Recognition of Emotion in Speech Using Structural and Temporal Information

Chung-Hsien Wu

Professor

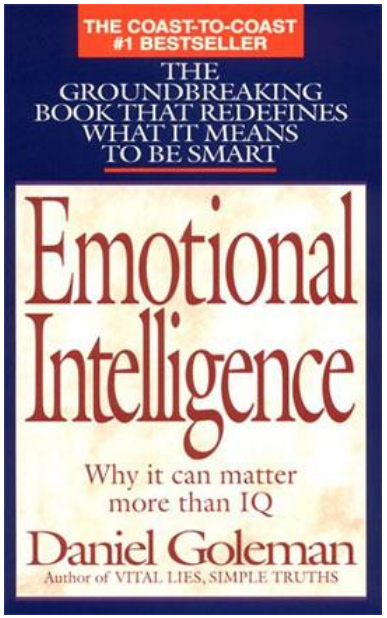
Department of Computer Science and Information Engineering
National Cheng Kung University, Tainan, TAIWAN

2013 Speech Workshop (APSIPA Distinguished Lecture), Chungli, Taiwan

Introduction

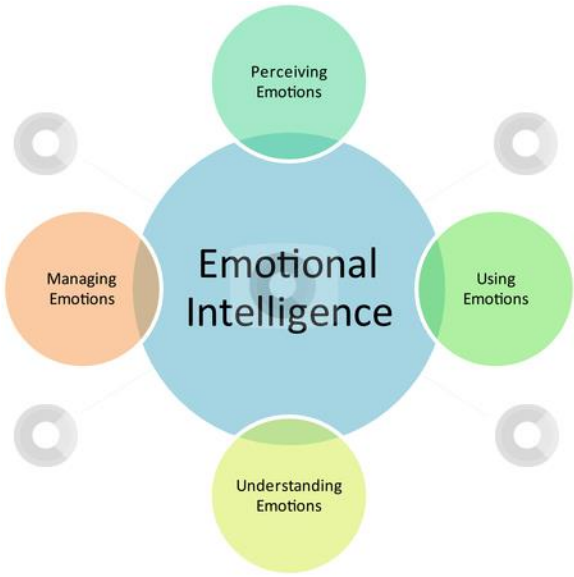
- Emotional intelligence has become more and more important to Social Interaction.
  - People with high emotional intelligence are usually successful in most things they do.

- *Emotional Intelligence* by Daniel Goleman was on *The New York Times* bestseller list for a year-and-a-half; with more than 5,000,000 copies in print worldwide in 40 languages.



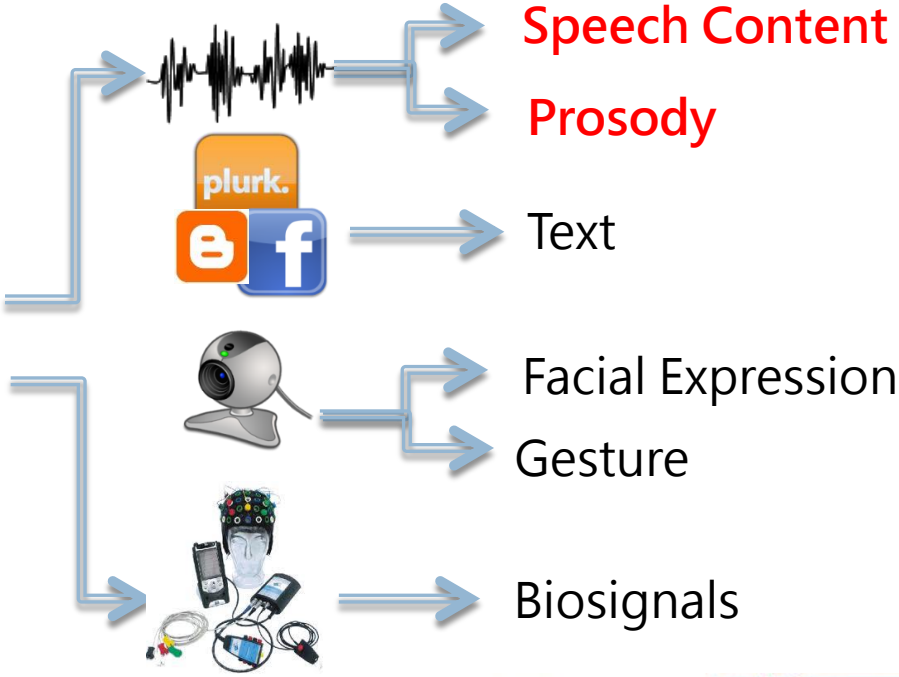
Introduction

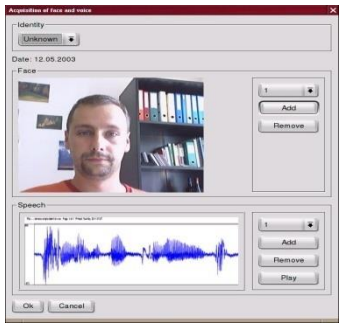
- Emotional Intelligence includes four types of abilities (From Wikipedia):
  - Perceiving emotions
  - Using emotions
  - Understanding emotions
  - Managing emotions

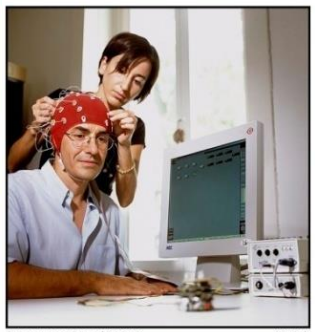


- A growing interest in the recognition of users' emotions in the interaction with machines can be observed.

Emotion Expression Through Social Behaviors







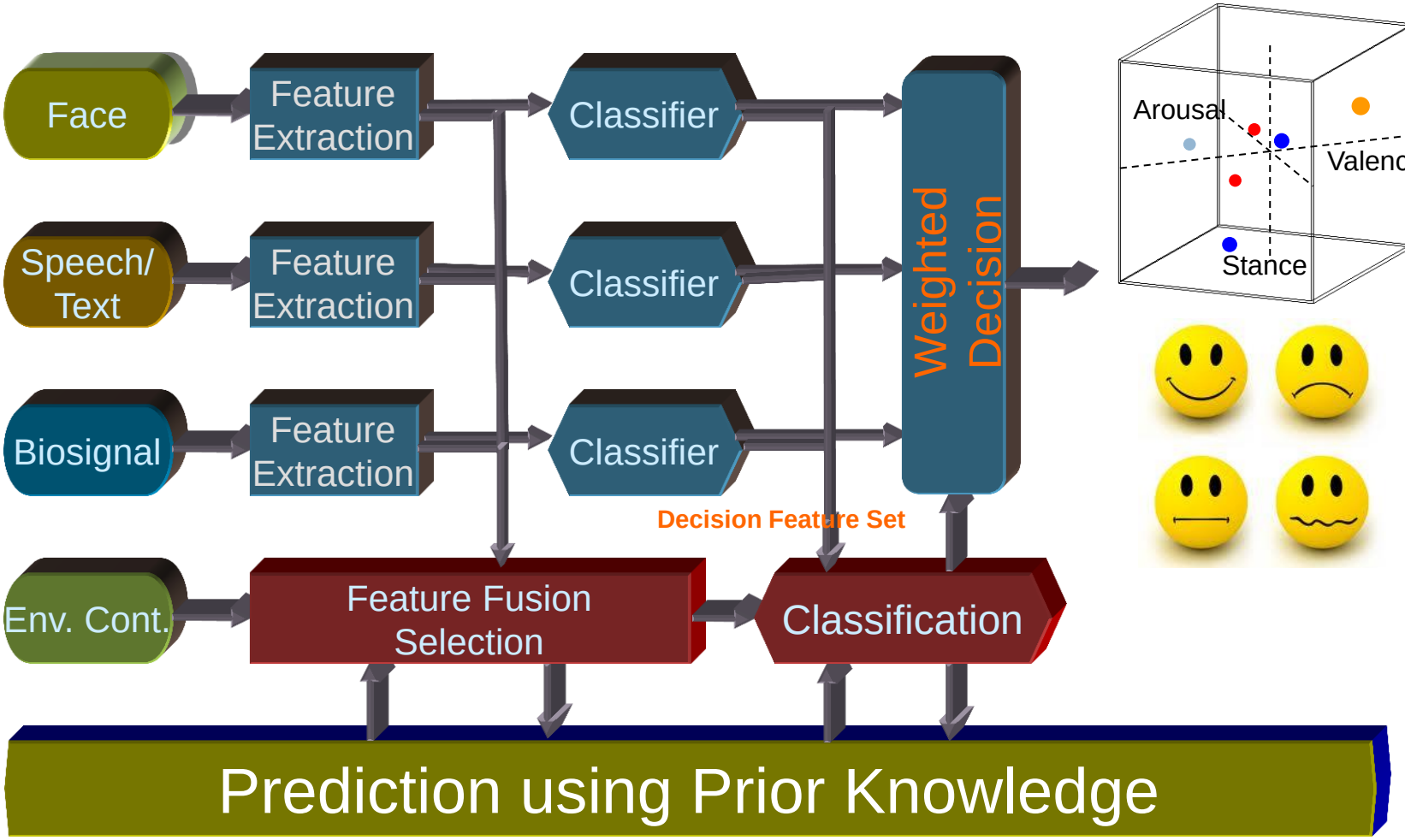
Why Speech Signal?

- “Speech” is the most natural way for people to interact with others
  - People can exchange information and emotions rapidly through speech signals
  - Speech signal can be received more easily through mobile devices and other media
- While speech analysis seems to be most promising, we focus on speech as input channel in this talk.

Applications

- Speech emotion recognition is particularly useful for applications which require natural man–machine interaction such as
  - Entertainment
  - e-Learning
  - In-car board system
  - Diagnostic tool for therapists
  - Call center applications
  - Mobile communication
  - ….

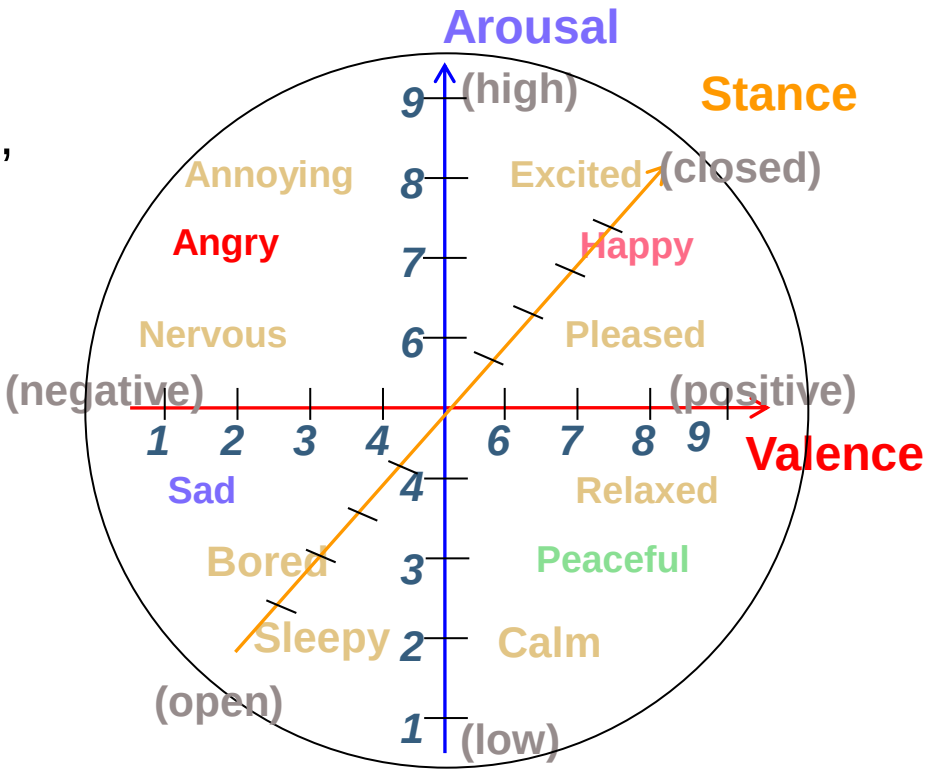
Block Diagram for Multisensory Emotion Recognition



Perspectives on Affection Description

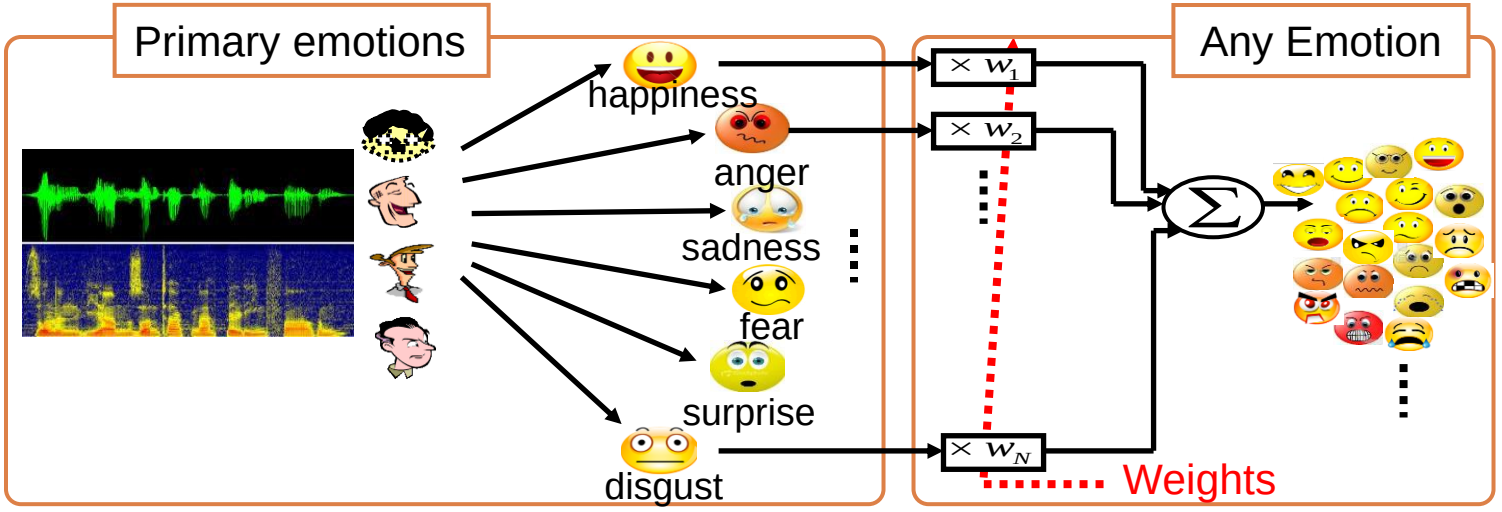
- Dimensional Description
  - Arousal, Valance, Stance (Dominance)

- Arousal: Excited at high arousal, Sleepy at low arousal
- Valence: Pleased at positive valence, Sad at negative valence
- Stance: Stern at closed stance, accepting at open stance



Perspectives on Affection Description

- Categorical Emotion Description
  - Happiness, anger, sadness, fear, surprise, disgust
  - "**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).



