

Mixed feelings: Expression of non-basic emotions in a muscle-based talking head

Irene Albrecht¹, Marc Schröder², Jörg Haber¹, Hans-Peter Seidel¹

¹ MPI Informatik, Saarbrücken, Germany, e-mail: {albrecht, haber.j, hpseidel}@mpi-sb.mpg.de

² DFKI GmbH, Saarbrücken, Germany, e-mail: schroed@dfki.de

Received: date / Revised version: date

Abstract We present an algorithm for generating facial expressions for a continuum of pure and mixed emotions of varying intensity. Based on the observation that in natural interaction among humans, shades of emotion are much more frequently encountered than expressions of basic emotions, a method to generate more than Ekman's six basic emotions (joy, anger, fear, sadness, disgust and surprise) is required. To this end, we have adapted the algorithm proposed by Tsapatsoulis et al. [1] to be applicable to a physics-based facial animation system and a single, integrated emotion model. A physics-based facial animation system was combined with an equally flexible and expressive text-to-speech synthesis system, based upon the same emotion model, to form a talking head capable of expressing non-basic emotions of varying intensities. With a variety of life-like intermediate facial expressions captured as snapshots from the system we demonstrate the appropriateness of our approach.

Key words continuous emotions – emotional speech synthesis – facial animation

1 Introduction

The naturalness of a talking head depends on a considerable number of factors related to the proper integration of visual and audio channel, i.e. of the (audible) synthetic speech and the (visible) facial model. One important factor is the generation of adequate lip movement in synchronisation to speech. Another factor is speech-related nonverbal facial expression, such as raised eyebrows or blinks related to the structure of the spoken utterance. If none of these are present in the synthesised animation, it is perceived as soulless, highly artificial, and plainly boring.

A third important factor for naturalness in a talking head is the expression of emotions [2]. Unfortunately, the “tool-box” for modelling emotions and their expression is not yet very well developed. While attempts are under way to organise the vocabulary and models used for the description of af-

fective states [3–5], much work on the expression of emotion has been limited to simple representations such as basic emotions (see e.g. [6]). Only recently, more flexible emotion representations have started to be explored in the domain of speech synthesis [7] and MPEG4-based facial animation [1].

The present paper follows this line of development in proposing a model for the integrated generation of speech and facial expression using an expressive text-to-speech (TTS) system in combination with a photo-realistic, muscle-based facial animation model. A representation of emotional states combining categorical and dimensional aspects is used for the prediction of vocal and facial expression of non-basic emotions, i.e., of low-intensity and intermediate emotional states.

The paper is structured as follows: we first refer to the relevant background related to speech synthesis, facial animation and emotion representations. After that, the building blocks required for our work are described, and our method for expressing emotions based on a dimensional representation of emotions is presented. Finally, we describe our plans to extend the work.

2 Background

2.1 Text-to-speech systems

Text-to-Speech synthesis [8] is a method for converting written text into audible speech. It consists of a text analysis part, generating a symbolic representation of a spoken utterance including a phonetic transcription of the words, followed by the actual speech synthesis part, in which the symbolic representation is converted into audible speech.

Speech synthesis systems that are to be used in conjunction with facial animation need to provide intermediate processing results such as timing information in addition to the resulting speech. New systems using XML-based internal data representations, such as BOSS [9] and MARY [10], make the output of partial processing results a straightforward task. The XML data can be further analysed by subsequent processing components using standard XML parsers. Emotions influence the speech audio signal to a great extent [11,6,

12]. People are able to identify emotions from the audio signal alone with an accuracy well above chance level [11, 12]. Hence an expressive talking head should not only include emotions in the facial animation, but must also be capable of emotional audible speech.

The modelling of emotion in speech synthesis relies on a number of parameters like, among others, fundamental frequency (F0) level, voice quality, or articulatory precision [13]. Different synthesis techniques provide control over these parameters to very different degrees. Formant synthesis [14], the most parametrisable synthesis technique, has been extensively used in emotional speech synthesis research (e.g., [15]), but is nowadays rarely used because of its low degree of naturalness. Unit selection synthesis [16], relying on the re-sequencing of units from large speech corpora, is the most recent and natural-sounding speech synthesis technique, but lacks the flexibility required for a general-purpose emotional expression tool. Only a small number of expressive categories can be modelled using this technique [17]. Diphone synthesis [18] is a compromise between naturalness and flexibility, and can be used for emotion expression if the diphone units concatenated are recorded in several voice qualities [19].

2.2 Facial animation systems

Approaches to facial animation in general can be divided into the following classes: physics-based [20, 21], example-based [22–24], and parameterised systems [25, 26]. Physically based approaches try to model the anatomical structure of the face as well as the underlying dynamics; facial movement is achieved by muscle contraction. Parameterised systems assign weighted vertices of the face mesh to every parameter. During animations, the vertices are displaced according to the parameter value. Example- or performance-based techniques usually reassemble frames from video footage or track movement of a real person to yield the desired new animation.

Maybe the most challenging application of facial animation is speech synchronisation. Due to coarticulation, designing a facial expression for every phoneme and then interpolating between the expressions according to the input phonemes does not produce realistic results. Several approaches to lip sync have emerged. Procedural methods try to synthesise lip movement for speech from scratch [27, 26]. Example-based systems reorder video frames of a speaking person so that they fit the new speech [22–24] or extract visemes from video and use them to animate 3D head models [28].

In addition, speech is accompanied by nonverbal facial expressions which serve to structure speech and to emphasise important parts of a sentence, thereby facilitating understanding. Eyebrow movement for instance can serve to accentuate important words or parts of a sentence, to structure speech, or to mark a sentence as question. There are several approaches to incorporate nonverbal speech-related facial expressions into animations of speech. Learning-based systems [23, 29] are trained to generate facial animations from speech that

include nonverbal speech-related facial expressions. Script-based systems [30–32] leave synchronisation of nonverbal facial expressions to speech to the user. Rule-based methods generate such expressions from analysis of content, utterance structure, and dialog state [33, 34], from an analysis of the speech signal [35], or from intermediate output of a coupled TTS [36].

