Vol.2009-MUS-80 No.5
2009/5/21

HMMに基づく歌声合成のための ビブラートモデル化

山田知彦[†1] 武藤聡[†1] 南角吉彦[†1]
酒向慎司[†1] 徳田恵一[†1]

HMM に基づく歌声合成は歌い手の特徴を歌声データと楽譜から自動学習し，任意のメロディからその特徴を再現した歌声を合成できる．その際，歌声の音色・発音と音高における歌い手の特徴を，それぞれスペクトルと基本周波数の時間変化としてHMM でモデル化している．本稿では，歌唱表現のひとつであるビブラートを音高の周期的な揺らぎと仮定し正弦波でモデル化する．そのパラメータをスペクトル及び基本周波数と同時に HMM でモデル化する．歌声の合成実験では，女性 1 名による童謡 60 曲の歌声データを学習し，主観評価実験によってビブラートモデルの導入による自然性の向上が確認できた．

Vibrato Modeling for HMM-based Singing Voice Synthesis

Tomohiko Yamada,[†1] Satoru Muto ,[†1]
Yoshihiko Nankaku ,[†1] Shinji Sako [†1]
and Keiichi Tokuda[†1]

HMM-based singing voice synthesis can automatically learn singer's features from singing voice waveform and musical scores and synthesize singing voice which the features are reflected in with any melody. The features in the signer's tone and pronunciation or pitch are modeled as a sequence of spectrum or fundamental frequency(F0) by HMM. In this report, we assume that vibrato is a periodic fluctuation of pitch and it is modeled by sinusoid. That parameters are modeled by HMM with spectrum and F0 simultaneously. In the experiments of subjective assessment, we confirmed that smooth and natural singing voice is synthesized.

1. まえがき

近年，コンピュータによる歌声合成が非常に注目を集めている．歌声合成は，任意のメロディの歌声を得ることができる技術である．この利便性から，高品質な歌声を合成する技術は，音楽制作やアミューズメント分野等での利用が期待される．既に様々な歌声合成ソフトが作られ，それを用いたコンテンツが多数作られており，歌声合成技術の進歩にはさらなる期待が持たれている．

歌声合成の手法のひとつに波形接続がある．波形接続は，歌い手の音声波形データを楽譜に基づいて接続し歌声を生成する手法である．この手法は生の波形を用いるためクリアな音声が得られるが，接続部分で歪みが発生しやすいという問題がある．

歌声合成のもうひとつの手法に隠れマルコフモデル（Hidden Markov Model; HMM）に基づく歌声合成[1] が挙げられる．これは HMM に基づく音声合成[2),3)] を歌声に応用したものであり，歌声データベースから歌い手の特徴を自動学習しその特徴を再現した歌声を合成する．HMM 歌声合成では，歌い手の歌声を統計的手法によりモデル化し，そのモデルを選択・接続して出力されるパラメータから歌声を合成する．その際，動的特徴量を併せて学習するので，モデルの接続部分で歪みのない歌声を合成することが可能である．ところで，歌声には楽譜の他に，歌うタイミングのずれ[4)] やビブラート等の楽譜にない表現が含まれていることがわかっている．これらは歌声に自然性や個人性を持たせる上で重要である．ところが，従来の HMM 音声合成ではビブラートのような微細な変動は，モデルの学習過程で平滑化されてしまうことから，合成される歌声の自然性を損ねていたと考えられる．

そこで本稿では，ビブラートをモデル化し HMM 歌声合成に導入することを提案する．この手法ではまず，ビブラートのパラメータを歌声データから抽出し従来の HMM 歌声合成の学習データに追加し学習を行う．その後，モデルと楽譜を基に得られるビブラートパラメータからビブラートを生成しそれを付加した歌声を合成する．

本節以降，2 節で HMM 歌声合成システム，3 節で本稿で導入するビブラートモデル，4 節で歌声の合成実験とその主観評価・考察について述べる．そして最後に 5 節でむすびとして，本稿のまとめと今後の課題について述べる．

