Speech Driven Automatic Facial Expression Synthesis

May 30, 2008

Outline

- Talking head applications & Project goal
- Emotional speech feature extraction & Classification techniques
- Decision fusion of emotion classifiers
- Experimental results
- Expression synthesis & Animation demos
- Conclusions

Emotionally Expressive Talking Head Applications

- e-Learning
  - foreign language learning
- Computer Games
  - life like characters
- 3D agent-based assistance
  - advanced human-computer relationship
- Information services
  - call center applications

Project Goal

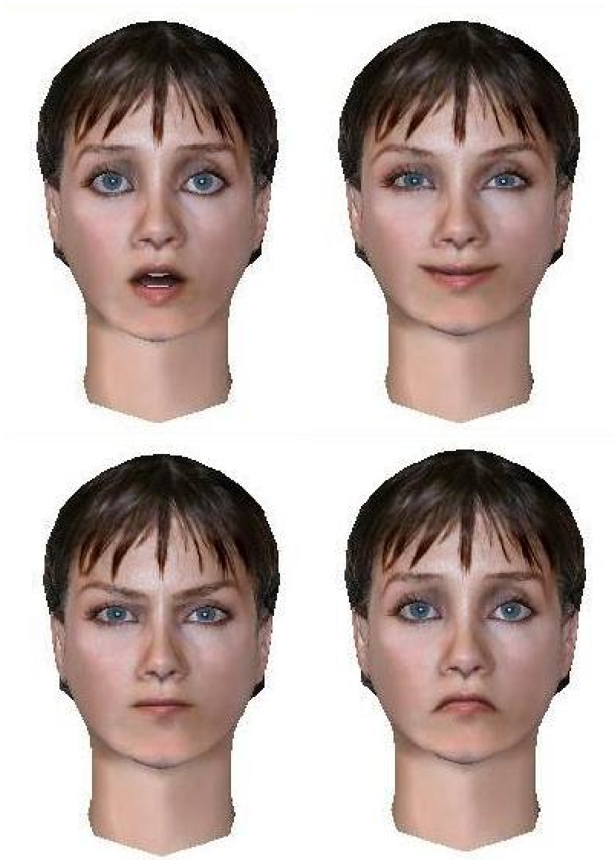
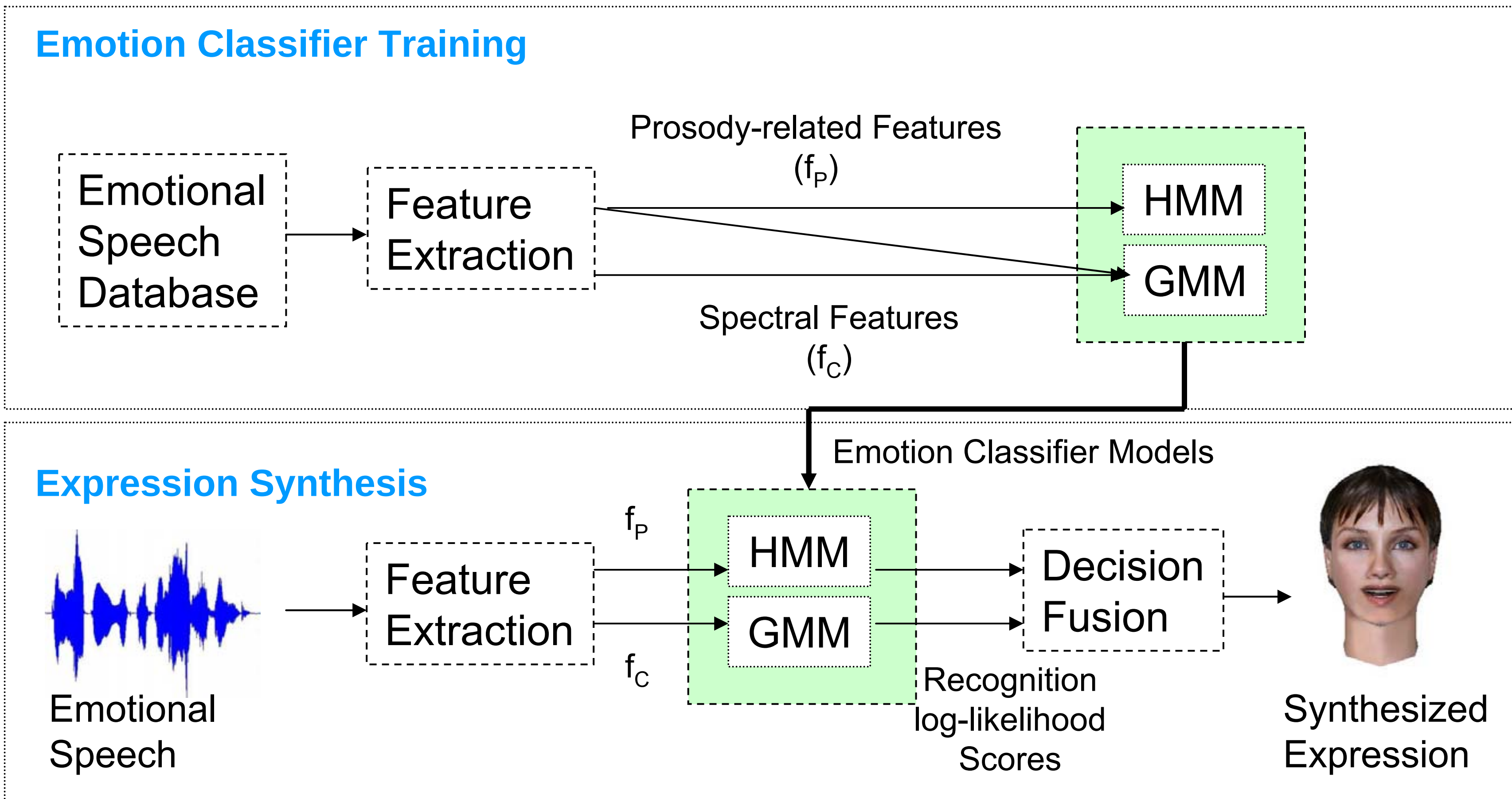


