

Why Emotional Awareness Matters:
**Recognizing the Power of Emotions from Facial Expression,
Speech and Language**

Emotion Recognition from Facial Expression and Speech

Chung-Hsien Wu

Professor

Department of Computer Science and Information Engineering
National Cheng Kung University, Tainan, TAIWAN

ICOT2013 & **APSIPA Distinguished Lecture**, Tainan, Taiwan

Emotion Recognition from Facial Expression

2

- Three aspects of facial emotion expression can be considered in emotion recognition:
 - ▣ Basic emotional states
 - ▣ Facial action units
 - ▣ Dimensional description

Emotion Recognition from Facial Expression

Basic emotional states

3

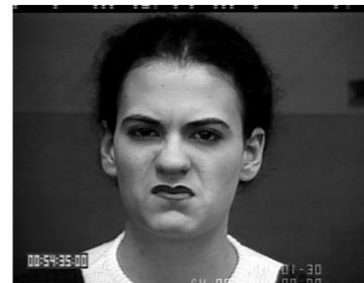
- Facial expressions are usually classified into one of the six basic emotional states (i.e., happiness, sadness, disgust, surprise, fear, and anger).



Happy



Sadness



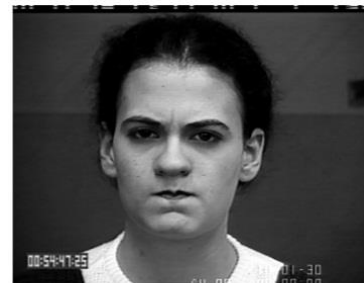
Disgust



Surprise



Fear



Anger

Ekman, P., Friesen, W. V., and Ellsworth, P. (1982). *Emotion in human face*. 2nd ed. Univ. of Cambridge Press., Cambridge.

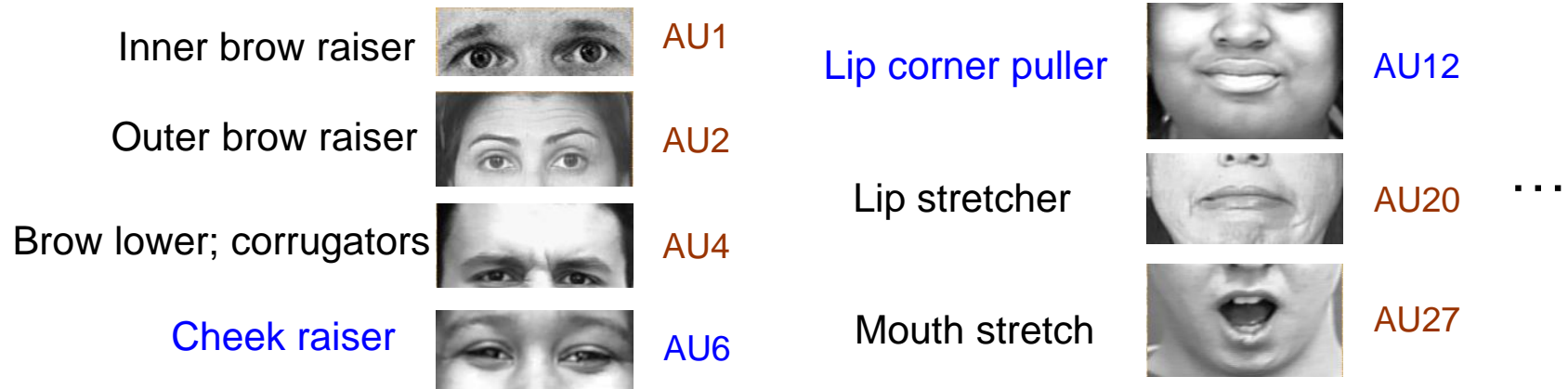
Kanade, T.; Cohn, J. F.; and Tian, Y., 2000. "Comprehensive database for facial expression analysis," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recog.*, pp. 46–53, 2000.

Emotion Recognition from Facial Expression

Facial Action Units

4

- Facial Action Coding System (FACS) was used to classify the atomic facial signals into **Action Units (AUs)** through analysis of facial muscle contractions.
- The AUs are then considered as building blocks of facial expressions for emotion recognition. (Happy: AU6+AU12)

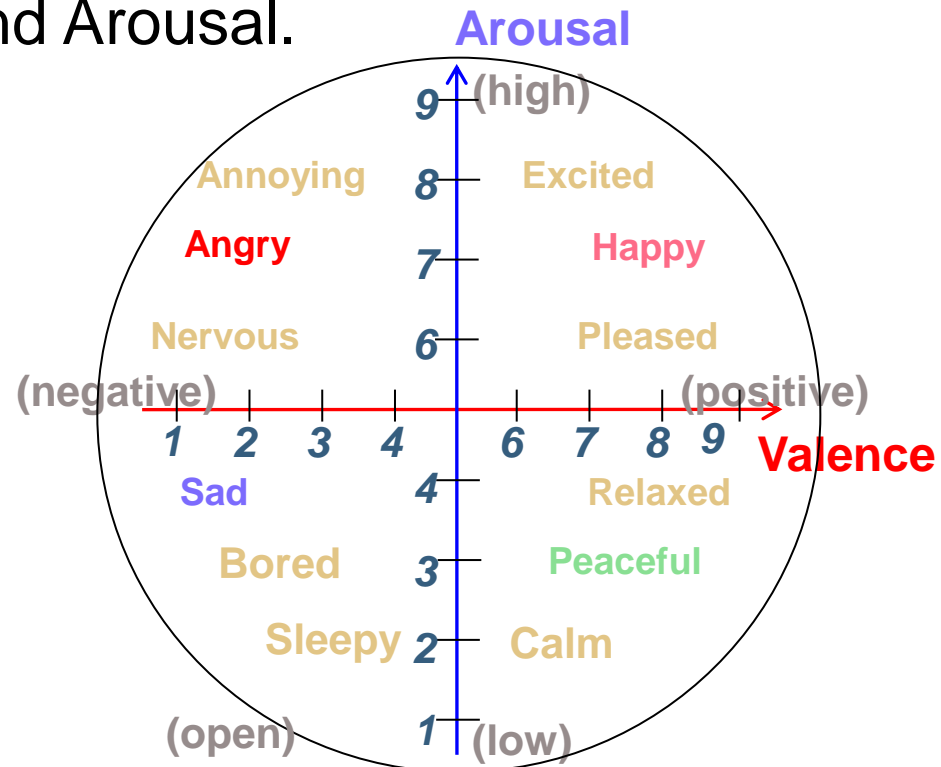


Emotion Recognition from Facial Expression

Dimensional Description

5

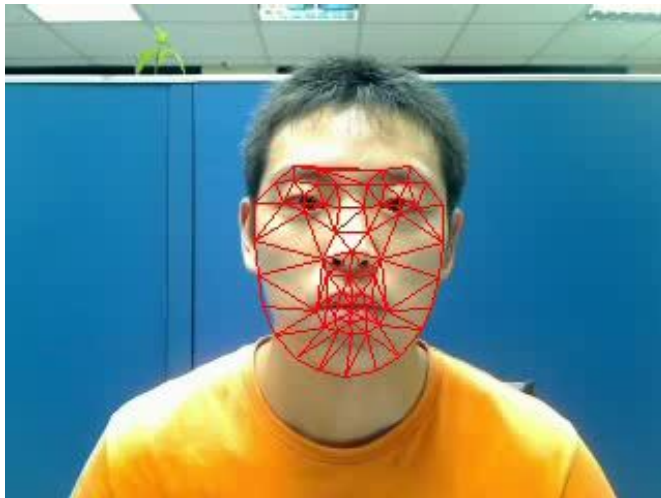
- Valence and Arousal (V-A) emotional plane is adopted to accommodate non-basic and subtle expressions.
- Dimensional and continuous emotion is predicted in terms of Valence and Arousal.



