Audiovisual Generation of Social Attitudes from Neutral Stimuli

Adela Barbulescu, Gérard Bailly, Rémi Ronfard, Maël Pouget

Audiovisual Generation of Social Attitudes from Neutral Stimuli

Adela Barbulescu[1,2]*, Gérard Bailly*[1]*, Rémi Ronfard*[2]*, Maël Pouget*[1]

[1]GIPSA-lab
[2]INRIA Grenoble

Abstract

The focus of this study is the generation of expressive audiovisual speech from neutral utterances for 3D virtual actors. Taking into account the segmental and suprasegmental aspects of audiovisual speech, we propose and compare several computational frameworks for the generation of expressive speech and face animation. We notably evaluate a standard frame-based conversion approach with two other methods that postulate the existence of global prosodic audiovisual patterns that are characteristic of social attitudes. The proposed approaches are tested on a database of "Exercises in Style" [1] performed by two semi-professional actors and results are evaluated using crowdsourced perceptual tests. The first test performs a qualitative validation of the animation platform while the second is a comparative study between several expressive speech generation methods. We evaluate how the expressiveness of our audiovisual performances is perceived in comparison to resynthesized original utterances and the outputs of a purely frame-based conversion system.

Index Terms: virtual actors, expressive speech animation, audiovisual prosody, GMM, superposition of functional contours

1. Introduction

With the goal of generating realistic expressive animation, we address the problem of converting neutral performances of a given speaker. While in previous work [2] we have studied the problem of audiovisual speaker conversion, this paper approaches the problem of expressivity conversion.

With respect to expressivity, we distinguish between the "push" and "pull" effects as noted by Scherer [3]: the push effects represent are triggered by underlying psychobiological mechanisms, and the pull effects, triggered by conventions, norms, cultural behaviours. Thus, affective expression in speech communications happens either involuntarily (expression of emotion) or voluntarily (expression of attitude) as Bolinger notably states: "..how we feel about what we say, or how we feel when we say" [4].

We are interested in the "pull" effect and explore the characteristics of controllable behaviors and the way it triggers speaker-specific prosodic *signatures* i.e. mental state-specific patterns of trajectories of audiovisual prosodic parameters. As we aim to analyze and model these specific prosodic signatures, our system deals with a discrete set of *socio-communicative attitudes* to highlight interactive dimensions of face-to-face communication in realistic social contexts.

Our goal is the generation of expressive speech animations, comprising facial expressions, head movements, gaze direction and voice. To this extent, the proposed approach is based on the idea that audiovisual performances can be described as a combination of segmental and suprasegmental features which can be converted separately.

2. State of the art

2.1. Expressive voice

Most of the existing approaches to recognition and generation of acoustic emotions and affects use both prosodic (pitch, energy, speech rate) and segmental (MFCC, cepstral features) features [5]. The early study performed by Vroomen et al [6] has shown that affective states can be expressed accurately by manipulating pitch and duration using rules. More recently, several statistical models have proposed neutral-to-expressive speech conversion using expression-specific durations and f_0 contours [7, 8]. Mori et al [9] proposed an f_0 synthesis method for using subspace constraint in prosody. Wo et al [10] proposed a hierarchical prosody conversion method where the pitch contour of the source is decomposed into a hierarchical prosodic structure consisting of sentence, prosodic word, and subsyllable levels.

The above approaches suppose that the conversion system has access or estimates parts of the underlying structure of the linguistic content. Some approaches perform voice conversion using a direct non-linear mapping between audio frames from neutral and expressive corpora. Gaussian Mixture Models (GMM) are widely used in spectrum conversion to modify non linguistic information such as voice characteristics [11, 12]. Currently, the most succesful technique for adding expressivity to neutral voice involves training GMMs and converting both spectrum and prosodic features [7, 13].

