# LW2A: an easy tool to transform voice wav files into talking animations

*Piero Cosi, Graziano Tisato*

ISTC-CNR
Istituto di Scienza e Tecnologie della Cognizione
Consiglio Nazionale delle Ricerche
Padova, ITALY
[piero.cosi, graziano.tisato]@pd.istc.cnr.it

## Abstract

LW2A (LuciaWav2Animation) is an easy tool to transform voice wav files into talking animations. LUCIA, an Italian animated talking face is provided with an automatic segmentation tool based on an Italian clean speech ASR system for creating realistic and effective lip-sync facial animations from audio files.
**Index Terms**: visual speech synthesis, speech-driven facial animation, audio-to-visual mapping

## 1 Introduction

In the development of human-machine interactive interfaces the use of synchronized talking animation is a natural and efficient way of communicating and it is quite usual to see virtual characters in many aspects of our everyday life. In order to synchronize the characters lip animation with its speech, the phonetic content, that is the phonemes and their duration, is needed. The sound can be synthesized from written text and in this case this information is generated by the speech synthesizer. However, the character has a much more natural appearance if its voice comes from a real person and in this case the phonetic information has to be obtained by performing audio analysis and lip-synching audio and visual animation. Most of the research done in this field concerns English, due to the lack of both resources and technology in minority languages, but the use of open source speech technologies for ASR and TTS should help reduce this gap.

Even if the term Lip-sync or Lip-synch (short for lip synchronization) can refer, according to "Wikipedia", to a technique often used for performances in the production of film, video and television programs, or the science of synchronization of visual and audio signals during post-production and transmission, or the common practice of people, including singers performing with recorded audio as a source of entertainment, within the present animation framework, with Lip-synch we refer to the technical term used to describe matching lip movements of animated characters to a prerecorded human speaking or singing voice.

Automating the lip-synch process, generally termed visual speech synthesis, has potential for use in a wide range of applications: from desktop agents on personal computers to language translation tools to provide a means for generating and displaying stimuli in speech perception experiments and it is often used in the production of films, cartoons, television programs, and computer games.

Generally speaking, when we consider talking heads, facial animation can be synchronized to speech in several ways. The employed method depends mainly on the kind of speech data which is available for synchronization, e.g. whether the audio signal, the phonemes and timing of an utterance or only a text representation are available. The text-driven approach (e.g. [1, 2]) receives a text as input, transcribes it into its phonemic representation and this information is then used to generate both synthetic audible speech and synchronized visible speech. The speech-driven method (e.g. [3]) considers prerecorded speech as input. The audio file is processed with an ASR to extract phonemes and timing information. These data are used to create the facial animation which is performed synchronously to the audio file playback. If both text and speech are available, a text-and-speech-driven hybrid approach (e.g. [4]) can be applied. The text and its phonemic representation are used for segmentation, that is the identification of segment boundaries and to compute timing information for the animation component, possibly aided by some rules for phoneme durations. In all of the above approaches the lip movements are determined using predefined goal lip shapes for the phonemes. These are obtained by measurements or by simple observation.

In this work we present LW2A (Lucia-Wav2Animation) a text-and-speech-driven hybrid software tool able to automatically compute the facial movements of LUCIA [5], an expressive animated talking face, given pre-recorded audio signals (wav files) and their corresponding orthographic transcriptions. LW2A is a specific module of a more complex tool called INTERFACE [6], designed and implemented in Matlab, for simplifying and automating many of the operation needed for building emotive/expressive talking heads from motion-captured data.

## 2 LUCIA: an Italian animated talking face

The system presented here is based on LUCIA a visual speech synthesizer that includes a control model to compute articulatory trajectories from a textual input, a simple shape model to animate the facial geometry from computed trajectories, and a quite simple graphic appearance model for rendering the animation by varying the colors of pixels. LUCIA is the culmination of years of experience working with lip-synchronization, 3D graphics, and facial animation. LUCIA's control model exploit a trajectory formation system based on a modified version of the Cohen-Massaro coarticulation model [7, 8] and utilizes a phonetic segmentation of the acoustic signal. LUCIA's shape model is an ad hoc parametric deformations of a 2D mesh while her appearance model is a quite simple texture rendering technique.

