CHAPTER 10

LUCIA, a New Emotive/Expressive Italian Talking Head

Piero Cosi and Carlo Drioli

Summary

In this chapter we present a study on audiovisual speech analysis and synthesis, for the creation of natural and expressive talking heads. The framework for the modeling of facial and lip movements, and the visual design of an animated MPEG-4 talking face, are first discussed. Then, the design of the voice synthesis modules aimed at producing expressive/emotive speech synthesis is described. The resulting audiovisual synthesis system integrates voice synthesis, facial animation, and a markup language-based interface that allows production of synchronized and emotionally coherent audio and video outputs from tagged text.

Introduction

" . . . Human communication technologies have matured to the point where it is now possible to conceptualize, develop and investigate computer systems that interact with people much like people interact with each other . . . " (OGI CSLU Reading Tutor Project[1]), and a new generation of intelligent and embodied animated virtual agents that engage users in natural face-to-face conversational interaction is not a futuristic flight of the imagination anymore (Cole et al., 2003). An intelligent virtual agent (IVA) is one that mimics the actions of real humans and behaves intelligently in the context of a specific application. A complete IVA

[1] OGI CSLU—Reading Tutor Project: http://cslr.colorado.edu/beginweb/reading/reading.html

must simultaneously interpret the user's auditory and visible speech, eye movements, facial expressions, and gestures to detect for example agreement, distress, trouble, confusion, desire to interrupt, and so on, and must also produce natural and expressive auditory and visible speech with facial expressions and gestures appropriate to the physical nature of language production, the context of the dialogue, and the goals of the task. In other words, an IVA will mimic the actions of real persons and behave intelligently and appropriately in the context of specific task domains.

The research in this field is relevant to a wide number of applications, in which human-computer interaction is characterized by natural face-to-face conversation: from dialog systems for information access and e-commerce services, to e-learning tutoring for teaching speech and language skills to children, to animation of avatars and characters in virtual environments and computer games.

The final goal of our work is to develop an Italian female IVA we named LUCIA (Cosi, Fusaro, & Tisato, 2003) that would be able to engage adults and children in face-to-face conversational interaction.

At this stage of LUCIA's development, only her expressive talking head capabilities have been taken into consideration. LUCIA is, in fact, a three-dimensional animated MPEG-4 (http://www.chiariglione.org/mpeg) computer talking head that produces emotive/expressive natural speech produced by an emotive/expressive version (Tesser, Cosi, Drioli, & Tisato, 2005) of the Italian Festival TTS (Cosi, Tesser, Gretter, & Avesani, 2001), as illustrated in the block diagram shown in Figure 10–1, and a wide variety of facial expressions and emotions.

LUCIA's Animation Engine

The knowledge that both acoustic and visual signal simultaneously convey linguistic, extralinguistic, and paralinguistic information is well accepted in the speech communication community, and this knowledge constitutes the basis for the work presented here.

Instead of imposing ad hoc expert rules, a data-driven procedure was utilized to build LUCIA both from an acoustic and a

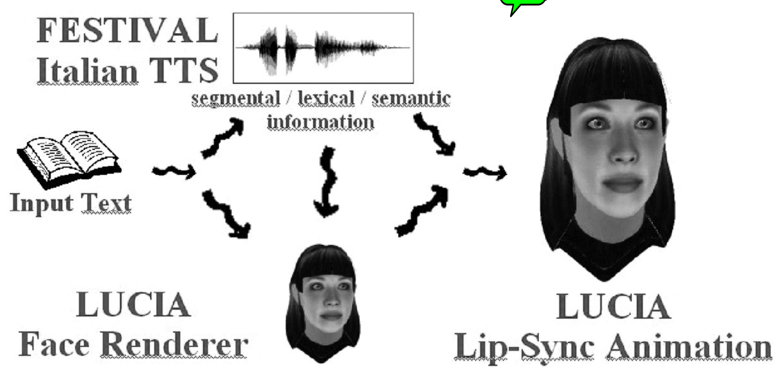


Figure 10–1. LUCIA's functional block diagram.

visual point of view. Our Italian talking head has been in fact directly driven by various style- and emotion-specific speech corpora together with their emotional visual counterparts, in the form of labial movements and facial expressions.

Visual Framework

Visual data are physically extracted by an automatic optotracking movement analyzer for 3D kinematics data acquisition called ELITE (Ferrigno & Pedotti, 1995). ELITE provides 3D coordinate reconstruction starting from 2D perspective projections by means of a stereophotogrammetric procedure, which allows a free positioning of the TV cameras. The 3D data coordinates of 28 reflecting markers positioned on the model subject face are then used to create a lips articulatory model and to drive directly, copying human facial movements, our expressive and emotive talking face (Figure 10–2).

All the movements of the 28 markers are recorded and collected, together with their velocity and acceleration, simultaneously with the coproduced speech, which is usually segmented and analyzed by means of PRAAT (http://www.fon.hum.uva.nl/praat) (Boersma, 1996), that computes also intensity, duration, spectrograms, formants, pitch synchronous F0, and various voice quality (VQ) parameters that are quite significant in characterizing emotive/expressive speech (Drioli, Tisato, Cosi, & Tesser, 2003; Magno Caldognetto, Cosi, Drioli, Tisato, & Cavicchio, 2003).

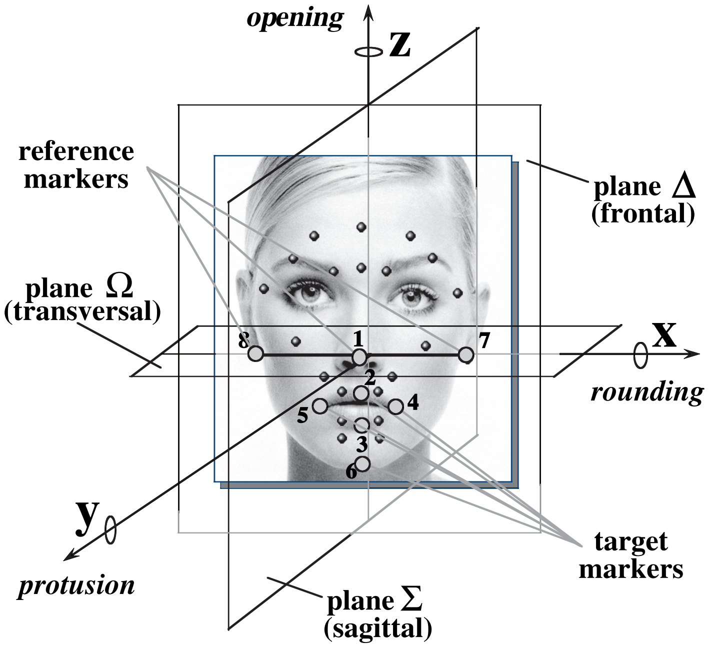


Figure 10–2. Position of reflecting markers and reference planes for the articulatory movement data collection on a real face.

Lips Articulation Model

The most common parameters selected to quantify the labial configuration modifications used in the analysis of the labial movements are listed in Table 10–1.

The parameter estimation procedure for LUCIA's lips articulation model is based on a least squared phoneme-oriented error minimization scheme with a strong convergence property, between real articulatory data Y(n) and modeled curves F(n) for the whole set of R stimuli belonging to the same phoneme set:

$$e=\sum_{r=1}^{R}\left(\sum_{n=1}^{N}\left(Y_r(n)-F_r(n)\right)^2\right) \tag{1}$$

where F(n) is generated by a modified version of the Cohen-Massaro coarticulation model (Cohen & Massaro, 1993) as introduced in Pelachaud et al. (2[illegible]) and Cosi et al. (2002a). This model implements Löfqvist's gestural theory of speech production (Löfqvist, 1990), and it is profoundly inspired by Browman and Goldstein's work on articulatory phonology (Browman & Goldstein, 1986). Each phoneme is specified in terms of speech control parameters (e.g., lip rounding, upper and lower lip displacement, lip protrusion) characterized by a target value and a dominance function.

Dominance functions of consecutive phonemes overlap in time and specify the degree of influence that a speech segment has over articulators in the production of preceding or following segments. The final articulatory trajectory of a specific parameter is the weighted average of the sum of all dominances scaled by the magnitude of the associated targets.

