MusicCommentator

Generating Comments Synchronized with Musical Audio Signals by a Joint Probabilistic Model of Acoustic and Textual Features

Kazuyoshi Yoshii   Masataka Goto

National Institute of Advanced Inductrial Science and Technology (AIST)

Background

- **Importance of expressing music in language**
  - **Language is an understandable common medium for human communication**

Free-form tags given to the entire clip

Good arrangement

Pretty cool! I am impressed

Time Comments

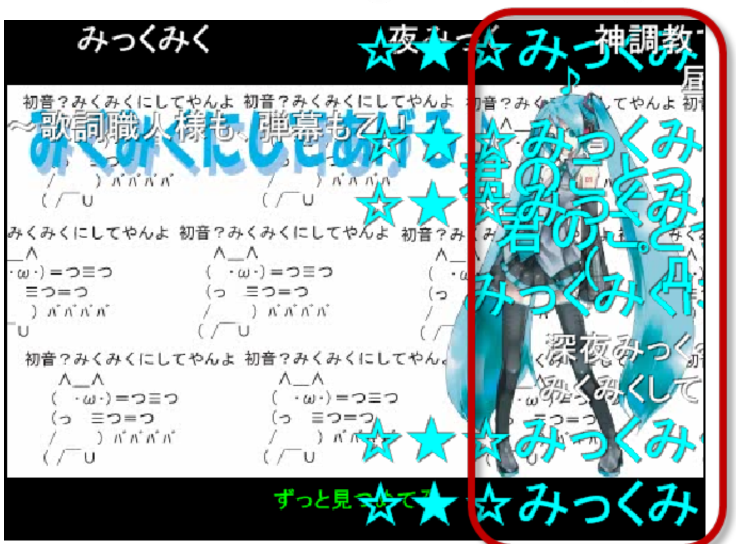
Short comments associated with temporal positions within the clip

Users can feel as if they enjoy together although they gave comments at different times in the real world

Snapshot from *Nico Nico Douga* (an influential video-sharing service in Japan)

Emerging Pehenomen in Japan

- **Commenting itself becomes entertainment**
  - **Commenting is an advanced form of collaboration**
    - **Users add effects to the video by giving comments**
    - **Commenting is a casual way of exhibiting creativity**
  - **Temporal comments strengthen a sense of togetherness**
    - **Users can feel as if they enjoy all together and collaborate to create something at the same time**
      - **Called pseudo-synchronized communication**

