Generation of Audiovisual Prosody for Expressive Virtual Actors

Adela Barbulescu

Advisor: Rémi Ronfard

Advisor: Gérard Bailly

Motivation

Theatrical performances

Motivation

Dramaturgic text

VLADIMIR:(*musingly*). The last moment . . . (*He meditates.*) Hope deferred maketh the something sick, who said that?

ESTRAGON: Why don't you help me?

VLADIMIR: Sometimes I feel it coming all the same. Then I go all queer. (*He takes off his hat, peers inside it, feels about inside it, shakes it, puts it on again.*) How shall I say? Relieved and at the same time . . . (*he searches for the word*) . . . appalled. (*With emphasis.*) AP-PALLED. (*He takes off his hat again, peers inside it.*) Funny. (*He knocks on the crown as though to dislodge a foreign body, peers into it again, puts it on again.*) Nothing to be done. (*Estragon with a supreme effort succeeds in pulling off his boot. He peers inside it, feels about inside it, turns it upside down, shakes it, looks on the ground to see if anything has fallen out, finds nothing, feels inside it again, staring sightlessly before him.*) Well?

ESTRAGON: Nothing.

VLADIMIR: Show me.

ESTRAGON: There's nothing to show.

VLADIMIR: Try and put it on again.

ESTRAGON:(*examining his foot*). I'll air it for a bit.

VLADIMIR: There's man all over for you, blaming on his boots the faults of his feet. (*He takes off his hat again, peers inside it, feels about inside it, knocks on the crown, blows into it, puts it on again.*) This is getting alarming. (*Silence. Vladimir deep in thought, Estragon pulling at his toes.*) One of the thieves was saved. (*Pause.*) It's a reasonable percentage. (*Pause.*) Gogo.

ESTRAGON: What?

VLADIMIR: Suppose we repented.

ESTRAGON: Repented what?

VLADIMIR: Oh . . . (*He reflects.*) We wouldn't have to go into the details.

ESTRAGON: Our being born? *Vladimir breaks into a hearty laugh which he immediately stifles, his hand pressed to his pubis, his face contorted.*

Motivation

Didascalia

VLADIMIR:(*musingly*). The last moment . . . (*He meditates.*) Hope deferred maketh the something sick, who said that?

ESTRAGON: Why don't you help me?

VLADIMIR: Sometimes I feel it coming all the same. Then I go all queer. (*He takes off his hat, peers inside it, feels about inside it, shakes it, puts it on again.*) How shall I say? Relieved and at the same time . . . (*he searches for the word*) . . . appalled. (*With emphasis.*) AP-PALLED. (*He takes off his hat again, peers inside it.*) Funny. (*He knocks on the crown as though to dislodge a foreign body, peers into it again, puts it on again.*) Nothing to be done. (*Estragon with a supreme effort succeeds in pulling off his boot. He peers inside it, feels about inside it, turns it upside down, shakes it, looks on the ground to see if anything has fallen out, finds nothing, feels inside it again, staring sightlessly before him.*) Well?

ESTRAGON: Nothing.

VLADIMIR: Show me.

ESTRAGON: There's nothing to show.

VLADIMIR: Try and put it on again.

ESTRAGON:(*examining his foot*). I'll air it for a bit.

VLADIMIR: There's man all over for you, blaming on his boots the faults of his feet. (*He takes off his hat again, peers inside it, feels about inside it, knocks on the crown, blows into it, puts it on again.*) This is getting alarming. (*Silence. Vladimir deep in thought, Estragon pulling at his toes.*) One of the thieves was saved. (*Pause.*) It's a reasonable percentage. (*Pause.*) Gogo.

ESTRAGON: What?

VLADIMIR: Suppose we repented.

ESTRAGON: Repented what?

VLADIMIR: Oh . . . (*He reflects.*) We wouldn't have to go into the details.

ESTRAGON: Our being born? *Vladimir breaks into a hearty laugh which he immediately stifles, his hand pressed to his pubis, his face contorted.*

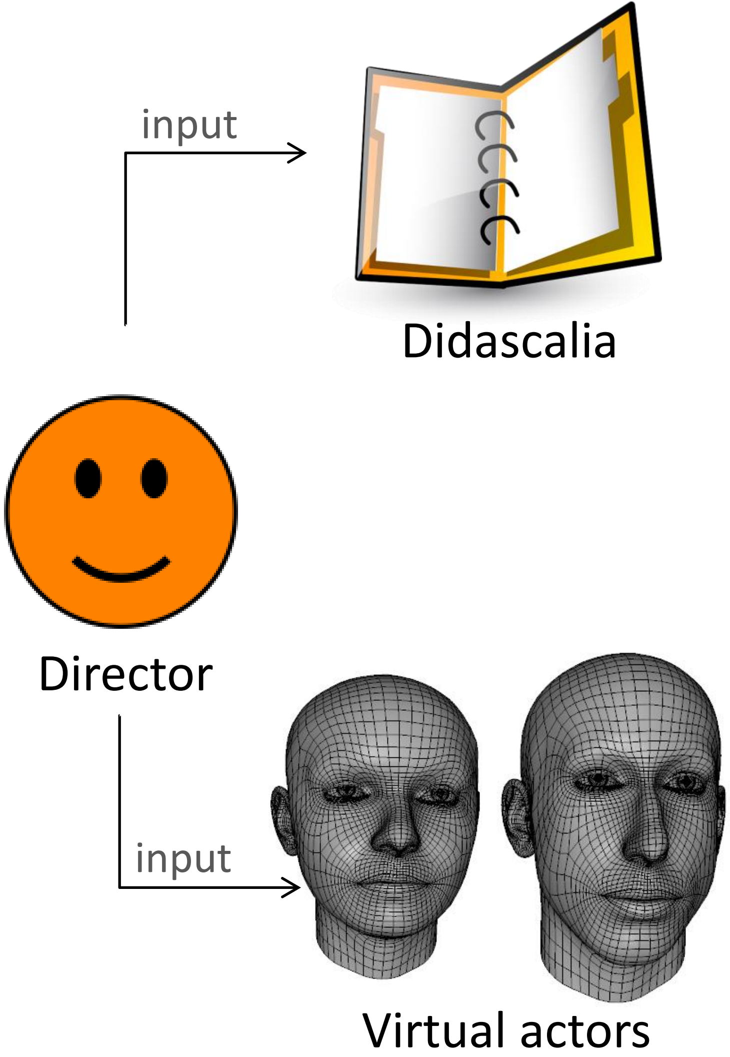
Motivation

Problems

- Choose the expressive style for each line
- Communicate the choice

Motivation

Our approach



input

Didascalia

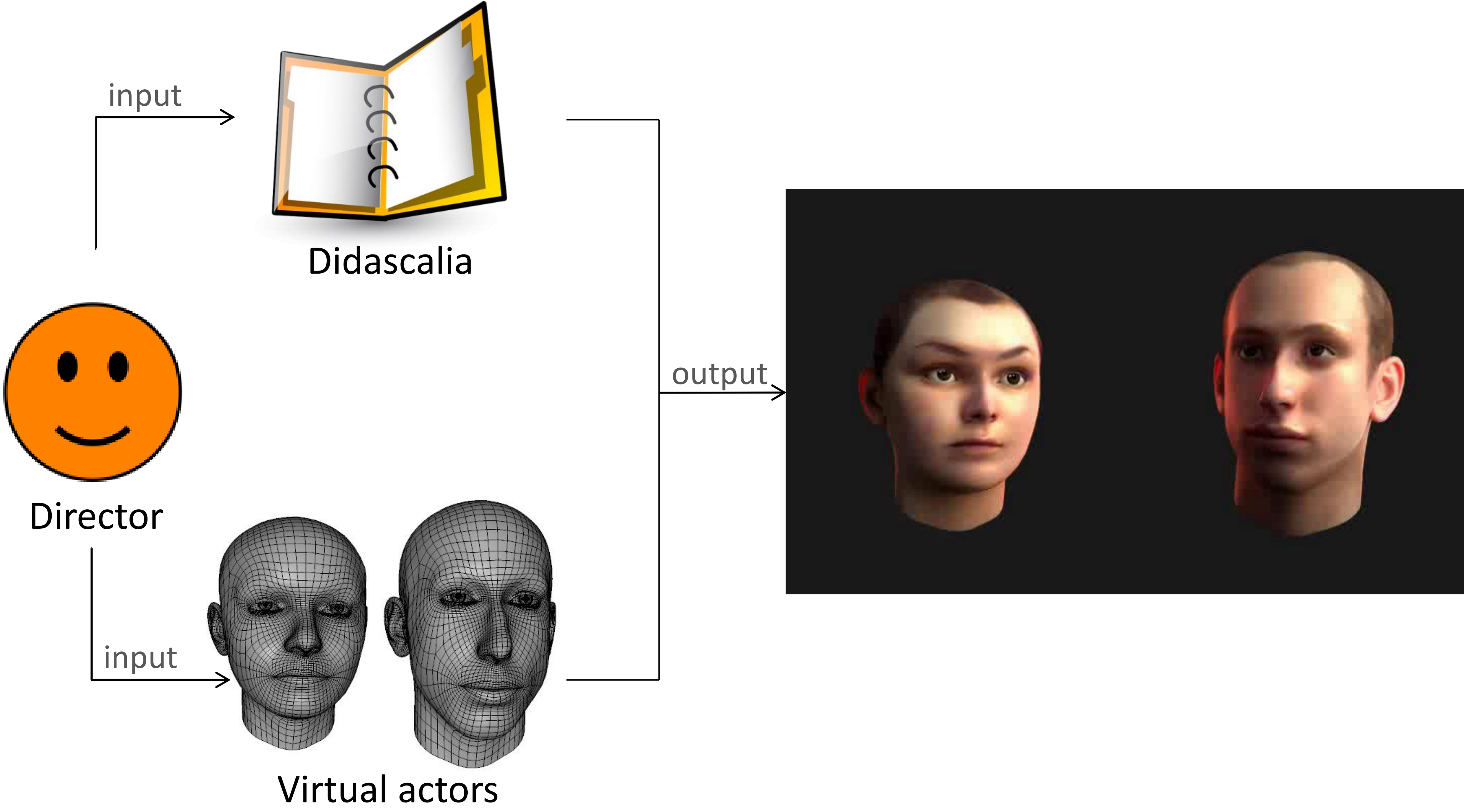
Director

input

Virtual actors

Motivation

Our approach



Plan

- Related work
  - Prosody
  - Expressive speech animation
- Dataset of dramatic attitudes & Analysis
- Generation of expressive performances & Evaluation
- Conclusion & Perspectives

Related work

Related work

Prosody

- □ Acoustic prosody [Hirst]

Speech = text + prosody

Neutral

Ironic

Related work

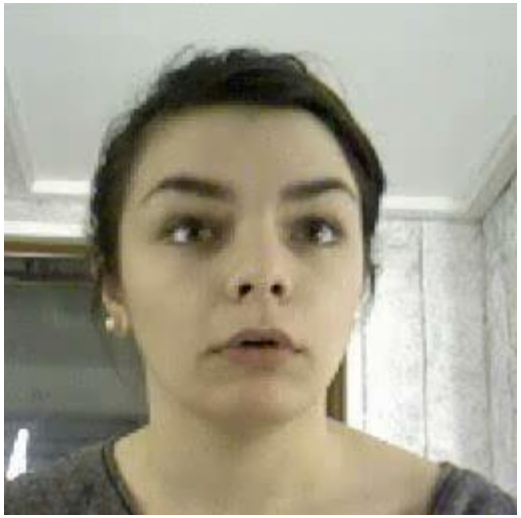
Prosody

- Acoustic prosody [Hirst]

Speech = text + prosody

- Audiovisual prosody

Neutral



Ironic

Related work

Prosody

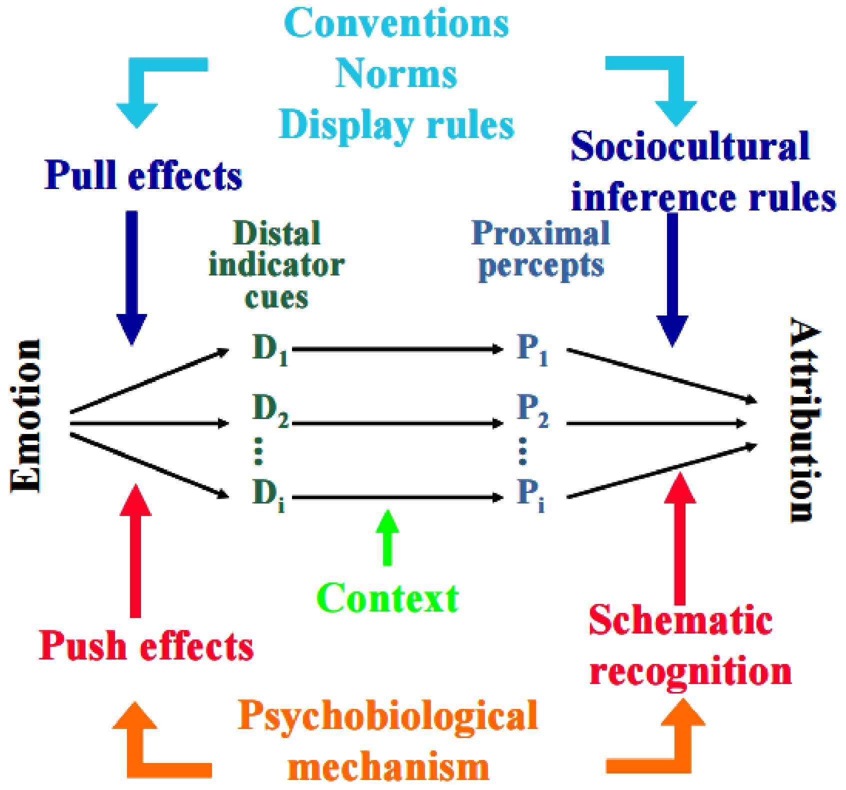
- Emotion vs attitude [Bolinger, 1989]

« How we feel when we say (emotions) and how we feel about what we say (attitudes) »

Related work

Prosody

- □ Emotion vs attitude [Bolinger, 1989]
- □ « Push / pull » effect [Scherer, 1986]

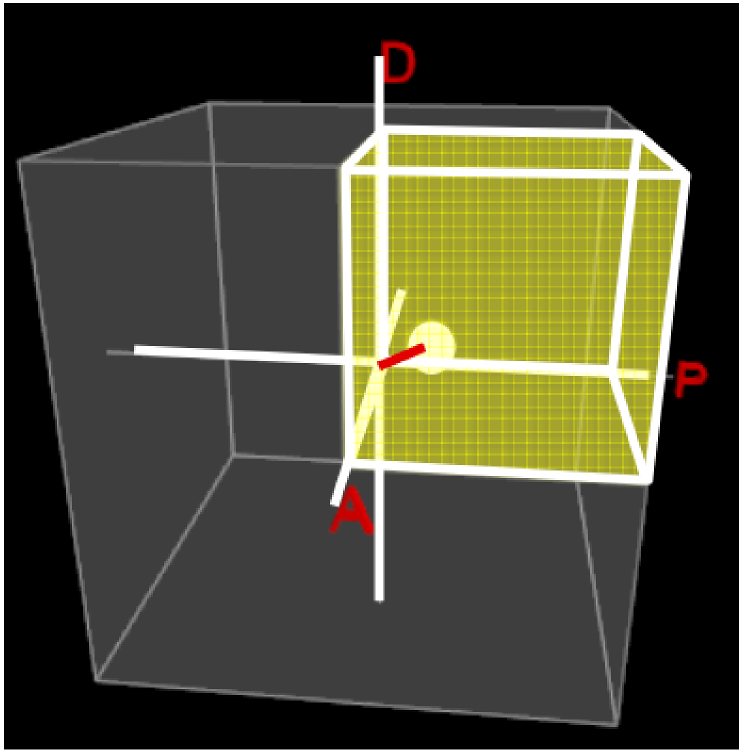


[Scherer, 2004]