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 Features

- 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.

Picard, R. 1995. *Affective Computing.* The MIT Press
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.

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 reconition in call-centres", Speech Commun., 2007

Rosalind W. Picard., Affective Computing MIT Press, 1997

Methods for Emotion Recognition

- 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.

Challenges

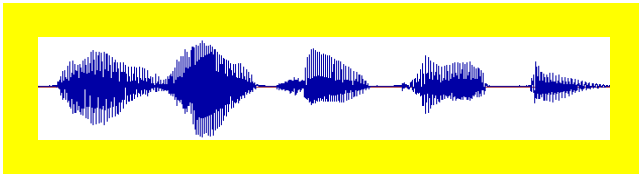
- Which speech features are most powerful in distinguishing between emotions
- Emotion is expressed generally depends on the speaker, his or her culture and environment
- The long-term emotion or the transient one
- Emotion does not have a commonly agreed theoretical definition on structure and time

➡ However, people know emotions when they feel them.

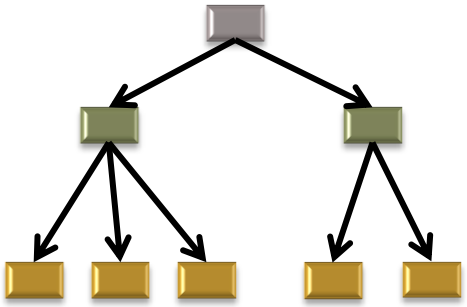
➡ For this reason, researchers were able to study and define different aspects of emotions.

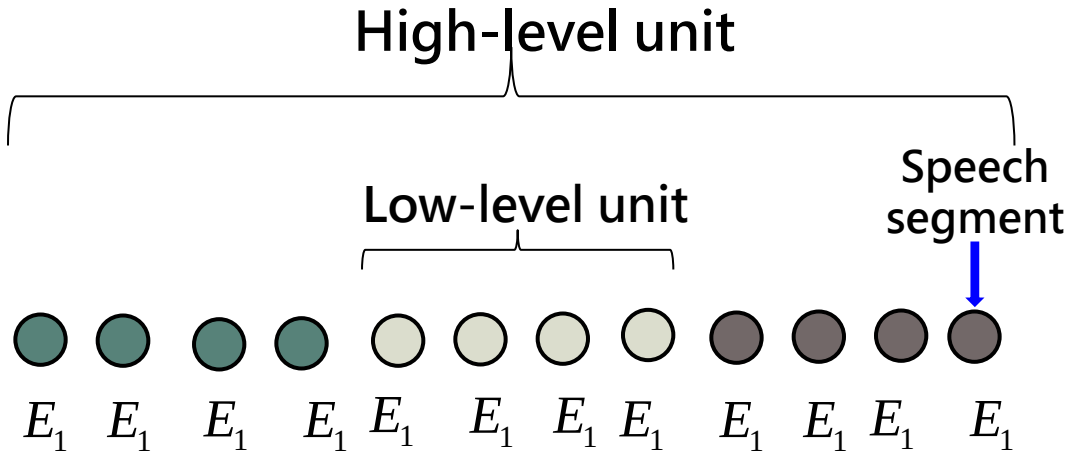
Interpretation of Emotion Expression

- A two-dimensional interpretation will be discussed in this talk
  - Temporal



  - Structural





Emotion Recognition from Temporal Interpretation

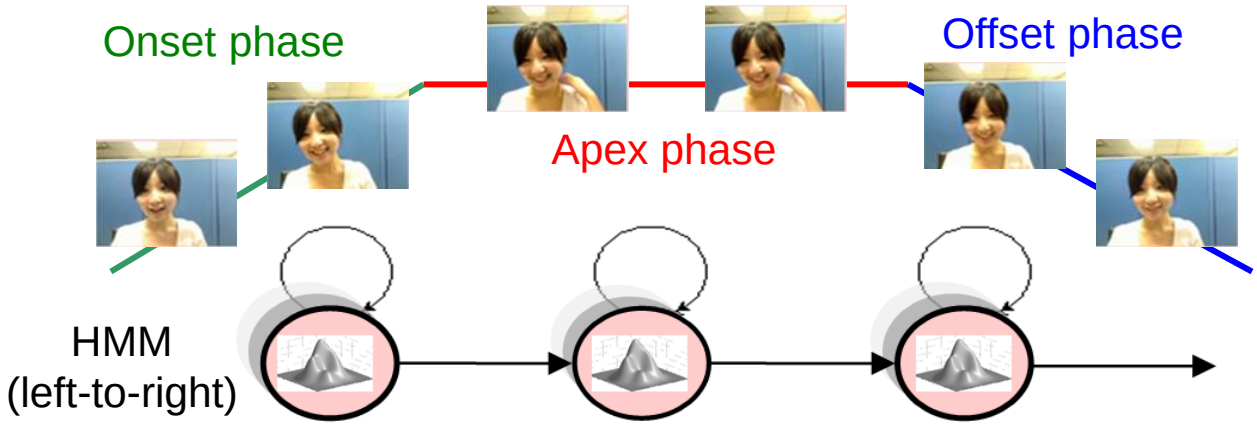
- Previous research showed that a complete emotional expression can be characterized in three sequential temporal phases:
  - onset (application), apex (release), and offset (relaxation), when considering the manner and intensity of an expression.

Temporal course of emotional expression

Ekman, P., Handbook of Cognition and Emotion. Wiley, 1999.
Picard, R. W., Affective Computing. MIT Press, 1997.
Valstar, M. F. and Pantic, M., "Fully automatic recognition of the temporal phases of facial actions", IEEE Trans. Systems, Man and Cybernetics–Part B, 42(1):28-43, 2012.

Motivation

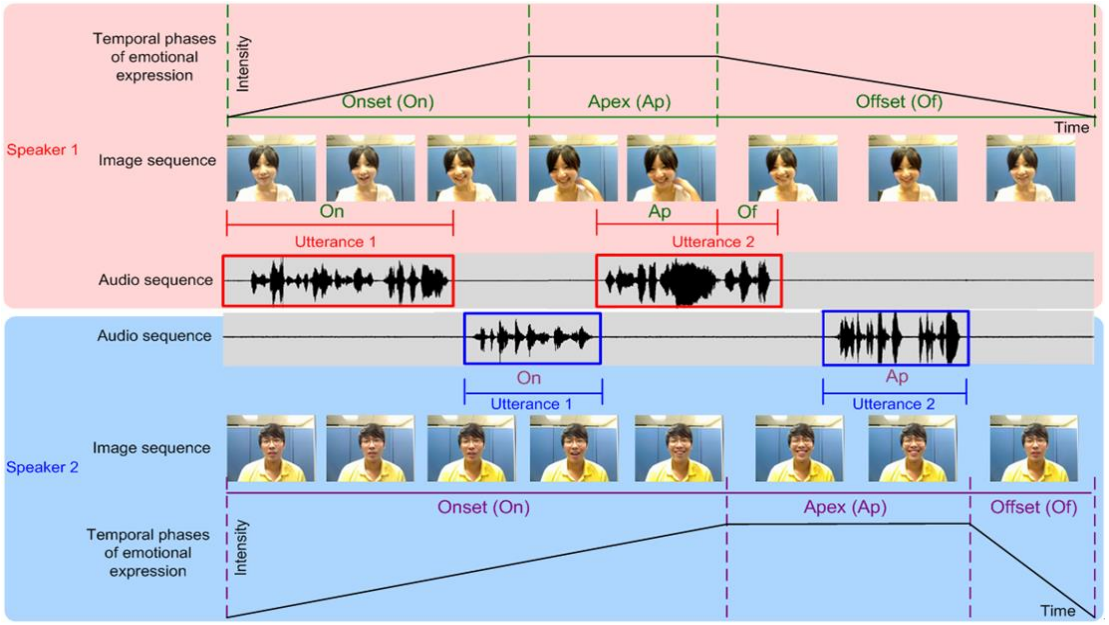
- To capture the temporal information, the hidden Markov model (HMM) have been investigated.
  - The left-to-right topology of the HMM structure was mostly used for describing the temporal courses of emotional expressions.

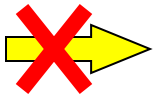


Schuller, B., Rigoll, G. and Lang, M., “Hidden Markov model-based speech emotion recognition”, Proc. Int’l Conf. Acoustics, Speech, and Signal Processing (ICASSP), II 1-4, 2003.
Ntalampiras, S. and Fakotakis. N., “Modeling the temporal evolution of acoustic parameters for speech emotion recognition”, IEEE Trans. Affective Computing, 3(1):116-125, 2012.
Lin, J. C., Wu, C. H. and Wei, W. L., “Error weighted semi-coupled hidden Markov model for audio-visual emotion recognition”, IEEE Trans. Multimedia, 14(1):142-156, 2012.

Temporal Information Modeled by a Single HMM

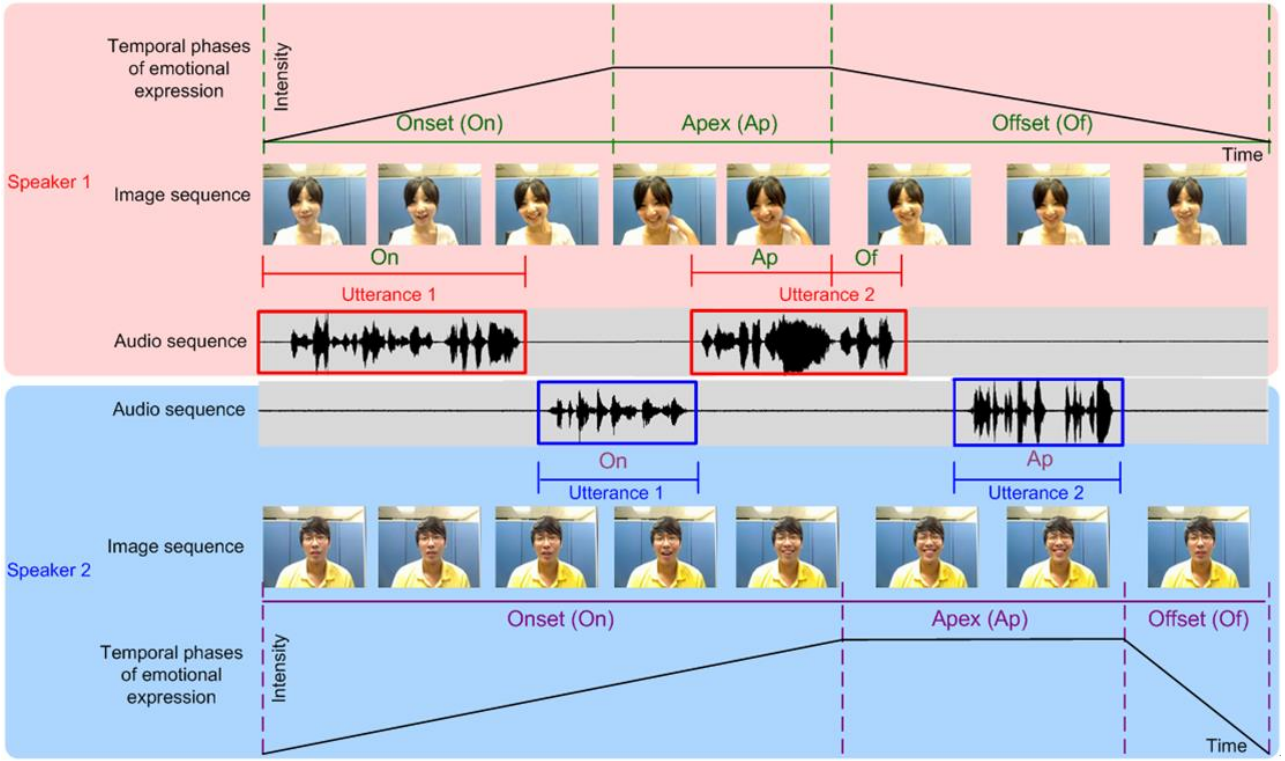
- A complete emotional expression is expressed by more than one utterance in natural conversation and each utterance may contain several temporal phases of emotional expression.