†1 名古屋工業大学大学院 工学研究科
Department of Computer Science and Engineering, Nagoya Institute of Technology

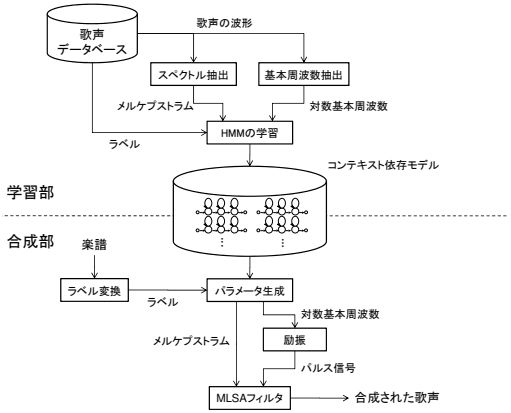


図 1 HMM に基づく歌声合成システムの概略図
Fig. 1 HMM-based singing voice system

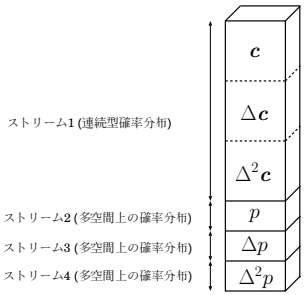


図 2 学習データにおける結合ベクトルの構造
Fig. 2 Vectorial structure of training data

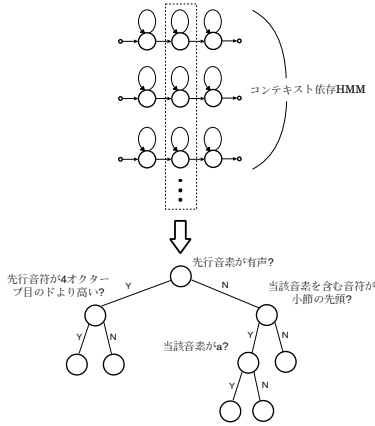


図 3 クラスタリングによる決定木の構築
Fig. 3 Constructing decision trees by clustering

2. HMM 歌声合成システム

本研究の基盤である HMM 歌声合成システムの概略図を**図 1** に示す．HMM 歌声合成システムは学習部・合成部の 2 つのパートで構成される．学習部では，歌声データベースから作成した学習データを HMM により音素単位でモデル化し学習を行う．合成部では，学習したモデルを楽譜に基に選択・接続し，そこから出力されるパラメータを用いて歌声を合成する．特徴量には歌声データベースから抽出したメルケプストラムと基本周波数を用いる．これらはそれぞれ音色・発音と音高に対応している．ここで，人間の聴覚は対数スケールであることから，基本周波数は対数基本周波数に変換する．また，特徴量の時間的変化を学習するためにメルケプストラムと対数基本周波数それぞれの 1 次動的特徴量及び 2 次動的特徴量（Δ，Δ^2）を求め，それらをすべて結合したベクトルを学習データとする． 学習データのベクトル構造を**図 2** に示す．ただし，$\boldsymbol{c}$，p はそれぞれメルケプストラム，対数基本周波数を表す．なお，対数基本周波数は有声区間（1 次元）と無声区間（0 次元）で次元が異なるため，多空間上の確率分布に基づく HMM （Multi-Space Probability Distribution HMM; MSD-HMM）[5] を用いて有声・無声による次元の変化に対応したモデル化を行う．

2.1 学 習 部

HMM 歌声合成システムでは音素単位で学習を行うが，同じ音素であっても当該音素や前後の音素の種類・音高・音長等の組み合わせによってその特徴量は変動する．一方，歌声情報の基本となる楽譜では，歌詞・音高・音長は音符単位で変動する．これらの変動要因を以降ではコンテキストと呼ぶことにする．本稿では以下の 4 種類のコンテキストを用いた．

