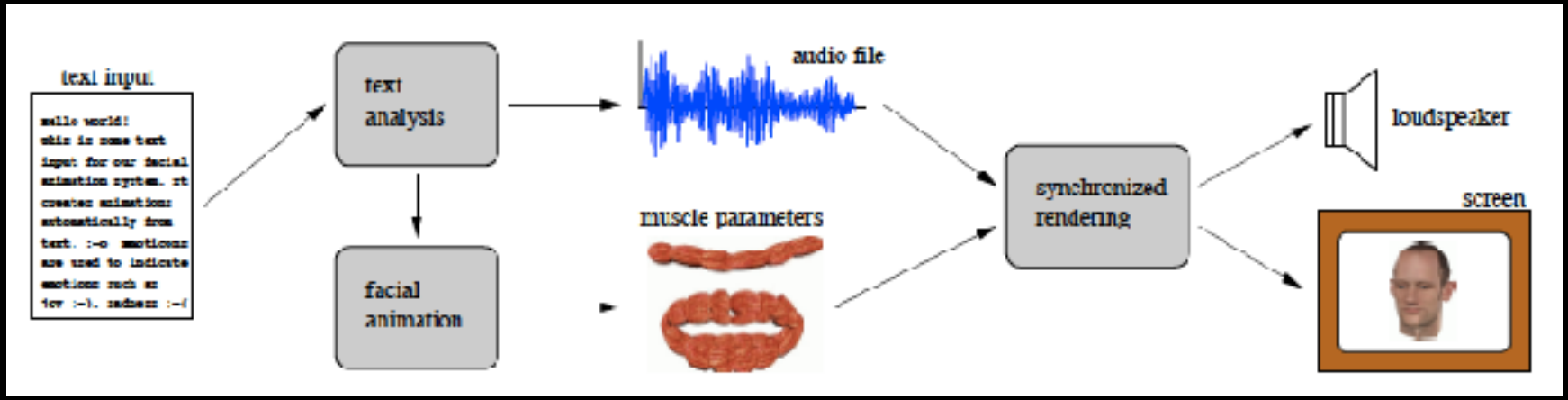
Generalized Speech Animation

- Sarah Taylor, Taehwan Kim, Yisong Yue, Moshe Mahler, James Krahe, Anastasio Garcia Rodriguez, Jessica Hodgins, and Iain Matthews. 2017. A deep learning approach for generalized speech animation. ACM Trans. Graph. 36, 4, July 2017.

Generalized Speech Animation

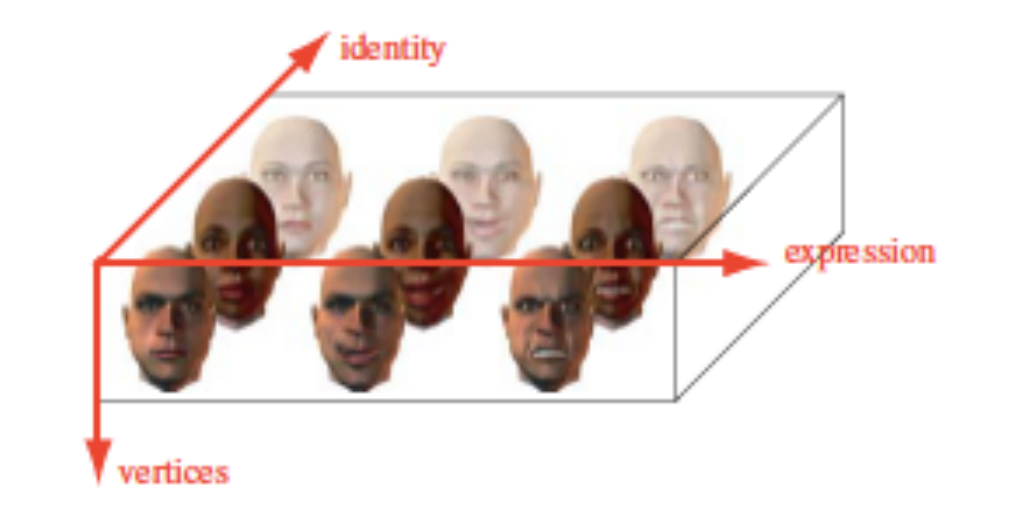


Text-driven animation



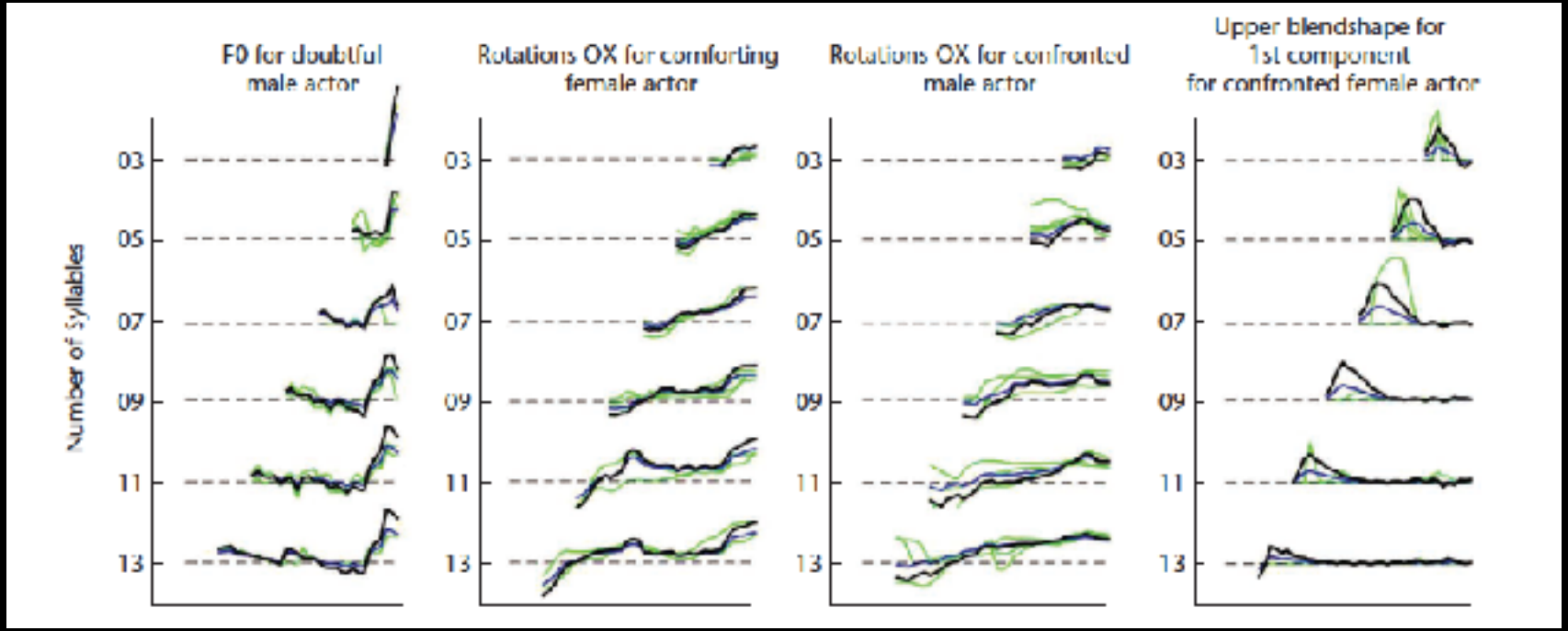
- Irene Albrecht, Jörg Haber, Kolja Kähler, Marc Schröder, and Hans-Peter Seidel. 2002. "May I talk to you?: -)" Facial Animation from Text. Pacific Graphics, 2002.