Facial Feature Extraction

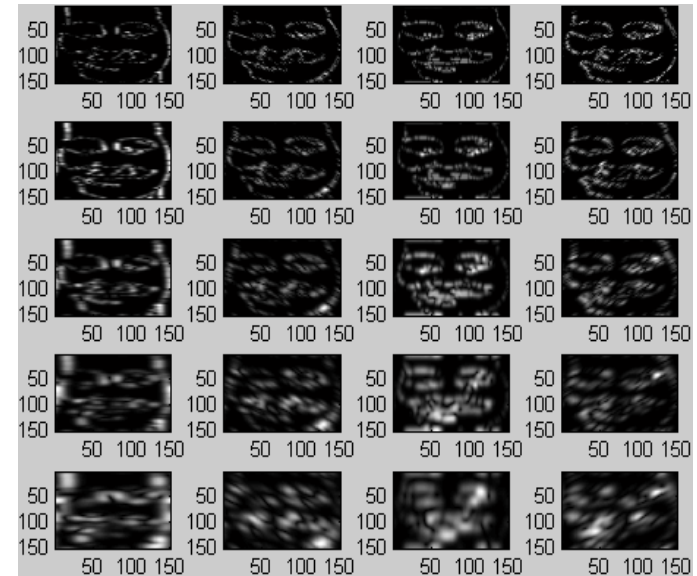
6

- Geometric features
- Shapes of facial components (eyes, mouth, etc.)
- Location of facial salient points (corners of the eyes, mouth)
- Appearance features (wrinkles, bulges, and furrows)



Geometric feature

(Active Appearance Model; AAM)



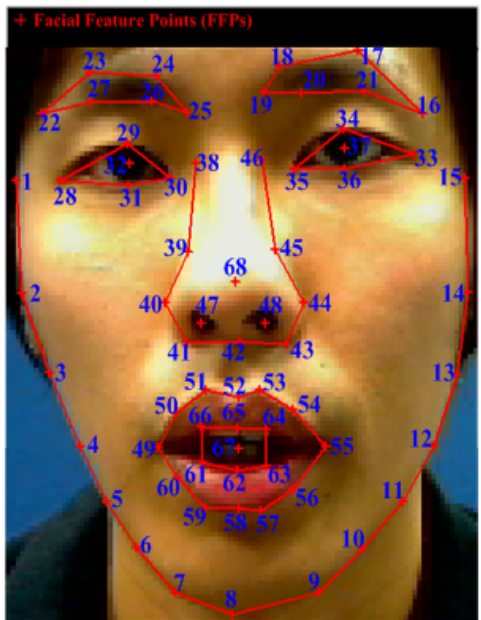
Appearance feature (Gabor wavelets)

Zeng, Z.; Pantic, M.; Roisman, G. I.; and Huang, T. S., 2009. "A survey of affect recognition methods: audio, visual, and spontaneous expressions," IEEE Trans. PAMI. 31, 1, 39-58, 2009.

Facial Feature Extraction

7

- **Feature selection or representation** method for increasing the recognition performance.
 - ▣ For example, the extracted facial feature points (through AAM) can be further represented as **Facial Animation Parameters (FAPs)**.

Extracted Facial Feature Points (FFPs) [↗]	Facial Regions [↗]	FAPs Num. [↗]	Euclidean Distance Between FFPs [↗]	Comparing FFPs Displacement with Neutral Frame [↗]
 <p>+ Facial Feature Points (FFPs)</p>	Eyebrows [↗]	1, 2 [↗] 3, 4 [↗] 5, 6 [↗] 7, 8 [↗] 9, 10 [↗] 11, 12 [↗] 13 [↗]	$D_{vertical,1}(22, 30), D_{vertical,2}(16, 35)^{↗}$ $D_{vertical,3}(25, 30), D_{vertical,4}(19, 35)^{↗}$ $D_{vertical,5}(22, 28), D_{vertical,6}(16, 33)^{↗}$ $D_{vertical,7}(23, 28), D_{vertical,8}(17, 33)^{↗}$ $D_{vertical,9}(25, 28), D_{vertical,10}(19, 33)^{↗}$ $D_{vertical,11}(23, 30), D_{vertical,12}(17, 35)^{↗}$ $D_{horizontal,13}(19, 25)^{↗}$	$D_{v,1_Neutral}-D_{v,1}, D_{v,2_Neutral}-D_{v,2}^{↗}$ $D_{v,3_Neutral}-D_{v,3}, D_{v,4_Neutral}-D_{v,4}^{↗}$ $D_{v,5_Neutral}-D_{v,5}, D_{v,6_Neutral}-D_{v,6}^{↗}$ $D_{v,7_Neutral}-D_{v,7}, D_{v,8_Neutral}-D_{v,8}^{↗}$ $D_{v,9_Neutral}-D_{v,9}, D_{v,10_Neutral}-D_{v,10}^{↗}$ $D_{v,11_Neutral}-D_{v,11}, D_{v,12_Neutral}-D_{v,12}^{↗}$ $D_{h,13_Neutral}-D_{h,13}^{↗}$
	Eyes [↗]	14, 15 [↗] 16, 17 [↗] 18, 19 [↗]	$D_{vertical,14}(29, 31), D_{vertical,15}(34, 36)^{↗}$ $D_{vertical,16}(28, 49), D_{vertical,17}(33, 55)^{↗}$ $D_{horizontal,18}(28, 30), D_{horizontal,19}(33, 35)^{↗}$	$D_{v,14_Neutral}-D_{v,14}, D_{v,15_Neutral}-D_{v,15}^{↗}$ $D_{v,16_Neutral}-D_{v,16}, D_{v,17_Neutral}-D_{v,17}^{↗}$ $D_{h,18_Neutral}-D_{h,18}, D_{h,19_Neutral}-D_{h,19}^{↗}$
	Nose [↗]	20, 21 [↗] 22, 23 [↗]	$D_{vertical,20}(52, 68), D_{vertical,21}(58, 68)^{↗}$ $D_{vertical,22}(49, 68), D_{vertical,23}(55, 68)^{↗}$	$D_{v,20_Neutral}-D_{v,20}, D_{v,21_Neutral}-D_{v,21}^{↗}$ $D_{v,22_Neutral}-D_{v,22}, D_{v,23_Neutral}-D_{v,23}^{↗}$
	Mouth [↗]	24, 25 [↗]	$D_{vertical,24}(52, 58), D_{horizontal,25}(49, 55)^{↗}$	$D_{v,24_Neutral}-D_{v,24}, D_{h,25_Neutral}-D_{h,25}^{↗}$
	Facial Contours [↗]	26, 27 [↗] 28, 29 [↗] 30 [↗]	$D_{horizontal,26}(5, 58), D_{horizontal,27}(11, 58)^{↗}$ $D_{horizontal,28}(2, 68), D_{horizontal,29}(14, 68)^{↗}$ $D_{vertical,30}(8, 68)^{↗}$	$D_{h,26_Neutral}-D_{h,26}, D_{h,27_Neutral}-D_{h,27}^{↗}$ $D_{h,28_Neutral}-D_{h,28}, D_{h,29_Neutral}-D_{h,29}^{↗}$ $D_{v,30_Neutral}-D_{v,30}^{↗}$

Methods for Emotion Recognition from Facial Expression

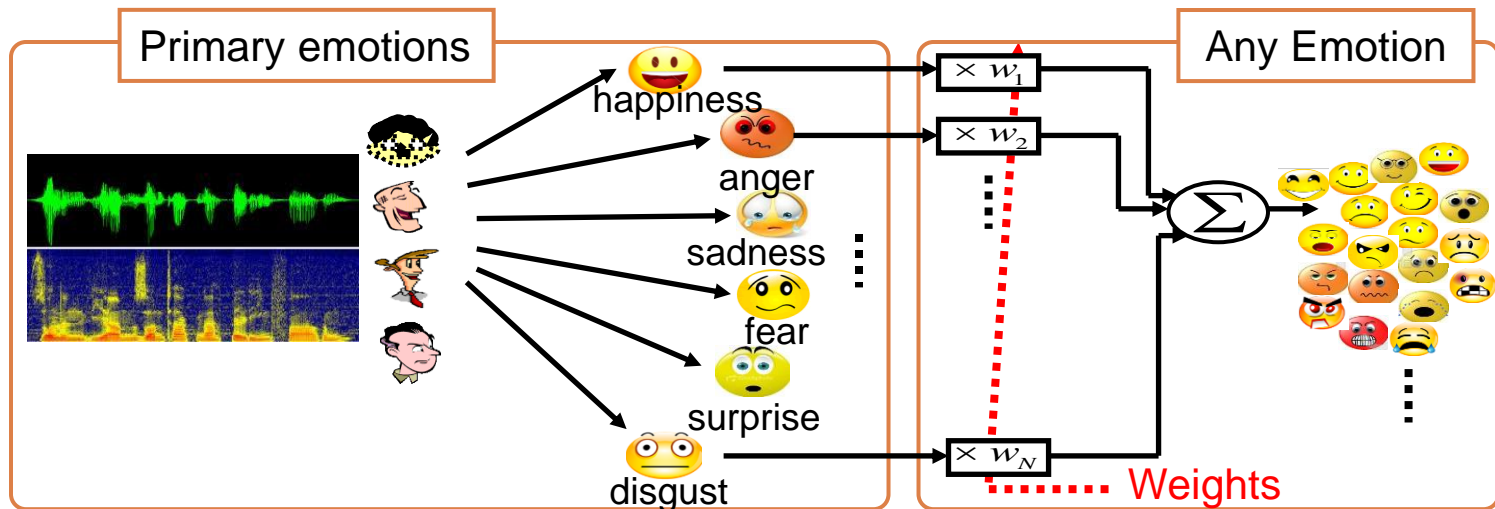
8

- The popularly used recognition methods include:
 - ▣ Support Vector Machine (SVM)
 - ▣ Gaussian Mixture Model (GMM)
 - ▣ Hidden Markov Model (HMM)
 - ▣ Dynamic Bayesian Network (DBN)
 - ▣ Neural Network (NN)
 - ▣ K-Nearest-Neighbor (KNN)
 - ▣ Rule-based method
 - ▣ Decision Tree

Emotion Recognition from Speech

9

- “**Palette Theory**” states that any emotion can be decomposed into primary emotions similar to the way that any color is a combination of basic colors (R,G,B).
- Primary emotions are anger, disgust, fear, happiness, sadness, and surprise.



