Speech-Driven Cartoon Animation with Emotions

Yan Li, Feng Yu* Ying-Qing Xu, Eric Chang, Heung-Yeung Shum
Microsoft Research China: 3F Sigma Center, No.49 Zhichun Road, Beijing 100080, China
{yli,yqxu,echang,hshum}@microsoft.com

ABSTRACT

In this paper, we present a cartoon face animation system for multimedia HCI applications. We animate face cartoons not only from input speech, but also based on emotions derived from speech signal. Using a corpus of over 700 utterances from different speakers, we have trained SVMs (support vector machines) to recognize four categories of emotions: neutral, happiness, anger and sadness. Given each input speech phrase, we identify its emotion content as a mixture of all four emotions, rather than classifying it into a single emotion. Then, facial expressions are generated from the recovered emotion for each phrase, by morphing different cartoon templates that correspond to various emotions. To ensure smooth transitions in the animation, we apply low-pass filtering to the recovered (and possibly jumpy) emotion sequence. Moreover, lip-syncing is applied to produce the lip movement from speech, by recovering a statistical audio-visual mapping. Experimental results demonstrate that cartoon animation sequences generated by our system are of good and convincing quality.

Keywords

Multimedia IICI, cartoon animation, speech emotion recognition, lip-syncing

1. Introduct ion

Facial animation has been an active research topic for computer graphics and multimedia. In particular, facial animation can be used as an effective communication channel for human-computer interface (I-ICI). For instance, MPEG4 has a face and body animation committee that defines how the human face and body should be modeled and transmitted. More and more commercial products have been developed recently in the forms of talking heads, virtual friends, face email, talking shows, virtual announcers, and so on [1, 2, 3, 4]. By animating human faces in cartoon forms, we can add some artistic styles [10, 24].

Directly animating human faces is challenging because there are so many parameters to be controlled for realistic facial expressions. To alleviate such difficulties for animators, speech-driven animation techniques (e.g., [8, 9]) have been proposed to learn the mapping between the voice and the facial motion, and then to drive the facial animation from the speech signals. However, most previous face animation systems, to the best of our knowledge, have not considered explicitly the emotions existing in the speech signal. While speaking the same content, people may have significantly different facial expressions depending on whether they are happy or sad. Emotions must be considered in facial animation systems [11, 22].

Although much work [15, 19, 20] has been done for emotion analysis in the field of speech recognition, there are some difficulties in analyzing emotions from speech. It is not clear how many categories of emotions should be considered for facial expressions and facial animation. Recognition of emotions has been shown to be much more difficult than recognition of phonemes or words. For example, recent work reports that human subjects can categorize five different emotions (normal, happy, angry, sad, and afraid) with the average accuracy of 63.5% [21]. In comparison, in the speech recognition field, character accuracy rates for large vocabulary continuous speech recognition tasks can be well over 90% [12].

In our work, we simplify emotion analysis in two ways. First, because our purpose of analyzing emotions is to animate cartoons, we can afford to classify the speech into only four different but representative emotions: neutral, sadness, happiness, and anger. Second, instead of classifying the utterance into one of these four emotional categories, we model the emotion in the utterance as a mixture of all emotions.

For each of the emotions, we generate a series of cartoon templates corresponding to the intensity of the emotion. For example, five cartoon templates are used to represent the emotion from neutral to happy. Given the recovered emotions from speech signal, facial expressions can be generated by morphing between these cartoon templates. A real-time lip-syncing algorithm is also developed in our system to make the cartoon animation more lively and believ-

*Visiting from Computer Science Department, Tsinghua University, China.