Information is often also relayed through emotional expressions. Ekman [37] identified a set of six basic emotional facial expressions that are valid throughout all cultures: joy, anger, fear, disgust, sadness, and surprise. Many facial animation systems can display these universal expressions of emotion. However, the human face is capable of displaying many more emotional expressions, but little research has been conducted in this direction so far, mainly due to the limited availability of data on other expressions. The FacEMOTE system [38] relies on the Laban Movement Analysis of body motion which has been transferred to the face. The method modifies an input facial animation stream to change its expressiveness. The four parameter pairs used to steer the process are direct-indirect, light-strong, sustained-quick, and free-bound. A direct mapping from these parameters to emotions is not provided. Ruttkay et al. [39] arrange the six basic emotions equidistantly on the border of a disc according to similarity. To every point on the disc a facial expression is associated which is computed by linear interpolation between the closest basic emotions. Distance from the circle center describes intensity. In the same paper, the authors present a second method to obtain new expressions based on PCA. However, they did not find the significant principal components to be as intuitive as one might expect. Tsapatsoulis et al. ([1], Section 4.2) have also developed a method to interpolate between affect displays to create new ones. They use a mixture of two emotion models, Whissell's activation-evaluation approach ([40], cf. Section 2.3.2) and Plutchik's emotion wheel [41].

Bui et al. [42] address the problem of combining different channels of facial expression, i.e. lip sync, conversational displays, emotional expressions, etc.

2.3 Emotion representations

Modelling of emotional expression needs to start from a suitable representation of the emotional states to be expressed.

2.3.1 Emotion categories The most straightforward description of emotions is the use of emotion-denoting words, or category labels. Human languages have proven to be extremely powerful in producing labels for emotional states: Lists of emotion-denoting adjectives were compiled that include at least 107 items [40]. Several approaches exist for reducing these to an essential core set, the most used in the literature being basic emotions, a Darwinian concept [3]. Based on the work by Ekman [37], basic emotions are usually used for modelling facial expression of emotions.

2.3.2 Emotion dimensions Many different approaches reported in the psychological literature have led to the proposal of dimensions underlying emotional concepts (see [43] for an overview). Different researchers came to propose two essential dimensions: *activation* (from active/aroused to passive/relaxed) and *evaluation* (from negative/bad to positive/good), sometimes complemented by a third dimension: *power* (from powerful/dominant to weak/submissive). These emotion dimensions are gradual in nature and represent the essential aspects of emotion concepts rather than the fine specifications of individual emotion categories. The names used for these dimensions were selected by the individual researchers *interpreting* their data, and did not arise from the data itself. This explains the large variation found in the literature regarding the names of the dimensions.

One concrete proposal for an emotion dimension model is the activation-evaluation space, proposed by Cowie et al. [44]. In accordance to Plutchik’s emotion wheel [41], they conceived of the space as circular; but they complemented the circle by a disk whose outer bounds represent maximally intense emotions, while its centre (the origin of the two-dimensional space) represents a “neutral”, unemotional state. The further a state is from the centre, the more intense it is, i.e., the radial distance from the centre is a measure of emotion intensity (see Figure 1). In accordance to Whissell [40], emotion categories can be located in that space.

2.3.3 Requirements for a natural emotionally expressive system Databases of naturally occurring emotions [45] show that humans usually express low-intensity rather than full-blown emotions, and complex, mixed emotions rather than mere basic emotions downscaled to a

needs to use an emotion representation capable of representing such states. Emotion dimensions are a suitable representation: they are naturally gradual, and are capable of representing low-intensity as well as high-intensity states. While they do not define the exact properties of an emotional state in the same amount of detail as a category label, they do capture its essential aspects.

2.3.4 Mappings between emotion representations Emotion categories can be *located* in emotion dimension space via rating tests [46]. The mapping from categories to dimensions is therefore a simple task, as long as the coordinates of the emotion category have been determined. The inverse, however, is not possible: as emotion dimensions only capture the most essential aspects of an emotion concept, they provide an under-specified description of an emotional state. For example, the coordinates for anger and disgust may be very close, because the two categories share the same activation/evaluation/power properties. The features distinguishing between the two categories cannot be represented using emotion dimensions, so that the corresponding region in space can only be mapped to “anger-or-disgust” rather than a specific category. One concrete proposal for a mapping from a list of emotion categories to emotion dimensions was brought forward as a working model by the NECA project [47] (see Table 1).

category	activation	evaluation	power
joy	17.3	42.2	12.5
distress	-17.2	-40.1	-52.4
happy-for	17.3	42.2	12.5
gloating	40.0	30.0	30.0
resentment	0.0	-40.0	-20.0
sorry-for	-17.2	-40.1	-52.4
hope	20.0	20.0	-10.0
fear	14.8	-44.4	-79.4
satisfaction	-14.9	33.1	12.2
relief	3.0	33.0	-3.0
fears-confirmed	-30.0	-50.0	-70.0
disappointment	2.4	-24.9	-37.2
pride	30.0	40.0	30.0
admiration	27.0	53.0	17.0
shame	4.6	-26.3	-62.3
reproach	-3.0	-30.0	43.0
liking	-14.9	33.1	12.2
disliking	15.0	-35.0	-10.0
gratitude	20.0	40.0	-30.0
anger	34.0	-35.6	20.0
gratification	-14.9	33.1	12.2
remorse	4.6	-26.3	-62.3
love	1.2	33.3	14.9
hate	60.0	-60.0	30.0

Table 1 Coordinates for a list of emotion categories on the three emotion dimensions activation, evaluation and power, as proposed as a first working model by the NECA project. All scales range from -100 (passive/negative/submissive) via 0 (neutral) to +100 (active/positive/dominant).

It should be kept in mind that, given methodological issues [3] as well as the limited empirical basis in existing studies [40, 46, 45], mappings between the currently existing emotion representations are necessarily imperfect.

3 Building blocks

The talking head system used in this paper (see Figure 2) consists of two main building blocks, a TTS system and a facial animation system.

The system input is plain text, which is passed to the TTS system. Here the text is transformed into a basic XML skeleton that is enriched continuously as the linguistic analysis of the input proceeds, until all necessary information has been assembled and the actual acoustic speech synthesis takes place. A detailed description of the TTS system is provided in Section 3.1.

