Acoustic and Visual Analysis of Expressive Speech: A Case Study of French Acted Speech

Slim Ouni, Vincent Colotte, Sara Dahmani, Soumaya Azzi

Acoustic and visual analysis of expressive speech: a case study of French acted speech

Slim Ouni[1,2,3]*, Vincent Colotte*[1,2,3]*, Sara Dahmani*[1,2,3]*, Soumaya Azzi* [4]

[1]Université de Lorraine, LORIA, UMR7503, Vandoeuvre-lès-Nancy, F-54506, France
[2]Inria, Villers-lès-Nancy, F-54600, France
[3]CNRS, LORIA, UMR 7503, Vandoeuvre-lès-Nancy, F-54506, France
[4]Polytech Clermont-Ferrand, Aubière, F-63178, France

FirstName.LastName@loria.fr

Abstract

Within the framework of developing an expressive audiovisual speech synthesis, an acoustic and visual analysis of expressive acted speech is proposed in this paper. Our purpose is to identify the main characteristics of audiovisual expressions that need to be integrated during synthesis to provide believable emotions to the virtual 3D talking head. 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. The acoustic and the visual data have been analyzed. The main finding is that although some expressions are not well identified, some expressions were well characterized and tied in both acoustic and visual space.

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

1. Introduction

Speech communication is consisting of the acoustic signal that carries the acoustic modality, and the facial deformation represents the visual modality. This audiovisual communication is not just expressing the phonetic content of spoken text but it allows conveying a mood or an emotion (joy, sadness, fear, etc.). Several researches have been conducted to study expressive audiovisual speech but mainly from perceptive point of view [1, 2, 3, 4]. The main addressed topic is the correlation between f_0 and eyebrow and head movements [5, 3, 6]. In acoustic domain, identifying emotion features has been mainly addressed in automatic speech recognition [7, 8, 9] and also in the synthesis domain [10, 11, 12]. As it has been highlighted in [9], general rules are difficult to define as different studies may have contradictory conclusions for a same given emotion. The emotion conveyed is dependent on language and culture [13], and also on the speaker [14].

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. In this context, we are focusing on finding the main acoustic and visual characteristic that should be provided during synthesis to convey convincing expressive audiovisual speech system. For this reason, we are investigating acted speech, uttered by an actor in 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. In fact, human expressions are not always visible, and in the majority of cases they are subtle and some human speakers are barely expressive. When developing a talking head, our goal is that expressions are easily perceived by the majority of the users.

In this paper, we present a study 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. The acoustic and the visual data have been analyzed.

2. Expressive audiovisual speech corpus

2.1. Corpus Acquisition

2.2. Setup

We have used a motion-capture system 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 face of the actor. The positions of these markers on the face are presented in Figure 1: we have placed 7 markers on the lips, 2 on each eyebrow, 3 around the nose, 4 on the forehead, 6 on the lower face (chin-jaw) and 5 markers on each side of the face between the cheek and the eye. Five extra markers have been used to remove the head movement, as we are not using the feature in this study. 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 that triggers simultaneously an infrared lamp captured by the Vicon system and high-pitched sound generated by a piezoelectric buzzer for audio.

2.3. Material

A semi-professional actor has been asked to utter 10 French sentences for 7 different emotions (neutral, joy, surprise, fear, anger, sadness, disappointment). 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 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. Data processing and analysis

The motion capture data were obtained directly as 3D coordinates for each marker using a proprietary software that process the data. We have developed a tool that allows the synchronization of the two streams (motion capture data and audio). For the visual data, we track the first frame where the infra-red light appears, and for audio, we track the first high-pitched acoustic signal. We used this information to align both streams.

3.1. Visual data analysis

The visual data of the recorded corpus consist of 34 points which represent a high-dimensional space, which may make the analysis lengthy and difficult to interpret. For this reason, we have performed a dimensionality reduction by applying a principal component analysis (PCA) on the data. The corpus has been divided into smaller corpora, where each one represents a set of ten sentences for a given emotion. We have applied the PCA on each corpus to identify what is the major direction when a given emotion is dominant.