Fig. 1. Example expressions: fear, happiness, anger and sadness

- Facial expressions and emotions of a person are related.
- We aim to build a speaker-independent speech emotion recognition system to drive fully automatic facial expression animation.

System Overview



Emotional Speech Dataset

- Berlin Emotional Speech Dataset (EMO-DB)
  - In German
  - 5 female, 5 male speakers
  - Totally 535 utterances
  - Emotions: *fear*, *disgust*, *happiness*, *boredom*, *neutral*, *sadness*, *anger*
  - 16 kHz, 16 bit Mono Windows PCM

Feature Extraction

- Prosody-related features
  - Short-time features like *pitch*, *1^{st} derivative of pitch* and *intensity*

- Spectral features
  - *Mel frequency cepstral coefficients* (MFCCs) with their delta and acceleration (1^{st} and 2^{nd} derivative) components

Prosody related features

- We use the autocorrelation method to extract prosody related features
- 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
- Speaker normalization
- Delta and acceleration coefficients are also used

Spectral Features

- MFCCs are obtained by processing speech recordings using 25 ms Hamming windows with overlapping frames of 10 ms

- Each spectral feature vector includes 12 cepstral coefficients and the energy term

- Delta and acceleration coefficients are also used

Emotion Classifier

- Gaussian Mixture Model (GMM)
  - Probability density function of the spectral feature space is modeled with a GMM for each emotion.
- Hidden Markov Model (HMM)
  - Temporal patterns of the emotion dependent prosody contours are modeled with an HMM based classifier.

GMM

- We use 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 GMM density with 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.

Speech Features modeled with GMMs

- f_P → Pitch – intensity
- f_C → MFCCs
- $f_{C\Delta}$ → MFCCs with delta and acceleration coefficients
- f_{PC} → Pitch – intensity and MFCCs
- $f_{PC\Delta}$ → Pitch – intensity and MFCCs with delta and acceleration coefficients