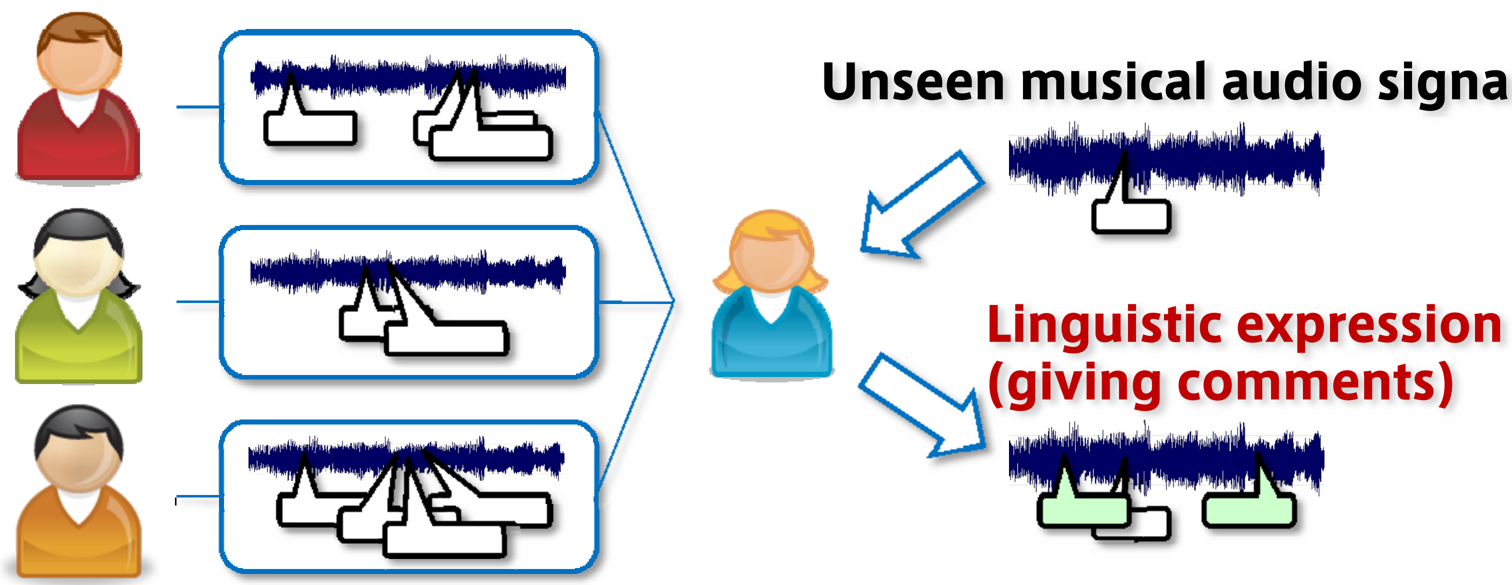


Temporal comments and *barrage*

Sophisticated ASCII art

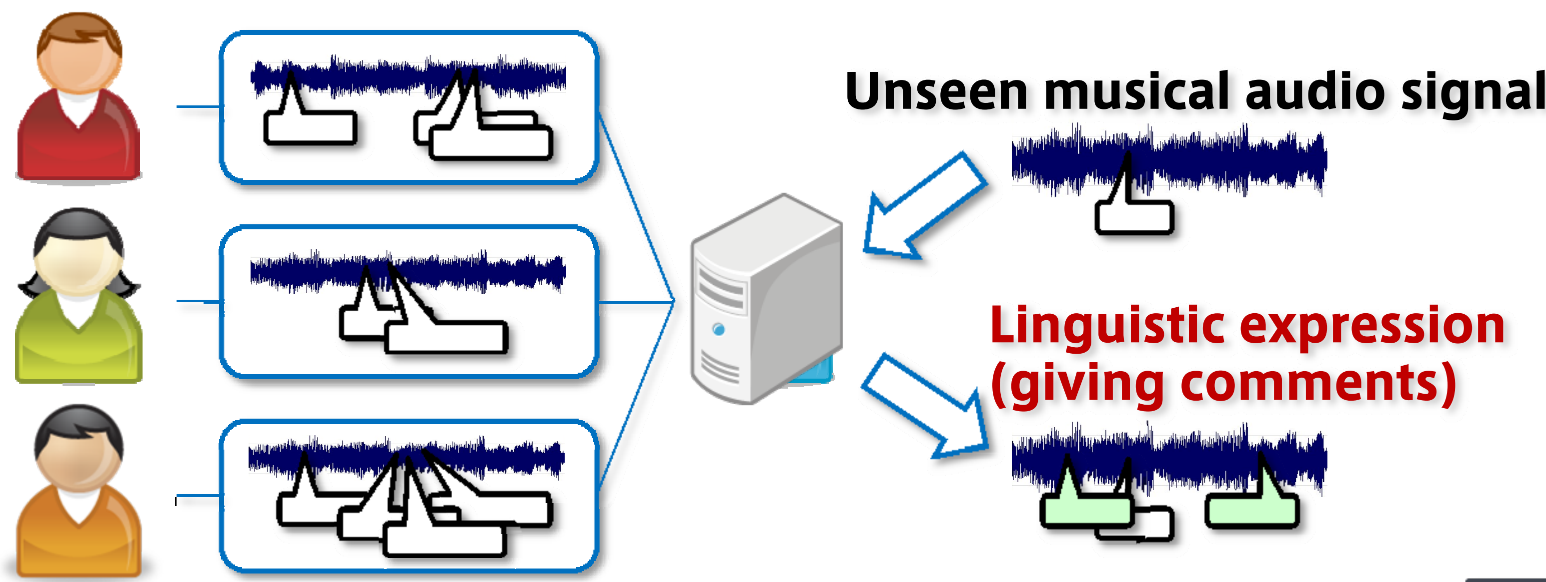
Motivation

- Facilitate human communication by developing a computer that can express music in language
  - Mediated by human-machine interaction
  - Hypothesis: Linguistic expression is based on learning
    - Linguistic expressions of various musical properties are learned through communication using language
    - Humans acquire a sense of what temporal events could be annotated in music clips



Approach

- **Propose a computational model of commenting that associates music and language**
  - **Give comments based on machine learning techniques**
    - **Train a model from many musical audio signals that have been given comments by many users**
    - **Generate suitable comments at appropriate temporal positions of an unseen audio signal**



Key Features

- **Deal with temporally allocated comments**
  - **Our study: Give comments to appropriate temporal positions in a target music clip**
  - **Conventional studies: Provide tags for an entire clip**
    - **Impression-word tags**
    - **Genre tags**

- **Generate comments as sentences**
  - **Our study: Concatenate an appropriate number of words in an appropriate order**
  - **Conventional studies: Only select words in a vocabulary**
    - **Word orders are not taken into account**
    - **Slots of template sentences are filled with words**

Ours [I] [am] [impressed] [with] [the] [cool] [playing] [!]

Conv. This is a [rock] song and has a [energetic] mood.