Expressive conversion

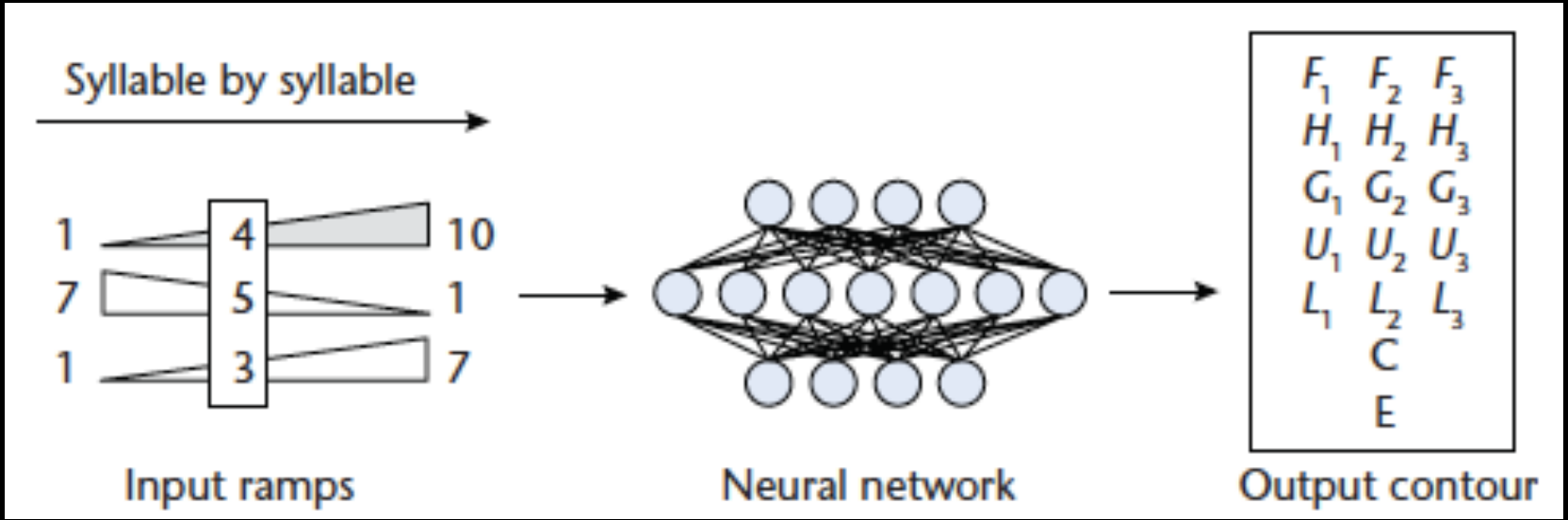


- Joint Gaussian Mixture Models of expression pairs
- Daniel Vlasic, Matthew Brand, Hanspeter Pfister, and Jovan Popovic. 2006. Face transfer with multilinear models. In ACM SIGGRAPH 2006 Courses (SIGGRAPH '06).

Our approach: prosodic contours

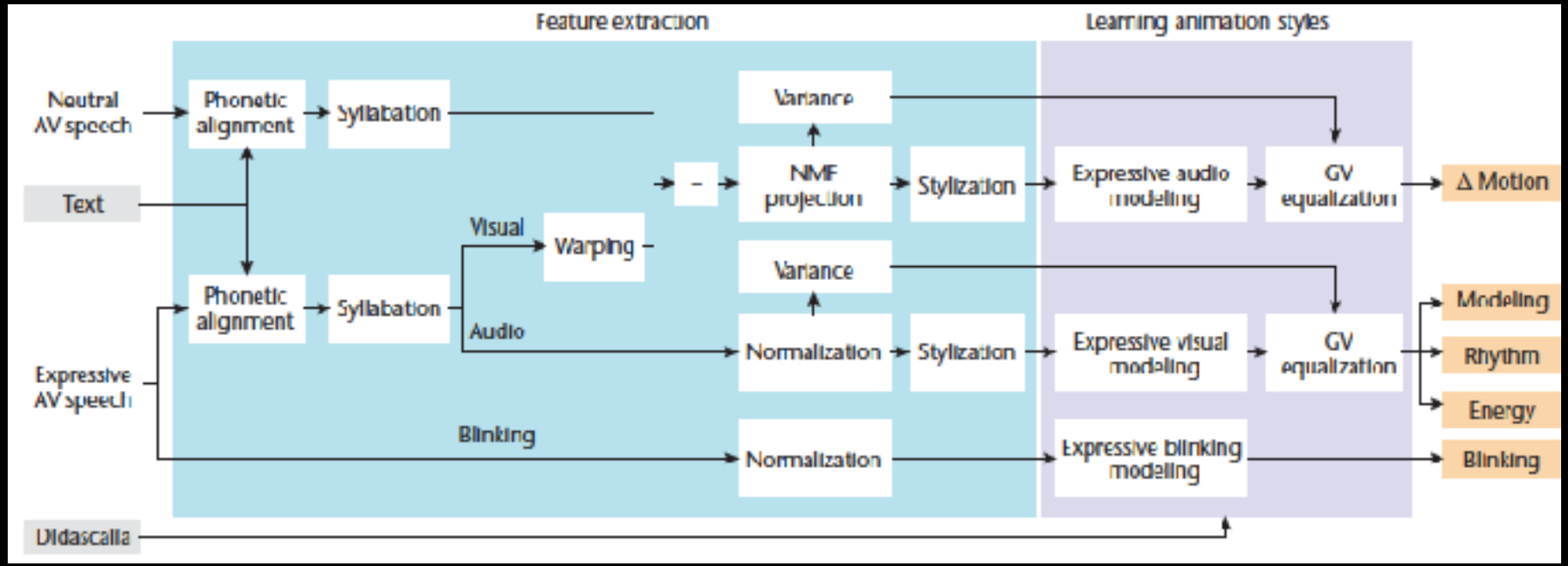


Our approach: prosodic contours

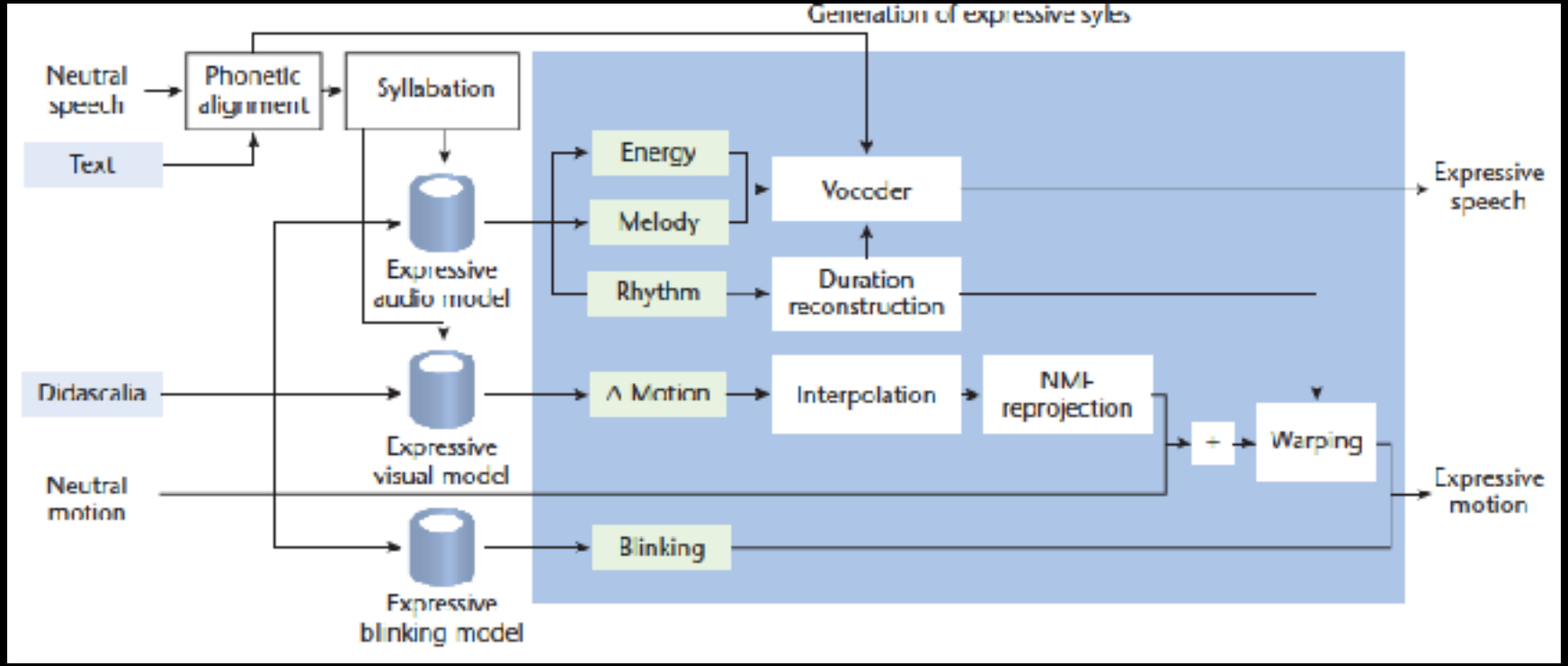


- F=voice pitch, H=head motion, G=gaze motion, U=upper-face, L=lower-face, C=rhythm, E=energy

