EAVA: A 3D Emotive Audio-Visual Avatar

Hao Tang, Yun Fu, Jilin Tu, Thomas S. Huang, and Mark Hasegawa-Johnson
University of Illinois at Urbana-Champaign*
405 North Mathews Avenue, Urbana, IL 61801, USA

Abstract

Emotive audio-visual avatars have the potential of significantly improving the quality of Human-Computer Interaction (HCI). In this paper, the various technical approaches of a novel framework leading to a text-driven 3D Emotive Audio-Visual Avatar (EAVA) are proposed. Primary work is focused on 3D face modeling, realistic emotional facial expression animation, emotive speech synthesis, and the co-articulation of speech gestures (i.e., lip movements due to speech production) and facial expressions. Experimental results clearly indicate that a certain degree of naturalness and expressiveness has been achieved by EAVA in both audio and visual aspects. Promising potential improvements can be expected by incorporating various data-driven statistical learning models into the framework.

1. Introduction

Avatars are virtual agents for Human-Computer Interaction (HCI). They introduce the presence of an individual to capture attention, mediate conversational cues, and communicate emotional affect, personality and identity in many application scenarios. They provide realistic facial expressions and natural sounding speech to help the understanding of the content and intent of a message. If computers are embodied and represented by avatars, the quality of HCI can be significantly improved. Customized avatars may also serve as assistant personal clones that enable individuals who have speech and hearing problems to participate in daily spoken communications.

The use of avatars is now emerging in the marketplace, while the understanding of human communication has not yet advanced to the point where it is possible to make avatars that demonstrate interactions with realistic emotive speech and emotional facial expressions. The key research issues to enable avatars to communicate subtle or complex emotions are how to model 3D faces, how to animate realistic emotional facial expressions, how to synthesize natural sounding emotive speech, and how to combine speech gestures (i.e. lip movements due to speech production) and facial expressions both naturally and realistically.

We have been conducting basic research leading to methodologies and algorithms for constructing a text-driven 3D Emotive Audio-Visual Avatar (EAVA). This is an interdisciplinary research that is related to speech, computer graphics and computer vision. The success of this project is believed to advance significantly the state-of-the-art of HCI.

2. Related work

2.1. Face modeling and animation

3D face modeling has been an active research topic for computer graphics and computer vision [2, 1]. People have used interactive tools to design geometric 3D face models under the guidance of prior knowledge. As laser-based range scanners such as the Cyberware^{TM} scanner have become commercially available, people have been able to measure the 3D geometry of human faces so that geometric 3D face models can be constructed [2]. Alternatively, some researchers have proposed to build 3D face models from 2D images using computer vision techniques [14, 2].

A geometric face model defines the 3D geometry of a static face. By deforming the geometric face model spatially and temporally, we can obtain different facial expressions. The free-form interpolation model is one of the most popular facial deformation models in the literature [14]. A set of control points are defined on the 3D geometry of the face model and facial deformation is achieved through proper displacements of these control points. The displacements of the rest of the facial geometry are obtained by a certain interpolation scheme. To approximate the facial deformation space, linear subspace methods have been proposed. One example is the Facial Action Coding System (FACS) [15] which describes arbitrary facial deformation as a linear combination of Action Units (AUs). Hong et al [1] applied Principal Component Analysis (PCA) to real facial motion captured data and derived a few bases called Motion Units (MUs). Any facial deformation can then be approxi-

*This work was supported in part by the U.S. Government VACE program and in part by the National Science Foundation Grant CCF 04-26627.
Email: {haotang2,yunfu2,jilintu,huang,hasegawa}@ifp.uiuc.edu.

mated by a linear combination of the MUs. Compared with AUs, MUs are automatically derived from data. Therefore, labor-intensive manual work can be avoided. In addition, MUs can yield smaller reconstruction error than AUs.

2.2. Emotive speech synthesis

Attempts to add emotions to synthetic speech have existed for over a decade. Cahn [6] and Murray [7] both used a commercial formant synthesizer (i.e., DECTalk) to generate emotive speech based on the emotion-specific acoustic parameter settings that they derived from the literature. Burkhardt [8] also employed a formant synthesizer to synthesize emotive speech. Instead of deriving the acoustic profiles from the literature, he conducted perception-oriented experiments to find optimal values for the various parameters. Although partial success was achieved, reduced naturalness was reported due to the imperfect rules inherent with formant synthesis.