The final XML structure is not only used by the TTS module, but also by the facial animation system. It extracts the phonemes and their durations for the lip sync, as well as additional linguistic information such as intonation and pauses that serve to generate the nonverbal parts of the animation,

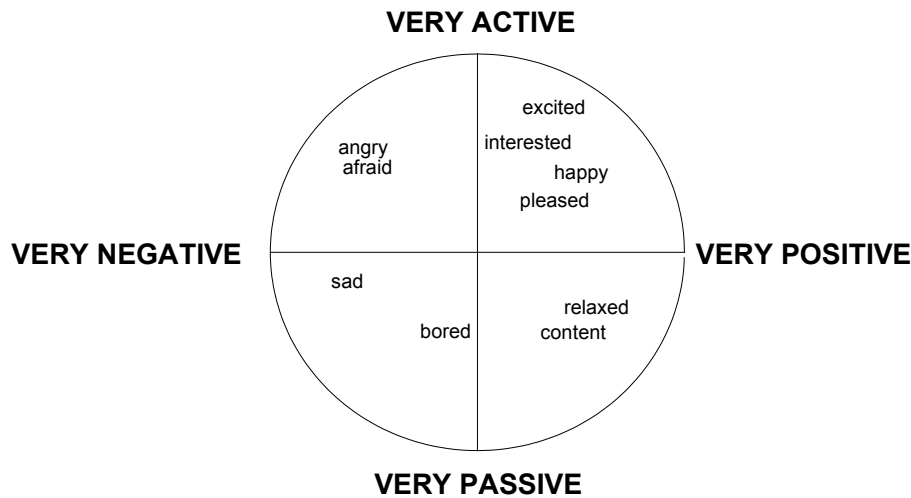


Fig. 1 The two-dimensional, disk-shaped activation-evaluation space proposed by Cowie et al.

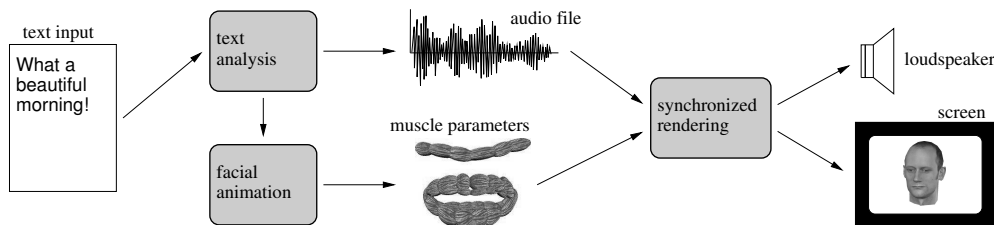


Fig. 2 System overview. The TTS system performs a text analysis on the input text. It generates the audio signal and passes intermediate results like phonemes etc. to the facial animation module.

as e.g. eyebrow raising on accented parts of the speech. The facial animation system will be discussed in Section 3.2.

Connecting both components leads to synthetic audio-visual speech. Section 4 will explain how emotional expressivity is added to this base system.

3.1 The text-to-speech component

The TTS system MARY [10] is a TTS server written in Java¹. It is a very flexible toolkit allowing for easy integration of modules from different origins. An XML-based representation language [48] is used in the system, which makes it easy to access the system's intermediate processing results, as is required for time-alignment with the visual component. The system uses the following modules.

Text normalisation consists of an optional input markup parser converting SSML into MaryXML; a tokeniser; a preprocessing component converting numbers, abbreviations etc. into pronounceable form; a part-of-speech tagger and chunker (local syntactic parser); and an information structure module recognising givenness and contrast based on text structure, optionally using a semantic database.

¹ A publicly accessible web interface can be found at <http://mary.dfki.de>

Phonemisation is performed using a pronunciation lexicon compiled into a finite state transducer, complemented with letter-to-sound rules.

Duration prediction is carried out using a set of rules predicting phoneme durations based on intrinsic phoneme properties and the surrounding context.

Intonation prediction is carried out in two rule-based steps. First, symbolic intonation labels are predicted; second, these symbolic labels are translated into frequency-time targets.

The synthesis module is instantiated using several synthesis engines, among them MBROLA [18].

3.2 The facial animation component

The physics-based facial animation system has been developed following human anatomy [21,49]. It includes skull, jaw, muscles and skin. Skin and skull consist of triangle meshes. The muscles are modeled as a volume, i.e. they have a geometrical shape that bulges during contraction and is elongated if the muscle is being stretched passively. The three components skull, muscles and skin are connected through a mass-spring network. Thence if a muscle contracts, the skin is moved along appropriately and the bulging of the muscle is propagated to the skin, which deforms accordingly. The model includes simple textured geometry for the

eyes, teeth and tongue. In addition to the muscle contraction parameters, parameters for head rotation, jaw rotation, eye and tongue movement are defined. Animations are specified through muscle contraction values and the above additional parameter values, varying over time. They are executed in real-time (40 fps for the simulation, ≈ 100 fps for rendering, on a Pentium 4 1.7 GHz dual processor PC).

3.2.1 Lip sync Lip sync is implemented following the approach by Cohen and Massaro [26]. It considers coarticulation by assigning a so-called dominance function $D_{s,p}(t)$ to every phoneme-facial animation parameter pair (s, p) . $D_{s,p}$ describes the influence of the phoneme s on the facial parameter p over time. All phonemes s have a target value $T_{s,p}$ for every parameter p assigned to them. The target value of one phoneme for all facial animation parameters give the target facial expression of the phoneme, i.e. the corresponding viseme. The trajectory $a_p(t)$ for parameter p over the entire utterance is obtained with the following formula:

$$a_p(t) = \sum_{s=0}^{n-1} \frac{D_{s,p}(t) * T_{s,p}}{D_{s,p}(t)}, \quad (1)$$

where n is the number of segments in the utterance. Since we obtain the phonemes and their durations directly from the speech processing system, visible and audible speech are synchronised inherently.

3.2.2 Nonverbal behaviour Apart from speech, several other channels of communication are open to humans, e.g. facial expressions, gestures, body posture, voice quality, or touching. Such nonverbal behaviour can be categorised into five groups [50]:

- *emblems*: nonverbal acts that have a well known verbal translation, e.g. nodding for "yes"
- *illustrators*: speech-accompanying movement that illustrates what is being said, e.g. eyebrow raising on accented syllables
- *affect displays*: facial expressions of emotions
- *regulators*: behaviour related to turn-taking during a conversation, e.g. looking towards the speaker at the end of questions to prompt for an answer
- *adaptors*: content-free behaviour such as touching or rubbing of oneself.

The animations generated with our system show behaviour pertaining to three of the five categories: illustrators, affect displays, and regulators. We decided to leave out the emblems, since they require some kind of semantic analysis, and the adaptors, since they do not convey any information. Synchronisation of these expressions to the speech at phoneme level is rendered possible by the TTS system which provides data on sentence, phrase², word, syllable, and phoneme boundaries as well as pitch information at phoneme level and

² A phrase is a part of a sentence delimited by grammatical pauses.

information on the type of sentence (e.g. question or informative). In our implementation, we followed the approach given in [36].

4 Emotional expressivity

The expression of emotions by means of a combination of emotion categories and emotion dimensions is explained in detail in the present section.

