複合ウェーブレットモデルと隠れマルコフモデルの統合モデルによるテキスト音声合成*

☆北条伸克[1]，吉里幸太[1]，亀岡弘和[1,2]，齋藤大輔[1]，嵯峨山茂樹[1]
([1]東大院 情報理工, [2]NTT CS 研)

1 はじめに

本稿では，テキスト音声合成を扱う．統計的モデルに基づくテキスト音声合成方式の基本戦略は，音声の確率的な生成モデルを立て，学習データからそのモデルパラメータを学習させ，当該モデルを用いて任意のテキスト入力に対して音声を生成するというものである．したがって，音声におけるさまざまな性質や挙動をいかに適切に生成モデルの形で記述できるかが合成音声の品質を左右する．特に音声の音韻に着目すると，スペクトル包絡特徴量の時系列をいかにモデル化するかが重要であると言えるが，従来の隠れマルコフモデル（Hidden Markov Model, HMM）による音声合成はこのような動機により考案された手法である．HMM は区分定常な系列を表現する確率モデルであり，状態遷移がスペクトル包絡系列の時間方向の確率的な伸縮現象に相当する．では，状態出力分布（スペクトル包絡の確率分布）については，どのようなものを仮定すべきであろうか．

従来の HMM 音声合成方式では，スペクトル包絡特徴量として，ケプストラム [3] や LSP[2] が扱われていた．ケプストラムを特徴量とした場合，スペクトル包絡がパワー方向に確率的に揺らぐ現象を表現したモデルに相当し，LSP 特徴量の場合，スペクトル包絡のピークが周波数方向に確率的に揺らぐ現象を表現したモデルに相当する．

スペクトル包絡ピークの周波数とパワーは，声道における共振の周波数とゲインに相当し，声道形状の物理的な変化に従い時間方向に連続に変化する．例えば，ある音素と後続音素の接続区間では，声道形状が後続音素の声道形状へ連続的に変化する過程にあるため，共振の周波数もゲインも大きく変化する．このような周波数方向，パワー方向双方の揺らぎを表現可能なスペクトル包絡特徴量の確率モデルを構築することで，品質の高いテキスト音声合成を実現できるのではないかというのが本研究における着眼点である．

ケプストラムを特徴量とした HMM 合成音声では，合成音声のスペクトル包絡が周波数方向に平滑化される傾向にあったが，これは生成モデルがスペクトルの周波数方向の揺らぎを上手く捉えられないモデルであったことが原因であると考えられる．スペクトル包絡が平滑化されると一般には buzzy な音になるが，これは従来の HMM 音声合成において良く知られた傾向である．そのため，例えばスペクトル包絡のピークとディップの間を強調する目的で，確率モデルに Global Variance(GV) のモデルを導入することにより改善が図られてきた [4]．これに対し，本研究は，スペクトル包絡の生成モデルをどのように組み立てるべきかという問題意識に立ち返り，上述の問題に対する根本的な解決を目指すものである．

スペクトル包絡の各ピークがガウス分布で近似可能という仮定に基づき，スペクトル包絡全体を混合ガウス関数モデル（Gaussian Mixture Model, GMM）によって表現した複合ウェーブレットモデル（Composite Wavelet Model; CWM）が提案されている [1]．CWM は，スペクトル包絡ピークの周波数とパワーの双方をパラメータとしてもつため，スペクトル包絡のパワー方向と周波数方向の双方の揺らぎを確率的な現象として記述可能なモデルとして適している．. なお，CWM パラメータから音声波形を合成する際は，周波数領域におけるガウス分布関数は時間領域では Gabor 関数に対応するため，この Gabor 関数を基本周波数に対応する時間間隔で配置することにより音声波形が合成される．CWM に基づく分析合成は，FIR フィルタによる合成手法であり，従来の LSP やケプストラムなどの巡回型フィルタによる合成手法に比べ，Q 値の高いフィルタであっても，基本周波数によらず時間特性のよい音声が合成可能であることが示されている [1]．