LUCIA, as illustrated in the block diagram shown in Figure 1, is an Italian talking head based on the mpeg-4 [9] standard,

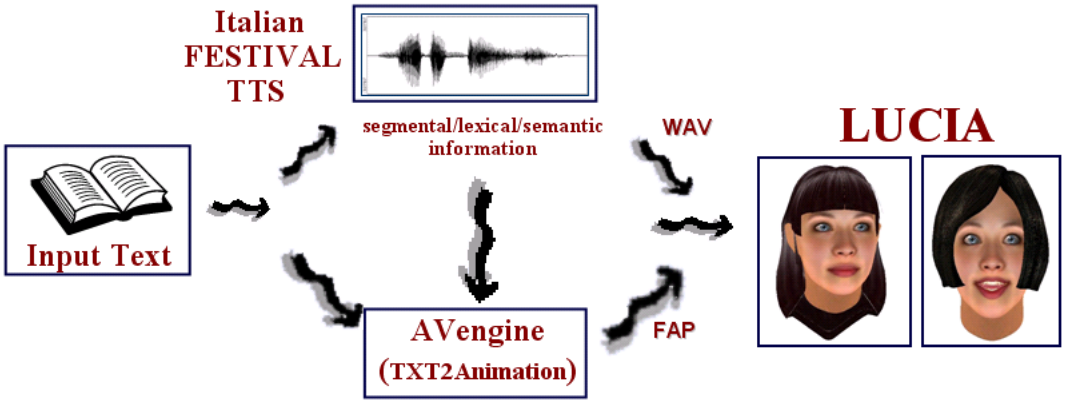


Figure 1: Lucia, an Italian talking head.

speaking with the Italian version of FESTIVAL TTS [10] and it is a graphic MPEG-4 compatible facial animation engine implementing a decoder compatible with the "Predictable Facial Animation Object Profile" [9].

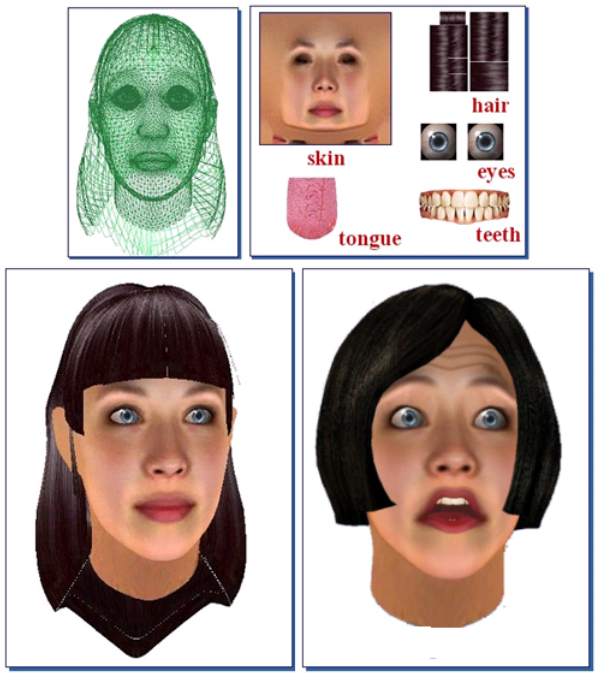


Figure 2: Lucia's wireframe and textures.

LUCIA is able to generate a 3D mesh polygonal model by directly importing its structure from a VRML file [11] and to build its animation in real time. At the current stage of development, as illustrated in Figure 2, LUCIA is a textured young female 3D face model built with 25423 polygons: 14116 belong to the skin, 4616 to the hair, 2688x2 to the eyes, 236 to the tongue and 1029 to the teeth respectively.

## 3 LW2A: Lucia Wav to Animation

At the core of LW2A technology is the ability to generate realistic animation data from an audio file; thus the first stage in the system is to process and segment input audio files into their phonemes. Two strategies can be followed to obtain this goal: the first exploits an ASR software, such as the CSLU Toolkit [12] or the SONIC [13] ASR system, in our case, to extract the phoneme sequence and boundaries, the second one uses also a text containing the orthographically transcribed text of the audio file to automatically determines where the corresponding phonemes occur in the sound file using ASR force alignment techniques. In particular, in this work, an automatic segmentation tool based on an Italian clean speech HMM ASR system [14] trained on the APSCI (FBK, Italy) corpus, was utilized. [15].