- 音素（a，i，u，$\cdots$）
- 音高（楽譜での音階）
- 音長（4 分音符を基準とした相対的な音の長さ）
- 小節内位置構造（32 分音符の 3 分の 1 の単位での位置）

これらのコンテキストを当該音素・音符とその前後の音素・音符に適用し，より詳細なモデル化を実現する．このとき，コンテキストの組み合わせは膨大であり学習データに存在しない，もしくはごく僅かしかないモデルの学習は十分に行われない．この問題に対処する手法にコンテキストクラスタリング[6] がある．この手法はコンテキストの組み合わせを分類する質問をノードとする決定木を構築する．以下にコンテキストクラスタリングの手順を示す．

(1) あらかじめコンテキストに関して yes または no で答えられる質問を用意する．
(2) 全てのモデルを状態ごとに 1 つのクラスタにまとめる．

Vol.2009-MUS-80 No.5
2009/5/21

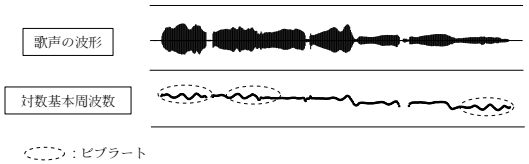


図 4 基本周波数に現れるビブラートの例
Fig. 4 Example of vibrato in F0

(3) 質問の中から分割前後の尤度変化に基づいて質問を選択し，クラスタに適用後 yes と no のクラスタに分割する．

(4) 分割された 2 つのクラスタそれぞれについて（3）を行い，停止条件を満たすまで繰り返す．

なお，停止条件には記述長最小化（Minimum Description Length; MDL）基準[7] が広く用いられている．**図 3** にコンテキストクラスタリングによる決定木の構築の例を示す．

構築された決定木のリーフノードには音響的特徴の類似した状態がまとめられ，パラメータの共有が行われる．これにより，各々のモデルに対して十分な量の学習データが与えられる．それに加えて，学習データに存在しないコンテキストの組み合わせを持つモデルに対しても，決定木を辿ることて類似したモデルのパラメータが適用されるため，このようなモデルを含む歌声の合成も可能になる．このようにして，60 曲程度の歌声データベースを学習すれば，歌い手の個人性を再現した歌声を合成可能である[1]．

2.2 合 成 部

図 1 に示すように，合成部では楽譜から作成したラベルに基づいて音素を選択・接続し，楽譜の音長の制約のもとパラメータ生成アルゴリズム[8] によってメルケプストラムと基本周波数系列を生成する．それらに MLSA フィルタ[9] を励振させることで歌声を合成する．なお，本研究では楽譜の記述形式として MusicXML[10] を利用した．

3. ビブラートモデル

ビブラートは音高や音量を周期的に揺らす歌唱表現である．その音高の例を**図 4** に示す．ビブラートのかかるタイミングや変動は歌い手によって異なり，これによって歌声に自然性や個人性が生まれると考えられる．しかし，従来の HMM 歌声合成システムでは，ビブラートのような音符内で発生する微細な変動は，モデルの学習過程で平滑化されてしまい，合成される歌声の自然性を損ねていた可能性があった．そこで，ビブラートを歌声データから自動学習し合成音声に再現するため，次節でビブラートモデルを導入する．

3.1 モ デ ル 化

本稿では簡単のため，ビブラートは歌声における音高の周期的な揺らぎであると仮定し，音量については考慮しないこととする．このとき，t 番目のフレームが i 番目（$i = 0, 1, \ldots, M-1$）のビブラート区間 $[t_i^{(0)}, t_i^{(1)}]$ に含まれるとき，ビブラート $v(\cdot)$（単位：cent）は次式でモデル化できる．

$$v(m_a(t), m_f(t), i) = m_a(t) \cdot \sin\left(2\pi \frac{m_f(t)}{f_r}(t - t_i^{(0)})\right) \quad (1)$$