以上の問題意識と CWM の利点より，本稿では CWM による HMM 音声合成の手法を提案する．

2 CWM と HMM の統合モデルによるスペクトル包絡系列の生成モデル

本章では，本研究で扱うスペクトル包絡系列の生成モデルについて述べる．

従来の HMM 音声合成方式 [3] では，ケプストラム特徴量系列を出力する HMM を立て，学習データから出力分布のパラメータを学習し，各状態での平均的なケプストラム特徴量が推定される．この方式で学習される，各状態の平均的なスペクトル包絡は周波数方向に平滑化される傾向にある．なぜなら，ケプストラムはスペクトル包絡の線形変換により得られるため，ケプストラムの平均を得ることはスペクトル包絡のパワー方向の平均を得ることと同等であるが，１章で見たようなスペクトル包絡ピーク周波数の揺らぎが存在すると，スペクトル包絡の山と谷が平均化され，なだらかな形状へ平滑化するためである．このようにスペクトル平滑化の原因は，ケプストラム特徴量の確率的な揺らぎを仮定し，スペクトル包絡のパワー方向のみの揺らぎをモデル化している点にあると考えられる．

１章でも述べたように，音声のスペクトル包絡に見られるゆらぎには，声道形状の物理的な変化に基づく共振周波数とパワーの変動が含まれると考えられるので，スペクトル包絡ピークの周波数とパワー

*Text-to-speech synthesis based on a combined system of the composite wavelet model and hidden Markov model. by HOJO, Nobukatsu, YOSHIZATO Kota, KAMEOKA Hirokazu, SAITO Daisuke, SAGAYAMA Shigeki (The University of Tokyo)

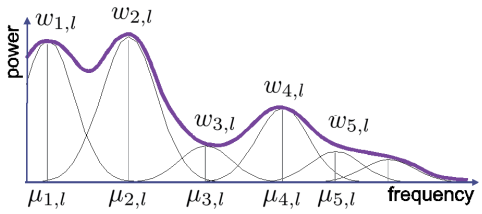


Fig. 1　CWM によるスペクトル包絡

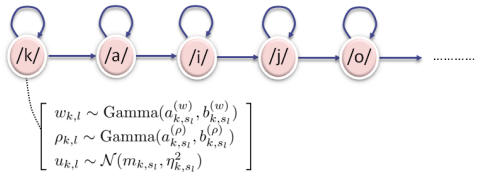


Fig. 2　提案 HMM の構成

の双方の揺らぎを表現できる確率的生成モデルを立てるべきである．

スペクトル包絡ピークの周波数とパワーをパラメータにもつ CWM により，この確率モデル化を行うことが可能である．CWM は，GMM によりスペクトル包絡を近似的に表現するモデルである．CWM によるスペクトル包絡 $f_{\omega,l}$ は，Fig. 1 のように，

$$f_{\omega,l} = \sum_{k=1}^{K} \frac{w_{k,l}}{\sqrt{2\pi}\sigma_{k,l}} \exp\left(- \frac{(\omega - \mu_{k,l})^2}{2\sigma^2_{k,l}} \right) \qquad (1)$$

のように表される．ただし，ω, l は周波数と時刻のインデックス，K は GMM の混合数である．また，$\mu_{k,l}, \sigma^2_{k,l}, w_{k,l}$ は，それぞれガウス関数を統計分布と見なした際の平均，分散，重みに対応し，スペクトル包絡のピーク周波数，ピークの鋭さ，パワーに対応するパラメータである．

以上の準備のもと，観測スペクトル系列が生成される過程を確率モデル化する．Fig. 2 のような，各離散時刻 l ごとに各スペクトルピークの周波数 $\mu_{k,l}$，拡がり $\sigma^2_{k,l}$ の逆数 $\rho_{k,l}$，パワー $w_{k,l}$ の CWM パラメータを出力する HMM を考える．HMM の各状態は，言語ラベルの一状態を表しており，例えば Fig. 2 中のようにそれぞれ一つの音素に対応させることもできるが，HTS[7] などの手法と同様に，音素状態に加え，前後の音素のやアクセント位置などの情報を用いたコンテキストラベルの一状態を対応させることも可能である．時刻 l における状態番号を s_l とし，本稿では状態出力分布（各状態における CWM パラメータの生成確率）を，後述のパラメータ推定アルゴリズムの導出の便宜上の都合により，

$$\mathrm{P}(\mu_{k,l}|s_l) = \mathcal{N}(\mu_{k,l}; m_{k,s_l}, \eta^2_{k,s_l}) \qquad (2)$$

$$\mathrm{P}(\rho_{k,l}|s_l) = \mathrm{Gamma}(\rho_{k,l}; a^{(\rho)}_{k,s_l}, b^{(\rho)}_{k,s_l}) \qquad (3)$$

$$\mathrm{P}(w_{k,l}|s_l) = \mathrm{Gamma}(w_{k,l}; a^{(w)}_{k,s_l}, b^{(w)}_{k,s_l}) \qquad (4)$$