Several later undertakings and smaller studies have made use of the diphone synthesis approach. Schroder [4] used a diphone synthesizer to model a continuum of intensity-varying emotional states under the emotion dimension framework. He searched for acoustic correlates of emotions by analysis of a carefully labeled emotional speech database and established a mapping of the points in the emotion space to their acoustic correlates. These acoustic correlates are then used to tune the prosodic parameters in the diphone synthesizer to generate emotive speech. The synthesized speech is observed to convey only non-extreme emotions and thus cannot be considered successful if it is used alone. Complimentary channels such as facial expressions are required to be used together in order for the user to fully comprehend the emotional state.

One might want to model a few "well-defined" emotion categories as close as possible, and it seems unit selection synthesis based on emotion-specific speech database can be a suitable choice for this purpose [9, 10]. Pitrelli et. al. at IBM [11] recorded a large database of neutral speech containing 11 hours of data as well as several relatively smaller databases of expressive speech (1 hour of data for each database). Instead of selecting units from one particular database at run time, they blended all the databases together and selected units from the blended database according to some criterion. They assumed that many of the segments comprising a sentence, spoken expressively, could come from the neutral database.

2.3. Co-articulation of speech and expressions

On the human faces, it is unclear how speech gestures are dynamically combined with facial expressions to ensure natural, realistic and coherent appearances. Some researchers [12, 13] have used techniques like Independent Component Analysis (ICA) and PCA to separate and

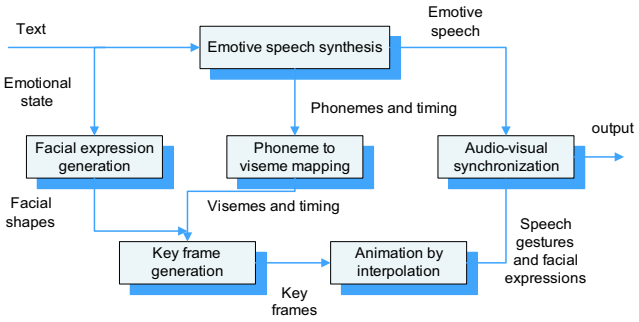

Figure 1. System framework of EAVA.

model the viseme and expression spaces. However, neither is based on the theoretical grounding for determining the interactions and relative contributions of the respective sources that together cause facial deformation.

3. Framework and approaches

In the research toward building a text-driven 3D emotive audio-visual avatar, we design a system framework as illustrated in Figure 1. In this framework, text is converted into emotive speech by emotive speech synthesis. Simultaneously, a phoneme sequence with timing is produced. The phoneme sequence is mapped into a viseme sequence that defines the speech gestures. The emotional states, extracted from the text (with emotion markers), decide the facial expressions, which will be combined with speech gestures synchronously and naturally. The key frame technique [14] is used to animate the sequence. Primary research is focused on emotional facial expression animation, emotive speech synthesis, and the co-articulation of speech gestures and facial expressions.

3.1. 3D face modeling

Our previous work on 3D face modeling (iFace) and animation [14, 1] lays a partial foundation of this research. iFace provides a research platform for 3D face modeling and animation. It takes as input the Cyberware^{TM} scanner data of a person's face and fits the data with a generic head model. The output is a customized geometric 3D face model ready for animation.

3.2. Emotional expression animation

We use the geometric 3D face model described earlier to animate facial expressions. On the face model, a control model is defined. The control model consists of 101 vertices and 164 triangles which cover the whole face region and divide it into local patches. By dragging the vertices of the control model, one can deform the 3D face model to arbitrary shapes. A set of basic facial shapes corresponding to different fullblown emotional facial expressions are designed and parameterized as the displacements

Table 1. Phonemes (Ph), examples (Ex), and visemes (Vi).