The values of the coefficients of the model have been determined starting from a corpus of real labial movements

Table 10–1. Meaning of some of the most commonly chosen articulatory parameters

Lip opening (LO), calculated as the distance between markers placed on the central points of the upper and lower lip vermillion borders [d(m2,m3)]; this parameter correlates with the *high-low* phonetic dimension.

Lip rounding (LR), corresponding to the distance between the left and right corners of the lips [d(m4,m5)], which correlates with the *rounded-unrounded* phonetic dimension: negative values correspond to the lip spreading.

Anterior/posterior movements (protrusion) of upper lip and lower lip (ULP and LLP), calculated as the distance between the marker placed on the central points of either the upper and lower lip and the frontal plane ?? containing the line crossing the markers placed on the lobes of the ears and perpendicular to ?? plane [d(m2, ?), d(m3, ?)]. These parameters correlate with the feature *protruded-retracted*: [illegible]ive values quantify the lip retraction.

Upper and lower lip vertical displacements (UL, LL), calculated as a distance between the markers placed on the central point of either upper and lower lip and the transversal plane ? passing through the tip of the nose and the markers on the ear lobes [d(m2, ?), d(m3, ?)]. Hence, positive values correspond to a reduction of the displacement of the markers from the ? plane. As told before, these parameters are normalized in relation to the lip resting position.

of an Italian speaker pronouncing VCV symmetrical stimuli, where V is one of the vowels /a/ /i/ or /u/, and C is one of the Italian consonant phonemes. The corpus represents spatio-temporal trajectories of labial parameters such as those specified in Table 10–1, and even if the number of parameters to be optimized is rather high, the size of the corpus is large enough to allow a meaningful estimation. However, due to the presence of several local minima, the optimization process has to be manually controlled in order to assist the algorithm convergence.

The mean total error between real and simulated trajectories for the whole set of parameters is lower than 0.3 mm in the case of bilabial and labiodental consonants in the /a/ and /i/ contexts (Perin, 2000, p. 63).

MPEG-4 Animation

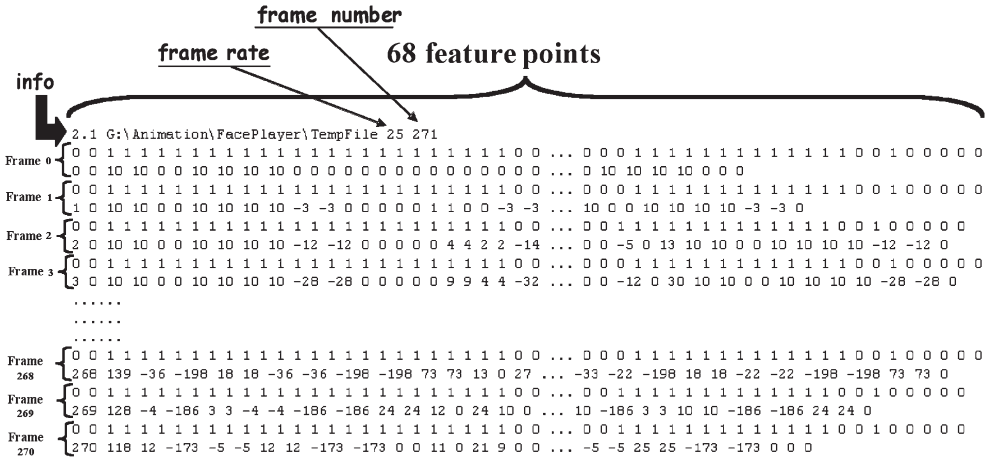
LUCIA is able to generate a 3D mesh polygonal model by directly importing its structure from a VRML file (Hartman & Wernecke, 1996) and to build its animation in real time.

LUCIA emulates the functionalities of the mimic muscles, by the use of specific *displacement functions* and of their following action on the skin of the face. The activation of such functions is determined by specific parameters that encode small muscular actions acting on the face, and these actions can be modified in time in order to generate the wished animation. Such parameters, in MPEG-4, take the name of *facial animation parameters,* and their role is fundamental for achieving a natural movement. Moreover, the muscular action is made explicit by means of the deformation of a polygonal reticule built around some particular key points called *facial definition parameters* (FDP) that correspond to the junction on the skin of the mimic muscles.

LUCIA is a graphic MPEG-4 compatible facial animation engine implementing a decoder compatible with the *predictable facial animation object profile*. FDPs define the shape of the model while FAPs define the facial actions and, given the shape of the model, the animation is obtained by specifying the FAP stream that is for each frame the values of FAPs (Figure 10–3). In a FAP stream, each frame has two lines of parameters. In the first line the activation of a particular marker is indicated (0, 1), while in the second, the target values are stored in terms of differences from the previous ones.

Moving only the FDPs is not sufficient to smoothly move the whole 3D model; thus, each *feature point* is related to a particular *influence zone* constituted by an ellipses that represents a zone of the reticule where the movement of the vertexes is strictly connected. Finally, after having established the relationship for the whole set of FDPs and the whole set of vertexes, all the points of the 3D model can be simultaneously moved with a graded strength following a raised-cosine function rule associated to each FDP.

Each feature point follows MPEG-4 specifications where a FAP corresponds to a minimal facial action. When a FAP is activated (i.e., when its intensity is not null) the feature point on which the FAP acts is moved in the direction signaled by the FAP itself (up, down, left, right, etc). Using the pseudomuscular approach, the facial model's points within the region of this particular feature point get deformed. A facial expression is characterized, not only by the muscular contraction that gives rise to it, but also by an intensity and a duration. The intensity factor is



```
2.1 G:\Animation\FacePlayer\TempFile 25 271
0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 ... 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 1 0 0 0 0 0
0 0 10 10 0 0 10 10 10 10 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 ... 0 10 10 10 10 0 0 0
0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 ... 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 1 0 0 0 0 0
1 0 10 10 0 0 10 10 10 10 -3 -3 0 0 0 0 0 1 1 0 0 -3 -3 ... 10 0 0 10 10 10 10 -3 -3 0
0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 ... 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 1 0 0 0 0 0
2 0 10 10 0 0 10 10 10 10 -12 -12 0 0 0 0 0 4 4 2 2 -14 ... 0 0 -5 0 13 10 10 0 0 10 10 10 10 -12 -12 0
0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 ... 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 1 0 0 0 0 0
3 0 10 10 0 0 10 10 10 10 -28 -28 0 0 0 0 0 9 9 4 4 -32 ... 0 0 -12 0 30 10 10 0 0 10 10 10 10 -28 -28 0
......
......
......
0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 ... 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 1 0 0 0 0 0
268 139 -36 -198 18 18 -36 -36 -198 -198 73 73 13 0 27 ... -33 -22 -198 18 18 -22 -22 -198 -198 73 73 0
0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 ... 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 1 0 0 0 0 0
269 128 -4 -186 3 3 -4 -4 -186 -186 24 24 12 0 24 10 0 ... 10 -186 3 3 10 10 -186 -186 24 24 0
0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 ... 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 1 0 0 0 0 0
270 118 12 -173 -5 -5 12 12 -173 -173 0 0 11 0 21 9 0 0 ... -5 -5 25 25 -173 -173 0 0 0
```

Figure 10–3. The FAP stream.

rendered by specifying an intensity for every FAP. The temporal factor is modeled by three parameters: onset, apex, and offset (Ekman & Friesen, 1978).

The FAP stream needed to animate a *facial animation engine* (FAE) could be completely synthesized by using a specific animation model, such as the lips coarticulation one used in LUCIA, or it could be reconstructed on the basis of real data captured by an optotracking hardware, such as ELITE.

At the current stage of development, as illustrated in Figure 10–4, LUCIA is a textured young female 3D face model built with 25,423 polygons: 14,116 belong to the skin, 4616 to the hair, 2688 × 2 to the eyes, 236 to the tongue, and 1029 to the teeth, respectively.

Currently the model is divided in two subsets of fundamental polygons: the skin on one hand and the inner articulators, such as the tongue and the teeth, or the facial elements such as the eyes and the hair, on the other. This subdivision is quite useful when animation is running, because only the reticule of polygons corresponding to the skin is directly driven by the pseudomuscles, and it constitutes a continuous and unitary element, while the other anatomical components move themselves independently and in a rigid way, following translations and rotations (for example, the eyes rotate around their center). According to this strategy the polygons are distributed in such a way that the resulting visual effect is quite smooth with no rigid "jumps" over the entire 3D model.