と仮定した．ここで，$\mathcal{N}(x; m, \eta^2)$ は正規分布，$\mathrm{Gamma}(x; a, b)$ はガンマ分布

$$\mathrm{Gamma}(x; a, b) = x^{a-1} \frac{\exp(-x/b)}{\Gamma(a)\, b^a} \qquad (5)$$

である．また，上述の HMM により生成された CWM パラメータの系列 $\boldsymbol{\mu} = \{\mu_k\}_{k,l}, \boldsymbol{\rho} = \{\rho_k\}_{k,l}, \boldsymbol{w} = \{w_k\}_{k,l}$ が与えられたとき，時刻 l において，観測スペクトル $y_{w,l}$ が生成される確率分布についても後述のパラメータ推定アルゴリズムの導出の便宜上の都合により，

$$\mathrm{P}(y_{w,l}|\boldsymbol{\mu}, \boldsymbol{\rho}, \boldsymbol{w}) = \mathrm{Poisson}(y_{w,l}|f_{w,l}) \qquad (6)$$

と定めた．ここで，$f_{w,l}$ は，CWM パラメータ系列 $\boldsymbol{\mu}, \boldsymbol{\rho}, \boldsymbol{w}$ が与えられたとき，時刻 l の CWM パラメータを用いて (1) で表されるスペクトル包絡モデルであり，$\mathrm{Poisson}(x; \lambda)$ はポアソン分布である．なお，この仮定の下での λ の最尤推定問題は，スペクトル間の近さを測る尺度の一つとして近年音響信号処理分野で多用される I ダイバージェンスと呼ぶ歪み尺度を規準とした x と λ の最適フィッティング問題と等価となることが知られている [6]。上記の生成モデルを定めることにより，次章のパラメータ推定アルゴリズムを導出する．

3　モデルパラメータの学習

スペクトル包絡系列生成モデルパラメータの推定は，スペクトル包絡生成モデルのいわば逆問題である．これは，観測スペクトル系列 $Y = \{y_{w,l}\}_{w,l}$ が与えられたときに，スペクトル系列生成モデルパラメータ Θ の事後確率 $\mathrm{P}(\Theta|Y)$ を最大化する問題として定式化される．推定すべきパラメータ Θ は，HMM の出力状態列 $(\boldsymbol{s})$，HMM の各状態 i の出力分布パラメータ $(\boldsymbol{\theta} = \{m_{k,i}, \eta_{k,i}, a^{(\sigma)}_{k,i}, b^{(\sigma)}_{k,i}, a^{(w)}_{k,i}, b^{(w)}_{k,i}\}_{k,i})$ と，CWM パラメータ系列 $(\boldsymbol{\mu}, \boldsymbol{\rho}, \boldsymbol{w})$ である．

さて，Θ の事後確率 $\mathrm{P}(\Theta|Y)$ を最大化する Θ を求めることは難しいが，各変数について局所最適化を繰り返すことは可能である．このとき $\mathrm{P}(\Theta|Y)$ は，

$$\log P(\Theta|\boldsymbol{Y}) \stackrel{c}{=} \log P(\boldsymbol{Y}|\Theta) + \log P(\Theta), \qquad (7)$$

$$\log P(\Theta) \stackrel{c}{=} \log P(\boldsymbol{s}) + \log P(\boldsymbol{\mu}|\boldsymbol{s}, \boldsymbol{\theta}) + \log P(\boldsymbol{\rho}|\boldsymbol{s}, \boldsymbol{\theta}) + \log P(\boldsymbol{w}|\boldsymbol{s}, \boldsymbol{\theta}) \qquad (8)$$