MM'01, Sept. 30-Oct. 5, 2001, Ottawa, Canada.
Copyright2001 ACM 1-58113-394-4/01/0009...$5.00

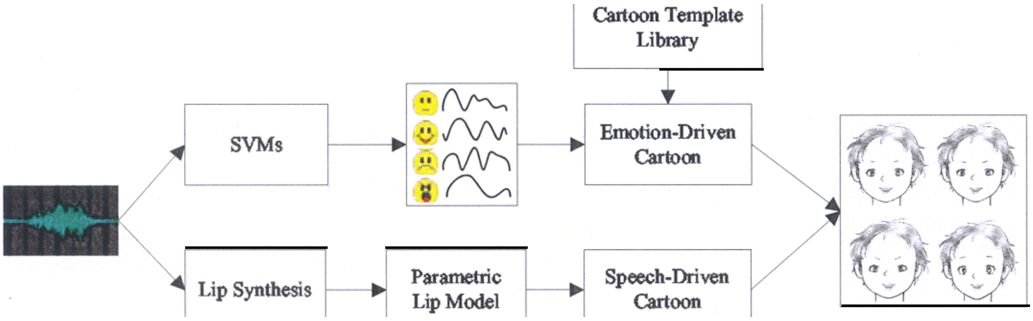

Figure 1: System overview. The system consists of two parts: emotion-driven cartoon animation and speech-driven cartoon animation

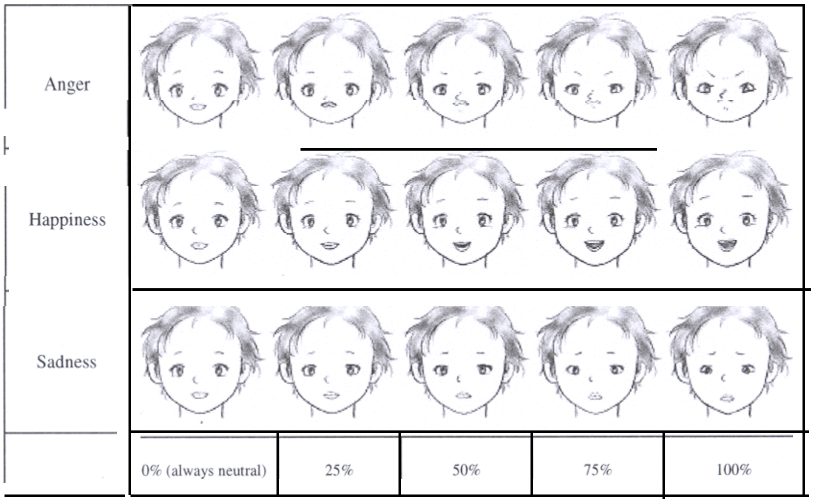

Figure 2: Four different levels (25%, 50%, 75% and 100%) of the cartoon face templates for three emotions (anger, happiness and sadness). The first levels of these three emotions (or 0% intensity) correspond to neutral emotion.

able. What is novel about our lip-syncing method, compared with conventional phoneme-viseme mapping, is that it is based on low-level acoustic speech signals. Therefore, it is language-independent, and applicable to multi-lingual translation agents.

The remainder of this paper is organized as follows. Section 2 outlines the framework of the emotion- and speech-driven cartoon animation system. Section 3 describes the emotion recognition algorithm. In Section 4, we discuss how the recognized results are used to drive the cartoon model. Section 5 explains our real-time lip-syncing algorithm. Some experimental results will be shown in Section 6. We conclude in Section 7 with discussion and future work.

2. System Overview

Figure 1 shows an overview of our system. Our system consists of two parts: speech-driven cartoon animation, and emotion-driven cartoon animation. First, a speech emotion recognition system is developed. To train the emotion classifier, we collected a corpus containing over 1000 utterances from different speakers appearing in a variety of videos and movie clips. Support vector machines (SVM's) are trained to classify each utterance into four categories: neutral, happiness, sadness and anger. In the recognition process, we segment the input utterance into phrases automatically. The trained SVM's are then applied to each phrase to yield a continuous emotion curve for every emotion.

The emotion-driven cartoon animation begins with a small number of hand-drawn sample sketches of annotated faces (templates), each of which corresponds to a specific facial expression (Figure 2). We have an artist draw four cartoons for each emotion. An appropriate cartoon template is selected from the library according to the recognized emotion intensity. Facial animation is generated by morphing between different cartoon templates.