Visual Emotions

Different from lip movements, at the present time, emotional visual configurations are not learned by specific model built on real data, such as the lips, but are designed and refined by means of visual inspection of real data. Emotional configurations are then superimposed to the lips' movement. As already underlined in the above section, in MPEG-4 animations, FDPs define the shape of the model while FAPs define the facial actions. The intensity and the duration of an emotive expression are driven by an intensity factor that is rendered by specifying an intensity for every FAP, and by a tempo-

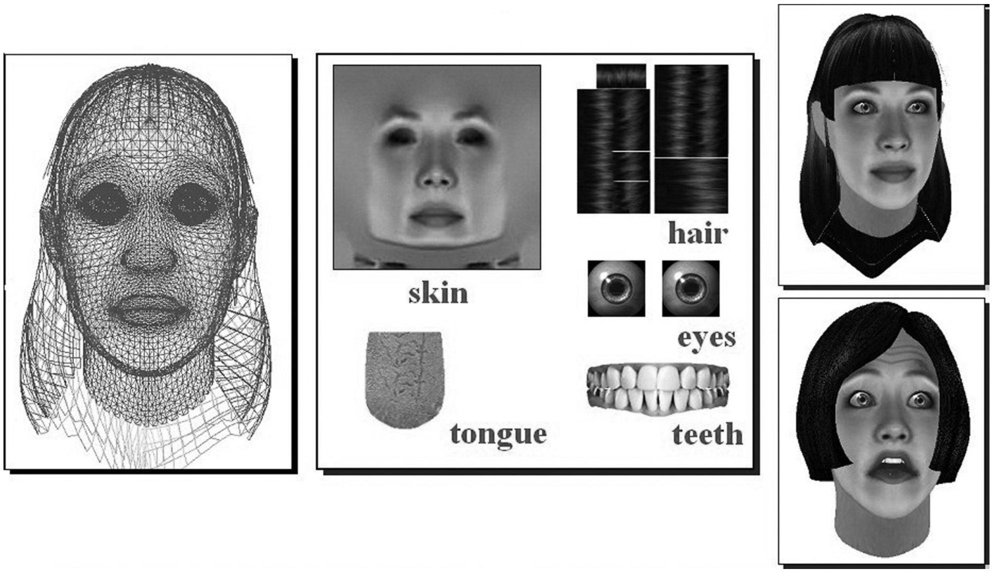


Figure 10–4. Lucia's wireframe, textures, and renderings.

ral factor which is modeled by onset, apex, and offset parameters, as explained in Ekman and Friesen (1978).

The onset and offset represent, respectively, the time the expression takes to appear and to disappear; the apex corresponds to the duration for which the facial expression is at its peak intensity value. These parameters are fundamental to convey the proper meaning of the facial expressions. In our system, every facial expression is characterized by a set of FAPs. Every set of FAPs allows for example the creation of the six facial expressions corresponding to the six basic primary emotions of Ekman's set (Table 10–2), chosen here for the sake of simplicity, and for every expression only three levels of intensity (low, medium, high) have been simulated.

As for LUCIA, "emotion basis" *EB(t)* and "emotion display" *ED(t),* both functions of the time *t,* are distinguished. An *EB(t)* involves a specific zone of the face such as the eyebrow, mouth, jaw, eyelid, and so on. *EB(t)* includes also facial movements such as nodding, shaking, turning the head, and movement of the eyes. Each *EB(t)* is defined as a set of MPEG-4 compliant FAP parameters:

$$EB(t) = \{ fap3 = v_1(t); \; ....; fap68 = v_k(t)\}$$

where $v_1(t),. . . ,v_k(t)$ specify the FAPs' function intensity value created by the user. An *EB(t)* can also be defined as a combination of *EB'(t)* by using the '+' operator in this way:

$$EB'(t) = EB_1'(t) + EB_2'(t)$$

The emotion display is finally obtained by a linear scaling:

$$ED'(t) = EB(t)*c = \{ fap3 = v_1(t)*c; \; ....; fap68 = \mathrm{v}_k(t)*c)\}$$

where *EB* is a *facial basis* and *c* a constant. The operator * multiplies each of the FAPs constituting the *EB* by the constant *c*. The onset, offset, and apex (i.e., the

Table 10–2. The six basic primary emotions of Ekman's set with corresponding facial expressions

Expression	Description
Anger	The inner eyebrows are pulled downward and together. The eyes are wide open. The lips are pressed against each other or opened to expose the teeth.
Fear	The eyebrows are raised and pulled together. The inner eyebrows are bent upward. The eyes are tense and alert.
Disgust	The eyebrows and eyelids are relaxed. The upper lip is raised and curled, often asymmetrically.
Happiness	The eyebrows are relaxed. The mouth is open and the mouth corners pulled back toward the ears.
Sadness	The inner eyebrows are bent upward. The eyes are slightly closed. The mouth is relaxed.
Surprise	The eyebrows are raised. The upper eyelids are wide open, the lower relaxed. The jaw is opened.

duration of the expression) of emotion is determined by the weighed sum of the functions $v_k(t)$ (k = 3,...,68) created by mouse actions. In Figure 10–5, two simple emotional examples for fear and happiness are illustrated.

LUCIA's Voice: Emotive Text-to-Speech Synthesis

LUCIA's voice relies on an emotive text-to-speech synthesis aimed at producing emotionally adequate speech starting from a tagged text. Emotive speech synthesis has recently gained much attention, as new expressive/emotive human-machine interfaces are being studied that try to simulate the human behavior while reproducing man-machine dialogs, and various attempts to incorporate the expression of emotions into synthetic speech have been made (Murray & Arnott, 1993; Schröder, 2001).

The speech characterization of a certain emotion must be defined by the measure of its associated acoustic correlates, which directly derive from the physiologic constraints. For example, when feeling fear or happiness, the heart beats, bloody pressure increases, mouth becomes dry and there are occasional muscle tremors, the voice increases in loudness and speech rate, and the spectrum becomes richer in high frequency components (Cahn, 1990). Many researches on the vocal expression of emotion have demonstrated that many features may be involved. They tended to be focused on prosody correlates such as pitch variables, especially F0 level and range, but also on the pitch contour and on the amount of jitter[2] or shimmer,[3] pausing

[2] Jitter is perceptual cycle-to-cycle pitch variations.

[3] Shimmer is the perceptual cycle-to-cycle amplitude variation.

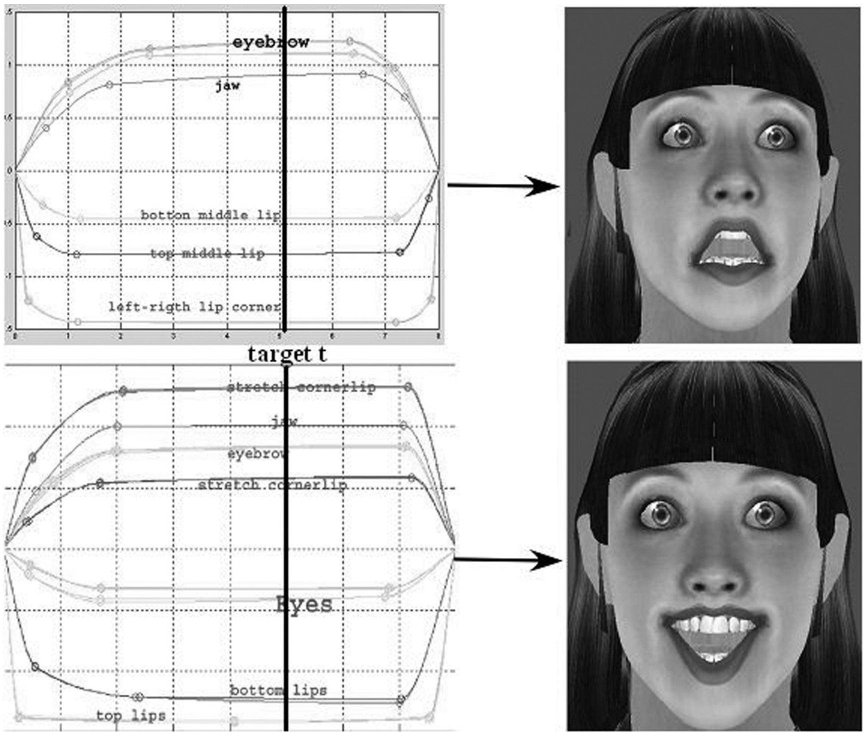


Figure 10–5. Fear (top) and happiness (bottom) emotional examples.

structure, speech rate, and intensity differences (Gobl & Chasaide, 2003). These parameters are relatively easy to measure and control (Scherer, 1986). However, other fundamental speech correlates of the emotions are based on the spectral and voice/speech quality[4] analysis.