Related work

Emotions

- Continuous
  - Arousal-Valence model [Russell, 1993]
  - Pleasure-Arousal-Dominance (PAD) model [Mehrabian, 1996]



Related work

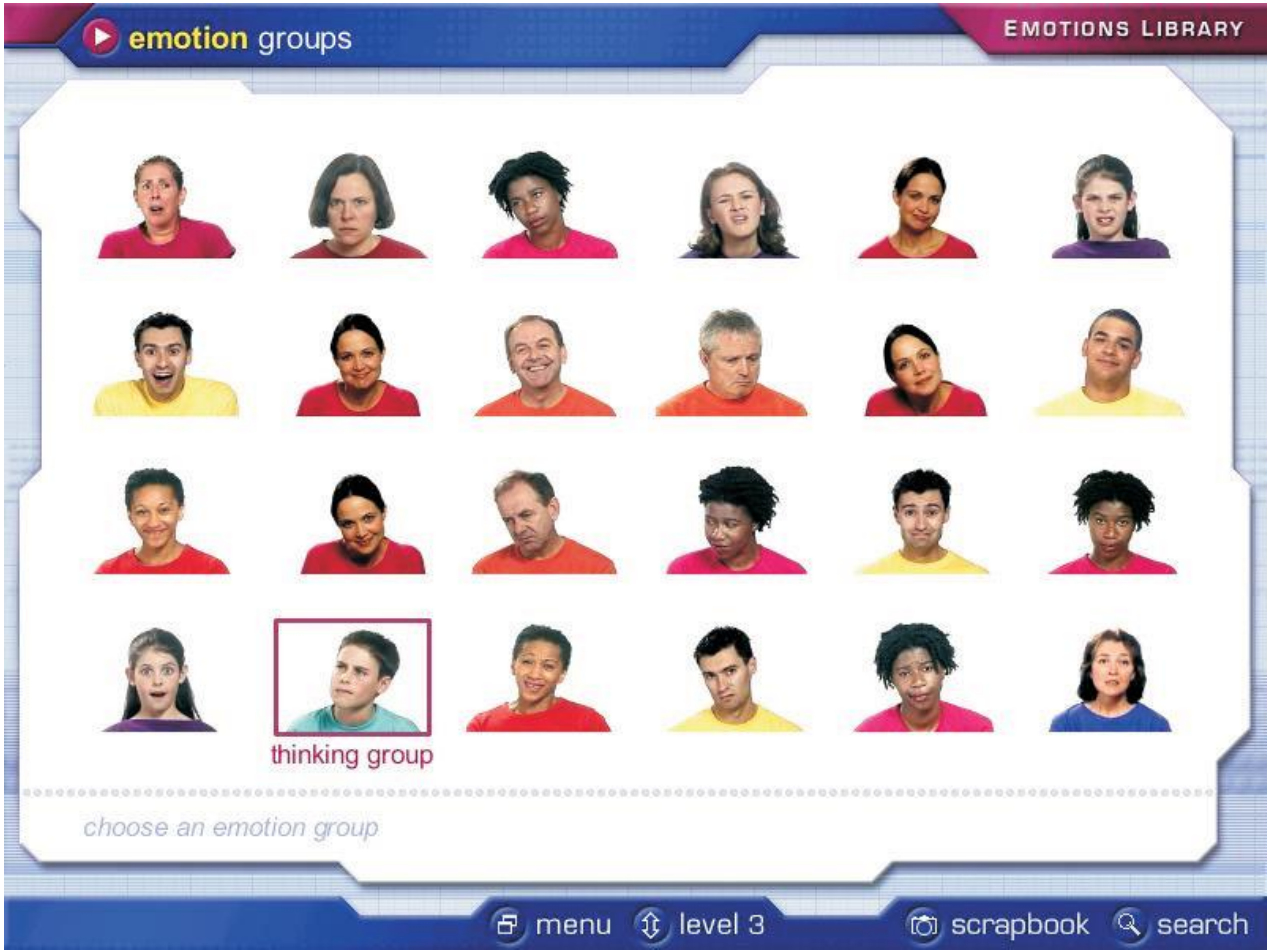
Emotions

- □ Discrete
  - □ Basic emotions [Ekman, 1971] : 6

Related work

Emotions

- Discrete
  - MindReading [Baron Cohen, 2004] : 412, under 24 categories

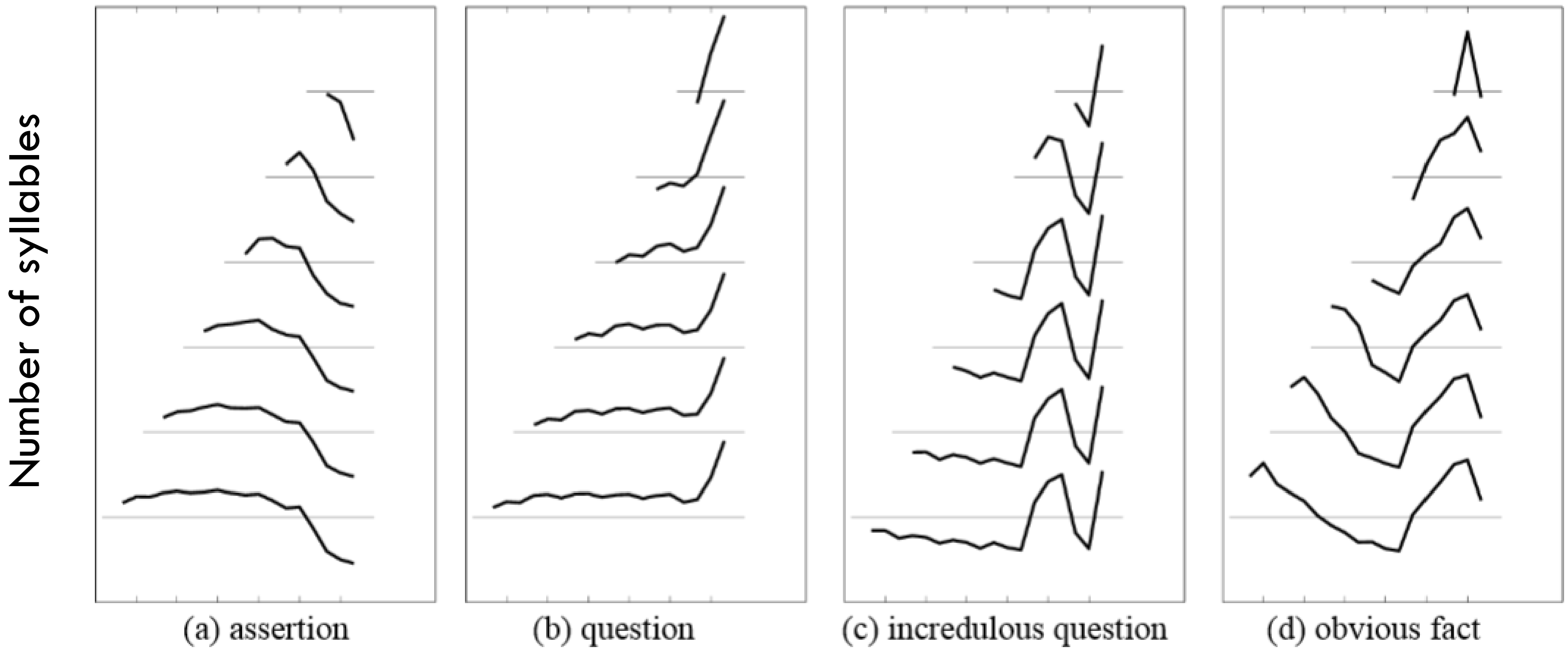


Related work

Attitudes

- Prosodic features present attitude-specific signatures which depend on the number of syllables [Fonagy, 1983] [Morlec, 2001] [Holm, 2005]

F0 contours



[Holm, 2005]

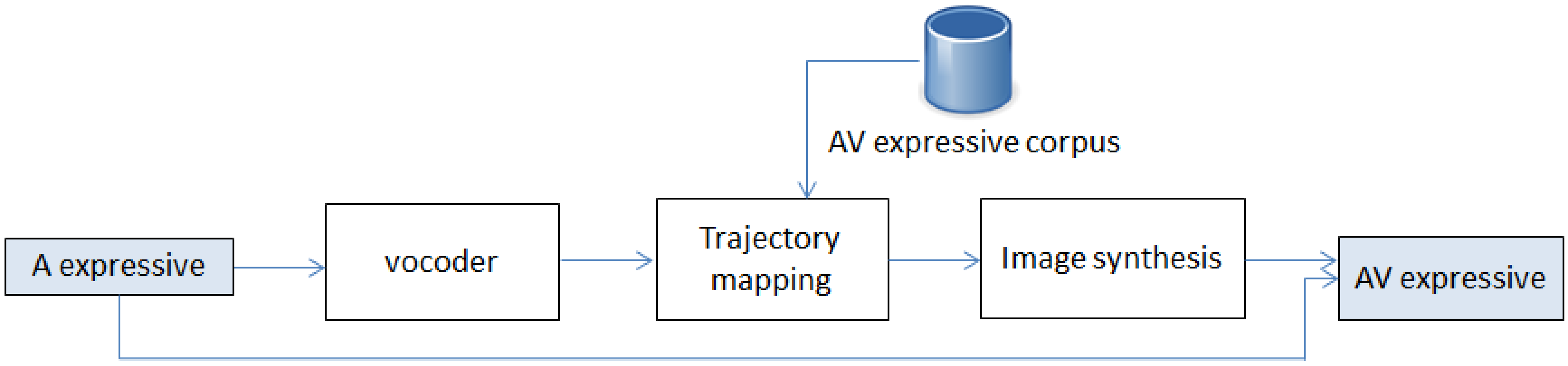
Related work

Expressive speech animation

- Speech-driven animation
- Visual text-to-speech
- Audio-visual conversion

Related work

Expressive speech-driven animation

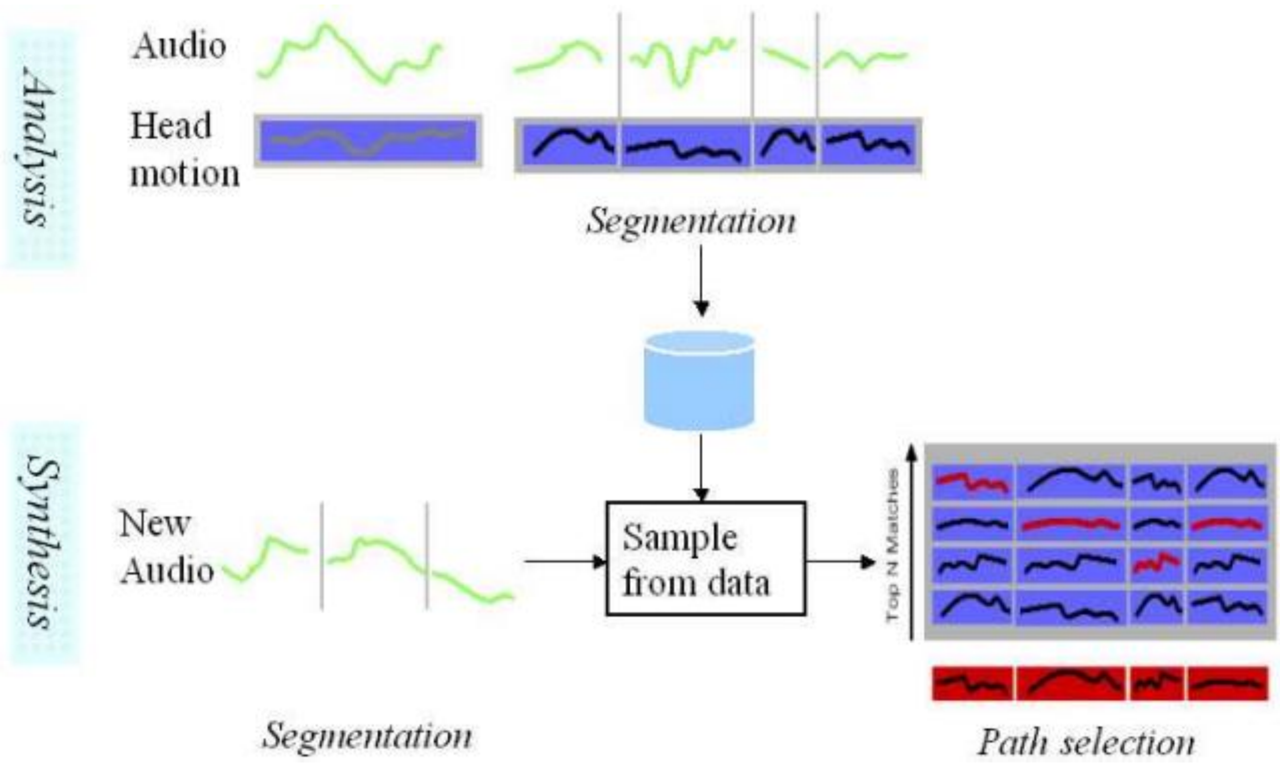


Author	Year	Features	Emotions / Attitudes
Bregler et al	**2005**	Facial expressions, head	Joy, Anger
Cao et al	2005	Facial expressions, head	Joy, Anger, Sadness, Frustration
Busso et al	2007	Head motion	Happiness, Sadness, Anger
Ding et al	2013	Eyebrows	Anger, Fear, Sadness, Surprise
Marsella et al	2013	Facial expressions, head, gaze, gestures	Uncertainty, awful etc