と書ける．ここで $\stackrel{c}{=}$ は定数部分を除いて一致することを意味する．

$P(\Theta|\boldsymbol{y})$ を最大化（または $-P(\Theta|\boldsymbol{y})$ を最小化）する Θ を解析的に求めることは難しいが，以下に述べるように補助関数法に基づき局所最適化アルゴリズムを導くことができる [6]．ここで，$-\log \mathrm{P}(Y|\Theta)$ は，各時刻の観測スペクトル包絡 $(y_{\omega,l})$ とスペクトル包絡モデル $(f_{\omega,l})$ の擬距離である I ダイバージェンスを，全時刻について足し合わせたものに相当する．Jensen の不等式を用いて，$-P(Y|\Theta)$ の上限関数を

$$-\log P(Y|\Theta) = \sum_{\omega,l} (y_{\omega,l} \log \frac{y_{\omega,l}}{f_{\omega,l}} - y_{\omega,l} + f_{\omega,l}) \qquad (9)$$

$$\leq \sum_{\omega,l} [y_{\omega,l} \log y_{\omega,l} - y_{\omega,l} - \sum_{k} (\lambda_{k,\omega,l} \log \frac{g_{k,\omega,l}}{\lambda_{k,l}} + g_{k,l})] \qquad (10)$$

により設計することができ，その上限関数を用いて立てられる $-P(\Theta|Y)$ の補助関数をパラメータ Θ と

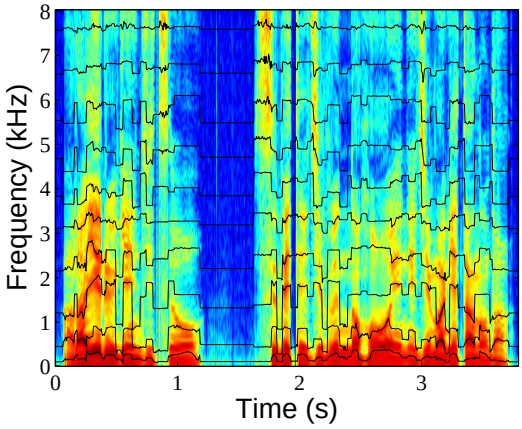


Fig. 3　音声「あらゆる現実をすべて自分の方へ捻じ曲げたのだ」のスペクトログラムと抽出した CWM 特徴量の平均パラメータの軌跡

補助変数 $\lambda_{k,\omega,l}$ に関して最小化するステップを繰り返すことで，$-P(\Theta|Y)$ を局所最小化する Θ を得ることができる．ただし，

$$g_{k,l} = \sqrt{\frac{\rho_{k,l}}{2}} \exp \frac{\rho_{k,l}(\mu_{k,l} - m_{k,i})^2}{2} \tag{11}$$

である．

補助関数を最小化する CWM パラメータ系列 $(\boldsymbol{\mu}, \boldsymbol{\rho}, \boldsymbol{w})$ についての更新式は，以下のように導かれる．

$$\mu_{k,l} = \frac{D_{k,l}\eta_{k,i}^2 + \frac{m_{k,i}}{\rho_{k,l}}}{C_{k,l}\eta_{k,i}^2 + \frac{1}{\rho_{k,l}}} \tag{12}$$

$$\rho_{k,l} = \frac{C_{k,l} - 2(a_{k,l}^{(\rho)} - 1)}{C_{k,l} - 2D_{k,l}\mu_{k,l} + E_{k,l} - \frac{1}{b_{k,l}^{(\rho)}}} \tag{13}$$

$$w_{k,l} = \frac{C_{k,l} - (a_{k,l}^{(w)} - 1)}{1 - \frac{1}{b_{k,l}^{(w)}}} \tag{14}$$

$$\lambda_{k,\omega,l} = \frac{w_{k,l}g_{k,l}}{\sum_{j=1}^{K} w_{j,l}g_{j,l}} \tag{15}$$

ただし，

$$C_{k,l} = \sum_{\omega} y_{\omega,l}\lambda_{k,\omega,l} \tag{16}$$

$$D_{k,l} = \sum_{\omega} \omega y_{\omega,l}\lambda_{k,\omega,l} \tag{17}$$

$$E_{k,l} = \sum_{\omega} \omega^2 y_{\omega,l}\lambda_{k,\omega,l} \tag{18}$$

である．HMM の出力状態列 $\boldsymbol{s}$, 出力分布パラメータ $\boldsymbol{\theta} = (\{m_{k,i}, \eta_{k,i}, a_{k,i}^{(\sigma)}, b_{k,i}^{(\sigma)}, a_{k,i}^{(w)}, b_{k,i}^{(w)}\}_{k,i})$ についての更新は，それぞれ Viterbi Alignment，Viterbi 学習により行うことができる．詳細は省略する．