- A single HMM  Temporal course of an emotional expression

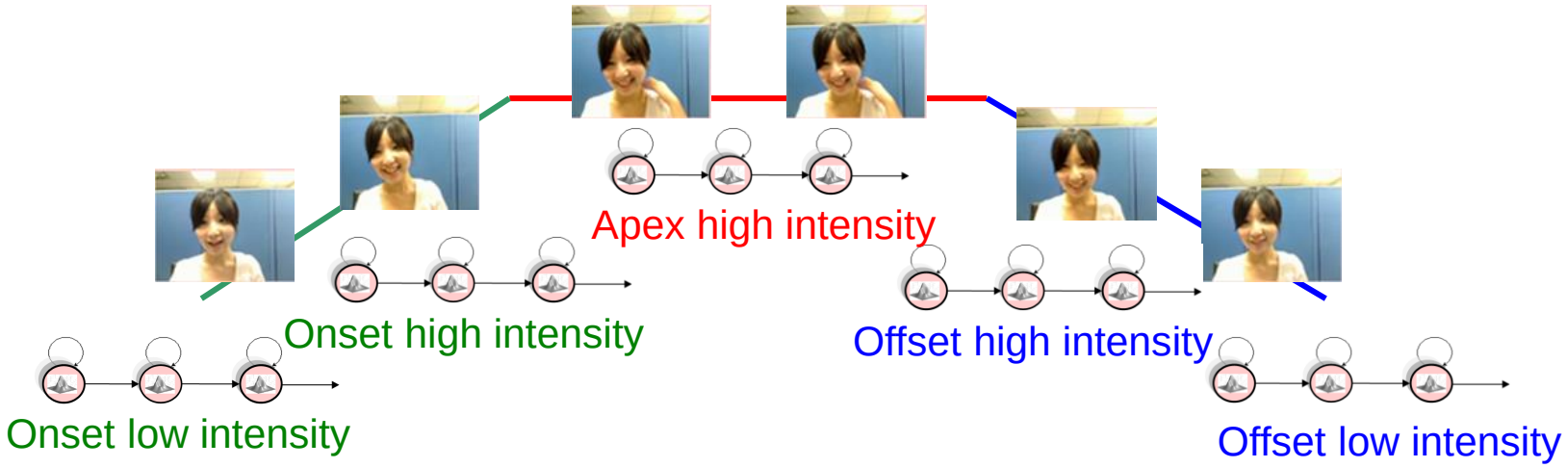
Temporal Information Modeled by a Single HMM

- When the emotional state of Speaker 1 is evoked through conversation, Utterance 1 only covers the temporal phase of onset, while the apex and offset phases are covered in Utterance 2.



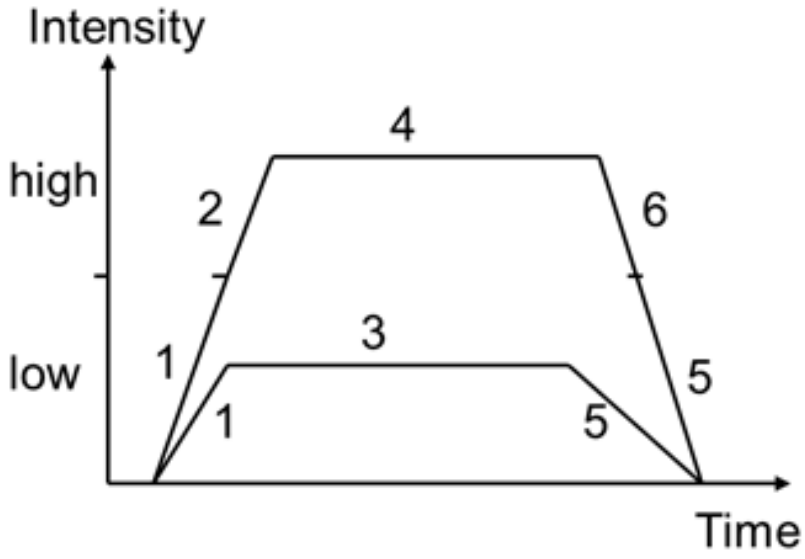
Temporal Course Modeling

- Each isolated sentence in a conversation can express one or several sub-emotional states, which are defined to represent the temporal phases (i.e., onset, apex, or offset with high or low intensity.
- An HMM is used to characterize one sub-emotional state, instead of the entire emotional state.



Intensity of Emotional Expression

- For each temporal phase (sub-emotional state), the low or high intensity is further considered on temporal course modeling.

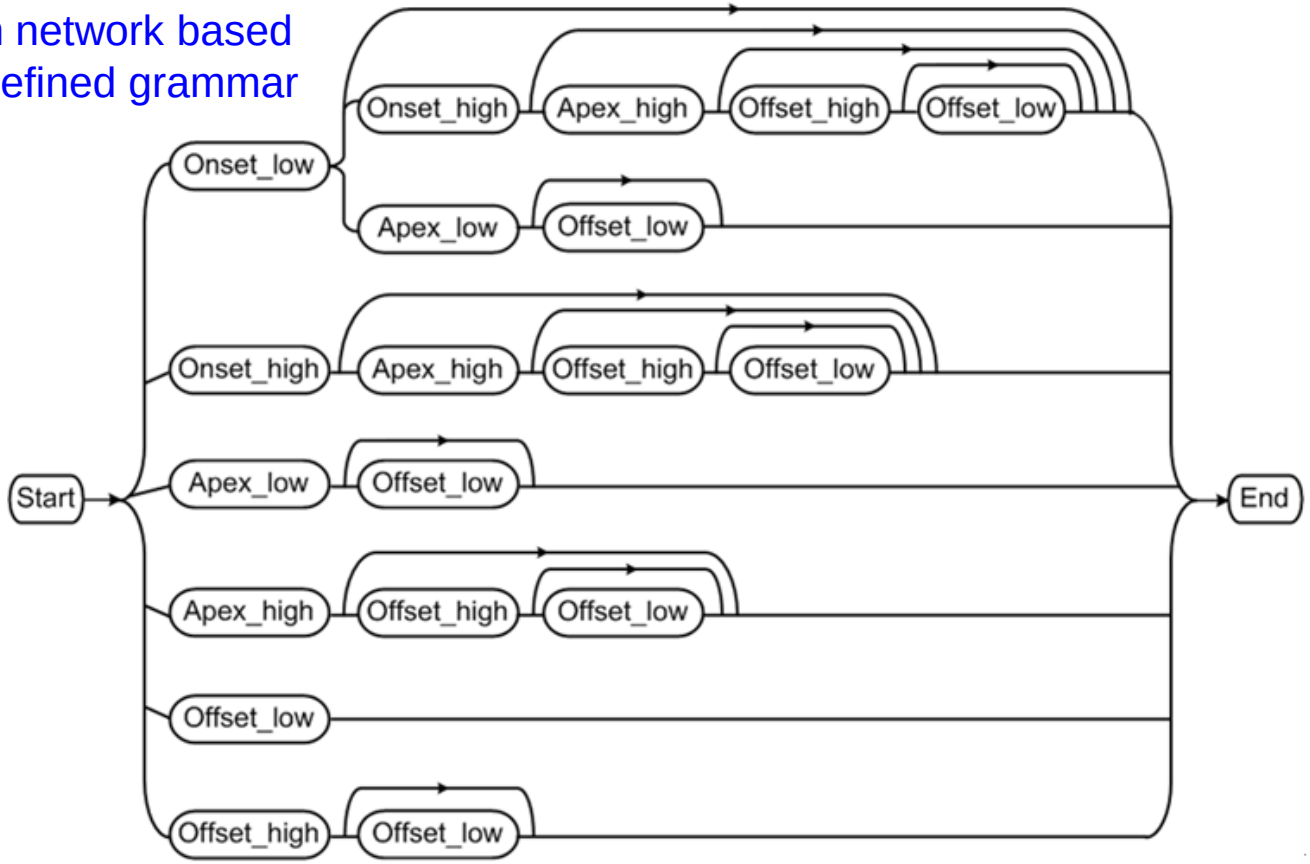


1 Onset_low
2 Onset_high
3 Apex_low
4 Apex_high
5 Offset_low
6 Offset_high

Temporal Course Modeling

- The sub-emotion language model is proposed and integrated with the recognition procedure.

Recognition network based on the predefined grammar



Model Derivation of Temporal Course Modeling

- The recognition task with four emotional states, happy, angry, sad, and neutral, represented by $E \in \{H, A, S, N\}$, is considered.

- Each emotional state of happy, angry, and sad for each isolated sentence is further expressed by an *M*-temporal phase sequence, denoted as $H = h_1^M = h_1, h_2, ..., h_M$, $A = a_1^M = a_1, a_2, ..., a_M$ and $S = s_1^M = s_1, s_2, ..., s_M$, where *M* ranges from 1 to 5 considering all possible transitions in the predefined grammar.

- Given the observation sequence $O = o_1^T = o_1, o_2, ..., o_T$, the probability of an emotional state with temporal phase sequence *E* can be estimated using (1),

$$\hat{E} = \arg\max_{E} P(E \mid O) \qquad (1)$$

Model Derivation of Temporal Course Modeling

- The *a posteriori* probability $P(E \mid O)$ can be further decomposed using the Bayes' rule as follows:

$$P(E \mid O) = P(O \mid E)P(E)/P(O) \qquad (2)$$

Likelihood; calculated using the corresponding sub-emotion HMM sequence

***Prior* probability;** estimated by the sub-emotion language model.

The same for all E, and can thus be omitted.

- Hence, (1) can be rewritten as (3) for emotion recognition using the proposed temporal course modeling.

$$\hat{E} = \arg\max_{E} P(O \mid E)P(E) \qquad (3)$$

  - where $P(E)$ is modeled by using a bigram language model as

$$P(E) = P(e_1, e_2, ..., e_M) = \prod_{k=2}^{M} P(e_k \mid e_{k\text{-}1}) \qquad (4)$$

Experimental Setup

- The MHMC conversation-based affective speech corpus was used for performance evaluation.
  - Provided by 53 students of both genders
  - A total of 2,120 utterances were collected
- The subjective tests were performed to set the ground truth of emotional expression for the recorded data.
  - Three annotators were recruited, and asked to give an opinion on the emotion label for the recorded data.
  - If less than two annotators reached an agreement, the data was not included in the experiment.
- Same as the process of emotion labeling, the subjective tests were also performed to set the ground truth of the temporal phases for the recorded data.

Experimental Setup

- A total of 1,114 data, which passed the evaluation (i.e., simultaneously passed the emotion and temporal phase labeling procedure), were regarded as the ground truth data for the ensuing experiments.

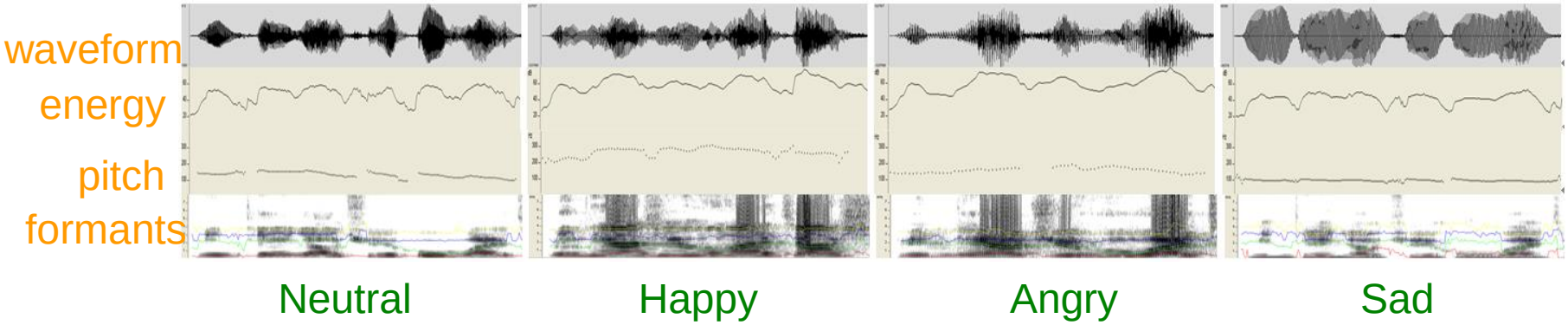
The number of ground truth utterances of the four emotional states

	Happy	Angry	Sad	Neutral
# (Utterances)	199	236	214	465

- Three classifiers were considered:
  - Support Vector Machine (SVM) with radial basis kernel function
  - Single HMM with left-to-right topology and eight states
  - Multiple HMMs: each HMM with left-to-right topology and three states for modeling each temporal phase.

Experimental Setup

- Three types of primary prosodic features, including pitch, energy and formants F1-F5 were used.
  - For SVM, the global features in which the minimum, mean, and maximum of the extracted prosodic features of the entire utterance were considered.
  - For HMM, the local features in which the prosodic features of every frame were used.



- In the experiments, 80% of the ground truth utterances were randomly selected for training, and the remaining utterances were selected for testing.

Experimental Results

- The average recognition accuracies for three approaches are shown in the following table

Average emotion recognition rates of four emotional states.

Models	SVM	Traditional HMM	Proposed
Accuracy	50.22%	56.50%	79.82%

best

- Two findings are summarized as follows:
  - Temporal phase model and the sub-emotion language model are able to better describe the complex temporal structure of emotional expression in natural conversation.
  - Expression intensity is helpful for reducing the variations in the statistical model parameters.
    - For example, since the introvert is often bashful, the expressed emotions are often accompanied with lower expression intensity.
    - The expression style may lead to the large variations in the statistical model parameters.

Summary

- This study presented an approach to automatic recognition of four emotional states from conversational speech signals using HMM-based temporal course modeling.

- Two findings are summarized from our experiments.
  - Temporal information is important for emotion recognition
    - Modeling complex temporal structure of emotional expression is useful to improve the recognition accuracy.
  - Expression intensity is important
    - Expression styles, such as expression intensity between introverts and extroverts, are significantly different for the same emotional state.

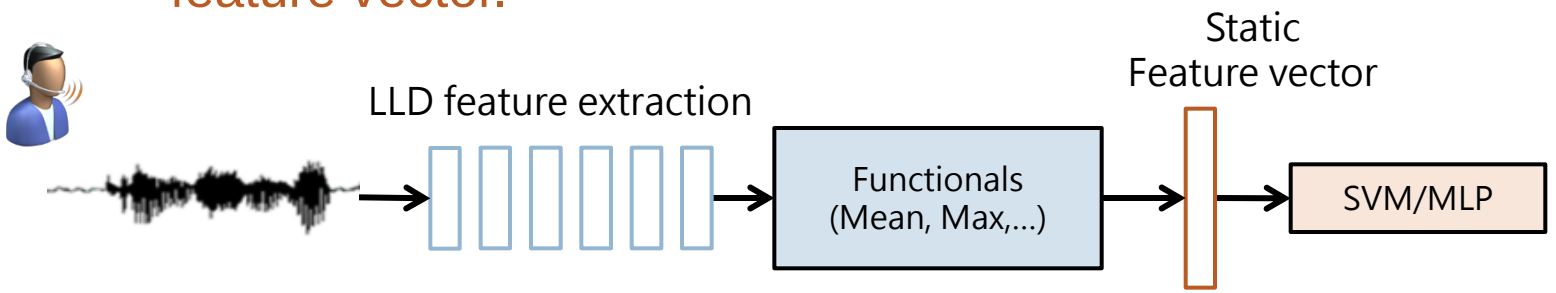
Temporal and Structural Interpretation

- Static vs. Dynamic Modeling
  - Turn-wise statistics of acoustic Low-Level-Descriptors followed by static classification has also been a dominant approach for speech emotion recognition.
  - It is well known that important information on temporal sub-turn layers exists.
  - It is desirable to integrate information on diverse time levels.

