An Italian Database of Emotional Speech and Facial Expressions

N. Mana[*], P. Cosi[†], G. Tisato[†], F. Cavicchio[†], E. Caldognetto Magno[†] , F. Pianesi[*]

[*]ITC-irst
Center for Scientific and Technological Research
via Sommarive, 18 – 38050 Povo (Trento), Italy
{mana, pianesi}@itc.it

[†]ISTC-CNR
Institute of Cognitive Sciences and Technology
via Anghinoni, 10 – 35121 Padova, Italy
{cosi, tisato, cavicchio, magno}@pd.istc.cnr.it

ABSTRACT

This paper presents an Italian database of acted emotional speech and facial expressions. New data regarding the transition between emotional states has been collected. Although acted expressions have intrinsic limitations related to their naturalness, this method can be convenient for speech and faces synthesis and within evaluation frameworks. Using motion capture is a good method to get precise information on data for playing back them on facial model and also to build specific animation engine. The procedure to adapt the recorded data to a MPEG-4 compliant facial animation model will be described.

Author Keywords

Emotional speech and facial animation, 3D motion capture.

ACM Classification Keywords

H5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

INTRODUCTION

During the last years there has been a growing interest in systems working on emotions. Many works addressed recognition and synthesis of emotions through vocal and facial signs. In this perspective, emotion corpora become fundamental to perform conceptual analyses, to develop emotion recognition and synthesis systems (especially in case of data-driven systems) both for speech and face, and to test emotion-oriented tools and applications [1].

In acquiring basic data, four main types of source are commonly used: a) spontaneous emotions; b) inducted emotions; c) acted emotions; and, d) application-driven emotions [2].

In this paper we present an audio/video database of acted emotional speech and facial expressions.

DATA COLLECTION

For the data collection a professional actor (male, 25 years old) was employed. He was instructed to utter short non-sense words with various emotional expressions and intensities.

Collected Data

The first part of the database includes "Isolated Emotions", i.e. a set of Italian non-sense words, acted with different emotions. These words, representing Vocal – Consonant – Vocal (VCV) sequences (specifically /aba/, /ada/, /aLA/, /adZa/, /ala/, /ana/, /ava/), cover the seven basic viseme[1] classes for Italian [3].

Each VCV sequence was acted with six emotional states, corresponding to the Ekman's set [4] – Anger, Disgust, Fear, Happiness, Sadness, and Surprise – plus the additional 'Neutral' state. Each emotion was acted with three different intensity levels (Low, Medium, High). Examples of the six emotional and neutral states are shown in Figure 1.

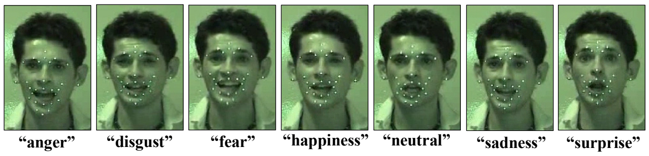


Figure 1. Examples of emotional facial expressions during speech.

The second part of the database includes "Combined Emotions", i.e. VCV-VCV sequences. In this case, in order to get significant examples of transitions from an emotional state to another during speech, the non-sense words were acted in pairs, each one with a different emotional state (Neutral, Anger, Happiness, Surprise), at medium intensity.

Finally, the third part includes examples of a long sentence with a good coverage of Italian phonemes ("Il fabbro lavora con forza usando il martello e la tenaglia"; lit. "the smith works with strength using the hammer and the tongs"), acted with the neutral and the six emotional states and three intensities.

[1] A viseme is a "visual phoneme", i.e. the visual equivalent of a phoneme (unit of sound) in spoken language. Phonemes can be clustered according to their visual similarity. So, e.g. /p/ and /b/ are phonetically different but they belong to the same viseme class given that, from a visual viewpoint, cannot be distinguished.

Globally 1,657 examples of emotional expressions during the speech have been collected.