### 3.1 Audio Analysis

LW2A makes use of a a speaker independent "general purpose" phonetic recognizer for Italian [14] trained by the CSLU Toolkit [12]. The recognizer, based on a frame-based hybrid HMM/ANN architecture trained on context-dependent categories to account for coarticulatory variation, recognizes 38 different phonemes (not including silence or closures), and can distinguish between stressed and unstressed vowels as well as open and closed vowels. The APASCI corpus [15], containing nearly 2500 sentences read by 100 speakers, where the sentences have been designed to maximize the number of phonemes occurring in different contexts, was used for training and testing. A phoneme-level accuracy of 82.90% on the development set and of 80.53% on the test set has been obtained. This level of accuracy is much greater than on a similar English-language corpus (with state-of-the-art performance of slightly better than 70%) and it represents the best performance obtained so far on this corpus.

As illustrated in Figure 3, in the lower branch, the application accepts in input a speech signal (WAV file) and its corresponding orthographic transcription (TXT file), or, in the upper branch, only the speech signal; in the first case the Italian version of FESTIVAL is used to phonetically transcribe the audio file starting from its orthographic TXT file and to map them to their corresponding visemes and the CSLU Toolkit force alignment procedure is used to identify the appropriate phoneme boundaries; in the second case both the CSLU Toolkit [12] or the SONIC [13] ASR system can be utilized.

Curves for speech targets are then automatically created by the LUCIA facial animation engine by taking into account coarticulation rules. Finally, curves for head movement, blinks, and

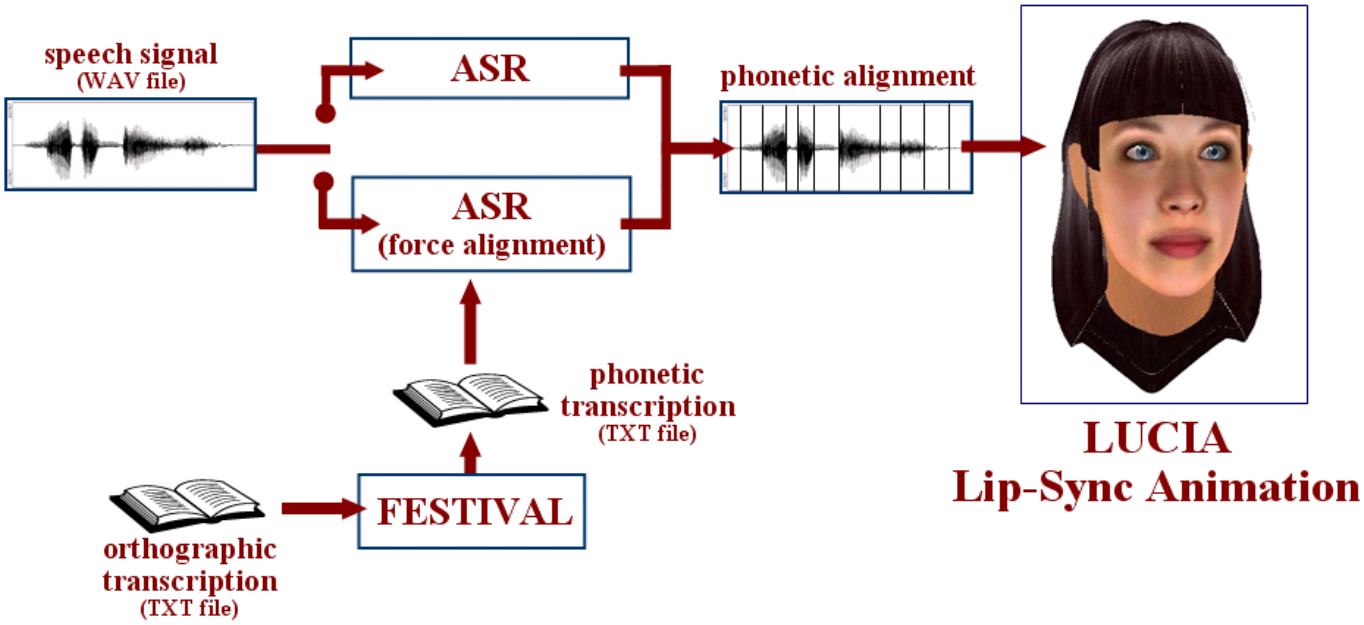


Figure 3: LW2A's block diagram. In the upper branch, the application accepts in input only the speech signal (WAV file) while in the lower branch a speech signal (WAV file) and its corresponding orthographic transcription (TXT file) are both needed.

eyebrow raises etc. are superimposed bu using ad hoc specified rules and LUCIA character is then animated in real time and synchronized with the speakers voice.