Ph	Ex	Vi	Ph	Ex	Vi	Ph	Ex	Vi
i	beet	IY	I	bit	AX	S	assure	SH
E	bet	AY	{	bat	AE	h	hope	AY
r=	above	AX	u	boot	W	v	vine	F
U	book	AX	V	above	AX	D	thy	TH
O	caught	AO	A	father	AA	z	resign	T
@	butter	AX	EI	bay	NE	Z	azure	SH
AI	bye	AY	OI	boy	T	tS	church	SH
aU	about	AY	@U	boat	OW	dZ	judge	SH
p	pan	M	t	tan	T	l	lent	LL
k	can	T	b	ban	M	r	rent	R
d	dan	T	g	gander	T	j	yes	T
m	me	M	n	knee	T	w	went	W
N	sing	AY	f	fine	F	_	(silent)	NE
T	thigh	TH	s	sign	T			

of all the vertices of the face model from their initial values corresponding to neutral state. Let those displacements be $\Delta\vec{V}_i, i = 1, 2, ..., N$, where N is the number of vertices of the face model. The vertices of the deformed face model are given by $\vec{V}_{0i} + \rho \times \Delta\vec{V}_i, i = 1, 2, ..., N$, where $\vec{V}_{0i}$ are the vertices of the neutral face and $0 \leq \rho \leq 1$ is the magnitude coefficient representing the intensity of the facial expression. $\rho = 1$ corresponds to the fullblown emotional facial expression and $\rho = 0$ corresponds to the neutral expression.

Speech gestures are implemented via visemes. We have defined 17 visemes, each of which corresponds to one or more of the 40 phonemes in American English. Table 1 lists all the phonemes and the corresponding visemes that they are mapped to. Phoneme-to-viseme mapping can thus be done by simple table look-up.

3.3. Emotive TTS synthesis

To synthesize emotive speech, we use a diphone synthesizer and adopt a rule-based approach for prosody modification. The ability of diphone synthesis to control the prosodic parameters for speech synthesis is a very nice property that enables the generation of emotional affect in synthetic speech by explicitly modeling the emotion. In order to produce high-quality synthetic speech, a TTS synthesis system usually consists of two main components: a Natural Language Processing (NLP) module and Digital Signal Processing (DSP) module. The NLP module performs necessary text analysis and prosody prediction procedures to convert orthographical text into proper phonetic transcription (i.e., phonemes) together with the desired intonation and rhythm (i.e., prosody). There are numerous methods that have been proposed and implemented for the NLP module. The DSP module takes as input the phonetic transcription and prosodic description which are the output of the NLP module, and transforms them into a speech signal. Likewise, various DSP techniques have been proposed for this purpose [3].

Figure 2 illustrates a general framework of emotive TTS synthesis using a diphone synthesizer. In this framework, an emotion transformer is inserted into the functional diagram of a general TTS synthesis system and plays a role that bridges the NLP and DSP modules. The NLP module generates phonemes and prosody corresponding to the neutral state (i.e., neutral prosody). The emotion transformer is aimed at transforming the neutral prosody into the desired emotive prosody as well as transforming the voice quality of the synthetic speech. The phonemes and emotive prosody are then passed on to the DSP module to synthesize emotive speech using a diphone synthesizer. While many aspects of this framework, especially the techniques, methods, and algorithms engaged in the NLP and DSP modules, have been thoroughly investigated in the literature [3], certain aspects of this framework must be singled out in this paper.

3.3.1 Prosody transformation by difference approach

Prosody transformation of the emotion transformer is achieved through a difference approach, which aims at finding the differences of emotional states with respect to the neutral state. The emotion transformer either maintains a set of prosody manipulation rules defining the variations of prosodic parameters between the neutral prosody and the emotive one, or trains a set of statistical prosody prediction models that predict these parameter variations given the features representing a particular textual context. The basic idea is formulated as $\Delta p = p_{\mathrm{e}} - p_{\mathrm{n}}$, where Δp denotes the parameter differences, p_{n} denotes the neutral prosodic parameters, and p_{e} denotes the emotive prosodic parameters.

At synthesis time, the rules or the prediction outputs of the models will be applied to the neutral prosody, obtained via the NLP module. The emotive prosody can thus be obtained by summing up the neutral prosody and the predicted variations. The basic idea is formulated as $\widehat{p_{\mathrm{e}}} = \widehat{p_{\mathrm{n}}} + \widehat{\Delta p}$, where $\widehat{\Delta p}$ denotes predicted parameter difference, $\widehat{p_{\mathrm{n}}}$ denotes the predicted neutral prosodic parameters, and $\widehat{p_{\mathrm{e}}}$ denotes the predicted emotive prosodic parameters.