The voice engine of LUCIA is based on the Festival text-to-speech synthesis framework developed at CSTR (Taylor, Black, & Caley, 1988), and on the MBROLA diphone concatenation acoustic back-end (Dutoit & Leich, 1993). In the following, two components that are particularly relevant to emotive speech synthesis will be discussed, that is, the prosodic control module and the voice quality control module.

The task of a prosodic module in a TTS synthesizer is that of computing the values of a set of prosodic variables, starting from the linguistic information contained in the text that has to be synthesized. In up-to-date TTS technologies, synthesis control has been mainly focusing on phoneme duration and pitch, which are the two main parameters conveying the prosodic information. Recently, data-driven machine learning techniques, such as CARTs (classification and regression trees; Breiman, Friedman, Olshen, & Stone, 1984), have proven to be effective for prosody modeling, and for providing substantial improvements over previously known rule-based approaches. Trees are

[4] Voice quality distinguishes the modality of the glottal signal production (tense voice, creaky voice, modal voice, breathy voice, harsh voice, whispery voice).

constructed by a data-driven training process, and are made of a set of yes/no or if/then questions relating to the structural linguistic data in order to predict the dependent prosodic variable.

Results on the use of simple phoneme duration and F0-based CARTs for learning different speaking styles for Italian have been reported in Cosi et al. (2002b) and Tesser et al. (2004), and the learning of emotive prosody is addressed in Tesser et al. (2005). This approach to the modeling of prosodic correlates of emotional speech will be discussed in the following.

Today, the speech synthesis community is also showing an increasing interest in the control of a broader class of voice characteristics. As an example, voice quality is known to play an important role in emotive speech, and some recent studies have addressed the exploitation of source models within the framework of articulatory synthesis to control the characteristics of voice phonation (d'Alessandro & Doval, 1998; Gobl & Chasaide, 2003). Intensity, for example, intended as the acoustical correlate of loudness, can in principle be roughly controlled by changing a gain factor uniformly across the spectrum. However, it is recognized that the result of such processing is not perceived as natural due to the lack of spectral balance modifications correlated to vocal effort variations that occur in real speech (Campbell, 1995; Sluijter, van Heuven, & Pacilly, 1997). Other voice quality cues are important as well in the characterization of nonmodal phonatory styles observed in emotive speech, and the control of the related acoustic cues in a diphone concatenation framework is required to embed a set of ad hoc signal processing techniques to modify the original timbre characteristics of recorded diphones.

Various approaches are known that can be used to convert the timbre of a neutral voice into that of an emotional one. A possible choice is to use voice conversion algorithms (Kain & Macon, 2001; Stylianou, Cappé, & Moulines, 1998), usually based on Gaussian mixture models, which are used to transform the spectral characteristics of the voice of a speaker into that of another one. Some experiments have been done with this regard (Drioli et al., 2003) but only for a small corpus of VCV[5] sequence such as ”aba” /’aba/ and ”ava” /’ava/ uttered with different emotions. Another possible technique is based on the study and analysis of the acoustic correlates of the emotions, aimed at providing simple signal processing manipulations (e.g., changing the spectral tilt, adding aspiration noise, or changing the shimmer and jitter), which will allow switching from a neutral timbre to an emotional one. Here we will discuss this last approach in combination with the CART-based prosodic modeling.

Analysis of Emotive Speech: Emotional Database

The best speech material suited for studying emotions would be that produced spontaneously during naturally occurring emotional events. However, quite serious methodological problems arise when collecting these data. Emotional voice samples obtained in natural situations are generally rare, very brief, and not infrequently suffering from bad recording quality (Scherer, 2003). In addition, there

[5] VCV means vowel-consonant-vowel.

are often severe labeling problems in determining the precise nature of the underlying emotion. Moreover, for training TTS prosodic models a large emotional corpus is required. For all these reasons, the simulated (portrayed) vocal expressions approach was chosen here. Professional actors were asked to produce vocal expressions of emotion (often using standard verbal content) as based on emotion labels and/or typical scenarios (Scherer, 2003). Using this method, as pointed out by Scherer (1986), it cannot be excluded that actors overemphasize the expression of emotion, and that emotions reflect socio-cultural norms or expectations more than the psychophysiological effects on the voice as they occur under natural conditions. Nevertheless, unlike for the emotional ASR framework, TTS synthesis is a field in which several applications oriented towards the traditional emotion archetypes exist (Douglas-Cowie, Cowie, & Campbell, 2003).

The Emotional-CARINI (E-Carini) database has been recorded for this study. It contains the recording of the novel *Il Colombre* by Dino Buzzati read and acted by C. Carini, a professional Italian actor, in different elicited emotions. According to Ekman's theory (Ekman, 1992) six basic emotions plus the neutral (narrative-style) one have been taken into consideration: anger, disgust, fear, happiness, sadness, and surprise. The duration of the database is about 15 minutes for each emotion.

Emotional Prosodic Data-Driven Modeling: A Differential Approach

A wide number of studies on speech and emotions investigate the differences of emotional states with respect to a "neutral" state (Anolli & Ciceri, 1997; Huang, Hon, & Acero, 2001; Murray & Arnott 1993), and the transformation of a neutral utterance (real or synthetic) into an emotional one has been attempted with various techniques. Here a CART data-driven approach will be used to design an emotional prosodic module that learns the differences between the neutral prosody and the emotional one. For each prosodic parameter x (i.e., F0 and duration), the parameter difference is given by $\Delta \mathrm{x} = \mathrm{x}_E - \mathrm{x}_N$, where x_E is the emotional value for parameter x, as given by the acoustic analysis of the emotional database, and x_N is the neutral one, as predicted by a prosodic module trained on a neutral database. In the synthesis stage, the emotional data will be obtained using the simple superimposing model $\mathrm{x}_E = \mathrm{x}_N + \Delta \mathrm{x}$. To be able to separate the macro-prosody factors from the micro-segmental prosody ones, and to reduce data sparseness, various solutions were adopted, including the use of z-scores, normalization with respect to value ranges, and the use of parametric models for intonation curve (Tesser, 2005; Tesser et al., 2005).

The training of differential CARTs was preferred over the training of emotion-specific CARTs, because this approach allowed us to use smaller databases for the different emotions (15 minutes each in our case, whereas the neutral one had a duration of about 50 minutes). Moreover, with this approach it is straightforward to implement smooth transitions from neutral to emotional speech for each emotion.

Duration E-Model

The macro-prosodic differences on duration are represented in Table 10–3, where the average statistics of the duration of

Table 10–3. Phoneme duration means (μ) and standard deviations (σ) calculated on the E-Carini database for different emotions

Emotion	μ (s)	σ (s)	μ_Δ (s)	σ_Δ (s)
Neutral	0.094	0.045	—	—
Anger	0.077	0.034	–0.017	–0.010
Disgust	0.103	0.055	0.009	0.011
Fear	0.078	0.036	–0.016	–0.009
Joy	0.076	0.032	–0.018	–0.013
Sadness	0.104	0.052	0.010	0.007
Surprise	0.076	0.033	–0.018	–0.012

Note. The columns μ_Δ and σ_Δ represent the mean and the standard deviation of the differences between the emotional durations that and neutral ones.

phones in the different emotions are shown.

Looking at the differences between the emotive durations and the neutral ones ($?_?$), it can be observed that the emotions with a negative $?_{??}$ might have a faster rhythm in comparison with the neutral one, while those with a positive value might be slowed down. This is only a broad analysis, but it is useful to have a general picture of the various rhythms used in the emotions. Details on the procedures for the training of the statistical model are given in Tesser et al. (2005). Using this approach the emotive duration prediction can be separate into a macro- and a segmental prosodic part. The macro-prosodic part is implemented by a table of the means and standard deviations for each phoneme and emotion, and the segmental prosodic part it is implemented by the differential CARTs.

Intonation E-Model

The macro-prosodic component of the intonation is the F0 mean and range values. Table 10–4 shows the average statistics of the pitch mean, lower bounds (LB) and upper bounds (UB), the pitch range (R = UB – LB), and the differences with the neutral for each emotions in the E-Carini database.