2.2. Facial expressions

Statistical approaches have also been applied to expressive video sequences [14] or to motion capture data [15, 16] in an attempt to generate facial movements related to speech and facial expression. Such approaches include bilinear [14, 17] and trilinear models [15], enabling the decomposition of expressive speech into linguistic, paralinguistic and non linguistic factors. Yehia et al [18] studied the correlation of head and eyebrow movements with pitch contours and exploited the correlation [19, 17] to synthesize head motion from expressive speech.

The work of Bregler [17] describes a method for creating expressive facial animations and head motion using information from expressive speech. A motion generation model is created based on speech segments that are defined from the onset of a voiced region and ending on the onset of the next voiced region. Segment characterisation is done using features that include statistics related to speech rhythm and pitch signal. However, the position of the analyzed segment within the utterance is not taken into consideration.

3. Our approach

Audiovisual features that compose an expressive performance can be split into two complementary sets of features: segmental and suprasegmental. Therefore, in order to generate expressive

performances, we separately use a prosodic model to explicitely generate suprasegmental features and a frame-based approach for converting segmental features. Figure 1 presents the outline of the proposed system.

3.1. Modeling audiovisual prosody

The generation of audiovisual prosody is based on the theoretical approach described in [20], which proposes that prosodic information is encoded via global multiparametric contours with prototypical shapes. These shapes are coextensive to linguistic units and only depend on the length of the units (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" [21]. Aubergé and Bailly [22] proposed a more general framework which 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 (Superposition of Functional Contours) model [23] proposes a method for extracting these elementary multiparametric contours from a training corpus given the set of linguistic functions and their scopes. Therefore, in order to make a prediction for a given attitude, the only input required is the position and number of syllables of the desired linguistic units. This model supposes that the set of training utterances randomly samples the possible functions and their positions, lengths and numbers of their overlapping contributions. As we focus on analyzing audiovisual prosody, this theoretical model is extended for the joint modeling of melody, syllabic duration, head motion, gaze direction and upper-face action units (which we represent as blendshapes).

3.2. Generating mouth movements and vocal-tract related acoustic parameters

Features such as lip movements and spectra are mostly dependent on the underlying phoneme pronounced at a certain position within the speech. For this reason, generating expressive segmental feature from neutral performances is done using frame-based conversion approaches. We chose the conventional GMM method [11] [12] for audio and visual features. The input necessary for a new prediction is thus the neutral audiovisual performance.

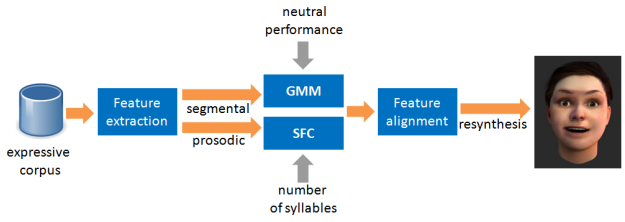

Figure 1: Segmental and prosodic features are extracted from the expressive corpus and are used in the generation of GMM and SFC models. For a given attitude, expressive contours are generated by providing a neutral phrase for the respective GMM model and the number of syllables for the SFC model. The predicted GMM contours are time-aligned with the prosodic contours so that the new audiovisual performance is resynthesized.

4. The expressive corpus

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.

4.1. Attitude recording

A starting point for the possible attitudes we considered was the Baron-Cohen's Mind Reading project [24]. The taxonomy proposed in this work gathers a total of 412 emotions grouped under 24 main categories, each consisting of several layers of subcategories. Due to the richness and complexity of the taxonomy, we chose a limited number of attitudes to be performed by two semi-professional native French actors under the active supervision of one director. They were asked to perform 35 utterances mainly restricted to one single sentence in the following attitudes: declarative (DC), exclamative (EX), question (QS), comforting (CF), fond-liking (FL), seductive (SE), fascinated (FA), jealous (JE), thinking (TH), doubtful-incredulous (DI), sneaky-humoring (SH), surprised-scandalised (SS), surprized-dazed (SD), responsible (RE), hurt-confronted (HC) and embarassed (EM). Throughout our study, we consider the DC attitude as "neutral".

