Vol.2009-MUS-81 No.20
2009/7/30

MusicCommentator: 音楽に同期したコメントを自動生成するシステム

吉井和佳[†1]　後藤真孝[†1]

本稿では，楽曲のライムライン上の適切な時刻に適切なコメントを自動付与するシステム MusicCommentator について述べる．近年，ユーザが動画全体に対してではなく，動画中のある時刻に対してコメントできるオンライン動画共有サービスが人気を博している．本研究では，音楽演奏の動画に含まれる音楽音響信号を対象とし，音響的特徴量とコメント特徴量との確率的同時生成モデルを提案する．システムはまず，多くの楽曲とそれに付与されたコメントから確率モデルを学習する．その後，別の楽曲が入力として与えられた場合に，どの時刻に対して，どのような単語を用いてどのくらいの長さのコメントを新たに付与できるかを確率モデルを用いて推定する．このとき，言語的制約として単語間の連接を考慮し，文の合成をおこなう．実験の結果，入力楽曲の音響的特徴量だけを用いてコメント生成した時に比べ，すでに付与されたコメントを参考にしてコメント生成を行うと精度が向上することがわかった．

MusicCommentator: A Computational System of Generating Music-Synchronized Comments

Kazuyoshi Yoshii[†1] and Masataka Goto[†1]

This paper presents a system called *MusicCommentator* that suggests suitable comments for appropriate temporal positions in a music clip. Recently, an online video sharing service in which users can provide comments for temporal events occurring in video clips not for entire clips has gained a lot of popularity. We focus on musical audio signals included in video clips of music performances and propose a probabilistic model that jointly generates acoustic features and comment features. The model can be trained by using many music clips and their corresponding comments. Given a new clip as input, the system then determines appropriate temporal positions of comments and estimates their content and length. Finally, comment sentences are generated by taking word concatenations into account as language constraints. Our experimental results showed that comment accuracy was improved when the system used not only acoustic features of an input clip but also users' comments in the clip.

1. はじめに

人と人とがコミュニケーションを行うためのメディアとして，音楽は重要な位置を占めている．例えば，ある楽曲のどこをどのように感じたかを友人どうしで語りあったり，両者が知っている楽曲をきっかけにほとんど接点がなかった人どうしでも会話がはずんだりする．近年は情報通信技術の発達により，多くの楽曲がデジタルデータとしてオンライン化されたのに伴い，音楽を介したコミュニケーションもインターネット上で行われるようになった．物理制約がなくなった結果，不特定多数の人による大規模なコミュニケーションが一般的になり,「音楽にコメントする」という行為がますます重要性を増している．例えば，多くのオンライン音楽配信サイトでは楽曲に関するコメント投稿機能が設けられ，多くのユーザが感想や批評を書き残している．さらに進んだ例として，オンライン動画共有サイトである「ニコニコ動画」[1)] では，動画中のある特定の時刻を指定して一言程度の短いコメントを付与できる．音楽演奏の動画であれば，局所的な音楽内容に対してコメントが可能である．

音楽を介して他者とコミュニケーションを行うために，人間は言語という手段を用いて自らが感じたことを語り合うことができる．それを可能にする知的な仕組みについては解明されていないが，いくつか手がかりはある．例えば，ジャズになじみがない人にとっては，どのジャズの曲も同じように聴こえてコメントしにくいということがしばしばある．一方，ジャズが好きな人は，この曲のここはこういう感じで，あちらは逆にこういう感じだなどと表現することができる．このときの表現は他の誰かが他の曲に対してコメントした内容に影響を受けていると考えられる．つまり，音楽にコメントするという行為は，言語を用いたコミュニケーション経験に基づいていると推察される．

我々は，人間は音楽の内容とコメントとの対応付けを無意識的に学習しているという仮説のもと，音楽に対するコメントを学習して生成するシステム MusicCommentator を提案する．システムはまず，多数の楽曲とそれに対してユーザが付与した多数のコメント例から，楽曲中のどの時刻に，どのくらいの密度で，どのような長さ・内容のコメントが付与され得るかを学習する．その後，学習時に現れなかった別の楽曲が与えられると，言語的制約のもとで適切な単語をつなぎ合わせ，適切な長さの一言程度のコメント文を生成する．このとき，コメントを付与すべき時刻や密度も推定する．このような過程を計算機上で実現しようとする試みは，音楽内容とその言語表現との関係を明らかにする上で学術的に興味深い．

†1 産業技術総合研究所 (AIST)

本研究で扱う音楽音響信号は,「ニコニコ動画」の音楽演奏に関する動画に含まれるものである（以降単に楽曲と呼ぶ）．すべてのコメントは楽曲中の時刻に対応付けられているが，楽曲の内容に無関係なコメントが非常に多く含まれる．このような実際的なデータからコメント生成モデルを学習できるかを検証することも本研究の課題の一つである．ニコニコ動画では，不特定多数のユーザによって付与されたコメントは，動画再生中に対応する時刻がくると動画に重ねて表示される．そのため，現実世界においてコメントを付与した時間は異なるにもかかわらず，同じ動画を多数のユーザでコメントしあいながら一緒に鑑賞しているような感覚が演出される（「擬似同期型アーキテクチャ」[2] と呼ばれる）．したがって，MusicCommentator の応用として，新作などのコメントの少ない動画に対してコメントを自動付与することで，ユーザ間コミュニケーションのきっかけを提供することも考えられる．

