On the quality of an expressive audiovisual corpus: a case study of acted speech

Slim Ouni[1]*, Sara Dahmani*[1]*, Vincent Colotte*[1]

[1]Université de Lorraine, LORIA, UMR7503, Vandoeuvre-lès-Nancy, F-54506, France

FirstName.LastName@loria.fr

Abstract

In the context of developing an expressive audiovisual speech synthesis system, the quality of the audiovisual corpus from which the 3D visual data will be extracted is important. In this paper, we present a perceptive case study on the quality of the expressiveness of a set of emotions acted by a semi-professional actor. We have analyzed the production of this actor pronouncing a set of sentences with acted emotions, during a human emotion-recognition task. We have observed different modalities: audio, real video, 3D-extracted data, as unimodal presentations and bimodal presentations (with audio). The results of this study show the necessity of such perceptive evaluation prior to further exploitation of the data for the synthesis system. The comparison of the modalities shows clearly what the emotions are, that need to be improved during production and how audio and visual components have a strong mutual influence on emotional perception.

Index Terms: expressive audiovisual speech, facial expressions, acted speech, audiovisual perception.

1. Introduction

Within the framework of developing audiovisual speech synthesis techniques, commonly known as the animation of a 3D virtual talking head synchronously with acoustics, providing an expressive talking head can highly increase the naturalness and the intelligibility of the audiovisual speech. These techniques are based on audiovisual corpus that should convey convincing expressive emotions. For this purpose, we are investigating acted speech, uttered by an actor with different emotions. In fact, as our purpose is not to investigate expressions for recognition neither for perception, characterizing speech in this context is valuable. In fact, the virtual talking head is in the same situation as an actor: making an effort to provide convincing and visible expressions, even though with some exaggeration [1]. In fact, human expressions are not always visible, and in the majority of cases they are subtle and some human speakers are barely expressive [2].

When developing a talking head, our goal is that expressions are easily perceived by the majority of the users. When recording acted speech, we should make sure that this acting is convincing. For this reason, it is reasonable to perform a perceptual evaluation, where during a human recognition task, the quality of the expressivity of an audiovisual corpus is assessed. Several researches have been conducted to study expressive audiovisual speech mainly from perceptive point of view [3, 4, 5, 6]. The main addressed topic is the correlation between f_0 an eyebrow and head movements [7, 5, 8]. The main purpose of our study is to show the importance of evaluating how an expressive audiovisual speech is perceived to try to quantify the quality of the expressiveness of the expressive audiovisual speech. We have observed different modalities: audio, real video, 3D-extracted data, as unimodal presentations and bimodal presentations (with audio).

In this paper, we present our recent research where we conducted a case study of a semi-professional actor who uttered a set of sentences for 6 different emotions in addition to neutral speech. We have recorded concurrently audio and motion capture data using a multimodal acquisition platform. This platform allows capturing the movement of reflective markers, electromagnetic sensors and the movement of the eyes without any markers. This platform has been designed to acquire 3D data adapted to expressive audiovisual speech. In the following paragraphs, we present the recorded expressive audiovisual speech corpus, then we present the conducted perceptive evaluation.

2. Expressive audiovisual speech corpus

2.1. Corpus Acquisition

2.1.1. Setup

We have used a multimodal acquisition system composed of a motion-capture system (VICON) using optical reflective markers, an articulograph (AG501) using electromagnetic sensors and a markerless motion capture (RealSense) that allows tracking the movement of the face without markers. The Vicon system is based on 4 Vicon cameras (MX3+) using modified optics for near range. The cameras were placed at *approx.* 150 cm from the speaker. Vicon Nexus software provides the 3D spatial position of each reflective marker at a sampling rate of 100 Hz. Reflective markers of 3 mm in diameter have been glued on the upper part of the actor face. They are aimed to capture facial expressions. The articulograph (EMA) sensors allow tracking mainly the lip movement. The EMA technique allows capturing finely these gestures at a sampling rate of 250 Hz and handle the occlusion problem, in the case of bilabial pronunciation (even hidden, it is always possible to track the sensors). The RealSense system was used to capture mainly the shape and movement of the eyes. As this system is markerless, it is appropriate for this task.