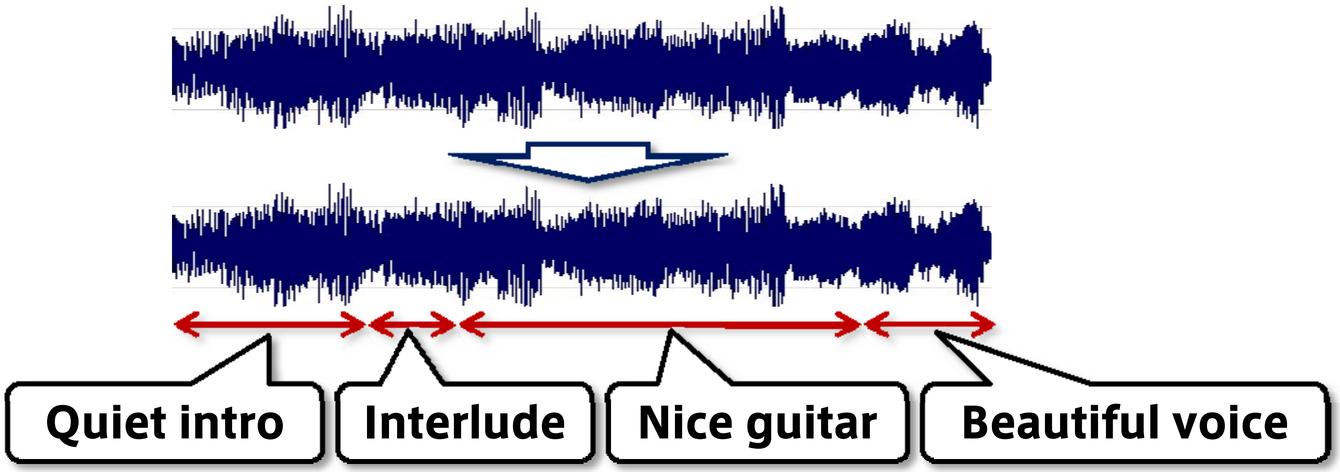
Applications to Entertainment

Semantic clustering & segmentation of music

- The performance could be improved
  by using features of both music and comments
- Users can selectively enjoy their favorite segments

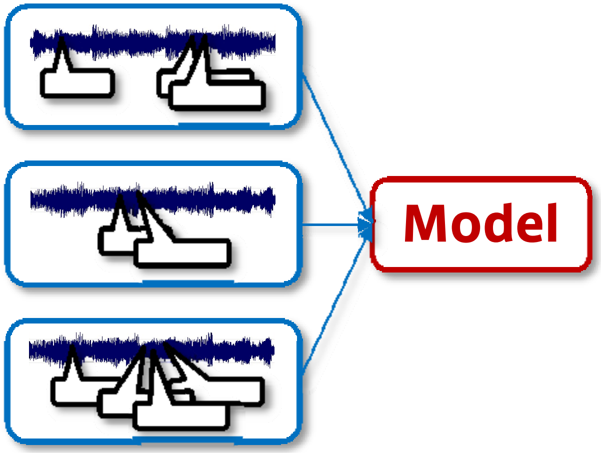
Linguistic interfaces for manipulating music

- Segment-based retrieval & recommendation
  could be manipulated by using language
- Retrieval & recommendations results
  could be explained by using language

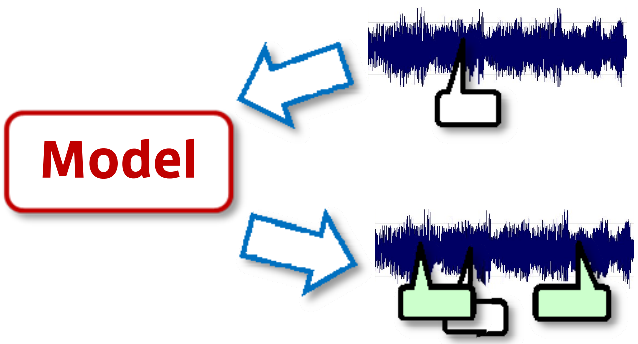


Problem Statement

- Learning phase
  - Input
    - Audio signals of music clips
    - Attached user comments
  - Output
    - Commenting model

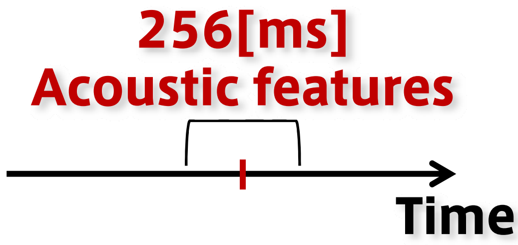


- Commenting phase
  - Input
    - Audio signal of a target clip
    - Attached user comments
    - Commenting model
  - Output
    - Comments that have suitable lengths and contents and are allocated at appropriate temporal positions