Zeng, Z.; Pantic, M.; Roisman, G. I.; and Huang, T. S., 2009. "A survey of affect recognition methods: audio, visual, and spontaneous expressions," *IEEE Trans. PAMI*. 31, 1, 39-58, 2009.
Ayadi, M. E.; Kamel, M. S.; and Karray, F., 2011. "Survey on speech emotion recognition: features, classification schemes, and databases," *Pattern Recognition*, pp. 572-587, 2011.

Emotion Related Speech Feature

10

- Speech features for emotion recognition can be grouped into four categories:
 - ▣ Prosodic features: **pitch**, **energy**, **formants**, etc.
 - ▣ Voice quality features: **harsh**, **tense**, **breathy**, etc.
 - ▣ Spectral features: **LPC**, **MFCC**, **LPCC**, etc.
 - ▣ Teager Energy Operator (TEO)-based features: **TEO-FM-var**, **TEO-Auto-Env**, etc.

Emotion Recognition

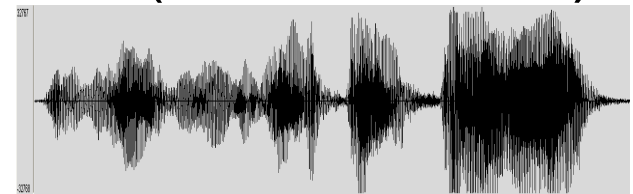
Feature Extraction

11

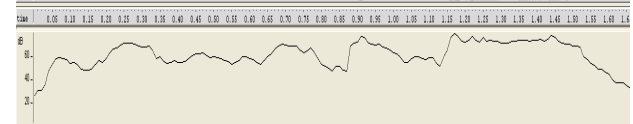
- Prosody has been proven to be the primary and classical indicators of a speaker's emotional state.
- Three kinds of primary prosodic features are extracted for each speech frame (local feature), including

- Pitch
- Energy
- Formants (F1-F5)

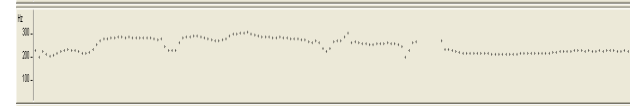
Original waveform



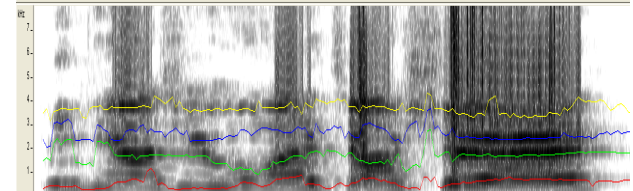
Pitch contour



Energy contour



Formant



Association between Speech Features and Emotion

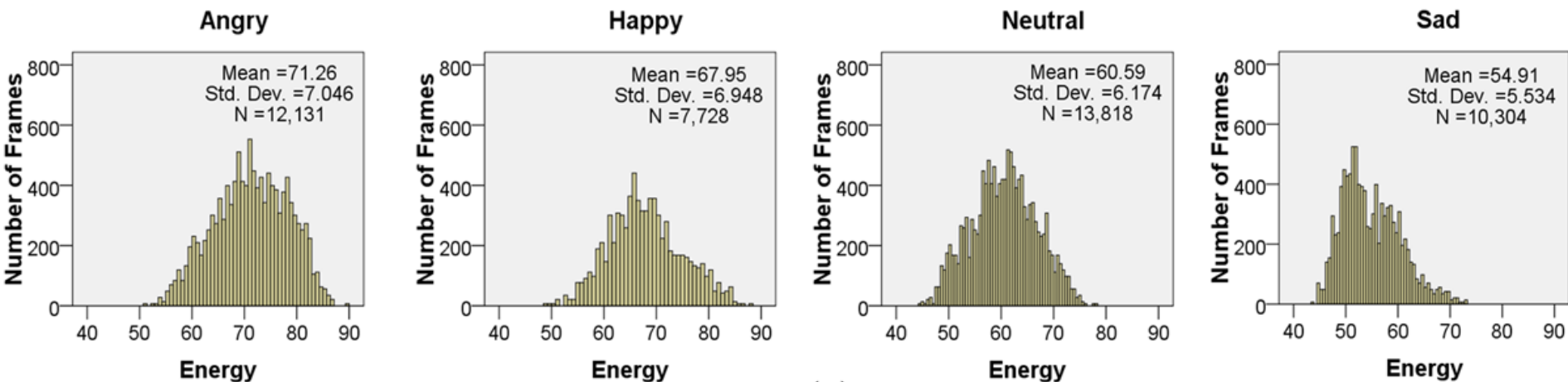
	Pitch mean	Pitch range	Energy/Intensity	Speaking rate	Formants
Anger	Increased (very much higher)	Wider (much wider)	Increased (higher)	High (slightly faster)	F1 mean increased; F2 mean higher or lower; F3 mean higher
Happiness	Increased (much higher)	Wider (much wider)	Increased (higher)	High (faster or slower)	F1 mean decreased & bandwidth increased
Sadness	Decreased (slightly lower)	Narrower (slightly narrower)	Decreased (lower)	Low (slightly slower)	F1 mean increased & bandwidth decreased; F2 mean lower
Disgust	Decreased (very much lower)	Wider or narrower (slightly wider)	Decreased or normal (lower)	Higher (very much slower)	F1 mean increased & bandwidth decreased; F2 mean lower
Fear	Increased or decreased (very much higher)	Wider or narrower (much wider)	Normal (normal)	higher or low (much faster)	F1 mean increased & bandwidth decreased; F2 mean lower

D. Morrison et al., "Ensemble methods for spoken emotion recognition in call-centres", Speech Commun., 2007
 Rosalind W. Picard., Affective Computing MIT Press, 1997

Emotion Related Speech Feature

13

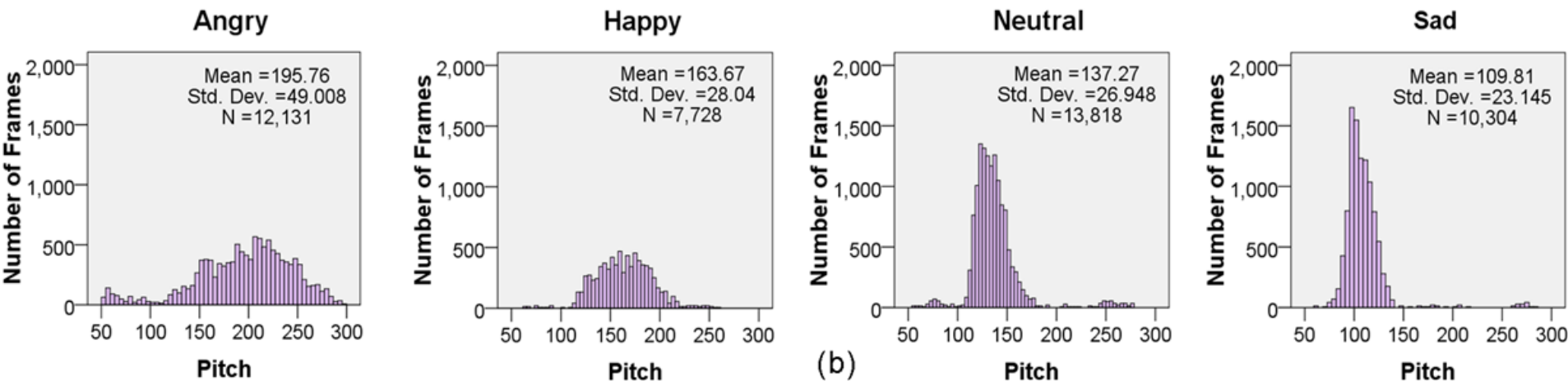
- Statistical analyses of the energy and pitch features were conducted for four emotional states using **MHMC** database.
- Speakers often increased their **energy** to emphasize high-active emotions.



Emotion Related Speech Feature

14

- The **pitch levels** and **pitch ranges** of sad emotion are lower and narrower than those of other emotional states
- The mean and standard deviation of pitch in sad emotion are smaller than those of other emotions.



Emotion Related Speech Feature

15

- The choice of proper features for speech emotion recognition highly depends on the classification task being considered.
- For example:
 - ▣ The TEO-based features are better for detecting stress in speech.
 - ▣ The prosodic features are better for classifying high or low arousal emotions.
 - ▣ The spectral features such as the MFCC are the most promising features for speech representation.