There are obvious advantages of using a difference approach over the use of a full approach that aims at finding the emotive prosody directly. One advantage is that the differences of the prosodic parameters between the neutral prosody and the emotive one have far smaller dynamic ranges than the prosodic parameters themselves and therefore the difference approach requires far less data to train the models than the full approach (e.g., 15 minutes versus several hours). Another advantage is that the difference approach makes possible the derivation of the prosody manipulation rules which are otherwise impossible to be obtained.

Practically, prosody transformation is achieved through applying a set of prosody manipulation rules. The global prosody settings that we use are given in Table 2, which are similar to those used in [19], and the prosody manipulation rules are given in Table 3.

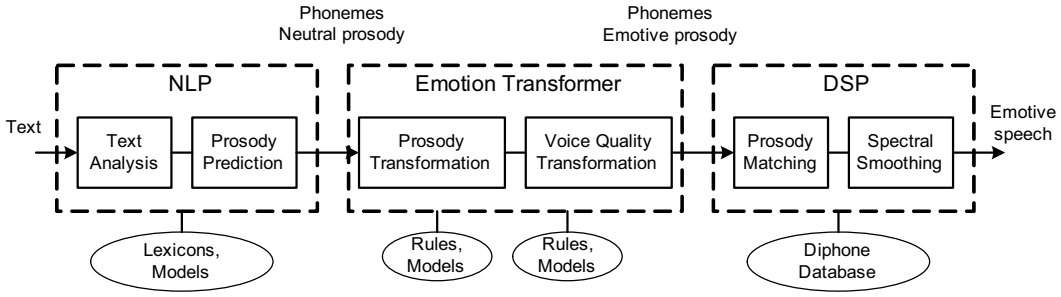

Figure 2. General framework of emotive TTS synthesis using a diphone synthesizer.

Table 2. Global prosody settings. f0 is the fundamental frequency of speech perceived as the pitch.

Parameter	Description
f0 mean	mean value of f0 contour, in Hz
f0 range	difference between max. and min. of f0 contour
f0 variability	degree of variation of f0 contour
f0 contour shape	shape of f0 contour (rising, falling)
speaking rate	duration of speech

3.3.2 Voice quality transformation

Some studies indicate that in addition to prosody transformation, voice quality transformation is important for synthesizing emotive speech and may be indispensable for some emotion categories [18]. In diphone synthesis, it is particularly not easy to control voice quality, as it is very difficult to modify the voice quality of a diphone database. However, one partial remedy of this is to record separate diphone databases with different vocal efforts [5]. During synthesis, the system switches among different voice-quality diphone databases and select the diphone units from the appropriate database. Another low-cost partial remedy is to use jitters to simulate voice quality transformation. Jitters are fast fluctuations of the f_0 contour. Thus, adding jitters is essentially equivalent to adding noise to the f_0 contour. By jitter simulation we can observe voice quality change in synthetic speech to a certain degree.

3.3.3 Prosody matching of diphone units

Diphone synthesis is performed by concatenating small segments of recorded speech called diphones together to create a speech signal. These segments of speech are recorded by a human speaker in a monotonic pitch to aid the concatenation process, which is carried out by employing one of the signal processing techniques including the residual excited linear prediction (RELP) technique, the time-domain pitch synchronous overlap-add (TD-PSOLA) technique, and the multi-band resynthesis pitch synchronous overlap-add (MBR-PSOLA) technique [3]. The prosody of the diphone units is also forced to match that of the desired specification at this point.

It is widely admitted that diphone synthesis inevitably introduces artifacts to synthetic speech due to prosody modification. However, in order to generate emotive speech, the diphone units are subject to extreme prosody modification. In order to alleviate this difficulty, we can record the diphone database with multiple instances for every diphone. The same diphone will be recorded monotonically but at multiple pitch levels and with multiple durations. During synthesis, given a target prosody specification, we can then choose the diphone unit whose prosody parameters are the closest to those of the target. In this way we believe that we can obtain higher-quality emotive speech by reducing the amount of signal processing required for prosody matching in a diphone synthesizer.

Table 3. Prosody modification rules.