本稿の構成は以下の通りである．まず，2 章で関連研究を紹介する．次に，3 章でコメント生成問題を定義し，4 章で MusicCommentator について説明する．5 章で評価実験について報告し，最後に 6 章でまとめを述べる．

2. 関 連 研 究

音楽音響信号に対して単語を付与する研究はいくつか行われてきた．典型的には，機械学習手法を用いて，楽曲に対して事前に用意した各単語がどのくらい強く関連しているかを学習・予測する．例えば，Whitman らは，カーネル手法を利用して，楽曲のレビュー文に現れる単語を予測した[3]．Turnbull らは，単語ごとに音響的特徴量に対する混合ガウス分布を学習して単語を予測する手法を提案した[4]．出力は音楽の内容を説明する文章であるが，事前に用意されたテンプレート文のスロットを，推定した単語で埋めることで生成していた．Bertin-Mahieux らはソーシャルタグと呼ばれる多数のユーザが付与したタグ（その多くはジャンル名や印象語など）をアンサンブル学習法の一種 AdaBoost を用いて予測した[5]．

本研究は，上記の従来研究と 2 点で異なる．第一に，我々は楽曲全体に対して与えられたコメントではなく，楽曲中の時刻に対応付けられたコメントを扱う点である．したがって，与えられた楽曲に対してコメントを生成するだけでなく，それらが付与され得る適切な時刻を決定する必要がある．第二に，テンプレートを用いずに自由形式の自然言語文を生成する点である．すなわち，適切な単語を適切な活用形・順序で接続しなければならない．

楽曲の各部に対してアノテーションを行うシステム（インタフェース）はいくつか提案されている[6),7]．例えば，梶らは楽曲中の区間を指定してユーザが感想や印象のアノテーションが可能なシステムを提案しているが[6]，システムによる自動生成は扱っていない．

3. コメント生成問題

本研究では，楽曲群およびそれらに付与されたコメントから抽出した特徴量を用いてコメント生成モデルを学習したあと，新たな楽曲が入力として与えられた時に，適切なコメントを適切な時刻に付与することを考える．入力となる楽曲にすでにいくつかのコメントが付与されている場合は，それらも考慮してコメントを追加する．いま，学習データとして N 個の音楽音響信号があるとし，n $(1 \leq n \leq N)$ を楽曲のインデックスとする．

3.1 音響的特徴量

音響的特徴量として，音楽音響信号中の各フレームから 13 次元のメル周波数ケプストラム係数 (MFCC) とエネルギー，およびそれらの動的変動成分を抽出する．MFCC は音楽音響信号を入力としたジャンル識別やムード判定に有効に利用されてきた特徴量である．楽曲 n のフレーム t から抽出した 28 次元の音響的特徴量を $\boldsymbol{a}_t^{(n)}$ とする．

3.2 コメント特徴量

ユーザによって付与されたコメント群を，その内容・密度・長さの観点で特徴量化する．

3.2.1 Bag-of-Words 素性

コメントの内容表現として，Bag-of-Words 素性を利用する．まず，自由形式で記述された日本語のコメントから記号やアスキーアートを除去し，形態素解析器 Mecab[8] を用いて単語に分かち書きする．次に，助詞・助動詞・接続詞・感動詞などの補助的な単語を除去する．さらに，同じ品詞および語幹を持つ意味的に同一である単語の区別は行わないことにしたうえで，全コメント中で一定回数以上使用されている自立語を抽出する．この結果，V 単語が語彙として得られたとする．これを用いて，あるフレーム中で 1 つのコメント中に語彙中の各単語が平均何回登場したかをカウントする．例えば，あるフレームに 3 つのコメント「愛すべき曲」「愛してる」「愛の歌」が付与されていたとする．このとき，このフレームにおける動詞「愛する」の平均使用回数は，3 つめのコメント中の名詞「愛」はカウントしないので 2/3 となる．楽曲 n のフレーム t における Bag-of-Words 素性を $\boldsymbol{w}_t^{(n)} = \{w_{t,1}^{(n)}, \cdots, w_{t,V}^{(n)}\}$ とすると，$w_{t,v}^{(n)}$ $(1 \leq v \leq V)$ は単語 v の 1 コメント当たりの平均使用回数である．

3.2.2 コメント密度

あるフレームにおけるコメントの密度（コメント数）は，そのフレームがどのくらいコメントされやすいのかを示す重要な指標である．例えば，ニコニコ動画で「弾幕」と呼ばれる現象では，多数のユーザがほとんど同一の大量のコメントを特定の時刻に付与しており，コメント密度が非常に高い．楽曲 n のフレーム t におけるコメント密度を $d_t^{(n)}$ とする．

Vol.2009-MUS-81 No.20
2009/7/30

3.2.3 コメント長

1 つのコメントを構成する単語数は，どのくらいの長さのコメントを生成すべきかを決定する際の重要な指標である．自然言語文であるコメントは Bag-of-Words 素性を計算する過程でスクリーニングされた補助的な単語を含むため，Mecab で単語に分かち書きした段階で単語数をカウントする．楽曲 n のフレーム t におけるコメント長を $l_t^{(n)}$ とする．