Bogdan Vlasenko, Björn Schuller, Andreas Wendemuth, and Gerhard Rigoll, “Frame vs. Turn-Level: Emotion Recognition from Speech Considering Static and Dynamic Processing,” ACII '07 Proceedings of the 2nd international conference on Affective Computing and Intelligent Interaction.

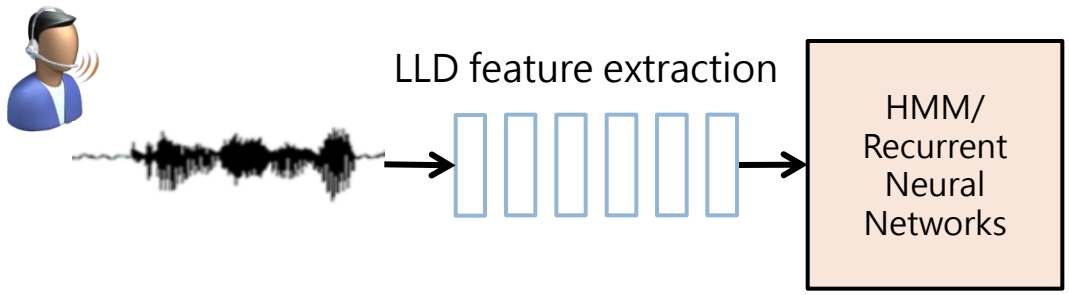
Related Work

- Static modeling approach [Ververidis, 2005][Schuller, 2009]
  - Pros:
    - Empirically, static approach has better performance in distinguishing between high-arousal emotions, e.g. anger versus sadness.
  - Cons:
    - Temporal information is lost.
    - It may be unreliable to train the classifier using the long static feature vector.



Related Work

- Dynamic modeling approach [Shuller, 2003][Nwe, 2003]
  - Pros:
    - Temporal information is considered.
    - Complex classifier can be trained reliably using large number of local feature vectors.
  - Cons
    - Due to the variation of the content in the utterance, it is often difficult to model the context well (over-modeling).



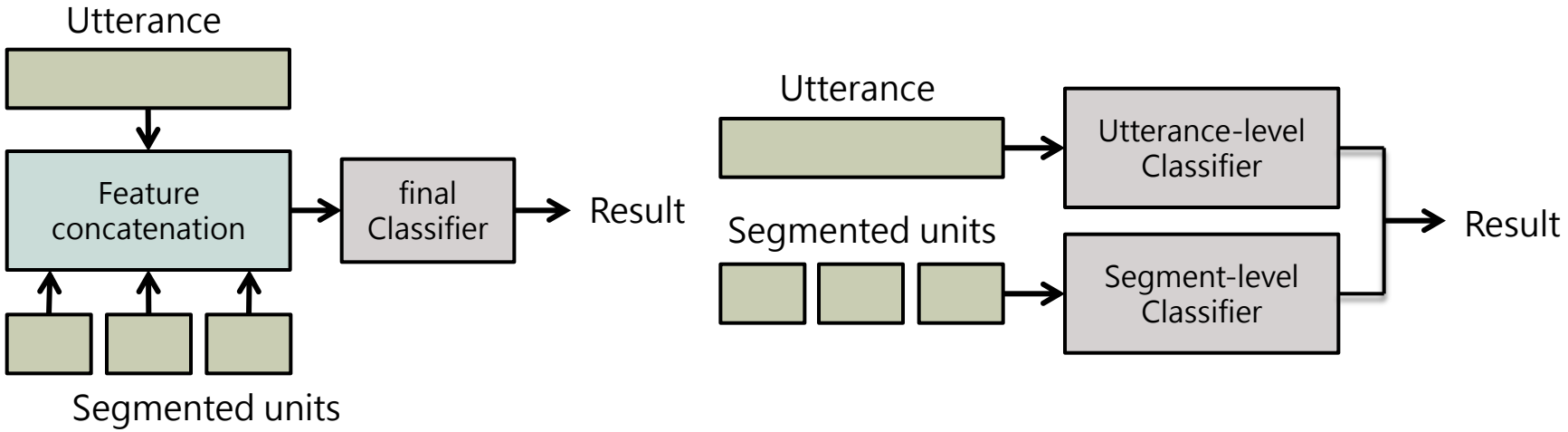
Related Work

- Different aspects of speech take place at different time scales [Fernandaz, 2011]
  - Each time scale is complementary in recognizing emotions [Han, 2009]
  - **Static features**, such as mean and range of F0, appear to be associated with the arousal aspect of emotion
  - **Temporal features** may be more relevant to the communication of valence, attitude, or intention [Jiang, 2004]

- Combination of the features from different time scales might improves the recognition performance

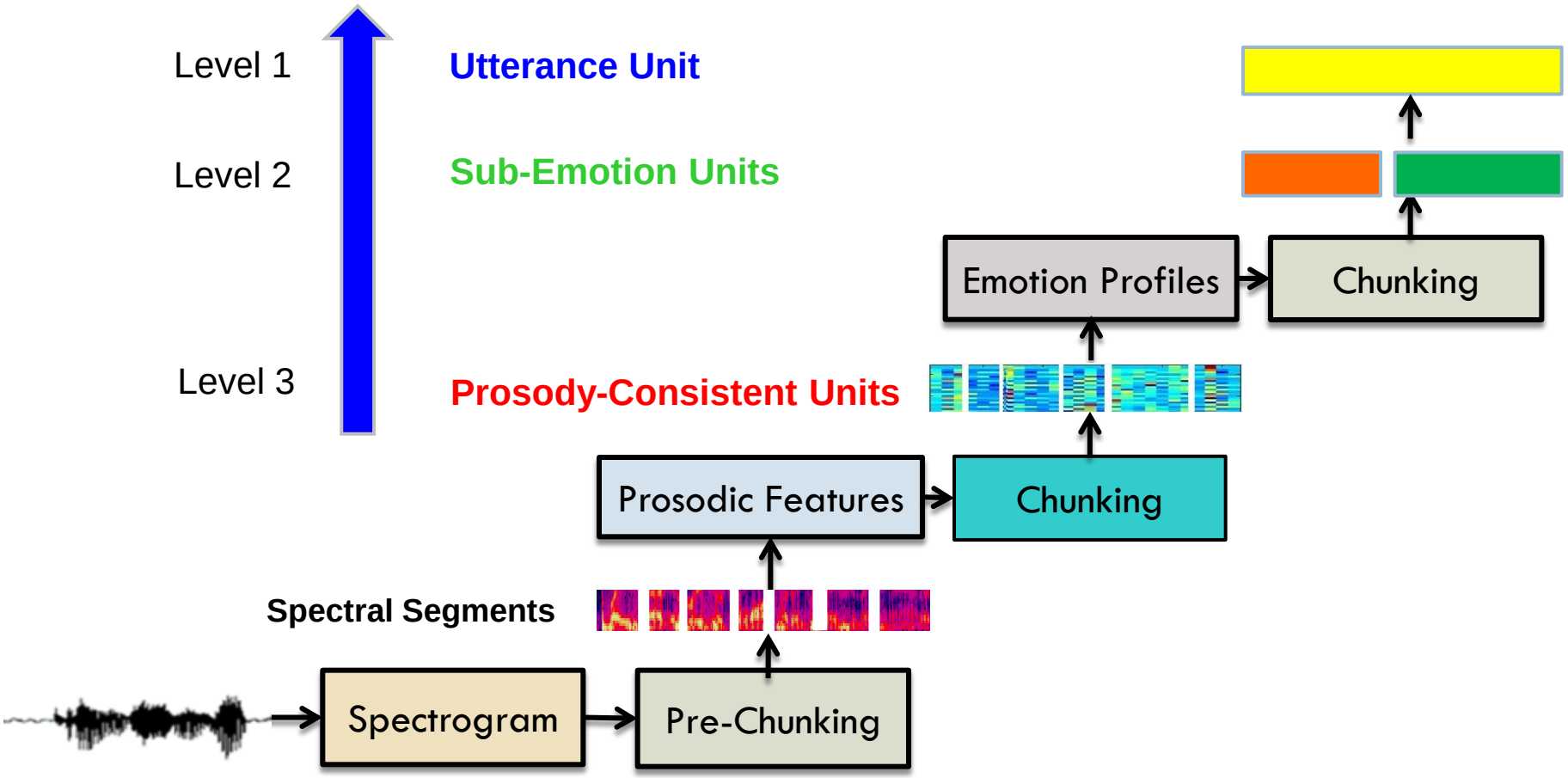
Related Work

- Problems in previous modeling strategy
  - Previous work just simply used **feature concatenation** or **decision-fusion** to exploit the information from different time scales
    - longer concatenated feature vectors would result in data sparseness.
    - Decision-fusion approaches assume that time scales are mutually independent, which is not convincing



Proposed Approach
Hierarchical Structure for Emotion

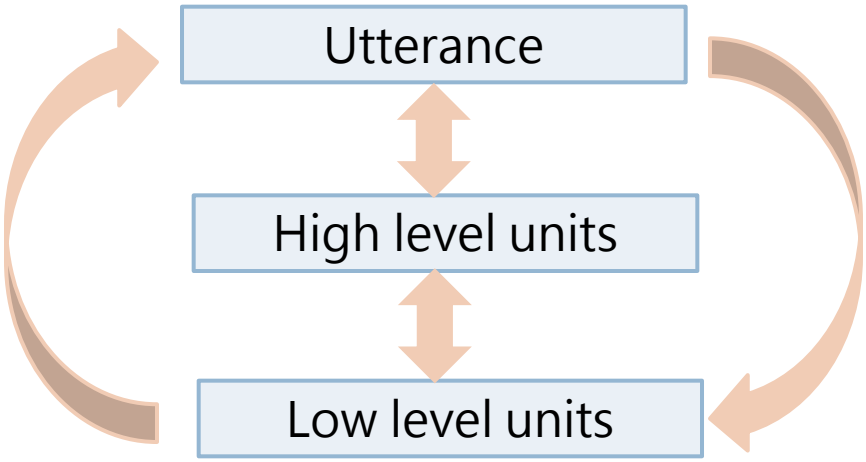
- Hierarchical Chunking for the input frame sequence



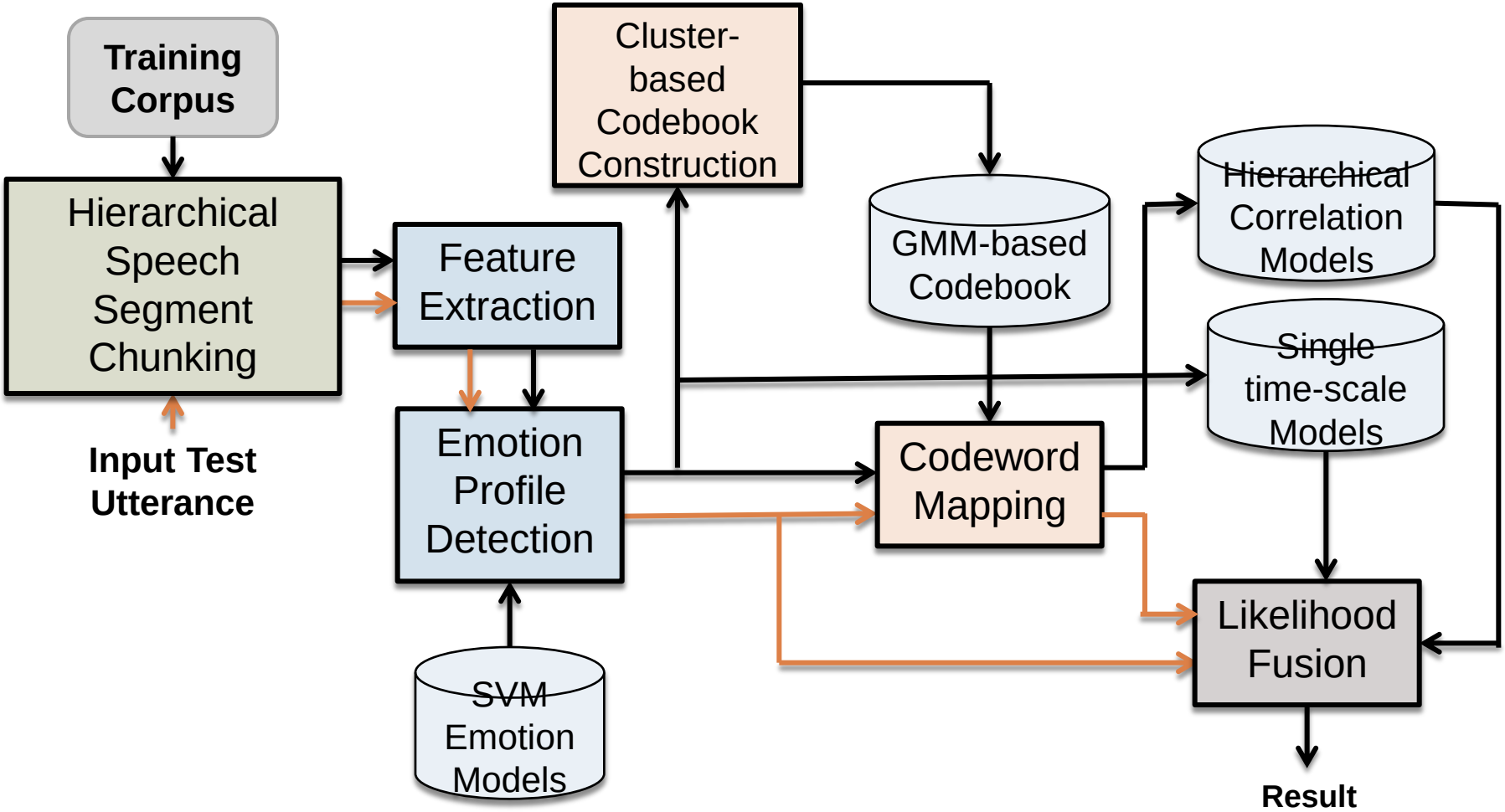
Proposed Approach

Hierarchical Correlation Model

- A Hierarchical Correlation Model to capture the correlation between each time scales