3.2. Acoustic data analysis

The acoustic data are recorded at the same time as the visual recordings. For each sentence, we carry out an automatic alignment at the phone level followed by a manual check. We also compute F0 and energy contours. As the corpus is relatively small, we have focused on global features, computed on the whole sentence, rather than local features (sub- and segmental level). We compute the most commonly features:

- F0 and energy features: mean, median, standard deviation, minimum, maximum, range.
- Duration: speech rate, absolute duration.

To evaluate the vocal characteristics, features as jitter and shimmer are commonly taken into account as features of an ASR system. Jitter (respect. Shimmer) measures the perturbation of the length (respect. amplitude) of two consecutive pitch periods. For synthesis, it is more difficult to take into account these features and to make a link between these acoustic features and the visual features. But we also compute these features to evaluate these invisible characteristics for our speaker.

These features are computed on each sentence and mean value is used for each emotion, excepted for the absolute duration which is the summation of all sentences. Other features have been computed but not used in the analysis, as number and length of pauses, due to the fact that those sentences are short and the number of pauses is too weak to obtain reliable values.

4. Results

4.1. Visual Results

Figure 2 and 3 summarizes the main finding of the analysis. We present the first 2 principal components of the facial data and their percentage of variance (Table 1 presents the values for the first 5 principal components). The deformation of the face is presented when the corresponding component has a value of -3 (blue) or +3 (red) standard deviations (we assume they are the lower and upper bound of the component variation).

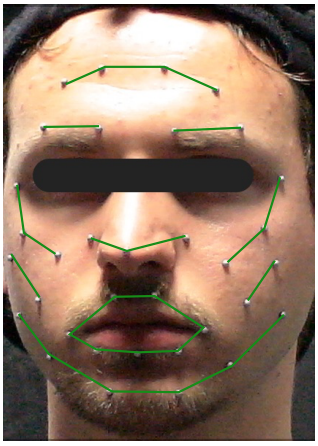

Figure 1: *The positions of the reflective markers on the face of the actor.*

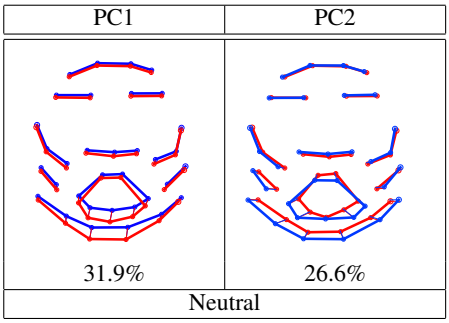

Figure 2: *The 2 first principal components of the facial data and their percentage of variance, for neutral expression. Each pair of colors shows the deformation of the face when the corresponding component assumes a value of -3 (blue) or +3 (red) standard deviations.*

The neutral case shows clearly that the main articulatory movements were performed by the lower part of the face (lips and jaw). The upper part of the face, often related to emotions, barely moves or does not move at all. For the joy case, we notice that the variation of the face expression is extreme for the whole face. The largest movement observed are those of eyebrows and forehead. The variation of the movement of the face related to speech is also important. It should be noted that the variance of the first component is 51% which means that the main features of the joy movements are represented by this component. The visual variation of the surprise emotion seems to be similar to the joy emotion, but the variation of the eyebrows and the forehead are slightly lower for the surprise case. The movement of the upper face is captured only by the first principal component. Similarly, anger emotion is mainly characterized by the first principal component (40% of the variance) and the facial expressions vary in a similar way as the surprise emotion. For disappointment and sadness cases, variation follows the same general trend observed with the previous emotion, but in more moderate way. In addition, we can notice a difference with other emotions that is the general form of eyebrows. This important difference is captured by the first 2 principal components and also slightly visible on the movement of the nasal region of the