ただし，f_r はフレーム周期の逆数であり，$m_a(t)$ と $m_f(t)$ はそれぞれ t 番目のフレームにおけるビブラートの振幅（単位：cent）と周波数（単位：Hz）である．また，無音・無声音区間及び非ビブラート区間においては $m_a(t) = 0$，$m_f(t) = 0$ とする．本稿では，$m_a(t)$ と $m_f(t)$ をビブラートパラメータとして学習及び合成に用いる．

3.2 パラメータの分析

ビブラートの抽出には中野らの手法[11] を参考にし，対数基本周波数系列からビブラート区間 $[t_i^{(0)}, t_i^{(1)}]$ を検出した．その際，ビブラートの振幅や周波数の制限範囲は中野らの手法に従い，それぞれ 30〜150cent，5〜8Hz とした．

得られたビブラート区間に対して $\Delta l f_0(t)$ の零交差点からピーク点を求め，パラメータの制限範囲に基づいて各ピークにおけるビブラートパラメータを求める．その様子を**図 5** に示す．ただし，$c = \log 2/1200$ であり，これは cent 単位から対数単位へのスケーリングのための定数である．次に，求めたピーク（$t_0, t_1, \cdots$）間のフレームにおけるビブラートパラメータを求める．パラメータは $m_a(t)$ と $m_f(t)$ のそれぞれについて線形補間を適用して求める．このとき，パラメータが制限範囲を超えた場合は範囲の上限及び下限をパラメータの値とする．ビブラートの振幅パラメータにおける線形補間の例を**図 6** に示す．

3.3 学習データ

モデルの学習は，3.2 節で述べた 2 つのパラメータ m_a，m_f を 2 次元のベクトルとし，図 2 の学習ベクトルと結合したものを用いて行う．ビブラートパラメータを結合して作成した学習データを**図 7** に示す．

3.4 HMM の出力確率

図 7 に示す通り，学習データ $\boldsymbol{o}_t$ は 3 種類の独立な観測データ（スペクトル，基本周波数，ビブラート）からなる．それぞれを $\boldsymbol{o}_t^{(spec)}, \boldsymbol{o}_t^{(f_0)}, \boldsymbol{o}_t^{(vib)}$ とすると，状態 s において $\boldsymbol{o}_t$ が出力される確率 $b_s(\boldsymbol{o}_t)$ は以下の式で与えられる．

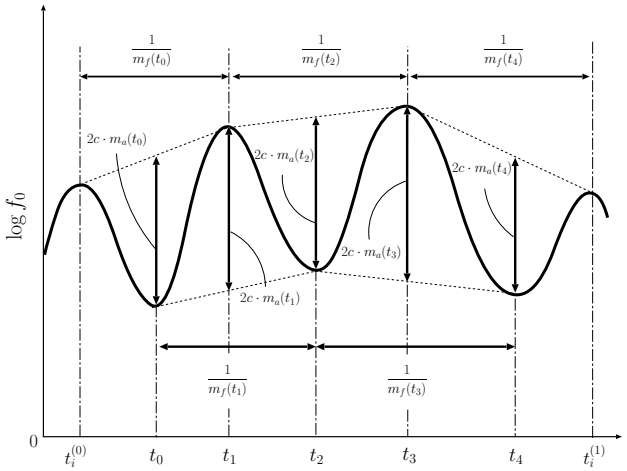


図 5 ビブラートパラメータの分析の概念図
Fig. 5 Analyzing vibrato parameters

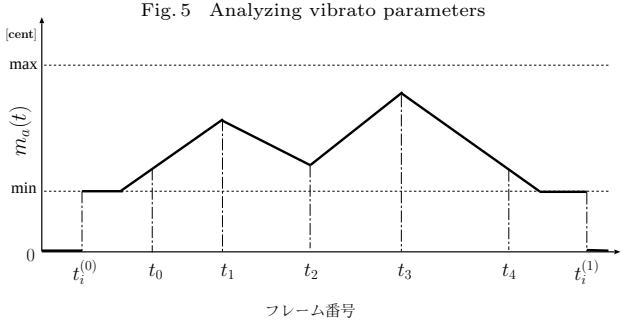


図 6 ビブラートパラメータの線形補間 (振幅パラメータ)
Fig. 6 Linear interpolation of vibrato parameters(amplitude parameter)