Related work

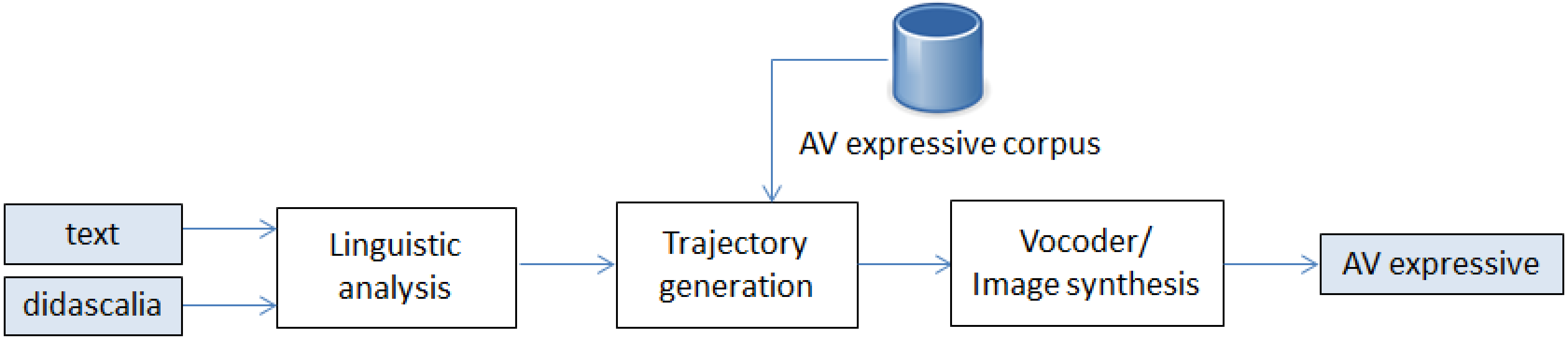
Expressive speech-driven animation

- □ Bregler et al, 2005
- □ Exemplar-based head motion synthesis: 67 phrases



Related work

Expressive visual text-to-speech



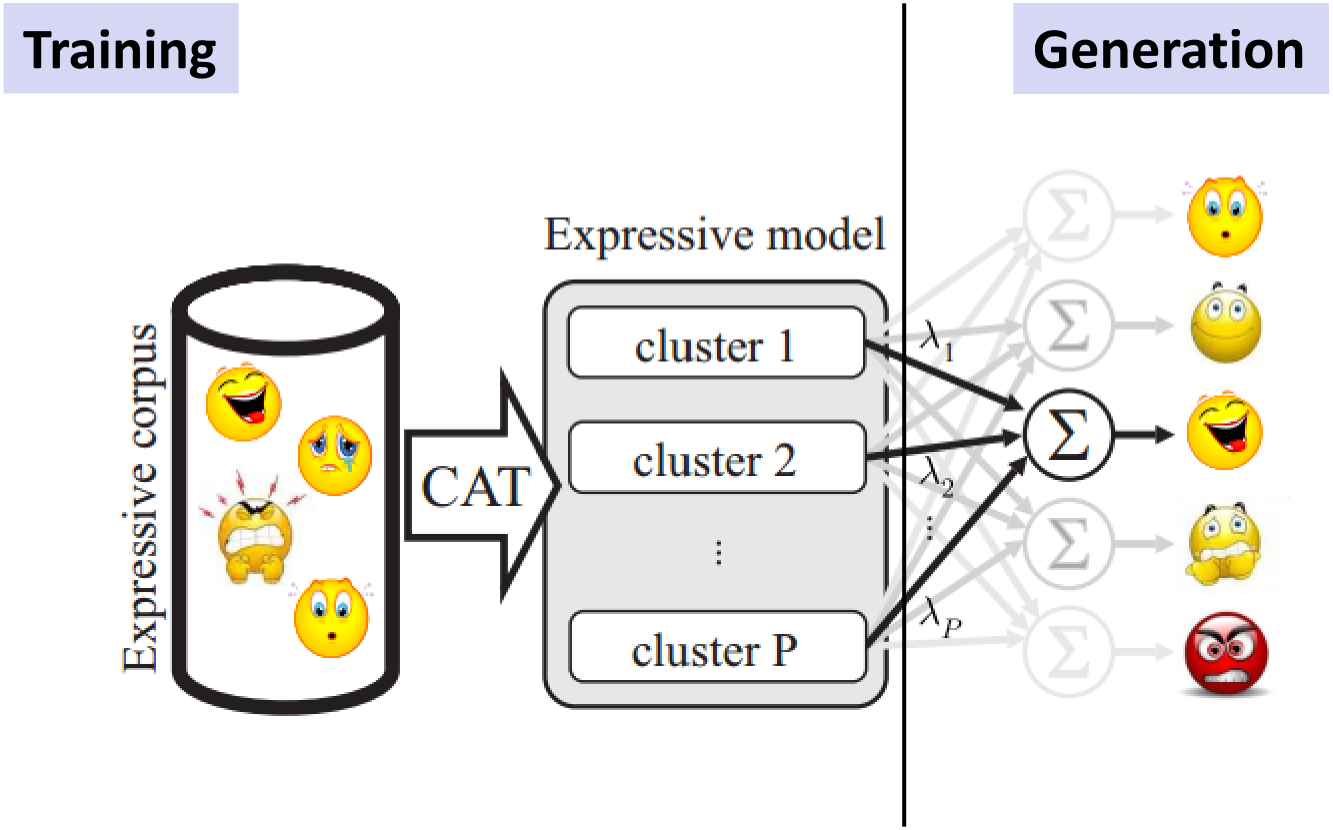
Author	Year	Features	Emotions / Attitudes
Pelachaud et al	1996	Facial expressions, head, gaze	Happiness, Sadness, Surprise, Anger, Fear, Disgust
Albrecht et al	2002	Facial expressions, head, gaze	Happiness, Sadness, Surprise, Anger, Kidding, Disgust
Liu et al	2011	Smiling	Happiness
Anderson et al	**2013**	Voice, facial expressions, head	Tenderness, Happiness, Fear Sadness, Anger
Jia et al	2014	Facial expressions, head	PAD model: 12 expressions (Happy, Surprise, Anxious etc)

Related work

Expressive visual text-to-speech

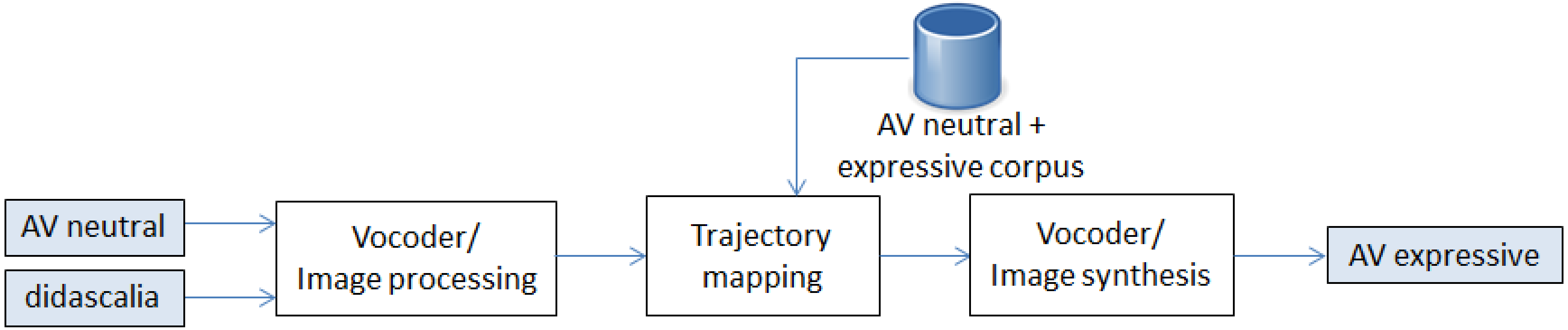
□ Anderson et al, 2013

□ Cluster Adaptive Training: >1000 sentences per style



Related work

Expressive audiovisual conversion



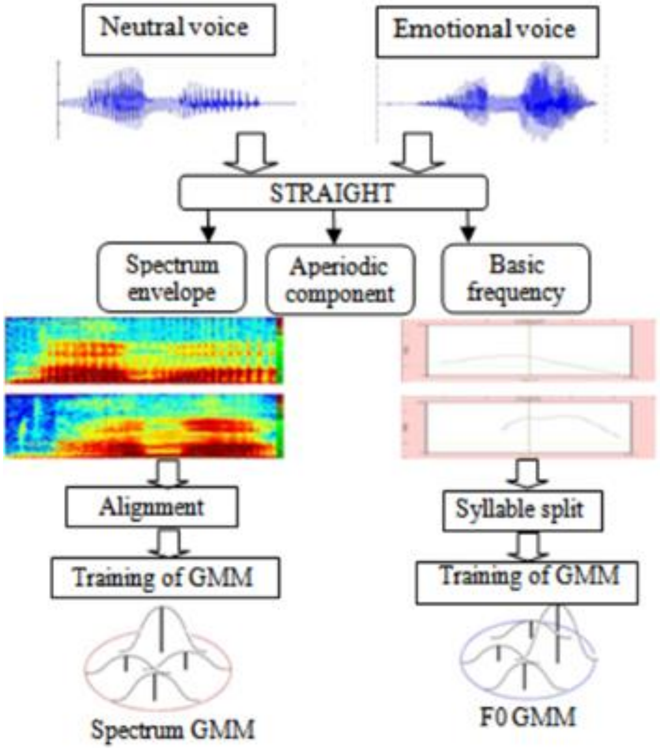
Author	Year	Features	Emotions / Attitudes
Mori et al	2006	Voice	Anger, Boredom, Depression
Veaux et al	2011	Voice	Joy, Fear, Sadness, Anger
Aihara et al	**2012**	Voice	Anger, Sadness, Joy
Ma et al	2009	Facial expressions, head	Anger, Joy
Shaw et al	2013	Facial expressions	undefined

Related work

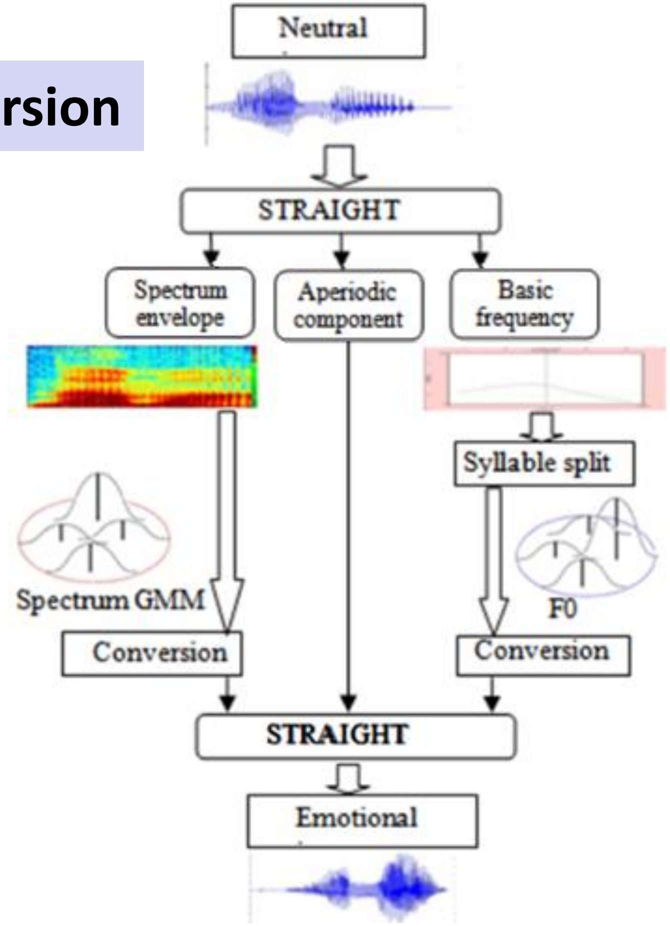
Expressive audiovisual conversion

- Aihara et al, 2012
- Gaussian Mixture Models (GMMs): 20 words per style

Training



Conversion



Related work

Critical review

- Expressive audiovisual speech
  - Taxonomies
  - Dynamic of contours
- Rhythm
  - Local, global
- Units
  - Frames, syllables

Related work

Critical review

- Expressive audiovisual speech
  - Taxonomies → Discrete attitudes
  - Dynamic of contours
- Rhythm
  - Local, global
- Units
  - Frames, syllables

Related work

Critical review

- □ Expressive audiovisual speech
  - □ Taxonomies → Discrete attitudes
  - □ Dynamic of contours → Visual contour signatures
- □ Rhythm
  - □ Local, global
- □ Units
  - □ Frames, syllables

Related work

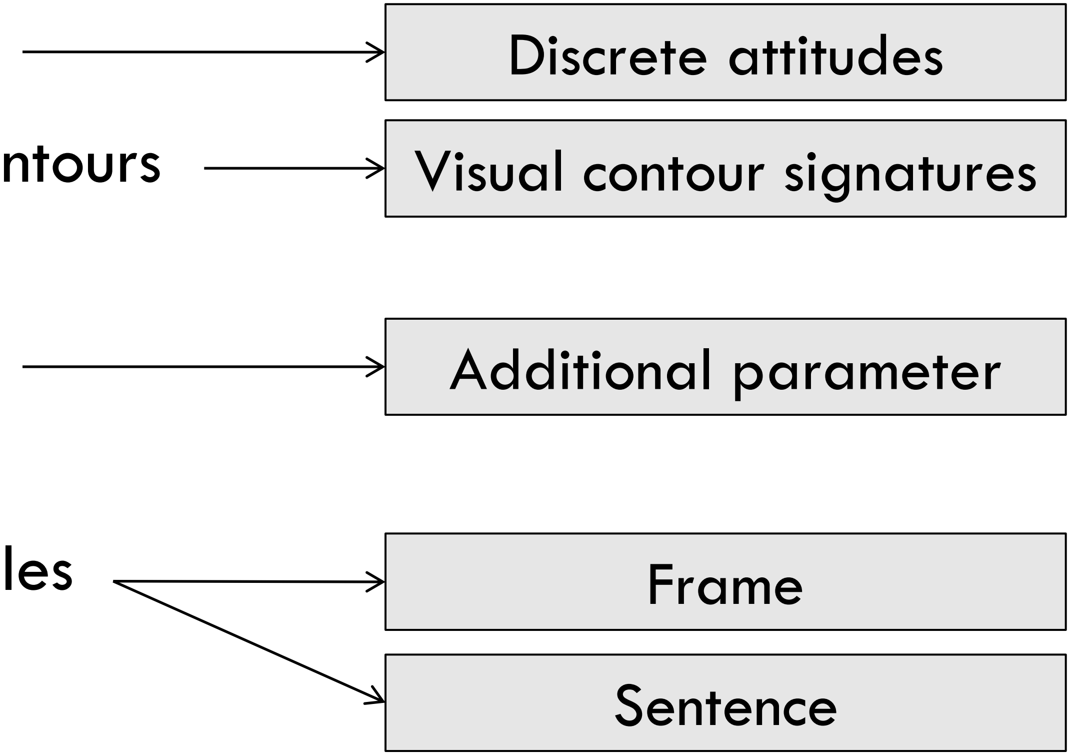
Critical review

- □ Expressive audiovisual speech
  - □ Taxonomies → Discrete attitudes
  - □ Dynamic of contours → Visual contour signatures
- □ Rhythm
  - □ Local, global → Additional parameter
- □ Units
  - □ Frames, syllables

Related work

Critical review

- □ Expressive audiovisual speech
  - □ Taxonomies
  - □ Dynamic of contours
- □ Rhythm
  - □ Local, global
- □ Units
  - □ Frames, syllables



Discrete attitudes

Visual contour signatures

Additional parameter

Frame

Sentence

Dataset of dramatic attitudes
&
Analysis

Dataset of dramatic attitudes & Analysis

Plan

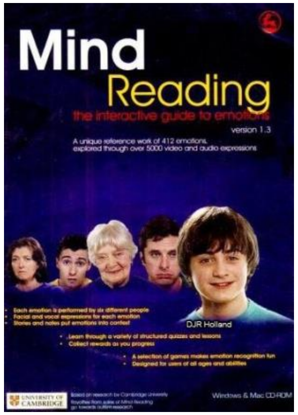
- Dataset of dramatic attitudes
  - Recording
  - Auto-evaluation
- Analysis
  - Frame-level
  - Syllable-level
  - Assessment of performances

Dataset of dramatic attitudes

Recording

- □ 35 sentences from « La ronde » [Arthur Schnitzler, 1920]
- □ 13 dramatic attitudes from Mind Reading [Baron Cohen, 2004]
  + modalities (assertion, interrogation, exclamation)
- □ 1 director (Georges) + 2 actors (Lucie and Greg)
- □ « Exercices in style » [Queneau, 1947]