Fig .3 に，ATR503 の A01 文「あらゆる現実を全て自分の方へ捻じ曲げたのだ」のスペクトログラムと，そこから抽出された CWM 特徴量のうち，平均パラメータの時間軌跡の例を示す．平均パラメータは共振周波数に相当するが，実際にスペクトル包絡ピーク付近に平均パラメータが対応づけられていることが確認される．パラメータが不連続に変化している時刻は，音素ラベルにより与えられた出力状態列 $(\boldsymbol{s})$ 初期値における音素境界である．

4　音声合成手法

本章では，２章，３章で述べたスペクトル包絡系列の生成モデルを用いて，テキスト音声合成を行う手法について述べる．

4.1　CWM 特徴量系列の合成

入力されたテキストから得られる音素状態系列に対し，尤度最大化規範により CWM 特徴量系列を出力することで，合成スペクトル包絡系列が得られる．

ただし，単に尤度最大化規範により得られる合成スペクトル包絡系列は，音素境界付近で不連続となり，合成音声品質の劣化の原因となる．先行研究 [7] では音素状態を細かく分割し，さらに動的特徴量（特徴量の１階，２階の時間差分量）のモデルを用いることにより連続的な特徴量系列の出力を実現しており，本研究もこのモデルを用いた．，一方で，例えば [8] のように，音素状態を細かく分割せず，２次系フィルタのフィルタリングにより CWM 特徴量の時間変化を記述するモデルを用いることにより，特徴量系列の生成過程をより自然に表現する手法も考えられる．

CWM 特徴量系列の出力の際には，音素状態の Duration に関するモデルが別途必要である．また，合成 CWM 特徴量系列から合成音声波形を得るためには，別途基本周波数に関するモデルが必要である．今回は，双方とも HTS-2.1[7] の手法を用いた．

4.2　音声波形合成

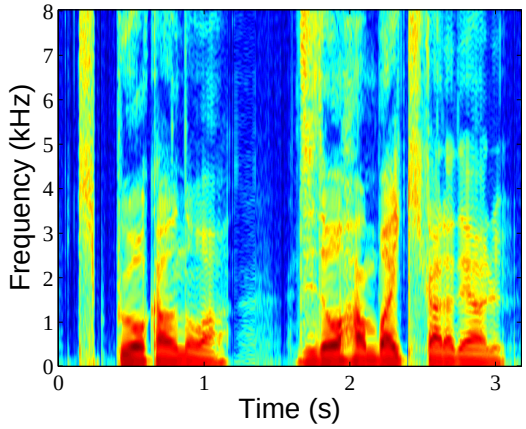
CWM 特徴量と基本周波数情報を用いて音声波形を合成する手法については，[1][9] の手法と同様である．

5　予備実験

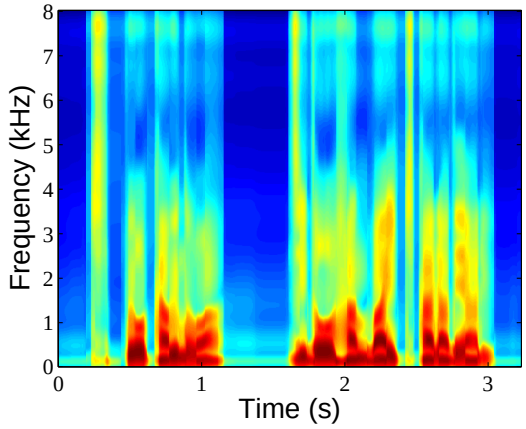
本章では，２章，３章，４章で述べたスペクトル包絡系列生成モデルを用いた音声合成手法に関し，適切に特徴量の推定・音声合成が実行可能であることの検証のために行った予備実験について述べる．

