SPEECH-DRIVEN AUTOMATIC FACIAL EXPRESSION SYNTHESIS*

Elif Bozkurt[1], Çiğdem Eroğlu Erdem[1], Engin Erzin[2], Tanju Erdem[1], Mehmet Özkan[1], A.Murat Tekalp[2]

[1]Momentum A. .
TÜB TAK-MAM-TEKSEB, A-205, Gebze, Kocaeli, Turkey
E-mail: {ebozkurt, cigdem.erdem, terdem, mozkan}@momentum-dmt.com
[2]Electrical and Electronics Engineering Department, Koç University, Istanbul, Turkey
E-mail: {eerzin, mtekalp}@ku.edu.tr

ABSTRACT

This paper focuses on the problem of automatically generating speech synchronous facial expressions for 3D talking heads. The proposed system is speaker and language independent. We parameterize speech data with prosody related features and spectral features together with their first and second order derivatives. Then, we classify the seven emotions in the dataset with two different classifiers: Gaussian mixture models (GMMs) and Hidden Markov Models (HMMs). Probability density function of the spectral feature space is modeled with a GMM for each emotion. Temporal patterns of the emotion dependent prosody contours are modeled with an HMM based classifier. We use the Berlin Emotional Speech dataset (EMO-DB) [1] during the experiments. GMM classifier has the best overall recognition rate 82.85% when cepstral features with delta and acceleration coefficients are used. HMM based classifier has lower recognition rates than the GMM based classifier. However, fusion of the two classifiers has 83.80% recognition rate on the average. Experimental results on automatic facial expression synthesis are encouraging.

Index Terms— Emotion recognition, facial expression synthesis, classifier fusion

1. INTRODUCTION

Facial expressions and emotional states of a person are related since gestures and facial expressions are used to express an emotion. In this work, we aim to build a correspondence model between emotional speech parameters and facial expressions for automatically animating our 3D talking head model proposed in [2]. In the literature, audio-visual mapping models for facial animation have been presented [3-5]. However, our goal is to build a language and speaker-independent emotion recognition system to drive a fully automatic facial expression animation.

Recognition of the emotional state from speech signal has become an important issue due to its benefits in the domain of the human-computer interaction. There are researchers that explore emotional content of speech for call center applications [6] or for developing toys that would advance human-toy interactions one further step by emotionally responding to humans [7]. Our target is to generate speech-driven expressional head and facial animations that has applications in computer games, e-learning, 3D agent-based assistance, etc.

Although extensively investigated, it is still an open problem to recognize emotions for computers. Cultural, gender-related factors may affect the way emotional speech is perceived. Researchers mostly focus on defining a universal set of features and classifiers that would provide high emotional speech classification rates. There are several approaches that try to classify the given audio into several emotions, which are based on HMMs (hidden Markov models), NNs (neural networks), k-NN (nearest neighbor) algorithm and SVM (support vector machines) [8-11].

2. SYSTEM OVERVIEW

People often can reveal the emotional state of the speaker from speech only. In our system, speech is the only input that drives the animation. We investigate short-term acoustic features like pitch and intensity in addition to Mel Frequency Cepstral Coefficients (MFCCs) with their delta and acceleration coefficients.

The emotions studied throughout this research are happiness, anger, fear, sadness, boredom, disgust, and neutral speech. The overall system is trained and tested on the Berlin Emotional dataset (EMO-DB) [1] using Gaussian

* This work is supported by EC within FP6 under Grant 511568 with the acronym 3DTV.

978-1-4244-1755-1/08/$25.00 ©2008 IEEE

Mixture Models (GMMs) and Hidden Markov Models (HMMs).

In the synthesis part, facial expressions corresponding to the seven emotions that are predefined by a graphic artist on the 3D talking head are linearly interpolated considering the emotion recognition results obtained from the input speech.

The rest of the paper is organized as follows. In Section 3, we discuss the methodologies used for feature extraction, emotion classification and classifier fusion. In Section 4, we present the overall emotion recognition results. In Section 5, we explain the synthesis steps. Finally in Section 6, we provide conclusions and discuss the results.

3. SPEECH-DRIVEN EMOTION RECOGNITION

In this work, we study GMM and HMM based classifiers with prosody related and spectral features for emotion recognition.