$$b_s(\boldsymbol{o}_t) = p_s^{\gamma_{spec}}(\boldsymbol{o}_t^{(spec)}) \cdot p_s^{\gamma_{f_0}}(\boldsymbol{o}_t^{(f_0)}) \cdot p_s^{\gamma_{vib}}(\boldsymbol{o}_t^{(vib)}) \tag{2}$$

ただし，$\gamma_{spec}, \gamma_{f_0}, \gamma_{vib}$ は 3 種類の観測データのそれぞれの重みである．本稿では上述の出力確率について，次に示す 2 つの手法でモデルの学習を行った．

(1) 出力確率にビブラートパラメータを考慮する手法
ビブラートの重み γ_{vib} を 1 に設定して出力確率 $b_s(\boldsymbol{o}_t)$ を計算する．この場合，モデルの学習時の尤度計算においてビブラートのパラメータが考慮される．

(2) 出力確率にビブラートパラメータを考慮しない手法
式（2）の出力確率の計算において，ビブラートの重み γ_{vib} を 0 に設定して出力確率 $b_s(\boldsymbol{o}_t)$ を計算する．この場合は学習時の尤度計算においてビブラートのパラメータは考慮されない．

なお，コンテキストクラスタリングにおいてはメルケプストラム，対数基本周波数，ビブラートについてそれぞれ個別の決定木を構築した．

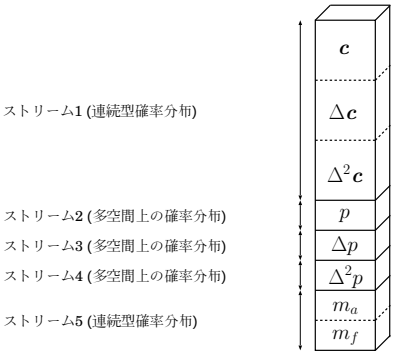


図 7 ビブラートパラメータを結合した学習データの構造
Fig. 7 Structure of training data combined vibrato parameters

3.5 合　　成

HMM で学習した対数基本周波数系列にビブラートパラメータ系列から計算される正弦波系列を重ね合わせることでビブラートを再現する．ある有声音区間におけるビブラートつき対数基本周波数系列 $lf_0(t)'$ は次式で表される．

$$lf_0(t)' = \begin{cases} lf_0(t) + c \cdot v(m_a(t), m_f(t), i) & (t_i^{(0)} \le t \le t_i^{(1)}) \\ lf_0(t) & (otherwise) \end{cases} \tag{3}$$

ただし，$lf_0(t)$ は基本周波数モデルから得られる対数基本周波数系列であり，$v(\cdot)$ は式（1）に学習したビブラートモデルのパラメータ $m_a(t)$，$m_f(t), i$ を代入して得た正弦波である．また，$v(\cdot)$ の係数は 3.2 節で説明したスケーリング定数である．

以上のようにして得られたビブラートつき対数基本周波数系列とメルケプストラム系列

表 1 歌声データベース
Table 1 Singing voice database

楽曲	童謡など 60 曲
総時間	64 分 17 秒
歌い手	女性 1 名
サンプリング周波数	16kHz
量子化ビット数	16bit short
チャネル	モノラル

表 2 歌声データの分析条件 (メルケプストラム)
Table 2 Experimental condition for Mel-cepstral analysis

サンプリング周波数	16kHz
FFT 長	1024point
分析周期	5ms
分析方法	24 次 STRAIGHT メルケプストラム[13]