Figure 1 shows the setup where Vicon markers and EMA sensors are placed on the face, and the RealSense is placed in front of the actor. We have placed 5 extra markers on the top of the head to remove the head movement. Figure 2 presents the layout of the markers, sensors and virtual markers on the face of the actor. The audio was acquired simultaneously with the spatial data using a unidirectional microphone. To synchronize the audio channel and the motion capture channel, we have used an in-house electronic device.It triggers simultaneously an infrared lamp captured by the Vicon system and RealSense, and a high-pitched sound generated by a piezoelectric buzzer for audio captured by the articulograph system. We have developed

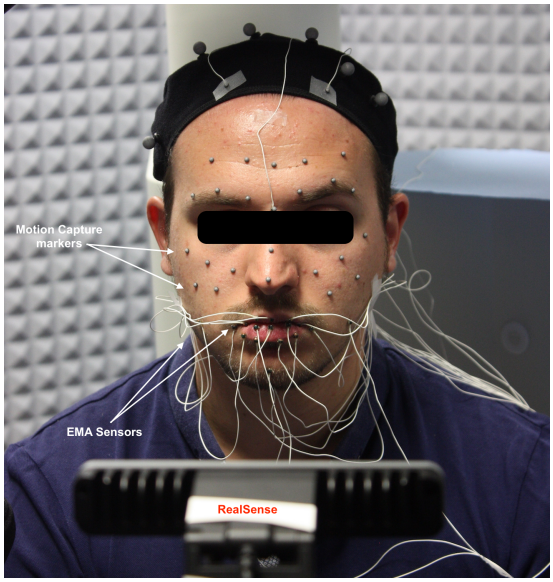


Figure 1: *The positions of the reflective markers, EMA sensors, and RealSense, relatively to the face of the actor.*

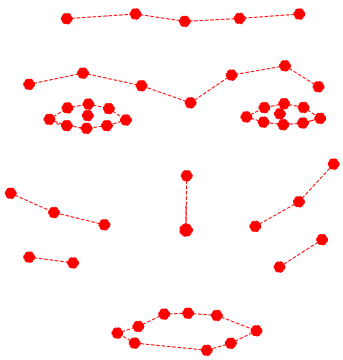


Figure 2: *The layout of the markers on the face of the actor.*

several tools and algorithmic techniques to process and merge the data accurately.

2.1.2. Material

A semi-professional actor has been asked to utter 10 French sentences for 7 different emotions (neutral, joy, surprise, fear, anger, sadness, and disgust). These sentences were 4 to 5 words in length. The actor has used a technique called *exercise in style*, where he dissociates the semantics of the syntax of the sentences and acts the same sentences in different styles.

The linguistic content of the sentences does not help to identify the expressed emotion (dissociation between semantics and syntax).

The ten sentences were presented one at a time on the screen in front of the actor who uttered them showing the same consistent emotion. In this context, the emotions should be considered as acted ones as they are a bit exaggerated as in the case of a play at the theater.

3. Perceptual Evaluation

The main question we are investigating here is whether the actor was able to convey the different expressions correctly the human receivers. This is essential to make a decision about how good is the acquired expressive audiovisual corpus. As the final data will be used to develop an expressive audiovisual speech synthesis. This data is based on 3D points (markers/sensors on the face). For this reason, evaluating this presentation is helpful to see whether a minimal presentation of the face is sufficient to model facial expressivity. Similar evaluation technique has been used in the past to evaluate an audiovisual speech synthesis system [9]. When using the 3D facial data, we usually remove head movement to make audiovisual synthesis process smooth and technically easier to concatenate segments. In this evaluation, we observed whether head movement can help to better communicate emotion.