+



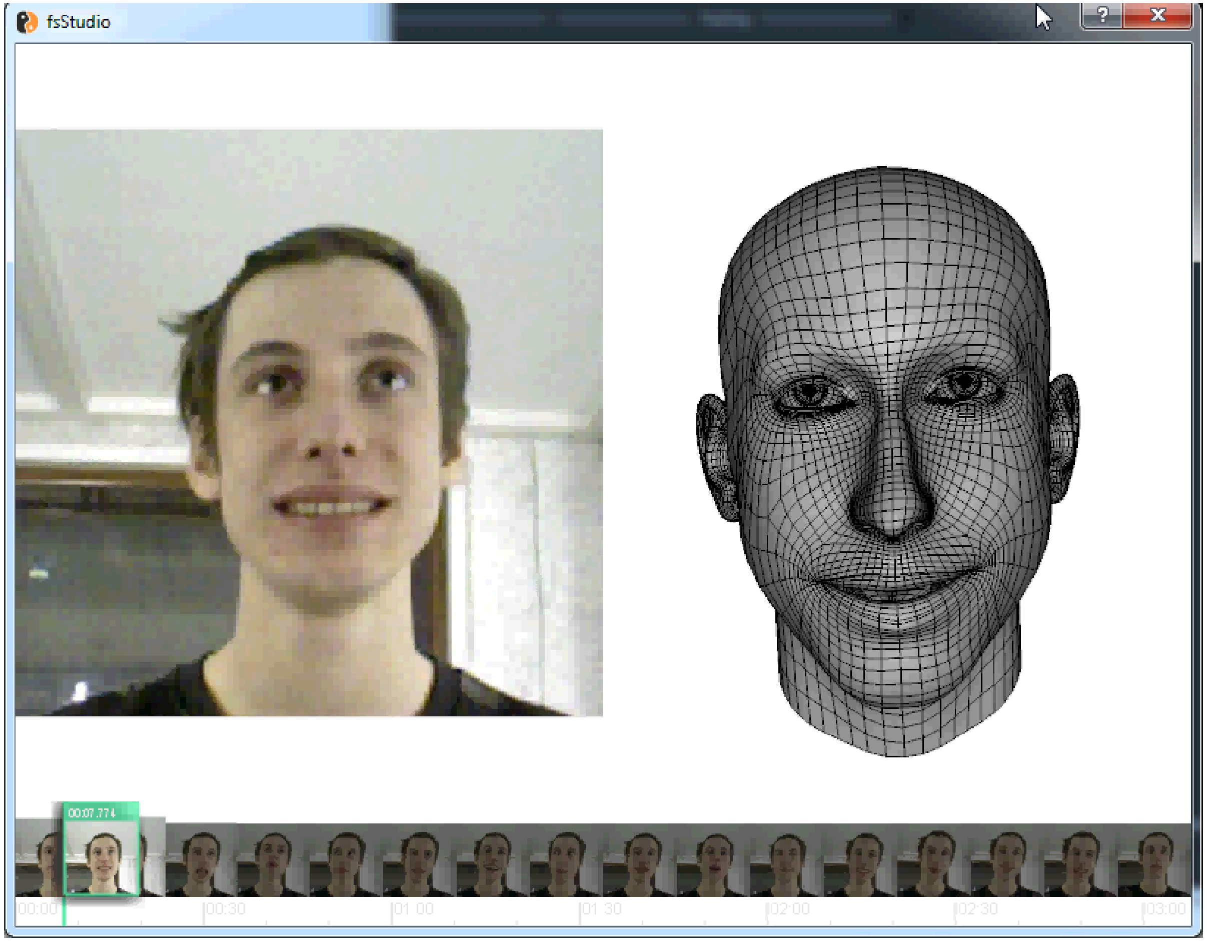
=

Dataset of dramatic attitudes

Faceshift

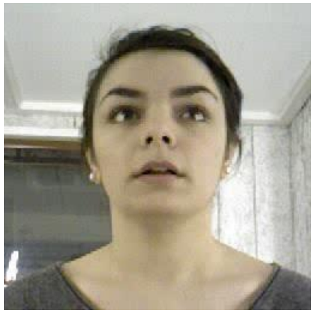
- □ Voice
- □ Head motion
- □ Facial expressions
- □ Eye gaze

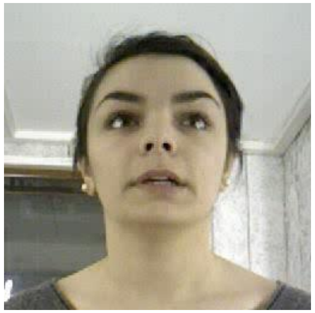
- □ No tongue

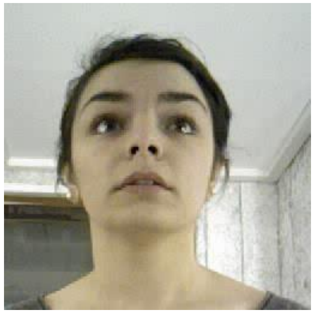


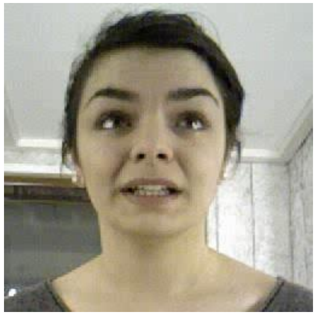
Dataset of dramatic attitudes

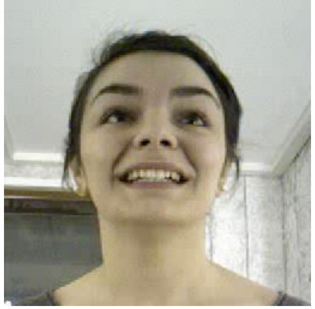
Attitudes in corpus

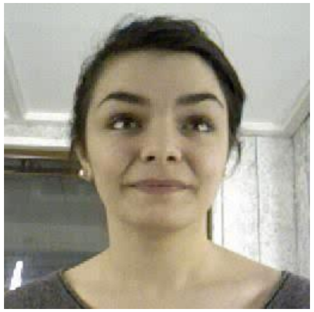

Declarative

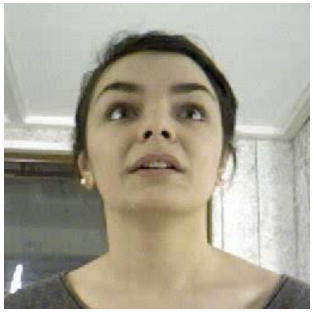

Exclamative

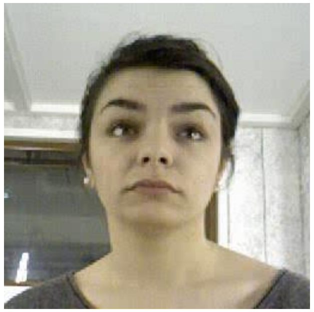

Interrogative

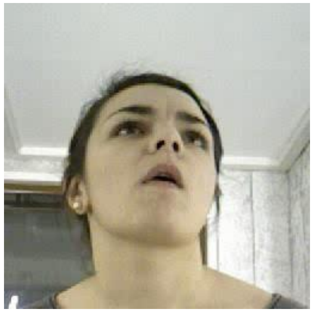

Comforting

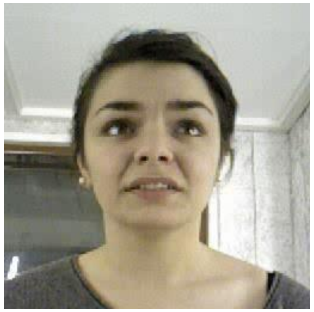

Tender

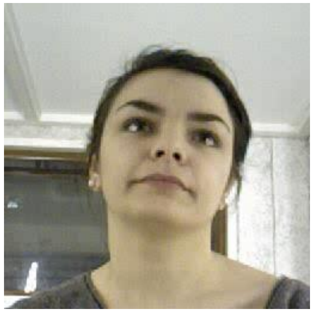

Seductive

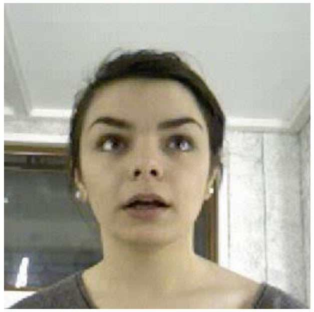

Fascinated

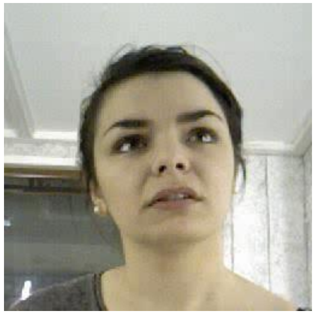

Jealous

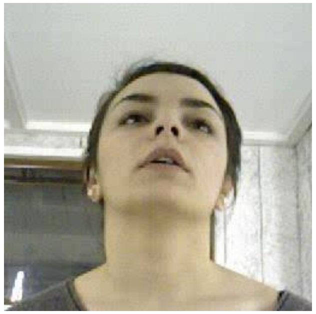

Thinking

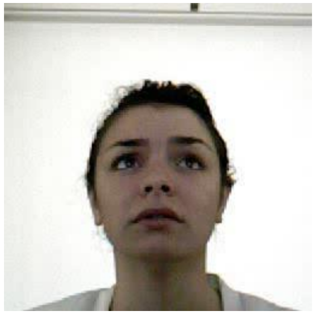

Doubtful

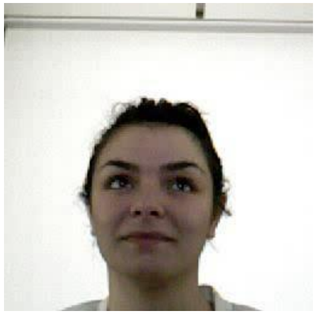

Ironic


Scandalized


Dazed


Responsible


Confronted

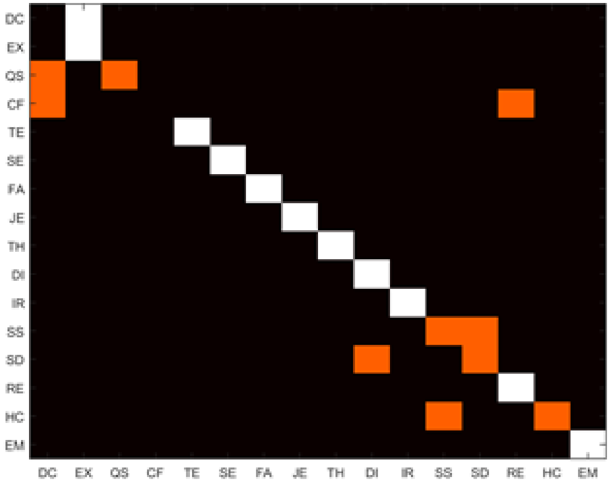

Embarrassed

Dataset of dramatic attitudes

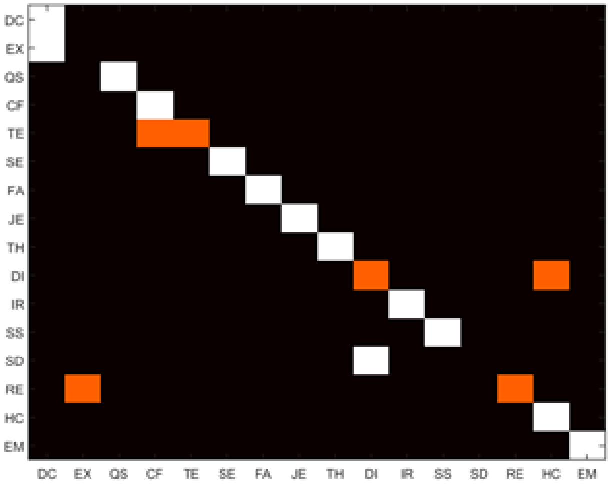
Auto-evaluation

□ Performances of Greg, Lucie, Georges: audio-only, video-only, audio-video

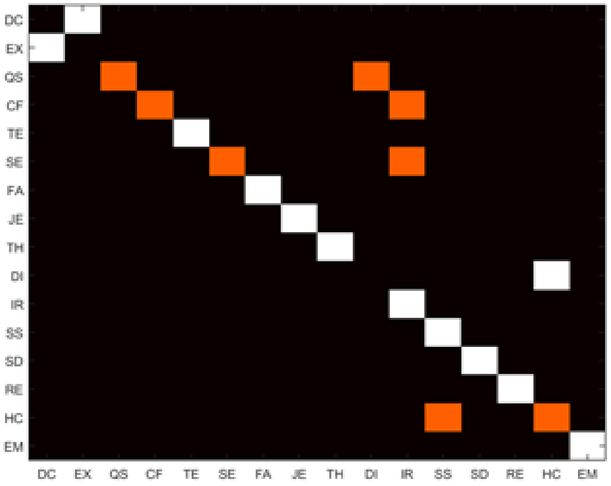
Greg 75%



Lucie 78,12%



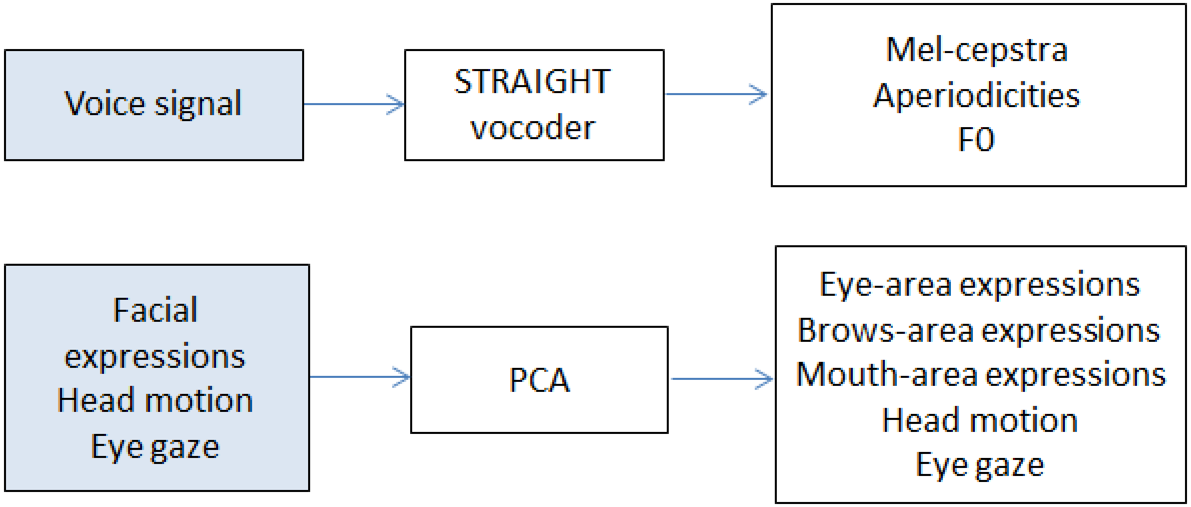
Georges 68,75%



Data analysis

Frame-level analysis

- Features extracted at each frame:



Data analysis

Feature characterization

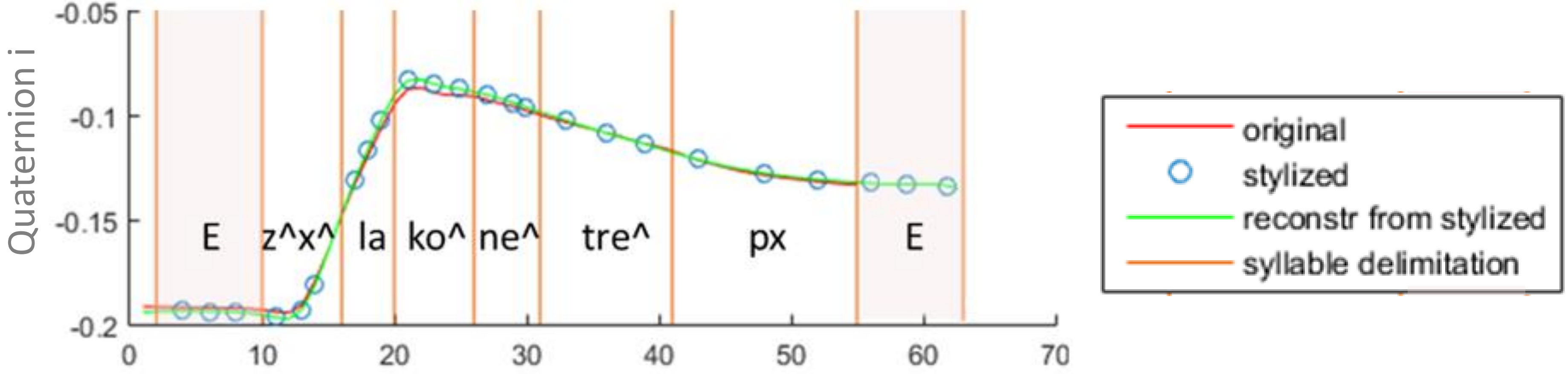
- □ Segmental / prosodic features

	Audio	Visual
Segmental	Mel-cepstra Aperiodicities	Mouth-area expressions
Prosodic	F0 Rhythm	Eye-area expressions Brows-area expressions Head motion Eye gaze Rhythm