The synchronized recording of voice signals and motion are performed using 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 were done in front of the camera, while seated, without additional markers such that the acting was not constrained.

4.2. Annotation and characterization

All utterances are automatically aligned with their phonetic transcription obtained using an automatic text-to-speech phonetizer [25]. The linguistic analysis (part-of-speech tagging, syllabation), the phonetic annotation and the automatic estimation of melody were further checked and manually corrected using Praat [26]. The subjects' performances are then caracterized at a suprasegmental level by the following parameters associated with each syllable:

- Melody: When the vocalic nucleus of the syllable is voiced, we sample the f_0 contour of the 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 supposes that the syllable compress/expands according to the intrinsic elasticity of its segmental constituents. Contingent pauses are included in the model by saturating the elastic lengthening [27].
- Motion: We sample the head and eye movements and eye-area expressions at syllable-level at three timestamps: 20%, 50% and 80% of the syllable duration. Principal component analysis is applied separately to

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

four facial segments: (a) rotation and translation of head movements, (b) brows, (c) eye blendshapes and (d) gaze direction, from which we keep 5, 3, 6 and 2 components respectively (explaining up to 80% of the data variance). In the following, we will refer to these sets of sampled PCA components as the stylized contours for motion.

For a given attitude, coherent sets of signatures for all variable-length utterances form families of contours. They are represented by stylized audiovisual parameter trajectories. Figures 2 and 3 illustrate families of contours obtained for the attitudes QS, FL and JE focusing on the following parameters: f_0 and second head motion component respectively. The sentences represented for each attitude contain 2, 3, 5, 7, 9 and 10 syllables. The figures also demonstrate the existence of prototypical contours that develop as the sentence lengths increase (rows) and the contrast between contours belonging to different attitudes (columns).

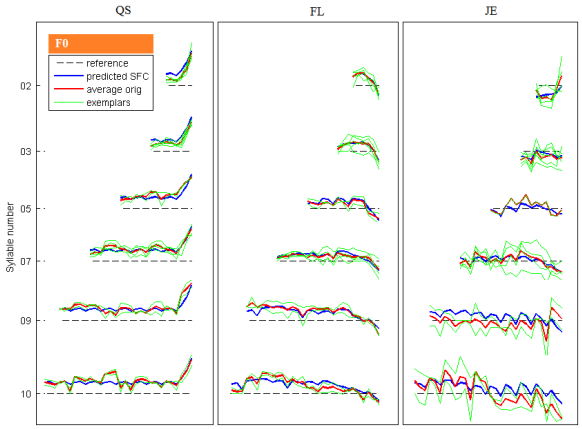

Figure 2: Family of contours for QS, FL, JE for the feature f_0. Rows represent different numbers of syllables and the family of contours is formed by the SFC predictions (blue trajectories). Overlaid are the stylized exemplars for each syllable-length (green trajectories), their computed average (red trajectory) and a reference line (dotted). The reference is considered at 210 Hz for female speakers and 110 Hz for male speakers. The trajectories show attitude-specific behaviors: rising towards the end (QS), lowering towards the end (FL), lowering towards the end for longer sentences (JE).

At the frame level, we use the STRAIGHT vocoder to extract frame-based features: mel cepstral coefficients and aperiodicities. The lower part of the face is represented by articulation blendshapes, from which we keep 8 components after applying PCA. All the features described are presented in Table 4.2.

5. Performance synthesis

An expressive performance is synthesized by converting the voice and movements of the neutral (DC) performance then enforcing prosodic features (temporal structure, pitch contour, upper-face motion, head and eye motion) using an attitude-specific prosodic model.