4.1 Emotional text-to-speech

The emotional text-to-speech synthesis system used in this work is the one developed by Schröder [43]. As it is based on linking emotion dimensions to their acoustic correlates, it integrates well with our approach to visual expression modelling presented in Section 4.2.

The key properties of the emotional speech synthesis system are reported below.

4.1.1 Emotional prosody rules Schröder [43] formulated emotional prosody rules on the basis of a literature review and a database analysis. His literature review brought about the following results. An unambiguous agreement exists concerning the link between the activation dimension and the most frequently measured acoustic parameters: activation is positively correlated with mean F0, mean intensity, and, in most cases, with speech rate. Additional parameters positively correlated with activation are pitch range, "blaring" timbre, high-frequency energy, late intensity peaks, intensity increase during a "sense unit", and the slope of F0 rises between syllable maxima. Higher activation also corresponds to shorter pauses and shorter inter-pause and inter-breath stretches.

The evidence for evaluation and power is less stable. There seems to be a tendency that studies which take only a small number of acoustic parameters into account do not find any acoustic correlates of evaluation and/or power.

The limited evidence regarding the vocal correlates of power indicates that power is basically recognised from the same parameter settings as activation (high tempo, high F0, more high-frequency energy, short or few pauses, large intensity range, steep F0 slope), except that sometimes, high power is correlated with lower F0 instead of higher F0, and power is correlated with vowel duration.

There is even less evidence regarding the acoustic correlates of evaluation. Positive evaluation seems to correspond to a faster speaking rate, less high-frequency energy, low pitch and large pitch range; a "warm" voice quality; and longer vowel durations and the absence of intensity increase within a "sense unit".

In a statistical analysis of the Belfast Naturalistic Emotion Database [45], perceptual ratings of the emotion dimensions activation, evaluation and power were correlated with

	Prosodic parameter	Coefficients		
		Activation	Evaluation	Power
fundamental frequency	pitch	0.3	0.1	-0.1
	pitch-dynamics	0.3%		-0.3%
	range	0.4		
	range-dynamics	1.2%		0.4%
	accent-prominence	0.5%	-0.5%	
	preferred-accent-shape		E<-20: falling -20<E≤40: rising E>40: alternating	
	accent-slope	1%	-0.5%	
	preferred-boundary-type			P≤0: high P>0: low
tempo	rate	0.5%	0.2%	
	number-of-pauses	0.7%		
	pause-duration	-0.2%		
	vowel-duration		0.3%	0.3%
	nasal-duration		0.3%	0.3%
	liquid-duration		0.3%	0.3%
	plosive-duration	0.5%	-0.3%	
	fricative-duration	0.5%	-0.3%	
	volume	0.33%		

Table 2 Emotion dimension prosody rules proposed by Schröder. Values on emotion dimensions range from -100 to 100, with 0 being the “neutral” value. The percentage values are *factors* – see text for details.

acoustic measures (see [43,6] for details). The study replicated the basic patterns of correlations between emotion dimensions and acoustic variables. It was shown that the acoustic correlates of the activation dimension were highly stable, while correlates of evaluation and power were smaller in number and magnitude and showed a high variability between male and female speakers. In addition, the analysis provided numerical linear regression coefficients which were used as a starting point for the formulation of quantified emotion prosody rules.

The effects found in the literature and in the database analysis were formulated in a quantified way (Table 2) and implemented in the MARY TTS system [10].

In Table 2, the columns represent the emotion dimensions, while the rows list all the acoustic parameters for which emotion effects are modelled. The numeric data fields represent the linear coefficients quantifying the effect of the given emotion dimension on the acoustic parameter, i.e. the change from the neutral default value. As an example, the value 0.5% linking *activation* to *rate* means that for an activation level of +50, rate increases by +25%, while for an activation level of -30, rate decreases by -15%.

4.1.2 Implementation The MARY system (see Section 3.1) was used as the platform for the implementation of the emotional prosody rules specified in Table 2. This system was

most suitable for the task because of the high degree of flexibility and control over the various processing steps, which arises from the use of the system-internal representation language MaryXML.

A major design feature in the technical realisation of the emotional speech synthesis system was that the acoustic effects of emotions should be specified in one single module. This module adds appropriate MaryXML annotations to the text which are then realised by the respective modules within the MARY system. As a consequence, all of the parameters are global in the sense that they will be applied to all enclosed text. This approach is considered the most transparent, as the link between emotions and their acoustic realisations is not hidden in various processing components, and the easiest to maintain and adapt, as all rules are contained in one document.

4.1.3 System evaluation The appropriateness of the generated emotional prosody and voice quality was assessed in a perception test. Due to the multimodal nature of any emotional utterance, this appropriateness was addressed in terms of coherence with other channels expressing the emotion, notably verbal content and the situational context.

Verbal situation descriptions with known activation and evaluation ratings were used as reference material. For each of the emotional states defined by the situation descriptions,

emotional speech prosody settings were calculated, and each of the texts was synthesised with each of the prosodic settings in a factorial design. In a listening test, subjects rated each stimulus according to the question: “How well does the sound of the voice fit with the content of the text?”

The results confirmed the hypothesis that the prosodic configurations succeed best at conveying the activation dimension. Moreover, the appropriateness of a prosodic configuration for a given emotional state was shown to depend on the *degree* of similarity between the emotional state intended to be expressed by the prosody and that in the textual situation description. In agreement with previous findings for human speech, the evaluation dimension was found to be more difficult to convey through the prosody. In summary, the speech synthesis system succeeded in expressing the activation dimension (the speaker “arousal”), but not the evaluation dimension. See [43] for a full account of the experiment.

4.2 Intermediate facial expressions

The human face is capable of displaying many more emotional expressions than just those of the six universal emotions joy, anger, fear, disgust, sadness, and surprise. However, little visual data is available on expressions of other emotions, and modelling them is hard, since differences between them are often subtle. Hence Tsapatsoulis et al. [1] have developed a method to interpolate between affect displays to create new ones. We present their original work before we describe our own model derived from theirs.

Tsapatsoulis et al. [1] modelled emotions using a combination of two emotion models: Plutchik [41] ordered 142 emotion words according to their similarity. He found that they can be arranged around a circle, the so-called *emotion wheel*. Hence the relative position of each emotion can be described by an angle [41, p.170]. This model does not consider activation or intensity, its goal was to establish similarity. In [40], Whissell describes the second model, a rating of emotion words according to their co-ordinates on the activation and evaluation dimensions. Tsapatsoulis et al. use the angles in the emotion wheel as a measure of similarity, while they use Whissell’s activation values to describe emotion intensity.

