Beyond Basic Emotions: Expressive Virtual Actors with Social Attitudes

Adela Barbulescu[1,2], Rémi Ronfard[1], Gérard Bailly[2], Georges Gagneré[3], and Hüseyin Cakmak[4]

[1]INRIA Grenoble
[2]GIPSA-lab
[3]Université Paris 8
[4]University of Mons

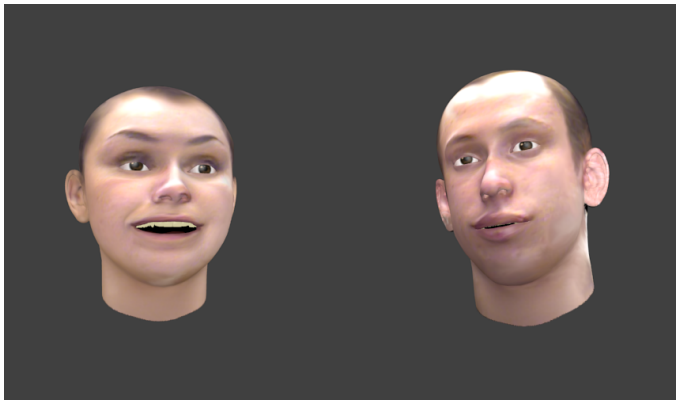

Figure 1: *Virtual actors with facial expressions transferred from theatre actors during a live performance of Schnitzler's Reigen*

Abstract

The purpose of this work is to evaluate the contribution of audio-visual prosody to the perception of complex mental states of virtual actors. We propose that global audio-visual prosodic contours - i.e. melody, rhythm and head movements over the utterance - constitute discriminant features for both the generation and recognition of social attitudes. The hypothesis is tested on an acted corpus of social attitudes in virtual actors and evaluation is done using objective measures and perceptual tests.

Keywords: audio-visual prosody, head motion, social attitudes

1 Introduction

This study represents preliminary work towards expressive audio-visual speaker conversion. Given a context in which speaker A creates a performance in an affective state X, we want to develop approaches to convert it to an expressive performance of speaker B, realised in affective state Y. While in previous work [Barbulescu et al. 2013] we have studied the conversion between speakers A and B, this paper approaches the problem of expressivity conversion and investigates the contribution of different prosodic parameters in expressive generation and recognition.

The term of prosody was initially introduced to denominate a general set of functions that convey the style in which speech is presented, such as intonation, rhythm or stress. As the combination between acoustic and visual cues has been shown to improve the perception of speech [Summerfield 1992], the term was generalized to include visual information. Therefore the concept of audio-visual prosody [Krahmer and Swerts 2009] refers to the use of multimodal cues for signaling and perceiving linguistic, paralinguistic and non linguistic functions in social communication.

The proper generation of expressive audio-visual prosody is an important factor in the perception of expressive speech animations. This topic is situated at the crossroads between computer graphics, vision and speech research areas, and it has received a considerable amount of interest as animated characters are now indispensable components of computer games, movies as well as smart human-computer interfaces.

In a majority of research papers [Zeng et al. 2009], expressivity is defined in terms of limited sets of categorical emotional states or using multi-dimensional emotional models such as the well-known Pleasure-Arousal-Dominance model (PAD). However, an expressive speech is expected to encode complex mental states, which usually go beyond the basic set of emotional states [Ekman 1992]. Scherer proposes that affective expression in speech communications happens either involuntarily (expression of emotion) or voluntarily (expression of attitude)[Scherer and Ellgring 2007]. Similar proposals have been done in the domain of intonation; Bolinger notably states: "Intonation [is] a nonarbitrary, sound-symbolic system with intimate ties to facial expression and bodily gesture, and conveying, underneath it all, emotions and attitudes... [it is] primarily a symptom of how we feel about what we say, or how we feel when we say" [Bolinger 1989]. Social attitudes (example: comforting, doubtful, ironic etc) directly concern carrier speech acts. These attitudes are highly conventionalized - entirely part of the language and the speech communication system - and socio-culturally built. If the production and evaluation of emotional content is highly dynamical, we will also show that social attitudes are also encoded into signals via complex dynamical multimodal patterns.

For this study, we choose a discrete set of social attitudes to high-

light interactive dimensions of face-to-face communication in realistic social contexts. Related work on perception and production of audio-visual prosody [De Moraes et al. 2010][Rilliard et al. 2008] has been carried with the goal of explaining the contribution of each modality for attitude recognition. Besides analyzing the influence of each modality, we study the contribution of specific prosodic parameters by objective measurements and analyze the correlation between these objective metrics and perceptual tests.

Most of the existing approaches to acoustic affect recognition and generation use spectral (MFCC, cepstral features) and prosodic features (related to pitch, energy, speech rate). Among these, pitch and energy are most used in affect recognition systems [Zeng et al. 2009]. Statistical models based on durations and F0 contours are used in neutral to expressive speech conversion [Tao et al. 2006] [Inanoglu and Young 2007]. The global statistical patterns of pitch are studied in [Vroomen et al. 1993] proving that affective states can be expressed accurately by manipulating pitch and duration in a rule-based way. Rule-based generation of animated dialog is also studied in [Cassell et al. 1994]. Recent studies have shown that specific visual cues such as head and eyebrow movements are correlated with pitch contours [Yehia et al. 2000]. This correlation is exploited in [Busso et al. 2007], [Chuang and Bregler 2005] and [Ben Youssef et al. 2013] to synthesize head motion for expressive speech animation.

Our work is preliminary to prosody generation as we first want to quantify the importance of the proposed prosodic parameters: *melody*, *rhythm* and *head movement*. We focus on a theoretical framework which proposes the size of the utterance as a dependent factor.

The paper is structured as follows. Section 2 describes our experimental setup, based on the paradigm of "exercises in style" [Queneau 2013] [Madden 2005]. Section 3 summarizes the methods we used to add expressive contents to neutral audio-visual performances. Subjective evaluations of the performances are presented in Section 4. Section 5 presents experimental results with subjective and objective evaluations of synthetic audio-visual performances obtained with our methods. Section 6 discusses those results and proposes directions for future work.