3.1. Material and Participants

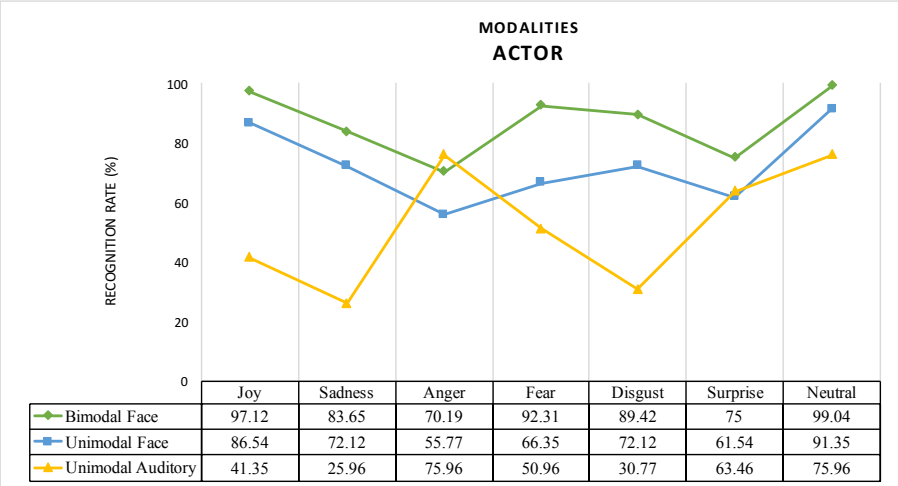
For each emotion, 10 sentences have been presented to thirteen participants (adults, 3 females, 10 males). They were asked to identify the emotion expressed by the actor. We studied two modalities : unimodal and bimodal. For the unimodal presentation, we considered the following channels : audio, 3D points without head movement, 3D points with head movement and the real video of the actor face. For the bimodal presentation, we combined the audio with different visual channels. The presentations were organized in 3 successive blocks : (1) unimodal audio, (2) unimodal visual only, and (3) bimodal audiovisual. Within each block the sentences of each emotion were presented randomly. For the visual channels during blocks (2) and (3), we presented 3D points without head movement, 3D points with head movement and finally, the video of the actor face. We should notice that we used the 3D points instead of a 3D talking head controlled by these 3D points to avoid any bias that might be introduced during the evaluation because of the quality of the 3D model.

The different stimuli were presented via a web application [10] where participants were asked to identify each expressed emotion by selecting the right one from a list of 7 emotions. This web application allows controlling the experiment and it makes several verification, to make sure that the experiment was performed in good conditions.

3.2. Results

We have analyzed participant answers of each presentation. The results are summarized in Figure 3, Figure 4 and Figure 5. In each figure, a table presents the average recognition rate for each emotion and each presentation.

Figure 3 presents the actor performance. The first presentation is the video with audio of the actor, the second is the video without audio, and the third is the audio alone. In bimodal presentation, which corresponds to the full information conveyed by the actor to express the different emotions, all the expressions were well recognized (recognition rate higher than 70%). This is reassuring that globally the different expressions were acted correctly. The unimodal presentation of the face is still having relatively good recognition rates, but lower than those of bimodal presentation. This shows that the visual modality (the actor face) provides reasonable good information to decode the expressed emotion. The unimodal auditory presentation result shows that some expressions were not well recognized as sadness (25%), disgust (30.77%), joy (41.35%)



	Joy	Sadness	Anger	Fear	Disgust	Surprise	Neutral
Bimodal Face	97.12	83.65	70.19	92.31	89.42	75	99.04
Unimodal Face	86.54	72.12	55.77	66.35	72.12	61.54	91.35
Unimodal Auditory	41.35	25.96	75.96	50.96	30.77	63.46	75.96

Figure 3: *Actor performance - 3 presentations: Bimodal actor face, unimodal actor face, and unimodal auditory. Table presenting recognition rates for each emotion across the different presentations.*

and fear (50.96%). Anger (75.96%) and neutral (75.96%) have the highest recognition rate in this presentation. Moreover, the recognition rate of anger in unimodal auditory is higher than that in unimodal face (55.77%) and even in bimodal face presentation.This indicates that the actor makes more emphasis on audio to convey anger and less on the face.