Tsapatsoulis et al. use the MPEG-4 facial animation parameters (FAPs) animate their head model. They identified eight fundamental emotions: acceptance, fear, surprise, sadness, disgust, anger, anticipation, and joy. These are the starting points for the interpolation. The facial expression e corresponding to an emotion E is described by the following parameters:

- the activation value a_E of the emotion E
- its angle on the emotion wheel ω_E
- the FAPs involved in forming the expression F_e
- for each contributing FAP $f \in F_e$ the range of variations of its value $R_e(f)$ associated with the expression.

There are two different ways to generate new expressions: if the new emotion E_n is very similar to the fundamental emotion E , i.e. if their facial expressions differ mainly in strength of muscle contraction, then the new expression e_n can be computed from the expression e in the following way:

$$F_{e_n} = F_e \quad (2)$$

$$R_{e_n}(f) = \frac{a_{E_n}}{a_E} \cdot R_e(f) \quad \forall f \in F_e \quad (3)$$

If the new emotion E_n does not clearly belong to a fundamental category, its facial expression is computed by interpolation between the shifted expressions of the two emotions E_1 and E_2 that are closest to E_n on the emotion wheel. For an interval $I = [i_1, i_2]$, let

$$\sigma(I) = \begin{cases} 1, & i_1 \leq i_2 \\ -1, & i_1 > i_2 \end{cases} \quad (4)$$

define the sign σ of I . Let $c(I)$ be the center of interval I and $s(I)$ be its length. Then e_n is determined by

$$F_{e_n} = F_{e_1} \cup F_{e_2} \quad (5)$$

$$R'_{e_1}(f) = \frac{a_{E_n}}{a_{E_1}} \cdot R_{e_1}(f)$$

$$R'_{e_2}(f) = \frac{a_{E_n}}{a_{E_2}} \cdot R_{e_2}(f)$$

$$c(R_{e_n}(f)) = \left(\frac{\omega_{E_n} - \omega_{E_1}}{\omega_{E_2} - \omega_{E_1}} \cdot c(R'_{e_2}(f)) + \frac{\omega_{E_2} - \omega_{E_n}}{\omega_{E_2} - \omega_{E_1}} \cdot c(R'_{e_1}(f)) \right)$$

$$s(R_{e_n}(f)) = \left(\frac{\omega_{E_n} - \omega_{E_1}}{\omega_{E_2} - \omega_{E_1}} \cdot s(R'_{e_2}(f)) + \frac{\omega_{E_2} - \omega_{E_n}}{\omega_{E_2} - \omega_{E_1}} \cdot s(R'_{e_1}(f)) \right)$$

$$R_{e_n}(f) = \left[c(R_{e_n}(f)) - \frac{1}{2} \cdot s(R_{e_n}(f)), c(R_{e_n}(f)) + \frac{1}{2} \cdot s(R_{e_n}(f)) \right] \\ \forall f \in F_{e_1} \cap F_{e_2} : \sigma(R_{e_1}(f)) = \sigma(R_{e_2}(f)) \quad (6)$$

$$R_{e_n}(f) = \frac{a_{E_n}}{a_{E_1}} \cdot R_{e_1}(f) \cap \frac{a_{E_n}}{a_{E_2}} \cdot R_{e_2}(f) \\ \forall f \in F_{e_1} \cap F_{e_2} : \sigma(R_{e_1}(f)) \neq \sigma(R_{e_2}(f)) \quad (7)$$

$$R_{e_n}(f) = \frac{a_{E_n}}{2 \cdot a_{E_1}} \cdot R_{e_1}(f) \quad \forall f \in F_{e_1} \setminus F_{e_2} \quad (8)$$

$$R_{e_n}(f) = \frac{a_{E_n}}{2 \cdot a_{E_2}} \cdot R_{e_2}(f) \quad \forall f \in F_{e_2} \setminus F_{e_1} \quad (9)$$

If now $R_{e_n}(f) = \emptyset$, then set $F_{e_n} := F_{e_n} \setminus \{f\}$.

In case FAP f is involved in the facial expressions of both generating emotions and its variation intervals have for both emotions the same sign (Equation (6)), i.e. it describes movement into the same direction, the variation intervals of the generating expressions e_1 and e_2 are first shifted, so that the resulting expressions have the same activation as E_n . Then from the centers and lengths of the shifted intervals the interval $F_{e_n}(f)$ can be computed through linear interpolation between the emotion wheel angles. If the variation intervals

of e_1 and e_2 have different signs (Equation (8)), the interval of the new expression is the intersection of the original ones. If f is present only in one generating expression (Equations (8) and (9)), say, e_1 , then its variation interval is averaged with the interval of the neutral face e_0 , for which $a_{E_0} = 0$ and $R_{e_0}(f) = [0] \quad \forall f \in F_{e_0}$.

We have modified the approach to work with our physics-based model. Instead of combining the data from the Whissell and Plutchik studies, as Tsapatsoulis et al. did, we use the set of emotion words with associated co-ordinates on the three dimensions activation, evaluation and power, as proposed by the NECA project [47] (see Table 1). In this first version of the system, we only use the first two dimensions from this Table (see Section 6).

We use Cowie et al.'s disk-shaped activation-evaluation space (see Figure 1) as our model of emotion dimensions. It appears natural to describe the states in the activation-evaluation space by means of polar coordinates, using angular orientation ω and radial distance from the centre r . Here again, the angle ω describes similarity. In contrast to Tsapatsoulis et al., we consider radial distance from the centre of the activation-evaluation space to be a better indicator of emotional intensity than activation (consider the case of despair, which would have high intensity but low activation), and therefore use this radial distance r rather than the activation level a for normalising the archetypal states' intensities to the intermediate state's intensity in our equations.

As our "basic" emotions, we use the closest correlates to the six Ekmanian emotions (joy, anger, fear, disgust, sadness, and surprise) that we can find in Table 1: joy, anger, fear, hate, sorry-for, and surprise (as a state with 100% activation, and 0% on evaluation and power). We are aware that these are crude approximations, which should be taken as illustrating the idea rather than as a final truth.

Since our animations are based mostly on muscle contractions instead of MPEG-4 FAPs, we had to adapt the approach to also work with muscles. We defined our expressions through single muscle contraction values v . They can be uniformly scaled by a number between 0 and 1 to achieve different intensities of the expressed emotion, but we leave it to the animator to decide how small the scaling value can be so that the resulting expression is still perceived as the same emotion. As a consequence, we have no means of deciding for a given muscle whether to use Equation (6) or Equation (8). Hence we identified facial muscles that operate in a roughly antagonistic fashion (see Table 3).