2 Theoretical framework and experimental data

Our study is based on the theoretical approach described in [Morlec et al. 2001], which proposes that prosodic information is encoded via global multiparametric contours that are organized as prototypical shapes depending on the length of the carrier part of speech (i.e. number of syllables). This model of intonation builds on the seminal work of Fónagy who first put forward the existence of protopypical melodic patterns in French for expressing attitudes, the so-called "melodic clichés" [Fónagy et al. 1983]. Aubergé and Bailly [Aubergé and Bailly 1995] proposed a more general framework that supposes that metalinguistic functions associated with various linguistic units are encoded via elementary global multiparametric contours that are coextensive to these units. The multiparametric prosodic contour of an utterance is then built by superposing and adding these elementary contours by parameter-specific operators. The SFC model [Bailly and Holm 2005] proposes a method for extracting these elementary multiparametric contours from arbitrary utterances given the set of attitudes and their scopes. This model supposes that the set of training utterances randomly samples the possible functions and the positions, lengths and numbers of their overlapping contributions. As we focus on analyzing audio-visual prosody, this theoretical model is extended for the joint modelling of melody, the pattern of syllabic lengthening and head motion trajectories.

As emphasized above, the extraction of prosodic shapes requires sufficient statistical coverage of the metalinguistic functions at varied positions and scope sizes. We have therefore designed and recorded an acted corpus of "pure" social attitudes, i.e. isolated sentences carrying only one attitude over the entire utterance.

2.1 Selected attitudes

A starting point for the possible attitudes considered was the Baron-Cohen's Mind Reading project [Baron-Cohen 2003]. The taxonomy proposed in this work gathers a total of 412 emotions grouped under 24 main categories, each comprizing several layers of subexpressions. Due to the complexity of the taxonomy, we choose a limited number of attitudes which are possibly expressed in one selected act from the play "Reigen" by Arthur Schnitzler [Schnitzler 1993], translated in French by Maurice Rémon et Wilhelm Bauer. This play nicely consists in a series of seduction pitches during dyadic face-to-face conversation. The set of short turns exhibit a large variety of emotions and attitudes, intimately associated with the verbal content.

Table 1 contains the list of attitudes we decided to study, to which the three basic modalities have been added: assertion (DC), exclamation (EX) and full question (QS)(see table 2). The correspondance to the Mind Reading emotional set is indicated and instructions given to actors to perform the attitudes are also included. Video snaphots of chosen social attitudes are illustrated in figure 2.

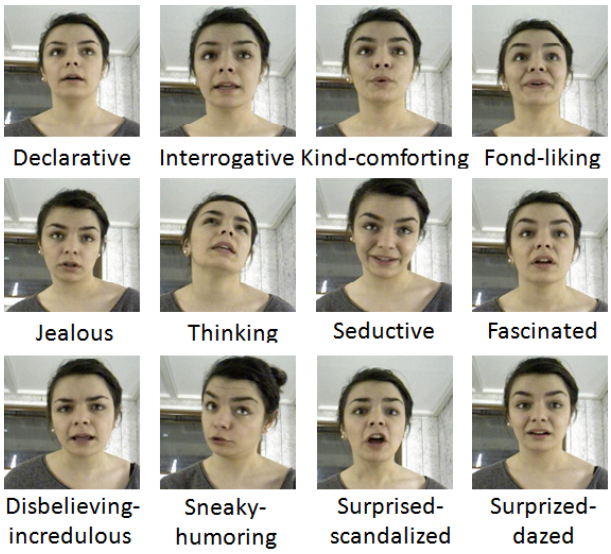

Figure 2: *Examples of attitudes interpreted by Actor 2.*

2.2 Corpus

Two semi-professional actors recorded two corpora under the active supervision of one director:

- dialogs: The two actors played the selected scene in a neutral and an expressive way.
- monologs: The director and each actor recorded 35 sentences uttered with each of the selected 16 attitudes. These sentences were selected from the target scene so that to span the distribution of lengths of sentences in the final text (from one syl-

Table 1: *Description of chosen social attitudes.*

Category	Subgroup	Abbr.	Definition
Kind	Comforting	**CF**	Making people feel less worried, unhappy or insecure
Fond	Liking	**FL**	Finding something or someone appealing and pleasant; being fond of something or someone
Romantic	Seductive	**SE**	Physically attractive
Interested	Fascinated	**FA**	Very curious about and interested in something that you find attractive or impressive
Wanting	Jealous	**JE**	Feeling resentment against someone because of that person's rivalry, success, or advantages
Thinking	Thoughtful	**TH**	Thinking deeply or seriously about something
Disbelieving	Incredulous	**DI**	Unwilling or unable to believe something
Unfriendly	Sarcastic	**US**	Using words to convey a meaning that is the opposite of its literal meaning
Surprised	Scandalized	**SS**	Shocked or offended by someone else's improper behavior
Surprised	Dazed	**SD**	So shocked and stunned that you can't think clearly
Sorry	Responsible	**RE**	Feeling that you are the cause of something that has happened and must therefore take the blame for its affects
Hurt	Confronted	**HC**	Approached in a critical or threatening way
Sorry	Embarrassed	**EM**	Worried about what other people will think of you

Table 2: *Attitudes added.*

Name	Abbr.	Definition
Declarative	**DC**	Making a declaration; neutral
Exclamative	**EX**	Making an exclamation
Interrogative	**QS**	Making an interrogation; question

lable up to 21-syllable sentences) and have a large variability of positions and lengths of constitutive grammar groups.

The recording session began with an intensive training of the actors, who received scenic indications from the theater director. The training consisted in fully understanding the interpreted attitudes and developing the ability to dissociate the affective state imposed by each attitude and the meanings of the phrases and to maintain a constant voice modulation, specific for each attitude, throughout uttering the 35 phrases. The actors did not receive any instruction related to head movements. The same set of sentences are then recorded by the director himself.