Methods for Emotion Recognition from Speech

16

- The popularly used recognition methods include:
 - ▣ Support Vector Machine (SVM)
 - ▣ Gaussian Mixture Model (GMM)
 - ▣ Hidden Markov Model (HMM)
 - ▣ Dynamic Bayesian Network (DBN)
 - ▣ K-Nearest-Neighbor (KNN)
 - ▣ Linear Discriminant Analysis (LDA)
 - ▣ CART tree

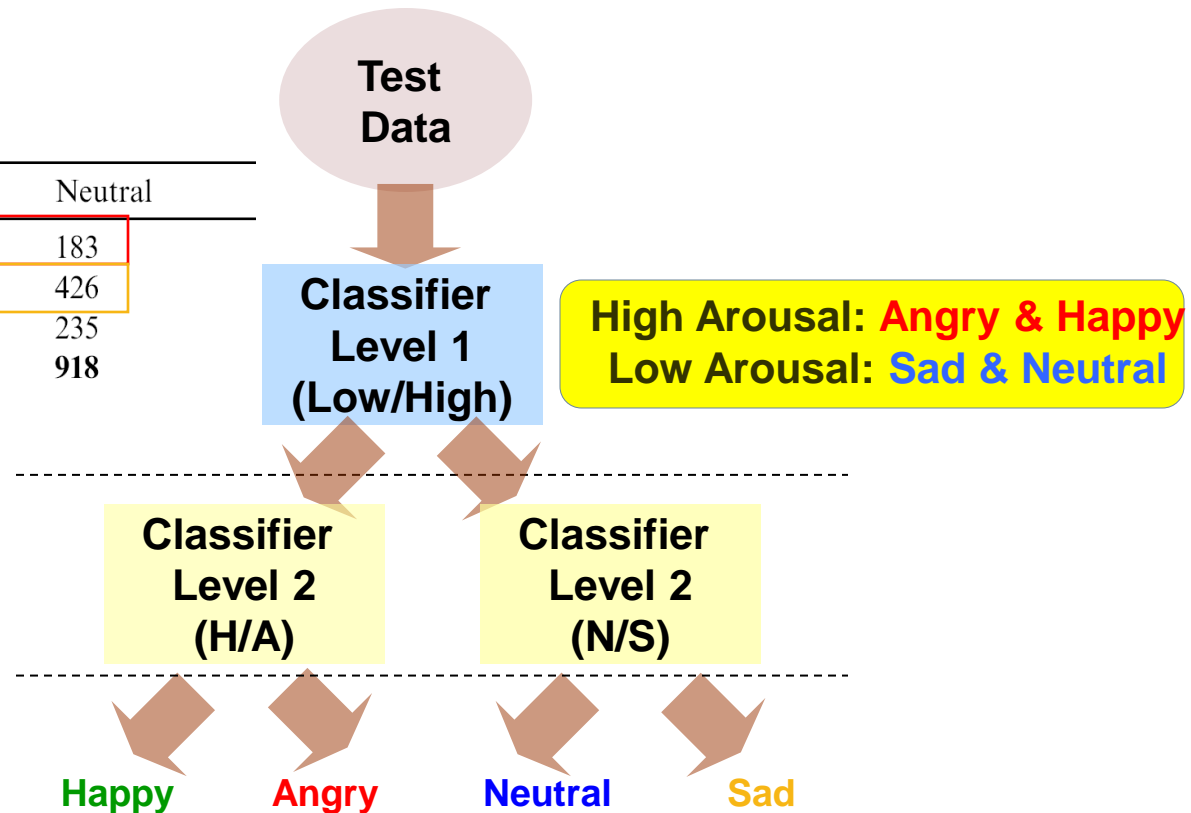
Zeng, Z.; Pantic, M.; Roisman, G. I.; and Huang, T. S., 2009. "A survey of affect recognition methods: audio, visual, and spontaneous expressions," *IEEE Trans. PAMI*. 31, 1, 39-58, 2009.
Ayadi, M. E.; Kamel, M. S.; and Karray, F., 2011. "Survey on speech emotion recognition: features, classification schemes, and databases," *Pattern Recognition*, pp. 572–587, 2011.

Speech Emotion Recognition

17

- A hierarchical classification structure is used for emotion recognition.

	Angry	Happy	Sad	Neutral
Angry	720	168	30	183
Happy	319	680	205	426
Sad	24	42	782	235
Neutral	116	256	394	918



Chi-Chun Lee et al., "Emotion recognition using a hierarchical binary decision tree approach",
Speech Commun., 2011

Bi-Modal Emotion Recognition from Facial Expression and Speech

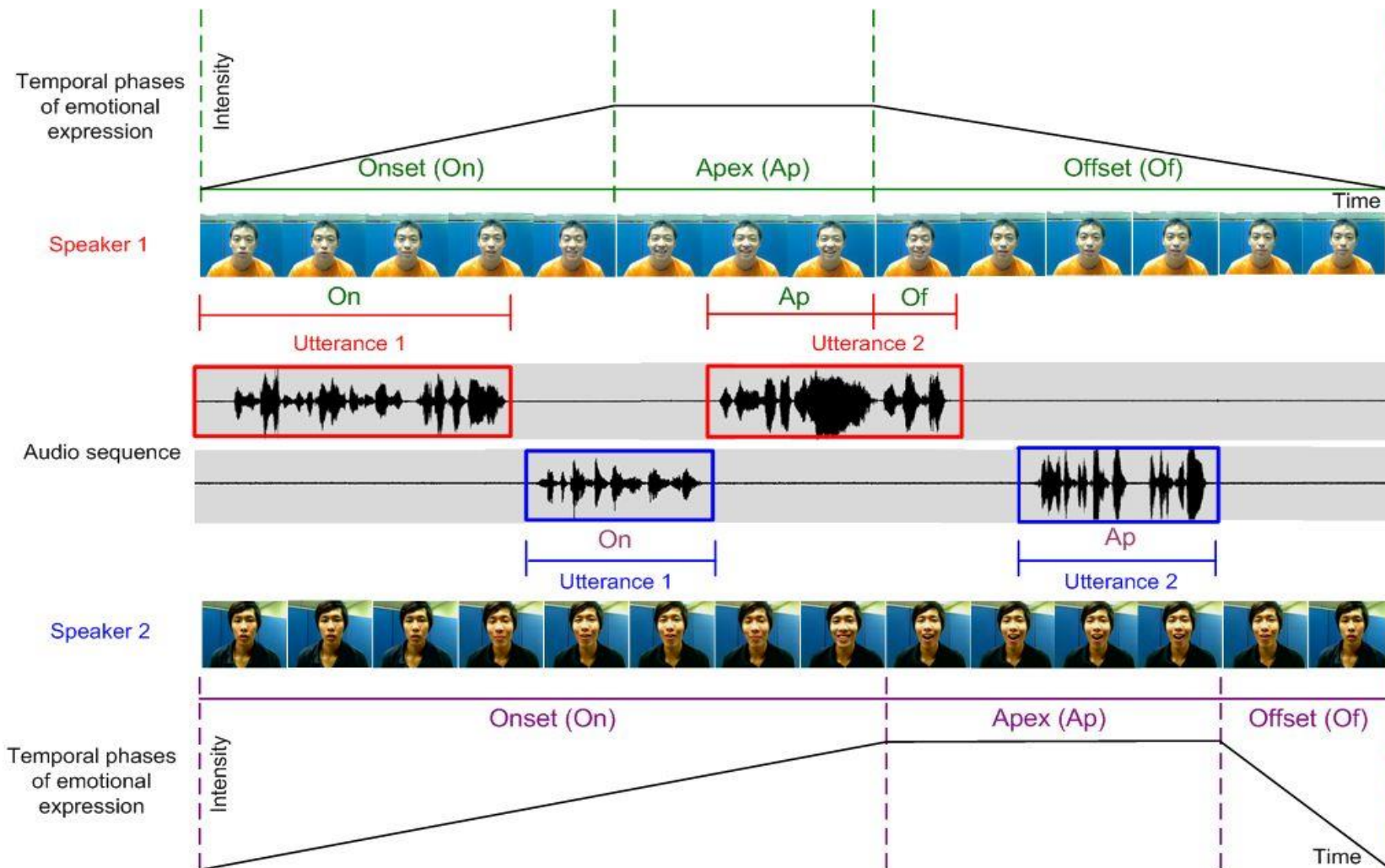
18

- In previous studies, emotion recognition was mostly focused on exploring single facial or vocal modality.
- Based on the psychological analysis, human emotions were mainly transmitted through “Face”, “Voice”, and “Speech Content” in verbal communication.
- Exploring data fusion strategy is a viable direction for emotion recognition.



Bi-Modal Emotion Expression

19

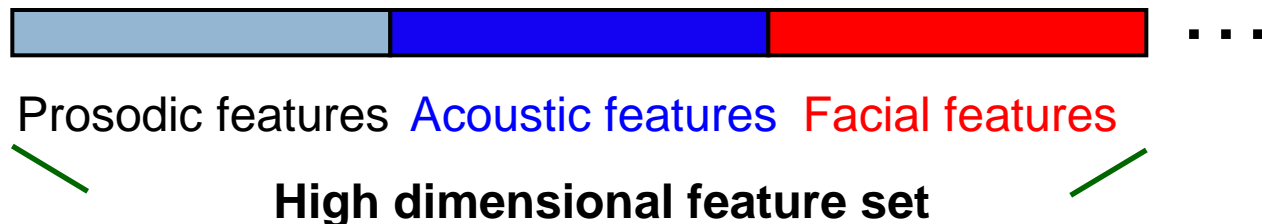


The State of The Art

Data Fusion Strategy

20

- The fusion operations reported can be classified into three major categories: fusion at feature-level, decision-level, or model-level for audiovisual emotion recognition.
- **Feature-level fusion**: facial and vocal features are concatenated to construct a joint feature vector, and are then modeled by a single classifier for emotion recognition.
 - ▣ **PROs**: Take advantage of combining various feature cues.
 - ▣ **CONs**: High dimensional feature set may easily suffer from the problem of data sparseness, and stress the computational resources.