Learning audiovisual prosody



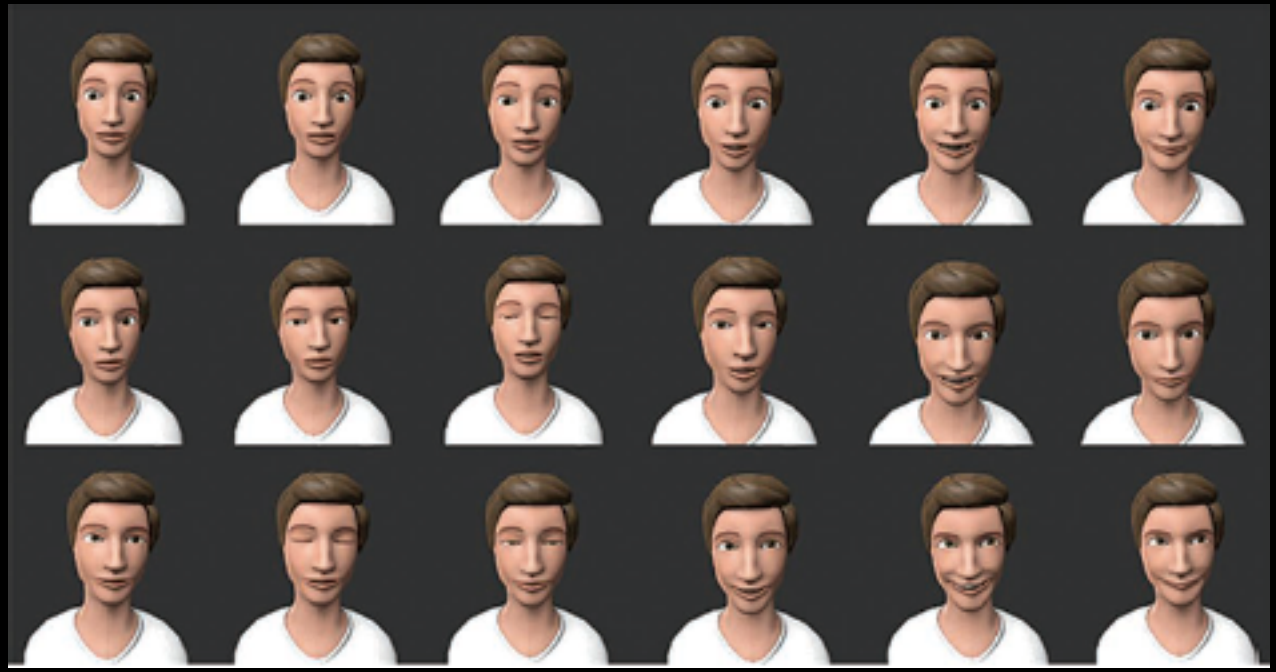
Generating audiovisual prosody



Experimental results

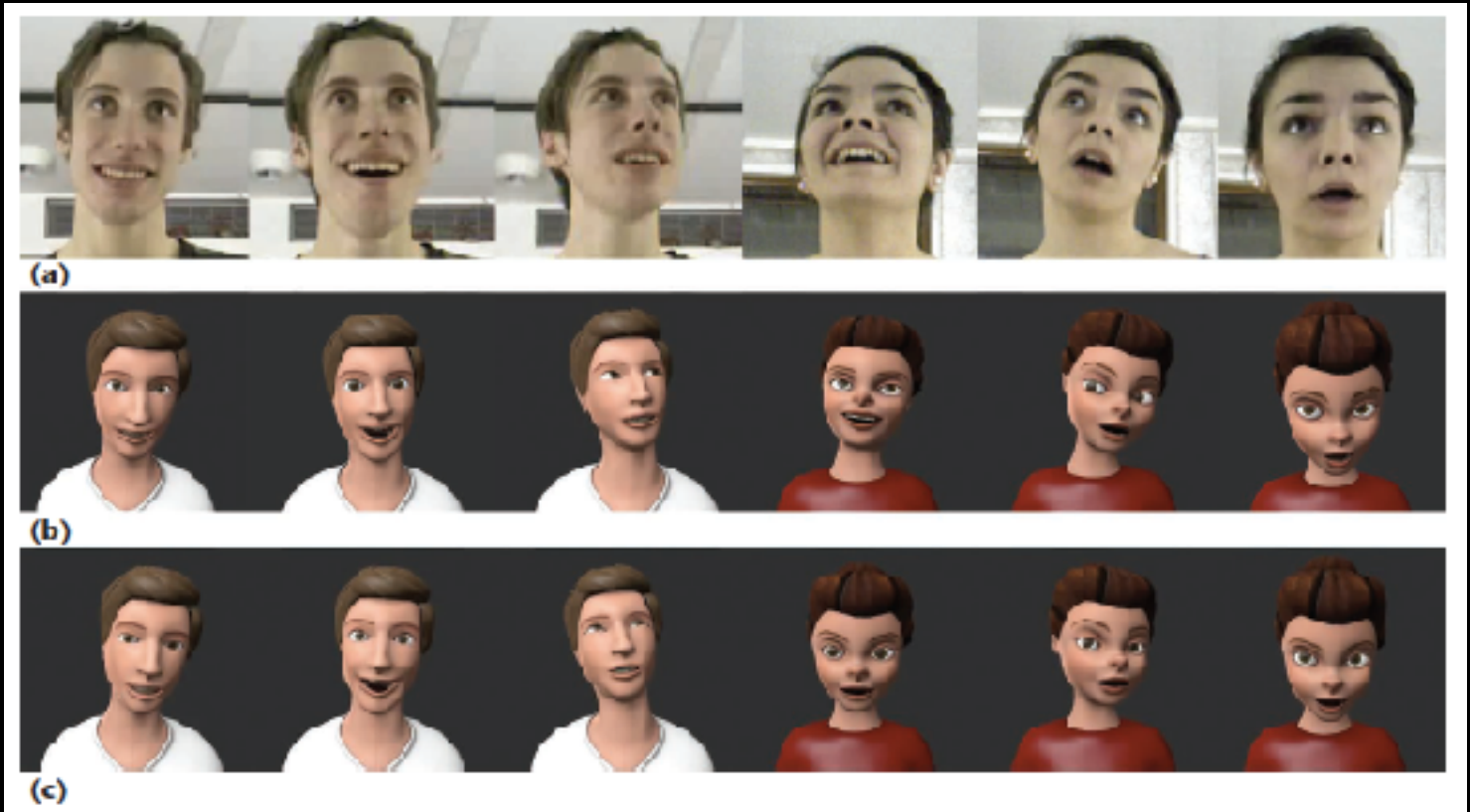
- Thank you for the lovely flowers: thinking, ironic, scandalized