Data analysis

Syllable-level analysis

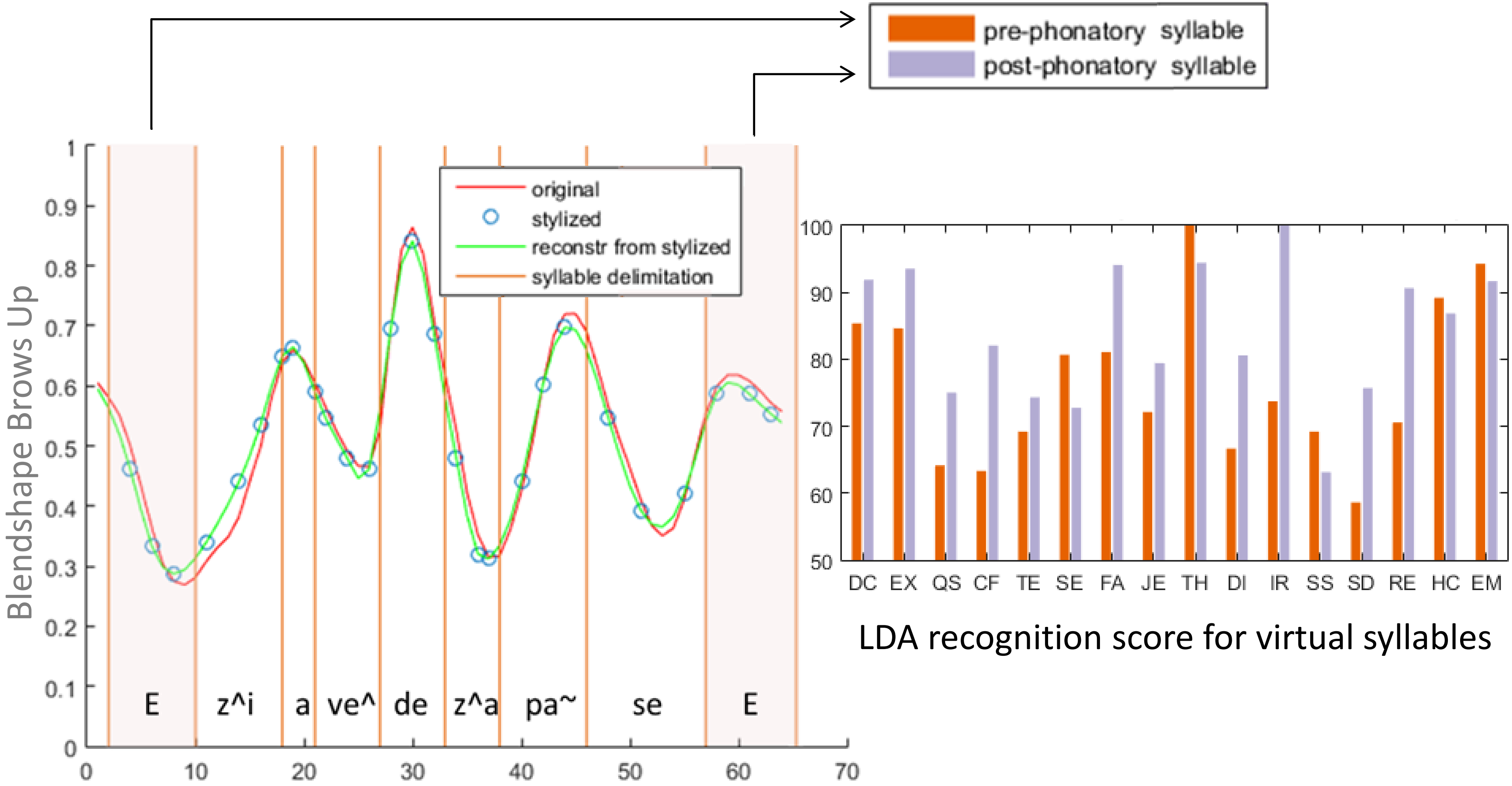
- Features extracted at each syllable (*stylization*):
  - **Melody:** 3 values extracted from the vocalic nucleus
  - **Motion:** 3 values extracted from the syllable
  - **Rhythm:** 1value, syllable elongation coefficient
- Motion for « virtual » silent syllables



Je la connais très peu (I know her very little).

Data analysis

Stylization + Virtual syllables



LDA recognition score for virtual syllables

J'y avais déjà pensé (I have already considered it).

Data analysis

Audio-visual vocoder + Reconstruction from stylization

- □ Audio-visual vocoder
- □ Rhythm: phonemic duration generation [Barbossa, 1997]
- □ Melody & motion: interpolation from stylized contours

Video

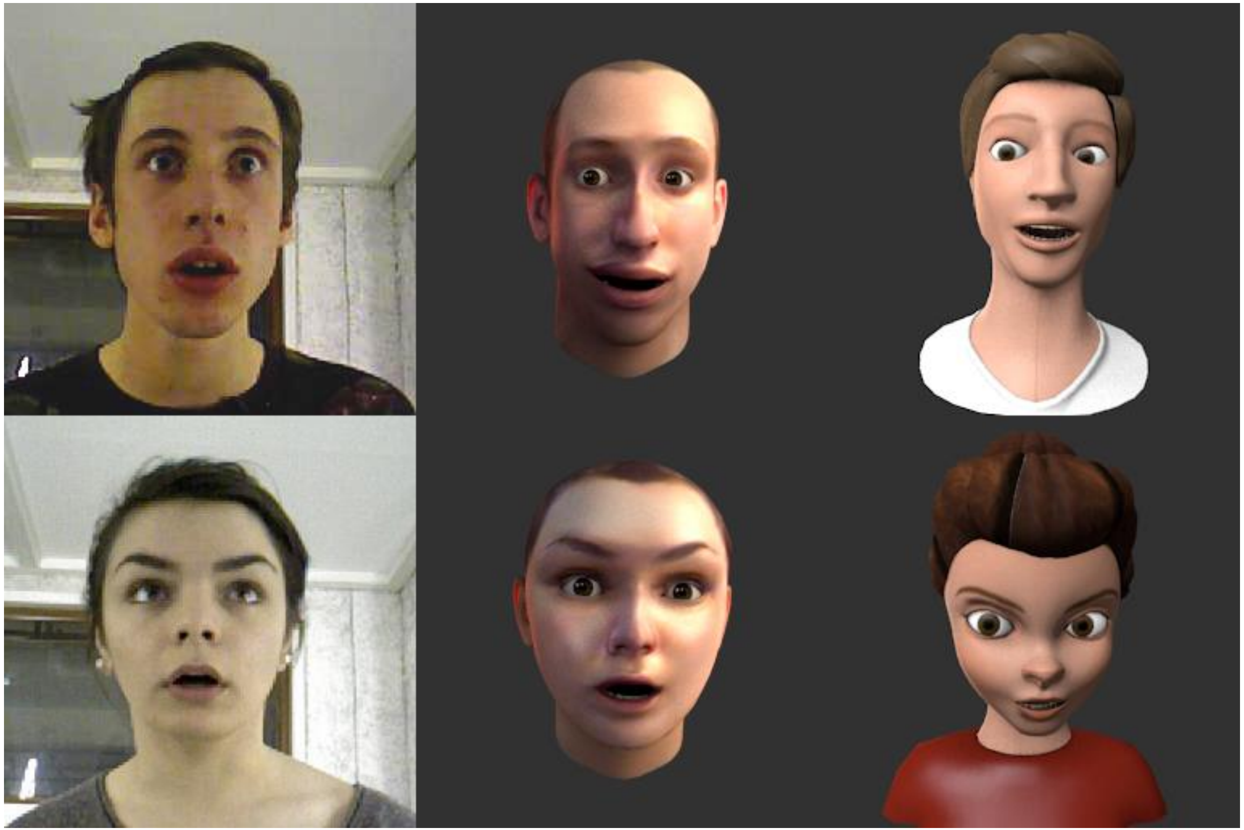
Original animation

Original reconstructed animation

Assessment of performances

Online evaluation

- □ Crowd-sourced platform
- □ Attitude recognition: video, animated (realistic and cartoon)
- □ Native French speakers



Assessment of performances

First perceptual test

- Performances of Lucie: audio-only, video-only, audio-video
- 80 participants

Audio-only

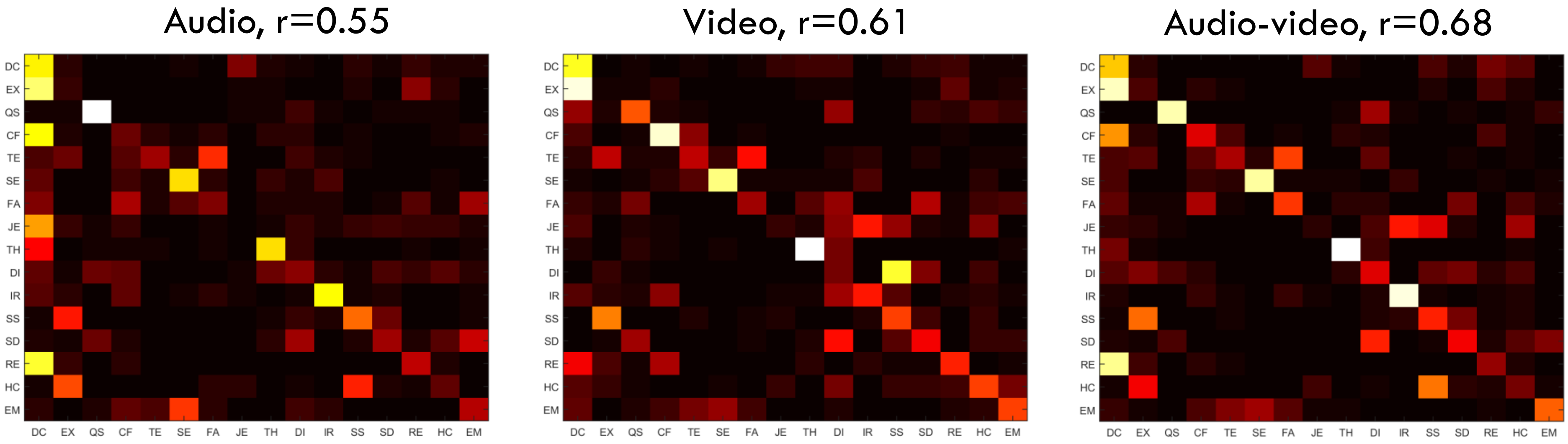
Video-only

Audio-Video

Assessment of performances

First perceptual test

- □ Performances of Lucie: audio-only, video-only, audio-video
- □ 80 participants



Audio, r=0.55

Video, r=0.61

Audio-video, r=0.68

Modality, cross-correlation with auto-evaluation matrix

Assessment of performances

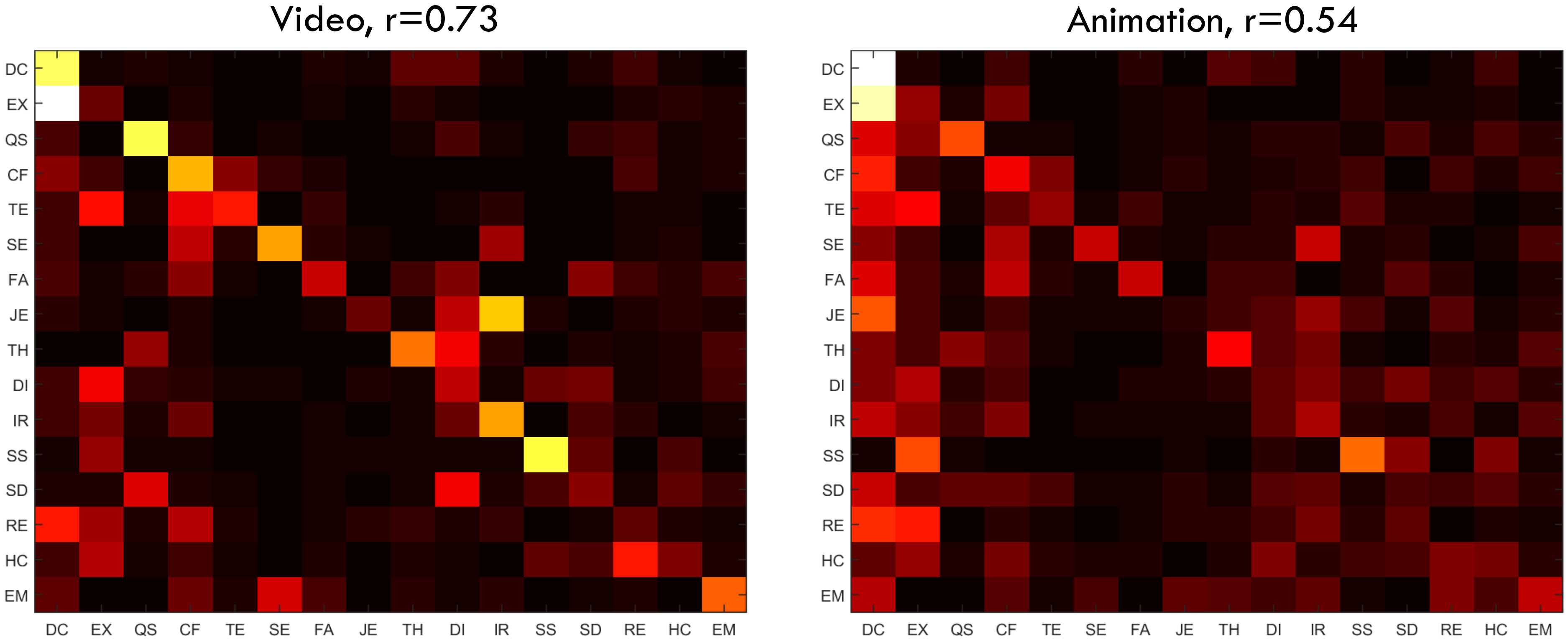
Second perceptual test

- □ Video and original reconstructed animations of Lucie
- □ 77 participants

Assessment of performances

Second perceptual test

- Video and original reconstructed animations of Lucie
- 77 participants

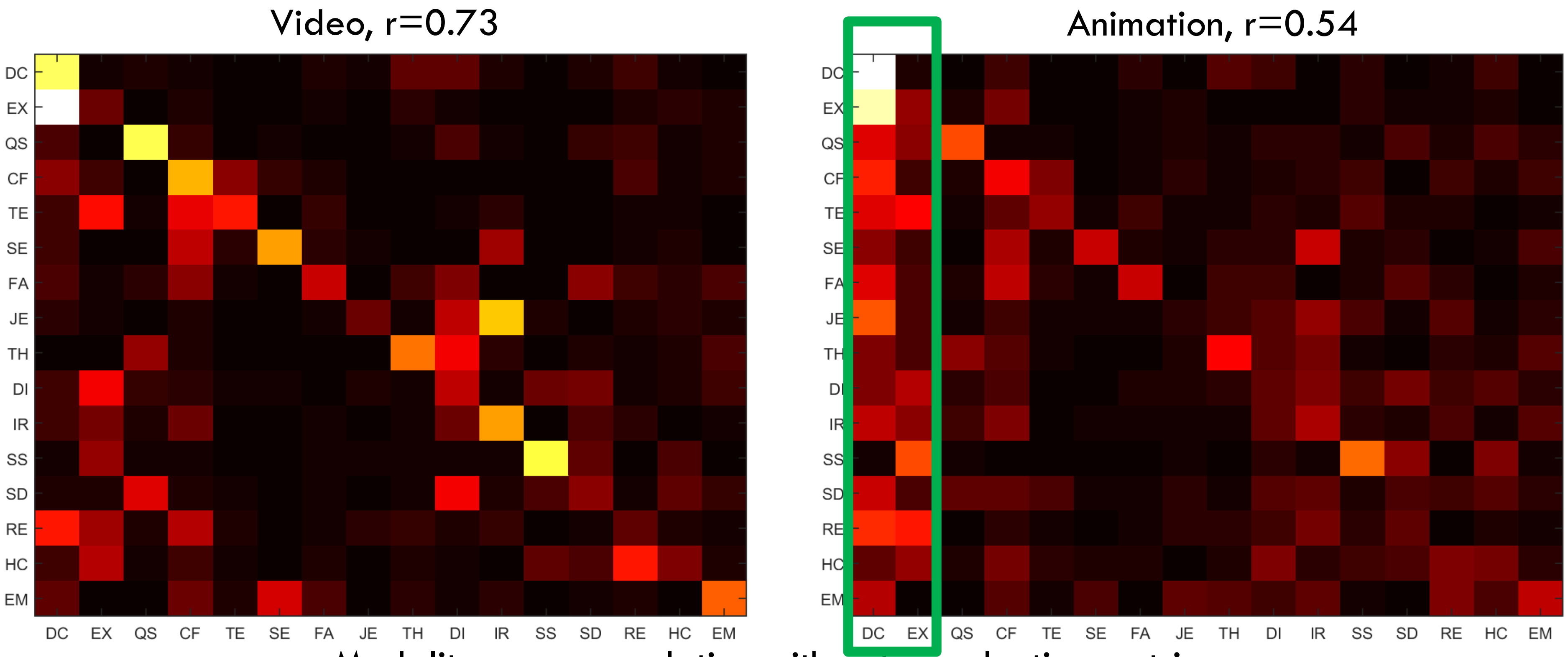


Modality, cross-correlation with auto-evaluation matrix

Assessment of performances

Second perceptual test

- □ Video and original reconstructed animations of Lucie
- □ 77 participants



Video, r=0.73

Animation, r=0.54

Modality, cross-correlation with auto-evaluation matrix

Assessment of performances

Third perceptual test

- □ Cartoon style original animations of Lucie and Greg
- □ 8 attitudes
- □ 53 participants