The key to speech-driven cartoon animation is a real-time lip-syncing algorithm. Instead of conventional phoneme-viseme mapping (e.g., [9, 18, 25]), our algorithm uses the acoustic feature vector (e.g. MFCC used in speech recognition) as system input. The advantage of using the acoustic

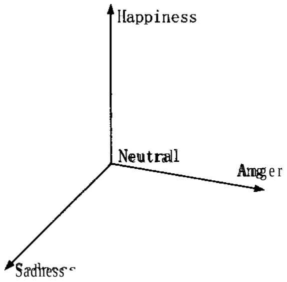

Figure 3: An emotion space where the origin is the neutral expression. Any motion is considered as a mixture of three emotions (sadness, anger and happiness).

feature vector is that different languages (e.g., Chinese and English) do not require training different models. In the training stage, all the lip configurations (training data) are clustered into a set of templates automatically. For each template, a generic model is trained to map from acoustic feature vector to visual feature vector in real time. In the synthesis stage, we employ Bayesian estimation to obtain the most probable lip configuration for a given acoustic vector.

Finally, the cartoon animation sequence is synthesized by combining morphed cartoon templates with the synchronized lip configurations.

3. Emotion Analysis

As an important human behavior for conveying psychological information, emotion has been studied for centuries. Many aspects of emotions have been studied for emotion-antecedent appraisal, emotion induction, physiological reaction and expression of emotion (facial and vocal), and emotional behavior on autonomous agents [5].

3.1 Emotion Modeling and Evaluation

From the psychological point of view, human emotions are often described as some subjective perceptions such as happiness, sadness, surprise, etc. From the experimental and computational perspective, however, we require emotion to be classified explicitly by some parameters. More specifically, the intensity of an emotion can be measured by a value, which can be further used to drive the cartoon face model.

Four emotions are recognized in our system: neutral, happiness, sadness and anger. Although they cannot encompass all emotions present in speech, it has been shown that humans can recognize them from utterances with less ambiguity. In addition, these four emotions are representative enough for a cartoon animation system, especially for applications such as talking shows, avatars and teleconferencing.

However, it is difficult to accurately recognize these four emotions. For the purpose of cartoon animation, rather than assigning a single emotion to an input utterance, we assume that the input utterance is a mixture of several emotions with different intensities. As shown in Figure 3, we model human emotion as a point in an emotion space. Neutral emotion is placed at the origin of this space because it can be used to describe the relative lack of other emotions. Three axes represent the intensity of sadness, happiness and anger, respectively. Note that the three axes do not need to be orthogonal in the emotion space. Modeling human emotion in such a three-dimensional continuous space is meant to be a simplification for recognition and animation, as we later show.

Table 1: Number of training samples for each emotion

Neutral	Anger	Happiness	Sadness
242	215	136	128

3.2 Training Data

We have collected more than 1000 movie clips and extracted the acoustic data from the clips as training data, where the average clip length is about 10 seconds. The utterances are carefully selected from speakers with different genders and ages. A trained experimenter classified the utterances into four categories: neutral, happiness, sadness and anger.

In a separate step, we had five adult subjects listen to each of the utterances and determine its emotion category. An utterance is considered to be valid as training data only when all the subjects agree unanimously with the initial classification. Table 1 shows the number of training samples for each category in our experiment.

3.3 Feature Extraction

We extract the acoustic features as used in [15]. For each utterance, a 16 dimensional vector with the following elements is calculated:

- Statistics related to rhythm: Speaking rate, Average length between voiced regions, Number of maxima / number of (minima + maxima), Number of up slopes / number of slopes.
- Statistics on the smoothed pitch signal: Min, Max, Median, Standard deviation.
- Statistics on the derivative of the smoothed pitch: Min, Max, Median, Standard deviation.
- Statistics over the individual voiced parts: Mean min, Mean max.
- Statistics over the individual slopes: Mean positive derivative, Mean negative derivative.

3.4 Continuous Speech Emotion Recognition

In our system, we use continuous input speech to drive a cartoon face model. We build multidimensional discriminators to classify each utterance into its proper category. Because the emotion recognizer is trained from a set of utterances and the emotion can only be stable within a short period, we need to segment the continuous speech into phrases. For each phrase, we apply the recognition algorithm that produces the proportions of each emotion in that phrase. The segmentation algorithm is based on the following assumptions:

- The intensity of an emotion is unchanging within a given phrase.
- The emotion is neutral when the phrase is silent or the classifier cannot determine the emotion category of that phrase.