Experimental results

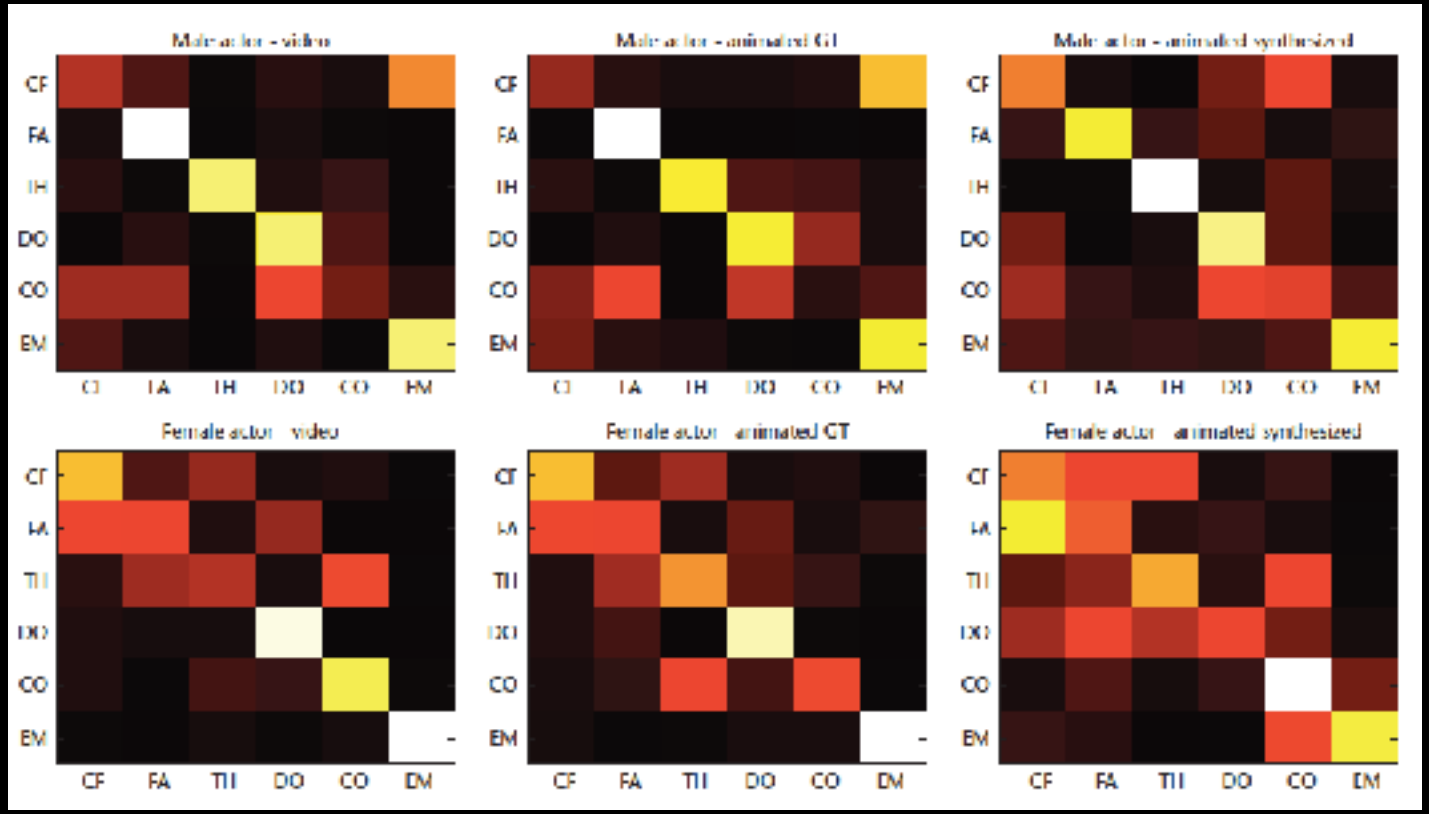


- You’re welcome (fascinated, doubtful, embarrassed)

Subjective evaluation



Subjective evaluation



- CF=comforting, FA=fascinated, TH=thinking, DO=doubtful, C0=confronted, EM=embarrassed

Subjective evaluation

Table 2. F1 scores obtained for the male actor for the video (C1), animated ground truth (C2), and our method (C3).*

Condition	CF	FA	TH	DO	CO	EM	Mean
C1	0.31	0.81	0.90	0.71	0.23	0.68	0.60
C2	0.27	0.79	0.80	0.68	0.08	0.58	0.53
C3	0.48	0.72	0.82	0.61	0.28	0.67	0.59

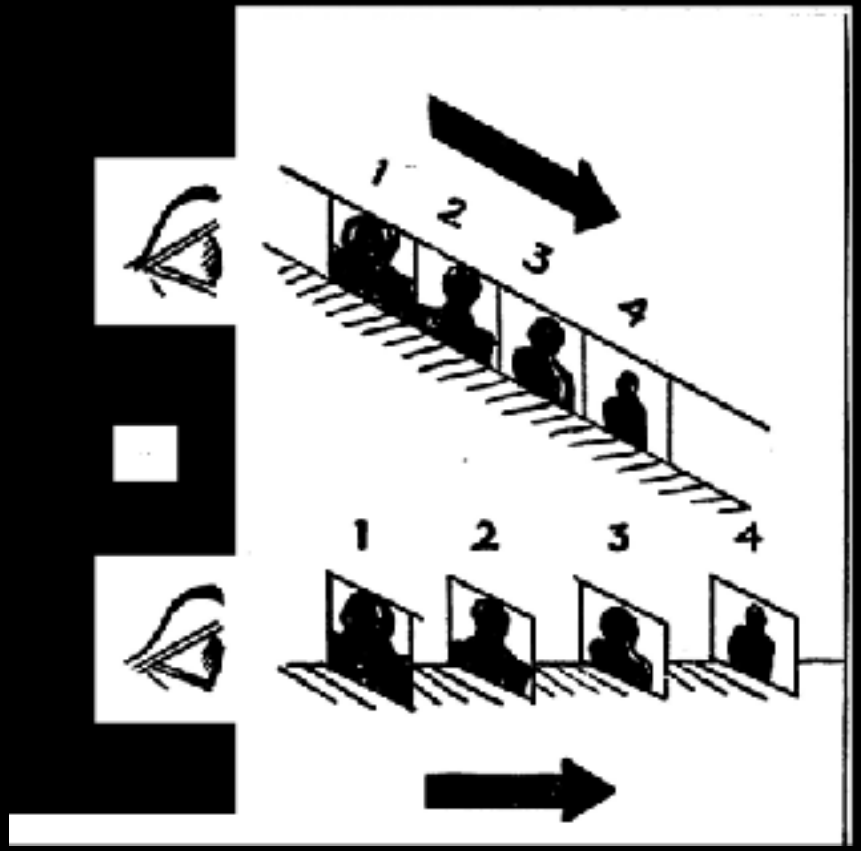
Table 3. F1 scores obtained for the female actor for the video (C1), animated ground truth (C2), and our method (C3).*

Condition	CF	TE	SE	FA	IR	SC	Mean
C1	0.58	0.41	0.31	0.82	0.68	0.98	0.63
C2	0.56	0.37	0.51	0.77	0.56	0.94	0.61
C3	0.35	0.33	0.44	0.36	0.56	0.67	0.45

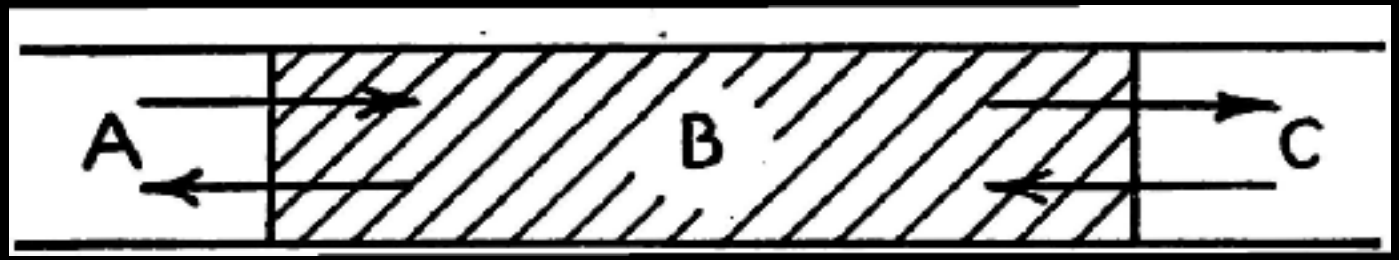
- CF=comforting, FA=fascinated, TH=thinking, DO=doubtful, C0=confronted, EM=embarrassed