At the present time, LW2A can create animation data using Italian language only and the Italian version of FESTIVAL is used to phonetically transcribe the audio file, but English will be added soon. Moreover text files can be optionally post-tagged with HTML-like expressions which can be used to automatically create more expressive visual animation curves at the appropriate time in the audio.

### 3.2 Tweaking Visual Animations

As graphically illustrated in Figure 4, tweaking animations is possible from the curve editor. The curve editor allows users to graphically modify all final software articulatory actuators. New modified curves can be easily saved to a new animation to control expressions or other targets.

### 3.3 Expressive Visual Animations

Expressive facial speech animation is a challenging topic of great interest to the computer graphics community and adding emotions to audio-visual speech animation is very important for realistic facial animation. Expressive TTS is a serious difficult task and even if a lot of studies have been advancing our knowledge its quality is still far from being acceptable for real applications, thus real speech is still used in the production of films, cartoons, television programs, and computer games when emotions and expressions need to be represented.

On the contrary, even if the inclusion of emotions and fluency effects in speech increases the complexity of neutral visual speech synthesis, because of the corresponding shape and timing modifications brought about in speech, and also even if speech is often accompanied by supportive visual prosodic elements such as motion of the head, eyes, and eyebrow, which improves the intelligibility of speech, expressive visual animation is now a little bit easier to be realized.

LUCIA, in fact, can be driven by a special APML (Affective Presentation Markup Language) tagging language and thus is able to show expressive behavior. The extension of such language is intended to support both facial and voice specific controls but with LW2A only the first were utilized. The text can be tagged with various specific semantic functions such as those exemplified in Figure 5 referring to the performative or affective ones and each of them is described by specific facial configurations and actions.

With LW2A, we thus present a technique to animate expressive real human audio and synthesize visual speech incorporating effects of emotion and fluency. The expressive visemes are blended using LUCIA ad-hoc co-articulation technique that can easily accommodate the effects of emotion. Moreover LW2A presents also a visual prosody model that exhibits non-verbal behaviors such as eyebrow motion and overall head motion.

## 4 Conclusions

Because expressive TTS is a serious, difficult and still unsolved problem and its quality is still far from being acceptable for real applications despite to the fact that expressive speech is required in the production of films, cartoons, television programs, and computer games when emotions and expressions need to be represented, animating real speech is quite an interesting field of research and has a lot of potential applications. LW2A is a quite promising starting point to attack this problem considering also the fact that various improvements can be conceived in the future mainly considering the expressive visual animation engine which could be highly improved by the use of new sophisticated graphic software.