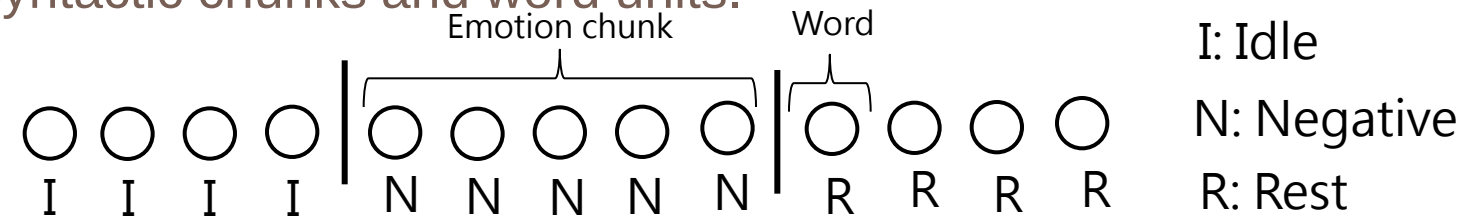


System Block Diagram



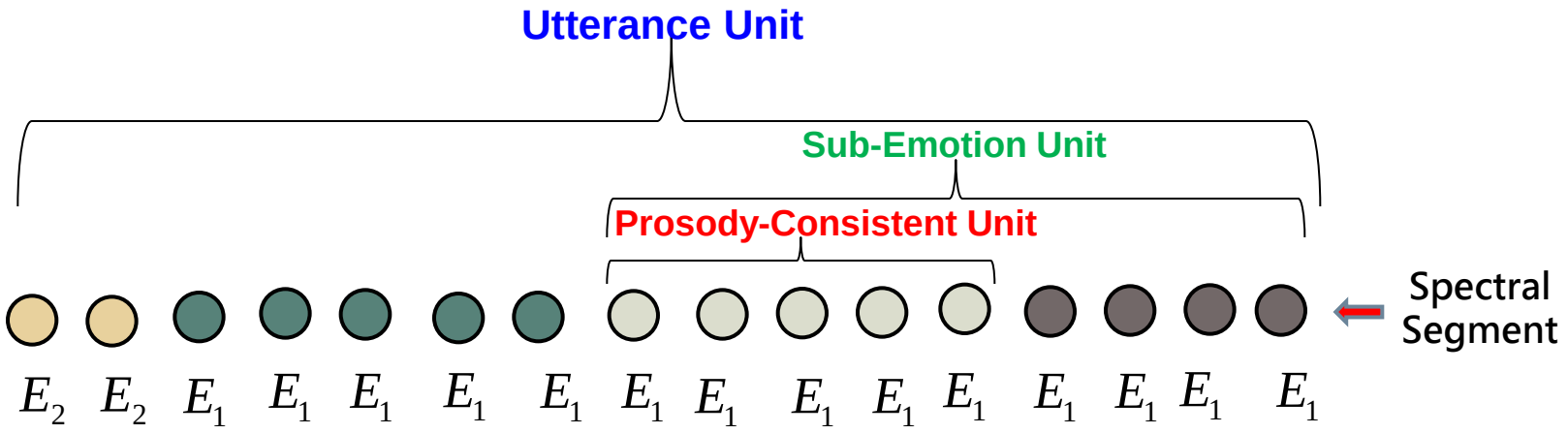
Hierarchical Speech Segment Chunking

- In some previous work, word or syllable is considered as the lowest-level basic units [Jiang, 2004][Batliner, 2010][Jeon, 2011]
  - They assume that the variation inside each basic unit is comparatively less, so the extraction of static features would not reduce too much temporal information
- In [Batliner, 2010], they used word-based annotation corpus and combined the word sequence into label-consistent chunks.
  - The use of emotion-consistent chunks performs better than syntactic chunks and word units.



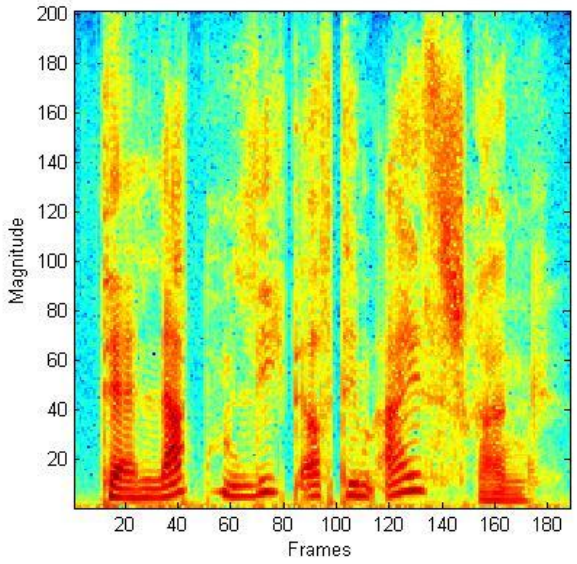
Hierarchical Speech Segment Chunking

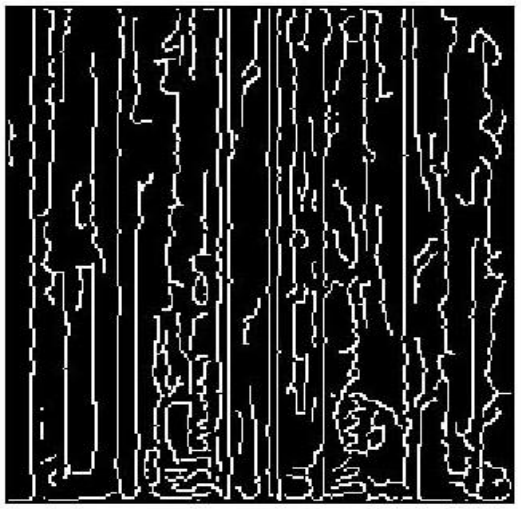
- There exist hierarchical structures in an emotion [Batliner, 2010].
  - Locate the spectral segments with similar spectral property
  - First, chunk the spectral segments with similar prosodic features into prosody-consistent units.
  - Second, chunk the prosody-consistent units with similar emotion expression into sub-emotion units.
  - Finally, utterance unit is considered.



Speech Segment Chunking

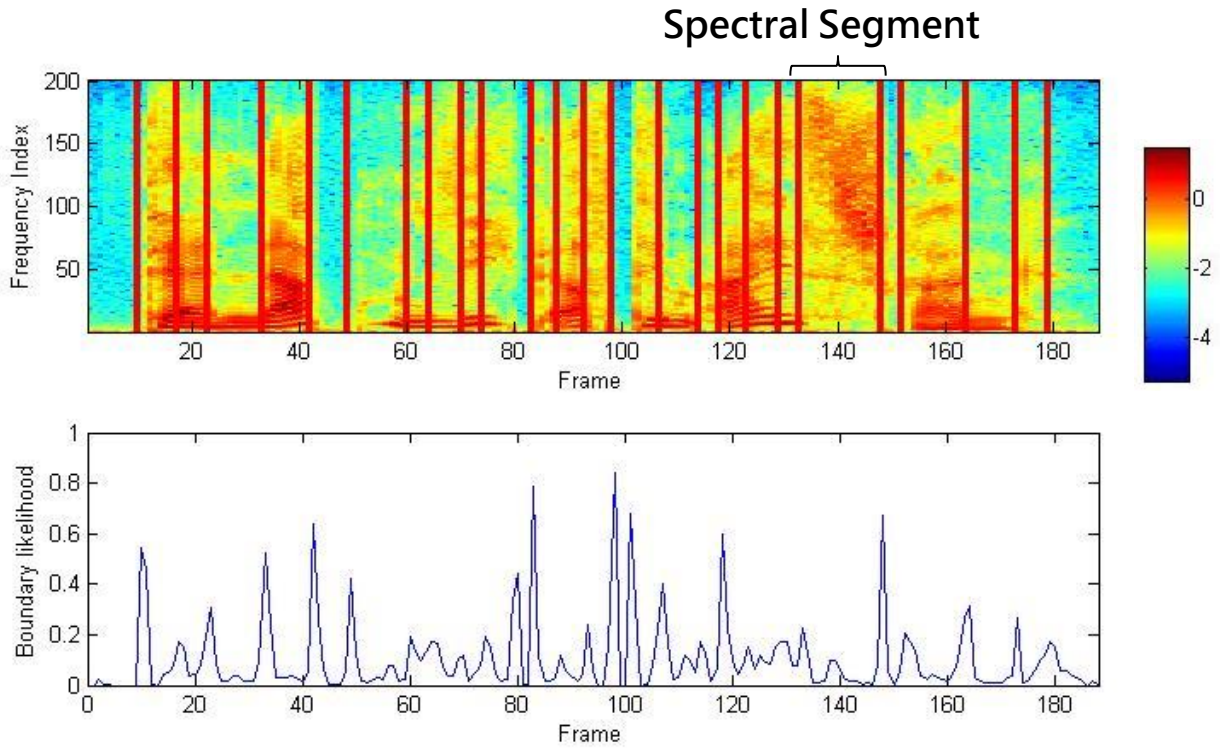
- Speech segment chunking using Canny's algorithm [Fernandaz, 2004]
  - Canny's algorithm [Canny, 1986]
    - Gradient Estimation
    - No-Maximum suppression
    - Hysteresis thresholding





Spectral Segment

- Spectral speech segment for pre-chunking
  - Consecutive speech frames with stationary spectral properties are regarded as a spectral segment

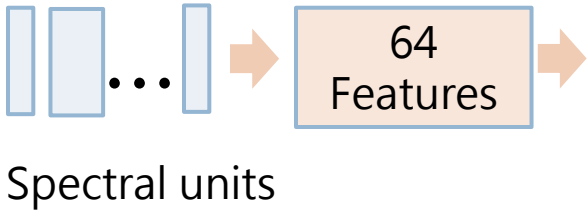


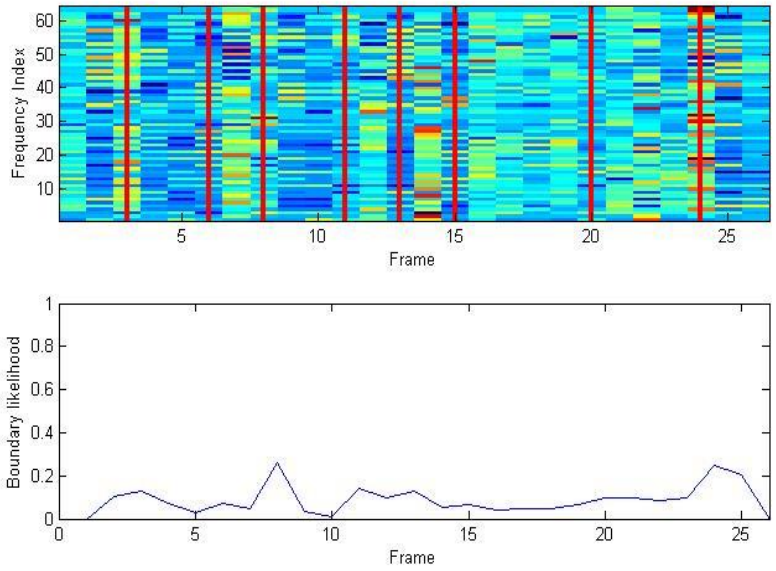
Hierarchical Speech Unit Chunking

- For each spectral segment, 64 prosody-related features were extracted

LLD (16*2)	Functionals (2)
ZCR, RMS Energy, F0, HNR, MFCC 1-12, and their Delta	Mean, standard deviation

- Chunk the utterance into prosody-consistent units based on 64 features

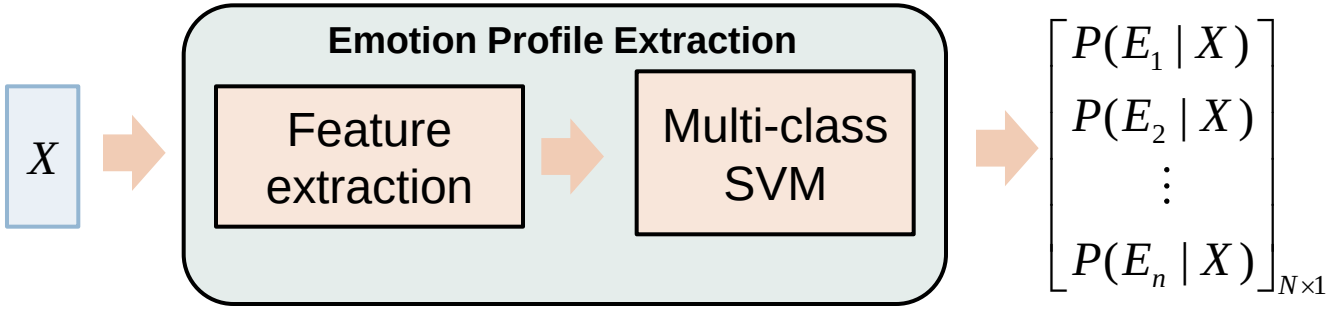




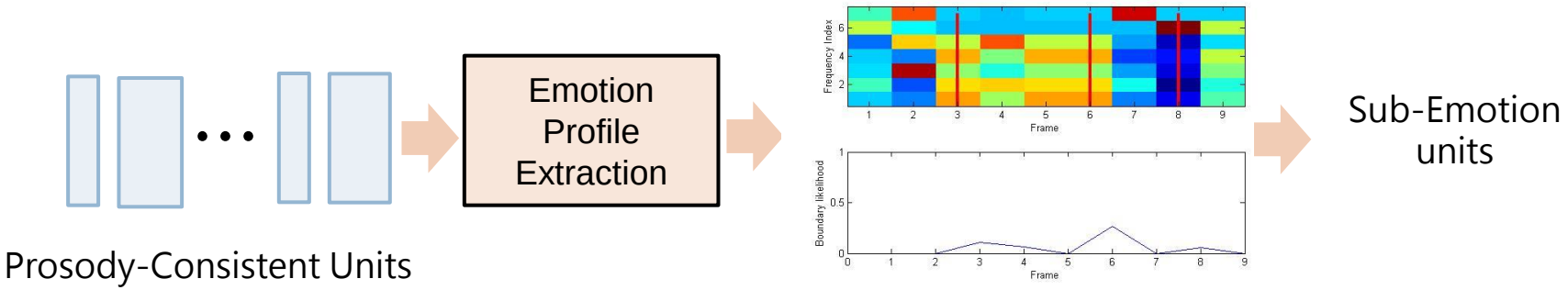
Prosody-Consistent units

Hierarchical Speech Unit Chunking

- For each prosody-consistent unit, an emotion profile is obtained using 64 prosody-related features