Exercises in style results

Eisenstein, synchronization of senses



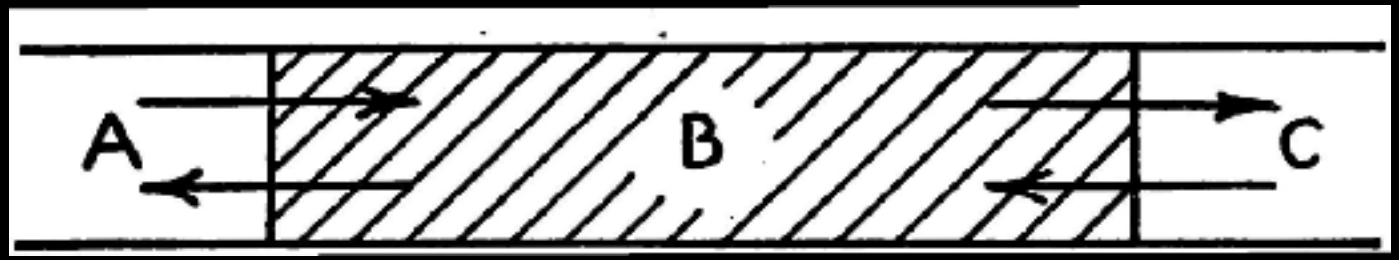
Eisenstein, synchronization of senses



MONTAGE defined as:

- Piece A, derived from the elements of the theme being developed
- Piece B, derived from the same source
- in juxtaposition give birth to the image in which the thematic matter is most clearly embodied.

Eisenstein, synchronization of senses



Representation A and representation B must be so selected from all the possible features within the theme that their juxtaposition shall evoke in the perception and feelings of the spectator the most complete image of the theme itself.

Eisenstein, synchronization of senses

- Transition from silent montage to sound-picture, or audio-visual montage changes nothing in principle. Our conception of montage encompasses equally the montage of the silent film and of the sound-film.
- However, this does not mean that in working with sound-film, we are not faced with new tasks, new difficulties, and even entirely new methods.
- On the contrary!

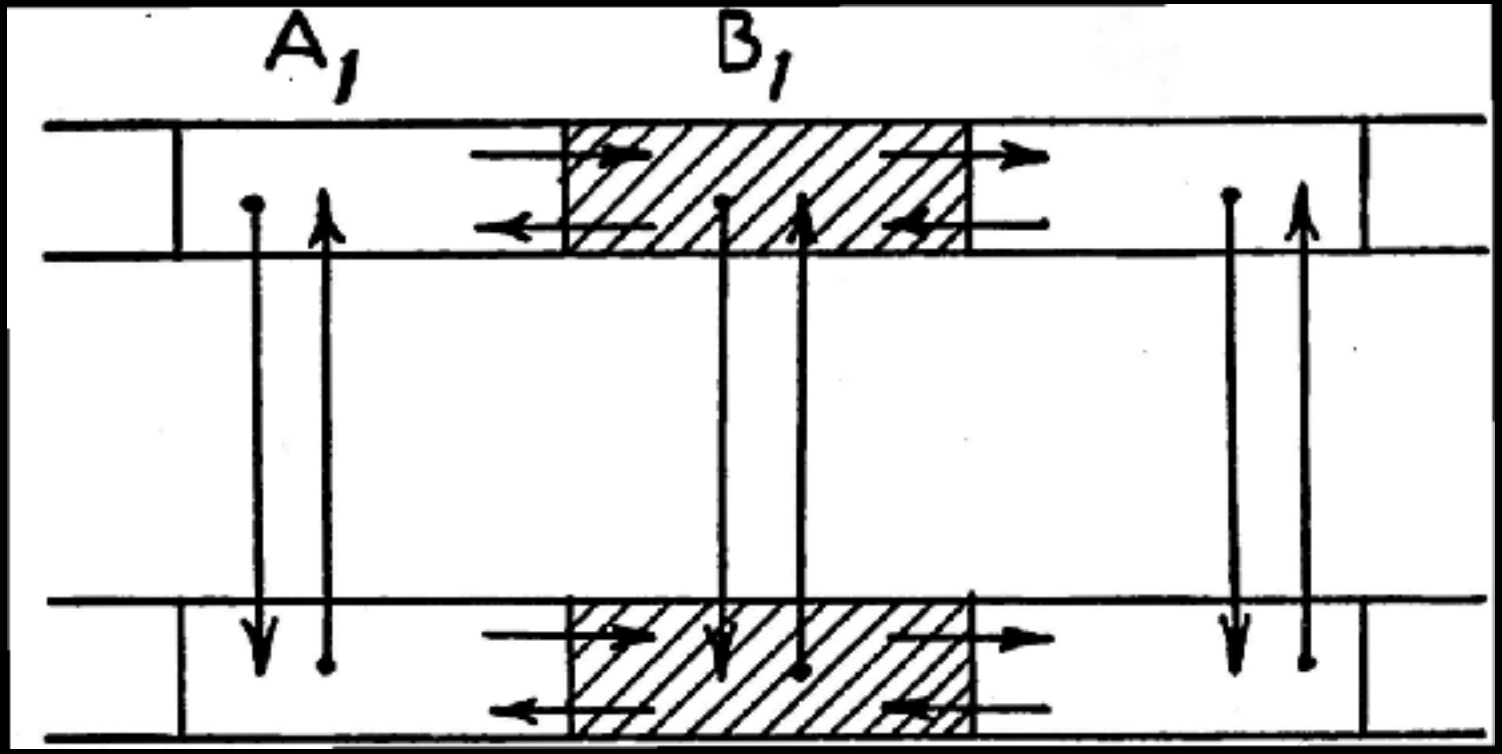
Eisenstein, synchronization of senses

- That is why it is so necessary for us to make a thorough analysis of the nature of audio-visual phenomena.
- Our first question is: Where shall we look for a secure foundation of experience with which to begin our analysis?

Eisenstein, synchronization of senses

– Man and the relations between his gestures and the intonations of his voice, which arise from the same emotions, are our models in determining audio-visual structures, which grow in an exactly identical way from the governing image.

Eisenstein, synchronization of senses



Eisenstein, synchronization of senses

- To relate image with sound, we find a natural language common to both-movement.
- Movement will reveal all the substrata of inner synchronization that we wish to establish in due course. Movement will display in a concrete form the significance and method of the fusion process.

Eisenstein, synchronization of senses

- Let us examine a number of different approaches to synchronization in logical order.
- The first is a purely factual synchronization: the sound-filming of natural things (a croaking frog, the mournful chords of a broken harp, the rattle of wagon wheels over cobblestone).