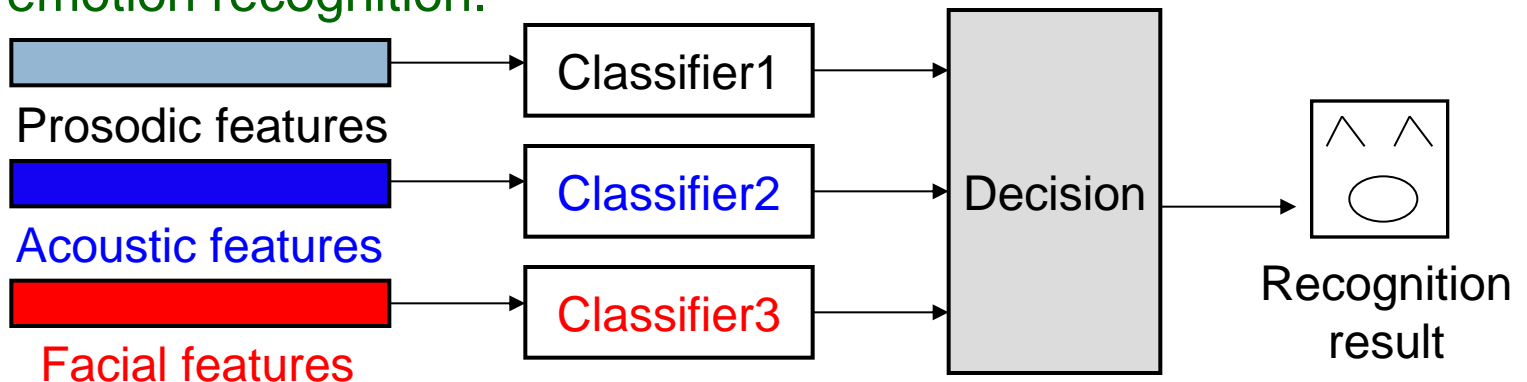


The State of The Art

Data Fusion Strategy

21

- **Decision-level fusion:** multiple features can be modeled by the corresponding classifier first, and then the recognition results from each classifier are fused in the end.
- **PROs:** Without increasing the dimensionality, this approach can combine various modalities by exploring the contributions of different emotional expressions.
- **CONs:** The assumption of conditional independence among multiple modalities is inaccurate in some applications such as emotion recognition.

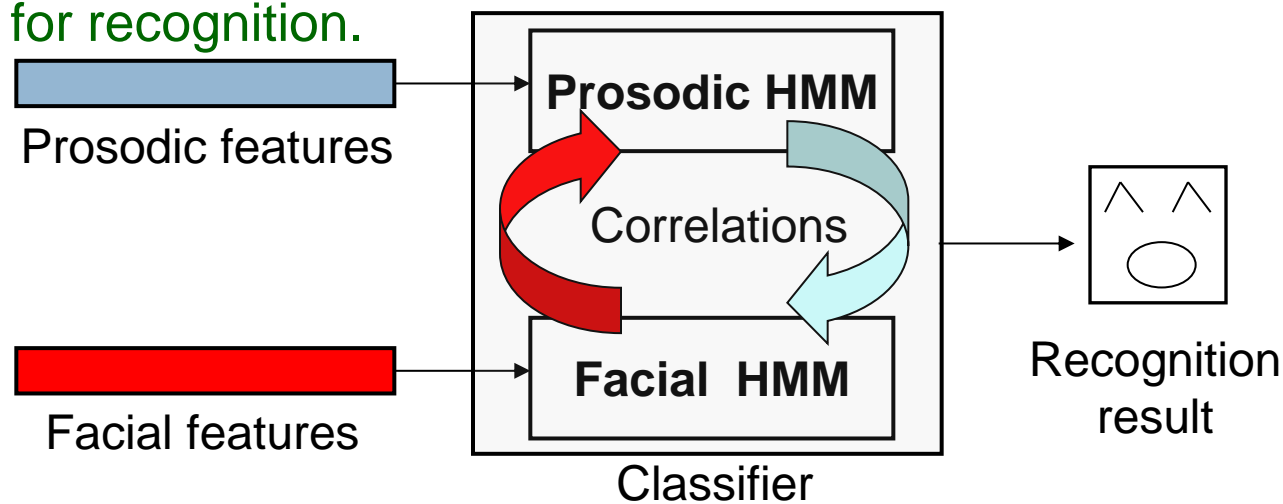


The State of The Art

Data Fusion Strategy

22

- **Model-level fusion**: multiple features can be modeled either dependently or independently through the proposed probabilistic models, which are (then) linked together through various criteria or assumptions.
 - ▣ **PROs**: Emphasizes the information of correlation among multiple modalities, and explores the temporal relationship between the various signal streams.
 - ▣ **CONs**: Without gaining insights into the role of multiple modalities for recognition.



The State of The Art

Data Fusion Strategy

23

- Recent studies have shown increasing attention to decision-level fusion and model-level fusion for multimodal pattern recognition.
- Notable examples include “Error Weighted Classifier Combination (EWC)” and “Coupled Hidden Markov Model (C-HMM).
- These Models were successfully used in various fields, such as “interest detection”, “human identification”, “hand gesture recognition”, “audiovisual speech recognition”, “speech animation”, and “**Emotion Recognition**”.

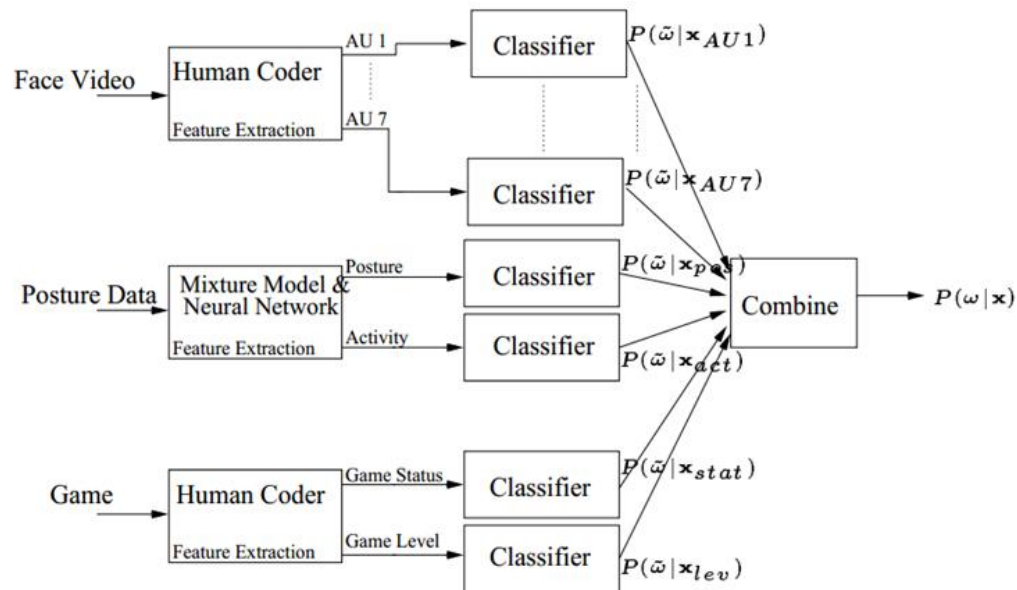
Ivanov, Y.; Serre, T.; and Bouvrie, j. 2005. “Error weighted classifier combination for multi-modal human identification,” Technical Report CBCL paper 258, Massachusetts Institute of Technology, Cambridge, MA, 2005.
Brand, M.; Oliver, N.; and Pentland, A. 1997. “Coupled hidden Markov models for complex action recognition,” *Proc. Int’l Conf. Computer Vision Pattern Recognition*, pp. 994–999, 1997.

The State of The Art

Data Fusion Strategy

24

- Error Weighted Classifier Combination (decision-level fusion)
 - ▣ In EWC, conditional error distribution was used. This distribution can be approximated from a confusion matrix for each classifier over all the training data.
 - ▣ The confusion matrix can be regarded as empirical weights, and are used to weight the output of each classifier in order to obtain an optimally combined decision.