表 3 歌声データの分析条件 (基本周波数)
Table 3 Experimental condition for F0 analysis

サンプリング周波数	16kHz
分析周期	5ms
抽出手法	TEMPO[14]

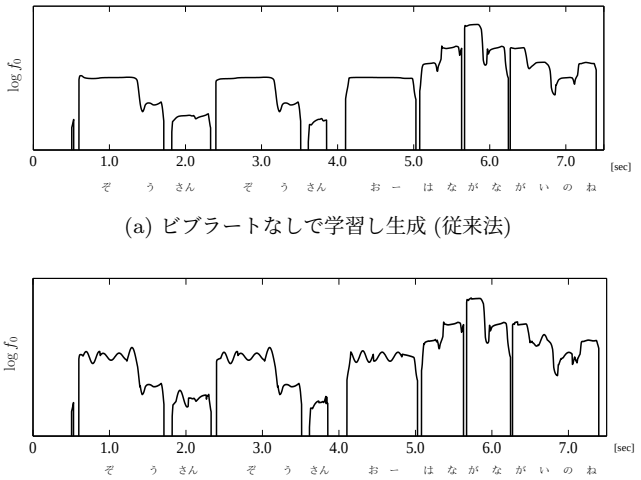
に対して，MLSA フィルタを駆動し歌声を合成する．

4. 歌声合成実験

HMM に基づく歌声合成システムにおいて，第 3.3 節で説明した学習データを用いてモデルを学習した．学習データに含まれない楽曲の MusicXML を用いて合成した．学習データに用いた歌声データベースを**表 1** に示す．

また，学習データの構成要素であるメルケプストラムおよび基本周波数の分析条件を**表 2**, **表 3** にそれぞれ示す．なお，今回の実験では時間構造をより精密に学習するために，明示的な状態継続長モデルを考慮した HMM である HSMM[12] を用いた．HSMM は状態数 5 の left-to-right 型のものを用いた．

童謡「ぞうさん」を合成した際の基本周波数の一部を**図 8** に示す．図中の（a）はビブラートなし（従来法）, (b）はビブラートを自動学習したモデルから生成されたものである．



(a) ビブラートなしで学習し生成 (従来法)

(b) ビブラートを自動学習し生成 (提案法)

図 8 童謡「ぞうさん」から合成された基本周波数
Fig. 8 Generated F0 contour for a Japanese song “ZOUSAN”

4.1 主観評価実験

ビブラートモデルの導入による自然性の向上を確かめるため，以下の 3 通りの方法で合成した歌声ついて評価実験を行った．

従来法 ビブラートパラメータを学習データに追加せずに歌声を合成（A）

提案法 1 ビブラートパラメータを学習データに追加し出力確率にビブラートを考慮する場合（B-1）

提案法 2 ビブラートパラメータを学習データに追加し出力確率にビブラートを考慮しない場合（B-2）

それぞれの手法について，歌声データベースに含まれない 10 曲の楽曲から歌声を合成して，5 小節程度で分割した 27 個のサンプルを用意し，その中から被験者ごとにランダムに選び出した 15 セットを用いた．被験者は 15 名で，歌声の自然性について 1〜5 の 5 段階の評価を行った．**図 9** に手法ごとの MOS 値[⋆1]を示す．

⋆1 被験者全員の評点の平均値

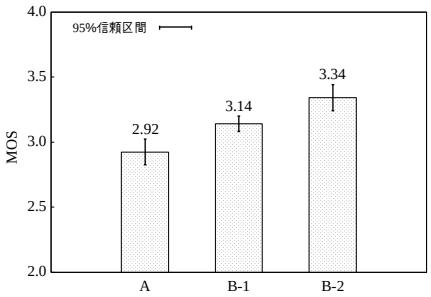


図 9 合成された歌声に対する主観評価値
Fig. 9 Result of MOS test for synthesized singing voices

4.2 考　　察

図 8 の対数基本周波数にビブラートが発生している．これは歌声データから抽出したビブラートのパラメータが HMM で正しく学習されたことを示している．また，図 9 に表されるように提案法が従来法に比べ高い評価を得た．これにより，ビブラートの付与が自然性の向上に有効であることが示された．