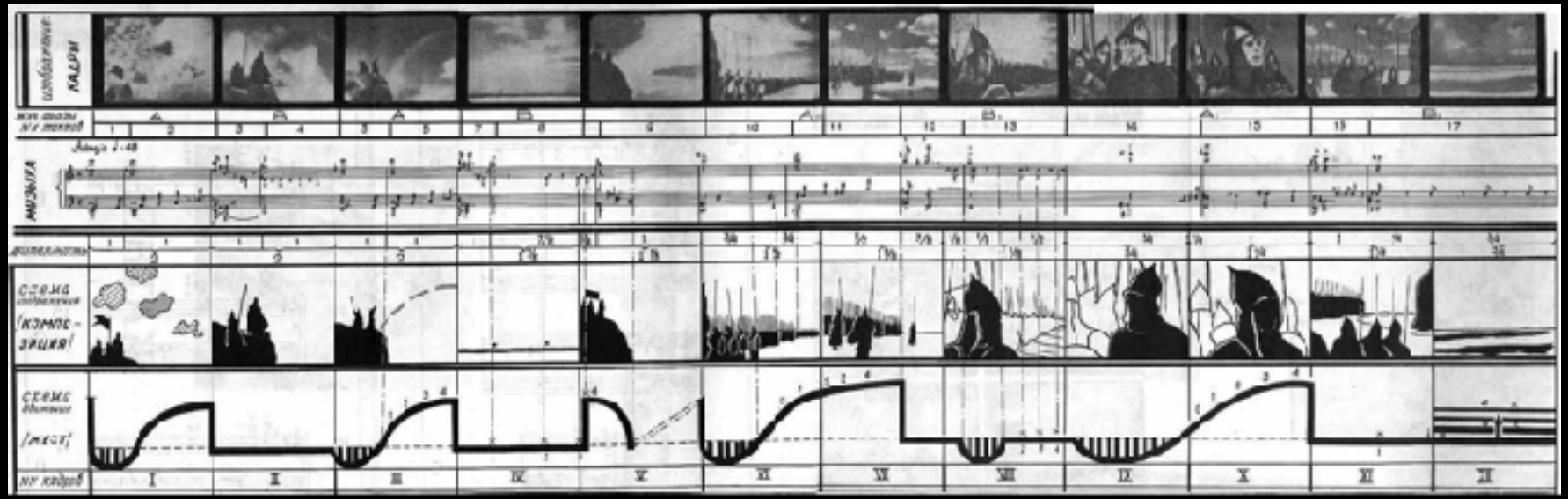
Eisenstein, synchronization of senses

- In the more rudimentary forms of expression both elements (the picture and its sound) will be controlled by an identity of rhythm, according to the content of the scene.
- This is the simplest, easiest and most frequent circumstance of audio-visual montage, consisting of shots cut and edited together to the rhythm of the music on the parallel sound-track.

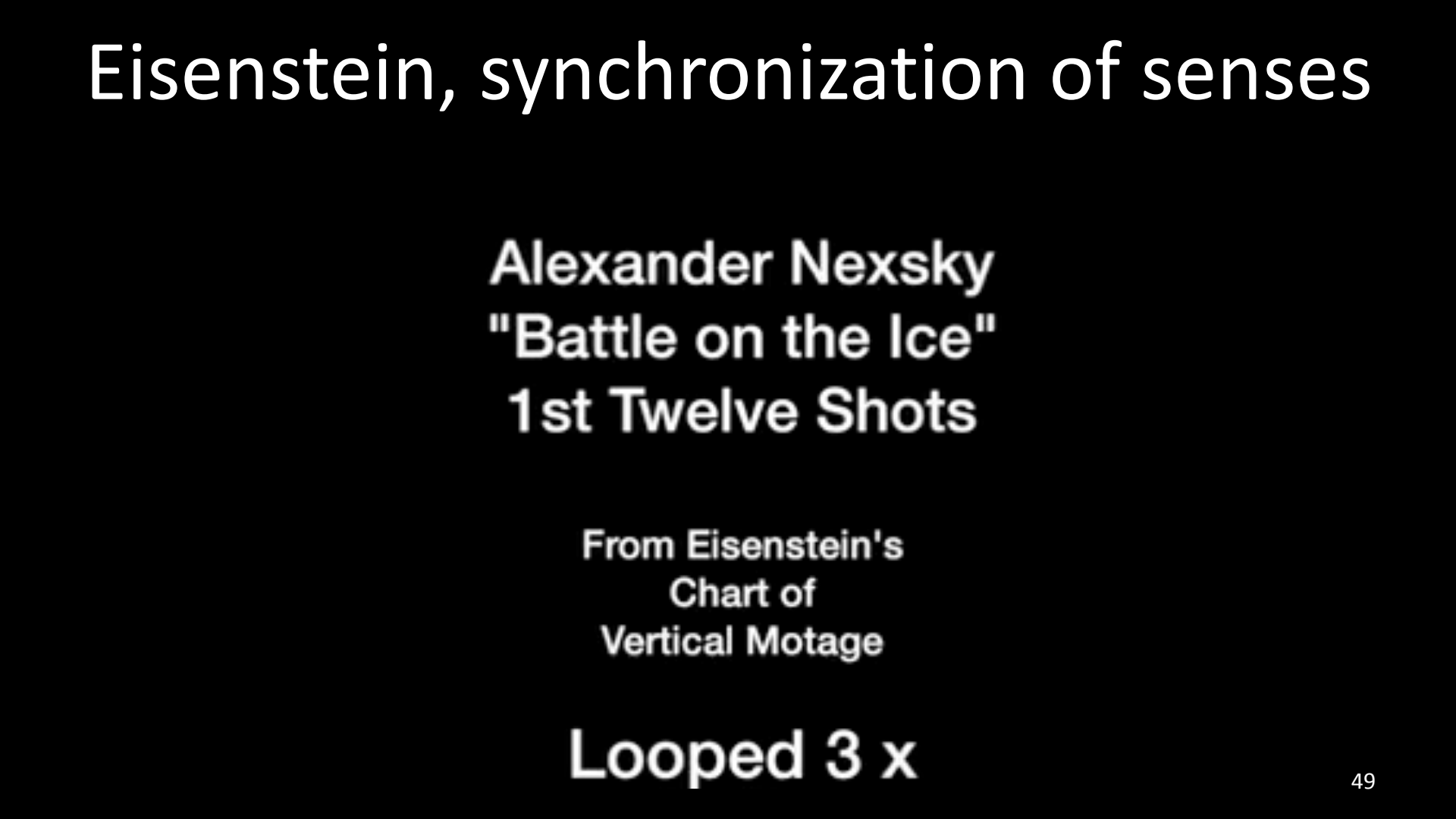
Eisenstein, synchronization of senses

- We can surely find a shot whose movement harmonizes not only with the movement of the rhythmic pattern, but also with the movement of the melodic line.
- (…)
- Synchronization can be natural, metric, rhythmic, melodic and tonal.