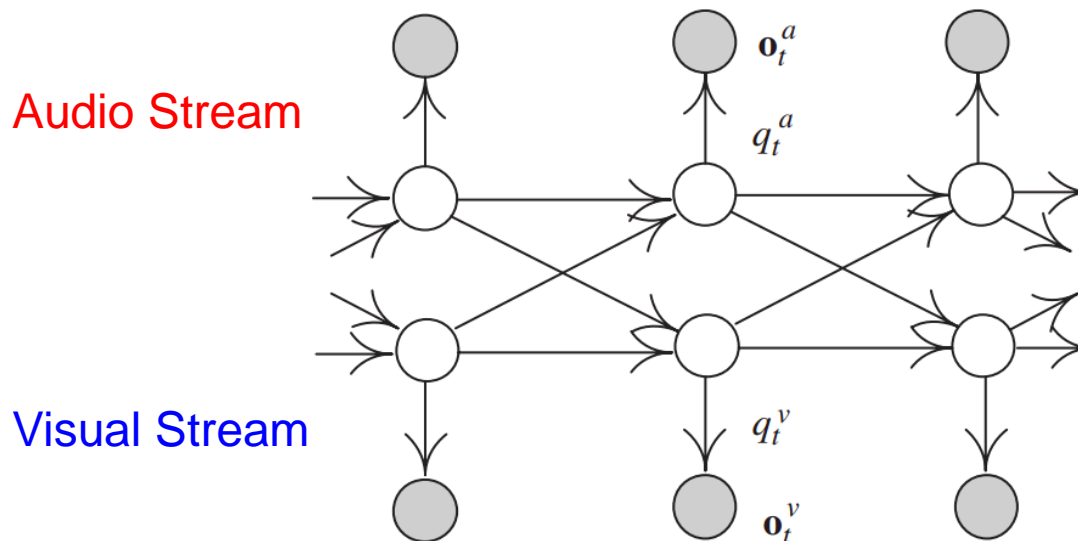


The State of The Art

Data Fusion Strategy

25

- Coupled Hidden Markov Model (model-level fusion)
 - ▣ C-HMM can be seen as a collection of two HMM chains coupled through cross-time and cross-chain conditional probabilities in which a state variable at time t is dependent on its two predecessors from two streams at time $t-1$.
 - ▣ This structure models the asynchrony of multimodal signals (e.g., audio and visual), and preserves their natural correlations over time.



Disadvantages of EWC and C-HMM

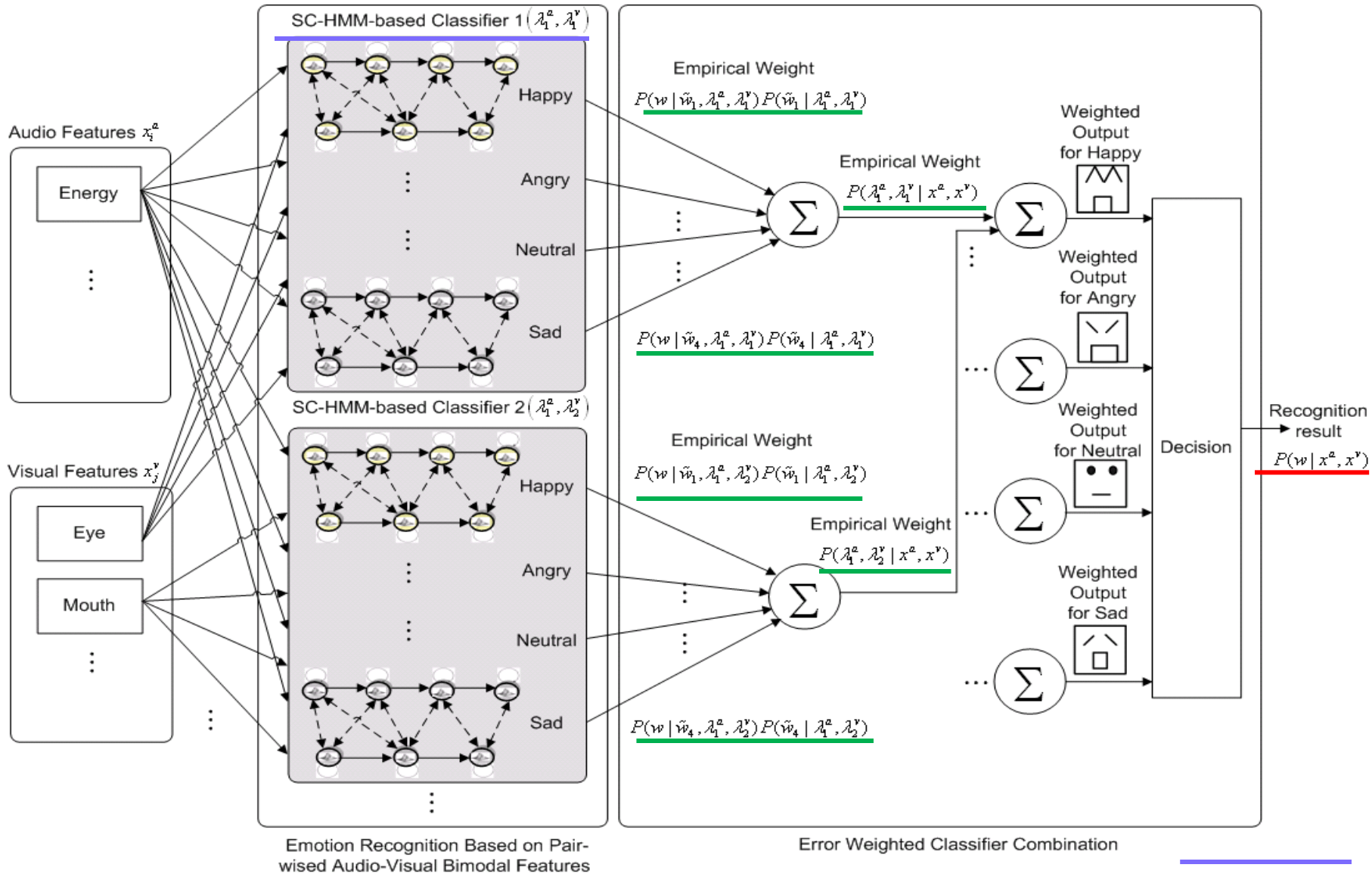
26

- Error Weighted Classifier Combination (EWC)
 - ▣ As the facial and vocal features are demonstrated complementary to each other in emotional expression, the assumption of conditional independence among multiple modalities in EWC is inappropriate.
- Coupled Hidden Markov Model (C-HMM)
 - ▣ To optimize all the parameters “globally” by iteratively refining the parameters of the component HMMs and their coupling parameters simultaneously may lead to an over-fitting effect in sparse data conditions and stress the computational resources.

Error Weighted Semi-Coupled HMM (EWSC-HMM)

27

- A hybrid approach named EWSC-HMM is proposed to take the advantage of model- and decision-level fusion to obtain a better recognition accuracy.
 - ▣ Not only consider the correlation of paired multiple streams but also explore their contributions for recognition.
- For relaxing the connection criterion, an SC-HMM based on a state-based bimodal alignment strategy is proposed.
 - ▣ Alleviate the over-fitting problem in sparse data conditions by providing a loosely coupled statistical dependency of states for linking two HMMs.



$$\begin{aligned}
 \underline{P(w | x^a, x^v)} &\approx \sum_{i=1}^C \sum_{j=1}^D \left\{ \sum_{k=1}^K \underline{P(w | \tilde{w}_k, \lambda_i^a, \lambda_j^v)} [\max_{S^a, S^v} \underline{P(x^a, S^a | \lambda_i^a, \tilde{w}_k)} P(S^v | S^a, \lambda^a, \tilde{w}_k)] \right. \\
 &\quad \left. \underline{P(S^a | S^v, \lambda^v, \tilde{w}_k)} P(x^v, S^v | \lambda_j^v, \tilde{w}_k)] P(\tilde{w}_k | \lambda_i^a, \lambda_j^v) \right\} \underline{P(\lambda_i^a, \lambda_j^v | x^a, x^v)}
 \end{aligned}$$

SC-HMM

Empirical weights

EWSC-HMM

Semi-Coupled HMM (SC-HMM)

29

- SC-HMM with state-based alignment strategy

$$\max_{S^a, S^v} \underbrace{P(x^a, S^a \mid \lambda_i^a, \tilde{w}_k)}_{\text{Audio HMM output}} \underbrace{P(S^v \mid S^a, \lambda^a, \tilde{w}_k) P(S^a \mid S^v, \lambda^v, \tilde{w}_k)}_{\text{State alignment probabilities}} \underbrace{P(x^v, S^v \mid \lambda_j^v, \tilde{w}_k)}_{\text{Visual HMM output}}$$

- In the training procedure (i.e. learning algorithm), the proposed SC-HMM consists of three parts.
 - Audio and visual HMMs were trained separately using the expectation-maximization (EM) algorithm.
 - The best state sequences of the audio and visual HMMs were obtained using the Viterbi algorithm, respectively.
 - The audio and visual state sequences were aligned based on the alignment strategy.