Manipulation	Description
f0 mean	shift the f0 mean by multiplying all f0 values by a factor
f0 range	widen or narrow the f0 range by shifting each f0 value by a percentile of its distance to the f0 mean
f0 variability	increase or decrease the global f0 variability by adjusting the f0 range of each syllable
f0 contour shape	reshape the f0 contour on either phrase or syllable level by making it rise or fall controlled by a gradient
speaking rate	lengthen or shorten the duration of speech at various levels: global phrase level; syllable level based on stress type (unstressed, word-stressed, phrase-stressed); sound level based on category (vowels, nasals, stops)

3.4. A rule-based emotive TTS synthesis system based on Festival-MBROLA architecture

In order to demonstrate the validity of our proposed general framework of emotive TTS synthesis, we have implemented a rule-based emotive TTS synthesis system using a diphone synthesizer under the guidance of the framework. The diagram is shown in Figure 3. This system has been built based on the Festival-MBROLA architecture. The Festival system [16] is used to perform text analysis and predict the neutral prosody targets from text. Emotive speech synthesis is then done by manipulation of prosodic parameters (primarily f0 and duration) using a difference approach and simulation of voice quality variation. First, a set of rules describing how prosodic parameters deviate from their initial values corresponding to neutral state to new values corre-

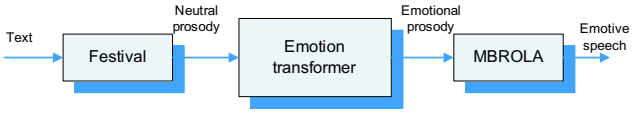

Figure 3. Diagram for emotive speech synthesis.

sponding to a specific emotional state are determined. Next, these rules are applied to the neutral prosody obtained from Festival to generate the desired emotional prosody (a sequence of phonemes, along with the associated prosodic parameters), which is then passed to MBROLA [17], a free-for-non-commercial-use diphone synthesizer developed by the TCTS Lab of the Faculte Polytechnique de Mons (Belgium), to create emotive speech.

The core part of the system, i.e., the emotion transformer, consists of a set of rules that define the various manipulations of the prosody and voice quality of speech. More specifically, the emotion transformer takes as input the phonemes and neutral prosody as well as other related information that are the output of the NLP module of Festival, applies appropriate parameterized rules for pitch transformation, duration transformation and voice quality transformation, and outputs an intermediate file containing the same phonemes and emotive prosody.

3.5. Co-articulation of speech and expressions

Due to the complex cooperation of facial muscular activities, the movements of the lower face are commonly controlled by both speech gestures and facial expressions. In situations where there is only neutral facial expression, the movements of the lower face can be assumed to be dominated by speech gestures. However, problems occur when emotional facial expressions are taken into account as in EAVA. Due to the highly dynamic nature of speech gestures and facial expressions, the exact interactions and contributions of these two sources that control facial movements are unknown. We propose a linear combination approach by assuming that speech gestures and facial expressions contribute equally (or weighted equally) to facial deformations. This assumption however can lead to faulty results (e.g., the mouth will never close while laughing and speaking). An ad hoc method may be used to remedy this problem. We first use the linear combination approach and then identify the top co-articulation difficulties between speech gestures and facial expressions. We can "fix" these difficulties by manually creating "emotive visemes". This approach, though not systematic, can work pretty well in practice.

4. Experiments

The system framework and approaches described in the previous section have resulted in a fully functional system of text-driven 3D emotive audio-visual avatar, EAVA. We have obtained preliminary but promising experiment results for rendering neutral as well as several basic emotions: happy, joyful, sad, angry, and fearful. Figure 4 displays two examples of the results: The animation sequence of emotional facial expressions along with the associated waveforms of synthetic emotive speech generated for happy and sad emotional states. That is, the avatar says "This is happy voice" and "This is sad voice" respectively.

Informal evaluations within our group show that for emotive speech, negative emotions (e.g., sad, angry, fear) are more successfully synthesized than positive emotions (happy, joyful). For some emotions such as joyful and fear, artifacts are easily observed in the synthesized speech due to extreme prosody modification by signal processing. While the synthetic emotive speech alone cannot be always recognized by a listener, it can be easily distinguished when combined with compatible emotional facial expressions.