Voice is synthesized in two steps: first the prosodic manipulation of the DC stimuli is performed using the TD-PSOLA technique [28] such that the resulting audio signal contains the pitch and rhythm that were predicted using SFC. The second

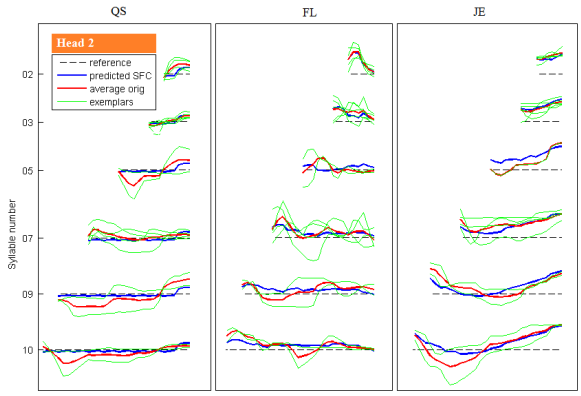

Figure 3: Family of contours for QS, FL, JE for the feature second head motion PCA component (corresponding to the head rising motion). The reference line is considered at the computed average of the component over all attitudes. The trajectories show attitude-specific behaviors: small rising of the head towards the end (QS), small head rising at the beginning (FL), lowering of the head at the beginning and rising towards the end (JE).

Table 1: Feature categorisation.

	Audio	**Visual**
Segmental	mel-cepstra aperiodicities	lower-face blendshapes (jaw open, puff, funnel etc)
Prosodic	f_0 rhythm	upper-face blendshapes (eyebrows up, squint etc) head movements eye gaze rhythm

synthesis step is carried using the STRAIGHT vocoder and consists in replacing the cepstral features with the time-aligned expressive mel-cepstra and aperiodicities that were generated using a GMM. Time-alignment for visual features is also done using the Dynamic Time Warping path [29] obtained by aligning the neutral and expressive performances.

Movements for the upper-part of the face are obtained by cubic spline interpolation of the reconstructed SFC stylized contours (3 landmarks per syllable) across all syllables of the target utterance. The movement reconstruction from SFC prediction is similar to the visual performance synthesis method described in [30]. The lower-face movements are then obtained using the GMM conversion along the DTW path.

The following list presents notations used throughout the experimental study. They describe the various methods used in audiovisual performance synthesis. The order of presentation reflects the expected similarity between converted and target stimuli, i.e. from the resynthesis of an original expressive performance to the frame-based conversion of all audiovisual features of neutral performances. Given an attitude and a target sentence, the evaluated animations are:

- *video*: stimuli combining the original voice and video of the performance (captured by the RGB camera of the Kinect)
- *self-transfer*: we feed the avatar of the speaker with an audio signal resynthesized by STRAIGHT from the orig-

inal expressive spectra, original lower-face expressive motion data, original time-structure and the reconstruction of the upper-face movements from the stylized original expressive motion data

- *exemplar-based*: given an expressive phrase with the same number of syllables, we feed the avatar of the speaker using the aligned resynthesized voice computed by STRAIGHT after the GMM conversion and the f_0 and syllable lenghtening factor (from which phoneme durations are obtained) of the target sentence, the reconstructed upper-face motions from the target sentence and the aligned lower-face motions of the GMM conversion from DC of the initial performance
- *prototype-based*: we feed the avatar of the speaker with the aligned resynthesized voice computed by STRAIGHT after the GMM conversion and the f_0 and syllable lengthening factor of the SFC prediction, the reconstructed upper-face motions from the SFC prediction and the aligned lower-face motions of the GMM conversion from DC of the initial performance
- *GMM with slope*: we feed the avatar of the speaker with both segmental and prosodic audiovisual features generated using the GMM conversion with prosodic feature approach described in [2]

We consider the results obtained in the *self-transfer* method as ground truth data as all segmental and prosodic features used in resynthesis are extracted from original expressive data. The following sections describe the perceptual tests carried in order to evaluate the methods we propose for expressive audiovisual speech generation: *prototype-based* and *exemplar-based*.