3.3 学習データの特徴量化

前述した 4 種類の特徴量をまとめて $\boldsymbol{o}_t^{(n)} = \{\boldsymbol{a}_t^{(n)}, \boldsymbol{w}_t^{(n)}, d_t^{(n)}, l_t^{(n)}\}$ で表すことにする．楽曲 n が T_n フレームで構成されているとすると，観測できる特徴量 $\boldsymbol{O}^{(n)}$ および $\boldsymbol{O}$ は $\boldsymbol{O}^{(n)} = \{\boldsymbol{o}_1^{(n)}, \cdots, \boldsymbol{o}_{T_n}^{(n)}\}$ および $\boldsymbol{O} = \{\boldsymbol{O}^{(1)}, \cdots, \boldsymbol{O}^{(N)}\}$ で与えられる．

4. MusicCommentator

図 1 に示す通り，MusicCommentator は音響的特徴量とコメント特徴量の発生過程を確率的にモデル化する学習フェーズと，得られたモデルを利用して楽曲にコメントを付与する生成フェーズから構成される．以降，モデルの構成法および各フェーズについて述べる．

4.1 モデル定義

我々は有用な確率モデルを構成する上で，以下の 3 つの要件を考慮する．

(1) 音響的特徴量とコメント特徴量を同時にモデル化できること：ユーザはある時刻にコメントを付与しようとする場合に，楽曲の内容だけでなく既存のコメントを参考にしている．そのため，両方の特徴量を同時に扱えることが望ましい．

(2) 音響的特徴量とコメント特徴量の時系列をモデル化できること：扱うデータは時系列メディアであるので，動的に変化するコンテクストに着目することが重要である．すなわち，特徴量の時間的変化を表現できるモデルが必要である．

(3) データの背後にあるコンテクストを通して双方の特徴量が関連付けられていること：各フレームにおいて，1 つの隠れた状態（トピックと解釈してもよい）が音響的特徴量とコメント特徴量の間で共有されている必要がある．

これらの要件を満たすため，**図 2** で示されるような，標準的な隠れマルコフモデル (HMM) を拡張した確率的同時生成モデルを提案する．いま全部で K 種類の隠れ状態があるとし，楽曲 n のフレーム t における隠れ状態を潜在変数 $\boldsymbol{z}_t^{(n)} = \{z_{t,1}^{(n)}, \cdots, z_{t,K}^{(n)}\}$ で表す．$\boldsymbol{z}_t^{(n)}$ は一対 K 表現，すなわち，実現された状態に対応する次元のみが 1 で他は 0 となるベクトルである．ここで，$\boldsymbol{Z}^{(n)} = \{\boldsymbol{z}_1^{(n)}, \cdots, \boldsymbol{z}_{T_n}^{(n)}\}$，$\boldsymbol{Z} = \{\boldsymbol{Z}^{(1)}, \cdots, \boldsymbol{Z}^{(N)}\}$ と定義しておく．

提案モデルはパラメータ $\boldsymbol{\theta} = \{\boldsymbol{\pi}, \boldsymbol{A}, \boldsymbol{\phi}\}$ で定義する．$\boldsymbol{\pi}$ は初期状態確率 $\{\pi_1, \cdots, \pi_K\}$ であり，$\pi_k \equiv p(z_{1,k}^{(\cdot)} = 1)$ で与えられる．$\boldsymbol{A}$ は遷移行列 $\{A_{jk}|1 \leq j,k \leq K\}$ であり，$A_{jk} \equiv p(z_{t,k}^{(\cdot)} = 1|z_{t-1,j}^{(\cdot)} = 1)$ で与えられる．$\boldsymbol{\phi}$ は出力確率分布のパラメータである．

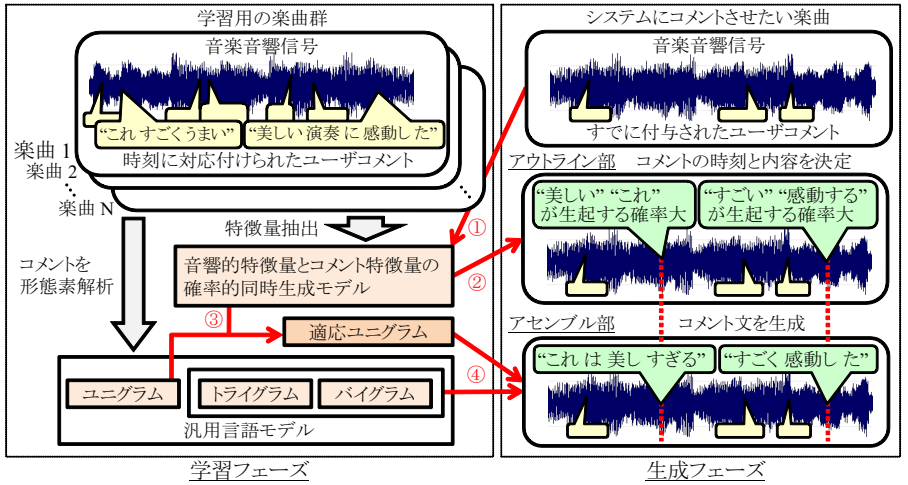

図 1 コメント学習・生成システム MusicCommentator の動作概要

本モデルはエルゴディック HMM であり，任意の状態から任意の状態への遷移を許容する．音声認識などの教師あり学習タスクでは学習データ中の状態遷移系列が与えられるが，本研究ではそれらが未知であるため教師なし学習となる．したがって，学習データ $\boldsymbol{O}$ をもっともよく説明できるような状態遷移系列 $\boldsymbol{Z}$ とパラメータ $\boldsymbol{\phi}$ を同時推定する．