2.3 Recordings

The synchronized recording of voice signals and motion are performed by the commercial system Faceshift [1] with a short-range Kinect camera and a Lavalier microphone. Faceshift enables the creation of a customized user profile consisting of a 3D face mesh and an expression model characterised by a set of predefined blendshapes that correspond to facial expressions (smile, eye blink, brows up, jaw open etc). The sampling rate for audio is 44.1 kHz. Recordings are done in front of the camera, while seated, without additional markers such that the acting is not constrained. The use of Faceshift requires soft, non-saturated light conditions. Therefore, the recordings are done in a sound-proof, uniformly illuminated studio.

For the monologs, the actors were asked to utter the set of 35 sentences for each attitude in raw. Due to the nature of the desired social attitudes and to the eye gaze tracker, the actors perform as if they are addressing to a person standing in front of them, at the same height. For the dialogs, the actors sat in front of each other across a table, where two kinect cameras were laying.

[1] http://www.faceshift.com/

2.4 Annotation and characterization

All utterances were automatically aligned with their phonetic transcription obtained by an automatic text-to-speech phonetizer [Bailly et al. 1991]. The linguistic analysis (part-of-speech tagging, syllabation), the phonetic annotation and the automatic estimation of melody were further checked and corrected by hand using a speech analysis software [Boersma 2002]. Figure 3 presents a snapshot of Praat and the video frames associated to the annotated utterance.

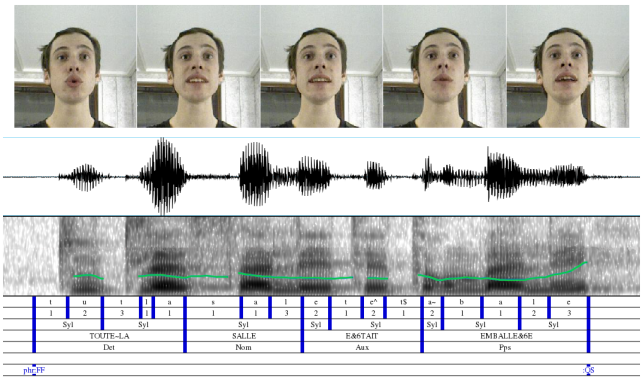

Figure 3: *Illustration of Praat usage for phonetic annotation of a the phrase: "Toute la salle était emballée." uttered by Actor 1 with the attitude Question. The contour highlighted over the spectrogram represents the F0 trajectory. The rising of the contour towards the end of phrase is characteristic for interogative phrases. Every eight video frames are selected from the performance and they illustrate the production of sounds: /t/, /a/, /e/, /a∼/ and /e/ respectively.*

The subjects' performances are then caracterized by the following parameters associated with each syllable:

- Melody: When the vocalic nucleus of the syllable is voiced, we sample the F0 contour of this vowel at three timestamps: 20%, 50% and 80% of its duration. The three values are left unspecified otherwise.
- Rhythm: A lengthening/shortening factor is computed by considering an elastic model of the syllable. This model that the syllable compress/expand according to the elasticity of its segmental constituents. Contingent pauses are included in the model by saturating the elastic lengthening [Barbosa and Bailly 1994].
- Head movements: We sample the head movements of the vocalic nucleus of the syllable at three timestamps: 20%, 50% and 80% of its duration. Principal component analysis is applied to rotation and translation and only the first three composants are kept for further analysis (explaining up to 80% of the information contained). Note that we considered one silent syllable (250 ms) before each utterance to gather preparatory head movements of prior to speech production.

2.5 Animation

As explained in the previous section, performances are recorded using a microphone and a Kinect camera. We synchronously collected audio, 2D (RGB images) and 3D data. Along with information regarding blendshapes, eye gaze and head movements, Faceshift can also generate RGB textures for each expression scanned during the training phrase.

We rendered the 3D performances of the actors using Blender [2]. This rendering consisted in morphing static face textures on the facial mesh animated by the blendshapes estimated by Faceshift. A simple mode of the neck deformation with fixed torso is applied (video sample of an expressive animated dialog between the two actors [3]). Figure 4 presents examples of rendered frames for different attitudes.

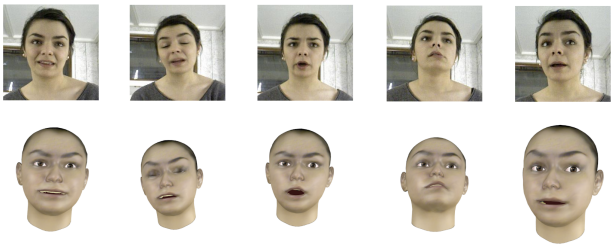

Figure 4: *Top: video images, bottom: rendered frames for animated performances for the respective attitudes: Seductive, Jealous, Scandalized, Thinking and Question.*

3 Expressive Synthesis

One approach to evaluate the proposed prosodic features is by copying the time structure, pitch contours and head movements from the emotional performance samples onto declarative performances. We synthesize expressive versions of performances by simply substituting prosodic parameters of the DC version with prosodic parameters of the target expressive version. Blendshapes trajectories - notably facial displays such as eyebrows, eye and lip movements - are kept from the DC version and only time-warped to comply with the target rhythm and syllabic durations.

Acoustic synthesis

Prosodic manipulation of the DC stimuli is performed by Time-Domain Pitch-Synchronous Overlap-and-Add (TD-PSOLA) technique that consists in controlling pitch and duration of sounds by moving, deleting or duplicating short-time signals [Moulines and Charpentier 1990].

Visual synthesis

Synthesized blendshapes and eye gaze are obtained by computing a Dynamic Time Warping path [Berndt and Clifford 1994] as we know the phoneme durations of DC and the analyzed attitude, and applying it to the movements of DC. The head movements are generated by interpolating between the movements at syllable landmarks of the analyzed attitude. In the interpolation step, movements are computed for 30 points per second to match the visual framerate. Rotations are used in quaternion representation so we use cubic interpolation (SQUAD) [Eberly 2001] to smoothly interpolate over a path of quaternions. This method is build upon the spherical linear interpolation (slerp) [Shoemake 1985]:

$$slerp(q_1, q_2, t) = \frac{\sin(1-t)\theta}{\sin\theta} q_1 + \frac{\sin t\theta}{\sin t} q_2 \quad (1)$$

$$SQUAD(q_1, q_2, q_3, q_4, t) = \\ = slerp(slerp(q_1, q_4, t), slerp(q_2, q_3, t), 2t(1-t)) \quad (2)$$

where q_i represent quaternions, θ is obtained by computing the dot product between quaternions and t is the desired interpolation position.