We have designed and conducted three subjective listening experiments on the results generated by EAVA. Experiment 1 uses six speech files corresponding to the six distinct emotions, synthesized by the system using the same semantically neutral sentence (i.e., a number). Experiment 2 also uses six speech files, but each of the files was synthesized using a semantically meaningful sentence appropriate for the emotion that it carries. Experiment 3 incorporates the visual channel based on experiment 1. That is, each speech file was provided with an emotive talking head showing emotional facial expressions and lip movements consistent and synchronized with the speech. Experiment 4 incorporates the visual channel based on experiment 2. 20 subjects (10 males and 10 females) were asked to listen to the speech files (with the help of other channels such as verbal content and facial expressions if possible), determine the emotion for each speech file (by forced-choice), and rate his or her decision with a confidence score (1: not sure, 2: likely, 3: very likely, 4: sure, 5: very sure). The results are shown in Table 4. In the table, recognition rate is the ratio of the correct choices, and average score is the average confidence score computed for the correct choices.

It is shown in the results of the subjective listening experiments that our system has achieved a certain degree of expressiveness despite of the relatively not perfect performance of the prosody prediction model of Festival. As can be clearly seen, negative emotions (e.g, sad, afraid, angry) are more successfully synthesized than positive emotions (e.g, happy). This observation is consistent with what has been found in the literature. In addition, we found that the emotional states "happy" and "joyful" are often mistaken for each other. So are "afraid" and "angry", mostly due to the close similarity of the respective pair of emotions. By incorporating other channels that convey emotions such as verbal content and facial expressions, the perception of emotional affect in synthetic speech can be significantly improved.

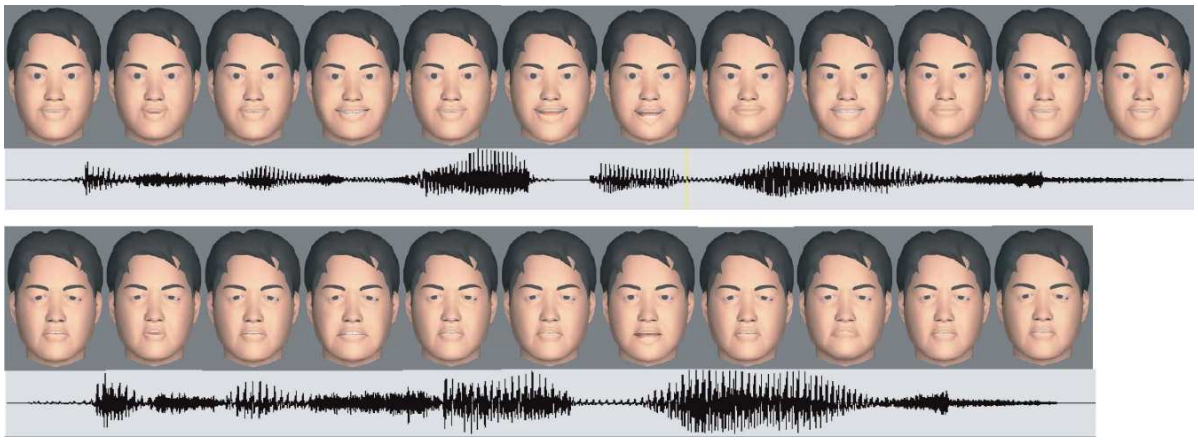

Figure 4. Examples of animation of emotive audio-visual avatar with associated synthetic emotive speech waveform. (Top): The avatar says "This is happy voice." (Bottom): The avatar says "This is sad voice."

Table 4. Results of subjective listening experiments. $R \longleftrightarrow$Recognition Rate. $S \longleftrightarrow$Average Confidence Score. Experiment 1 to 4 are speech only, speech+verbal content, speech+facial expression, and speech+verbal content+facial expression, respectively.