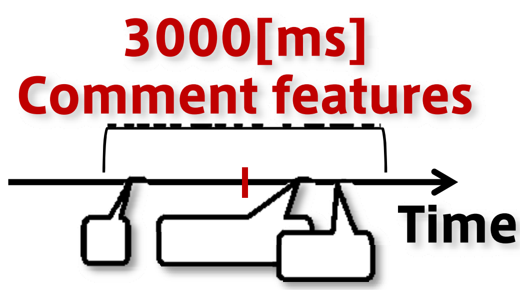


Feature Extraction

- **Extract features from each frame**
  - **Acoustic features**
    - **Timber feature: 28 dim**
      - **Mel-frequency cepstrum co-efficients (MFCCs): 13 dim.**
      - **Energy: 1 dim.**
      - **Dynamic property: 13+1 dim.**



  - **Textual features**
    - **Comment content: 2000 dim.**
      - **Average bag-of-words per comment**
    - **Comment density: 1 dim.**
      - **Number of user comments**
    - **Comment length: 1dim.**
      - **Average number of words per comment**



Bag-of-Words Feature

1. **Morphological analysis**
   - Identify
     - Part-of-speech
     - Basic form
2. **Remove auxiliary words**
   - Symbols / ASCII arts
   - Conjunctions, interjections particles, auxiliary verbs
3. **Assimilate same-content words**
   - Do not distinguish words that have same part-of-speech and basic form
   - Example: “take” = “took” = “taken
4. **Count number of each word**
   - The dimension of bag-of-words features is equal to vocabulary size

He played the guitar (^_^)

⇩ 1. Morph. analysis

He+played+the+guitar+(^_^)