EWSC-HMM

Semi-Coupled HMM (SC-HMM)

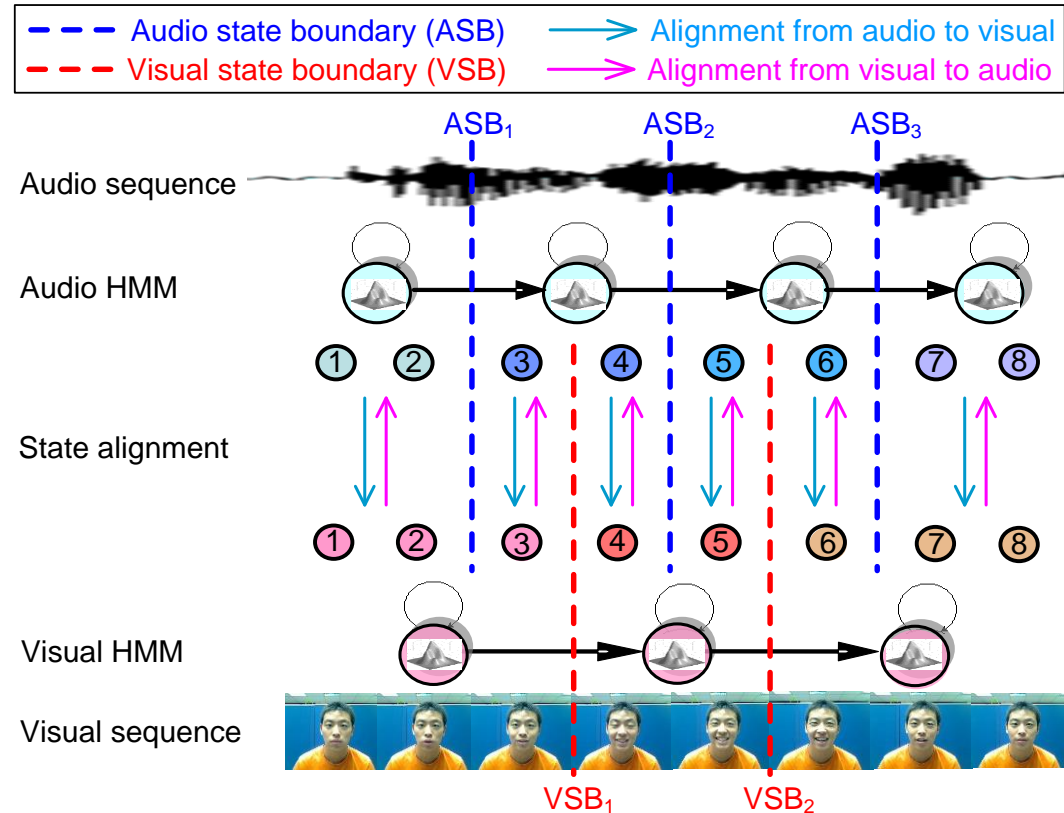
30

State-based alignment strategy

Training Phase

$$P(S_p^v | S_q^a, \lambda^a, \tilde{w}_k) = \frac{1}{N} \sum_{n=1}^N \frac{\text{Count}_{aq \Rightarrow vp}^n}{N_{aq}^n}$$

$$P(S_q^a | S_p^v, \lambda^v, \tilde{w}_k) = \frac{1}{N} \sum_{n=1}^N \frac{\text{Count}_{vp \Rightarrow aq}^n}{N_{vp}^n}$$



N_{aq}^n represents the number of frames of the q^{th} audio state of n^{th} training data.

$\text{Count}_{aq \Rightarrow vp}^n$ denotes the number of frames of the q^{th} audio state aligned to the p^{th} visual state for the n^{th} training data.

N denotes the total number of training data.

EWSC-HMM

Semi-Coupled HMM (SC-HMM)

31

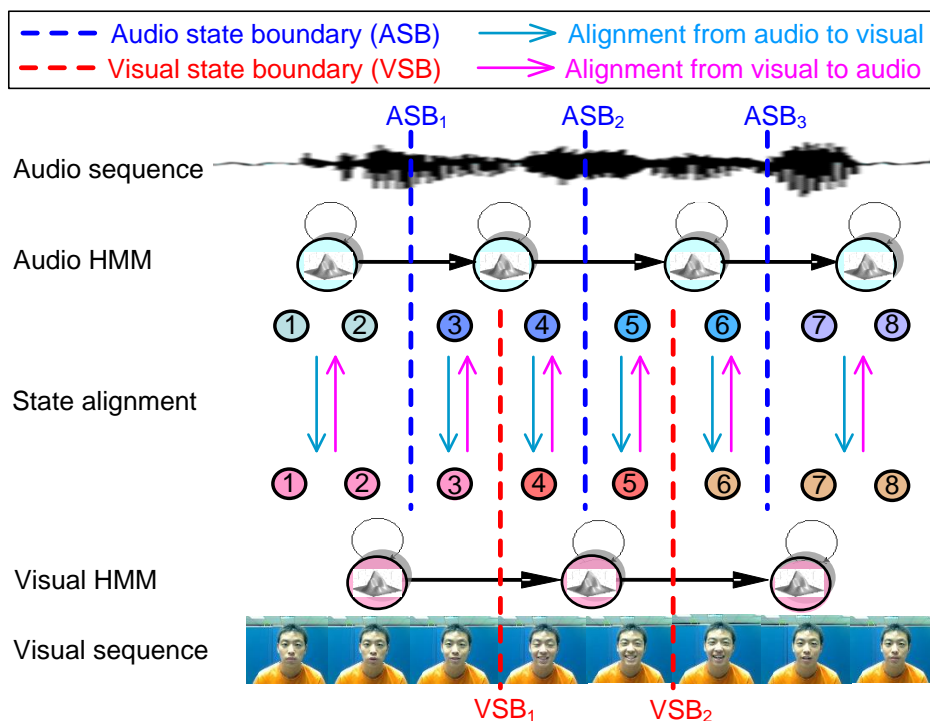
- Test Phase: The constructed alignment probability is applied for linking the two component HMMs.

Audio state aligned to visual state

$$P(S^v | S^a, \lambda^a, \tilde{w}_k) = \prod_{t=1}^T P(S_{p(t)}^v | S_{q(t)}^a, \lambda^a, \tilde{w}_k)$$

Visual state aligned to audio state

$$P(S^a | S^v, \lambda^v, \tilde{w}_k) = \prod_{t=1}^T P(S_{q(t)}^a | S_{p(t)}^v, \lambda^v, \tilde{w}_k)$$



EWSC-HMM

Empirical Weights Estimation

32

$$P(w | x^a, x^v) \approx \sum_{i=1}^C \sum_{j=1}^D \left\{ \sum_{k=1}^K \underbrace{P(w | \tilde{w}_k, \lambda_i^a, \lambda_j^v)}_{P(S^a | S^v, \lambda^v, \tilde{w}_k) P(x^v, S^v | \lambda_j^v, \tilde{w}_k)} [\max_{S^a, S^v} P(x^a, S^a | \lambda_i^a, \tilde{w}_k) P(S^v | S^a, \lambda^a, \tilde{w}_k)] \right. \\ \left. \underbrace{P(\tilde{w}_k | \lambda_i^a, \lambda_j^v)}_{P(\lambda_i^a, \lambda_j^v | x^a, x^v)} \right\} P(\lambda_i^a, \lambda_j^v | x^a, x^v)$$

Confusion Matrix of SC-HMM-based Classifier

SC-HMM \ Data	H	A	S	N
H	40	5	0	5
A	0	30	5	15
S	5	10	25	10
N	0	0	0	50
	45/200	45/200	30/200	80/200

$P(\lambda_i^a, \lambda_j^v | x^a, x^v)$ representing the confidence of the decision of classifier

$P(\tilde{w}_k | \lambda_i^a, \lambda_j^v)$ representing the probability of predicted emotion classes of classifier \tilde{w}_k

$P(w | \tilde{w}_k, \lambda_i^a, \lambda_j^v)$ representing the probability of true emotion class under each SC-HMM

Experimental Setup

33

- The MHMC database was used for performance comparison.
 - The FAPs are extracted from five facial regions: eyebrow, eye, nose, mouth, and facial contour as facial features.
 - The prosodic features including pitch, energy, and formants F1-F5 are extracted as audio features.

In total, 15 kinds of models including unimodal approaches and bimodal fusion approaches are considered for experiments

LIST AND DESCRIPTION OF ABBREVIATIONS^o

Abbreviations	Descriptions ^o
EB ^o	eyebrow HMM ^o
EY ^o	eye HMM ^o
NO ^o	nose HMM ^o
MO ^o	mouth HMM ^o
FAC ^o	facial contour HMM ^o
WF ^o	whole facial (feature-level fusion) HMM ^o
EN ^o	energy HMM ^o
PI ^o	pitch HMM ^o
FO ^o	formant HMM ^o
WP ^o	whole prosodic (feature-level fusion) HMM ^o
WFP ^o	whole facial-prosodic (feature-level fusion) HMM ^o
EWC ^o	error weighted classifier combination (decision-level fusion) ^o
C-HMM ^o	coupled HMM (model-level fusion) ^o
SC-HMM ^o	semi-coupled HMM (model-level fusion) ^o
EWSC-HMM ^o	combines semi-coupled HMM and error weighted classifier combination (combines model-level and decision-level fusion approaches) ^o

Experimental Setup

34

- For the MHMC database, four emotional states including neutrality (NEU), happiness (HAP), anger (ANG) and sadness (SAD) were considered.