Also in the intonation case for effective comparison between neutral and emotional values it is necessary to find a good representation of the data. A valuable representation for intonation curves is the PaIntE (Parametric Representation of Intonation Events) model (Cosi et al., 2002c; Möhler, 1998; Möhler & Conkie, 1998). A pitch range normalization was performed in order to get rid of the influence of different pitch range levels in the different emotions. This normalization was done using the LB and UB values of Table 10–4. If we call PN_{Ereal} the real normalized PaIntE parameter vector in the emotional case and PN_{Npred} the predicted normalized PaIntE vector in the neutral case, the difference is given by ?PN = $PN_{Ereal} – PN_{Npred}$. The whole procedure for the design of the emotional intonation module is given in Tesser et al.

Table 10–4. Pitch boundaries and means for the different emotions in the E-Carini data-base

Emotion	*LB* (Hz)	μ (Hz)	*UB* (Hz)	*R* (Hz)	LB_Δ (Hz)	μ_Δ (Hz)	UB_Δ (Hz)	R_Δ (Hz)
Neutral	62	105	213	169	—	—	—	—
Anger	66	122	258	192	4	17	45	23
Disgust	53	81	238	185	–9	–24	25	16
Fear	66	114	223	157	4	9	10	–12
Joy	63	129	308	245	1	24	95	76
Sadness	53	89	208	155	–9	–16	–5	–14
Surprise	66	136	250	184	4	31	37	15

Note. The LB column represents the lower boundary, UB the upper boundary, and R the pitch range. The columns with the subscript symbol Δ represents the same entity calculated on the differences between the emotional F0 and the neutral one.

(2005). Using this approach the emotive intonation prediction can be separate in a macro-prosodic part and a segmental one. The macro-prosodic part is implemented by Table 10–4 of the UB and LB means of each emotion and the segmental prosodic part is implemented by the differential CART.

Intensity E-Model

Due to the current state of art diphone synthesizer, the segmental intensity model does not have a great perceptual relevance. Nevertheless, if we examine the intensity means of different emotions (Table 10–5), we can notice the differences in the average statistics of the intensity evaluated for nonsilence speech.

Table 10–5. Intensity means (μ) for different emotions in the E-Carini database

Emotion	μ (dB)	μ_Δ (dB)
Neutral	70.1	—
Anger	71.9	1.8
Disgust	68.5	–1.6
Fear	70.1	0
Joy	72.7	2.6
Sadness	62.2	–7.9
Surprise	70.1	0

Note. The column with Δ represent the differences between the emotional intensity and the neutral ones.

Joy (72.7 dB) and anger (71.9 dB) are characterized by a high intensity mean, which is a characteristic of emotions with a high degree of physiological activation. By contrast, there are sadness (62.2 dB) and disgust (68.5 dB) with a low intensity mean, while for surprise and neutral we have a medium intensity (70.1 dB). For fear, the intensity mean values collected from the E-Carini database lie around the same values obtained for neutral and surprise (70.1 dB). This is in contrast with what can be found in other studies (Anolli & Ciceri, 1997), and is due principally to the choice of considering the intensity only at the macro-prosodic level using a simple perceptual energy filter on the whole phrase.

Voice Quality (VQ) E-Model

In order to be able to control voice quality in diphone concatenative synthesis, it is necessary to embed opportune signal processing routines into the synthesis framework. The Festival-MBROLA speech synthesizer, which originally provides controls only for pitch and phoneme duration, has been further extended to allow for control of a set of low level acoustic parameters that can be combined to produce the desired voice quality effects. Time evolution of the parameters can be controlled over the single phoneme by instantaneous control curves. Here we give a rough description of the implementation of some of the low level acoustic controls:

- Spectral tilt ("SpTilt"): the spectral balance is changed by a reshaping function in the frequency domain that enhances or attenuates the low and mid-frequency regions, thus changing the overall spectral tilt;
- F0 flutter ("F0Flut"): random low frequency fluctuations of the pitch; the low frequency fluctuations are obtained by random noise band-pass filtering, the second order band-pass filter being tuned in the 4 Hz–10 Hz range;
- Spectral warping ("SpWarp"): the rising or lowering of upper formants is obtained by warping the frequency axis of the spectrum (through a bilinear transformation), and by interpolation of the resulting spectrum magnitude with respect to the DFT frequency bins.

A three-level hierarchic model was designed, in which the affective high level attributes (e.g., <anger>, <joy>, <fear>, etc.) are described in terms of medium level voice quality attributes defining the phonation type (e.g., <modal>, <soft>, <pressed>, etc.). These medium level attributes are in turn described by a set of low level acoustic attributes defining the perceptual correlates of the sound (e.g., <spectral tilt>, <shimmer >, <jitter>, etc.). The low level acoustic attributes correspond to the acoustic controls that the extended MBROLA synthesizer can render through the sound processing procedures described above. In Figure 10–6, an example of a qualitative description of high level attributes through medium and low level attributes is shown.

Given the hierarchical structure of the acoustic description of emotive voice, we performed preliminary experiments

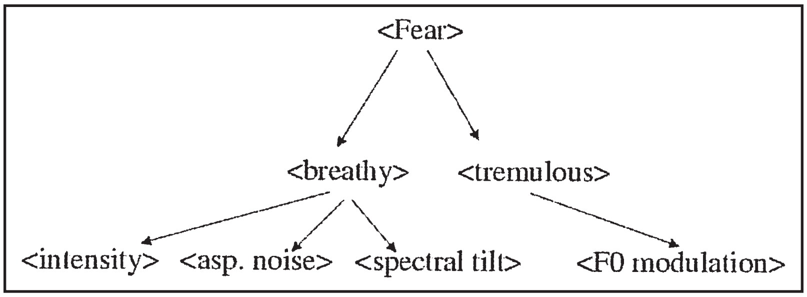


Figure 10–6. Qualitative description of voice quality for fear in terms of acoustic features.

focused on the definition of speaker-independent rules to control voice quality within a text-to-speech synthesizer. Different sets of rules describing the high and medium level attributes in terms of low level acoustic cues where designed, based on acoustic analysis of the E-Carini database and on previous studies (Drioli et al., 2003). Table 10–6 shows an example of the mapping between emotions, voice quality, and low level acoustic parameters that was implemented and adopted in our experiments. Values are in the range [0,1], and have different meanings for the different parameters. For example, SpTilt = 0 means maximal deemphasis of higher frequency range, whereas SpTilt = 0 means maximal emphasis; AspNoise = 0 means absence of noise component, whereas AspNoise = 1 means absence of voiced component, thus letting aspiration noise component alone; for F0Flut, Shimmer, and Jitter, value = 0 means effect is off, whereas value = 1 means effect is maximal; SpWarp = 0 means maximal spectrum shrinking, and SpWarp = 1 means maximal spectrum stretching.

Anger is characterized by a loud, harsh voice implemented by an attenuation of the low-mid frequency (SpTilt = 0.3) and a lowering of upper formants (SpWarp = –0.4). To realize the fear voice quality (tremulous and breathy) a random F0 fluctuation is added (F0Flut = 0.7). Joy and surprise (loud and breathy) are realized by a high attenuation of the low-mid frequency (SpTilt = 0.4). Disgust has a harsh voice quality that has been realized by an attenuation of the low-mid frequency (SpTilt = 0.3) and a raising of upper formants (SpWarp = 0.35). The breathy voice quality of sadness is implemented by a substantial lowering of upper formants (SpWarp = –0.5).

Emotional Festival TTS

A general overview of the Festival-MBROLA architecture framework for emotional TTS synthesis in shown in Figure 10–7: for a given emotion and a given input text, the Natural Language Processing (NLP) module operates to produce a phonetic-linguistic representation of the text.