Translations are obtained using cubic spline interpolation. Examples of synthesized trajectories are given in Figure 5 (video sample of animations obtained from original and synthesized from neutral performances [4]).

4 Subjective evaluation

For evaluation purposes, a series of subjective and objective tests are conducted on for three separate modalities: audio, visual and audio-visual. All subjective evaluations consist in forced-choice identification tests. Subjects are asked to label heard and/or viewed performances with one of the 16 attitudes.

4.1 Auto-evaluation

The persons who participated in the recording of our corpus (the two actors and director) are asked to perform an auto-evaluation test in which they were asked to label samples from their own performances, a few days after having been recorded. They are presented with random sets of 32 performances for each modality, which they can play several times. No explanation is given for the labels. Table 3 shows the accuracy rates for all performers on all modalities used:

The audio channel seems to encode slightly less discriminant information than the video channel. Note however that the audio/video hierarchy is strongly attitude-dependent. Overall, the recognition rate is high above the chance level (6.25%) and in the case of the

[2] http://www.blender.org/

[3] http://youtu.be/Se_P26Tg6-s/

[4] http://youtu.be/B_BjeBnaC7I/

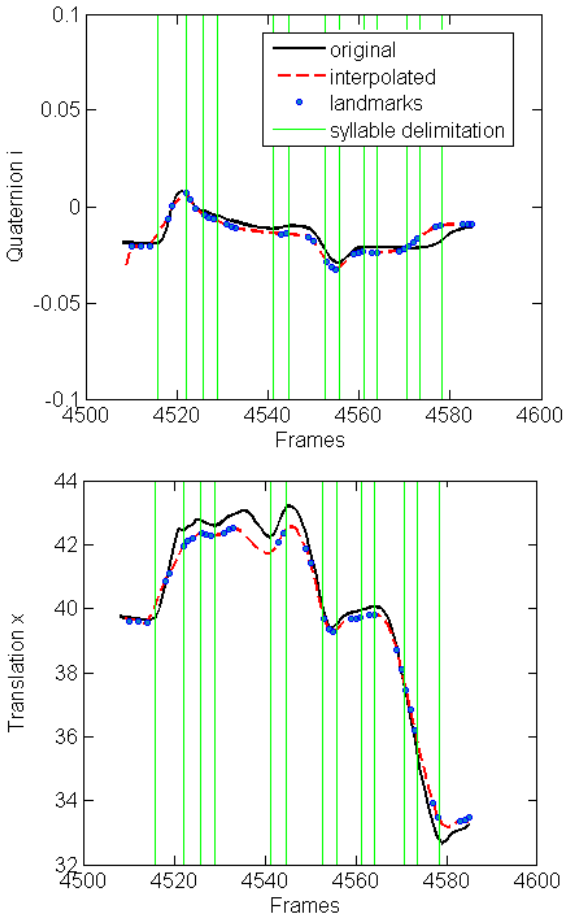

Figure 5: *Interpolated trajectories of prosodic parameters for phrase 30, attitude Question. Top: Interpolated and original trajectories of first dimension of the rotation quaternion, bottom: interpolated and original trajectories of translation along axis X of the coordinate system used. Prosodic landmarks are marked as circles and syllables are delimited by green lines. Note that for some syllables, only two landmarks appear. This happens when the vocalic duration is very short and at the conversion to the visual framerate, some landmarks are attributed to the same frame.*

actors, is generally higher than 60%. Actor 2 has the highest consistency. We thus choose performances made by this female speaker for a crowdsourcing subjective test.

4.2 Evaluation of original material

Eighty anonymous participants accessed the evaluation test which was carried using a normal web browser. Before starting the test, user information is collected, of which the most important is whether french is the mother tongue. Participants were given the definitions of all the attitudes and asked to identify the attitudes performed by the chosen actor (Actor 2) using audio files, video-only files and audio-video files. For each modality, the participants have to label 32 random performances. Each set is obtained by a random function which picks 2 performances from each attitude with the condition of not retrieving identical adjacent phrases. Performances last up to 5 seconds and can be played several times until they are labeled. Only the answers provided by native speakers are taken into account. After test completion, users have the option of leaving comments. Evaluation tests took on average 20 minutes to be completed.

Table 3: *Accuracy rates obtained for all modalities and performers for the auto-evaluation test.*

Performer	Gender	**Audio**	**Video**	**Audio-Video**
Director	male	43.75%	68.75%	68.75%
Actor 1	male	65.62%	71.87%	75.00%
Actor 2	female	**78.12%**	**81.25%**	**78.12%**

The test containing original material consists in original audio files and video files obtained from sequences of 2D images which are synchronized with sound. Confusion matrices are built based on the choices made in the subjective tests for each modality. For representation of attitudes, we use the abbreviations indicated in Tables 1 and 2. The recognition rates for all modalities are presented in Table 4.

Table 4: *Recognition rates for the online test containing original audio and video performances.*

	Audio	**Video**	**Audio-Video**
DC	0.148	0.251	0.170
EX	0.048	0.014	0.066
QS	**0.725**	0.412	**0.788**
CF	0.130	**0.563**	0.290
FL	0.481	0.340	0.368
SE	**0.506**	**0.731**	**0.761**
FA	0.192	0.289	0.391
JE	0.000	0.250	0.167
TH	**0.597**	**0.722**	**0.814**
DI	0.180	0.083	0.214
US	**0.672**	0.383	**0.640**
SS	0.435	0.250	0.264
SD	0.325	0.294	0.357
RE	0.314	0.462	0.273
HC	0.216	0.361	0.150
EM	0.263	0.464	**0.617**

The best recognized attitudes are: Interrogative(QS), Seductive(SE), Thinking(TH) and Unfriendly-Sarcastic(US), while the lowest result values are obtained for: Exclamative(EX), Jealous(JE), Doubt-Incredulity(DI) and Hurt-Confrunted(HC). Except for EX, most attitudes have a recognition score high above chance level (0.0625%). Declarative (DC) also presents low scores as it was confused with EX and RE in all modalities. According to the comments left by users, the meaning of the HC attitude was not well understood, thus explaining the low score obtained. For this reason and for the fact that many users considered the test was too long, HC is removed from the set of attitudes in the following tests. A factor which may explain why for some attitudes (DC, JE, RE, HC) the scores obtained for video-only were higher than the ones including audio cues is the difficulty of dissociating the attitude from the semantic content of the carrier sentence.