⇩ 2. Screening

He played guitar

⇩ 3. Assimilation

he play guitar

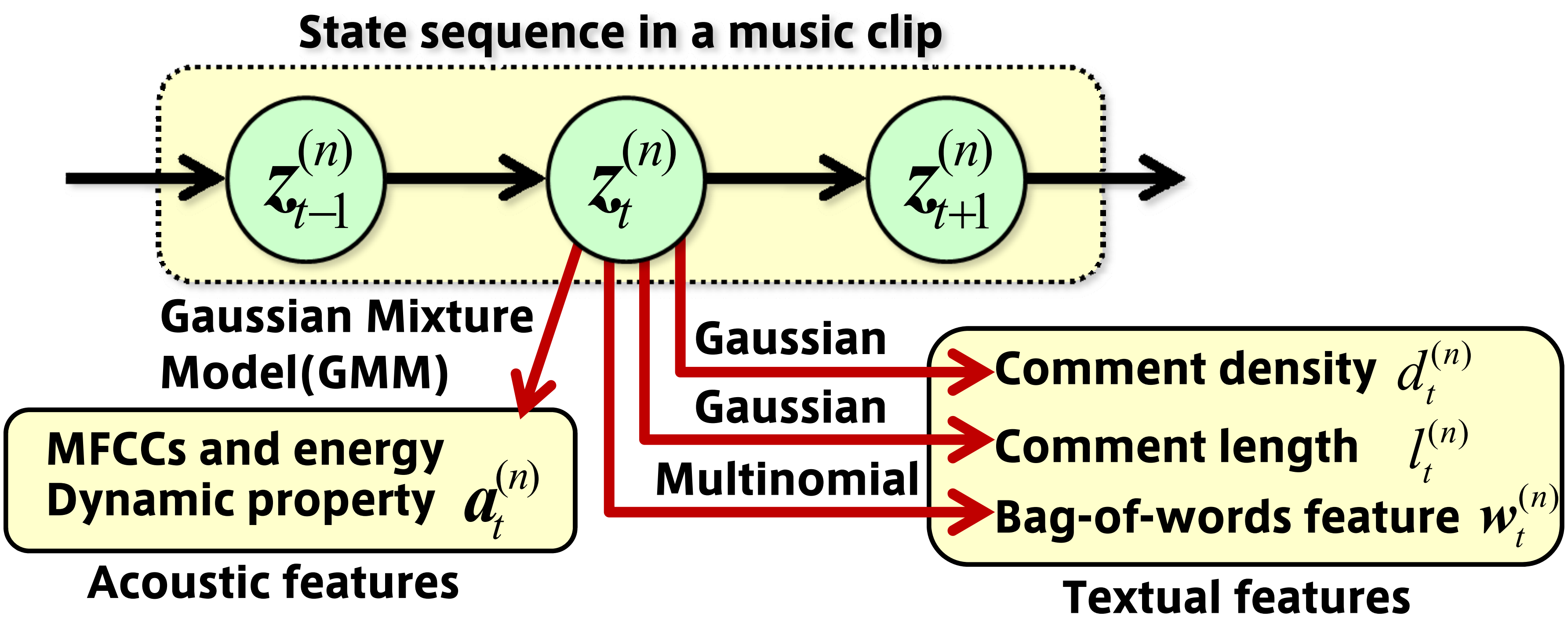
⇩ 4. Counting

he:1 play:1 guitar:1

COMMENTING MODEL

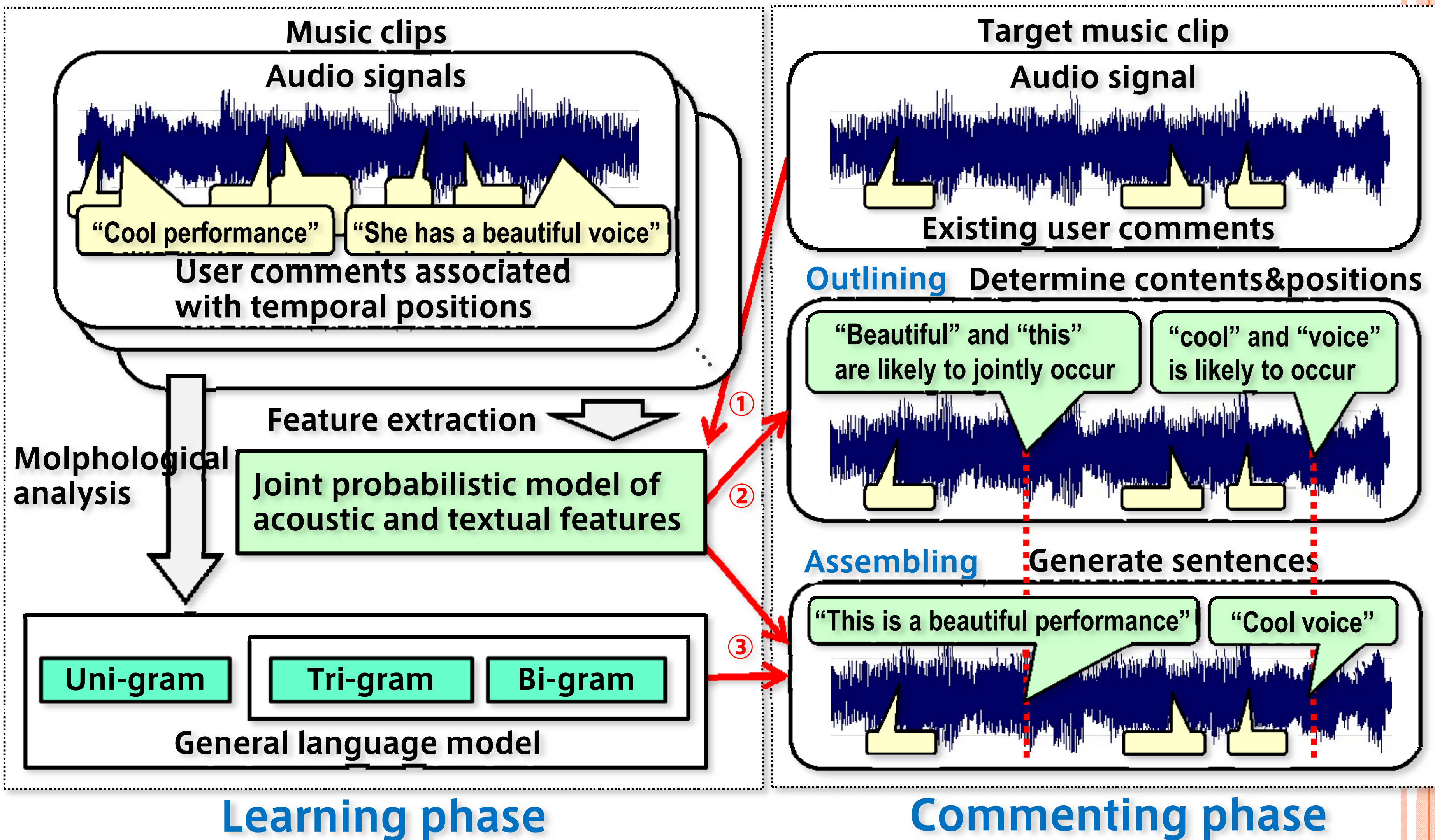
- **Three requirements**
  - All features can be simultaneously modeled
  - Temporal sequences of features can be modeled
  - All features share a common dynamical behavior

→ Extend Hidden Markov Model (HMM)



MusicCommentator

- Comment generation based on machine learning
  - Consistent in a maximum likelihood (ML) principal



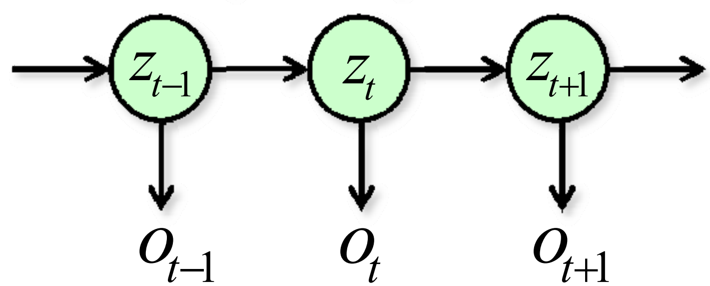
Learning Phase

- **ML Estimation of HMM parameters**
  - **Three kinds of parameters**
    - **Initial-state probability** $\{\pi_1, \cdots, \pi_K\}$
    - **Transition probability** $\{A_{jk} \mid 1 \le j,k \le K\}$
    - **Output probability** $\{\phi_1, \cdots, \phi_K\}$
  - **E-step: Calculate <u>posterior probabilities of latent states</u>**
  - **M-step: Independently update <u>output probabilities</u>**