3.1. Feature Extraction

While determining emotional status from speech, two types of information sources are available: one is acoustic content and the other is linguistic content. We avoid using the latter, which is a drawback for multi-language emotion recognition purposes and only consider acoustic content. Within acoustic content, we use prosody related features and spectral features to model emotional states. It is well known that in different emotional states speech signal carries different prosodic patterns. Hence prosody features such as pitch and speech intensity can help us model prosodic patterns of different emotions. For example, high values of pitch appear to be correlated with happiness, anger, and fear, whereas sadness and boredom seem to be associated with low pitch values. Similarly, the spectral features, such as MFCC, are expected to model varying nature of speech spectral under different emotions. The first and second derivative components of the prosody and spectral features are also considered to model temporal dynamic changes in the speech signal.

Pitch and intensity features of the emotional speech are calculated using the auto-correlation method. Since pitch values differ for each gender and we aim a speaker-independent system, we need speaker normalization. For each segment with non-zero pitch values, the mean removed pitch and intensity segments are extracted and used as normalized prosody features. We also use pitch-intensity features with their delta and acceleration parameters over non-zero pitch segments.

Other features that we utilize are mel-frequency cepstral coefficients with their delta and acceleration components (first and second derivatives) for GMM classification. Speech files are processed using a 25 ms Hamming window with overlapping frames of 10 ms. Each spectral feature vector includes 12 cepstral coefficients and the energy term.

3.2. GMM based Classification

In the GMM based classifier, probability density function of the feature space is modeled with a GMM for each emotion. We conveniently pick 25 mixtures with diagonal covariance matrices for all GMM based density functions. All the features that belong to a certain emotion are used to train the GMM using the iterative expectation maximization technique. In the recognition phase, posterior probability of the features of a given speech utterance is maximized over all emotion GMM densities. The emotion GMM density that maximizes posterior probability of the utterance is set as the recognized emotion class.

3.3. HMM based Classification

In the HMM based classification, we model the temporal patterns of the emotion dependent prosody contours with the HMM structures. The HMM structure is set to two parallel branches, where each branch has five left-to-right states with a possible loop back. We assume that one branch models emotion dependent prosody patterns in the utterance and the other branch models emotion independent patterns.

We use pitch, first derivative of pitch, and intensity for HMM classifiers to model the prosody patterns. The regions between utterances without a valid pitch are filled with zero mean unit variance Gaussian noise to avail proper training of the HMM classifiers.

3.4. Classifier Fusion

We consider weighted summation based decision fusion technique to combine GMM and HMM based classifiers [12]. The GMM and HMM based classifiers output likelihood scores for each emotion and utterance. Likelihood streams need to be normalized prior to the fusion process. First, for each utterance likelihood scores of both classifiers are mean-removed over emotions. Then, sigmoid normalization is used to map likelihood values to the [0, 1] range for all utterances [12]. After normalization, we have two likelihood score sets for GMM and HMM based classifiers for each emotion and utterance. Let us denote log-likelihoods of GMM and HMM based classifiers respectively as

$$\rho_G(\lambda_{gn}), \text{ for } n=1, 2, \ldots, N \qquad (1)$$
$$\rho_H(\lambda_{hn}), \text{ for } n=1, 2, \ldots, N$$

where λ_{gn} and λ_{hn} are the GMM and HMM models for the n-th emotion; and N is the total number of emotions. The decision fusion then reduces to computing a single set of joint log- likelihood ratios, $\rho(\lambda_1), \rho(\lambda_2), \ldots \rho(\lambda_N)$ for each emotion class. Assuming the two classifiers are statistically independent, we fuse the two classifiers by computing the weighted average of the normalized likelihood scores

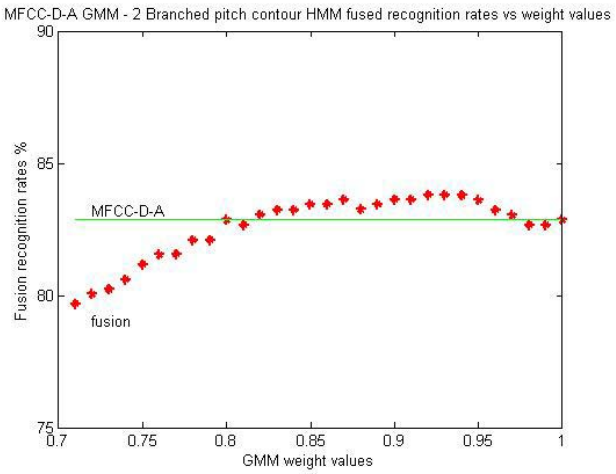

Figure 1. Comparison of MFCC-D-A GMM and MFCC-D-A GMM fused with pitch contour HMM recognition results for varying weight values of MFCC-D-A GMM in the fusion.

$$\rho(\lambda_n) = \alpha\rho_G(\lambda_{gn}) + (1-\alpha)\,\rho_H(\lambda_{hn}) \quad (2)$$

where the weight value α is selected to maximize the recognition rate in the interval [0, 1]. Figure 1 depicts the fusion recognition rate versus weight value (α) relationship where the GMM weight ranges from 0.7 to 1.