These data are used by the prosodic modules to predict the emotive prosody. Both the duration and intonation modules use the differential approach: the internal data are used by both the neutral

Table 10–6. Voice quality modifications and low- level acoustic parameters implementation for different emotions

Emotion	Voice quality	Low-level acoustic parameters
Anger	loud and harsh	SpTilt = 0.3, SpWarp = 0.4
Disgust	harsh	SpTilt = 0.3, = SpWarp = 0.35
Fear	tremulous and breathy	SpTilt = 0.3, F0Flut = 0.7, SpWarp = 0.4
Joy	loud and breathy	SpTilt = –0.4
Sadness	breathy	SpWarp = –0.5
Surprise	loud and breathy	SpTilt = 0.4

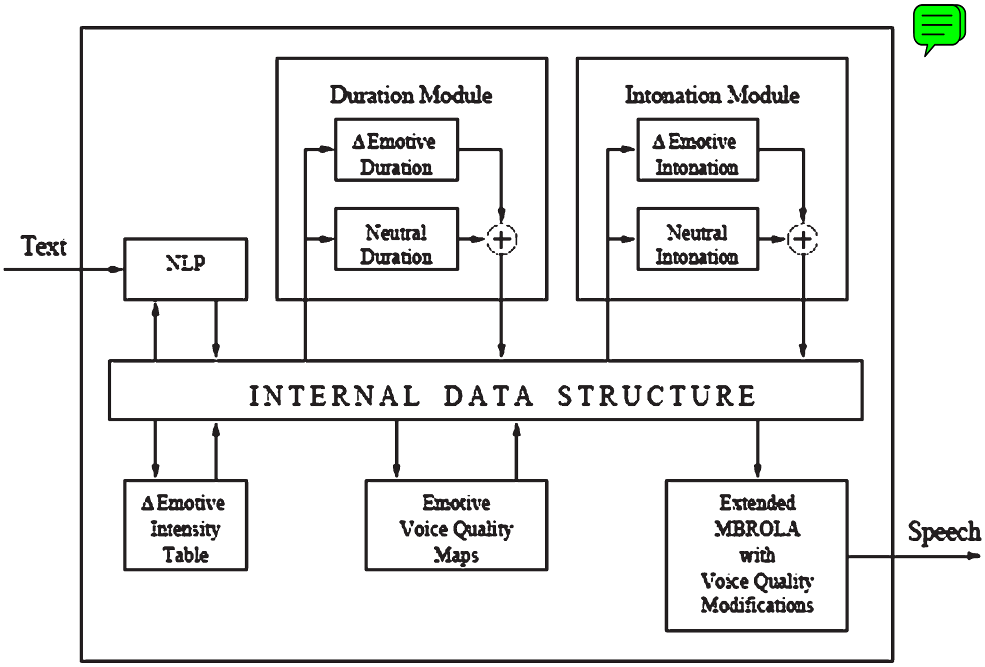


Figure 10–7. Overall functional diagram of the implemented Festival-MBROLA E-TTS.

module and emotional differential one and subsequently summed to provide the emotional prosody pattern.

Evaluation

The prosody prediction models were assessed both with an objective and a subjective evaluation. Moreover, prosody prediction and voice quality modifications were assessed together and separately with a subjective evaluation.

Objective Evaluation

An objective evaluation of the prosodic modules was performed by splitting both the Carini and E-Carini database in a training set (90%) and a test set (10%), and measuring the differences between the synthetic prosody and the actual prosody in the test set. An indication of the performance can be given by the RMSE and correlation ρ between the original prosodic signal and the predicted one, and by the absolute error |e| between the two prosody patterns. Table 10–7 shows the RMSE and the correlation ρ between the original and the predicted values computed by the duration module for the different emotions. The mean and the variance of the absolute error |e| are also given. The values on the first three columns are expressed in *z*-score units, while the values on the last three columns are expressed in seconds.

In prosody prediction the most significant values are the RMSE and the correlation: looking at the *z*-score RMSE and correlation columns the best performance is obtained by the neutral duration module (lowest RMSE 0.88 and highest correlation 0.64). This can be due to

Table 10–7. Duration prediction results for the different emotions

Emotion	RMSE	$\mu_{\|e\|}$	$\sigma_{\|e\|}$	ρ	RMSE(s)	$\mu_{\|e\|}$(s)	$\sigma_{\|e\|}$(s)
Neutral	0.88	0.66	0.58	0.64	0.039	0.039	0.026
Anger	1.20	0.79	0.90	0.46	0.041	0.027	0.031
Disgust	0.99	0.65	0.74	0.50	0.054	0.035	0.041
Fear	1.03	0.70	0.75	0.56	0.037	0.025	0.027
Joy	0.88	0.62	0.63	0.61	0.028	0.020	0.020
Sadness	0.89	0.60	0.66	0.59	0.046	0.031	0.034
Surprise	0.93	0.65	0.66	0.62	0.030	0.022	0.022

the fact that the emotional database is smaller than the neutral one, and to the fact that the errors in the emotional modules are the sum of the errors on the neutral module and on the differential one. The worst result has been obtained by anger, with a resulting very simple differential CART, and best results were obtained for the surprise and joy modules, with more complex corresponding CARTs.

Looking again at the *z*-score RMSE column, sadness has a good performance too (0.89), and surprise (0.93), disgust (0.99), and fear (1.03) have mid-low scores. As for the correlation coefficient, the duration module of disgust, fear, and sadness have mid values (all above 0.5). Looking at the values expressed in seconds it can be noticed that joy and surprise have the best performance, but it is necessary to underline that these values are intrinsically correlated with the speech rate of the given emotion, and then joy and surprise that have the highest speech rate will have the lowest values of RMSE if express in seconds. However, these values can be useful to having a dimensional idea of the prediction errors.

Intonation. Table 10–8 shows the results for the different emotions in the objective evaluation test for the intonation module. The values on the first three columns are expressed in pitch normalized units,[6] while the values on the last three columns are expressed in Hz.

Also in the intonation case the best performance has been obtained by the neutral intonation module (lowest RMSE 0.13 and highest correlation 0.43). Making a comparison with the correlation values for the durations on Table 10–7, it can be noticed that the correlation values are much lower in the intonation case. This is due to different aspects that make the intonation prediction more difficult than the duration one. Moreover there is a difference between the correlation coefficients of the neutral intonation module (0.43) and the emotive ones (0.28, 0.22, 0.16, 0.23, 0.19, 0.22) and

[6] In pitch normalized scale 0 and 1 correspond respectively to the lower bound and the upper bound of F0 in the given emotion.

Table 10–8. Intonation prediction results for the different emotions on the test set

Emotion	RMSE	$\mu_{\|e\|}$	$\sigma_{\|e\|}$	ρ	RMSE(Hz)	$\mu_{\|e\|}$(Hz)	$\sigma_{\|e\|}$(Hz)
Neutral	0.13	0.09	0.07	0.43	29	20	15
Anger	0.20	0.16	0.12	0.28	38	30	24
Disgust	0.15	0.11	0.10	0.22	28	21	19
Fear	1.25	0.18	0.17	0.16	39	28	26
Joy	0.22	0.18	0.13	0.23	54	43	32
Sadness	0.20	0.14	0.14	0.19	31	22	22
Surprise	0.27	0.21	0.17	0.22	49	39	30

this is probably due also to the decreasing size of the data available for learning the emotional intonation with respect to the neutral case.

As for the emotions looking at the pitch normalized RMSE values, the best performances are obtained by the disgust (0.15). As the opposite, the worst result has been obtained by surprise (0.27), due to the fact this emotion is characterized by a large pitch variability. The Hz RMSE values are obviously influenced by the pitch range of the given emotion; the worst score is obtained by joy (54 Hz) and the best by the low degree of physiological activation emotions: disgust (28 Hz) and sadness (31 Hz). Also in this case the significance of these scores is doubtful, but these values can be useful for having a dimensional idea of the prediction errors.

Subjective Evaluation

The effectiveness of the prosodic modules and of the voice quality modifications was also assessed with perceptual tests aimed at evaluating: (a) the single contribution on the emotional expressiveness carried out separately by the emotional prosodic modules and the emotive voice quality modifications, and (b) the synergistic contribution given by the union of these two correlates of the emotive speech. Four types of test sentences were generated:

(A) *neutral* prosody *without* emotive *VQ* modifications;

(B) *emotive* prosody *without* emotive *VQ* modifications;

(C) *neutral* prosody *with* emotive *VQ* modifications;

(D) *emotive* prosody *with* emotive *VQ* modifications.