4.3 Evaluation of animation

As the main focus of our work involves animated performances, a necessary step is the evaluation of animations obtained using data from original performances. Our expectation is that the level of recognition will be lower than in the case of the original video sources. This loss of expressivity has been studied in works as [Afzal et al. 2009] and in our case may be due to several causes:

systematical errors of the data capture system, inability of capturing very fine facial expressions, audio-visual data synchronisation, unrealistic effects of the face texture, complex nature of the attitudes studied etc. The audio data is original as in the previous section.

Forty two anonymous participants were recruited to label performances via an online test. The performances are obtained using the original audio files and animations obtained from the original blendshape, eye gaze and head movement values. The total number of attitudes is 15 and each participant has to label 30 performances per modality. Tables 5, 6 and 7 show the confusion matrices obtained.

Table 5: *Confusion matrix obtained for audio-only.*

	DC	EX	QS	CF	FL	SE	FA	JE	TH	DI	SH	SS	SD	RE	EM
DC	**42**	13	1	3	1	0	1	4	7	0	0	1	3	5	3
EX	43	**11**	0	8	0	0	3	3	2	1	1	6	1	4	1
QS	5	12	**51**	1	0	0	0	0	0	5	1	2	1	1	4
CF	22	5	0	**16**	9	5	7	1	5	1	2	0	2	4	4
FL	6	21	0	8	**17**	3	15	0	1	2	5	0	4	1	1
SE	5	0	3	2	1	**35**	6	1	9	4	10	2	1	3	2
FA	7	3	0	4	1	14	**20**	1	5	5	3	2	6	7	6
JE	34	4	3	5	3	1	0	**2**	2	8	7	2	1	7	5
TH	14	5	3	3	2	0	5	1	**22**	12	2	0	2	3	9
DI	7	18	4	12	3	2	5	0	3	**10**	1	3	6	5	5
SH	12	8	4	5	2	0	4	2	3	6	**15**	2	3	8	9
SS	1	23	5	0	0	0	0	1	0	2	1	**34**	16	1	0
SD	7	9	20	0	1	2	4	2	2	11	2	1	**13**	4	6
RE	35	17	0	8	0	0	2	3	5	1	3	5	1	**3**	1
EM	7	1	1	2	3	9	6	0	1	6	3	0	0	8	**37**

Table 6: *Confusion matrix obtained for video-only.*

	DC	EX	QS	CF	FL	SE	FA	JE	TH	DI	SH	SS	SD	RE	EM
DC	**29**	1	3	1	0	0	6	1	7	8	0	0	6	5	4
EX	40	**8**	3	1	2	2	1	2	3	4	2	1	0	2	0
QS	16	3	**19**	1	1	1	1	0	4	7	0	3	8	2	6
CF	8	16	3	**12**	8	3	6	0	4	1	5	1	3	1	0
FL	5	19	0	6	**24**	0	10	2	0	0	4	0	1	0	1
SE	3	12	2	7	11	**6**	10	1	2	1	10	2	2	1	2
FA	14	1	2	0	0	1	**13**	1	2	12	3	1	13	5	4
JE	11	7	9	2	1	2	0	**2**	11	4	11	3	3	4	2
TH	4	8	7	3	1	0	0	0	**29**	9	7	1	1	0	2
DI	5	15	3	0	0	0	1	6	0	**3**	5	27	6	0	0
SH	10	10	6	5	1	3	1	0	3	4	**5**	10	7	2	5
SS	3	12	4	2	0	0	1	8	1	5	3	**17**	13	3	0
SD	5	8	14	0	2	1	1	1	4	15	2	9	**7**	2	1
RE	4	1	13	0	0	0	3	2	9	16	3	2	12	**4**	2
EM	7	1	3	7	3	11	2	3	5	3	7	2	0	5	**13**

Similarly to the previous test, DC is confused with EX and RE in all modalities, leading to small recognition rate for DC. The best recognition scores are obtained by QS, SE, SS and EM, especially for the modalities which include audio cues. As expected, the biggest difference between the scores obtained in this test and the ones from the previous test are caused by a general lower recognition rate for the video-only modality.

4.4 Evaluation of expressive synthesis

The last subjective experiment evaluates the relevance of the characteristics of the audio-visual prosody we chose to extract and control. We used the approaches presented in Section 3 to synthesize expressive performances for all modalities. Two modal attitudes are removed in order to shorten the test duration: EX and QS. Up to the time of this submission, thirteen native speakers participated in the online evaluation test and were presented a total of 26 performances per modality. Table 8 shows the recognition scores obtained for all attitudes:

Table 7: *Confusion matrix obtained for audio-video.*

	DC	EX	QS	CF	FL	SE	FA	JE	TH	DI	SH	SS	SD	RE	EM
DC	**43**	4	0	0	1	0	7	1	5	3	1	1	4	2	0
EX	41	**14**	0	5	0	0	2	3	2	0	1	1	1	1	1
QS	2	5	**48**	2	0	0	0	1	0	7	0	2	3	1	1
CF	12	6	0	**21**	8	6	5	0	3	2	1	1	2	2	3
FL	3	20	0	13	**16**	2	12	1	1	0	1	0	3	0	0
SE	6	3	2	4	4	**27**	14	0	3	0	3	0	1	4	1
FA	4	3	1	5	3	2	**24**	0	5	2	2	0	13	4	4
JE	27	0	1	1	0	0	1	**6**	5	7	13	1	0	6	4
TH	11	4	4	2	0	2	0	1	**35**	11	0	0	0	1	1
DI	3	21	3	4	1	0	3	2	2	**12**	2	7	5	4	3
SH	10	5	1	4	2	1	5	0	4	7	**16**	0	4	9	4
SS	0	20	2	1	0	0	0	0	0	1	0	**32**	15	1	0
SD	2	7	12	1	1	1	3	1	2	10	9	4	**16**	2	1
RE	27	8	0	3	0	0	0	3	9	8	1	2	4	**2**	3
EM	3	0	1	0	3	10	6	0	4	3	2	0	1	8	**31**