4. EXPERIMENTAL RESULTS

In our experimental studies, we used the Berlin Emotional Speech dataset [1]. The Berlin Emotional Speech dataset includes 5 male and 5 female speakers producing 10 German utterances that add up to 535 emotional speech recordings. The recordings have a sampling rate of 16 kHz and we analyze speech signals over 25 ms windows for every 10 ms frame.

We employed the 5 fold stratified cross validation (SCV) method where the dataset is divided into five groups and one is used for testing purposes in turn, while the rest four are spared for training GMMs or HMMs.

GMMs use different feature sets, which are (i) pitch-intensity (PI), (ii) MFCCs, (iii) pitch-intensity and MFCCs (PI-MFCC), (iv) pitch-intensity and MFCCs with delta and acceleration coefficients (PI-MFCC-D-A), and (v) MFCCs with their delta and acceleration coefficients (MFCC-D-A). Given input test emotional speech recordings, the GMM system has a maximum overall recognition rate of 82.85% when MFCCs with delta and acceleration coefficients are used (Table 1).

Since common emotional speech databases are scarce, it is a problem to compare performances of the most systems. In [9], Vlasenko, et al., model the Berlin Emotional Speech corpus with single-state HMMs by parametrizing the speech with MFCCs, added their delta and acceleration parameters.

TABLE I
5 FOLD SCV EMOTION RECOGNITION RATES WITH GMM BASED CLASSIFIER

Feature set	Recognition Rate (%)
PI	47.11
MFCC	76.48
PI-MFCC	75.20
PI-MFCC-D-A	77.82
MFCC-D-A	82.85

TABLE II
CONFUSION MATRIX FOR 5 FOLD SCV EMOTION RECOGNITION RATES AFTER DECISION FUSION OF GMM AND HMM CLASSIFIERS FOR α =0.92

	F	**D**	**H**	**B**	**N**	**S**	**A**
Fear	**72.4**	1.4	17.4	0	7.2	0	1.4
Disgust	0	**86.6**	0	0	6.6	2.2	4.4
Happiness	6.8	1.4	**67.9**	6.6	0	0	17.1
Boredom	0	2.5	0	**86.3**	7.3	3.7	0
Neutral	0	2.5	0	6.4	**89.8**	1.2	0
Sadness	0	0	0	3.2	6.6	**90.1**	0
Anger	2.3	0	8.5	0	0	0	**89.1**

Their 3 SCV test results perform in average 80.6% when HMMs with 25 Gaussian mixture components are used.

Our pitch contour HMM recognition results are higher than random classification. However, in comparison to GMM results, they are lower (37.59% in average). For fair comparison to GMMs, we use grammar model that matches each utterance only to one emotion.

Since MFCC-D-A GMM has the highest recognition rate among our GMM classifiers, we choose this method for fusion with pitch contour HMMs. Above Table 2, shows the confussion matrix of decision fusion results of the two classifiers with average recognition rate 83.80% (Figure 1) when weight value for MFCC-D-A GMM is 0.92.

5. EXPRESSION SYNTHESIS

Our expression synthesizer uses the head model proposed in [2] to visualize the results. In real life, a speaker's emotional status may vary during the speech so we need to determine a minimum duration value that is long enough to satisfy high recognition rates while it is short enough to capture emotional changes. The Berlin dataset recordings have duration 2.7 s in average. Starting from 100 ms up to 3 s, we obtain changes in the recognition rates. For the maximum likelihood approach the more the number of features the best the recognition results are. Observing the plot in Figure 2 below that shows the recognition rate versus window size relationship, we decide that at least 2 s is an appropriate interval for the recognition process.

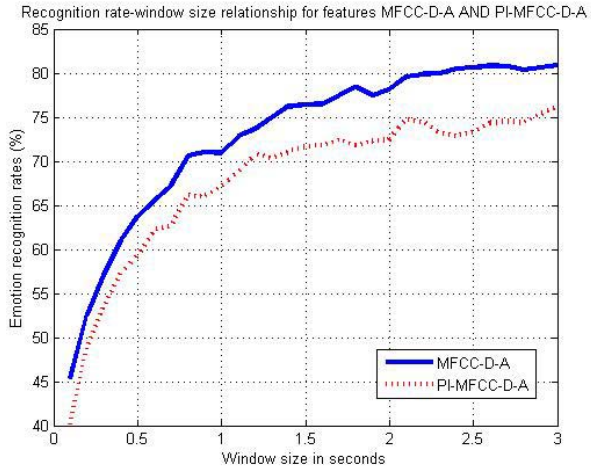

Figure 2. Recognition rate changes for features MFCC-D-A and PI-MFCC-D-A for varying decision window sizes.