Let M_e be the set of muscles involved in expression e of emotion E and $v_e(m)$ the contraction value of m . For simplicity, the animation parameters of eyes, head and jaw rotation, and tongue are included in M_e . This leads to the following modified algorithm for the case where the new emotion E_n is similar to a fundamental emotion E :

$$M_{e_n} = M_e \quad (10)$$

$$v_{e_n}(m) = \frac{r_{E_n}}{r_E} \cdot v_e(m) \quad \forall m \in M_e. \quad (11)$$

Zygomaticus major left	Depressor anguli oris left
Zygomaticus major right	Depressor anguli oris right
Orbicularis oris	Risorius left, Risorius right
Mentalis	Depressor labii inferioris left, Depressor labii inferioris right

Table 3 Facial muscles operating in a roughly antagonistic fashion.

Since several muscles can be antagonistic to others, e.g. the *orbicularis oris* to both the *risorius left* and the *risorius right*, we define for every muscle m the set of its antagonists as $A_-(m)$ and the set of muscles that share these antagonists as $A_+(m)$. For $m = risorius left$ for instance, $A_+(m) = \{risorius left, risorius right\}$ and $A_-(m) = \{orbicularis oris\}$. If the facial expression for E_n is computed from two fundamental expressions E_1 and E_2 , we get:

$$M_{e_n} = M_{e_1} \cup M_{e_2} \quad (12)$$

$$v'_{e_1}(m) = \frac{r_{E_n}}{r_{E_1}} \cdot v_{e_1}(m)$$

$$v'_{e_2}(m) = \frac{r_{E_n}}{r_{E_2}} \cdot v_{e_2}(m)$$

$$v_{e_n}(m) = \left(\frac{\omega_{E_n} - \omega_{E_1}}{\omega_{E_2} - \omega_{E_1}} \cdot v'_{e_2}(m) + \frac{\omega_{E_2} - \omega_{E_n}}{\omega_{E_2} - \omega_{E_1}} \cdot v'_{e_1}(m) \right) \quad \forall m \in M_{e_n} : A_+(m) \cup A_-(m) = \emptyset \quad (13)$$

$$S_+ = \sum_{m' \in A_+(m)} \left(\frac{r_{E_n}}{r_{E_1}} \cdot v_{e_1}(m') + \frac{r_{E_n}}{r_{E_2}} \cdot v_{e_2}(m') \right)$$

$$S_- = \sum_{m' \in A_-(m)} \left(\frac{r_{E_n}}{r_{E_1}} \cdot v_{e_1}(m') + \frac{r_{E_n}}{r_{E_2}} \cdot v_{e_2}(m') \right)$$

$$v_{e_n}(m) = \begin{cases} 0, & \text{if } S_+ \leq S_- \\ (S_+ - S_-) \cdot \frac{1}{S_+} \cdot \left(\frac{r_{E_n}}{r_{E_1}} \cdot v_{e_1}(m) + \frac{r_{E_n}}{r_{E_2}} \cdot v_{e_2}(m) \right), & \text{else} \end{cases} \quad \forall m \in M_{e_n} : A_+(m) \cup A_-(m) \neq \emptyset \quad (14)$$

Equation (12) is analogue to Equation (5). The differences in Equation (13) stem from the use of a single value instead of an interval. This obviates the need to compute the center and length of the interval. Instead we can scale and interpolate the contraction values directly. Since they do not describe a direction but a value, no conflict arises. The main difference lies in Equation (14). $S_+(m)$ and $S_-(m)$ are the summed, scaled contraction values of all muscles in the same and different antagonistic class, respectively. The overall scaled contraction for all muscles in the same and the antagonistic class of m is $S_+(m) - S_-(m)$. This is distributed to the individual muscles of the set with the stronger scaled overall contraction according to their contribution to that value. If we assign contraction values $v_{e_1}(m) = 0 \quad \forall m \in$

$M_{e_2} \setminus M_{e_1}$ and $v_{e_2}(m) = 0 \quad \forall m \in M_{e_1} \setminus M_{e_2}$, this obviates the need for Equations (8) and (9).

In Figure 3, the new expressions *anxiety* and *panic fear* have been generated as a scaled version of *fear*, following the method in Equations (10) and (11). In the examples in Figures 4 and 5, new expressions (center of each row), have been generated from the fundamental expressions to the left and right as described in Equations (12) to (14).

5 Conclusions

We have presented a flexible approach to generating non-basic, mixed emotional states in the facial expressions of an anatomically based talking head. This has been achieved by modifying the work of Tsapatsoulis et al. [1], aimed at an MPEG4-based face model, to a physics-based facial animation system. In extension to their work, we have used data in a single emotion model, the activation-evaluation space [44], for indicating both emotion quality and intensity. As a result, our system is able to generate emotional facial expressions of various intensities, and to show mixed emotions by a gradual blending of facial configurations of basic emotions.

We have combined the facial animation model with an emotional text-to-speech synthesis system which is also based on emotion dimensions. In combining these two components, we have created photo-realistic animations of a talking head capable of expressing a continuum of shades of emotion.

6 Future work

There are several possible directions for future work. The most exciting is the planned extension to 3D emotion space, i.e. to not only consider activation and evaluation, but also power to allow for a more fine-grained model. Since emotions are arranged inside a sphere in this space, we propose to project the individual emotions onto the sphere's surface and to interpolate between expressions on the surface of the sphere. This would permit interpolation between more than two expressions. The resulting expression is then projected back to the desired distance from the origin.

A next step should be the evaluation of the system. This could be done in a similar manner as described in [43].

Since emotion categories are more intuitive for most people than positions in activation-evaluation-power space, we require coordinates for more emotion words to enhance the user-friendliness of the system.

Another pressing issue is the extension of the system to include several different emotions in a single utterance, allowing for transitions between emotions over time.

Adapting the frequency and strength of the nonverbal speech-related facial expressions to the current emotion could enhance the realism of the animations, e.g. look downwards more often and in general show movement with less amplitude when sad. As an additional visible effect of emotion the artificial face should be capable of blushing. Frequency and

intensity of breathing is also an indicator of the emotion currently felt.

Acknowledgements Part of this research is supported by the EC Project HUMAINE (IST-507422).