6. Evaluation

A series of subjective tests were conducted on the data generated using the four methods outlined above. The perceptive tests were carried on a crowdsourcing online platform. Only data from the native French participants are considered here.

6.1. Attitude recognition

We first performed a test to assess our low-level audiovisual coders/encoders, i.e. the STRAIGHT vocoder for audio and the Faceshift motion capture system and the Blender animation system for facial animation. We conducted a subjective test to compare the recognition performance of subjects watching original video data (cf. *video*) vs. the recreated audiovisual animations (cf. *self-transfer*). Participants were instructed to label 16 original *video* stimuli and then 16 synthetic *self-transfer* stimuli, with the appropriate attitude. Stimuli were randomly selected from a subset of 7 sentences out of the 35 exercise sentences. The online test can be found at [2].

A total of 77 French native speakers participated in this experience. Table 6.1 presents the precision and recall obtained for each attitude for the original videos and the animations.

For the video test, the best recognized attitudes are: DC, QS, CF, SE, TH, SH, SS and EM. High recognition rates are also obtained for DC, QS and SS for the animation test. However the low precision values for DC and EX show that these attitudes have a higher chance of being chosen when another attitude is played. The least recognized attitudes are: EX, DI, RE, HC. These attitudes are subtle in audiovisual changes (as opposed to SS for example) and have meanings that are more difficult to grasp and dissociate from the sense of the phrase. All these attitudes are generally confused with DC in both tests. Some notable results: EX is generally confused with DC, DI with EX, SD with DC and DI, RE with EX. As expected, all recognition rates decrease as the animation version is used and it is most apparent for attitudes such as: DI, SH, SD, RE and EM.

[2] http://www.gipsa-lab.fr/ adela.barbulescu/test1/

Table 2: Precision and recall obtained for videos and animations for all attitudes. Values above 0.3 are outlined in bold.

	Video		**Animation**	
	Precision	**Recall**	**Precision**	**Recall**
DC	0.25	**0.61**	0.18	**0.66**
EX	0.07	0.09	0.07	0.14
QS	**0.50**	**0.60**	**0.40**	**0.31**
CF	0.26	**0.45**	0.16	0.23
FL	**0.51**	0.29	0.29	0.15
SE	**0.60**	**0.45**	**0.56**	0.19
FA	**0.39**	0.21	**0.43**	0.20
JE	**0.50**	0.10	0.10	0.04
TH	**0.50**	**0.39**	**0.34**	0.25
DI	0.13	0.18	0.09	0.08
SH	**0.32**	**0.43**	0.14	0.16
SS	**0.66**	**0.58**	**0.38**	**0.35**
SD	0.18	0.14	0.09	0.06
RE	0.11	0.09	0.00	0.00
HC	0.23	0.13	0.15	0.10
EM	**0.47**	**0.36**	0.25	0.18

6.2. Comparative mean opinion scores (MOS)

Evaluation of the proposed conversion methods was carried using a ranking test paradigm similar to [31] in which the animated stimuli is obtained from following methods: *self-transfer*, *exemplar-based*, *prototype-based* and *GMM with slope*. As we have a small expressive corpus, we performed a leave-one-out GMM training per attitude for audio and visual parameters separately and 7 test sentences.

Participants to the test are asked to rate 4 animations per trial and then to specify the level of their expressivity relative to an indicated attitude. Instead of choosing from a limited list of possible scores to describe the perceived expressity of one animation, the participant is able to retrieve relative and absolute perceptual information by placing symbolic icons in a ranking rectangular-shaped grid. The horizontal sides of the grid effectively represent quality ratings, from Bad to Excellent. The vertical axis has no dimension and just ease the layout of icons by avoiding messy superpositions. Subjects can play animations on demand by clicking on the four icons representing the tested systems and can then move the icons anywhere within the grid, considering that verticals represent identical quality options. Ranking an animation is equivalent to associating it with a numerical value which ranges continuously from 0 to 5 such that $Bad = 0$, $Average = 2.5$ and $Excellent = 5$ (see figure 4).