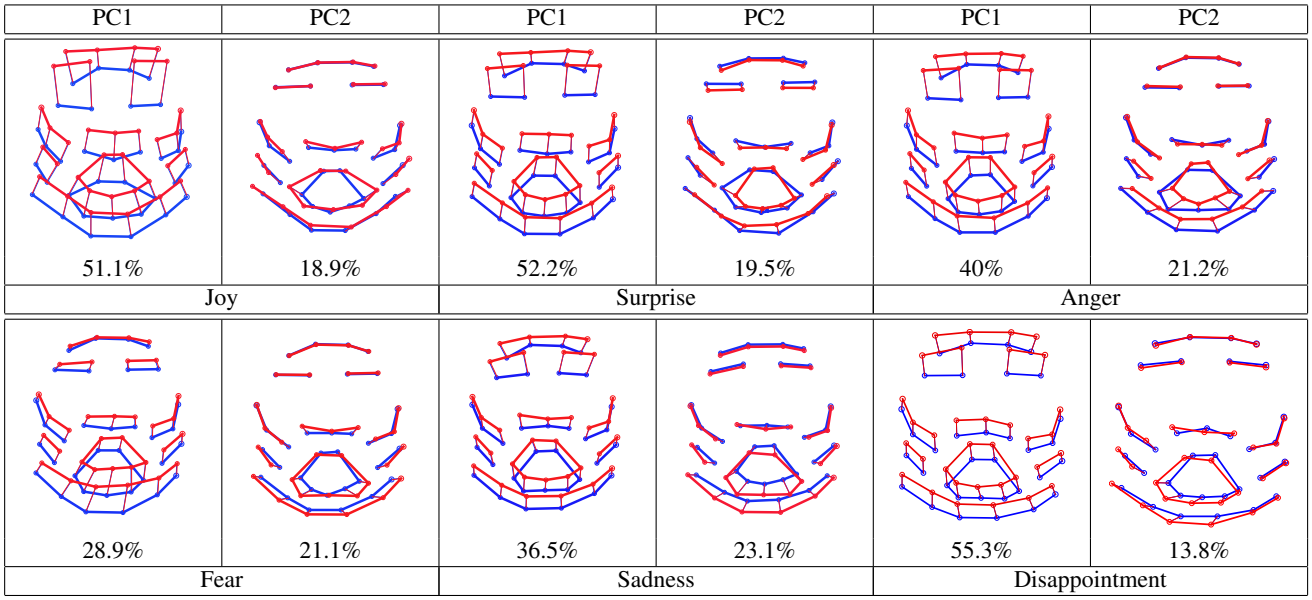

Figure 3: *The 2 first principal components of the facial data and their percentage of variance. Each pair of colors shows the deformation of the face when the corresponding component assumes a value of -3 (blue) or +3 (red) standard deviations.*

face. Finally, the fear emotion variation is the lowest (close to neutral case) as shown in Figure 4. The other noticeable difference is that the variance percentage of the first component of this emotion is lower than that of the other emotions (28.9% vs. more than 40%) and so the first 2 components just explain 50% of the variance. The Table 1 confirms this trend with other components. We can notice the existence of the eyebrows movement, but there is no movement of forehead. Figure 4 shows the eyebrow range variation for each emotion (the difference between the eyebrows value at +3 standard deviation and at -3 standard deviation for the first component, for each emotion). It is clear that the eyebrows movement seems to be the most prominent movement to express the different emotions, but that was not the case for fear emotion, and neutral emotion.

Table 1: *Percentages of variance for the first five principal components for the 7 emotions. The number in parenthesis is the cumulative percentage of variance.*

Emotion	PC1	PC2	PC3	PC4	PC5
Neutral	32 (*32*)	26 (*58*)	16 (*74*)	10 (*84*)	4 (*89*)
Joy	51 (*51*)	19 (*70*)	12 (*82*)	7 (*89*)	3 (*92*)
Surprise	52 (*52*)	19 (*71*)	9 (*80*)	5 (*85*)	3 (*89*)
Anger	40 (*40*)	21 (*62*)	10 (*73*)	9 (*82*)	4 (*87*)
Fear	28 (*28*)	21 (*49*)	17 (*67*)	10 (*77*)	5 (*82*)
Sadness	36 (*36*)	23 (*59*)	15 (*75*)	6 (*82*)	5 (*87*)
Disap.	55 (*55*)	13 (*69*)	9 (*78*)	6 (*84*)	4 (*88*)

4.2. Acoustic Results

It is difficult to compare results from different languages. Here, the term *commonly* refers to the results obtained in the literature for similar languages as French, English and German, and in opposition to Chinese and Japanese.