For each emotion and for each of these four conditions, two utterances were produced by the new emotional TTS system for a total of 48 sentences, which were presented in a randomized order to 40 listeners, who judged, knowing the target emotion and the level of acceptability of the emotional synthesis, within a MOS scale (5 = excellent, 4 = good, 3 = fair, 2 = poor, 1 = bad). Results are summarized in Figure 10–8. Cases B, C, and D had always better results than those obtained for case A, signifying that emo-

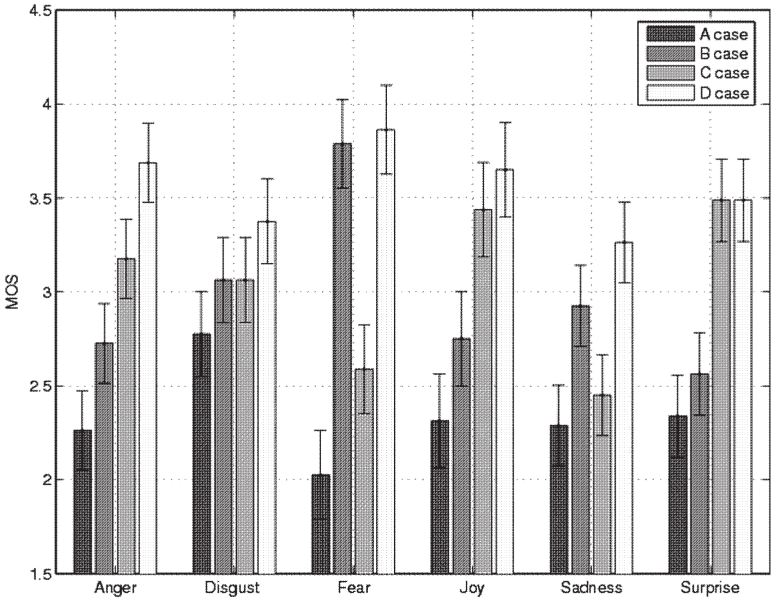


Figure 10–8. Subjective evaluation results (for A,B,C, and D; see text).

tive modules were quite successful. Case D always shows better MOS values and this is an indication that the created emotive prosodic modules quite improve the acceptability of the emotional TTS. Emotive VQ modifications alone were superior to the neutral case, except for fear and sadness. This can be the consequence of the fact that the contribution of prosody and voice quality might differ between different emotions, or might indicate that the chosen VQ acoustic processing should be modified for these emotions.

It is immediately evident that in B, C, and D cases the results are always better than those obtained in the A case, and this indicates that emotive modules were quite successful. D case always shows better MOS values and this gives an indication that the created emotive prosodic modules quite improve the acceptability of the emotional TTS. Emotive VQ modifications alone were superior to the neutral case apart from the fear and sadness case, indicating that with these emotional moods the chosen VQ acoustic modification should be modified.

Emotion Audiovisual Markup Language

The affective presentation markup language (APML) for behavior specification allows one to specify how to mark up the verbal part of a dialog so as to add to it the "meanings" that the visual and the speech generation components of an animated agent need to produce the required expressions (De Carolis et al., 2004). So far, the language defines the components that may be useful to drive a face animation through the facial animation parameters and facial display functions. A scheme for the extension of a previously developed APML has been studied. The extension of such language is intended to support voice-specific controls. An extended version of the APML has been included in the Festival speech synthesis environment, allowing the automatic generation of the MBROLA control file from an APML tagged text with emotive tags. This module implements the three-level hierarchy introduced before, in which the affective high level attributes (e.g., <anger>, <joy>, <fear>) are described in terms of medium level voice quality attributes (phonation type), and low level acoustic attributes (the perceptual correlates of sound). This descriptive scheme has been implemented within Festival as a set of mappings between high level and low level descriptors. This APML extension allows the generation of emotive facial animation and speech synthesis, starting from tagged text such as:

```
<performative type = “inform”>

<voqual type = “modal” level =
“1.0’>This is my modal voice.
</voqual><voqual type =
“tremulous” level = “1.0”>This is
my tremulous voice.</voqual>

</performative>.
```

LUCIA’s System Architecture

FAP stream generation components and the audio synthesis components have been integrated into a unique system able to produce the facial animation, including emotive audio and video cues, from tagged text. The facial animation framework relies on previous studies for the realization of Italian talking heads (Cosi et al., 2003; Magno Caldognetto et al., 2004). A schematic view of the whole system is shown in Figure 10–9.

The modules used to produce the FAP control stream (AVENGINE), and the speech synthesis phonetic control stream (Festival), are synchronized through the phoneme duration information. The output control streams are in turn used to drive the audio and video rendering engines (i.e., the MBROLA speech synthesizer and the face model player).

Conclusions and Future Work

In this chapter, we discussed audio/visual speech analysis and synthesis issues, for the creation of natural and expressive talking heads. An automatic optotracking 3D movement analyzer was implemented, in order to build up the animation engine based on the Cohen-Massaro coarticulation model, and also to create the correct WAV and FAP files needed for the animation of LUCIA, an animated MPEG-4 talking face.

LUCIA’s voice is based on an Italian version of Festival-MBROLA speech synthesis environment, modified for expressive/emotive synthesis. A data-driven method for the design of emotive prosodic modules was illustrated, as well as a rule-based Festival-MBROLA voice quality modification module, designed for control of temporal and spectral characteristics of the synthesis.

Even if emotional synthesis still remains an attractive open issue, our preliminary evaluation results underline the effectiveness of the proposed solutions. From the results of perceptual tests, it can be concluded that adequate emotional

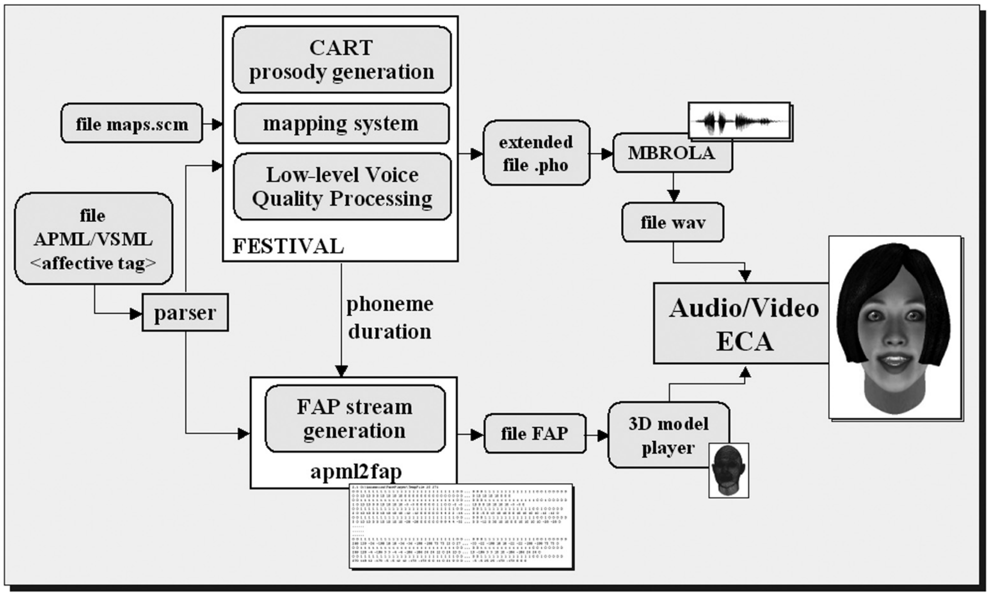


Figure 10–9. Block scheme of the system designed to produce the facial animation with emotive audio and video cues from tagged text.

speech can be obtained within a diphone-based approach, and that emotional prosodic modeling based on data-driven approaches produces appreciable results. Moreover, subjective tests demonstrated that voice quality processing increases the quality of the perceived emotion.

Finally, an interface to the speech synthesizer was described that gives the user the possibility of specifying a desired emotion for the text through a markup language. The audiovisual synthesis system driven by the tagged text integrates voice synthesis and facial animation, so to provide a speech synthesis control stream (PHO file) and a facial animation control stream (FAP file), which are used in turn to produce synchronized and emotionally coherent audio and video outputs.

For the future, improvements are foreseen for some specific aspects, such as the modeling of smooth transition from one emotion to the other, both on the visual and on the acoustic side. Refinements on the evaluation procedures will be conducted such as, for example, perceptual tests for comparing human movements and talking head animations, thus giving us the possibility to gain insights about where and how the animation engine could be improved. Similarly, the modeling of prosody and of voice quality will be improved through the use of more advanced signal processing techniques and specifically designed perceptual tests.

Acknowledgements. Part of this work has been sponsored by PF-STAR (Preparing Future multiSensorial inTerAction Research, European Project IST- 2001-37599, http://pfstar.itc.it), and TICCA (Tecnologie cognitive per l'interazione e la cooperazione con agenti artificiali, joint "CNR - Provincia Autonoma Trentina"

Project 2001–2003). We wish to thank the MBROLA team for providing the source code of their synthesis engine.