各フレームの音響的特徴量とコメント特徴量は同じ状態から発生すると考える．b_k を状態 k の出力確率分布とすると，特徴量 $\boldsymbol{o}_t^{(n)}$ の尤度は $b_k(\boldsymbol{o}_t^{(n)})$ で与えられる．これは，4 種類の特徴量 $\{\boldsymbol{a}_t^{(n)}, \boldsymbol{w}_t^{(n)}, d_t^{(n)}, l_t^{(n)}\}$ がどれくらい同時に発生しやすいかを示す．各特徴量の状態に関する条件付き独立性を仮定すると，出力確率分布は以下の通り分解できる．

$$b_k(\boldsymbol{o}_t^{(n)}) = b_{a,k}(\boldsymbol{a}_t^{(n)})\ b_{w,k}(\boldsymbol{w}_t^{(n)})\ b_{d,k}(d_t^{(n)})\ b_{l,k}(l_t^{(n)}) \tag{1}$$

ここで，$b_{a,k}$ は標準的な音声認識用 HMM と同様に音響的特徴量に対する混合ガウス分布 (GMM) である．GMM の混合数を M とし，m 番目 $(1 \leq m \leq M)$ のガウス分布の混合比，平均，分散を $g_{a,k,m}$，$\boldsymbol{\mu}_{a,k,m}$，$\boldsymbol{\Sigma}_{a,k,m}$ とする．$b_{w,k}$ は bag-of-words 素性に対する多項分布であり，そのパラメータを $\boldsymbol{p}_k = \{p_{k,1}, \cdots, p_{k,V}\}$ とする．$b_{d,k}$ はコメント密度に対するガウス分

布であり，その平均と分散は $\mu_{d,k}$，$\Sigma_{d,k}$ で与える．$b_{l,k}$ はコメント長に対するガウス分布であり，その平均と分散は $\mu_{l,k}$，$\Sigma_{l,k}$ で与える．ここで，状態 k における出力確率分布のパラメータを $\boldsymbol{\phi}_k$ とすると，$\boldsymbol{\phi}_k = \{\{g_{a,k,m}, \boldsymbol{\mu}_{a,k,m}, \boldsymbol{\Sigma}_{a,k,m} | 1 \leq m \leq M\}, \boldsymbol{p}_k, \mu_{d,k}, \Sigma_{d,k}, \mu_{l,k}, \Sigma_{l,k}\}$ となる．したがって，全 K 個の隠れ状態のパラメータは $\boldsymbol{\phi} = \{\boldsymbol{\phi}_1, \cdots, \boldsymbol{\phi}_K\}$ となる．

4.2 学習フェーズ

学習フェーズでは，EM アルゴリズムを用いて，状態遷移系列 $\boldsymbol{Z}$ およびパラメータ $\boldsymbol{\theta}$ を E ステップおよび M ステップで反復的に最適化する．まず，完全データの尤度は

$$p(\boldsymbol{O}, \boldsymbol{Z}|\boldsymbol{\theta}) = \prod_{n=1}^{N} p(\boldsymbol{z}_1^{(n)}|\boldsymbol{\pi}) \left[\prod_{t=2}^{T_n} p(\boldsymbol{z}_t^{(n)}|\boldsymbol{z}_{t-1}^{(n)}) \right] \prod_{t=1}^{T_n} p(\boldsymbol{o}_t^{(n)}|\boldsymbol{z}_t^{(n)}) \tag{2}$$

で与えられる．ここで，$p(\boldsymbol{z}_1^{(n)}|\boldsymbol{\pi}) = \prod_{k=1}^{K} \pi_k^{z_{1,k}^{(n)}}$ である．これを用いて Q 関数は，

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}_{old}) = \sum_{\mathbf{Z}} p(\boldsymbol{Z}|\boldsymbol{O}, \boldsymbol{\theta}_{old}) \log p(\boldsymbol{O}, \boldsymbol{Z}|\boldsymbol{\theta}) \tag{3}$$

で定義できる．ここで，$\boldsymbol{\theta}_{old}$ は現在のパラメータ推定値である．

E ステップでは以下で定義される潜在変数の事後確率を計算する．

$$\gamma(\boldsymbol{z}_t^{(n)}) \equiv p(\boldsymbol{z}_t^{(n)}|\boldsymbol{O}, \boldsymbol{\theta}_{old}),\ \xi(\boldsymbol{z}_{t-1}^{(n)}, \boldsymbol{z}_t^{(n)}) \equiv p(\boldsymbol{z}_{t-1}^{(n)}, \boldsymbol{z}_t^{(n)}|\boldsymbol{O}, \boldsymbol{\theta}_{old}) \tag{4}$$

$$\gamma(\boldsymbol{y}_{t,k}^{(n)}) \equiv p(\boldsymbol{y}_{t,k}^{(n)}|\boldsymbol{O}, \boldsymbol{\theta}_{old}) = p(\boldsymbol{y}_{t,k}^{(n)}|\boldsymbol{z}_t^{(n)}) \gamma(\boldsymbol{z}_t^{(n)}) \tag{5}$$