Complete Likelihood

$$p(O,Z \mid \theta) = p(z_1 \mid \pi) \prod_{t=2}^{T} p(z_t \mid z_{t-1}) \prod_{t=1}^{T} p(o_t \mid z_t)$$

K =200 (#states)



$o_t = \{a_t, w_t, d_t, l_t\}$

Timber, Content, Density, Length

Objective (Q fuction)

$$Q(\theta;\theta_{\text{old}}) = \sum_{Z} p(Z \mid O, \theta_{\text{old}}) \log p(O, Z \mid \theta)$$

$$= \sum_{k=1}^{K} \underbrace{\gamma(z_{1,k})}_{\text{Posterior}} \log \pi_k + \sum_{t=2}^{T} \sum_{j=1}^{K} \sum_{k=1}^{K} \underbrace{\xi(z_{t-1,j}, z_{t,k})}_{\text{Posterior}} \log A_{jk}$$

$$+ \sum_{t=1}^{T} \sum_{k=1}^{K} \underbrace{\gamma(z_{t,k})}_{\text{Posterior}} \boxed{\log p(o_t \mid \phi_k)}$$

$$\log p(a_t \mid \phi_{a,k}) + \log p(w_t \mid \phi_{w,k}) + \log p(d_t \mid \phi_{d,k}) + \log p(l_t \mid \phi_{l,k})$$

Commenting Phase

- **ML Estimation of comment sentences**
  - **Assume a generative model of word sequences**

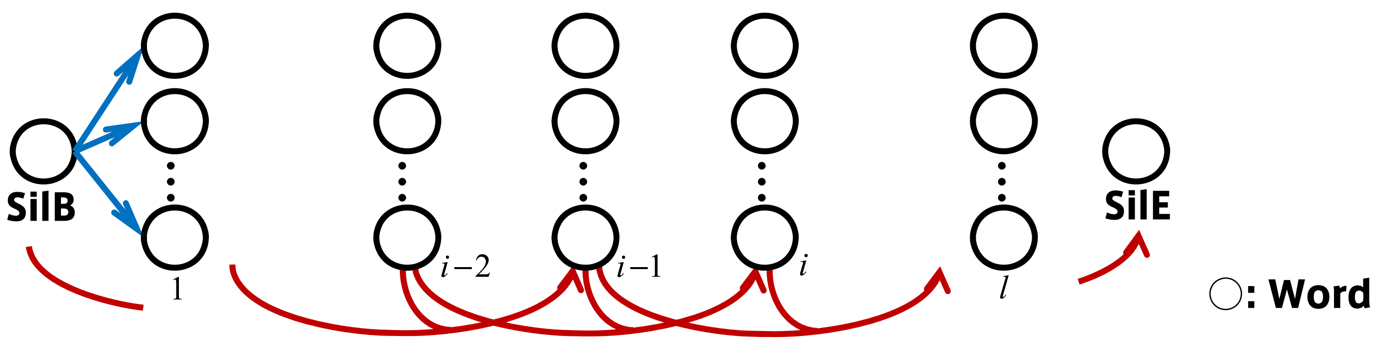
$$\{\hat{c},\hat{l}\} = \arg\max_{\{c,l\}} p(c,l) = \arg\max_{\{c,l\}} p(c\,|\,l)\,p(l)$$