Assessment of performances

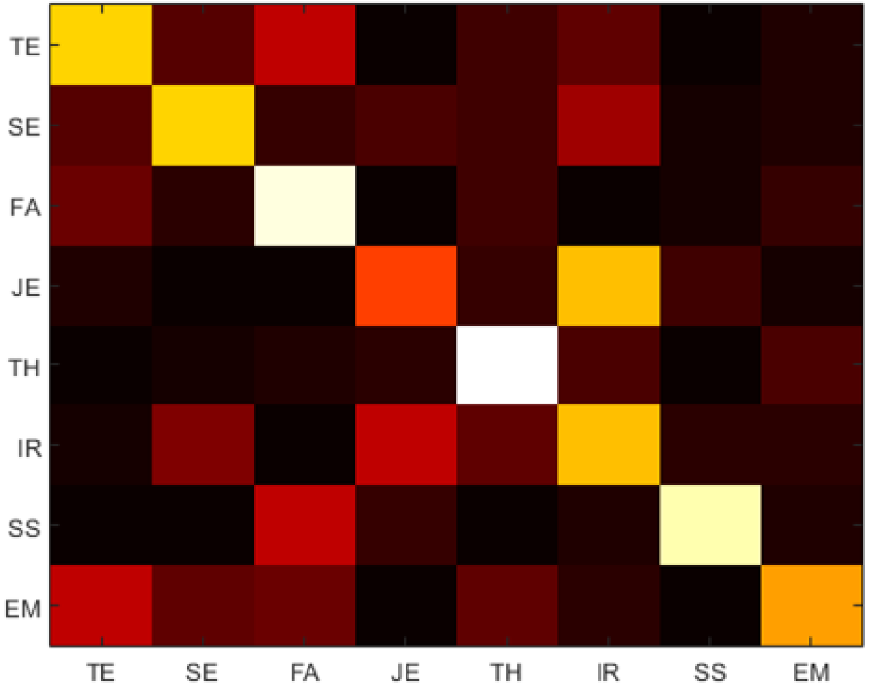
Third perceptual test

□ Cartoon style animated performances of Lucie and Greg

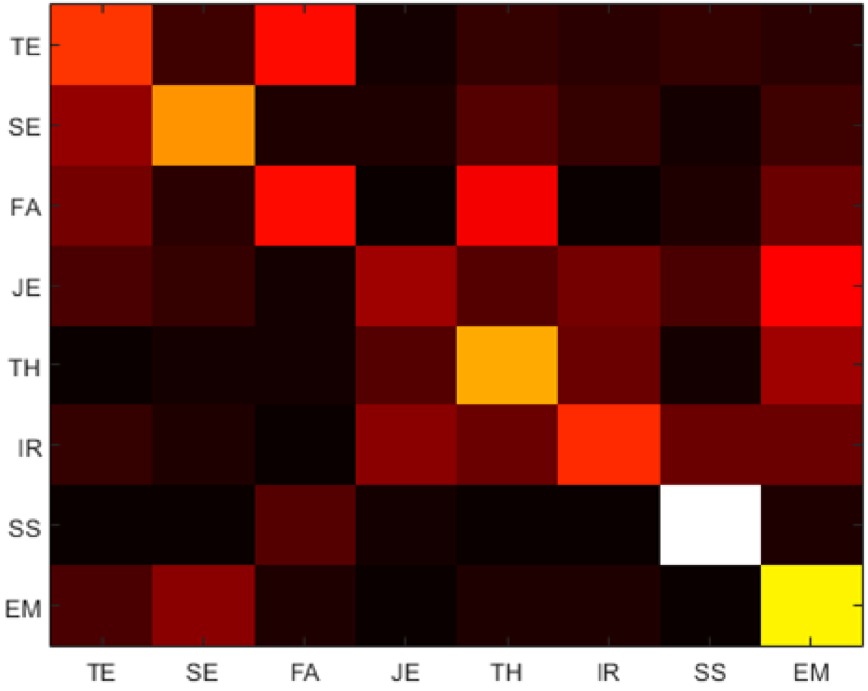
□ 8 attitudes

□ 53 participants

Greg



Lucie



Dataset of dramatic attitudes & Analysis

Summary

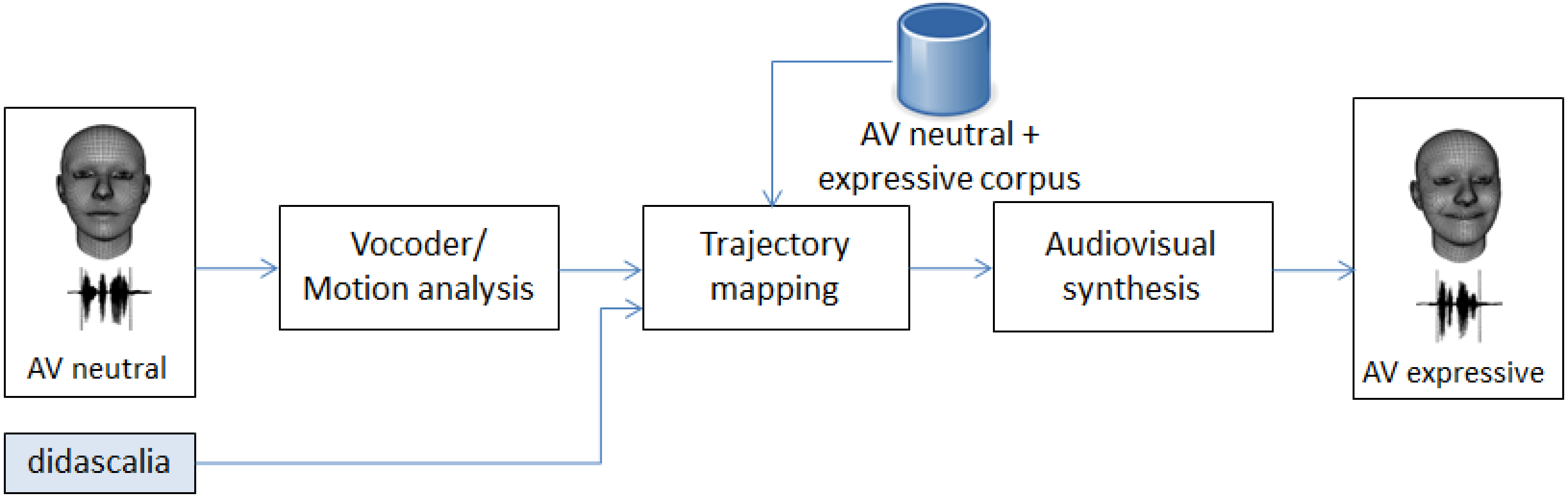
- Dataset of 16 dramatic attitudes, 35 sentences, 3 actors
- 1 hour of AV speech/actor
- High-dimensional feature space: Voice (31), head motion (5), facial expressions (24) and eye gaze (2), rhythm (1)
- Feature characterization: segmental (30+8) and prosodic (1+16+1)
- Stylization of syllabic units (1*3+16*3+1)
- 3 perceptual tests
- 7 attitudes: Comforting, Seductive, Fascinated, Thinking, Ironic, Scandalized, Embarrassed

Generation of expressive performances
&
Evaluation

Generation of expressive performances

Outline

- Expressive conversion from neutral performances
- Trajectory mapping
  - Frame-based
  - Model-based
  - Exemplar-based

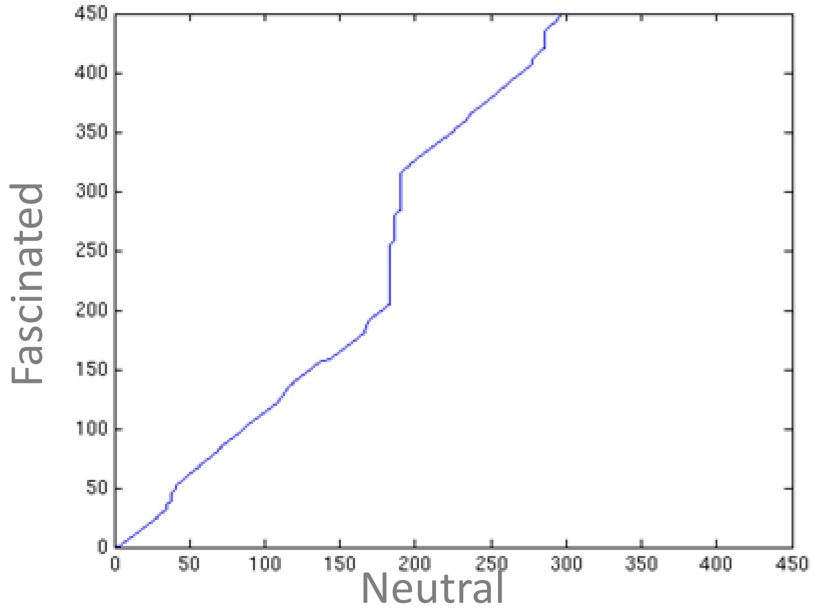


Generation of expressive performances

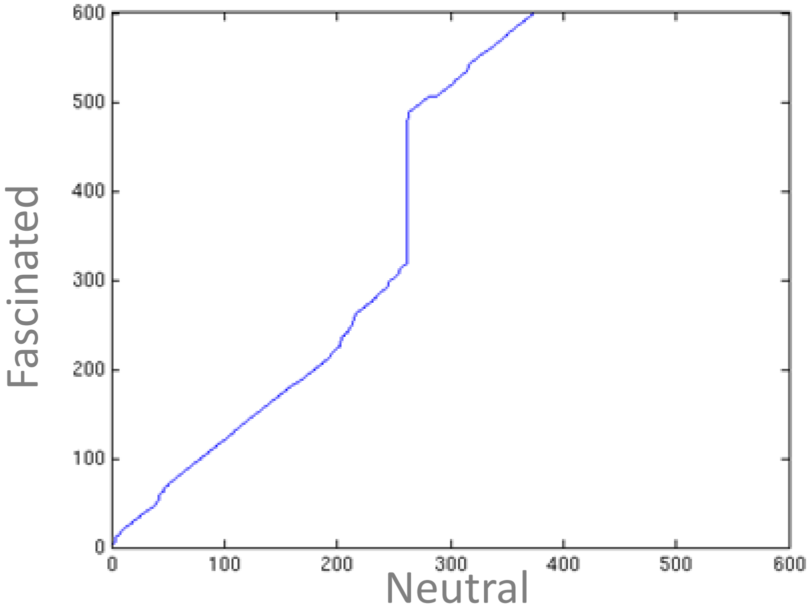
Frame-based

- Voice conversion: Gaussian Mixture Model (GMM) regression
- Frame-level unit
- Local speech rate prediction: slope feature
- Dynamic Time Warping (DTW) alignment

Greg

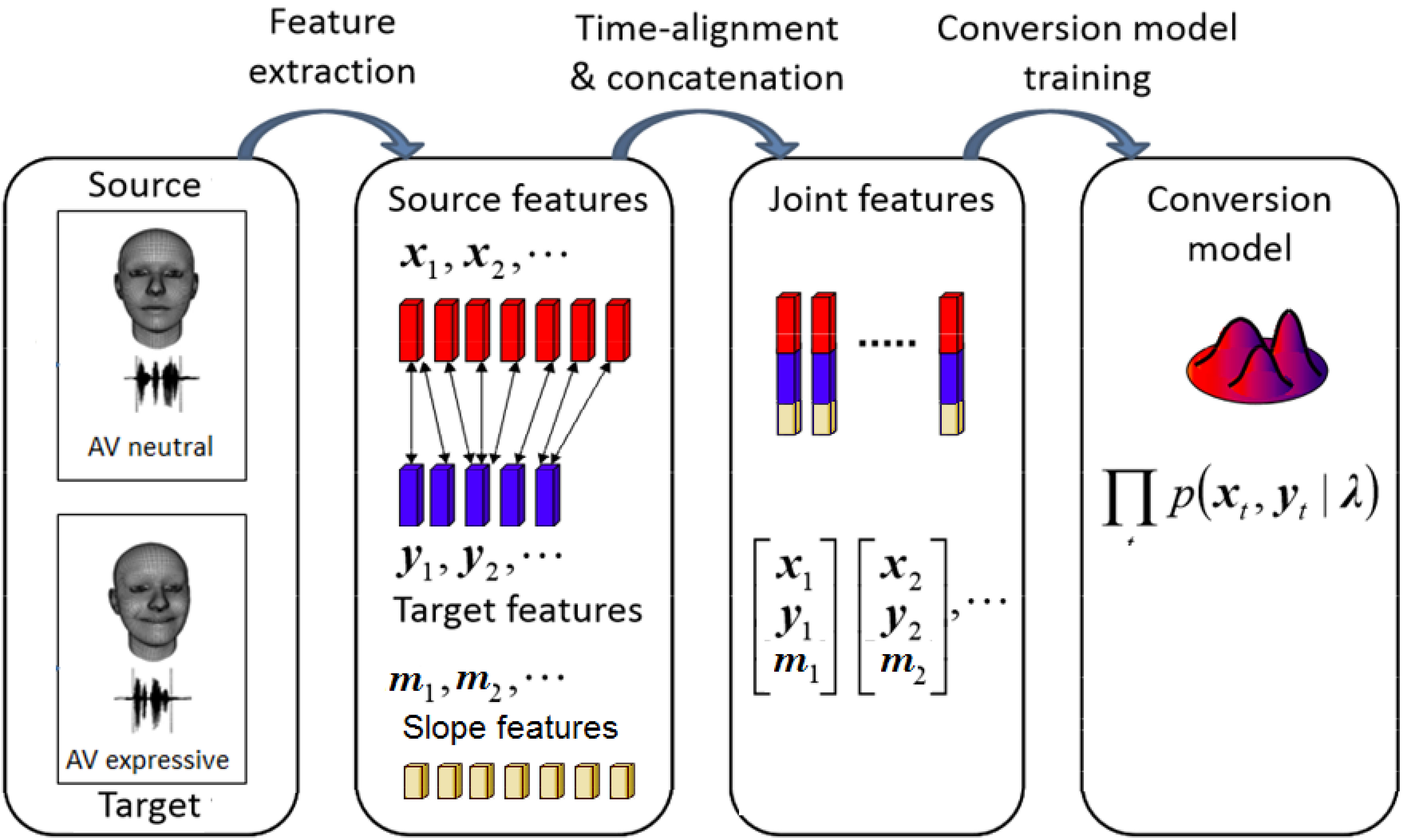


Lucie



Generation of expressive performances

Frame-based GMM training



Feature extraction

Time-alignment & concatenation

Conversion model training

Source

AV neutral

AV expressive

Target

Source features

$x_1, x_2, \cdots$

$y_1, y_2, \cdots$

Target features

$m_1, m_2, \cdots$

Slope features

Joint features

$\begin{bmatrix} x_1 \\ y_1 \\ m_1 \end{bmatrix} \begin{bmatrix} x_2 \\ y_2 \\ m_2 \end{bmatrix}, \cdots$