As this test requires a longer time for completing one validation, only a limited subset of attitudes is chosen. The choice is based both on the results of the previous test and the level of discrimination existing in the attitude subset. The attitudes chosen are: QS, CF, SE, TH, DI, SH and EM. Each of 7 trials

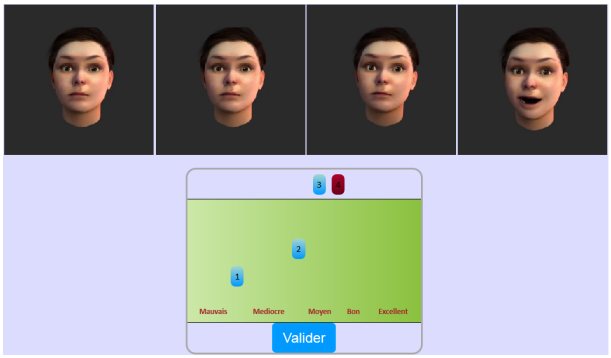

Figure 4: Snapshot of the ranking area. The 4 animation stimuli are placed above the ranking grid, where the first two animations are ranked, the last two are not ranked and the 4th is being played.

presents 4 animated versions for a given attitude of a random sentence (out of the 7 test sentences). The associations between icons and versions is randomized within each trial. The online test can be found at [3].

A total of 41 native French subjects thus evaluated 28 animations. The statistical significance was assessed using an ANOVA test which considers the *X position* of the ranking as a continuous variable and *version* as factor. The main effect of the factor *version* was significant ($p < 0.005$), thus allowing for further analysis of the results obtained for each version.

All version averages are situated between $Mediocre$ and $Good$. As expected, the best results were obtained for the self-transfer version with an average and standard deviation of 2.87 ± 1.03, while the worst were obtained for the GMM with slope version with 1.80 ± 1.07. A paired t-test for the prototype (with 2.16 ± 1.04) vs. exemplar-based (with 2.33 ± 1.09) versions showed that there is no significant statistical difference between the observations. This can be explained by the intrinsic quality of our speech resynthesis system. Table 6.2 presents the average values obtained for each attitude and version used in this test:

Table 3: Average ranking values of the four methods in all attitudes.

	Slope GMM	Prototype	Exemplar	Self-transfer
QS	1.18	2.69	2.52	3.28
CF	2.28	2.12	1.97	2.29
SE	2.12	2.02	2.07	2.78
TH	1.62	2.36	2.57	2.69
DI	2.01	2.03	2.40	2.85
SH	1.53	1.98	2.61	3.29
EM	1.91	2.01	2.43	3.18

7. Discussion and future work

The ranking test results show that on average the GMM slope method is outperformed by the two proposed methods while the self-transfer method presents the highest test scores. The lowest scores obtained for the GMM with slope version are observed for attitudes in which f_0 and speech rhythm have a big impact on perception: QS, TH, SH. According to comments retrieved by subjects, a few animations presented unnatural audio which lead to a lower ranking. These samples can be both attributed to the prototype or exemplar-based methods, due to the phoneme duration generation algorithm, thus explaining the bigger scores obtained by the *GMM with slope* method in the case of CF and SE.

All version averages are situated between $Mediocre$ and $Good$. The general lower scores obtained in the case of CF and DI can also be explained by the fact that these attitudes are more difficult to be recognized as shown by the attitude recognition test results (see Table 6.1).

There are many opportunities to improve the presented methods: one proposition is that of combining the two proposed approaches. We show that the two proposed methods are complementary by computing the best recognition scores i.e. choosing the maximum between the two methods for each subject and test sentence. The selected method for one test is either exemplar or prototype-based and it is noted EOP. In this case, the results obtained are presented in Figure 5.