Because the acoustic features are extracted from pitch, the phrases can be separated at regions where the pitch value appears to be zero for a period ***T***. ***T*** can be used as a parameter to tune the smoothness of the emotion. For instance, ***T*** should be small when speech emotion changes drastically. On the other hand, a large ***T*** is used to generate smooth emotion transitions. In our experiments, we set T=40ms.

For analysis, we process each phrase as a whole. Because different features extracted from the audio data are correlated, nonlinear classifiers need to be designed. Existing emotion recognition algorithms are mostly based on K-nearest neighbor (KNN) or neural networks. In our system, we use support vector machines (SVMs) that can be trained quickly without considering the correlations between different features. Moreover, training data are separated from the recognition process after we obtain the classifiers. We have implemented SVMs using Gaussian kernels:

$$K(x_i, x_j) = e^{-\|x_i - x_j\|^2/2\sigma^2}. \tag{1}$$

Four ***l-v-r*** (one-versus-rest) SVMs are trained to distinguish one class from the three others After the training process, we obtain four two-class classifiers $S_i(v)$ $(i = 1\ .4)$, where v is the feature vector. The recognition process can be represented by

$$\begin{cases} v \in \text{class} & (\text{if S,(v)} > 0) \\ v \text{ is rejected by classifier } i & (\text{if S,(v)} \leq 0) \end{cases} \tag{2}$$

Because the discriminative plane tends to approach the class with more training samples, the performance of SVM will be influenced by the numbers of positive and negative training samples. For each discriminator, we select training samples randomly from data sets and choose the numbers of positive and negative training samples that are nearly equal. For example, in order to train an SVM that recognizes the emotion "anger", we select 150 anger-labeled utterances as positive samples. The negative samples meanwhile consist of 50 happiness-labeled, 50 neutral-labeled and 50 sadness-labeled samples. Table 2 shows the performance of the SVMs and their corresponding training samples. The performance of the classifier is evaluated by judging whether the automatic classifications yield the same label as the subjective ones. In our experiment, the most clearly recognizable emotion is neutral, and the recognition performance of happiness and anger is relatively poor. However, what we have found is that even for human listeners, it is very easy to confuse between happiness and anger. Our results are comparable with the state-of-art results recently obtained by other researchers [15, 19].

Table 2: Performance of SVMs and training sample numbers

	Number of training samples				Accuracy
	A	H	N	S	
Anger	162	46	49	52	77.16%
Happiness	31	102	31	32	65.64%
Neutral	61	68	194	64	83.73%
Sadness	31	34	31	96	70.59%

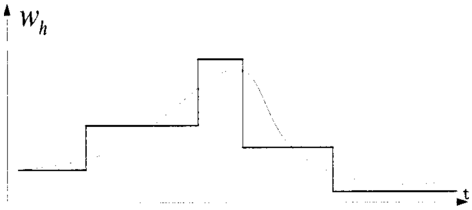

Figure 4: Filtering the emotion curve. The solid lines are the emotion recognition results. The dash line is the smoothed emotion curve.

One benefit of an SVM discriminator is that it will provide a recognition confidence, which can be considered as the proportion of a specific emotion in that phrase. We will describe in Section 4.2 that the proportion can be used to morph between different cartoon templates.

4. Animating Facial Expressions

The emotion recognition results are then used to generate cartoon animation, using cartoon templates drawn for all four emotions.

4.1 Cartoon Face Templates

As shown in Figure 2, the cartoon face model begins with a set of hand-drawn images. For each emotion, we have an artist draw four cartoons that correspond to different emotion levels or intensities ranging evenly from 25% to 100%. Since neutral emotion is the origin in the emotion space, we need only one template for it. In other words, the first level (or 0% intensity) of three emotions (sadness, anger and happiness) always corresponds to the neutral emotion.

4.2 Facial Expression Animation

Because the emotion discriminators produce a deterministic metric for each emotion in a phrase, the emotion will change abruptly between neighboring phrases. Directly animating the facial expression from those outputs will result in frequent face jitter. Therefore, we apply Gaussian filters to the emotion curves to generate a smooth emotion transition. For example, Figure 4 shows the emotion curve of happiness before and after low-pass filtering.