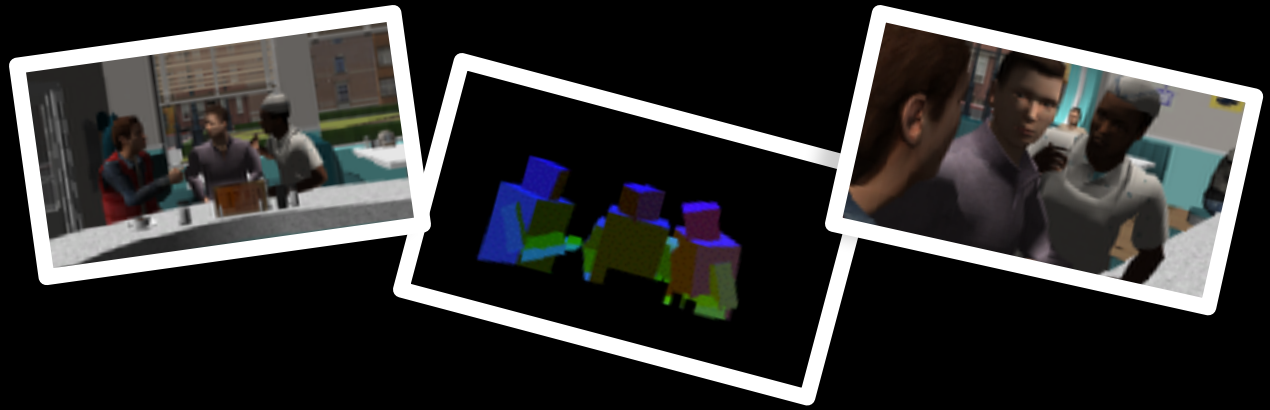
Eisenstein, synchronization of senses



Eisenstein, synchronization of senses



Continuity Editing for 3D Animation



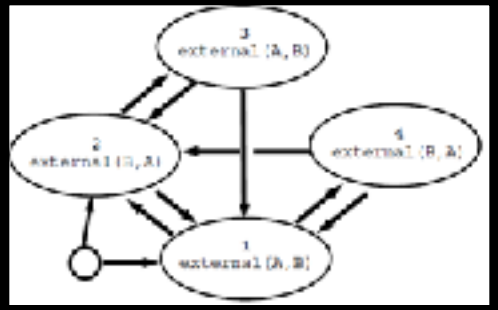
Twenty-Ninth AAAI Conference
2015

Quentin Galvane
Rémi Ronfard
Christophe Lino
Marc Christie

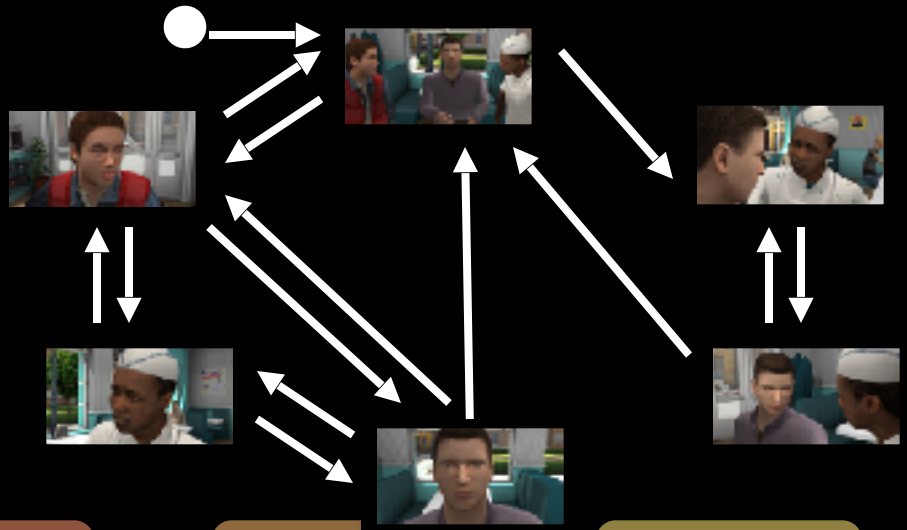
Objectives

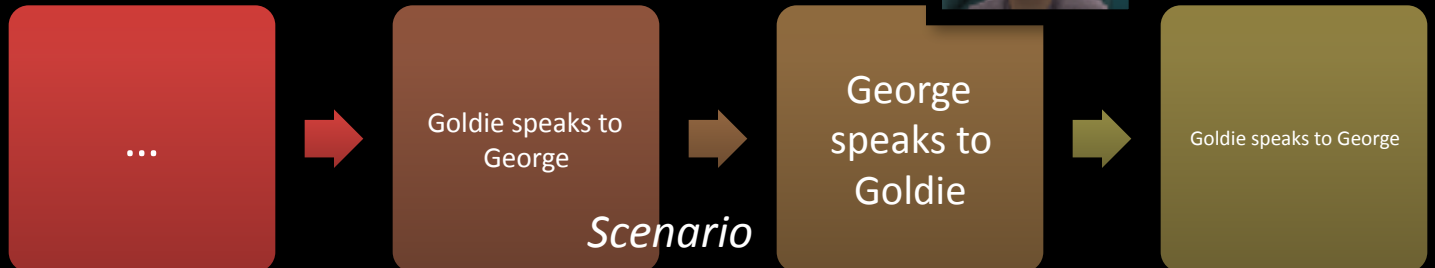
- Read actions and dialogues from script
- Generate speech and animation
- Place cameras and lights, generate rushes
- Edit the rushes into a movie

Idiom based solutions



Virtual cinematographer
[Christianson et al. 1996]





Optimization based approach
Dynamic programming

[Riedl, M. et al., 2008]

All cameras evaluated over the entire beat

All transitions evaluated at beat changes

...

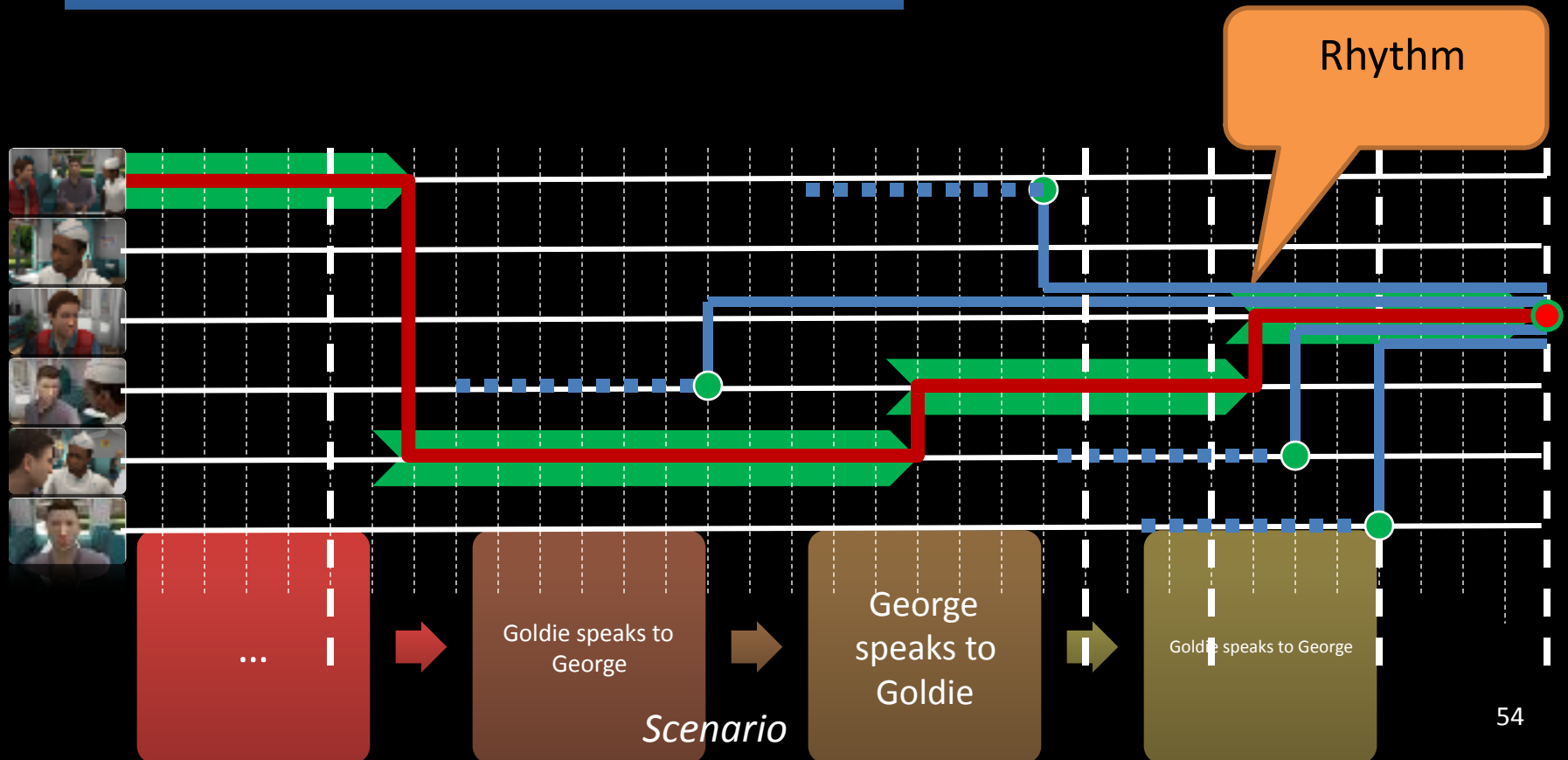
Goldie speaks to George

George speaks to Goldie

Goldie speaks to George

Scenario

Evaluate all possible transitions



Outline

- Film editing as an optimization problem
  - Semi-Markov chains
- Create an editing graph that evaluates 3 aspects:
  - Shot quality
  - Cut quality
  - Rhythm