$\gamma(\boldsymbol{z}_t^{(n)})$ は $\boldsymbol{z}_t^{(n)}$ の事後確率である．$\gamma(z_{t,k}^{(n)})$ を $z_{t,k}^{(n)} = 1$ となる事後確率とすると，$\gamma(\boldsymbol{z}_t^{(n)})$ は和が 1 となる K 次元ベクトルで表現できる．同様に，$\xi(\boldsymbol{z}_{t-1}^{(n)}, \boldsymbol{z}_t^{(n)})$ は $\boldsymbol{z}_{t-1}^{(n)}$ から $\boldsymbol{z}_t^{(n)}$ に遷移する事後確率であり，和が 1 となる $K \times K$ 行列で表現できる．これらの確率は Forward-Backward アルゴリズムを用いて効率的に求めることができる．$\boldsymbol{y}_{t,k}^{(n)}$ は音響的特徴量に対する GMM である $b_{a,k}$ 中のどの要素分布から $\boldsymbol{a}_t^{(n)}$ が発生したかを示す潜在変数である．これは，$\boldsymbol{z}_t^{(n)}$ と同様に一対 M 表現 $\{y_{t,k,1}^{(n)}, \cdots, y_{t,k,M}^{(n)}\}$ で表せる．したがって，$p(\boldsymbol{y}_{t,k}^{(n)}|\boldsymbol{z}_t^{(n)})$ は，$z_{t,k}^{(n)} = 1$ である場合に $y_{t,k,m}^{(n)} = 1$ となる確率であり，$K \times M$ 行列で表現できる．

M ステップでは Q 関数を展開し，これを最大化するパラメータを計算する．

$$\begin{aligned} Q(\boldsymbol{\theta}|\boldsymbol{\theta}_{old}) = &\sum_{n=1}^{N} \sum_{k=1}^{K} \gamma(z_{1,k}^{(n)}) \log \pi_k + \sum_{n=1}^{N} \sum_{t=2}^{T_n} \sum_{j=1}^{K} \sum_{k=1}^{K} \xi(z_{t-1,j}^{(n)}, z_{t,k}^{(n)}) \log A_{jk} \\ &+ \sum_{n=1}^{N} \sum_{t=1}^{T_n} \sum_{k=1}^{K} \gamma(z_{t,k}^{(n)}) \log p(\boldsymbol{o}_t^{(n)}|\boldsymbol{\phi}_k) \end{aligned} \tag{6}$$

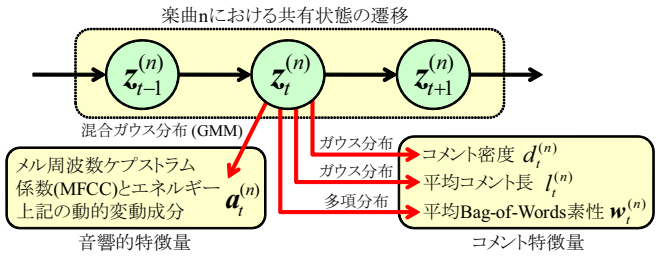

図 2 HMM に基づく音響的特徴量とコメント特徴量の確率的同時生成モデル

ここで，最後の項が $\log p(\boldsymbol{o}_t^{(n)}|\boldsymbol{\phi}_k) = \log b_{a,k}(\boldsymbol{a}_t^{(n)}) + \log b_{w,k}(\boldsymbol{w}_t^{(n)}) + \log b_{d,k}(d_t^{(n)}) + \log b_{l,k}(l_t^{(n)})$ と分解できるので，各分布のパラメータは独立に更新可能である．