$p(l)$ **Probability that length is** l ← **Gaussian**

$p(c\,|\,l)$ **Probability that sequence is** c **when length is** l ← **? ? ?**

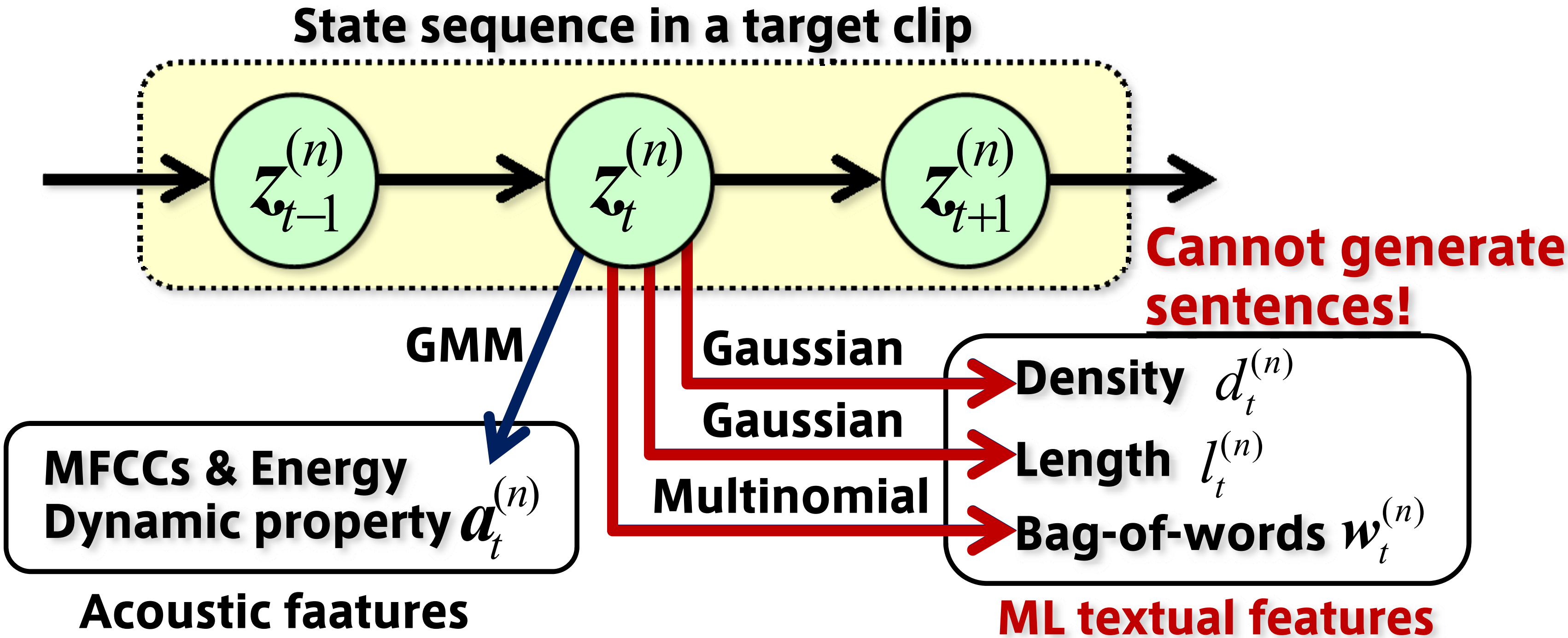
Computed by the Viterbi algorithm using bi- and tri-grams

$$p(c\,|\,l) = \left(p(w_1\,|\,\mathrm{SilB}) \prod_{i=2}^{l} p(w_i\,|\,w_{i-2}, w_{i-1})\, p(\mathrm{SilE}\,|\,w_{l-1}, w_l) \right)^{\frac{1}{l}}$$



Outlining Stage

- Determine content and positions of comments
  - **Input acoustic and textual features**
    - Input only acoustic features if there are no existing user comments in a target clip
  - **Estimate a ML state sequence**
    - Use the Viterbi algorithm
    - Calculate ML textual features at each frame



Problems and Solusions

- No probabilities of words required for sentences
  - Bag-of-words feature=Reduced uni-gram
    - Verb conjugations are not taken into account
    - Auxiliary words are removed

- No probabilities of word concatenations
  - Bi- and tri- grams are not taken into account

This performance is good ↔ This is a good performance

Which is more suitable?

⇩

Use general bi- and tri-grams learned from all user comments

All words required for composing sentences are contained

Assembling Stage

- **Adaptation of general language models**
  - **Adaptation of general uni-gram**
    - **ML bag-of-words feature is embedded**

General uni-gram + **ML Bag-of-words feature at a frame** = **Adapted uni-gram**

$$p(w_i) \qquad w_t^{(n)} \qquad p'(w_i)$$

  - **Adaptation of general bi- and tri- grams**
    - **Linear combination with adapted uni-gram**

$$p'(w_i \mid w_{i-1}) \propto p(w_i \mid w_{i-1}) + p'(w_i)$$

$$p'(w_i \mid w_{i-2}, w_{i-1}) \propto p(w_i \mid w_{i-2}, w_{i-1}) + p(w_i \mid w_{i-1}) + p'(w_i)$$

⇨ **Search for ML word sequence (Viterbi path)**

$$\{\hat{c}, \hat{l}\} = \arg\max_{\{c,l\}} p(c,l) = \arg\max_{\{c,l\}} p(c \mid l)\, p(l)$$

EXAMPLE 1

- ML comment sentences with respect to lengths

Length	Log-Likeli.	Sentence
1	-10.1036	☺
2	-7.99174	Play well ☺
3	-6.33792	Play very well ☺
4	-5.30383	Very funny guitar playing ☺
5	-4.90632	Play well but very funny ☺
6	-5.04090	Play well but waste of talent ☺
7	-5.95158	Play well but brought …(cannot be translated) ☺
8	-7.39973	Play well, but very funny strap ☺
9	-9.43043	Play well but brought …(cannot be translated) ☺
10	-12.3661	Play well but brought …(cannot be translated) ☺