HMM

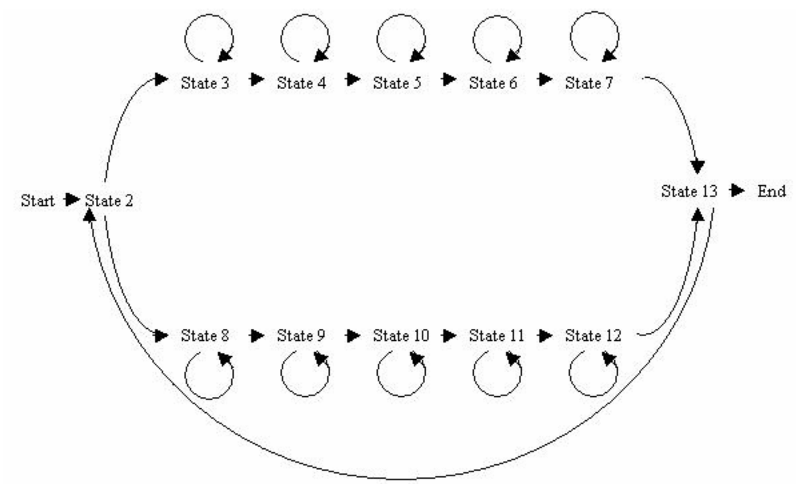


Fig. 2. 2-branched HMM structure, each branch with five left-to-right emitting states

- Pitch, 1st derivative of pitch, and intensity
- 2 branched HMM structure where each branch has 5 left-to-right emitting states with a possible loop back
- Emotion-dependent and emotion-independent characteristics are modeled

Decision Fusion of Classifiers (1)

- Weighted summation based decision fusion technique is used to combine GMM and HMM based classifiers.

- The GMM and HMM based classifiers output likelihood scores are sigmoid normalized and mapped into [0, 1] range.

- After normalization, we have two likelihood score sets for GMM and HMM based classifiers for each emotion and utterance.

Decision Fusion of Classifiers (2)

- Let us denote log-likelihoods of GMM and HMM based classifiers respectively as

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

λ_{gn} : n^{th} emotion GMM
λ_{hn} : n^{th} emotion HMM
N : number of emotions

Decision Fusion of Classifiers (3)

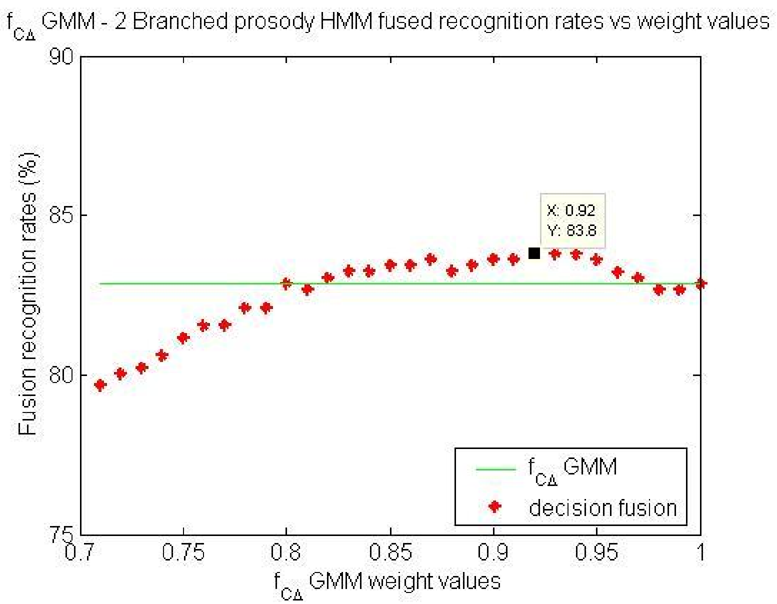


Fig. 3. Comparison of $f_{C\Delta}$GMM and $f_{C\Delta}$ GMM fused with prosody HMM recognition results for varying weight values of GMM in the decision process.

- Assuming the two classifiers are statistically independent

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

- Maximum recognition rate after the decision fusion is 83.80 % for α value 0.92.