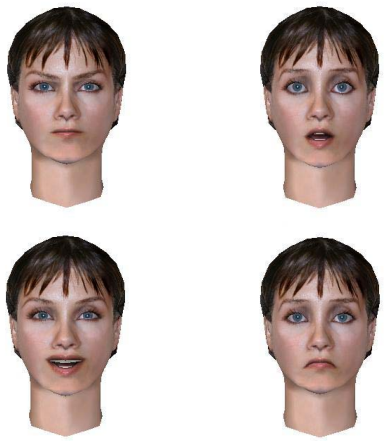

Figure 3. Synthesized expressions anger, fear, happiness and sadness of the 3D head model.

The 3D head model contains facial expressions modeled by a graphic artist (Figure 3). So, for the synthesis, facial expressions are linearly interpolated synchronous to emotion recognition results over windows of two seconds. During expression interpolation, we assume a transition duration of 50 ms between consecutive expressions and the saturation duration is chosen as 1.9 s.

6. DISCUSSION & CONCLUSIONS

We evaluate our emotion recognition system on the Berlin dataset using prosody related and spectral features, all modeled by GMMs and HMMs. For all classifiers and speech features, the misclassification of emotions like happiness and anger is still a problem. As a solution to this problem, history behind the current speech can be used during emotion recognition.

MFCC-D-A feature set among other GMM features outperforms (82.85%) hence we decided to use it for synthesis purposes. Adding the first and second order dynamic features improves overall recognition rates for all kinds of features.

MFCC parameters perform better than pitch parameters because they span the whole spectrum. Since pitch is a short-term feature we explore its contour and use HMMs, which are appropriate for modeling short period patterns.

Additionally, we analyze performance of three and four branched pitch contour HMM structures. However, they do not improve the fusion recognition results. Use of pitch in classification does not help the performance of the spectral features much. Animated expressions are available at http:\\www.momentum-dmt.com\3DTVCon08.

Language is also another factor that determines the recognition result. Since the Berlin dataset is in German, it is less preferable to test the system in other languages. Multi language emotion recognition requires more recordings in more languages.

REFERENCES

[1] Berlin Emotional Speech Database, available online at http://www.expressive-speech.net/

[2] A.T. Erdem, "A New Method for Generating 3D Face Models for Personalized User Interaction," *13th European Signal Processing Conference*, Antalya, September 4-8, 2005.

[3] M. Brand, "Voice Puppetry," *Proceedings of the SIGGRAPH*, 21-28, 1999.

[4] J. Cassell, T. Bickmore, L. Campbell, K. Chang, H. Vilhjlmsson and H. Yan, "Requirements for an architecture for embodied conversational characters," *Proceedings of Computer Animation and Simulation,* 109-120, 1999.

[5] 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," *Proceedings of the SIGGRAPH*, 28(4): 413-420, 1994.

[6] C.M. Lee and S.S. Narayanan, "Toward Detecting Emotions in Spoken Dialogs," *IEEE Transactions on Speech and Audio Processing*, vol. 13, No. 2, March 2005.

[7] P.Y. Oudeyer "The Production and Recognition of Emotions in Speech: Features and Algorithms," *International Journal of Human-Computer Studies*, vol. 59, pp. 157-183, 2003.

[8] B. Schuller, G. Rigoll and M. Lang "Hidden Markov Model-Based Speech Emotion Recognition," *ICASSP*, 2003.

[9] B. Vlasenko and A. Wendemuth, "Tunning Hidden Markov Models for Speech Emotion Recognition," *33rd German Annual Conference on Acoustics*, Stuttgart, Germany, March, 2007.

[10] Y.H. Cho, K.S. Park and R.J. Pak, *Speech Emotion Pattern Recognition Agent in Mobile Communication Environment Using Fuzzy-SVM*, Springer Berlin, Heidelberg, July, 2007.

[11] M. El Ayadi, M. Kamel, and F. Karray, "Speech Emotion Recognition Using Gaussian Mixture Vector Autoregressive Models," *ICASSP*, 2007.

[12] E. Erzin, Y. Yemez and A.M. Tekalp, "Multimodal Speaker Identification Using an Adaptive Classifier Cascade Based on Modality Realiability," *IEEE Transactions on Multimedia*, vol. 7, no.5, pp. 840-852, October 2005.