Emotion	Experiment 1		Experiment 2		Experiment 3		Experiment 4	
	$R(\%)$	S	$R(\%)$	S	$R(\%)$	S	$R(\%)$	S
neutral	54.54	2.83	100	4.36	90.90	4.10	100	4.81
happy	18.18	2.50	63.63	4.57	63.63	4.43	54.54	4.50
joyful	45.45	2.80	63.63	4.86	63.63	4.29	45.45	5.00
sad	36.36	2.75	81.81	4.67	100	4.18	100	4.81
angry	18.18	2.00	90.90	4.30	90.90	4.00	100	4.63
afraid	45.45	2.40	81.81	4.89	72.72	4.00	81.81	5.00

5. Conclusion and future work

This paper reported the various aspects of research leading to a text-driven 3D emotive audio-visual avatar, EAVA. Primary work is focused on 3D face modeling, realistic emotional facial expression animation, emotive speech synthesis, and the co-articulation of speech gestures and facial expressions. We have obtained preliminary but promising experiment results as well as built a fully functional system. Particularly, our ongoing work is to pursue data-driven methodologies in resolving the different but related aspects of this research. Instead of using prosody manipulation rules, potential improvements to the current EAVA system can be made within the presented general framework by using statistical prosody prediction models trained with diphone databases of different vocal efforts and multi-pitch multi-duration.

References

[1] P.Y. Hong, Z. Wen, and T.S. Huang, "Real-time Speech-driven Face Animation with Expressions Using Neural Networks," *IEEE Trans. on NN*, vol.13, no.4, pp. 916-927, 2002.

[2] Z. Wen and T.S. Huang, *3D Face Processing: Modeling, Analysis and Synthesis (1^{st} edition)*, Springer, 2004.

[3] T. Dutoit, *An Introduction to Text-to-Speech Synthesis*, Kluwer Academic Publishers, Dordrecht, 1997.

[4] M. Schroder, *Speech and emotion research: an overview of research frameworks and a dimensional approach to emotional speech synthesis* (Ph.D thesis), vol. 7 of Phonus, TR of the Institute of Phonetics, Saarland University, 2004.

[5] M. Schroder, and M. Grice, "Expressing vocal effort in concatenative synthesis," *Proc. of the 15th Int'l Conf. of Phonetic*, 2003.

[6] J.E. Cahn, *Generating expression in synthesized speech* (Master thesis), MIT Media Lab, 1989.

[7] I.R. Murray, *Simulating emotion in synthetic speech* (Ph.D thesis), University of Dundee, UK, 1989.

[8] F. Burkhardt, *Simulation emotionaler sprechweise mit sprachsyntheseverfahren* (Ph.D thesis), TU Berlin, 2000.

[9] A. Iida, *Corpus-based speech synthesis with emotion* (Ph.D thesis), University of Keio, Japan, 2002.

[10] G. Hofer, *Emotional Speech Synthesis* (Master thesis), University of Edinburgh, 2004.

[11] J.F. Pitrelli, R. Bakis, E.M. Eide, R. Fernandez, W. Hamza, and M.A. Picheny, "The IBM expressive text-to-speech synthesis system for American English," *IEEE Trans. on Audio, Speech and Language Processing*, vol.14, no. 4, pp. 1099-1108, 2006.

[12] Y. Cao, P. Faloutsos, and F. Pighin, "Expressive speech-driven facial animation," *ACM Trans. on Graph.*, vol. 24, no. 4, 2005.

[13] S. Kshirsagar, T. Molet, N. Magnenat-Thalmann, "Principal Components of Expressive Speech Animation," *Proc. of Int'l conf. on Computer Graphics*, pp 38-44, 2001.

[14] P. Hong, Z. Wen, T.S. Huang, "iFace: a 3D synthetic talkingn face," *Int'l J. of Image and Graphics*, vol. 1, no.1, pp. 19-26, 2001.

[15] P. Ekman, and W.V. Friesen, *The Facial Action Coding System*, Palo Alto, Calif.: Consult. Psychological Press, 1977.

[16] The Festival Project, http://www.cstr.ed.ac.uk/projects/festival/.

[17] The MBROLA Project, http://mambo.ucsc.edu/psl/mbrola/.

[18] Montero, J.M., Gutirrez-Arriola, J., Cols, J., Macłas, J., Enrłquez, and E., Pardo, J., "Development of an emotional speech synthesiser in spanish," *Proc.of Eurospeech'99*, 1999.

[19] F. Burkhardt, "Emofilt: the Simulation of Emotional Speech by Prosody-Transformation," *Proc. INTERSPEECH*, pp. 509-512, 2005.