Neutrality



Happiness



Anger



Sad

- The experiments were performed based on seven-fold cross-validation.
 - For each fold, each emotional state contains 360 sentences (from six subjects) for training, and 60 sentences (from the remaining subject) for testing.
 - The left-to-right topology of the HMM structure with eight hidden states was used in the approaches for comparison.

Experimental Results

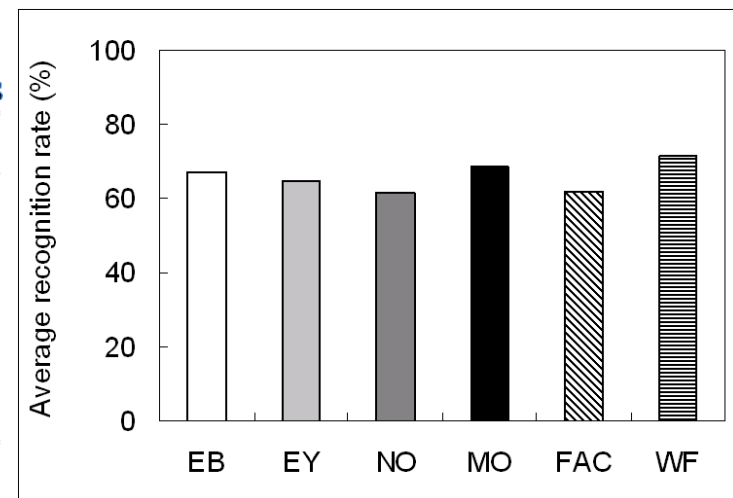
Facial Expression

35

- The recognition rates for each facial model as well as for the whole facial model are shown in the following
 - ▣ From the results different facial regions have different levels of contributions in different emotional states.
 - For example, EY had a better performance for happiness and anger, and MO had a better performance for happiness and sadness.
 - ▣ The performance of individual facial model was comparably to that the WF (feature-level fusion) model. (caused by sparse data problem; high dimensional feature sets)

RECOGNITION RATES (%) OF FOUR EMOTIONAL STATES FOR FACIAL MODELS

Model \ Emotion	NEU	HAP	ANG	SAD
EB	50.48	83.33	75.71	59.05
EY	48.81	80.48	82.62	46.43
NO	63.33	74.05	48.33	60.24
MO	66.90	83.81	46.19	77.62
FAC	43.33	75.48	58.33	70.48
WF	61.43	87.14	70.00	66.90



Average emotion recognition rates of facial models

Experimental Results

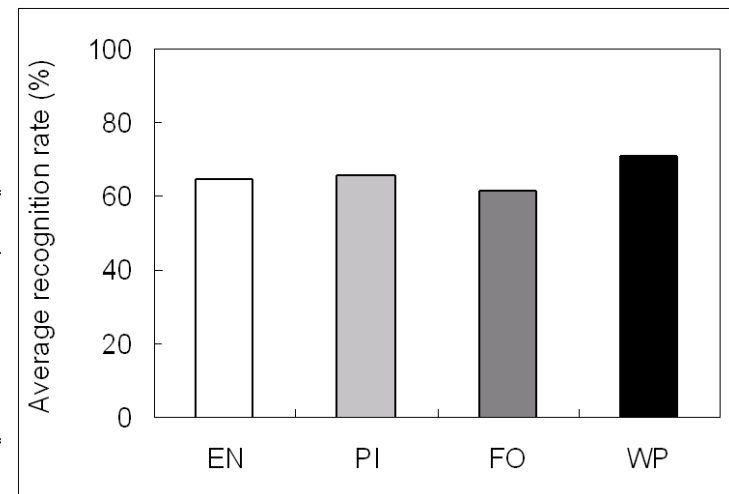
Speech

36

- The recognition rates for each prosodic model as well as for the whole prosodic (WP) model are shown in following
 - ▣ Similarly, experiments also demonstrated that different prosodic features have distinct contributions to different emotional states.
 - For example, energy (EN) contributes more to neutrality and sadness, and pitch (PI) contributes more to anger and sadness.
 - ▣ For the average emotion recognition accuracy, the performance of the whole prosodic (WP) model is also limited. (caused by sparse data problem; high dimensional feature sets)

RECOGNITION RATES (%) OF FOUR EMOTIONAL STATES FOR PROSODIC MODELS

Model \ Emotion	NEU	HAP	ANG	SAD
EN	77.38	46.90	60.71	73.33
PI	65.95	50.71	77.86	69.05
FO	58.57	67.14	56.19	63.33
WP	71.43	65.71	70.24	76.67



Average emotion recognition rates of prosodic models

Experimental Results

37

- The results demonstrated that the performance of the unimodal approach is limited, that is, recognizing human emotional state just from single facial or vocal modality is not enough.
- **The most important finding** from unimodal features analysis suggests that different features have distinct contributions to different emotional states.
 - ▣ These findings conclude that a bimodal fusion approach should be considered (e.g., decision-level fusion) in order to utilize the characteristics of different input modalities.

Experimental Results

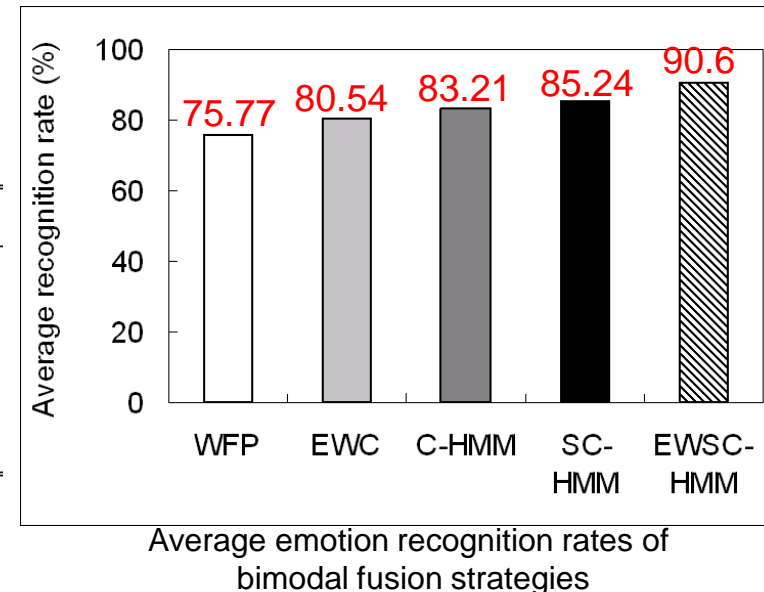
Fusion

38

- Based on the above analyses, the proposed SC-HMM is further combined with EWC to achieve a better recognition result.
- The proposed EWSC-HMM can achieve a better recognition accuracy than the current fusion strategies.
 - ▣ The findings support the claim that considering the correlated multiple streams between audio and visual sequences and the contribution of different feature pairs simultaneously are useful.

RECOGNITION RATES (%) OF FOUR EMOTIONAL STATES FOR BIMODAL FUSION COMPARISON[⌘]

Model \ Emotion [⌘]	NEU [⌘]	HAP [⌘]	ANG [⌘]	SAD [⌘]
WFP [⌘]	71.67 [⌘]	82.14 [⌘]	72.38 [⌘]	76.90 [⌘]
EWC [⌘]	76.90 [⌘]	88.81 [⌘]	75.95 [⌘]	80.48 [⌘]
C-HMM [⌘]	80.71 [⌘]	87.38 [⌘]	82.62 [⌘]	82.14 [⌘]
SC-HMM [⌘]	82.38 [⌘]	88.57 [⌘]	84.29 [⌘]	85.71 [⌘]
EWSC-HMM [⌘]	86.90 [⌘]	95.71 [⌘]	88.81 [⌘]	90.95 [⌘]



Discussion

39

- Three findings are summarized from the experiments:
 - ▣ The results from unimodal features analysis indicate that exploring the role of different facial and prosodic features for further bimodal fusion is important; different features have distinct contributions to different emotions.
 - ▣ Combining audio and visual cues is useful to improve the performance of emotion recognition based on decision-level fusion, model-level fusion or hybrid approach (i.e., EWSC-HMM).
 - ▣ Since the collection of emotional data is not easy and the high dimensional feature set coming from audiovisual modalities, data sparseness is a significant problem for emotion recognition.
- Future research to recognize an expanded set of emotion categories and explore naturalistic database are envisioned.

Acknowledgements

40

- Special thanks to my students supporting related research, especially **Jen-Chun Lin** and **Wen-Li Wei** for their great contribution.



Conclusions

41

- Multi-modal features are helpful to improve the detection performance.
 - Facial Expression
 - Speech/Vocal expression
 - Text
 - Gesture
 - Bio-signals
 - Brain signal, skin temperature, blood pressure, heart rate, respiration rate
- Personalization
- Collaboration among affection researchers from different disciplines



Thank
you

iStockphoto

Questions?