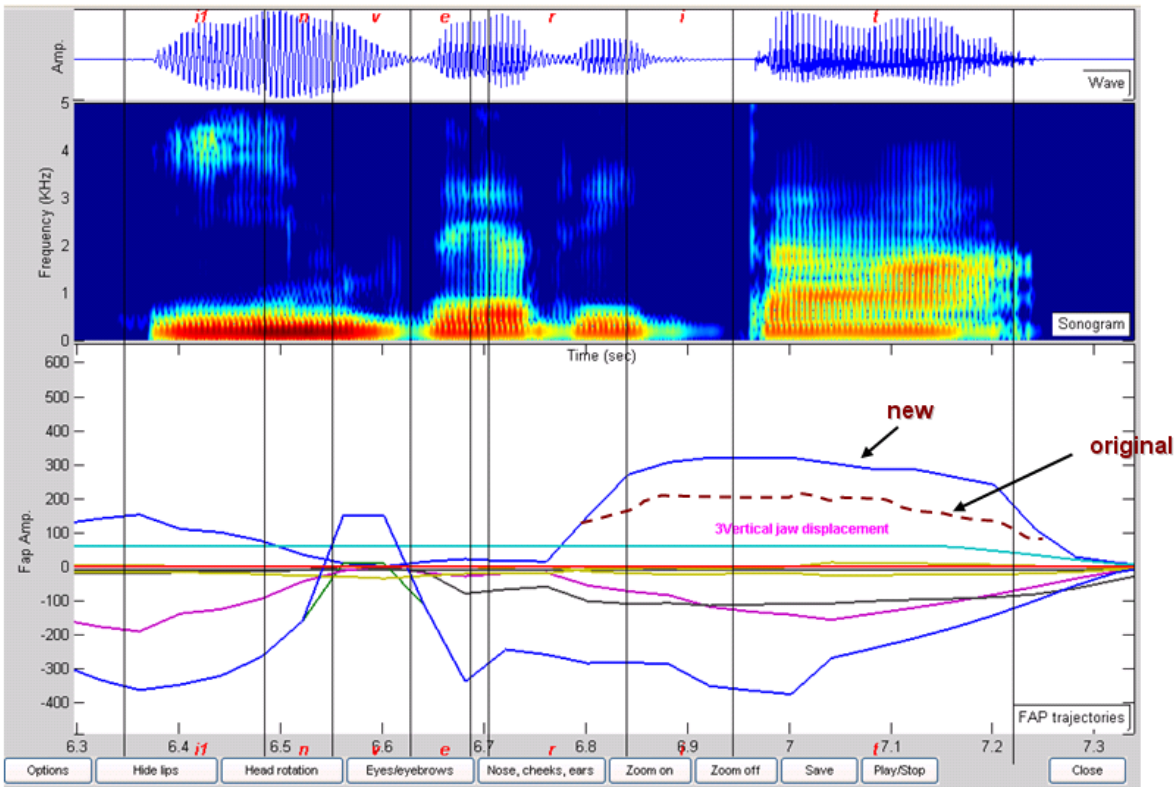


Figure 4: graphic illustration example of tweaking visual animation parameters.

## References


[1] K. Waters and T. Levergood, "Decface: An automatic lip-synchronization algorithm for synthetic faces," in *Tech. Report 93-4*, C. R. Laboratories, Ed., 1993.

[2] E. E. Vatikiotis-Bateson, K. Munhall, M. Hirayama, Y. Kasahara, and H. Yehia, "Physiology-based synthesis of audiovisual speech," in *Proceedings of 4th Speech Production Seminar: Models and Data*, 1996, pp. 241–244.

[3] S. Kshirsagar and N. Magnenat-Thalmann, "Lip synchronization using linear predicitve analysis," in *Proceedings of IEEE International Conference on Multimedia and Expo, August 2000*, 2000.

[4] S. Morishima and H. Harashima, "Facial animation synthesis for human-machine communication systems," in *Proceedings of 5th International Conference on Human-Computer Interaction*, 1993, pp. 1085–1090.

[5] P. Cosi, A. Fusaro, and G. Tisato, "Lucia a new italian talking-head based on a modified cohen-massaro's labial coarticulation model," in *Proceedings of Eurospeech 2003, Geneva, Switzerland*, 2003, pp. 127–132.

[6] G. Tisato, P. Cosi, C. Drioli, and F. Tesser, "Interface: a new tool for building emotive/expressive talking heads," in *Proceedings of Interspeech 2005, Lisbon, Portugal*, 2005, pp. 781–784.

[7] D. Massaro and M. Cohen, "Modeling coarticulation in synthetic visual speech," pp. 139–156, 1993.

[8] D. Massaro, M. Cohen, J. Beskow, and R. Cole, "Developing and evaluating conversational agents," pp. 287–318, 2000.

[9] R. Koenen, "Wg11 - mpeg-4 overview - (v.21 jeju version)," 2002.

[10] P. Cosi, F. Tesser, R. Gretter, and C. Avesani, "Festival speaks italian!" in *Proceedings of Eurospeech 2001, Aalborg, Denmark, September 3-7 2001*, 2001, pp. 509–512.

[11] J. Hartman and J. Wernecke, *The VRML Handbook*. Addison Wessley, 1996.

[12] S. Sutton, R. Cole, J. de Villiers, J. Schalkwyk, P. Vermeulen, M. Macon, Y. Yan, E. Kaiser, B. Rundle, , and C. M., "Universal speech tools: the cslu toolkit," in *ICSLP 98, Sydney, Australia, November, Vol. 7*, 1998, pp. 3221–3224.

[13] B. Pellom, "Sonic: The university of colorado continuous speech recognizer," in *Technical Report TR-CSLR-2001-01, University of Colorado, USA*, 2001.

[14] P. Cosi and J. Hosom, "High performance general purpose phonetic recognition for italian," in *ICSLP-2000, International Conference on Spoken Language Processing, Beijing, Cina, 16-20 October, 2000, Vol. II*, 2000, pp. 527–530.