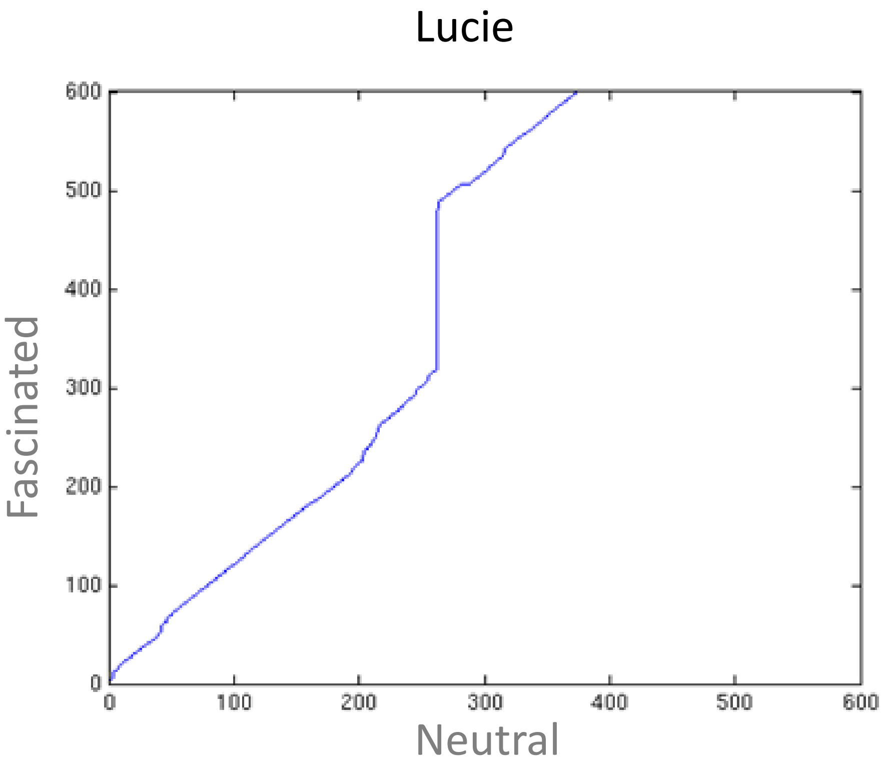
Conversion model

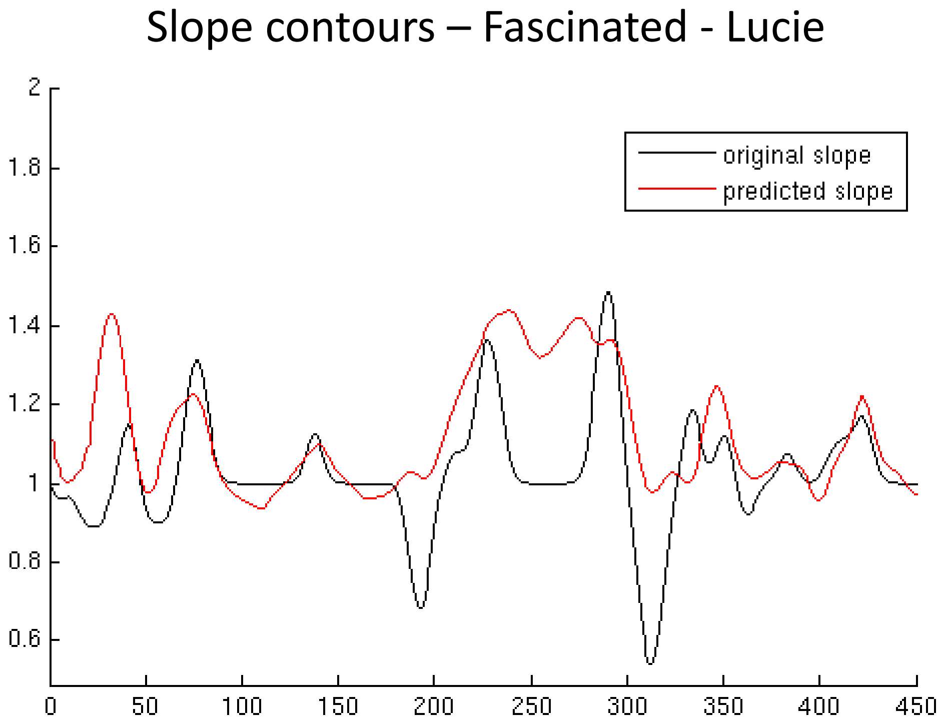
$\prod_t p(x_t, y_t \mid \lambda)$

Generation of expressive performances

Frame-based GMM regression

MMSE estimate: $\hat{\boldsymbol{y}}_t = \int \boldsymbol{y}_t p(\boldsymbol{y}_t \mid \boldsymbol{x}_t, \lambda)\,\mathrm{d}\boldsymbol{y}_t = \sum_{m=1}^{M} p(m \mid \boldsymbol{x}_t, \lambda)\boldsymbol{\mu}_{m,t}^{(y|x)}$





Generation of expressive performances

Frame-based results

Désormais, vous dînerez plus tôt (From now on, you will dine earlier)

Seductive

Neutral source

Expressive target

Frame-based

✓

Thinking

Neutral source

Expressive target

Frame-based

✗

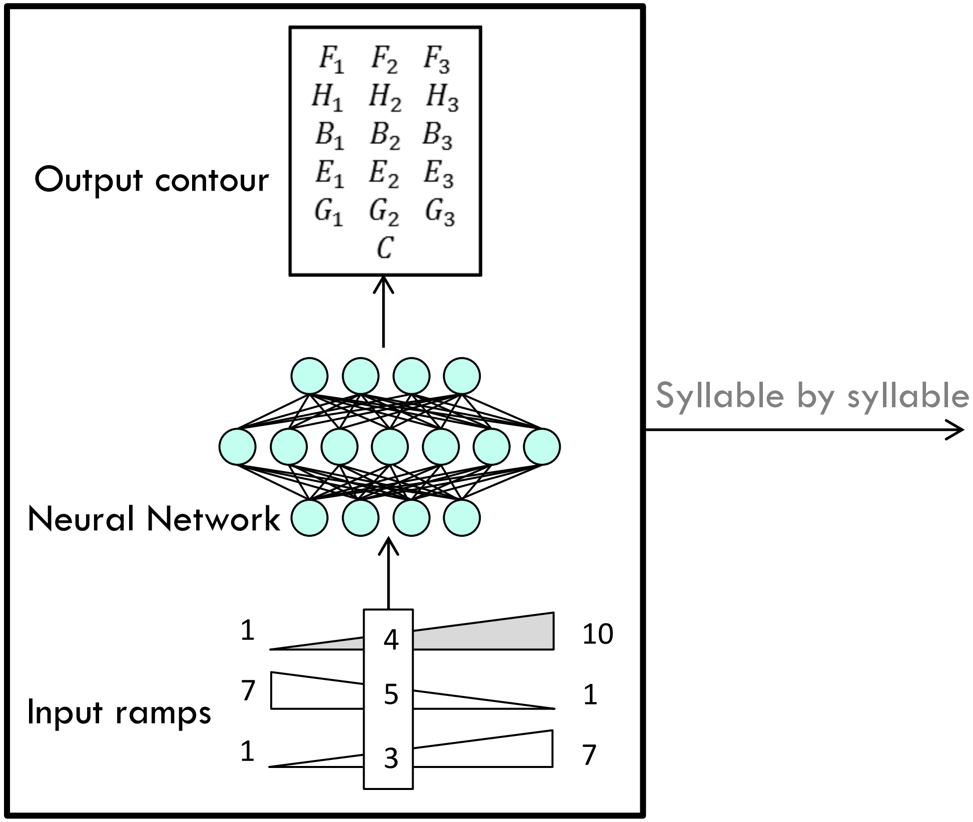
- F0

Generation of expressive performances

Model-based

- □ Separation segmental (GMMs)/ prosodic features
- □ Extend the SFC model [Holm, 2005] to include motion component
- □ Sentence-level unit

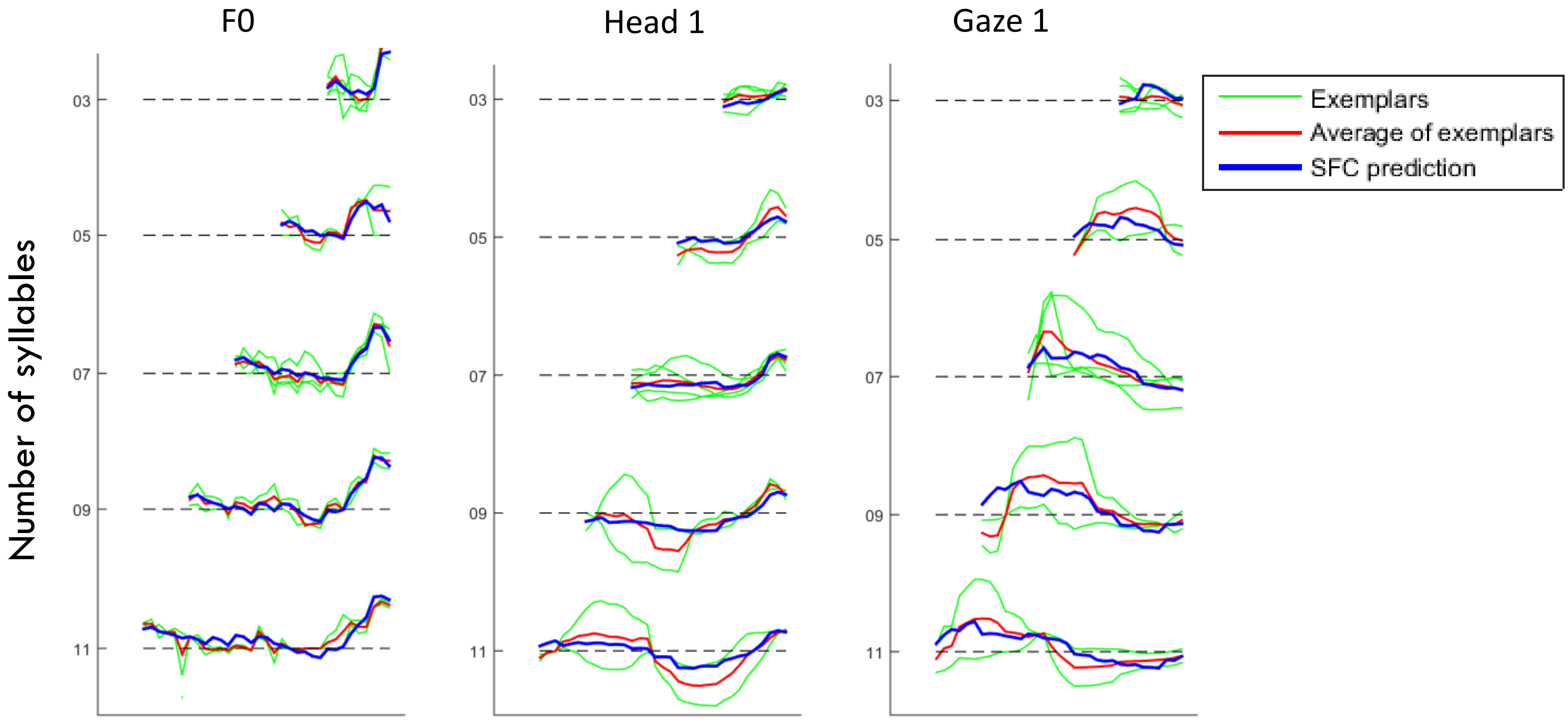
- □ Contour generator
  - □ Neural network
  - □ Input: linear ramps
  - □ Output: stylized prosody



Contour generator (eg. *7 syllables sentence*)

Generation of expressive performances

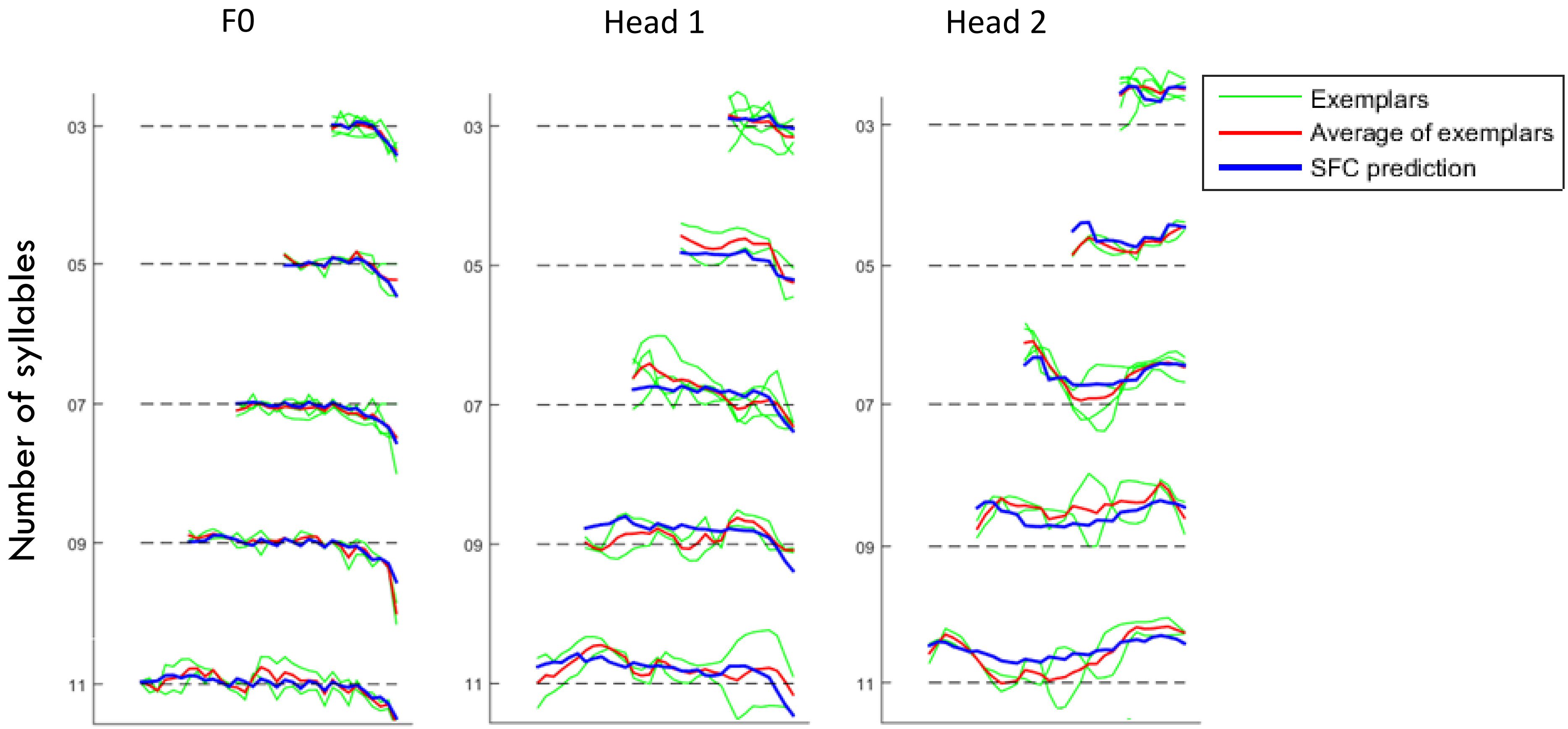
SFC results – Greg - Doubtful



Melody and motion encoded at sentence level: 3, 5, 7, 9 and 11 syllables

Generation of expressive performances

SFC results – Lucie - Comforting



Melody and motion encoded at sentence level: 3, 5, 7, 9 and 11 syllables

Generation of expressive performances

Model-based results

Ce n'est pas possible (It is not possible).