Table 8: *Recognition rates obtained for the online test on synthesized performances.*

	Audio	Video	Audio-Video
DC	0.120	0.101	0.124
CF	0.140	0.162	0.105
FL	0.130	0.125	0.143
SE	0.000	0.000	0.091
FA	0.133	0.158	**0.231**
JE	0.000	0.136	0.067
TH	0.194	**0.257**	**0.379**
DI	0.036	0.149	0.135
US	0.120	0.045	0.222
SS	**0.350**	0.033	**0.333**
SD	0.130	0.071	0.200
RE	0.059	0.000	0.000
EM	**0.313**	0.182	0.192

The best recognition rates are obtained by FA, TH, SS and EM, showing results higher than chance level (7.69%). As expected, DC has a low recognition rate as it was confused with all other attitudes for all modalities. Attitudes such as SE and RE were not recognized at all for the video-only modality. This proves that the chosen prosodic parameters are more discriminant for certain attitudes; for example, there are distinguishable head movement patterns (for FA and TH, there is a tendency in rising the head) or speech patterns (for SS the rythm is much faster than in EM, and the pitch is higher). However, for more subtle attitudes such as SE, JE and RE, the visual cues rely more on eye gaze and eyebrow movements. As commented by a few participants in the study, confusion is induced by the neutral appearance of the eye gaze and expressions generated in the upper part of the face.

5 Objective evaluation

The objective evaluation refers to inter-class distances between all attitudes. These are computed for all modalities considering as metric the euclidian distance between the prosodic parameters for

equal-sized utterances:

$$d(att_i, att_j) = \sum_{k=1}^{35} \sqrt{\sum_{l=1}^{n} \sum_{m=1}^{3} (P_{iklm} - P_{jklm})^2} \qquad (3)$$

where n is the number of syllables of the phrase k and P represents the value of the prosodic element.

Depending on the modality chosen, the distance is computed at syllable level between the respective three values of prosodic parameters: F0 for audio-only, the first 3 PCA components over rotation and translation for video-only or all of them for audio-video. A confusion matrix is obtained for each modality using a K-nearest framework which enables the formation of attitudinal clusters. The accuracy rates obtained from the confusion matrix for all performers and modalities are presented in Table 9.

Table 9: *Accuracy rates obtained for all modalities and actors for the objective test.*

Performer	Audio	Video	Audio-Video
Director	13.02%	26.35%	33.48%
Actor 1	**32.53%**	**31.47%**	**46.98%**
Actor 2	30.60%	24.14%	40.95%

The results show that Actor 1 presents overall more discriminant prosodic features for each modality. Next we will focus on the results obtained by Actor 2 in order to study the correlation between the objective measures and perceptual tests. Table 10 presents the recognition rates for each attitude:

Table 10: *Recognition rate for the objective measure.*

	Audio	Video	Audio-Video
DC	**0.514**	0.294	0.442
EX	0.125	0.471	**0.625**
QS	**0.714**	0.250	0.455
CF	0.100	0.143	0.458
FL	**0.773**	0.182	**0.783**
SE	0.290	0.200	0.244
FA	0.154	0.241	0.304
JE	0.154	0.438	**0.714**
TH	0.250	0.120	0.292
DI	0.440	0.250	**0.556**
US	0.235	0.310	0.476
SS	0.333	0.182	0.238
SD	0.333	0.073	0.344
RE	0.184	0.321	0.304
HC	0.385	0.333	0.389
EM	0.150	0.158	0.167

For the audio-only modality the most distinguishable attitudes are DC, QS, FL and DI. For video-only these are DC, EX, JE, US, RE and HC. Audio-video modality generally receives the highest scores and especially for EX, FL, JE and DI. The video-only modality receives the lowest scores among modalities reinforcing the results obtained from the perceptive tests. Overall, the recognition rates are higher than the chance level (6.25%) proving that there are distinguishable attitude characteristics in the analyzed prosodic features.

6 Discussion

The accuracy rates for all modalities and tests performed are presented in Table 11. The recognition rates for the three crowdsourced tests for the Audio-Video modality are presented in figure 6. Depending on the material used for each crowdsourced test we will consider the following notations: Material 1 for original video performances (Section 4.2), Material 2 for original animated performances (Section 4.3) and Material 3 for synthesized animated performances (Section 4.4).

Table 11: *Accuracy rates obtained for all modalities and tests for Actor 2.*

Test type	Audio	Video	Audio-Video
Auto-evaluation	78.12%	81.25%	78.12%
Material 1	30.98%	35.47%	36.90%
Material 2	26.00%	17.73%	31.72%
Material 3	15.58%	11.65%	16.96%
Objective	30.60%	24.14%	40.95%

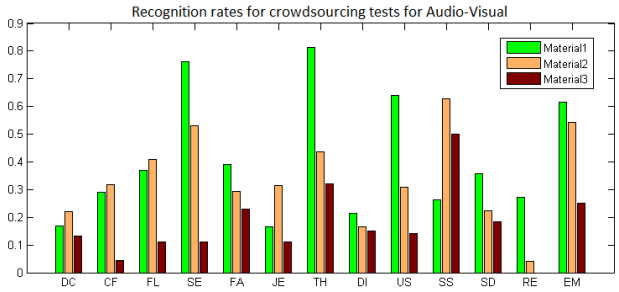

Figure 6: *For each attitude, the bars represent recognition rates for the audio-video modality.*

We notice that the accuracy rates for objective test and the subjective where the original audio and video were used are comparable only for the audio channel. The subjective tests reveal general higher recognition rates for original video and lower rates for synthesized from neutral performances. A higher recognition in the test using original video means that we need to include other visual parameters. Considering only head movements is not sufficient to obtain an objective signature of our set of social attitudes. F0 and rhythm are sufficient for obtaining discriminating audio contours.