Recording: Technical set-up and Procedure

The dynamics of every facial expression during the speech, as well as the respective articulatory and acoustic data, was captured with high precision by means of ELITE [5]. ELITE is an automatic opto-tracking movement analyser for 3D kinematics data acquisition, which also allows synchronous recording of the acoustic signal.

The system records the acoustic signal and tracks the infrared light reflected by 28 small (only 2 mm diameter) passive markers, glued on the actor's face in different positions. For each marker, displacement and transition speed from a position to another, are recorded every 1/10th on a second by two cameras. In this way the dynamics of every facial expression is captured with high accuracy (100 Hz sampling rate, maximal error of 0.1 mm for a 28x28x28 cm cube).

A speaker announced to the actor the short word to be played, followed by the corresponding emotional state and intensity, according to the following scheme:

< short word><emotional state><intensity>

For example: *Ava, Disgust, High*.

To ensure an easier detection of the starting and ending point of the emotional expression, the actor uttered two additional Italian words, "chiudo" (lit. I close) and "punto" (point), respectively before and after the short word played emotionally. For example: "*chiudo*" $[ava]_{\text{Disgust, High}}$ "*punto*"

The order of the sequences to act was generated randomly, so that the actor could not be conditioned. Only the intensity sequence (Low, Medium and High) was kept fixed so that the actor could use the first intensity as reference and generate the second and the third expression increasing the intensity progressively

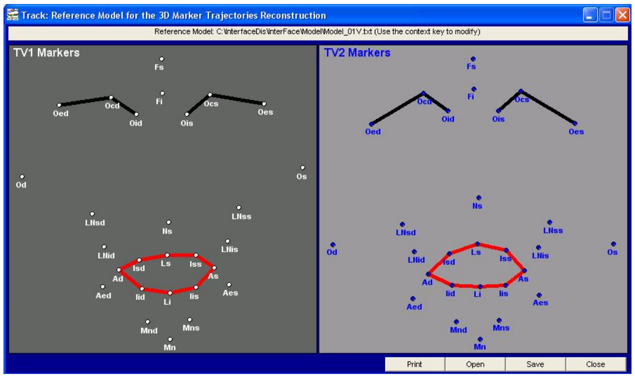
DATA PROCESSING

Given the recordings of the marker movements for each emotional expression, two kinds of data processing were accomplished: a) a 3D reconstruction of the marker trajectories; and b) a conversion of these trajectories into Facial Animation Parameters.

3D reconstruction

Starting from the recordings captured by the two cameras of the ELITE system, it was necessary to build the 3D trajectories corresponding to the articulatory movements of the 28 markers on the face. For this task a MatLab© software (called "Track" [6]), specifically developed by ISTC-CNR, has been used.

Track allows to construct the 3D trajectories of markers, frame by frame, starting from the 2D recordings. For doing that, it needs a "reference model", i.e. a file in which a name is associated to each marker, so that an unambiguous linking between marker and its position is created for each camera (TV1 and TV2), as showed in Figure 2.



Track: Reference Model for the 3D Marker Trajectories Reconstruction

Reference Model: C:\InterfaceDis\InterFace\Model\Model_01V.txt (Use the context key to modify)

Figure 2. Definition of the "Reference Model" for the 3D reconstruction in the Track system

The system output is a set of values corresponding to the x,y,z coordinates for each marker, frame by frame.

The marker identification and the reference space deformation problem have been exceeded with an algorithm based on the Singular Value Decomposition (SVD) which has the intrinsic advantage to operate an error minimization while calculating the roto-translation, even independently from using a perfect undeformable reference space. The produced FAP-stream takes into account the roto-translation and the scale factors of the head that has to be animated, thus allowing a correct data-driven synthesis of whichever MPEG-4 compatible agent.