Similarly, we can obtain smooth emotion curves for sadness and anger. At any time instant, we have three values that measure the intensity of emotional happiness, sadness, and anger in a speech phrase, represented by w_h, w_s and w_a. We only use three emotions here because we assume the emotion is neutral when the phrase is silent or the discriminator cannot determine the emotion category of that phrase.

In the animation process, we first quantify the emotion intensity into five levels. For each emotion, appropriate cartoon templates are chosen from the library according to the emotion level. The final facial expression is generated by morphing between the cartoon face templates:

$$I = \frac{w_h I_h + w_s I_s + w_a I_a}{w_h + w_s + w_a}, \qquad (3)$$

where I_h, I_s and I_a are cartoon templates for each emotion.

Before creating the animation, we have the artist place control lines upon key facial features such as the eyes, mouth, chin and nose. The control lines are placed in the same order so that we can build the correspondence between different templates. In our system, we use the field morphing technique proposed in [6] to animate the face templates. Compared with other morphing techniques, field morphing has the advantage of easy-control and smooth transition. Figure 6(A) shows an example sequence generated by our system. Note that the facial expression changes smoothly. The lip shape does not change in this sequence because we only animate the facial expressions in this step.

5. Lip-Syncing

In order to enhance the realism of facial animation, we have developed an algorithm that synthesizes lip configurations from speech [13, 17]. Three problems should be considered for lip-syncing:

- how to represent the audio and visual signals
- how to define the audio-visual mapping
- how to train optimal model parameters

Generally speaking, speech signals in a speech recognition system can be represented at three different levels: front end (or signal level), acoustic model (or phoneme level) and language model (or word level). Although each of the three levels can be applied within a lip-syncing system, much consideration should be taken for a specific application. Because a high-level signal representation is associated to more context cues, better results can be achieved using the latter two methods. At the same time, a higher input signal level necessitates a more complex system.

For example, phoneme-viseme mapping has been used to generate very exciting results in the "Video Rewrite" system [8]. But there are some drawbacks for phoneme-viseme mapping. First, there is no standard definition for a phoneme set, and a variety of phoneme definitions have been adopted for different languages. Second, to obtain a correct segmentation of phonemes for an individual's speech, many parameters need to be tuned to maximize the performance of the acoustic model according to the speaker's gender, dialect or co-articulation. Last, in order to obtain a phoneme sequence for a given speech, we have to incorporate a speech engine in the lip-syncing system — a severe hindrance for real-time implementation.

In our lip-syncing system, we propose a new algorithm that directly maps low-level acoustic signals to visual signals. It is a very important design decision to use low-level acoustic signals. Because the low-level acoustic signals come from the front-end output, our approach has the benefits of real-time processing and simple and language-independent audio-visual mapping relations. Details of the lip-syncing system are described in this section.

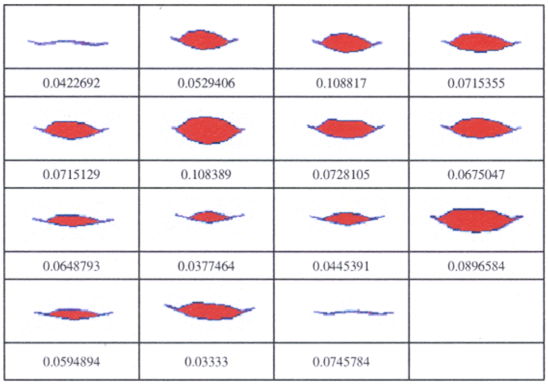

Figure 5: More than 1200 training frames are clustered into 15 classes. The value below each proto-lip represents its relative proportion.

5.1 Audio and Visual Signal Representing

In our system, we calculate the Mel-Frequency Cepstrum Coefficients (MFCC) and the delta coefficients [23]. These coefficients are commonly used for speech recognition and are robust and reliable to variations according to speakers and recording conditions. In our system, we also use these coefficients as features to find the mapping between audio and visual signals. To simplify the synthesis process, we have also adjusted some parameters that are typically used in speech recognition systems to produce a more intuitive mapping.