References

1. N. Tsapatsoulis, A. Raousaiou, S. Kollias, R. Cowie, and E. Douglas-Cowie. Emotion Recognition and Synthesis Based on MPEG-4 FAPs. In *MPEG-4 Facial Animation - The standard, implementations, applications*, pages 141–167. John Wiley & Sons, Hillsdale, NJ, USA, 2002.
2. E. André, L. Dybkyær, W. Minker, and P. Heisterkamp, editors. *Proceedings of the Tutorial and Research Workshop on Affective Dialogue Systems (ADS04)*, volume 3068 of *Lecture Notes in Artificial Intelligence*, Kloster Irsee, Germany, June 2004. Springer.
3. R. Cowie and R. Cornelius. Describing the emotional states that are expressed in speech. *Speech Communication Special Issue on Speech and Emotion*, 40(1–2):5–32, 2003.
4. K. Scherer. Psychological models of emotion. In J. Borod, editor, *The Neuropsychology of Emotion*, pages 137–162. Oxford University Press, New York, 2000.
5. The humane network portal. <http://emotion-research.net>.
6. M. Schröder, R. Cowie, E. Douglas-Cowie, M. Westerdijk, and S. Gielen. Acoustic Correlates of Emotion Dimensions in View of Speech Synthesis. In *Proc. Eurospeech'01*, volume 1, pages 87–90, 2001.
7. M. Schröder. Dimensional emotion representation as a basis for speech synthesis with non-extreme emotions. In *Proc. Workshop on Affective Dialogue Systems*, pages 209–220, Kloster Irsee, Germany, 2004.
8. Th. Dutoit. *An Introduction to Text-to-Speech Synthesis*. Kluwer Academic Publishers, Dordrecht, 1997.
9. E. Klabbers, K. Stöber, R. Veldhuis, P. Wagner, and S. Stefan Breuer. Speech synthesis development made easy: The Bonn Open Synthesis System. In *Proceedings of Eurospeech 2001*, pages 521–524, Aalborg, Denmark, 2001.
10. M. Schröder and J. Trouvain. The German text-to-speech synthesis system MARY: A tool for research, development and teaching. *International Journal of Speech Technology*, 6:365–377, 2003.
11. R. Banse and K. Scherer. Acoustic Profiles in Vocal Emotion Expression. *J. Personality and Social Psychology*, 70(3):614–636, 1996.
12. L. Yang. Prosody as expression of emotion. In Ch Cavé, editor, *Oralité et gestualité, Proc. ORAGE 2001*, pages 209–212, 2001.
13. M. Schröder. Emotional speech synthesis: A review. In *Proceedings of Eurospeech 2001*, volume 1, pages 561–564, Aalborg, Denmark, 2001.
14. J. Allen, S. Hunnicutt, and D. H. Klatt. *From Text to Speech: The MITalk System*. Cambridge University Press, Cambridge, UK, 1987.
15. J. Cahn. The generation of affect in synthesized speech. *Journal of the American Voice I/O Society*, 8:1–19, July 1990.
16. A. W. Black and N. Campbell. Optimising selection of units from speech databases for concatenative synthesis. In *Proceedings of Eurospeech 1995*, volume 1, pages 581–584, Madrid, Spain, 1995.

**anxiety**

$$\begin{aligned} a &= 8 \\ e &= -24.1 \\ r &= 25.3 \\ \omega &= 288.4 \end{aligned}$$

fear

$$\begin{aligned} a &= 8 \\ e &= -24.1 \\ r &= 46.8 \\ \omega &= 288.4 \end{aligned}$$

panic fear

$$\begin{aligned} a &= 8 \\ e &= -24.1 \\ r &= 63.4 \\ \omega &= 288.4 \end{aligned}$$

Fig. 3 Anxiety and panic belong to the same fundamental class as fear, but differ in intensity. Therefore their facial expressions can be generated from fear by scaling with the ratio of the radii. The angle on the emotion disc is kept fixed for both new expressions, while the radii are varied, thereby yielding new values for activation and evaluation.

17. W. Johnson, S. Narayanan, R. Whitney, R. Das, M. Bulut, and C. LaBore. Limited domain synthesis of expressive military speech for animated characters. In *Proceedings of the 7th International Conference on Spoken Language Processing*, Denver, Colorado, USA, 2002.
18. Th. Dutoit, V. Pagel, N. Pierret, F. Bataille, and O. van der Vrecken. The MBROLA project: Towards a set of high quality speech synthesizers free of use for non commercial purposes. In *Proceedings of the 4th International Conference of Spoken Language Processing*, pages 1393–1396, Philadelphia, USA, 1996.
19. M. Schröder and M. Grice. Expressing vocal effort in concatenative synthesis. In *Proceedings of the 15th International Conference of Phonetic Sciences*, Barcelona, Spain, 2003. to appear.
20. Y. Lee, D. Terzopoulos, and K. Waters. Realistic face modeling for animation. In *Proc. SIGGRAPH'95*, pages 55–62, 1995.
21. K. Kähler, J. Haber, and H.-P. Seidel. Geometry-based Muscle Modeling for Facial Animation. In *Proc. Graphics Interface 2001*, pages 37–46, June 2001.
22. Ch. Bregler, M. Covell, and M. Slaney. Video Rewrite: Driving Visual Speech with Audio. In *Proceedings of SIGGRAPH '97*, pages 353–360. ACM Press, 1997.
23. M. Brand. Voice Puppetry. In *Proc. SIGGRAPH '99*, pages 21–28, 1999.
24. T. Ezzat, G. Geiger, and T. Poggio. Trainable Videorealistic Speech Animation. In *Proc. SIGGRAPH'02*, pages 388–398, 2002.
25. F. Parke. *A Parametric model for Human Faces*. PhD thesis, University of Utah, Salt Lake City, UT, 1974.
26. M. Cohen and D. Massaro. Modeling Coarticulation in Synthetic Visual Speech. In N. Magnenat-Thalmann and D. Thalmann, editors, *Models and Techniques in Computer Animation*, pages 139–156. 1993.
27. C. Pelachaud, N. Badler, and M. Steedman. Linguistic Issues in Facial Animation. In N. Magnenat-Thalmann and D. Thalmann, editors, *Computer Animation'91*. 1991.
28. G. Kalberer, P. Müller, and L. Van Gool. A Visual Speech Generator. In *Proc. Videometrics VII 2003, IS&SPIE*, pages 173–183, 2003.
29. S. Lee, J. Badler, and N. Badler. Eyes Alive. In *Proc. SIGGRAPH'02*, pages 637–644, 2002.
30. A. Pearce, B. Wyvill, G. Wyvill, and D. Hill. Speech and Expression: A Computer Solution to Face Animation. In *Proc. Graphics Interface '86*, pages 136–140, May 1986.
31. H. Ip and C. Chan. Script-Based Facial Gesture and Speech Animation Using a NURBS Based Face Model. *Computers & Graphics*, 20(6):881–891, November 1996.
32. P. Kalra, A. Mangili, N. Magnenat-Thalmann, and D. Thalmann. SMILE: A Multilayered Facial Animation System. In *Proc. IFIP WG 5.10, Tokyo, Japan*, pages 189–198, 1991.
33. J. Cassell, C. Pelachaud, N. Badler, M. Steedman, B. Achorn, T. Becket, B. Douville, S. Prevost, and M. Stone. ANIMATED CONVERSATION: Rule-based Generation of Facial Expression, Gesture & Spoken Intonation for Multiple Conversational Agents. In *Proc. SIGGRAPH '94*, pages 413–420, 1994.
34. C. Pelachaud, N. Badler, and M. Steedman. Generating Facial Expressions for Speech. *Cognitive Science*, 20(1):1–46, 1996.
35. I. Albrecht, J. Haber, and H.-P. Seidel. Automatic Generation of Non-Verbal Facial Expressions from Speech. In *Proc. CGI 2002*, pages 283–293, 2002.
36. I. Albrecht, J. Haber, K. Kähler, M. Schröder, and H.-P. Seidel. "May I talk to you? :-)" – Facial Animation from Text. In *Proc. Pacific Graphics 2002*, pages 77–86, 2002.
37. P. Ekman and D. Keltner. Universal Facial Expressions of Emotion: An Old Controversy and New Findings. In U. Segerströle and P. Molnár, editors, *Nonverbal Communication: Where Nature Meets Culture*, pages 27–46. Lawrence Erlbaum Associates, Inc., Mahwah, NJ, 1997.
38. M. Byun and N. Badler. FacEMOTE: qualitative parametric modifiers for facial animations. In *Proc. SCA'02*, pages 65–71, 2002.