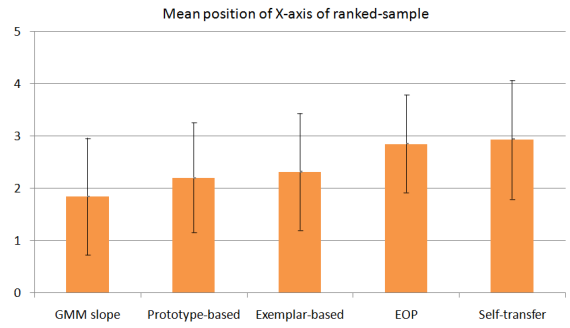

Figure 5: Results of the perceptual ranking test including EOP: the maximum score between *prototype-based* and *exemplar-based* results. The bars represent the rank value per version averaged across subjects and attitudes.

We observe that the EOP average (2.85) is higher than the averages obtained by the *prototype-based* and the *exemplar-based* methods separately (2.20 and 2.32 respectively) and close to the *self-transfer* average (2.94). There is a significant statistical difference between each of the two proposed methods and EOP thus concluding that the two methods can be combined to retrieve better results. In fact, a paired t-test shows that there is no statistical difference between the EOP and the *self-transfer* method.

Another direction of improvement for the visual component is that of choosing a better dimensionality reduction method. A particular mention is that of encoding left-right head rotation and translation such that the prosodic model only considers the absolute displacement and not the direction of the movement. Future experiments should focus on the best recognized attitudes in the video version (QS, FL, SE, FA, JE, TH, SH, SS, EM) and should also investigate the contributions of audio and video modalities separately.

[3] http://www.gipsa-lab.fr/ adela.barbulescu/test2/

8. Conclusion

We described and evaluated two methods for combining segmental and suprasegmental features in the conversion from neutral to expressive audiovisual speech. Experimental results show that both approaches outperform previous work [2] that used a unique frame-based conversion approach for all audiovisual features. The prototype and the exemplar-based methods do not present significant statistical differences.

Considering that we used a small corpus with validated exemplars, the results show promise that using SFC to predict prosody can outperform exemplar-based methods in a larger corpus where exemplars may fail due to lack of exemplars with the appropriate lengths or low quality data. Future work includes combining the two proposed methods and testing on utterances with greater expressive and size variability.

9. Acknowledgements

This work is supported by the LabEx PERSYVAL-Lab (ANR–11-LABX-0025). 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.

10. References

[1] R. Queneau, *Exercises in style*. New Directions Publishing, 2013.

[2] A. Barbulescu, T. Hueber, G. Bailly, R. Ronfard *et al.*, "Audiovisual speaker conversion using prosody features," in *International Conference on Auditory-Visual Speech Processing*, 2013.

[3] K. R. Scherer, "Vocal affect expression: a review and a model for future research." *Psychological bulletin*, vol. 99, no. 2, p. 143, 1986.

[4] D. Bolinger, *Intonation and its uses: Melody in grammar and discourse*. Stanford University Press, 1989.

[5] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, no. 1, pp. 39–58, 2009.

[6] J. Vroomen, R. Collier, and S. J. Mozziconacci, "Duration and intonation in emotional speech." in *Eurospeech*, 1993.

[7] J. Tao, Y. Kang, and A. Li, "Prosody conversion from neutral speech to emotional speech," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 4, pp. 1145–1154, 2006.

[8] Z. Inanoglu and S. Young, "A system for transforming the emotion in speech: combining data-driven conversion techniques for prosody and voice quality." in *INTERSPEECH*, 2007, pp. 490–493.

[9] S. Mori, T. Moriyama, and S. Ozawa, "Emotional speech synthesis using subspace constraints in prosody," in *Multimedia and Expo, 2006 IEEE International Conference on*. IEEE, 2006, pp. 1093–1096.

[10] C.-H. Wu, C.-C. Hsia, C.-H. Lee, and M.-C. Lin, "Hierarchical prosody conversion using regression-based clustering for emotional speech synthesis," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 18, no. 6, pp. 1394–1405, 2010.