We recall that when developing the expressive talking head, we are not going to use the video of the actor directly, but the 3D points corresponding to the markers and sensors on the face of the actor. The purpose of this perceptual evaluation is also to see whether the landmarks are sufficient to express the different emotions. Although this information is minimalist, it can give a preview of the quality of the talking head rendering, as it will be based on this data. Figure 4 shows the results of the recognition rates for the different emotions when we only present the 3D points (as presented in Figure 2). We present two modalities (3D points with audio, 3D points without audio) and with or without head movement. Overall, bimodal presentations have higher recognition rate than unimodal 3D point presentations, except that of sadness, which is expected as in the unimodal auditory presentation, the emotion sadness has the lowest recognition rate. The results of unimodal presentations show that some expressions have very low recognition scores as anger (5.77%), fear (21.15%), disgust (17.31%) and to a lesser extent surprise (40.38%). This shows that the 3D points representing the face are not sufficient to convey these expressions, and that the audio channel is an essential complementary modality to convey emotions. We notice that auditory modality increases well for the emotions joy, fear, disgust and surprise.

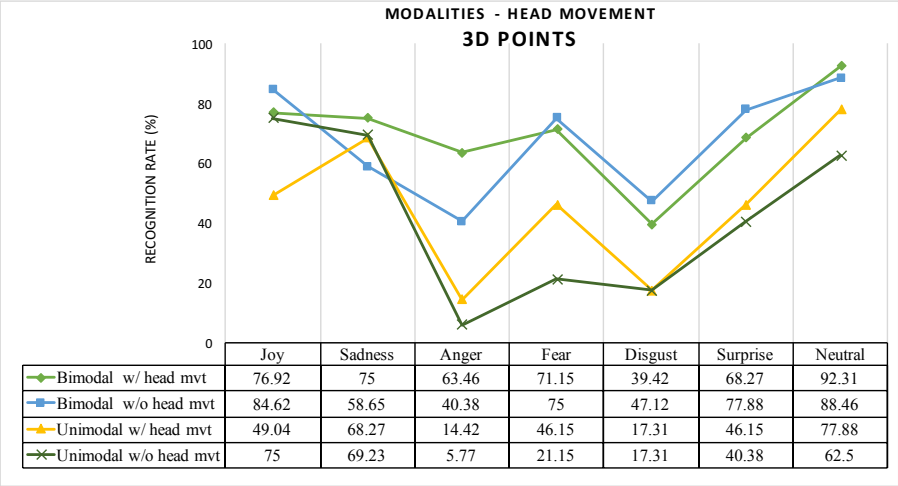
The impact of the head movement during audiovisual expressive speech is not clear. In fact, during bimodal 3D point presentation, the head movement has improved the recognition rate of a set of expressions (neutral, anger, sadness) but not the others (surprise, disgust, fear, joy). The highest recognition rate improvement (compared to the presentations without head movement) was that of sadness and anger. During the unimodal 3D point presentation (without audio), the result shows that the head movement improved the recognition rate of some expressions (neutral, surprise, fear and anger) but decreased the recognition rate of joy. We notice that head movement improved the recognition of fear during unimodal 3D points (from 21.15% to 46.15%).

Figure 5 helps to compare each channel as a unimodal presentation. Thus, we present the recognition rate per emotion when during the auditory unimodal presentation, the actor-face unimodal presentation, 3D point unimodal presentation with and without head movement. Across all the unimodal presentations, the actor face presented the highest recognition rate (except that of anger). During the 3D point unimodal presentations, performances are lower than during the actor-face presentation. We also notice that unimodal auditory has better performances than 3D point unimodal presentations (except for joy and sadness).