Film editing as optimization

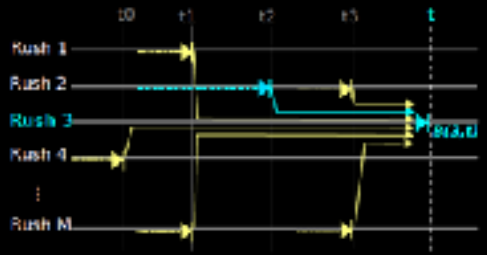
- Search over semi-Markov chains $s = (r_j, d_j)$ given actions $a(t)$
- Minimize cost function:

$$C(s,a) = \sum_j \sum_{t_j \le t \le t_j + d_j} C^A(r_j, t) + \sum_{1 \le j} C^T(r_{j-1}, r_j, t_j) + \sum_j C^R(d_j)$$

Action cost (Shot quality)

Transition cost (Cut quality)

Rhythm cost (Rhythmic Quality)



The final editing is given by the shortest path in the editing graph

➢Shot quality:

- Hitchcock principle

The size of a character on the screen should be proportional to its narrative importance in the story.

- Narrative importance from script
- Visible area V = S – O for each rush

$$C_H^A(r,t) - \sum_{c \in C} \left| \frac{I(c,t)}{\sum_{c'} I(c',t)} - \frac{V(c,r,t)}{\sum_{c'} V(c',r,t)} \right|$$

Actors and actions

Back to the Future

R.Zemeckis (1985)

Results

Input

Computed Editing

Camera : 6

Last shot cost: 13.50s
Value: 0.47

Last cut from 15 to 6 at 7.54s
Value: 0.03

Pacing from 7.54s to 15.50s (8.00s)
Value: 0.06

Limitations

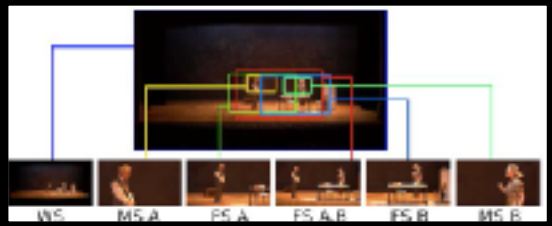
- Audio tracks and cameras must be pre-computed
- Cannot handle ellipsis or flashbacks
- Cannot handle book-ending
  - Context free grammar

Future work

- Optimize over camera positions and movements

[Galvane et al., 2014]

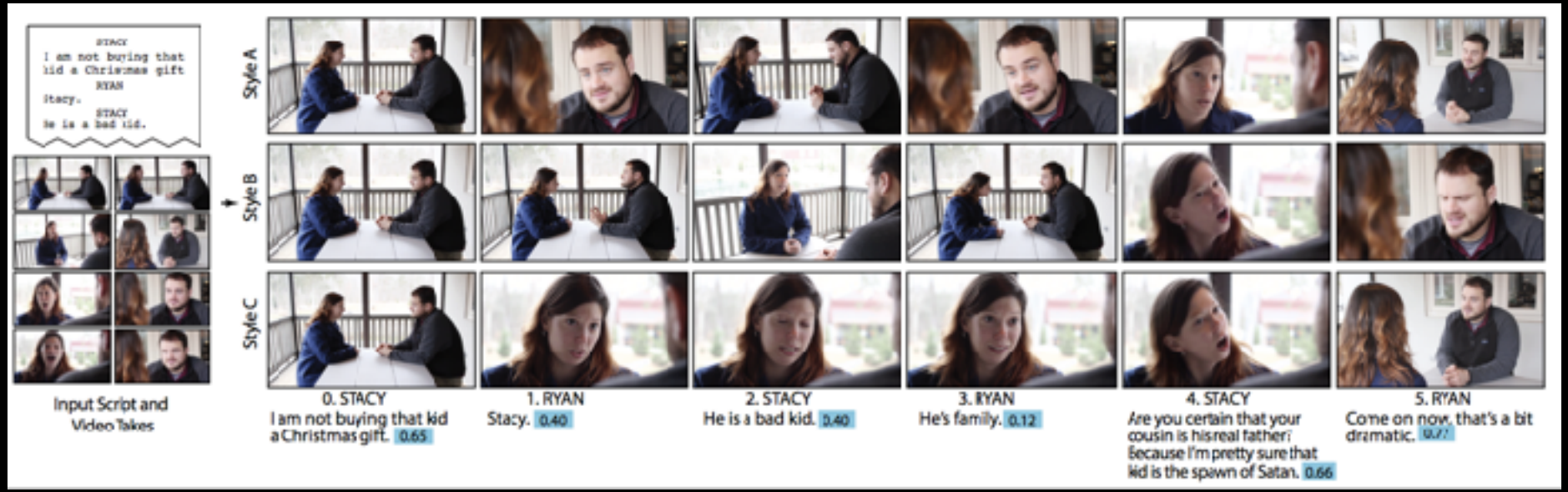
- Extend to live action video



- Learn other editing styles from real movies

[Gandhi et al., 2014]

Computational Video Editing for Dialogue-Driven Scenes
Mackenzie Leake, Abe Davis, Anh Truong, Maneesh Agrawala



What about vertical editing ?

- Can we learn statistical models of Eisenstein style of montage ?
  - Harder than continuity editing
  - Semi-Markov model still relevant
  - Vertical relations between sound and picture
  - Vertical relations between virtual camera shots

What about vertical editing ?

- Semi-Markov models can be useful !
  - syllable and sentence durations
  - shot and scene durations
  - action durations
- Multimodal Semi-Markov models needed !

Conclusion

- Generative audiovisual models
  - expression of emotions and attitudes with prosody
  - expression of narratives with video editing
- Motivated by text to movie conversion
- Also important for movie description

References

- The Prose Storyboard Language: A Tool for Annotating and Directing Movies. *Rémi Ronfard, Vineet Gandhi, Laurent Boiron. 2nd Workshop on Intelligent Cinematography and Editing part of Foundations of Digital Games - FDG 2013.*
- Narrative-Driven Camera Control for Cinematic Replay of Computer Games. *Quentin Galvane, Rémi Ronfard, Marc Christie, Nicolas Szilas. MIG 2014.*
- Beyond Basic Emotions: Expressive Virtual Actors with Social Attitudes. *Adela Barbulescu, Rémi Ronfard, Gérard Bailly, Georges Gagneré, Huseyin Cakmak. 7th International ACM SIGGRAPH Conference on Motion in Games 2014.*

References

- Continuity Editing for 3D Animation. *Quentin Galvane, Rémi Ronfard, Christophe Lino, Marc Christie. AAAI Conference on Artificial Intelligence, Jan 2015.*
- Camera-on-rails: Automated Computation of Constrained Camera Paths. *Quentin Galvane, Marc Christie, Christophe Lino, Rémi Ronfard. ACM SIGGRAPH Conference on Motion in Games, Nov 2015.*
- Implementing Hitchcock - the Role of Focalization and Viewpoint. *Quentin Galvane, Rémi Ronfard. Eurographics Workshop on Intelligent Cinematography and Editing, Apr 2017.*

References

- Five Challenges for Intelligent Cinematography and Editing. Rémi Ronfard. *Eurographics Workshop on Intelligent Cinematography and Editing, Apr 2017.*
- Which prosodic features contribute to the recognition of dramatic attitudes? *Adela Barbulescu, Rémi Ronfard, Gérard Bailly. Speech Communication, Aug. 2017.*
- A Generative Audio-Visual Prosodic Model for Virtual Actors. *Adela Barbulescu, Rémi Ronfard, Gérard Bailly. IEEE Computer Graphics and Applications, Nov./Dec. 2017.*