[11] Y. Stylianou, O. Cappe, and E. Moulines, "Statistical methods for voice quality transformation." in *EUROSPEECH*, 1995.

[12] T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 8, pp. 2222–2235, 2007.

[13] R. Aihara, R. Takashima, T. Takiguchi, and Y. Ariki, "Gmm-based emotional voice conversion using spectrum and prosody features," *American Journal of Signal Processing*, vol. 2, no. 5, pp. 134–138, 2012.

[14] E. Chuang and C. Bregler, "Performance driven facial animation using blendshape interpolation," *Computer Science Technical Report, Stanford University*, vol. 2, no. 2, p. 3, 2002.

[15] D. Vlasic, M. Brand, H. Pfister, and J. Popović, "Face transfer with multilinear models," in *ACM Transactions on Graphics (TOG)*, vol. 24, no. 3. ACM, 2005, pp. 426–433.

[16] Y. Cao, P. Faloutsos, E. Kohler, and F. Pighin, "Real-time speech motion synthesis from recorded motions," in *Proceedings of the 2004 ACM SIGGRAPH/Eurographics symposium on Computer animation*. Eurographics Association, 2004, pp. 345–353.

[17] E. Chuang and C. Bregler, "Mood swings: expressive speech animation," *ACM Transactions on Graphics (TOG)*, vol. 24, no. 2, pp. 331–347, 2005.

[18] H. Yehia, T. Kuratate, and E. Vatikiotis-Bateson, "Facial animation and head motion driven by speech acoustics," in *5th Seminar on Speech Production: Models and Data*. Kloster Seeon, Germany, 2000, pp. 265–268.

[19] C. Busso, Z. Deng, M. Grimm, U. Neumann, and S. Narayanan, "Rigid head motion in expressive speech animation: Analysis and synthesis," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 3, pp. 1075–1086, 2007.

[20] Y. Morlec, G. Bailly, and V. Aubergé, "Generating prosodic attitudes in french: data, model and evaluation," *Speech Communication*, vol. 33, no. 4, pp. 357–371, 2001.

[21] I. Fónagy, E. Bérard, and J. Fónagy, "Clichés mélodiques," *Folia linguistica*, vol. 17, no. 1-4, pp. 153–186, 1983.

[22] V. Aubergé and G. Bailly, "Generation of intonation: a global approach." in *EUROSPEECH*, 1995.

[23] G. Bailly and B. Holm, "Sfc: a trainable prosodic model," *Speech Communication*, vol. 46, no. 3, pp. 348–364, 2005.

[24] S. Baron-Cohen, *Mind reading: the interactive guide to emotions*. Jessica Kingsley Publishers, 2003.

[25] G. Bailly, T. Barbe, and H.-D. Wang, "Automatic labeling of large prosodic databases: Tools, methodology and links with a text-to-speech system," in *The ESCA Workshop on Speech Synthesis*, 1991.

[26] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot international*, vol. 5, no. 9/10, pp. 341–345, 2002.

[27] P. Barbosa and G. Bailly, "Characterisation of rhythmic patterns for text-to-speech synthesis," *Speech Communication*, vol. 15, no. 1, pp. 127–137, 1994.

[28] E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech communication*, vol. 9, no. 5, pp. 453–467, 1990.

[29] D. J. Berndt and J. Clifford, "Using dynamic time warping to find patterns in time series." in *KDD workshop*, vol. 10, no. 16. Seattle, WA, 1994, pp. 359–370.

[30] A. Barbulescu, R. Ronfard, G. Bailly, G. Gagneré, and H. Cakmak, "Beyond basic emotions: expressive virtual actors with social attitudes," in *Proceedings of the Seventh International Conference on Motion in Games*. ACM, 2014, pp. 39–47.

[31] M. Pouget, T. Hueber, G. Bailly, and T. Baumann, "Hmm training strategy for incremental speech synthesis," in *Interspeech*, 2015.