References

Anolli, L., & Ciceri, R. (1997). La voce delle emozioni. *Franco Angeli s.r.l*,.

Boersma, P. (1996). PRAAT, a system for doing phonetics by computer. *Glot International*, *5*(9/10), 341–345.

Boula de Mareil, P., Clrier, P., & Toen, J. (2002). Generation of emotions by a morphing technique in English, French and Spanish. In (Ed.), *Proceedings of Speech & Prosody Workshop 2002*, [CD-ROM].

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and regression trees*. Wadsworth and Brooks.

Browman, C. P., & Goldstein, L. (1986). Towards an articulatory phonology. *Phonology Yearbook*, *3*, 219–252.

Cahn, J. E. (1990). The generation of affect in synthesized speech. *Journal of the American Voice I/O Society*, *8*, 1–19.

Campbell, N., & Isard, S. (1991). Segment durations in a syllable frame. *Journal of Phonetics*, [illegible]-47.

Campbell, W. (1995). Loudness, spectral tilt, and perceived prominence in dialogues. In (Ed.), *Proceedings of ICPhS* (pp. 676–679).

Cohen, M., & Massaro, D. (1993). Modeling coarticulation in synthetic visual speech. In N. Magne-nat-Thalmann & D. Thalmann (Eds.), *Models and techniques in computer animation* (pp. 139–156). Tokyo: Springer-Verlag.

Cole, R., van Vuuren, S., Pellom, B., Hacioglu, K., Ma, J., Movellan, J., et al. (2003). Perceptive animated interfaces: First steps toward a new paradigm for human computer interaction. *Proceedings of the IEEE: Special Issue on Multimodal Human Computer Interface*, [illegible]

Cosi, P., Avesani, C., Tesser, F., Gretter, R., & Pianesi, F. (2002). On the use of cart-tree for prosodic predictions in the Italian Festival TTS. In P. Cosi, E. Magno, & A. Zamboni (Eds.), *Voce, canto, parlato-studi in onore di Franco Ferrero* (pp. 73–81). Padova, Italy: UNIPRESS.

Cosi, P., Fusaro, A., & Tisato, G. (2003). LUCIA, a new Italian talking-head based on a modified Cohen-Massaro's labial coarticulation model. In (Ed.), *Proceedings of EUROSPEECH 2003* (pp. 127–132).

Cosi, P., Magno Caldognetto, E., Perin, G., & Zmarich, C. (2002). Labial coarticulation modeling for realistic facial animation. In (Ed.), *Proceedings of 4th IEEE International Conference on Multimodal Interfaces ICMI 2002* (pp. 505–510).

Cosi, P., Tesser, F., Gretter, R., & Avesani, [illegible] (2001). Festival speaks Italian! In (Ed.), *Proceedings of EUROSPEECH 2001* (pp. 509–512).

Cosi, P., Tesser, F., Gretter, R., & Pianesi, F. (2002). A modified "PaIntE Model" for Italian TTS. In (Ed.), *Proceedings of IEEE Workshop on Speech Synthesis*. [CD-ROM].

d'Alessandro, C., & Doval, B. (1998). Experiments in voice quality modification of natural speech signals: The spectral approach. In (Ed.), *Proceedings of 3rd ESCA Workshop on Speech Synthesis* (pp. 277–282).

Douglas-Cowie, E., Cowie, R., & Campbell, N. (2003). Speech and emotion. *Speech Communication*, *40*(1-2), 1–3.

Drioli, C., Tisato, G., Cosi, P., & Tesser, F. (2003). Emotions and voice quality: Experiments with sinusoidal modelling. In (Ed.), *Proceedings of VOQUAL ESCA/Workshop* (pp. 127–132).

Dutoit, T., & Leich, H. (1993). MBR-PSOLA: Text-to-speech synthesis based on an MBE re-synthesis of the segments database. *Speech Communication*, *13*(3–4), 167–184.

Ekman, P. (1992). An argument for basic emotions. In N. L. Stein, & [illegible] Oatley (Eds.), *Basic emotions* (pp.). Hove, UK: Lawrence Erlbaum.

Ekman, P., & Friesen, W. (1978). *Facial Action Coding System*. Palo Alto, CA: Consulting Psychologist Press.

Ferrigno, G., & Pedotti, A. (1985). ELITE: A digital dedicated hardware system for movement analysis in real-time TV signal processing. In (Ed.), *IEEE Transactions on Biomedical Engineering, BME-32* (pp. 943–950).

Gobl, C., & Chasaide, A..N. (2003), The role of the voice quality in communicating emotions, mood and attitude. *Speech Communication*, *40*, 189–212.

Hartman, J., & Wernecke, J. (1996). *The VRML handbook*. Addison Wessley.

Huang, X., Hon, H. W., & Acero, A. (2001). *Spoken language processing: A guide to theory, algorithm, and system development*. Prentice Hall.

Kain, A., & Macon. M. W. (2001). Design and evaluation of a voice conversion algorithm based on spectral envelope mapping and residual prediction. In (Ed.), *Proceedings of ICASSP 2001* (Vol. 2, pp. 813–816).

Löfqvist, A. (1990). Speech as audible gestures. In W. J. Hardcastle & A. Marchal (Eds.), *Speech production and speech modeling* (pp. 289–322). Dordrecht, Netherlands: Kluwer Academic.

Magno Caldognetto, E., Cosi, P., Drioli, C., Tisato, G., & Cavicchio, F. (2003). Coproduction of speech and emotions: Visual and acoustic modifications of some phonetic labial targets. In (Ed.), *Proceedings of AVSP 2003, ISCA/Workshop* (pp. 209–214).

Möhler, G. (1988). Describing intonation with a parametric model. In (Ed.), *Proceedings of ICSLP 1988* (pp. 2581–2584).

Möhler, G., & Conkie, A. (1998). Parametric modeling of intonation using vector quantization. In (Ed.), *Proceedings of Third International Workshop on Speech Synthesis*.

Murray, I..R., & Arnott, J. L. (1993). Toward the simulation of emotion in synthetic speech: A review of literature on human vocal emotion. *Journal of the Acoustical Society of America*, *93*(2), 1097–1108.

Pelachaud, C., Magno Caldognetto, E., Zmarich, C., & Cosi. P. (2000). Modelling an Italian talking head. In (Ed.), *Proceedings of AVSP 2001, ISCA/Workshop* (pp. 72–77).

Perin, G. (2000). *Facce parlanti: sviluppo di un modello coarticolatorio labiale per un sistema di sintesi bimodale*. Master's thesis, University of Padova, Italy.

Scherer, K. R. (1986). Vocal affect expression: A review and a model for future research. *Psychological Bullettin*, *99*, 143–165.

Scherer, K. R. (2003). Vocal communication of emotions: A review of research paradigms. *Speech Communication*, *40*(2–3), 227–256.

Schröder, M. (2001). Emotional speech synthesis: A review. In (Ed.), *Proceedings of EUROSPEECH 2001* (Vol. 1, pp. 561–564).

Sluijter, A., van Heuven, V., & Pacilly, J. (1997). Spectral balance as a cue in the perception of linguistic stress. *Journal of the Acoustical Society of America*, *101*(1), 503–513.

Stylianou, Y., Cappé, O., & Moulines, E. (1998). Continuous probabilistic transform for voice conversion. *IEEE Transactions on Speech and Audio Processing*, *6*(2), 131–142.

Taylor, P., Black, A., & Caley, R. (1988). The architecture of the Festival speech synthesis system. In (Ed.), *Proceedings of 3rd ESCA Workshop on Speech Synthesis* (pp. 147–151).

Tesser, F. (2005). *Emotional speech synthesis: From theory to applications.* Doctoral dissertation, DIT—University of Trento, Trento, Italy.

Tesser, F., Cosi, P., Drioli, C., & Tisato, G. (2004). Prosodic data driver modelling of a narrative style in Festival TTS. In (Ed.), *Proceedings of 5th ISCA Speech Synthesis Workshop* [CD-ROM].

Tesser, F., Cosi, P., Drioli, C., & Tisato, G. (2005). Emotional FESTIVAL-MBROLA TTS synthesis. In (Ed.), *Proceedings of INTERSPEECH 2005.* [CD-ROM].