- Chunk the utterance into Sub-Emotion units based on emotion profiles



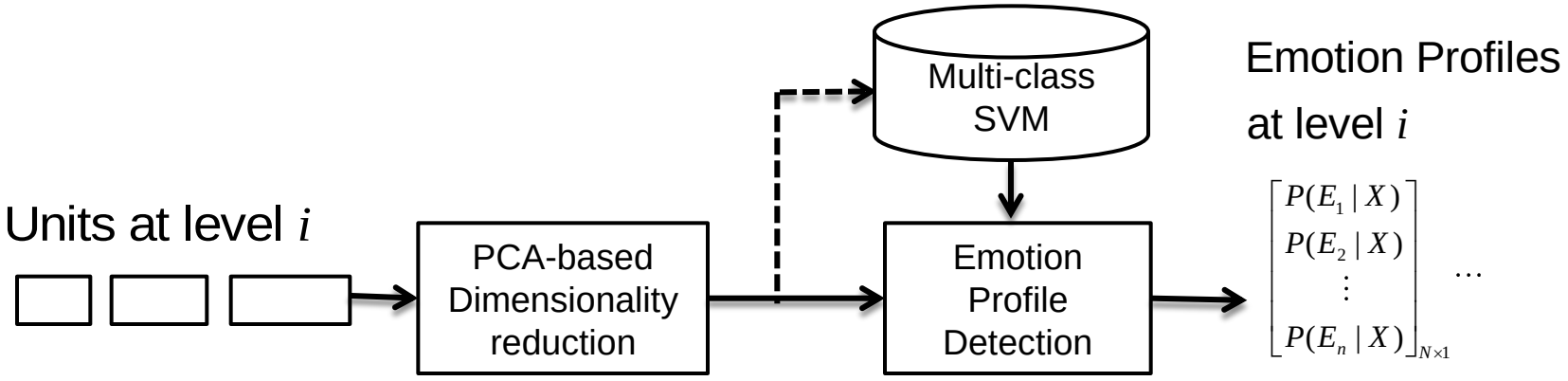
Feature Extraction for Emotion Recognition

- Feature Extraction for emotion recognition
  - For each utterance or unit, 384 static features are extracted through the functionals of its frames.

LLD (16*2)	Functionals (12)
(Δ) ZCR	mean
(Δ) RMS	standard deviation
(Δ) F0	kurtosis, skewness
(Δ) Harmonic-Noise-Ratio	extremes: value, relative position, range
(Δ) MFCC 1-12	linear regression: offset, slope, MSE

Codebook Construction

- Dimensionality reduction
  - The size of original feature vector is too large to utilize
  - Principal Component Analysis (PCA)
    - Top N Eigenvector were chosen for the Eigenspace, which consists of more than 90% of total variation.
- Emotion profile vector
  - In this phase, we will use the whole feature to train the multi-class SVM to obtain the emotion profile vector.



Hierarchical Correlation Model

- Emotion recognition
  - Let *O* be the feature vector of the test utterance, the recognition result is based on the maximum a posteriori criterion

$$E_e = \arg\max_e P(\lambda_{E_e} \mid O) = \arg\max_e \frac{P(O \mid \lambda_{E_e}) p(\lambda_{E_e})}{P(O)} = \arg\max_e P(O \mid \lambda_{E_e}) \quad (1)$$

  - The observation *O* can be regarded as the collection of the observations from each level.

$$E_e = \arg\max_e P(O \mid \lambda_{E_e}) = \arg\max_e P(O^{(1)}, O^{(2)}, O^{(3)} \mid \lambda_{E_e}) \quad (2)$$

where $O = \{O^{(1)}, O^{(2)}, ..., O^{(i)}\}$ and $\lambda_{E_e} = \{\lambda_{E_e}^{(1)}, \lambda_{E_e}^{(2)}, ..., \lambda_{E_e}^{(i)}\}$ are the observation and recognition model of level i, respectively.

Hierarchical Correlation Model

- For loosely coupled observations, we firstly assume the observations between levels are dependent, and (2) can be re-written as

$$P(O^{(1)},O^{(2)},O^{(3)} \mid \lambda_{E_e}) \approx P(O^{(1)} \mid \lambda_{E_e}^{(1)})P(O^{(2)} \mid \lambda_{E_e}^{(2)})P(O^{(3)} \mid \lambda_{E_e}^{(3)}) \quad (3)$$

- However, the observation from each level should be tightly coupled. In [Pan, 2001], they estimated tightly coupled signals using the following equation:

$$P(X,Y) = P(X)P(Y)\frac{p(w,v)}{p(w)p(v)} \quad (4)$$

where $w = f_X(X)$ and $v = f_Y(Y)$ that f_X and f_Y are mapping functions

Hierarchical Correlation Model

- Therefore, we follow the joint probability estimation method from [Pan, 2001] and obtain

$$E_e = \arg\max_e P(O \mid \lambda_{E_e}) = \arg\max_e P(O^{(1)}, O^{(2)}, O^{(3)} \mid \lambda_{E_e})$$

$$\approx \arg\max_e (P(O^{(1)} \mid \lambda_{E_e}^{(1)}) P(O^{(2)} \mid \lambda_{E_e}^{(2)}) P(O^{(3)} \mid \lambda_{E_e}^{(3)}))$$

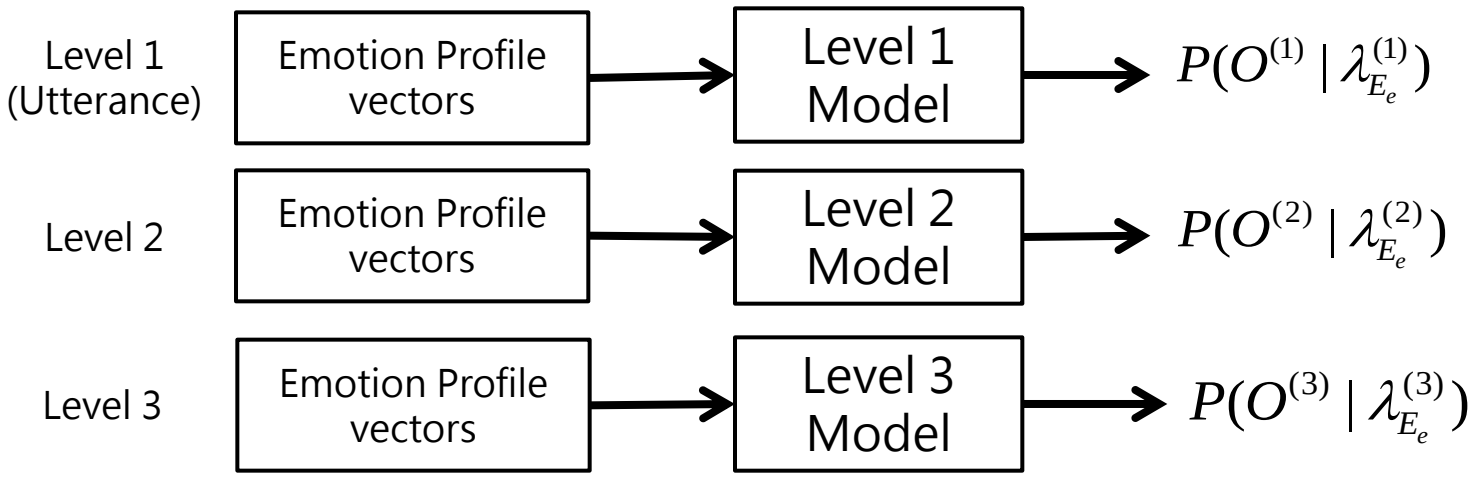
$$\times \left(\frac{P(D^{(1)}, D^{(2)}, D^{(3)} \mid E_e)}{P(D^{(1)} \mid E_e) P(D^{(2)} \mid E_e) P(D^{(3)} \mid E_e)} \right) \quad (5)$$

where $O^{(i)} = \left\{ o_1^{(i)}, o_2^{(i)}, ..., o_{N_i}^{(i)} \right\}$ is the observation sequence at level i, and

$D^{(i)} = \{ d_1^{(i)}, d_2^{(i)}, ..., d_{N_i}^{(i)} \}$ is the codeword sequence at level i

Hierarchical Correlation Model

- For the Marginal probability of each level



- For the model of each level, a four mixture Gaussian mixture model is used to model the emotion profile vectors

Hierarchical Correlation Model

- For the choice of mapping functions, we want to discretize the observation for facilitating the calculation of the correlation term in equation (5)

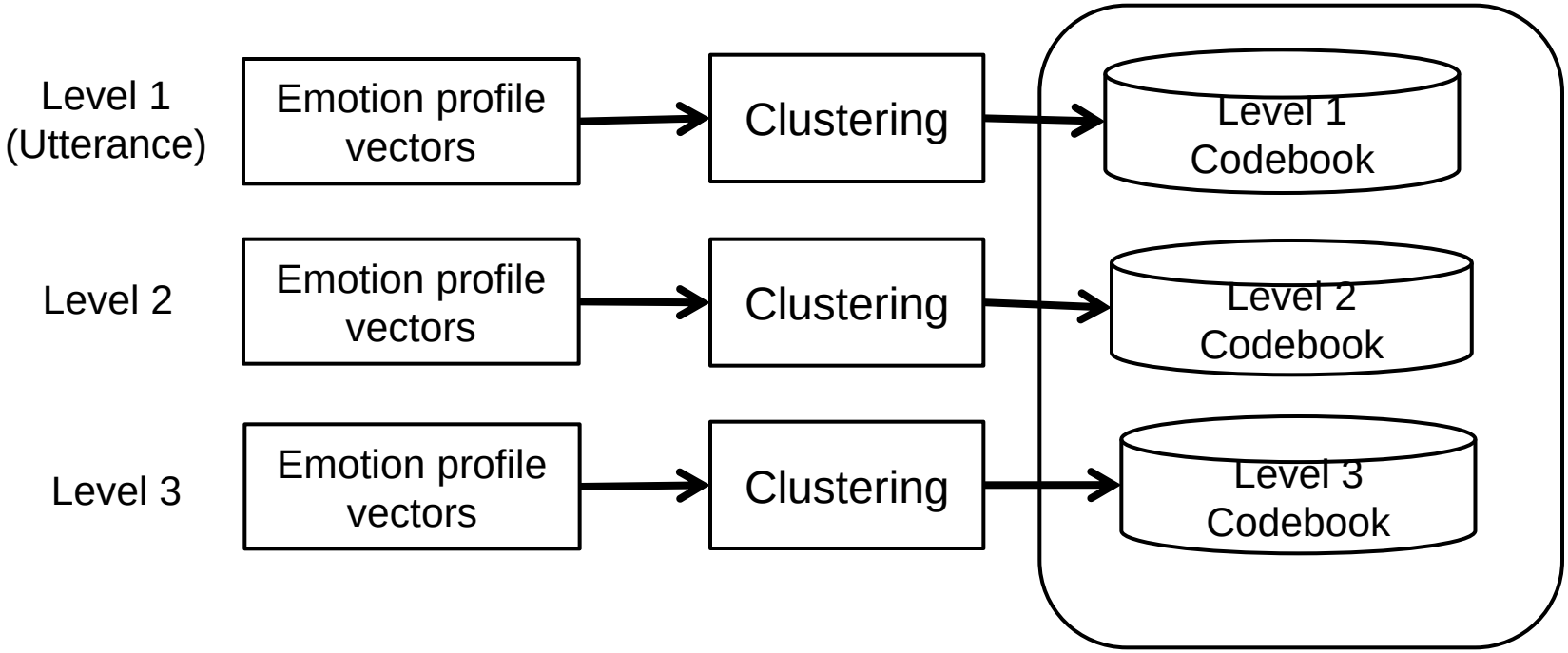
- For our mapping functions

$$D^{(i)} = \arg\max_{D^{(i)} \in C} P(\lambda_C^{(i)} \mid O^{(i)}) = \arg\max_{D^{(i)} \in C} P(O^{(i)} \mid \lambda_C^{(i)}) \qquad (6)$$