4. Discussion

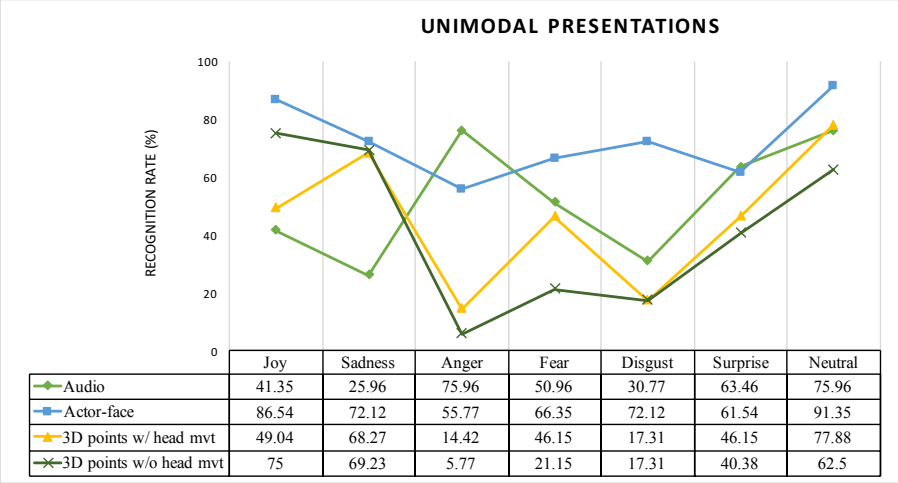
The quality of the recorded corpus can influence the quality of the audiovisual speech synthesis. When dealing with expressive audiovisual speech, it is important to make sure that expressivity is well perceived. The results of our current study show the importance of evaluating the quality of the expressive recorded data at an earlier stage. The first benefit is to see how good is the audiovisual acting of the actor. In this case study, and overall, the result seems to be good and the acting is convincing. When combining audio channel and visual channel, this study show improvement in recognition task scores when the two channels were combined. The audio alone does not seem to convey the right expression all the time. In our study the linguistic content of the sentences does not help participants in the perceptive evaluation to identify the expressed emotion, to avoid any bias. Overall, the bimodal presentations are better than unimodal ones. This confirms the importance of considering expressive audiovisual speech as a bimodal signal [10]. Since our earlier work in audiovisual speech synthesis, we consider both channels acoustic and visual together [10, 11]. This case study seems to confirm that this is still the best way to convey emotions correctly.

The bimodal 3D point presentation with head movement seems to provide good results. However, we cannot confirm that head movement helps in improving expressiveness. When



	Joy	Sadness	Anger	Fear	Disgust	Surprise	Neutral
Bimodal w/ head mvt	76.92	75	63.46	71.15	39.42	68.27	92.31
Bimodal w/o head mvt	84.62	58.65	40.38	75	47.12	77.88	88.46
Unimodal w/ head mvt	49.04	68.27	14.42	46.15	17.31	46.15	77.88
Unimodal w/o head mvt	75	69.23	5.77	21.15	17.31	40.38	62.5

Figure 4: *3D point presentations - bimodal 3D points with and without head movement, unimodal 3D points with and without head movement. Table presenting recognition rates for each emotion across the different presentations.*



	Joy	Sadness	Anger	Fear	Disgust	Surprise	Neutral
Audio	41.35	25.96	75.96	50.96	30.77	63.46	75.96
Actor-face	86.54	72.12	55.77	66.35	72.12	61.54	91.35
3D points w/ head mvt	49.04	68.27	14.42	46.15	17.31	46.15	77.88
3D points w/o head mvt	75	69.23	5.77	21.15	17.31	40.38	62.5

Figure 5: *Unimodal presentations - Unimodal auditory, unimodal actor faces, Unimodal 3D points with and without head movement. Table presenting recognition rates for each emotion across the different presentations.*

observing Figure 3, we may suggest to the actor improving facial expressions to better perform the emotions with lower recognition rates (for instance, anger and surprise) and to improve voice performance for sadness and disgust emotions, for instance. Each low recognition score should help to identify expressivity-related problem. The problematic expressions will be presented to the actor and a debriefing will be held to make the acting more convincing.