Figure 5 shows the results for minimum, maximum, mean and range of F0, for each emotion. These values are the means

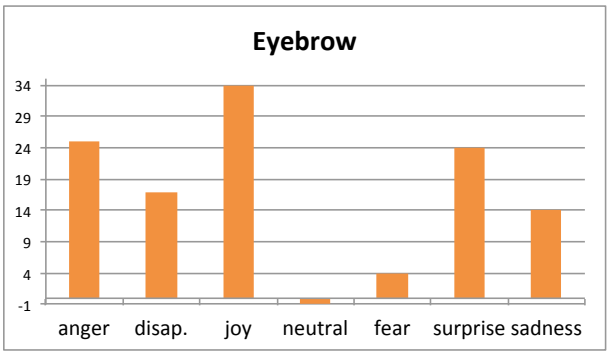

Figure 4: *Range of the eyebrow variation (in mm) for each emotion: Range is the position difference between +3 standard deviation and-3 standard deviation for the first component, for each emotion.*

on the 10 sentences. In the following, *minimum* (resp. *maximum, mean*) is used instead of *mean of minimums* (resp. mean of maximums, mean of means). They are computed in Hz without normalization, as the corpus is uttered by one speaker. The error bars represent the confident interval (95%) of the values of the mean. The value between brackets is the mean of the range value (maximum - minimum). For the mean and maximum value of F0, we can see that anger, joy, fear and surprise are higher than neutral, as commonly expected [9]. But sadness has a higher mean value and even a relatively high maximum value (similar to anger and fear) compared to neutral. The difference with common results (for instance in [10]) could be explained by the fact that corpus is acted speech and depending on the affective degree desired by the speaker during his acting (sadness-grief, sadness-depression, sadness-regret). Disappointment, in particular for the range value, is lower than the neutral emotion. Joy and surprise cases are particularly highlighted by the maximum and range values. Anger and fear are

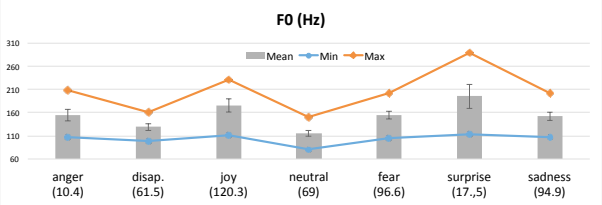

Figure 5: *F0 (in Hz) statistics, for each emotion: minimum and maximum in solid lines, mean with bars (and confident interval error bars). The number in parenthesis is the range of F0.*

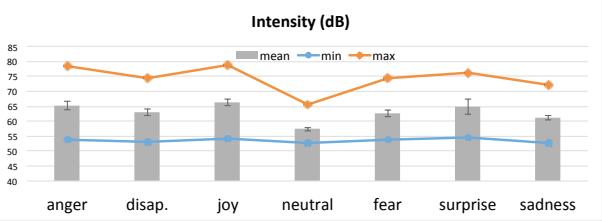

Figure 6: *Intensity (in dB) statistics for each emotion: minimum and maximum in solid lines, mean with bars (and confident interval error bars).*

slightly lower than joy case, but higher than the neutral case. It seems that the minimum is not a discriminative feature. This could enhance the argument that the speaker seems to give a minimum of activation (or arousal) even for disappointment and sadness.

Figure 6 shows the statistics of the intensity (in dB): mean (with error bars/confident intervals), minimum and maximum. Joy and anger are much higher than the other cases, just before fear and surprise (as in [10] and [12]). Sadness is the lowest (except neutral case). The disappointment case is similar to fear and surprise. In all cases, the means are higher than neutral. This illustrates that during acted speech, the conveyed emotion is expressed by the actor more intense than in natural and spontaneous case.

Figure 7 show the speech rate (phones per second). Anger, joy, fear and surprise have similar values and are very much faster than the neutral case (as proposed in explicit prosody rules for synthesis systems referenced in [12]). Sadness and disappointment are close to neutral cases (with wide confident intervals) . The absolute duration (summation on all sentences) is fully correlated to the speech rate, because pauses are very rare (and very short) for this corpus (excepted pauses before and after the sentence which do not take into account for statistics).