where $\lambda_C^{(i)}$ is the Gaussian mixture codebook in level i

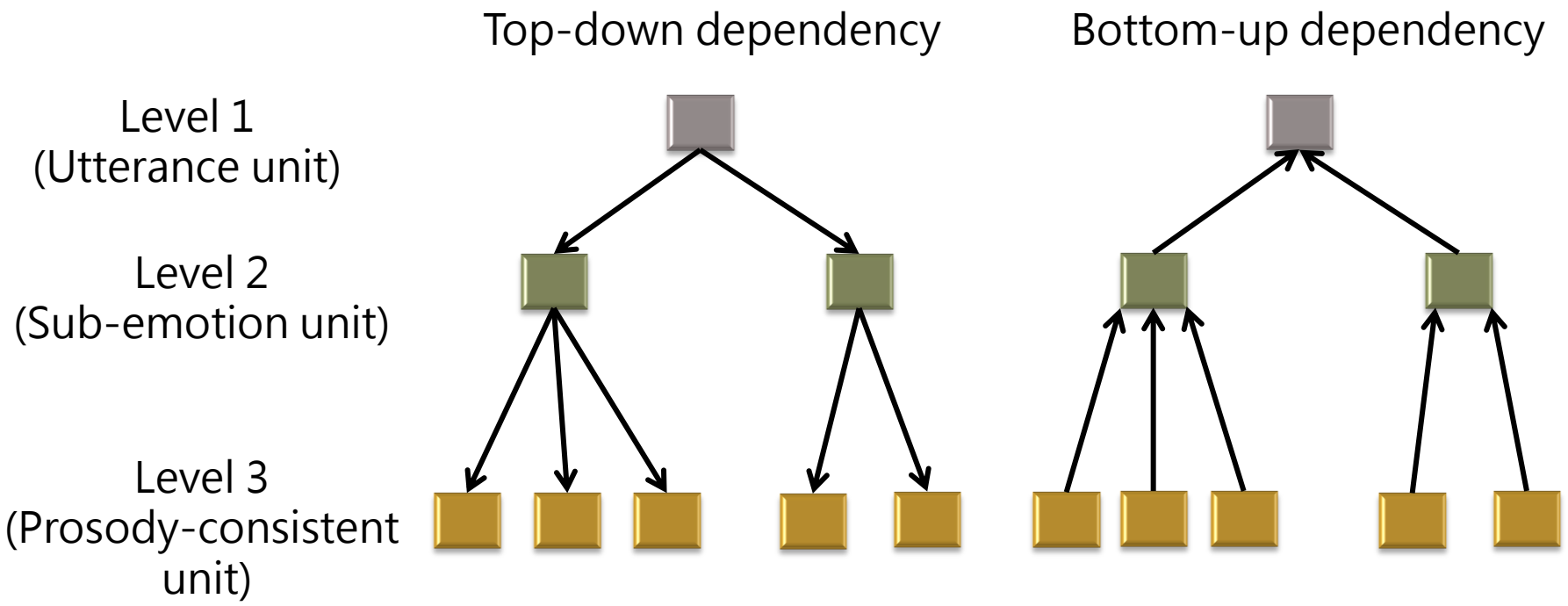
Codebook Construction

- Codebook Construction
  - K-means clustering is utilized to obtain the clusters
  - Each cluster is used for training a GMM as a codeword



Hierarchical Correlation Model

- For simplifying the term $P(D^{(1)}, D^{(2)}, D^{(3)} \mid E_e)$, we assume there exist two ways of dependencies



Hierarchical Correlation Model

- We first assume there exist top-down dependencies between levels

$$\frac{P(D^{(1)},D^{(2)},D^{(3)}\mid E_e)}{P(D^{(1)}\mid E_e)P(D^{(2)}\mid E_e)P(D^{(3)}\mid E_e)}=\frac{P(D^{(3)}\mid D^{(1)},D^{(2)},E_e)P(D^{(2)}\mid D^{(1)},E_e)P(D^{(1)}\mid E_e)}{P(D^{(1)}\mid E_e)P(D^{(2)}\mid E_e)P(D^{(3)}\mid E_e)}$$

$$\approx\frac{P(D^{(3)}\mid D^{(2)},E_e)P(D^{(2)}\mid D^{(1)},E_e)}{P(D^{(2)}\mid E_e)P(D^{(3)}\mid E_e)} \qquad (7)$$

- In equation (7)

$$P(D^{(i)}\mid D^{(i-1)},E_e)=P(d_1^{(i)},d_2^{(i)},...,d_{N_i}^{(i)}\mid d_1^{(i-1)},d_2^{(i-1)},...,d_{N_{i-1}}^{(i-1)},E_e) \qquad (8)$$

where $\{k(1),k(2),...,k(M_k)\}$ is the index of children of $d_k^{(i)}$

Hierarchical Correlation Model

- Suppose each node is only related to its parent

$$P(d_1^{(i)}, d_2^{(i)}, ..., d_{N_i}^{(i)} \mid d_1^{(i-1)}, d_2^{(i-1)}, ..., d_{N_{i-1}}^{(i-1)}, E_e)$$

$$= \prod_{k=1}^{N_{i-1}} P(d_{k(1)}^{(i)}, d_{k(2)}^{(i)}, ..., d_{k(M_k)}^{(i)} \mid d_k^{(i-1)}, E_e)$$

- Each node in the same level is independent to each other

$$P(d_{k(1)}^{(i)}, d_{k(2)}^{(i)}, ..., d_{k(M_k)}^{(i)} \mid d_k^{(i-1)}, E_e) = \prod_{m=1}^{M_k} P(d_{k,m}^{(i)} \mid d_k^{(i-1)}, E_e)$$

Corpus Analysis

- Emotional Speech Databases

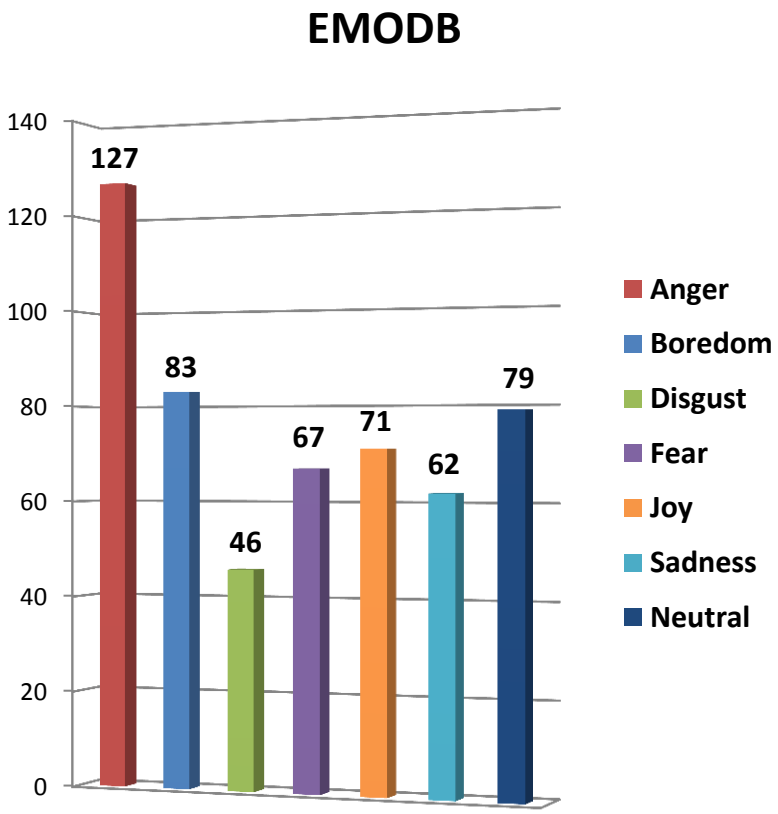
Database	**EMO-DB** [Burkhardt, 2005]	**eNTERFACE** [Martin, 2006]
No. of speakers	10 (5 male, 5 female)	43 (34 male, 9 female)
Language	German	English
Emotions	*Anger, boredom, disgust, fear, joy, neutral, sadness*	*Anger, disgust, fear, joy, sadness, surprise*
No. of utterances	535	1257
Style	Acted	Acted

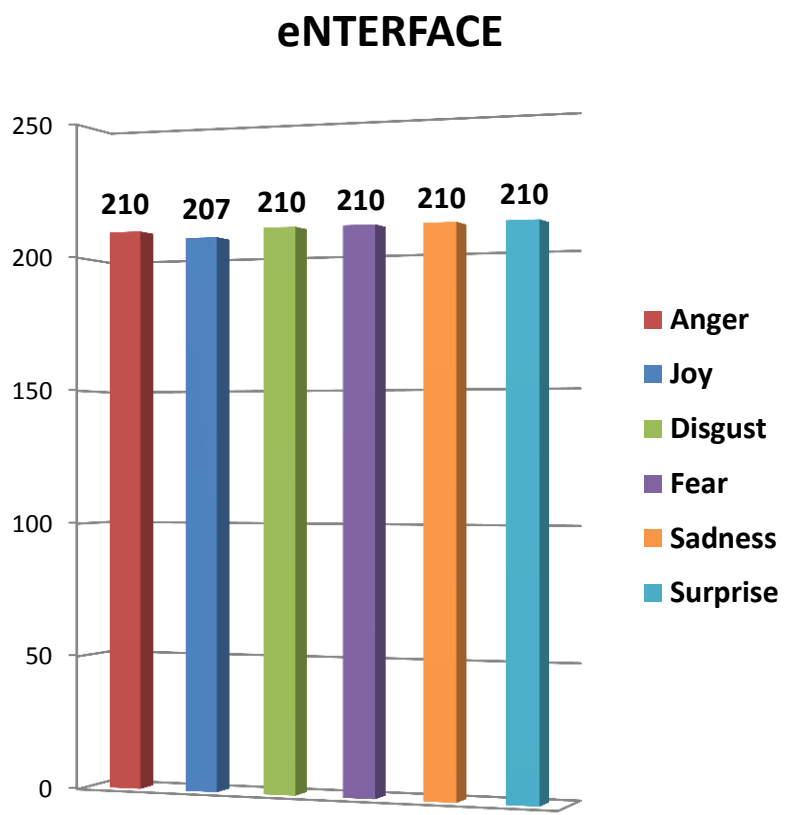
- Toolkits

Model	**HMM/GMM**	**Support Vector Machine**
Toolkits	HTK [Young, 2002]	LibSVM [Fan, 2005]

Corpus Analysis

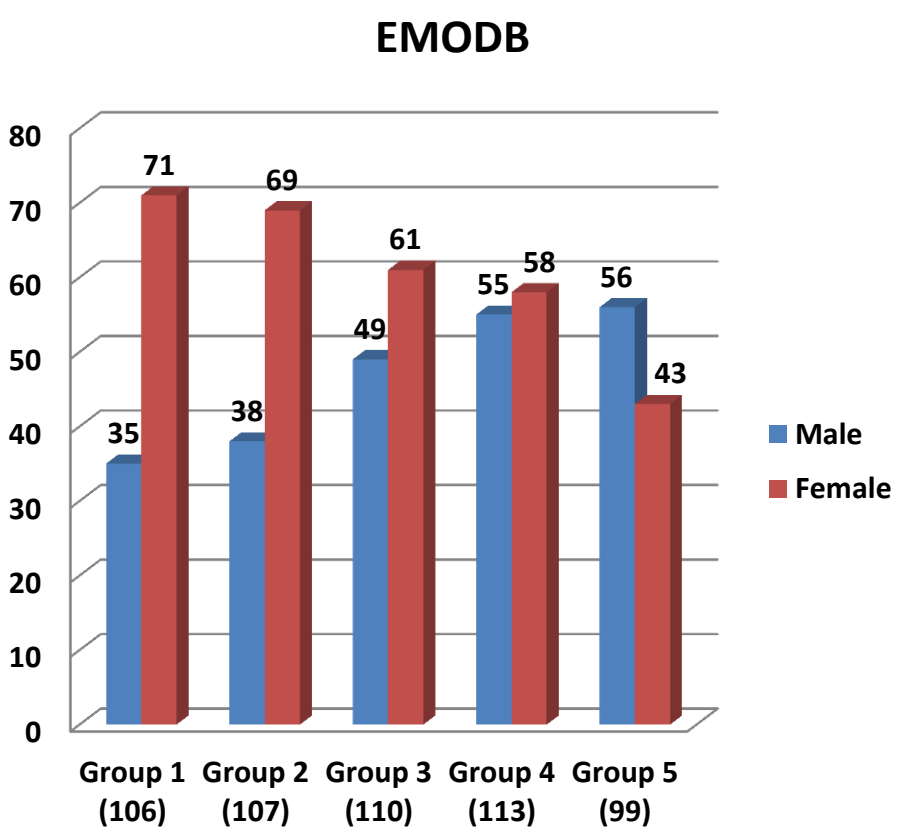
- The number of utterance for each emotion