$$\begin{aligned}
\pi_k &= \frac{\sum_{n=1}^{N} \gamma(z_{1,k}^{(n)})}{\sum_{n=1}^{N} \sum_{k=1}^{K} \gamma(z_{1,k}^{(n)})},\ A_{jk} = \frac{\sum_{n=1}^{N} \sum_{t=2}^{T_n} \xi(z_{t-1,j}^{(n)}, z_{t,k}^{(n)})}{\sum_{n=1}^{N} \sum_{l=1}^{K} \sum_{t=2}^{T_n} \xi(z_{t-1,j}^{(n)}, z_{t,l}^{(n)})}, \\
g_{a,k,m} &= \frac{\sum_{n=1}^{N} \sum_{t=1}^{T_n} \gamma(y_{t,k,m}^{(n)})}{\sum_{n=1}^{N} \sum_{t=1}^{T_n} \sum_{m=1}^{M} \gamma(y_{t,k,m}^{(n)})},\ \boldsymbol{\mu}_{a,k,m} = \frac{\sum_{n=1}^{N} \sum_{t=1}^{T_n} \gamma(y_{t,k,m}^{(n)}) \boldsymbol{a}_t^{(n)}}{\sum_{n=1}^{N} \sum_{t=1}^{T_n} \gamma(y_{t,k,m}^{(n)})}, \\
\boldsymbol{\Sigma}_{a,k,m} &= \frac{\sum_{n=1}^{N} \sum_{t=1}^{T_n} \gamma(y_{t,k,m}^{(n)}) (\boldsymbol{a}_t^{(n)} - \boldsymbol{\mu}_{a,k,m})^2}{\sum_{n=1}^{N} \sum_{t=1}^{T_n} \gamma(y_{t,k,m}^{(n)})},\ \boldsymbol{p}_k = \frac{\sum_{n=1}^{N} \sum_{t=1}^{T_n} \gamma(z_{t,k}^{(n)}) \boldsymbol{w}_t^{(n)}}{\sum_{n=1}^{N} \sum_{t=1}^{T_n} \gamma(z_{t,k}^{(n)})}, \\
\mu_{d,k} &= \frac{\sum_{n=1}^{N} \sum_{t=1}^{T_n} \gamma(z_{t,k}^{(n)}) d_t^{(n)}}{\sum_{n=1}^{N} \sum_{t=1}^{T_n} \gamma(z_{t,k}^{(n)})},\ \Sigma_{d,k} = \frac{\sum_{n=1}^{N} \sum_{t=1}^{T_n} \gamma(z_{t,k}^{(n)}) (d_t^{(n)} - \mu_{d,k})^2}{\sum_{n=1}^{N} \sum_{t=1}^{T_n} \gamma(z_{t,k}^{(n)})}, \\
\mu_{l,k} &= \frac{\sum_{n=1}^{N} \sum_{t=1}^{T_n} \gamma(z_{t,k}^{(n)}) l_t^{(n)}}{\sum_{n=1}^{N} \sum_{t=1}^{T_n} \gamma(z_{t,k}^{(n)})},\ \Sigma_{l,k} = \frac{\sum_{n=1}^{N} \sum_{t=1}^{T_n} \gamma(z_{t,k}^{(n)}) (l_t^{(n)} - \mu_{l,k})^2}{\sum_{n=1}^{N} \sum_{t=1}^{T_n} \gamma(z_{t,k}^{(n)})}
\end{aligned} \tag{7}$$

4.3 生成フェーズ

生成フェーズでは，確率モデルに基づいてコメントを合成・付与する．学習データと同様に，コメントを付与したい音響信号は $\boldsymbol{a}' = \{\boldsymbol{a}'_1, \cdots, \boldsymbol{a}'_{T'}\}$ で，すでに付与されたユーザコメントの内容・密度・長さは $\boldsymbol{w}' = \{\boldsymbol{w}'_1, \cdots, \boldsymbol{w}'_{T'}\}$，$\boldsymbol{d}' = \{d'_1, \cdots, d'_{T'}\}$，$\boldsymbol{l}' = \{l'_1, \cdots, l'_{T'}\}$ で与えられるとする．ここで，T' はフレーム数である．本フェーズは，各フレームに対して，どのようなコメントが生成され得るかを推定するアウトライン部と，その推定結果に基づき言語制約を考慮しながら単語を連結して文を生成するアセンブル部から構成される．

4.3.1 アウトライン部

最尤の状態系列 $\boldsymbol{z}' = \{\boldsymbol{z}'_1, \cdots, \boldsymbol{z}'_{T'}\}$ はビタビアルゴリズムを用いて推定できる．ユーザが付与したコメントを参考にしない場合は，出力確率を $p(\boldsymbol{o}'_t|\boldsymbol{\phi}_k) = b_{a,k}(\boldsymbol{a}'_t)$ とする．あるフレーム t において状態 k をとる（$z'_{t,k} = 1$）とすると，最尤のコメント密度 $\hat{d}_t$ はガウス分布 $b_{d,k}$ のモード（最頻値）であり $\hat{d}_t = \mu_{d,k}$ となる．したがって，全フレームに対して最尤密度を求めれば，合計が指定した総コメント数になるように各フレームに付与すべきコメント数を決定できる．同様に，最尤のコメント内容 $\hat{\boldsymbol{w}}_t$ は $b_{w,k}$ のモード $\boldsymbol{p}_k$ で与えられる．

4.3.2 アセンブル部

最尤のコメント内容である $\hat{\boldsymbol{w}}_t$ はスクリーニング後の V 単語の生起確率（縮退したユニグラム確率）であり，これだけでは以下の 3 つの理由で文を生成することはできない．

(1) 助詞や接続詞などの補助的な単語の生起確率が推定されていない．
(2) 活用語に対しては基本形以外の活用形の生起確率が推定されていない．
(3) 単語の連接確率が推定されていない．

例えば，コメント内容として「これ」や「すごい」が生起しやすく，コメント長が 3 だと推定されても,「これ+は+すごい」や「これ+すごく+好き」などの文は生成できない．

これらの問題を解決するため，スクリーニング前の全コメントから学習した汎用言語モデル（ユニ・バイ・トライグラム）を利用する．汎用モデルでは，品詞と語幹が同じでも活用形が違えば異なる単語とみなされ，単語の定義が縮退ユニグラムとは異なっている．汎用モデルを用いれば後述する方法で全楽曲に対するコメント文を生成できる．しかし，いまは縮退ユニグラム $\hat{\boldsymbol{w}}_t$ 中の単語生起確率を反映させて，ある楽曲のある時刻に対するコメント文を生成したい．したがって，汎用モデルを縮退ユニグラム $\hat{\boldsymbol{w}}_t$ に適応させる必要がある．まず，**図 3** に，汎用ユニグラムを縮退ユニグラム $\hat{\boldsymbol{w}}_t$ に適応させる方法を示す．最初に，$\hat{\boldsymbol{w}}_t$ 中の各単語の生起確率で汎用ユニグラム中の対応する単語の生起確率を更新する．ここで，$\hat{\boldsymbol{w}}_t$ の単語が活用語であれば一対多の対応となる．こうすると，縮退ユニグラムの単語はすべて汎用ユニグラムに出現しているので，更新した汎用ユニグラム中の生起確率の和は 1 を超えてしまう．そこで，確率を更新した単語の生起確率の和が α，もとのままの単語の生起確率の和は $1-\alpha$ になるよう正規化する．α を大きくすると，$\hat{\boldsymbol{w}}_t$ 中で生起しやすい単語（とその活用形）が出現しやすくなる．汎用バイ・トライグラムの適応方法はあとで述べる．