Conversion to FAPs

Given the marker trajectories in the 3D space (x,y,z coordinates), these were converted into FAPs (Facial Animation Parameters), according to MPEG4 standard[2]. FAPs are based on the study of minimal perceptible actions and are closely related to muscle actions involved in a facial expression. MPEG4 animation uses a pseudo-muscle approach, in which the muscle contractions are obtained through the deformation of the polygonal network around feature points. Each feature point corresponds to skin muscle attachment. The deformation is performed in a zone of influence that has an ellipsoid shape whose centroid is a feature point. The displacement of points within this area of influence obeys to a deformation function that is function of the distance between the points and the feature point. MPEG-4 defines 84 Feature Points (FPs) that describe the shape of a standard face and should be defined for every face model. These points are used for defining animation

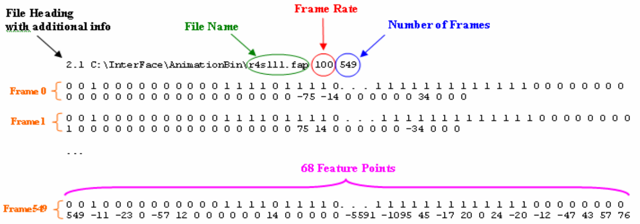
[2] MPEG4 is an ISO/IEC standard developed by MPEG (Moving Picture Experts Group) [7], widely used in 3D animation. More details may be found at MPEG4 web page: http://www.chiariglione.org/mpeg/standards/mpeg-4/mpeg-4.htm

parameters as well as calibrating the models when exchanged between different players. Two sets of parameters describe and animate the 3D facial model: facial animation parameter set (FAPs) and facial definition parameters (FDPs). FAPs define the facial actions, while FDPs define the shape of the model.

FAPs have to be calibrated prior to use them on a specific face model. For this reason, FAPs are expressed in normalized units called FAPUs (Facial Animation Parameter Units) which are defined as fractions of distances between key facial features (e.g. eye-nose separation). Only FAPUs are specific to the actual 3D face model that is used, while FAPs are independent. That means they can drive different face models, regardless of geometry. As a result, by coding a face model using FPs and FAPUs, developers can freely exchange face models without the problem of calibration and parameterization for animation, and FAPs can be used for different 3D facial models.

When the model has been characterized with its FDPs (namely the model shape has been defined), the animation is obtained by specifying the FAP-stream, i.e. the values of FAPs frame by frame.

The 68 FAPs values, specified in the FAP-stream frame by frame, cause the facial animation by defining the deformation between two frames of animation. Of these 68 values, the first 2 are high level parameters representing visemes and emotions. The remaining 66 are low level FAPs, dealing with specific regions on the face (e.g. bottom of chin, left corner lip, right corner of left eyebrow, etc.). Most of FAPs correspond to an FP, and define translation or rotation on that FP along an axis in three dimensions, while some of the FAPs represent rotation of the head and eyes.



```
2.1 C:\InterFace\AnimationBin\r4s111.fap 100 549
0 0 1 0 0 0 0 0 0 0 0 0 0 1 1 1 1 0 1 1 1 1 0. . . 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 -75 -14 0 0 0 0 0 0 34 0 0 0
0 0 1 0 0 0 0 0 0 0 0 0 0 1 1 1 1 0 1 1 1 1 0 . . . 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0
1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 75 14 0 0 0 0 0 0 -34 0 0 0
...
0 0 1 0 0 0 0 0 0 0 0 0 0 1 1 1 1 0 1 1 1 1 0. . . 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0
549 -11 -23 0 -57 12 0 0 0 0 0 0 14 0 0 0 0 0 -5591 -1095 45 -17 20 0 24 -20 -12 -47 43 57 76
```

Figure 3. A FAP stream

As shown in Figure 3, in a FAP-stream the relevant information for animation are distributed on two lines: the first line indicates which point is specifically activated in that moment (activation or not is expressed by 0 and 1), while the second one contains the target values, in terms of differences from the previous frame target values. When a FAP is activated (i.e. when its value is not null), the feature point on which the FAP acts is moved in the direction indicated by the FAP itself (up, down, left, right, etc).