Specifically in our system, PAL video (25fps) sequences are used as training data. The sampling rate of speech is 44.1 KHz with 16-bit resolution. Under this configuration, the speech signal is blocked into windows of 40ms each that correspond to a 25-Hz sampling rate in the visual domain. For each audio frame, an 18-dimensional feature vector (with 9 MFCC and 9 delta MFCC values), $a = (a_1, a_2, \ldots, a_{18})^T$, is calculated to represent the speech signal.

In our 2D animation system, the lip of the cartoon character is modeled by a closed Catmull-Rom spline [16] determined by ten control points. The lip can then be animated by manipulating the control points to different positions.

5.2 Model Training

Since each control point is a 2D point, we model the lip configuration as a 20-dimensional random vector $v = (v_1, v_2, \ldots, v_{20})^T$. The random vector is assumed to have a distribution formed by a mixture of n Gaussian distributions. Each cluster, or component of the mixture distribution, is parameterized by its relative proportion π_i, its mean μ_i and its covariance R_i. In order to find the mapping between the audio and visual feature vectors, we assume the following:

- All the lip configurations of a cartoon character can be clustered into several classes called, proto-lips.
- Any given lip configuration can be represented by a linear combination of these proto-lips.

From the above assumptions, the first step in the training process is to obtain the key lip templates and their weights (or relative proportions). Although the training data can be obtained by some vision-based techniques such as eigen points [14], creating more artistic work for cartoon face animation is more important than depicting the character realistically; so training data can also be derived from manual labeling results by an artist.

After we have obtained the training data, an unsupervised algorithm is adopted to model the Gaussian mixtures [7]. The training data is finally clustered into n classes (n=15 in our case). Figure 5 shows the mean (proto-lip) and relative proportion for each class. Note that $(\mu_1, \mu_2, . . , \mu_n)$ is not the linear decomposition in vector space because v may not lie in the span of $(\mu_1, \mu_2, . , \mu_n)$. Since these templates represent the lip configuration well, any new configuration can be approximated by a linear combination of the proto-lips.

Given the proto-lips, the next step is to classify each lip configuration in the training data into different classes. Here we use the *Mahalanobis* distance as the similarity measure, i.e.,

$$v_i \in \text{class}$$

if

$$\hat{k} = \arg\min_k (v_i - \mu_k)^T R_i^{-1} (v_i \quad \mu_k) \; (k = 1.. \; n). \tag{4}$$

Because each lip configuration corresponds to an 18-dimensional acoustic feature vector $a = (a_1, a_2, . , a_{18})^T$, all the samples in the audio vector space are also classified into n classes, each of which is associated with a proto-lip template. We further assume that the random vector a, for class **i** has a Gaussian distribution and each dimension in the vector distributes independently. By regression, we can compute the mean $\bar{a}_{ij}$ and covariance σ_{ij} for each Gaussian model (for class i, dimension j). After the training process, we have the following model parameters:

- Proto-lip templates μ_i and their relative proportion π_i ☎)(= 1 n)
- Mean $\bar{a}_{ij}$ and covariance σ_{ij} for the j-th dimension of i-th class for the acoustic feature vector (i = 1. n, $j = 1 \ldots 18$).

5.3 Audio to Visual Mapping

Given a new audio clip, we first segment the audio signal into frames of 40ms each. Then the acoustic feature vector a for each frame is calculated as the system input. Since we do not have any information about the lip configuration when the speaker is silent, we assume that the mouth is closed when the speech energy is below a predefined threshold. Otherwise, we approximate the lip configuration by a linear combination of the proto-lips. Since we assume each dimension of the acoustic feature vector distributes independently, the likelihood $p(a|\mu_i)$ can be represented by:

$$p(a|\mu_i) = \prod_{j=1}^{18} \frac{1}{\sqrt{2\pi}\sigma_{ij}} \exp(-\frac{(a_j - \bar{a}_{ij})^2}{2\sigma_{ij}^2}). \tag{5}$$

According to Bayesian estimation, the posterior probability $p(\mu_i|a)$ is

$$p(\mu_i|a) = \frac{p(a|u_i)p(\mu_i)}{\sum_{i=1}^{n} p(a|\mu_i)p(\mu_i)} \tag{6}$$

where $p(\mu_i) = \pi_i$ is the prior. Then the mapping result becomes

$$v = \sum_{i=1}^{n} \mu_i p(\mu_i|a). \tag{7}$$

Due to mapping error and the existence of noise, the synchronized sequence will appear to flutter open and closed inappropriately. Thus, we apply a Gaussian filter to the synthesized sequence to achieve a smooth transition between neighboring frames.