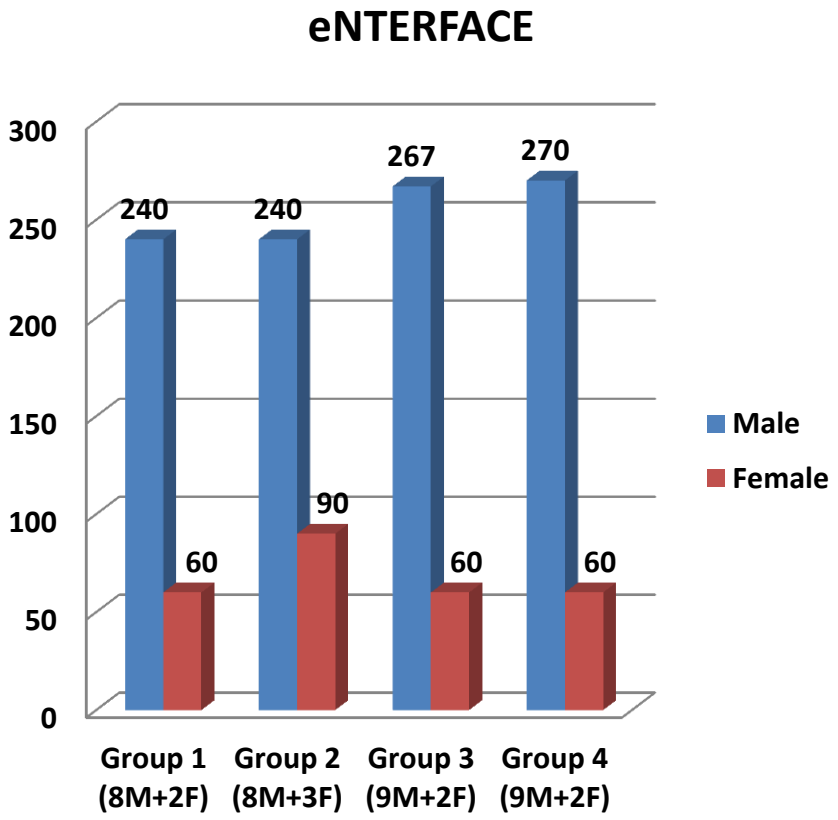




Corpus Analysis

- Leave-One-Group-Out (LOGO) evaluation for Speaker Independent condition





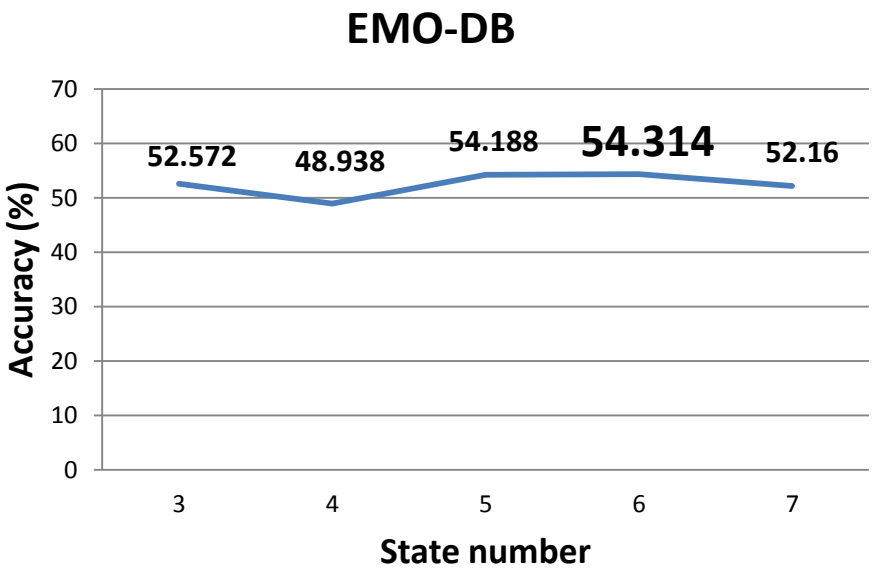
Experimental result

- Evaluation on analysis unit
  - Dynamic modeling
  - Static modeling
  - Fixed window segmentation [Jeon'11]

- Evaluation on fusion strategy
  - SVM-based fusion
  - Linear-weighting fusion
  - Hierarchical correlation model

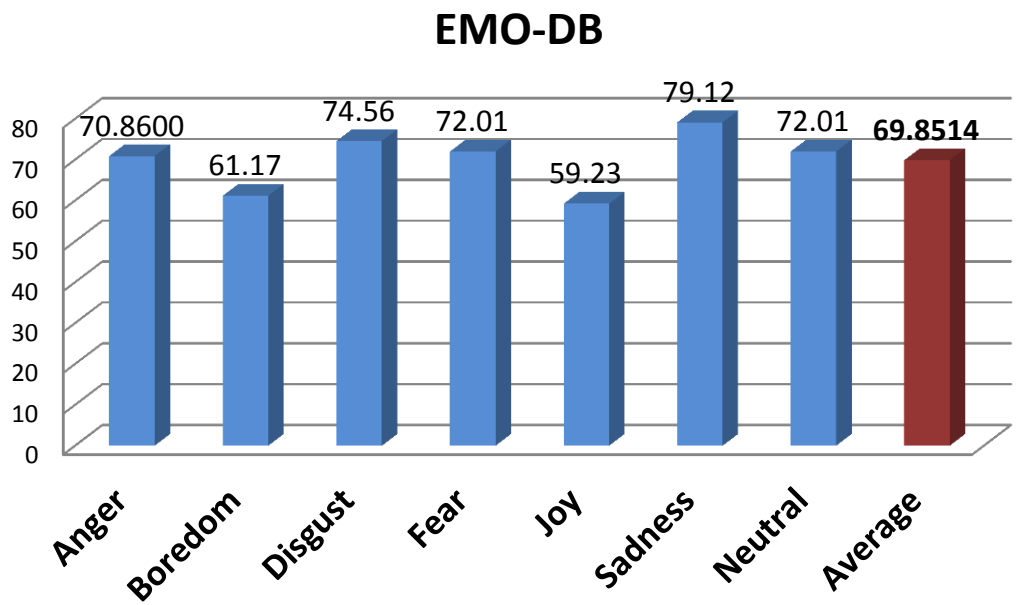
Experimental result

- Dynamic modeling baseline
  - Hidden Markov Model
  - One model per emotion, with two mixture GMM for each state
  - The Best result is from 6-state HMM



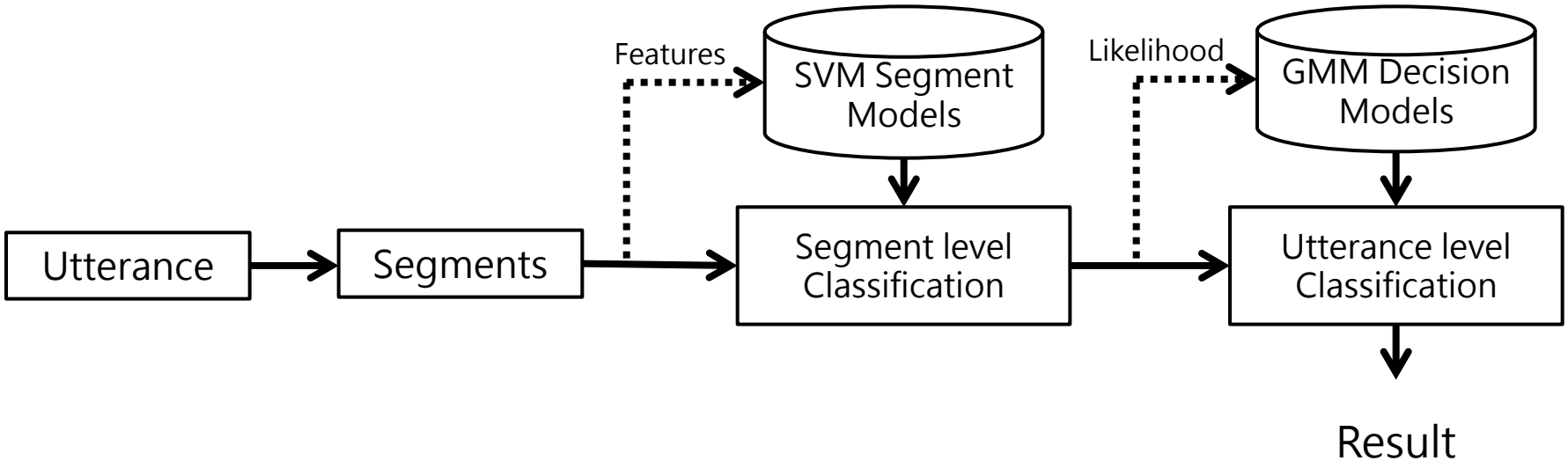
Experimental result

- Static Modeling baseline
  - Support Vector Machine with RBF kernel
  - Pairwise multi-class discrimination



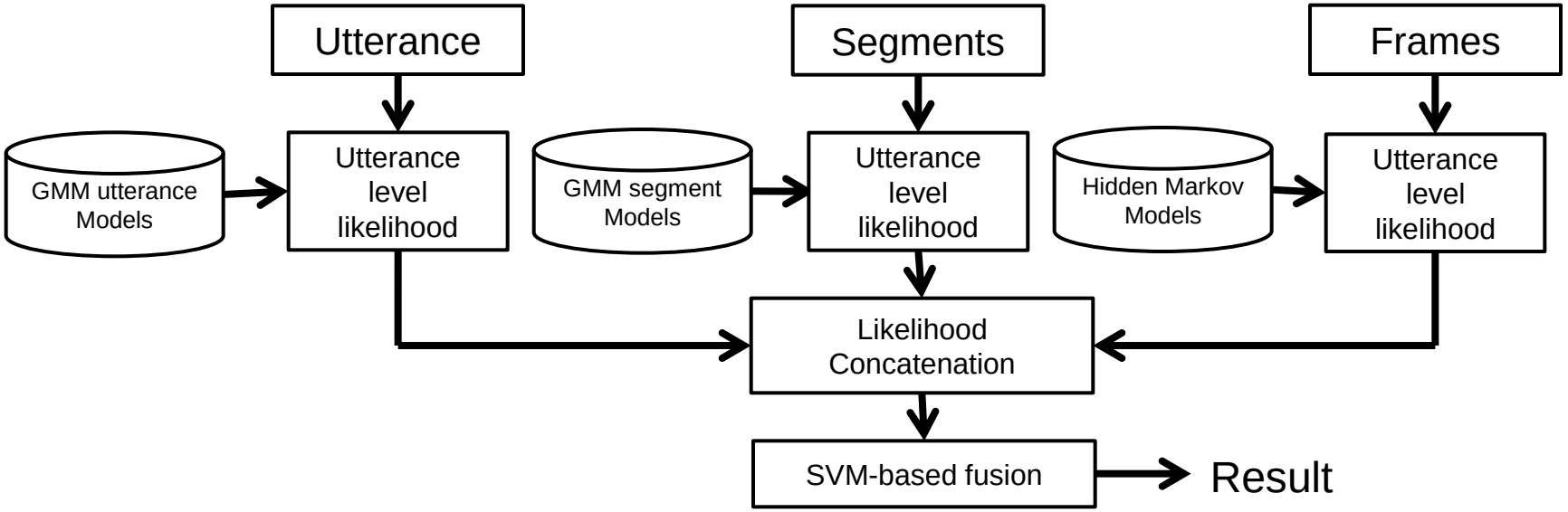
Baseline system

- Segmentation approach Baseline [Jeon, 2011]
  - The utterance level recognition is based on decisions from sub-sentence segments
  - Settings
    - Fixed length segment: 1 second with 0.2 second overlap
    - Segment model: SVM with RBF kernel
    - Decision model: 4 mixture GMM



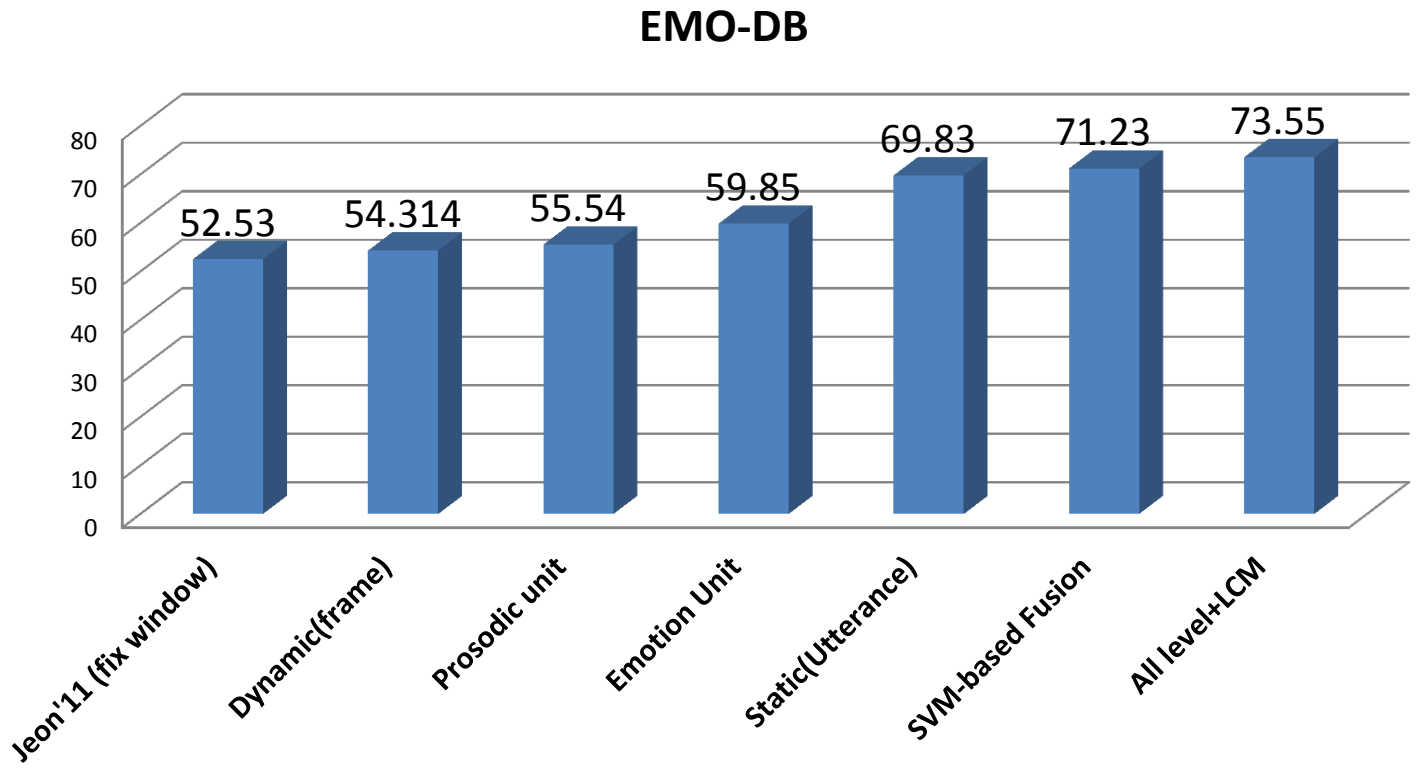
Baseline system

- Fusion approach baseline [Lee, 2010]
  - Lee combined different time-scale features for emotion recognition
  - Settings
    - Each utterance is equally divided into three segments.
    - Four mixture GMM for utterance- and segment-level models
    - Three state HMM with two mixture GMM for frame-level models



Experimental results

- Experimental results of the approaches for comparison



Conclusions

- Temporal and structural information is helpful to improve emotion detection performance.
- Temporal course modeling with intensity can effectively characterize the complex temporal structure in emotion expression.
- Hierarchical correlation modeling based multi-level units is beneficial to recognition of emotion in speech with long-term and transient expression.
- Collaboration among affection researchers from different disciplines is mandatory to achieve a satisfactory performance for real applications.

Questions?