By converting the 3D marker trajectories into FAP streams, it is possible to animate the emotional facial expressions played by the actor on any MPEG4-compliant synthetic face, as depicted in Figure 4.

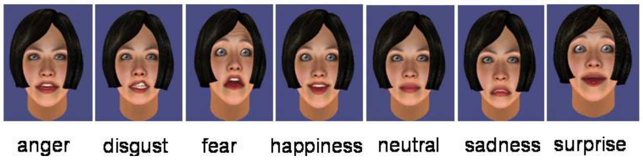


Figure 4. Basic emotions on a synthetic face

DATA ANALYSIS

Emotional Facial Expressions

In order to represent graphically the dynamics of a facial expression, i.e. the marker trajectories frame by frame, we converted the x,y,z coordinate values for each marker in a unique value. For this purpose we used the 3D vector module:

$$|V| = \sqrt{\left(x^2 + y^2 + z^2\right)}$$

where x,y,z represent the distance on the three Cartesian axis of a marker with respect to the axis origin.

We can see, for example, the dynamics[3] of "Neutral" in Figure 5 and that one of "Surprise" acted with medium intensity by the actor when uttering "aba" in Figure 6.

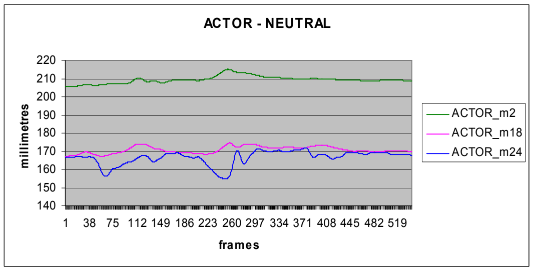


Figure 5. Dynamics of a "Neutral" facial expression

By using this graphical representation, the "prototypical behaviour" of each emotional expression becomes evident. Note, for example, in Figure that the m2 curve presents a pick in correspondence to the maximum value of the surprise expression, as typically it happens[4].

[3] Note that for the sake of graph readability, the dynamics of not all of the 28 markers is showed. The analysis is focused only on marker 2 (m2), marker 18 (m18) and marker 24 (m24), corresponding respectively to "Left Central Eye", "Left Lip Corner" and "Middle Lower Lip" because in some way they are the most significant in capturing facial expression changes.

[4] As well known, one of the features characterizing the facial expression of "surprise" is "risen eyebrows".

By graphs it is also possible to analyze how emotions affect the speech production. This is particular evident by comparing the dynamics of emotional expressions for a specific viseme with the dynamics of the same viseme uttered without emotion (neutral state).

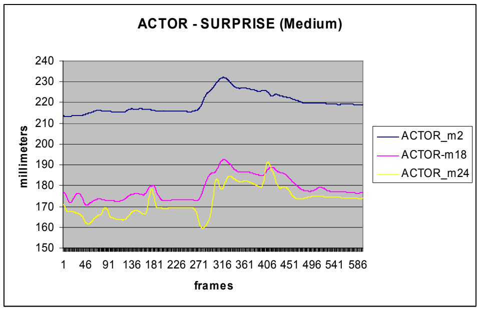


Figure 6. Dynamics of a "Surprise" facial expression

Figure 7 shows, for example, the dynamics of marker 2 for "aba" viseme in neutral and surprise condition[5]. As it is evident by the graph, the curves are strongly affected by the emotion (see the central peak). Furthermore, we can see that the emotional expression tends to be longer than the neutral one.