6. Animation Results

After we obtain the facial expression and lip-syncing results, the two components are combined to generate the final animation. Figure 6 shows three animation sequences generated by our system: emotion-driven cartoon animation (A), speech-driven cartoon animation (B) and combination of both (C). It can be seen that convincing facial expression can be generated according to the emotion recognition results. In addition, lip configuration is synchronized with the character's speech.

7. Conclusions and Future Work

In this paper, we have proposed a system to animate cartoon faces from speech with emotions. Our system consists of two components: emotion-driven cartoon animation, and speech-driven cartoon animation. First, a speech emotion recognition system is developed. Using a corpus of over 1000 utterances from different speakers, we have trained SVM's to recognize four categories of emotions: neutral, happiness, anger and sadness. Given an input speech, the emotion recognizer can generate a smooth transition curve for every emotion, which is further used to drive a cartoon face model. The cartoon face model consists of a small number of hand-drawn templates. Facial expressions are then animated by morphing between these templates. Moreover, the lip shape of each cartoon frame is synthesized from the input audio. Our lip-syncing algorithm uses acoustic signals rather than conventional phoneme-viseme mapping and therefore is language-independent and can run in real-time. The lip shapes are then composed with the morphed facial expression images to form the final cartoon animation. We believe our emotion and speech-driven cartoon animation system will be very useful for HCI applications on desktop PCs and on the Internet.

While our cartoon results are encouraging, there remain a number of areas to be further explored. First, the emotion recognition can be improved. And more emotions such as

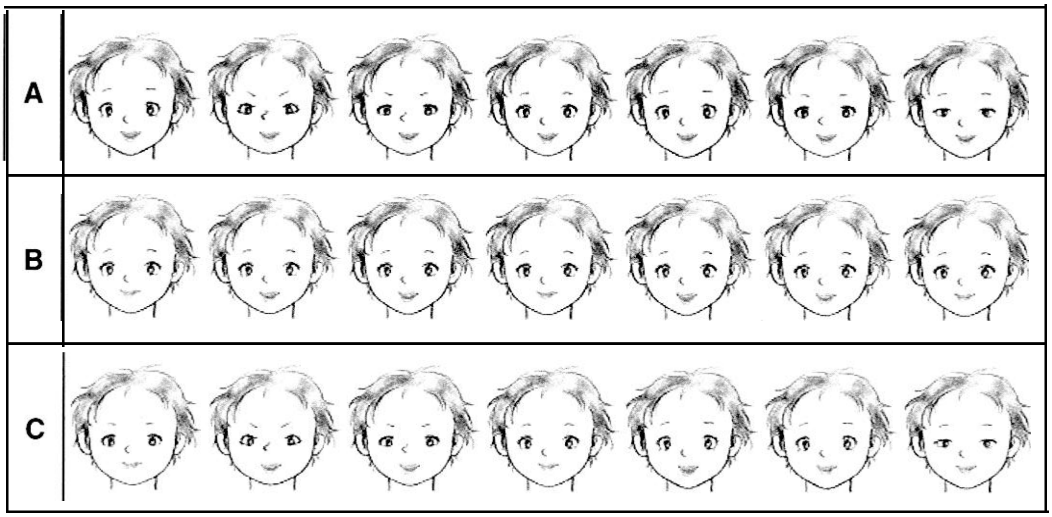

Figure 6: An example animation sequence. A. Facial expressions without lip-syncing (emotion-driven cartoon animation). B. Lip-syncing without facial expressions (speech-driven cartoon animation). C. Combination of A. and B.

surprise, disgust and fear can be used to enhance the realism of the cartoon animation. Currently, our system can only deal with frontal faces. We plan to investigate cartoon animation with different head poses. Both image-based and model-based techniques will be considered.