また，合成された歌声の基本周波数を調べたところ，音長の長い母音によくビブラートがかかる傾向が見られた．

5. む　す　び

HMM 歌声合成システムで合成される歌声の自然性を向上させるために，ビブラートモデルを導入した．従来法では，ビブラートのような音符内の微細な変動は学習において平滑され，合成される歌声に再現できなかった．それに対して提案法では，歌声データからビブラートパラメータを抽出して学習し，合成された歌声にビブラートが付与された．実験ではビブラートを付与した歌声の方が自然性が高いという印象を人間に与えることが確認された．

今後の課題として，複数名の合成実験を行い，声質だけでなくビブラート等の歌唱表現の個人性も合成音声に再現されるか検討することが挙げられる．

参　考　文　献

1) 酒向慎司，宮島千代美，徳田恵一，北村　正：隠れマルコフモデルに基づいた歌声合成システム，情報処理学会論文誌， Vol.45, No.3, pp.719–727 (2004).
2) 益子貴史，徳田恵一，小林隆夫，今井　聖：動的特徴を用いた HMM に基づく音声合成，信学論 (D-II)， Vol.J79-D-II, No.12, pp.2184–2190 (1996).
3) 吉村貴克，徳田恵一，益子貴史，小林隆夫，北村　正：HMM に基づく音声合成におけるスペクトル・ピッチ・継続長の同時モデル化，信学論 (D-II)， Vol.J83-D-II, No.11, pp.2099–2107 (2000).
4) Saino, K., Zen, H., Nankaku, Y., Lee, A. and Tokuda, K.: An HMM-based Singing Voice Synthesis System, *Proc. of Interspeech 2006*, Vol.1, pp.1–4 (2006).
5) 徳田恵一，益子貴史，宮崎　昇，小林隆夫：多空間上の確率分布に基づいた HMM，信学論 (D-II)， Vol.J83-D-II, No.7, pp.1579–1589 (2000).
6) Odell, J.J.: The use of context in large vocabulary speech recognition, *PhD dissertation, Cambridge University* (1995).
7) Shinoda, K. and Watanabe, T.: Acoustic modeling based on the MDL criterion for speech recognition, *Proc. EuroSpeech*, pp.99–102 (1997).
8) Tokuda, K., Kobayashi, T. and Imai, S.: Speech parameter generation from HMM using dynamic features, *Proc. ICASSP-95*, pp.660–663 (1995).
9) 今井　聖，住田一男，古市千恵子：音声合成のためのメル対数スペクトル近似（MLSA）フィルタ，信学論（A）， Vol.J66-A, No.2, pp.122–129 (1983).
10) Recordare: *MusicXML 2.0 Tutorial*. http://www.recordare.com/xml.html.
11) 中野倫晴，後藤真孝，平賀　譲：楽譜情報を用いない歌唱力自動評価手法，情報処理学会論文誌， Vol.48, No.1, pp.227–236 (2007).
12) Zen, H., Masuko, T., Tokuda, K., Kobayashi, T. and Kitamura, T.: A Hidden Semi-Markov Model-Based Speech Synthesis System, *IEICE Trans. Inf. & Syst.*, Vol.90D, No.5, pp.825–834 (2007).
13) Kawahara, H., Masuda-Katsuse, I. and Cheveigne, A.: Speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based f0 extraction: possible role of a repetitive structure in sound, *Speech Communication*, Vol.27, pp.187–207 (1999).
14) 河原英紀，Zolfaghari, P.，Cheveigne, A.，Patterson, R.D.：周波数から瞬時周波数への写像の不動点を用いた音源情報の抽出について，信学技報， Vol.SP99, No.40 (1999).