In Figure 5, an emotion presented by 3D points (without audio) that has a low recognition rate when compared with the same emotion during actor-face presentation (without audio) indicates that the 3D points were not sufficient to express that given emotion accurately (for instance, anger, disgust and surprise). This may mean that the layout presented in Figure 2 should be improved or modified to better capture facial expressions. We should not forget that the 3D point presentation is an important simplification of the facial information. However, this simplification should allow capturing the important expressive facial features. Currently, we are continuing investigating the results of the analysis. In particular, we are considering studying the confusion of a given expression with another one, and what that can tell us about the expression itself. We also intend to study whether there is a correlation between perceptual results of the experiment and anatomical analysis of the actor face.

5. Acknowledgements

This work was supported by Region Lorraine (COREXP Project) and the EQUIPEX Ortolang. We also thank Florian Sietzen for his participation in this work.

6. References

[1] S. Ouni, V. Colotte, S. Dahmani, and S. Azzi, "Acoustic and Visual Analysis of Expressive Speech: A Case Study of French Acted Speech," in *Interspeech 2016*, vol. 2016. San Francisco, United States: ISCA, Nov. 2016, pp. 580 – 584. [Online]. Available: https://hal.inria.fr/hal-01398528

[2] U. Hess and P. Thibault, *Why the Same Expression May Not Mean*

the Same When Shown on Different Faces or Seen by Different People. London: Springer London, 2009, pp. 145–158.

[3] E. Vatikiotis-Bateson, K. G. Munhall, Y. Kasahara, F. Garcia, and H. Yehia, "Characterizing audiovisual information during speech," in *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, vol. 3, Oct 1996, pp. 1485–1488 vol.3.

[4] B. Granstrom, D. House, and M. Lundeberg, "Prosodic cues in multimodal speech perception," in *ICPhS*, San Francisco, USA, 1999, pp. 655–658.

[5] K. G. Munhall, J. A. Jones, D. E. Callan, T. Kuratate, and E. Vatikiotis-Bateson, "Visual prosody and speech intelligibility head movement improves auditory speech perception," *Psychological science*, vol. 15, no. 2, pp. 133–137, 2004.

[6] M. Swerts and E. Krahmer, "Facial expression and prosodic prominence: Effects of modality and facial area," *Journal of Phonetics*, vol. 36, no. 2, pp. 219 – 238, 2008.

[7] C. Cave, I. Guaitella, R. Bertrand, S. Santi, F. Harlay, and R. Espesser, "About the relationship between eyebrow movements and fo variations," in *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, vol. 4, Oct 1996, pp. 2175–2178 vol.4.

[8] J. Beskow, B. Granstrom, and D. House, "Visual correlates to prominence in several expressive modes," in *Proc. Interspeech*, Pittsburg, PA, USA, 2006, p. 12721275.

[9] G. Bailly, G. Gibert, and M. Odisio, "Evaluation of movement generation systems using the point-light technique," in *Speech Synthesis, 2002. Proceedings of 2002 IEEE Workshop on*. IEEE, 2002, pp. 27–30.

[10] S. Ouni, V. Colotte, U. Musti, A. Toutios, B. Wrobel-Dautcourt, M.-O. Berger, and C. Lavecchia, "Acoustic-visual synthesis technique using bimodal unit-selection," *EURASIP Journal on Audio, Speech, and Music Processing*, no. 2013:16, Jun. 2013.

[11] A. Toutios, U. Musti, S. Ouni, V. Colotte, B. Wrobel-Dautcourt, and M.-O. Berger, "Setup for Acoustic-Visual Speech Synthesis by Concatenating Bimodal Units," in *Interspeech*, Makuhari, Japan, 2010.