8. Acknowledgments

We would like to thank Jianlai Zhou for providing the pitch extraction code, and Chao Huang for helpful discussions on the lip-syncing algorithm.

References

[1] http://www.ananova.com.

[2] http://www.famous3d.com.

[3] http://www.hapteck.com.

[4] http://www.intoon.com.

[5] http://www.unige.ch/fapse/emotion.

[6] T. Beier and S. Neely. Feature-based image metamorphosis. In *Proc. ACM Siggraph92*, pages 35–42, 1992.

[7] C. A. Bouman. CLUSTER: An unsupervised algorithm for modeling Gaussian mixtures. available at http://dynamo.ecn.purdue.edu/~ bouman/.

[8] M. Brand. Voice puppetry. In *Proc. ACM Siggraph99*, pages 21–28, 1999.

[9] C. Bregler, M. Covell, and M. Slaney. Video Rewrite: driving visual speech with audio. In *Proc. ACM Siggraph97*, pages 353–360, 1997.

[10] I. Buck, A. Finkelstein, A. Klein, D.H.Salesin, J. Seims, R. Szeliski, and K. Toyama. Performance-driven hand-drawn animation. In *Proc. The 1st International Symposium on Non-Photorealistic Animation and Rendering*, pages 101–108, 2000.

[11] J. Cassell, C. Pelachaud, N. I. Badler, M. Steedman, B. Achorn, T. Beckett, B. Douville, S. Prevost, and M. Stone. Animated conversation: rule-based generation of facial display, gesture and spoken intonation for multiple conversational agents. In *Proc. ACM Siggraph94*, pages 413–420, 1994.

[12] E. Chang, J. L. Zhou, S. Di, C. Huang, and K. F. Lee. Large vocabulary mandarin speech recognition with different approaches in modeling tones. In *Proc. ICSLP 2000*, pages 983–986, 2000.

[13] T. Chen and R. R. Rao. Audio-visual integration in multimodel communication. *IEEE Proceedings*, pages 837–852, May 1998.

[14] M. Covell. Eigen-points: Control-point location using principal component analysis. In *International Workshop on Automatic Face and Gesture Recognition*, pages 122–127, 1996.

[15] F. Dellaert, T. Polzin, and A. Waibel. Recognizing emotion in speech. In *Proc. ICSLP 1996*, Oct 1996.

[16] J. D. Foley, A. van Dam, S. K. Feiner, and J. F. Hughes. *Computer Graphics: Principles and Practice (Second Edition in C)*. Addison-Wesley Publishing Company, 1996.

[17] Y. Li and H. Y. Shum. Animating cartoon face from video. In *6th International Conference on Control, Automation, Robotics and Vision*, 2000.

[18] S. Morishima, K. Aizawa, and H. Harashima. An intelligent facial image coding driven by speech and phoneme. In *Proc. IEEE ICASSP*, 1989.

[19] R. Nakatsu, J. Nicholson, and N. Tosa. Emotion recognition and its application to computer agents with spontaneous interactive capabilities. In *Proc. The Third Conference on Creativity and Cognition*, pages 135–143, 1999.

[20] A. Paeschke and W. F. Sendlmeier. Prosodic characteristics of emotional speech: Measurements of fundamental frequency movements. In *Proc. ISCA-Workshop on Speech and Emotion*, 2000.

[21] V. Petrushin. Emotion recognition in speech signals: Experimental study, development, and application. In *Proc. ICSLP 2000*, 2000.

[22] F. Pighin, J. Hecker, D. Lischinski, R. Szeliski, and D. Salesin. Synthesizing realistic facial expressions from photographs. In *Proc. ACM SIGGRAPH98*, pages 75–84, 1998.

[23] L. Rabiner and B. H. Juang. *Fundamentals of Speech Recognition*. Prentice Hall, 1993.

[24] Z. Ruttkay and H. Noot. Animated chartoon faces. In *Proc. The 1st International Symposium on Non-Photorealistic Animation and Rendering*, pages 91–100, 2000.

[25] K. Waters and T. M. Levergood. DECface: an automatic lip-synchronization algorithm for synthetic faces. Technical report, DEC Cambridge Research Lab, 1993.