次に，最尤のコメント（単語列）$\hat{\boldsymbol{c}}_t$ およびコメント長 $\hat{\boldsymbol{l}}_t$ の生成モデルを次式で与える．

$$\{\hat{\boldsymbol{c}}_t, \hat{\boldsymbol{l}}_t\} = \underset{\boldsymbol{c},l}{\operatorname{argmax}}\, p(\boldsymbol{c}, l; \boldsymbol{\theta}_k) = \underset{\boldsymbol{c},l}{\operatorname{argmax}}\, p(\boldsymbol{c}|l; \boldsymbol{\theta}_k) p(l; \boldsymbol{\theta}_k) \tag{8}$$

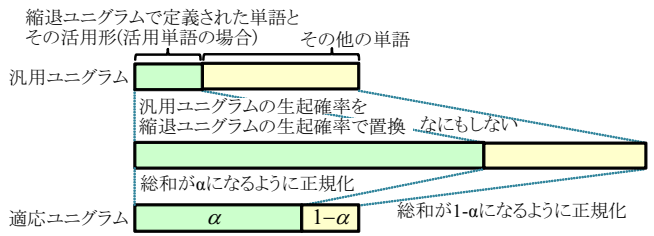

図 3 適応ユニグラムの作成：汎用ユニグラム確率への縮退ユニグラム確率の取り込み

ここで，$p(l; \boldsymbol{\theta}_k)$ は状態 k においてコメント長が l である確率であり，ガウス分布 $b_{k,l}$ で与えられる．$p(\boldsymbol{c}|l; \boldsymbol{\theta}_k)$ はコメント長が l であった場合に，単語列 $\boldsymbol{c}$ が生成される確率である．最尤コメント $\hat{\boldsymbol{c}}_t$ を求めるには, 各 l に対して $\operatorname{argmax}_{\boldsymbol{c}} p(\boldsymbol{c}|l; \theta_k)$ を計算する必要がある．

本研究では，ビタビアルゴリズムを用いて単語トレリス上の最尤単語経路を推定する手法を提案する．通常，HMM を用いた最尤状態経路の推定ではトレリスの各ノードが状態に対応するが，単語トレリスでは各ノードが単語に対応する．SilB および SilE をコメントの始端と終端を表す特別な記号とすると，コメント $\boldsymbol{c}$ の尤度は次式で与えられる．

$$p(\boldsymbol{c}|l) = p(w_1|\mathrm{SilB}) \left(\prod_{i=2}^{l} p(w_i|w_{i-2}, w_{i-1}) \right) p(\mathrm{SilE}|w_{l-1}, w_l) \tag{9}$$

w_i はコメント中で i 番目の単語であり，w_0 は SilB とする．$p(w_i|w_{i-2}, w_{i-1})$ は適応トライグラムであり，線形結合 $p(w_i|w_{i-2}, w_{i-1}) \propto \beta_t p_t(w_i|w_{i-2}, w_{i-1}) + \beta_b p_b(w_i|w_{i-1}) + \beta_u p'_u(w_i)$ で得られる．ここで，β_t，β_b および β_u は汎用トライグラム，汎用バイグラム，適応ユニグラムの重みである．適応バイグラム $p(w_1|\mathrm{SilB})$ も同様に $p(w_1|\mathrm{SilB}) \propto \beta_b p_b(w_1|\mathrm{SilB}) + \beta_u p'_u(w_i)$ とできる．最後に，コメント長で正規化して $p(\boldsymbol{c}|l) \leftarrow p(\boldsymbol{c}|l)^{\frac{1}{l}}$ とする．

5. 評価実験

評価実験として，システムが人間のコメントと似たコメントを生成できるのかを検証した．

5.1 実験条件

実験データとして，ニコニコ動画の音楽カテゴリからタイトルに「演奏してみた」を含む動画を，コメント数が多い順に 100 個収集した．次に，各動画から抽出可能な最大値である 1100 個のコメントを古い順に抽出した．得られた語彙数は $V = 2082$ であった．同様に，タイトルに「弾いてみた」を含む動画を 100 個収集し，各動画から 2400 個のコメントを抽