Experimental Results

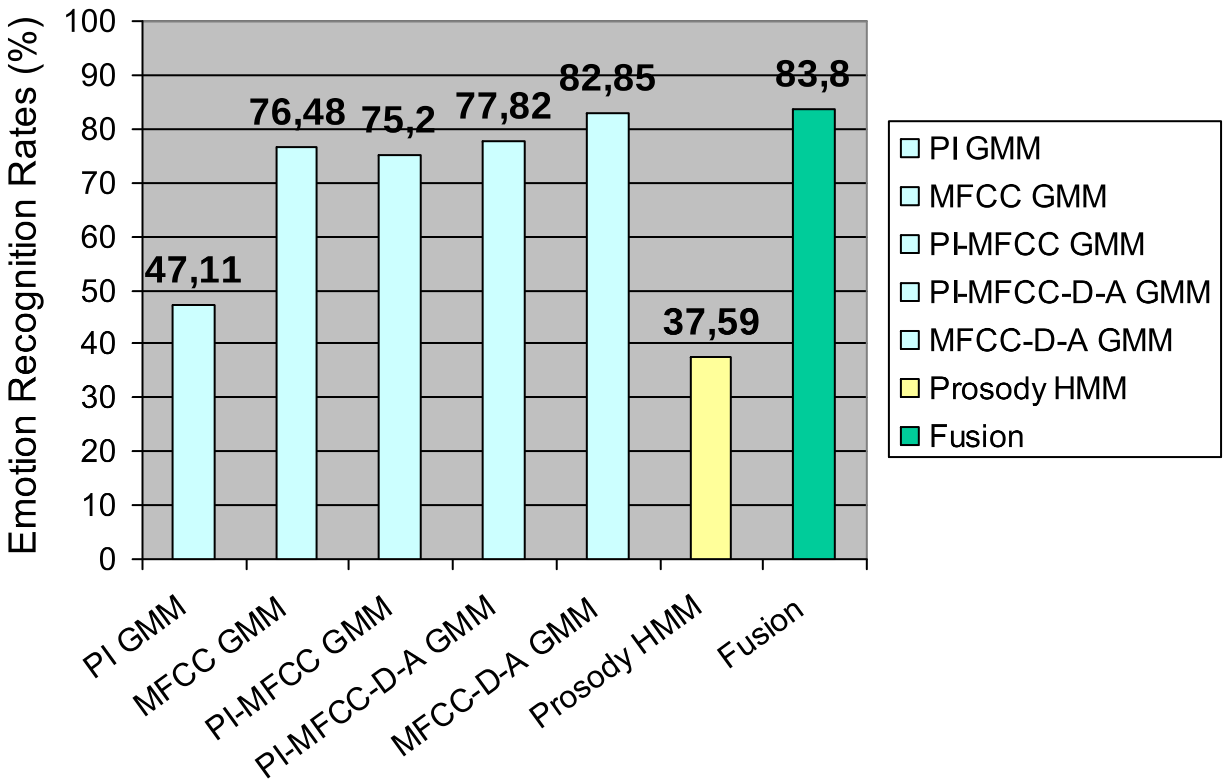


Fig. 4. 5 fold SCV emotion recognition rates for prosody related and spectral speech features classified with HMMs, GMMs and decision fusion of these two classifiers.

Expression Synthesis

Fig. 5. Recognition rate changes of $f_{C\Delta}$ using GMM classifier for varying decision window sizes

- EMO-DB has 2.7 s of average speech recording duration.
- Observing the plot on the left we select decision window size as 2 s.

Animation Player

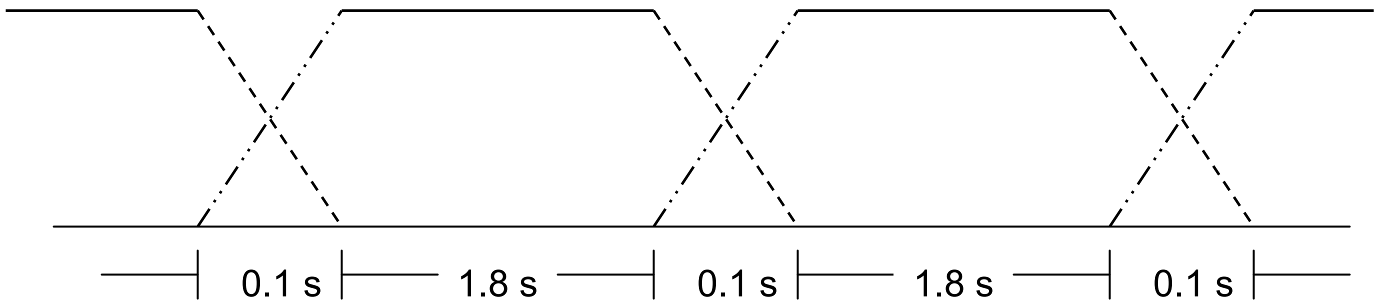


Fig. 6. Linear interpolation of consecutive expressions

- Linear interpolation
- 100 ms transition duration between consecutive expressions and 1.8 s saturation duration

Conclusions

- Prosody related and spectral features are all modeled with GMMs and HMMs.
- Spectral features perform better than prosody related features since they span the whole spectrum.
- Decision fusion of the classifiers increase the recognition results up to 83.80 %.

Thank you