Thinking

Expressive target

Model-based

✓

Doubtful

Expressive target

Model-based

✗

- □ Head motion

Generation of expressive performances

Exemplar-based

- Separation segmental (GMMs) / prosodic features
- Impose prosody from random exemplar
- Sentence-level unit

Generation of expressive performances

Exemplar-based results

Vous savez (You know) -> Mon cher comte (My dear count): 3 syllables

Doubtful

Expressive source

Expressive target

Exemplar-based

✓

Seductive

Expressive source

Expressive target

Exemplar-based

✗

□ Rhythm

Generation of expressive performances

Objective evaluation

- RMSE: (1) frame-based (2) model-based (3) exemplar-based
- Reflects spatial similarity

	Head rotation (deg)			Brow-area expressions (cm)		
	(1)	(2)	(3)	(1)	(2)	(3)
Interrogative	0.87	1.08	**0.74**	0.50	0.67	**0.37**
Comforting	1.27	**0.98**	1.21	0.36	0.41	**0.16**
Seductive	**1.46**	1.77	1.77	0.50	0.26	**0.21**
Thinking	0.65	0.78	**0.57**	0.45	0.49	**0.21**
Doubtful	0.69	1.73	**0.57**	**0.12**	0.63	0.22
Ironic	1.14	**0.86**	1.12	0.40	0.44	**0.35**
Embarrassed	2.65	**1.58**	2.25	0.33	0.39	**0.16**

Generation of expressive performances

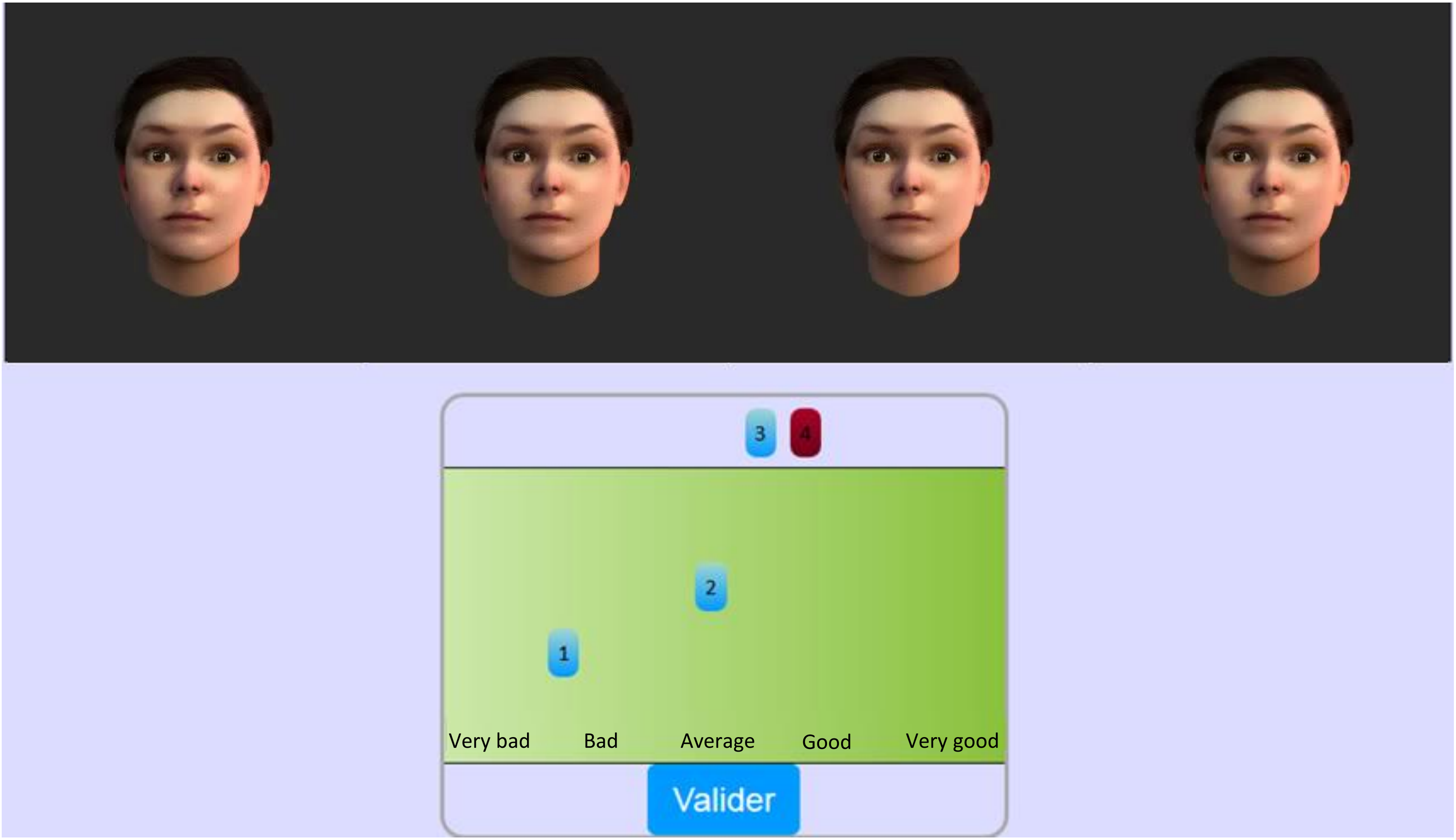
Ranking test

- □ Crowd-sourced ranking test with animations
- □ Baseline methods: frame-based, original reconstructed
- □ Methods to evaluate: exemplar-based, model-based
- □ 7 test sentences, 7 attitudes
- □ 41 native French participants

Generation of expressive performances

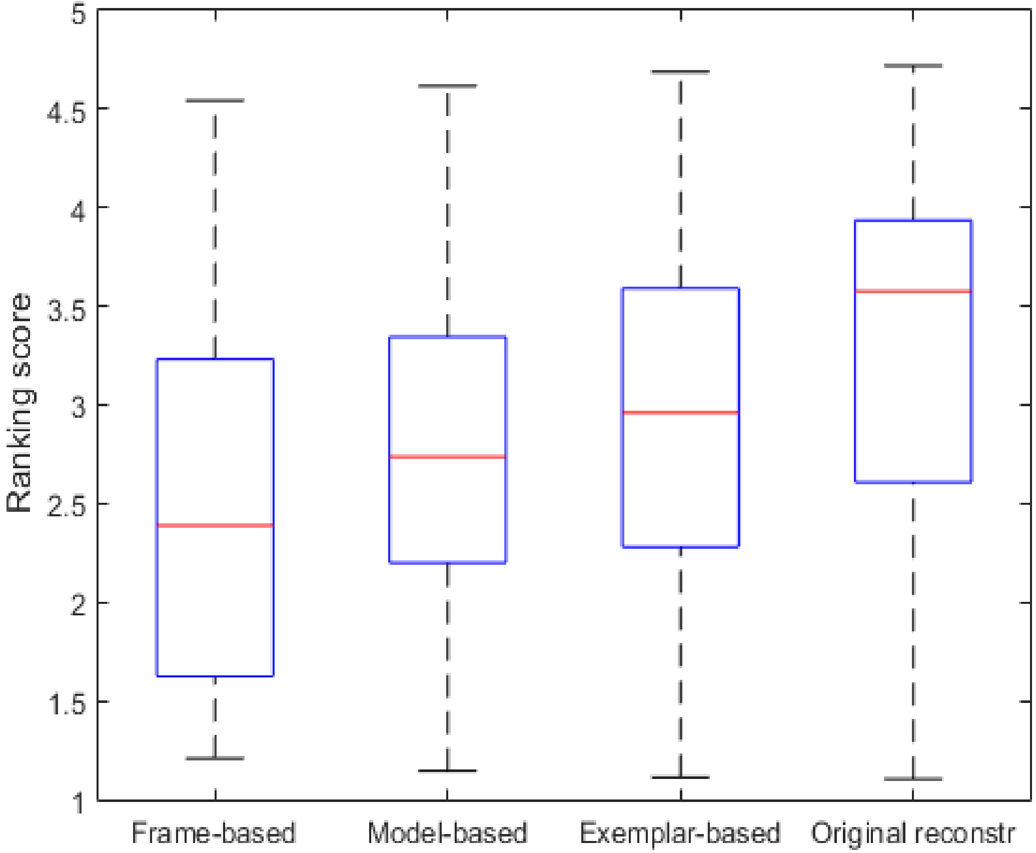
Ranking test

Interrogation



Generation of expressive performances

Ranking test results



	(1)	(2)	(3)	(4)
Interrogative	1.87	3.30	3.11	**3.59**
Comforting	3.02	2.62	2.45	**3.04**
Seductive	2.84	2.59	2.54	**3.34**
Thinking	2.28	2.96	**3.11**	**3.11**
Doubtful	2.69	2.67	2.87	**3.29**
Ironic	2.35	2.46	3.04	**3.58**
Embarrassed	2.45	2.74	2.91	**3.51**

Generation of expressive performances

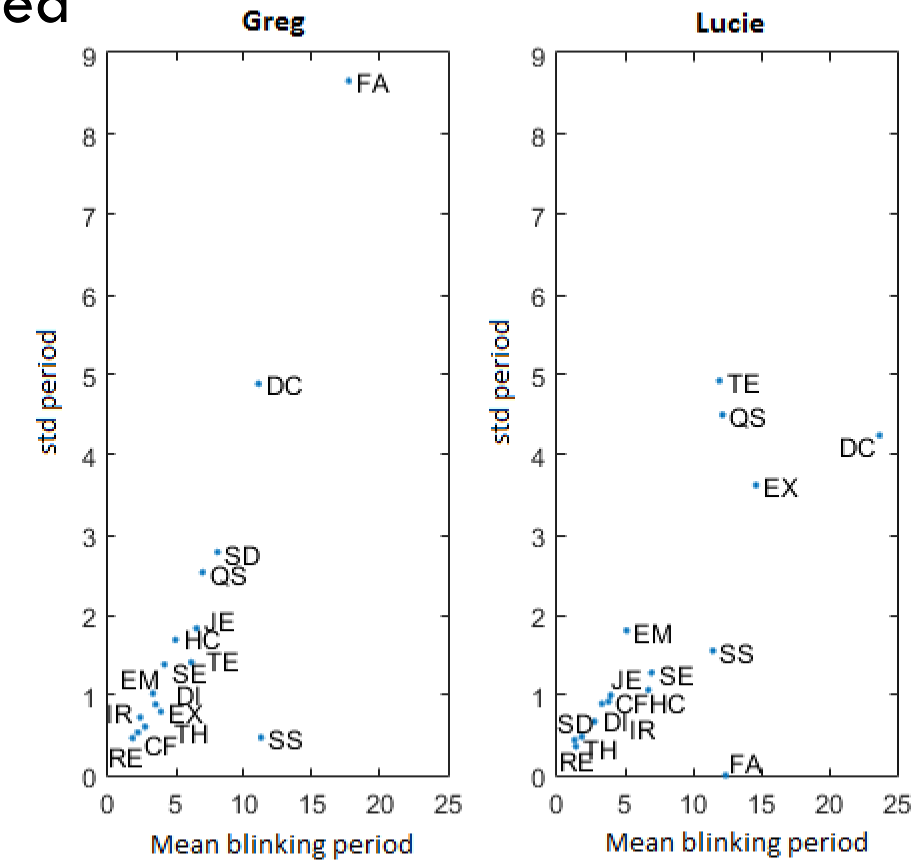
Limitations

- Size and structure of database: few exemplars
- Ranking test is done on few attitudes, only on Lucie, only realistic, A and V not separated
- Missing features
  - Fixations not contingent
  - Virtual syllables
  - Blinking

Generation of expressive performances

Limitations

- □ Size and structure of database: few exemplars
- □ Ranking test is done on few attitudes, only on Lucie, only realistic, A and V not separated
- □ Missing features
  - □ Fixations not contingent
  - □ Virtual syllables
  - □ Blinking



Conclusions

- Framework for the generation of audiovisual expressive performances from didascalia
- Database of 16 interactive « dramatic » attitudes
- There are attitude-specific signatures in visual prosody
  - Extended SFC
- But:
  - Recording « Exercices in style » is difficult even for semiprofessional actors
  - Expressiveness evaluation is difficult
  - Naturalness is harder to achieve in voice synthesis than in facial animation

Perspectives

- Discriminative training
- Intra-sentence structure
  - Model-based: modulation
  - Exemplar-based: phonological matching
- Dialog modeling
  - Gaze
  - Backchannels (Listening to attitudes)
- Extended vocabulary
  - Bigger number of exemplars
  - Choice of didascalia, drama

Dialog generation

Dramaturgic text

SHE: (*Interrogative*). C'est vous, comte?

HE: (*Embarrassed*) Madame votre mère m'a autorisé... autrement je ne me serais pas...

SHE: (*Tender*) Asseyez-vous, mon cher comte.
HE: (*Comforting*) Madame votre mère m'a dit que vous étiez souffrante... mais j'espère que ce ne sera rien.

SHE: (*Scandalized*) Rien!
HE: (*Doubtful*) Et hier encore vous avez joué comme un archange.
SHE: (*Fascinated*) Oui, ç'a été un vrai triomphe.
HE: (*Fascinated*) Toute la salle était emballée. Je ne parle pas de moi.
SHE: (*Seductive*) Merci de vos jolies fleurs.

Dialog generation

Dramaturgic text + result

SHE: (*Interrogative*). C'est vous, comte?

HE: (*Embarrassed*) Madame votre mère m'a autorisé... autrement je ne me serais pas...

SHE: (*Tender*) Asseyez-vous, mon cher comte.
HE: (*Comforting*) Madame votre mère m'a dit que vous étiez souffrante... mais j'espère que ce ne sera rien.

SHE: (*Scandalized*) Rien!
HE: (*Doubtful*) Et hier encore vous avez joué comme un archange.
SHE: (*Fascinated*) Oui, ç'a été un vrai triomphe.
HE: (*Fascinated*) Toute la salle était emballée. Je ne parle pas de moi.
SHE: (*Seductive*) Merci de vos jolies fleurs.

Thank you!

Director & actors

- □ Georges Gagneré
- □ Lucie Carta
- □ Grégoire Gouby

3D artists

- □ Laura Paiardini
- □ Estelle Charleroi
- □ Romain Testylier

Publications

- Audio-Visual Speaker Conversion using Prosody Features, *AVSP 2013*
- Reactive Statistical Mapping: Towards the Sketching of Performative Control with Data, *Springer 2013*
- Beyond Basic Emotions: Expressive Virtual Actors with Social Attitudes, *MIG 2014*.
- Directing virtual actors by interaction and mutual imitation, *doctoral symposium IEEE VR 2015*
- Audiovisual Generation of Social Attitudes from Neutral Stimuli, *FAAVSP 2015*