The correlation between the objective and perceptual confusions on different modalities may help determining if the chosen audio-visual characteristics can explain the perceptual judgements and how much of the variance of the perceptual results they capture. The correlations between the objective and subjective confusion matrices are computed using the Pearson coefficients:

$$R_{i,j} = \frac{C_{i,j}}{\sqrt{C_{i,i} C_{j,j}}} \qquad (4)$$

where C is the covariance matrix of $[CM_{online}\ CM_{objective}]$ and CM are the two confusion matrices.

Figure 7 presents the coefficient values obtained between each of the online perceptual tests and the objective test for all attitudes:

The first caption suggests correlations for all modalities between the results of the objective test and the recognition scores given by the first online test. Attitudes that are well correlated are: DC, QS,

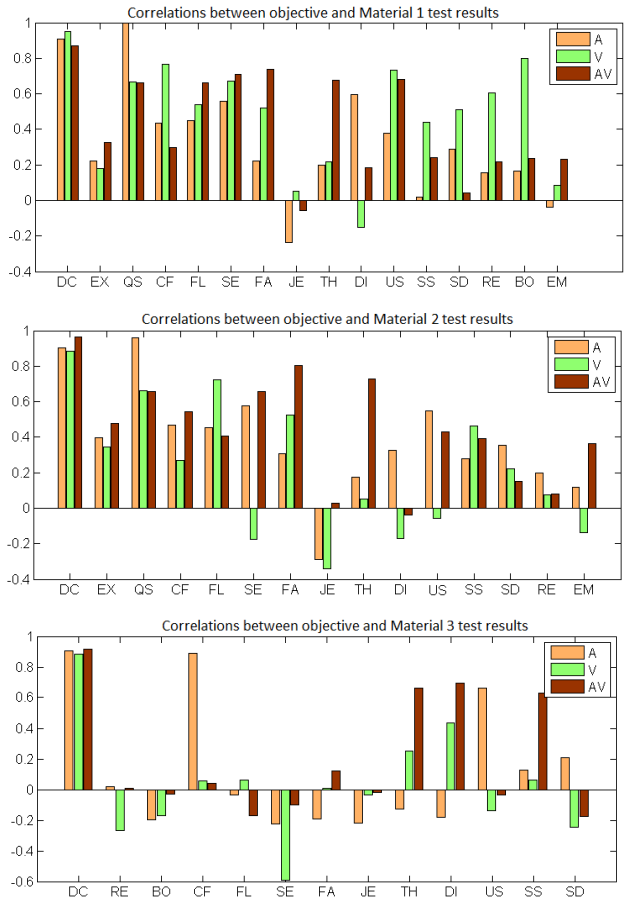

Figure 7: *For each attitude, the bars represent the correlation coefficients obtained for audio, video and audio-video.*

FL, SE, US. Video-only presents a general correlation higher than 0.5, except for EX, JE, DI, EM which have low correlations for all modalities. Audio-only has very strong correlations with DC and QS (higher than 0.9), proving that the F0 shapes have a high influence on the auditive perception of attitude.

The second caption presents correlations between the objective test and the second online test. The correlations obtained in the audio-only modality are very similar to the ones obtained in the previous test, which is expected, as both experiments use random samples of original audio-data. The video-only correlations are generally lower for this test, showing that perceptual recognition scores are degraded overall. Audio-visual correlations are also similar to the ones obtained in the previous test.

The third online test presents much lower correlations with the objective test; strong correlations exist only for DC for all modalities, then for CF and US for audio-only and TH, DI and US for audio-visual. The low scores obtained on the third experiment (see Table 11) show that modifying the proposed parameters is not sufficient for obtaining an acceptable perceptive score for all attitudes. Therefore it is necessary to include more prosodic features - more subtle and surely more segment-dependent - in our analysis.

7 Conclusion

We have examined the contribution of audio-visual prosodic parameters to the evaluation of expressive performances. The prosodic features studied are voice pitch, audio-visual rhythm and head motion. Our dataset of social attitudes is validated through auto-evaluation and crowdsourcing evaluation tests. Using objective measures and perceptual tests, we have proved that the parameters reveal distinguishable attitude characteristics. Comparable recognition scores for perceptive tests and objective measures show that F0 and rhythm are sufficient for obtaining attitude signatures, especially for Question, Fond-Liking, Disbelieving-Incredulous and Surprised-Scandalized. However, considering only head motion as video cues is not sufficient for attitude recognition.

The lower scores obtained when using animations show that they can be improved (texture, eye and lips movements). Another reason for obtaining lower scores in the perceptual tests may be the amount and nature of the attitudes studied, which require a good understanding of their definitions. When evaluating audio data, participants dissociate from the meaning of the sentence with difficulty.

The test using synthesized data shows that there are more parameters needed to better exploit the signatures of the recorded attitudes. At the acoustic level such parameters are intensity and timbre. For visual cues the first parameters we will look into eye gaze and eyebrow movements.

Future work will include further analysis of the new prosodic features, improvement of the animation platform and the development of the prosodic model sustained by the theoretical framework described in Section 2. The visual prosody model will be extended to other parameters including eye movements, eye blinks, eyebrow movements and related FACS parameters. Idiosincracies will also be studied with the goal of building person-specific models of prosodic contours to enable the expression transfer across speakers and utterances. This study represents an initial step towards audio-video expressive speaker conversion, ultimately targeting numerous applications within the domains of video games, intelligent user interfaces and expressive speech animations.