For shimmer and jitter (local amplitude and length pitch perturbation), any emotion does not seem to be discriminated (the anger is barely the highest). The confident interval is very wide. These kinds of features are very sensitive to the pitch period algorithm and the corpus is too small to obtain reliable values.

To sum up, we obtain coherent acoustic results. Anger, surprise, fear, and surprise results are similar to those of literature. Sadness and disappointment are in general higher than neutral. This can be explained by the nature of the task: acted speech can convey a stronger degree of affect (or activation) than spontaneous speech/emotion.

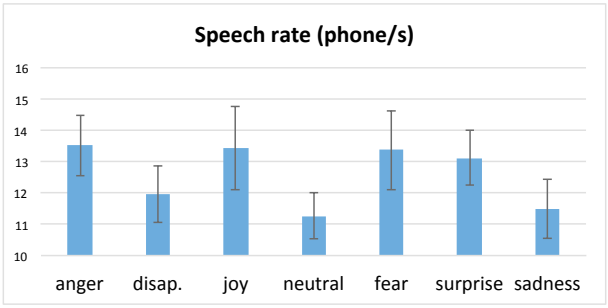

Figure 7: *Speech rate feature (in phone/second) for each emotions.*

5. Discussion

In this study, we presented a characterization of expressive acted speech. This approach is different from the classical analyses where expressive speech is either spontaneous or acted by normal human speakers, and is mainly addressed in the field of speech recognition or perception. The acted speech by a professional actor is a choice that can be justified in the field of audiovisual speech synthesis, as the purpose is to make sure to convey perfectly the right emotion, even with some exaggeration. In both acoustic and visual domains, joy, surprise, anger and sadness present the most noticeable features. One feature can discriminate some emotions, but it seems that there is no universal feature for all the emotions. For instance, we have noticed that anger, joy, fear and surprise have similar speech rates. For F0, anger and fear are similar but slightly different than joy and surprise. The facial movements are more important for joy, surprise and anger. It is probably reasonable to consider combining several features, in acoustic and visual domains to enhance the discrimination. For unit-selection audiovisual synthesis, this may suggest that it is possible to consider units in the audio-visual domain. For instance, in HMM-based synthesis, the acoustic-visual features can be considered as one vector. This combination at every step of the synthesis process (training or selection, and generation) may lead to believe that the emotion will be better perceived (even though, it is not necessarily the case when each modality is considered independently).

It should be noted that, in this study and for sake of simplification, we did not consider head movement. This feature will be considered in future work. We expect that head movement can help to better characterize a given emotion. As additional analysis, we will consider investigating the timing of each visual movement in comparison with an acoustic event. The work presented in this paper was a case study, where one actor has been considered. We are planning to consider 4 actors to try to extract common feature patterns during acted speech.

6. Acknowledgements

This work was supported by Inria (ADT Plavis) and Region Lorraine (Corexp).

7. References

[1] 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.

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

[3] 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.

[4] 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.

[5] 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.

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

[7] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor, "Emotion recognition in human-computer interaction," *Signal Processing Magazine, IEEE*, vol. 18, no. 1, pp. 32–80, 2001.

[8] B. Schuller, G. Rigoll, and M. Lang, "Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture," in *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*, vol. 1. IEEE, 2004, pp. I–577.

[9] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recogn.*, vol. 44, no. 3, pp. 572–587, Mar. 2011.

[10] I. R. Murray and J. L. Arnott, "Implementation and testing of a system for producing emotion-by-rule in synthetic speech," *Speech Communication*, vol. 16, no. 4, pp. 369 – 390, 1995.

[11] 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.

[12] M. Schröder, *Affective Information Processing*. London: Springer London, 2009, ch. Expressive Speech Synthesis: Past, Present, and Possible Futures, pp. 111–126.

[13] K. R. Scherer, "A cross-cultural investigation of emotion inferences from voice and speech: implications for speech technology." in *INTERSPEECH*, 2000, pp. 379–382.

[14] I. R. Murray and J. L. Arnott, "Applying an analysis of acted vocal emotions to improve the simulation of synthetic speech," *Computer Speech & Language*, vol. 22, no. 2, pp. 107–129, 2008.