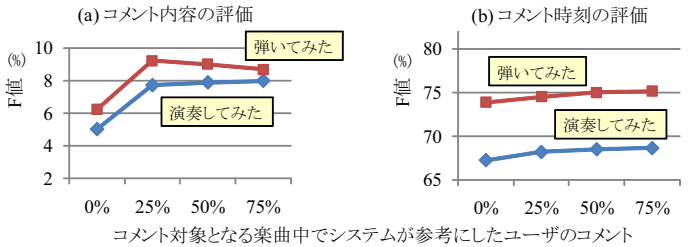

図 4 実験結果：生成したコメントとユーザコメントとの内容および付与時刻の一致率

出した．このとき $V = 2278$ となった．動画から抜き出した音響信号は 16 [kHz]・モノラルの PCM WAV 形式に変換し，フレームシフト長は 256 [ms] として特徴量抽出を行った．パラメータは $K = 200$，$M = 8$, $\alpha = 0.9$ および $\beta_t = \beta_b = \beta_u = 1.0$ とした．

実験は各タイトルごとに 4 クロスバリデーションで行った．すなわち，75 個の楽曲でモデルを学習し，残り 25 個の楽曲をテストデータとした．テストデータの各楽曲にコメントを付与する際に，既存のユーザコメントの分量を 0%，25%，50%，75%と変化させて与えた．

生成されたコメント内容は，F 値 ($= \frac{2PR}{P+R}$) で評価した．適合率 P と再現率 R を計算するため，システムがあるフレームにコメントを付与したとすると，その中の各単語に対して，付与された時刻の前後 5 秒以内にユーザが同じ単語を付与しているかをチェックした．ここで，単語とは縮退したユニグラムにおける V 種類を意味し，活用形は考慮しなかった．

$$P = \frac{\#\text{一致した単語}}{\#\text{システムが生成した単語}},\quad R = \frac{\#\text{一致した単語}}{\#\text{ユーザが付与した単語}} \tag{10}$$

5.2 実験結果

図 4 に示すように，25%のユーザコメントを利用すると，F 値が大きく向上した．依然 10%以下であるが，人間でも他人のコメントを単語レベルで正確に予測するのは難しいことを考えると，コメント生成の初の試みとして十分見込みがある成果であると考える．コメント付与時刻のみを評価した場合，F 値は 70%程度であった．また，コメント生成に利用するユーザコメントを 25%から増加させても，F 値は向上しなかった．

この原因として，現状の決定的な手法では各状態につき最尤のコメントのみが生成されるので，コメント種類数は高々状態数に制限されて多様性を表現しきれなかったことが考えられる．実際の生成結果を観察すると，ユーザが頻繁に使う表現が多く生成される傾向があり,「すごい」「うまい」「かっこいい」といった楽曲を肯定的に評価する頻出単語で 40%程度が占められていた．人間が見て有用なコメントとは，楽曲中のある内容をよく表す特徴的な単語で構成されるべきであり，コメントの的確性と多様性を向上するための改良は今後の課題である．ただし，現段階でも「この曲泣けてきたw」「タンバリンうめえええw」「ギターひどいw」「アレンジすごいと思うよw」「テンション上がったw」「この才能の無駄遣いw」などの興味深いコメントが生成された．

6. おわりに

本稿では，与えられた楽曲に対して，コメント文を生成し，それらを適切な時刻に付与するシステム MusicCommentator について述べた．本システムは音響的特徴量とコメント特徴量との同時的生成 HMM を基礎としている．HMM における状態遷移は，音楽のムードの遷移としてだけでなく，コメントのトピックの遷移としても解釈できる．学習データである多数の楽曲とコメントのペアから HMM のパラメータを最尤推定にて求めたあと，新たに与えられた楽曲に対して単語間の連接を考慮してコメントを生成した．

実験の結果，人間のように音楽に対してコメントを行える計算機をつくるという究極の目標に到達するには，未だ多くの課題があることが明らかとなった．コメントするという行為は人間の高度な能力の一つであるだけでなく，文化的な影響もあり，機械学習の手法だけでは実現が難しいかもしれない．しかし，我々の試みは学術的に重要なチャレンジであったと考えている．今後は音楽の特徴量として MFCC だけでなくリズムや歌唱に関する内容を取り入れたり，映像特徴量も考慮するなどしてシステムを改良していきたい．

謝辞: 本研究の一部は CrestMuse プロジェクト (JST CREST) の支援を受けた．

参考文献

1) ニコニコ動画: http://www.nicovideo.jp/
2) 濱野智史:「ニコニコ動画」をめぐる冒険—「擬似同期型アーキテクチャ　複製技術 II」のアーキテクチャ分析」. InterCommunication No.65 Summer 2008, Vol.17, No.3, NTT 出版, pp.90-95, 2008.
3) Whitman, B. and Rifkin, R.: Musical Query-by-Description as a Multiclass Learning Problem. MMSP, pp.153–156, 2002.
4) D. Turnbull, et al.: Semantic Annotation and Retrieval of Music and Sound Effects. IEEE Trans. on ASLP, Vol.16, No.2, pp.467–476, 2008.
5) T. Bertin-Mahieux, et al.: Autotagger: A Model for Predicting Social Tags from Acoustic Features on Large Music Databases. JNMR, Vol.37, No.2, pp.115–135, 2008.
6) 梶 克彦, 長尾 確: 楽曲に対する多様な解釈を扱う音楽アノテーションシステム. 情報処理学会論文誌, Vol.48, No.1, pp.258-273, 2007.
7) X. Amatriain, et al.: The CLAM Annotator: A Cross-platform Audio Descriptors Editing Tool. ISMIR, pp.426–429, 2005.
8) T. Kudo, et al.: Applying Conditional Random Fields to Japanese Morphological Analysis. EMNLP, 2004.