[15] A. B., F. Brugnara, D. Falavigna, D. Giuliani, R. Gretter, and M. Omologo, "Automatic segmentation and labeling of english and italian speech databases," in *Proceedings of EUROSPEECH93, Berlin, Germany, 1993, Vol. 1*, 1993, pp. 653–656.

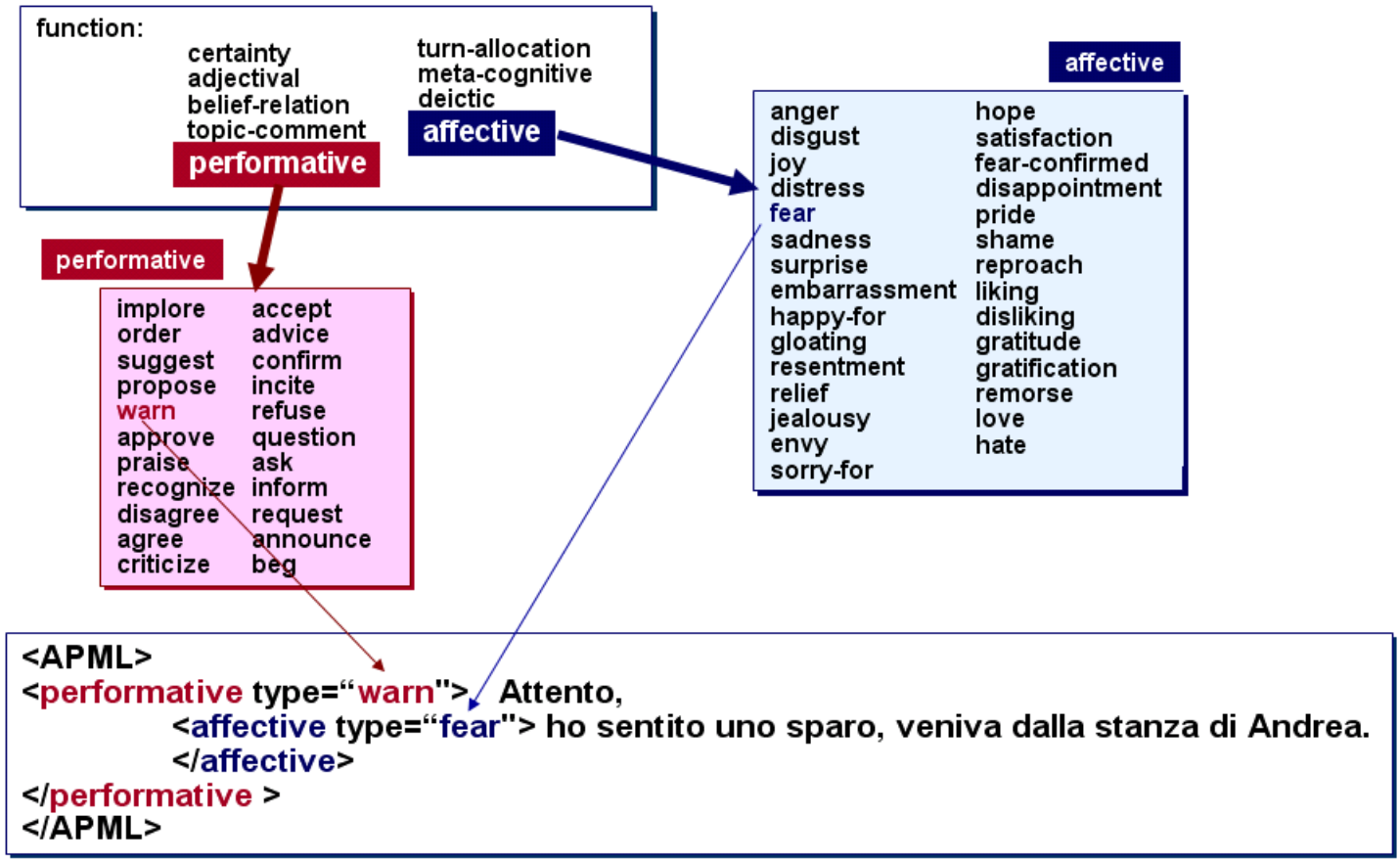


Figure 5: list of some of the available APML-like facial animation TAGs followed by a synthesis text example written with APML syntax.