**sadness**

$a = -17.2$
 $e = -40.1$
 $r = 43.6$
 $\omega = 246.8$

**remorse**

$a = 4.6$
 $e = -26.3$
 $r = 26.7$
 $\omega = 279.9$

**fear**

$a = 14.8$
 $e = -44.4$
 $r = 46.8$
 $\omega = 288.4$

**joy**

$a = 17.3$
 $e = 42.2$
 $r = 45.6$
 $\omega = 67.7$

**gratification**

$a = -14.9$
 $e = 33.1$
 $r = 36.3$
 $\omega = 114.2$

**sadness**

$a = -17.2$
 $e = -40.1$
 $r = 43.6$
 $\omega = 246.8$

Fig. 4 The emotional expression in the middle has been obtained from those at the left and right using the blending algorithm. The radius r and the angle on the emotion disc ω determine the influence of each generating expression and hence the degree of similarity to the new one. The coordinates in emotion space have been obtained from the NECA data.

39. Z. Ruttkey, H. Noot, and P. ten Hagen. Emotion Disc and Emotion Squares: tools to explore the facial expression space. *Computer Graphics Forum*, 22(1):49–53, 2003.
40. C. Whissell. The Dictionary of Affect in Language. In R. Plutchik and H. Kellerman, editors, *Emotion: Theory, Research, and Experience*, volume 4: The Measurement of Emotions, chapter 5, pages 113–131. Academic Press, Inc., San Diego, CA, 1989.
41. R. Plutchik. *Emotions: A Psychoevolutionary Synthesis*. Harper & Row, New York, 1980.
42. T. Bui, D. Heylen, and A. Nijholt. Combination of facial movements on a 3D talking head. In *Proc. CGI'04*, pages 284–291, 2004.
43. M. Schröder. *Speech and emotion research: An overview of research frameworks and a dimensional approach to emotional speech synthesis*. PhD thesis, Vol. 7 of Phonus, Research Report of the Institute of Phonetics, Saarland University, 2004.
44. R. Cowie, E. Douglas-Cowie, S. Savvidou, E. McMahon, M. Sawey, and M. Schröder. 'FEELTRACE': An instrument for recording perceived emotion in real time. In *Proceedings of the ISCA Workshop on Speech and Emotion*, pages 19–24,

**fear**

$a = 14.8$
 $e = -44.4$
 $r = 46.8$
 $\omega = 288.4$

**upset**

$a = 25$
 $e = -50$
 $r = 55.9$
 $\omega = 296.6$

**anger**

$a = 34$
 $e = -35.6$
 $r = 49.2$
 $\omega = 313.7$

**surprise**

$a = 100$
 $e = 0$
 $r = 100$
 $\omega = 0$

**positive surprise**

$a = 80$
 $e = 15$
 $r = 81.4$
 $\omega = 10.6$

**joy**

$a = 17.3$
 $e = 42.2$
 $r = 45.6$
 $\omega = 67.7$

**disgust**

$a = 60$
 $e = -60$
 $r = 84.9$
 $\omega = -45$

**negative surprise**

$a = 80$
 $e = -15$
 $r = 81.4$
 $\omega = -10.6$

**surprise**

$a = 100$
 $e = 0$
 $r = 100$
 $\omega = 0$

Fig. 5 The emotional expression in the middle has been obtained from those at the left and right using the blending algorithm. The radius r and the angle on the emotion disc ω determine the influence of each generating expression and hence the degree of similarity to the new one. The first example is a not too active, but rather negative emotion, while the second one could be pleasant surprise, and the last one unpleasant surprise.

- Northern Ireland, 2000.
45. E. Douglas-Cowie, N. Campbell, R. Cowie, and P. Roach. Emotional speech: Towards a new generation of databases. *Speech Communication Special Issue Speech and Emotion*, 40(1–2):33–60, 2003.
 46. R. Cowie, E. Douglas-Cowie, B. Appolloni, J. Taylor, A. Romano, and W. Fellenz. What a neural net needs to know about emotion words. In N. Mastorakis, editor, *Computational Intelligence and Applications*, pages 109–114. World Scientific & Engineering Society Press, 1999.
 47. B. Krenn, H. Pirker, M. Grice, P. Piwek, K. van Deemter, M. Schröder, M. Klesen, and E. Gstrein. Generation of multimodal dialogue for net environments. In *Proceedings of Konvens*, Saarbrücken, Germany, 2002.
 48. M. Schröder and S. Breuer. XML representation languages as a way of interconnecting TTS modules. In *Proc. ICSLP*, Jeju, Korea, 2004.
 49. K. Kähler, J. Haber, and H.-P. Seidel. Head Shop: Generating animated head models with anatomical structure. In *Proc. SCA'02*, pages 55–64, 2002.
 50. P. Ekman and W. Wallace. The repertoire of nonverbal behavior: Categories, origins, usage, and coding. *Semiotica*, 1:49–98, 1969.