Acknowledgements

This work is supported by the LabEx PERSYVAL-Lab (ANR–11-LABX-0025). H. Cakmak receives a PhD grant from the Fonds de la Recherche pour l'Industrie et l'Agriculture (FRIA), Belgium. We strongly thank Georges Gagneré, Lucie Carta et Grégoire Gouby for providing us with these invaluable experimental resources. This research will not be possible without the numerous anonymous contributions of crowdsourced participants.

References

AFZAL, S., SEZGIN, T. M., GAO, Y., AND ROBINSON, P. 2009. Perception of emotional expressions in different representations using facial feature points. In *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on*, IEEE, 1–6.

AUBERGÉ, V., AND BAILLY, G. 1995. Generation of intonation: a global approach. In *EUROSPEECH*.

BAILLY, G., AND HOLM, B. 2005. Sfc: a trainable prosodic model. *Speech Communication 46*, 3, 348–364.

BAILLY, G., BARBE, T., AND WANG, H.-D. 1991. Automatic labeling of large prosodic databases: Tools, methodology and links with a text-to-speech system. In *The ESCA Workshop on Speech Synthesis*.

BARBOSA, P., AND BAILLY, G. 1994. Characterisation of rhythmic patterns for text-to-speech synthesis. *Speech Communication 15*, 1, 127–137.

BARBULESCU, A., HUEBER, T., BAILLY, G., RONFARD, R., ET AL. 2013. Audio-visual speaker conversion using prosody features. In *International Conference on Auditory-Visual Speech Processing*.

BARON-COHEN, S. 2003. *Mind reading: the interactive guide to emotions*. Jessica Kingsley Publishers.

BEN YOUSSEF, A., SHIMODAIRA, H., AND BRAUDE, D. A. 2013. Articulatory features for speech-driven head motion synthesis. *Proceedings of Interspeech, Lyon, France*.

BERNDT, D. J., AND CLIFFORD, J. 1994. Using dynamic time warping to find patterns in time series. In *KDD workshop*, vol. 10, Seattle, WA, 359–370.

BOERSMA, P. 2002. Praat, a system for doing phonetics by computer. *Glot international 5*, 9/10, 341–345.

BOLINGER, D. 1989. *Intonation and its uses: Melody in grammar and discourse*. Stanford University Press.

BUSSO, C., DENG, Z., GRIMM, M., NEUMANN, U., AND NARAYANAN, S. 2007. Rigid head motion in expressive speech animation: Analysis and synthesis. *Audio, Speech, and Language Processing, IEEE Transactions on 15*, 3, 1075–1086.

CASSELL, J., PELACHAUD, C., BADLER, N., STEEDMAN, M., ACHORN, B., BECKET, T., DOUVILLE, B., PREVOST, S., AND STONE, M. 1994. Animated conversation: Rule-based generation of facial expression, gesture & spoken intonation for multiple conversational agents. In *Proceedings of the 21st Annual Conference on Computer Graphics and Interactive Techniques*, ACM, New York, NY, USA, SIGGRAPH '94, 413–420.

CHUANG, E., AND BREGLER, C. 2005. Mood swings: expressive speech animation. *ACM Transactions on Graphics (TOG) 24*, 2, 331–347.

DE MORAES, J. A., RILLIARD, A., DE OLIVEIRA MOTA, B. A., AND SHOCHI, T. 2010. Multimodal perception and production of attitudinal meaning in brazilian portuguese. *Proc. Speech Prosody, paper 340*.

EBERLY, D. H. 2001. *3D game engine design*. San Francisco: Morgan Kaufmann Publishers, Inc.

EKMAN, P. 1992. An argument for basic emotions. *Cognition & Emotion 6*, 3-4, 169–200.

FÓNAGY, I., BÉRARD, E., AND FÓNAGY, J. 1983. Clichés mélodiques. *Folia linguistica 17*, 1-4, 153–186.

INANOGLU, Z., AND YOUNG, S. 2007. A system for transforming the emotion in speech: combining data-driven conversion techniques for prosody and voice quality. In *INTERSPEECH*, 490–493.

KRAHMER, E., AND SWERTS, M. 2009. Audiovisual prosody-introduction to the special issue. *Language and speech 52*, 2-3, 129–133.

MADDEN, M. 2005. *99 ways to tell a story: exercises in style*. Chamberlain Bros.

MORLEC, Y., BAILLY, G., AND AUBERGÉ, V. 2001. Generating prosodic attitudes in french: data, model and evaluation. *Speech Communication 33*, 4, 357–371.

MOULINES, E., AND CHARPENTIER, F. 1990. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech communication 9*, 5, 453–467.

QUENEAU, R. 2013. *Exercises in style*. New Directions Publishing.

RILLIARD, A., MARTIN, J.-C., AUBERGÉ, V., SHOCHI, T., ET AL. 2008. Perception of french audio-visual prosodic attitudes. *Speech Prosody, Campinas, Brasil*.

SCHERER, K. R., AND ELLGRING, H. 2007. Multimodal expression of emotion: Affect programs or componential appraisal patterns? *Emotion 7*, 1, 158.

SCHNITZLER, A. 1993. *Reigen*. Invito alla lettura. Impresor.

SHOEMAKE, K. 1985. Animating rotation with quaternion curves. In *ACM SIGGRAPH computer graphics*, vol. 19, ACM, 245–254.

SUMMERFIELD, Q. 1992. Lipreading and audio-visual speech perception. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences 335*, 1273, 71–78.

TAO, J., KANG, Y., AND LI, A. 2006. Prosody conversion from neutral speech to emotional speech. *Audio, Speech, and Language Processing, IEEE Transactions on 14*, 4, 1145–1154.

VROOMEN, J., COLLIER, R., AND MOZZICONACCI, S. J. 1993. Duration and intonation in emotional speech. In *Eurospeech*.

YEHIA, H., KURATATE, T., AND VATIKIOTIS-BATESON, E. 2000. Facial animation and head motion driven by speech acoustics. In *5th Seminar on Speech Production: Models and Data*, Kloster Seeon, Germany, 265–268.

ZENG, Z., PANTIC, M., ROISMAN, G. I., AND HUANG, T. S. 2009. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *Pattern Analysis and Machine Intelligence, IEEE Transactions on 31*, 1, 39–58.