Naked Guitarist is playing

Appropriately generate comment sentences

Example 2

- ML comment sentences with respect to lengths

Length	Log-likeli.	Sentence
1	-236.545	☺
2	-70.2561	Good work ☺
3	-3.51156	G O D bo
4	-37.0469	Well done work ☺
5	-170.226	Well done work! ☺
6	-403.678	G O D bra bo ☺
7	-712.145	G O D bra bo he is cool ☺
8	-712.091	G O D bra bo he is cool … ☺
9	-712.225	G O D bra bo he is cool … ☺
10	-712.324	G O D bra bo he is cool good work☺

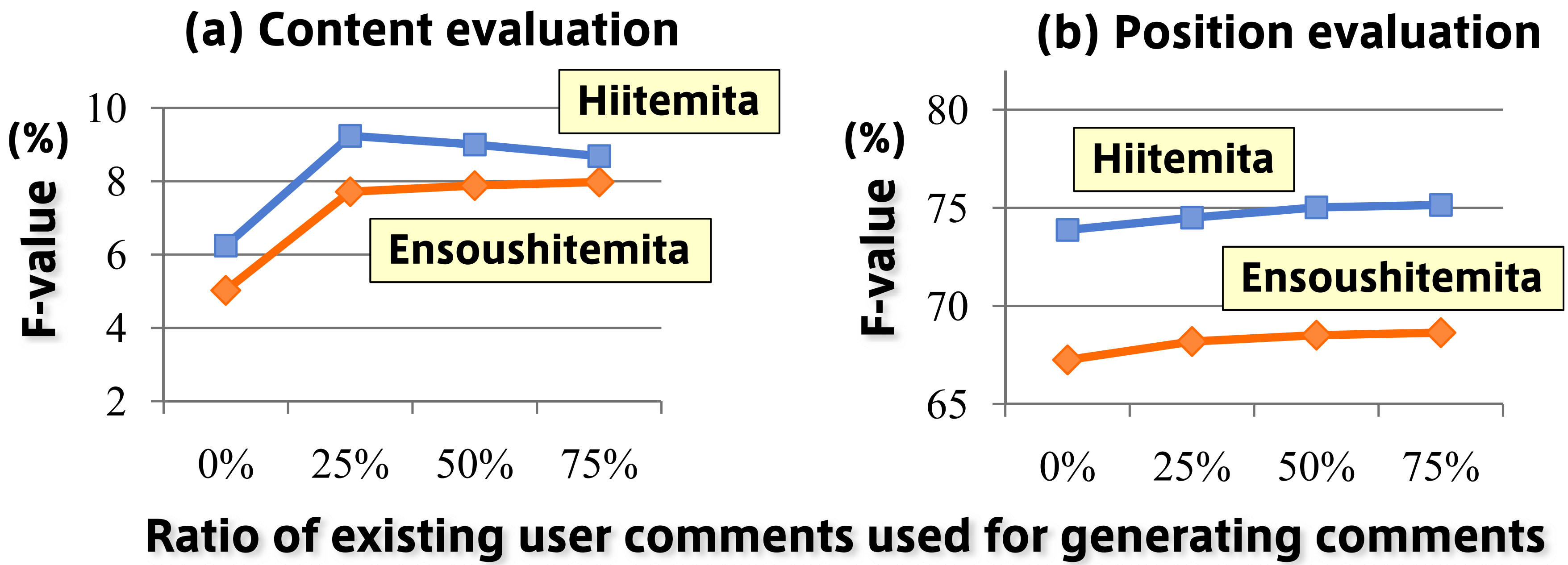
End of piano performance

The system can synthesize unique phrases that not included in vocabulary by using language models

Experiments

- Datasets
  - Collected from Nico Nico Douga
    - 100 clips whose titles include "Ensoushitemita"
      - "I played something, not limited to musical instruments, e.g., music box and wooden gong"
      - Extracted 1100 comments from each clip
    - 100 clips whose titles include "Hiitemita"
      - "I played piano or stringed instruments, e.g., violin and guitar"
      - Extracted 2400 comments from each clip
- Evaluation metric
  - 4-cross fold validation
    - Train a model by using 75 clips
    - Generate comments for 25 clips
      - 0,25,50,75% of existing user comments was used
      - Remaining 25% was used as the ground truth
  - F-values
    - Harmonic means of Precision and recall rates)
    - The error tolerance is 5[s]

Results



- **Performance was improved by using existing comments in target clips**
  - **Effective for estimating the content of music clips**
- **Performance improvement was hardly gained if we use more existing comments**
  - **Diversity was spoiled**

Conclusion and Future Directions

- Proposed a computational model of associating acoustic features with textual features
  - HMM-based probabilistic comment generation
  - A model is learned from many user comments
  - Language constraints are taken into account for generating sentences by using language models

- Future works
  - Use various kinds of features
    - High-level musical features other than MFCCs
      - Vocal, rhythm, tempo
    - Visual features
  - Improve our commenting model
    - Avoid the over-fitting problem
  - Refine word screening
    - TF*IDF

End of Presentation