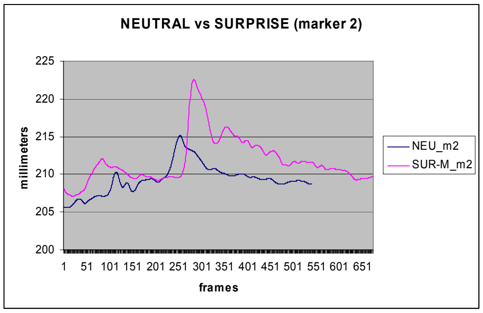


Figure 7. Dynamics of marker-2: neutral vs surprise

Emotional Audio/Visual Speech

All these data have been used for building LUCIA [8], an emotional audio/visual talking head that uses 3D polygon models, which are parametrically articulated and deformed by a data/driven-based animation engine [2], and speaks with the Italian version of the FESTIVAL diphone TTS synthesizer [9], appropriately modified with emotive and expressive capabilities.

CONCLUSIONS

In this paper we have presented an audio/video database of acted emotional speech and facial expressions. Although acted expressions have intrinsic limitations related to their naturalness with respect to spontaneous ones, they can be convenient and suited for specific tasks, such as speech and faces synthesis and within evaluation frameworks.

ACKNOWLEDGMENTS

The data collection was carried out within the project PF-STAR (**P**reparing **F**uture multi**S**ensorial in**T**er**A**ction **R**esearch), supported by the EU grant IST-2001-37599 (for more details see http://pfstar.itc.it). The authors would like to thank all the people involved in the data collection, and in particular the actor, Fabio Fusco, for his patient cooperation.

REFERENCES

1. Cowie R., Douglas-Cowie E. and Cox C., Beyond emotion archetypes: Databases for emotion modelling using neural networks. In *Neural Networks* - Special Issue Emotion and Brain: Understanding Emotions and Modelling their Recognition, 18 (4), (2005), 371-388.
2. Cosi P., Fusaro A., Grigoletto D., and Tisato G., Data-Driven Tools for Designing Talking Heads Exploiting Emotional Attitudes. In *Proceedings of Tutorial and Research Workshop "Affective Dialogue Systems"*, Germany (2004).
3. Magno Caldognetto E., Zmarich C., Cosi P. and Ferrero F., Italian Consonantal Visemes: Relationships Between Spatial/temporal Articulatory Characteristics and Coproduced Acoustic Signal. In *Proceedings of AVSP-97*, Tutorial & Research Workshop on Audio-Visual Speech Processing: Computational & Cognitive Science Approaches, Rhodes, Greece (1997).
4. Ekman P. , An Argument for Basic Emotions. In N.L. Stein, and K. Oatley, editors, *Basic Emotions,* (1992), 169-200.
5. Ferrigno G., and Pedotti A.m ELITE: A Digital Dedicated Hardware System for Movement Analysis via Real-Time TV Signal Processing. In *IEEE Transactions on Biomedical Engineering, BME-32*, (1985), 943-950.
6. Tisato G., Cosi P., Drioli C., Tesser F., INTERFACE: a New Tool for Building Emotive/Expressive Talking Heads. In CD Proceedings INTERSPEECH 2005, Lisbon, Portugal, 2005, pp. 781-784.
7. Doenges P., Lavagetto F., Ostermann J., Pandzic I.S., and Petajan E., MPEG-4: Audio/Video and Synthetic Graphics/Audio for Mixed Media. In *Image Communications Journal*, 5(4), (1997).
8. Cosi P., Fusaro A., Tisato G., LUCIA a New Italian Talking-Head Based on a Modified Cohen-Massaro's Labial Coarticulation Model. In *Proceedings of Eurospeech 2003*, Geneva, Switzerland, September 1–4, 2003, Vol. III, pp. 2269-2272.
9. Tesser F., Cosi P., Drioli C., Tisato G., Emotional Festival-Mbrola TTS Synthesis. In CD *Proceedings INTERSPEECH 2005*, Lisbon, Portugal, 2005, pp. 505-508.

[5] More examples of several emotional expressions will be showed during the presentation.