5.1　実験手法

今回行った予備実験では，スペクトル包絡生成モデルパラメータの推定手法として，３章で述べたようなパラメータの反復推定は行わず，音素境界位置と標準的な CWM 特徴量について適当な初期値を定め，各時刻の CWM 特徴量の抽出と，標準的な CWM 特徴量の推定を一度ずつ行った．音素境界位置の初期値としては HTS[7] の音素ラベルを活用した．各状態の出力分布パラメータの初期値は，[9] の手法により各音素に対応する時刻のスペクトル系列全体に対する CWM フィッティングを行い定めた．また，出力分布の推定には Viterbi 学習ではなく，HTS-2.1[7] と同様の手法により，動的特徴を含め，Baum-Welch 学習を行い，音声合成を行った．スペクトル包絡の解析は，STRAIGHT[5] により行った．学習には，HTS 2.1 のデモスクリプト [10] に同梱された男性話者 1 名の音声（サンプリング周波数 16 kHz ・サンプルサイズ 16



(a) サンプル音声



(b) 合成音声

Fig. 4　スペクトログラム（「切符を買うのは自動販売機からである」）

bit ）のうち 450 文を用い，評価文章として 53 文を用いた．

5.2　実験結果と考察

ATR503 の J04 文「切符を買うのは自動販売機からである。」の (a) サンプル音声のスペクトログラムと (b) 合成音声のスペクトログラムを Fig. 4 に，冒頭「切符」の音素/i/の中央部のスペクトル包絡を，予備実験の手法（赤），従来法（青），肉声（黒）についてそれぞれ Fig. 5 に示す．ただし，ここで従来法とは，24 次メルケプストラムによる手法 [3] である．Fig. 4 に示された，今回の手法により合成されたスペクトログラムは，肉声のものと類似しており，本手法によりテキスト音声合成が可能であることを示している．今回の手法で再現されたスペクトル包絡は，おもに 4kHz から 7kHz の周波数において，スペクトル包絡のディップを上手く再現する傾向があった．これは，CWM パラメータがスペクトル包絡ピークの周波数とパワーの両方のゆらぎをとらえたため，従来法に比べスペクトル包絡が平滑化しにくくなった結果であると考えることができる．一方で，1kHz 以下の低周波数において，複数のスペクトル包絡ピークがなだらかな曲線で再現されており，共振周波数が不明瞭となり，品質劣化の原因となっていると考えられる．

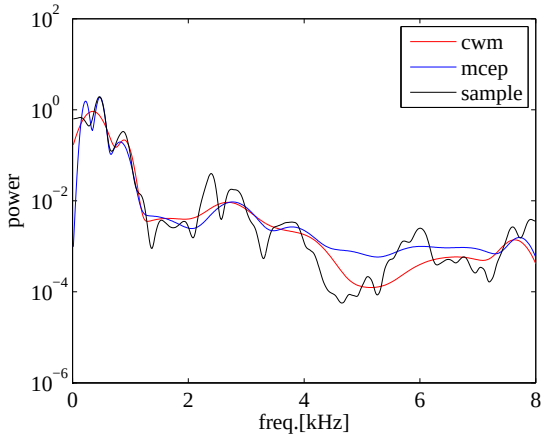


Fig. 5　「切符を買うのは自動販売機からである。」先頭音素/i/中央部のスペクトル包絡再現結果

これは，特徴量抽出の際，複数のスペクトル包絡ピークを少数のガウス関数の和で近似しているためであると考えられる．共振周波数がより明瞭な音声を合成するためには，例えば GMM の混合数を増やすなど，スペクトル包絡の各ピークに対し，精緻にガウス関数を対応付けるなどの方法が可能であると考えられる．

6　おわりに

本稿では，HMM 音声合成の手法として，CWM 特徴量を用いたスペクトル包絡系列生成モデルを用いる手法を考案し，その有効性を検証するための予備実験を行った．合成音声のスペクトルは，スペクトル包絡ピークがより鮮明になる傾向にあり，合成音声の品質向上のために有効な手法であることが確認された．今後の課題は，考案手法の，特徴量学習学習・Viterbi Alignment・確率分布の推定を反復することにより，合成音声の品質向上を達成することである．合成音声の品質向上に取り組むとともに，合成音声の主観評価実験を行い，従来法と比較するなど，品質に対する定量評価を行う予定である．

参考文献

[1] 槻他，音講論（春），2-11-7，2006.
[2] 管村他，電子通信学会論文誌，Vol. J64-A，pp. 599–606，1981.
[3] 徳田他，日本音響学会誌，Vol. 53, No. 3, pp. 192–200，1997.
[4] T. Toda and K. Tokuda, *IEICE Transactions*, Vol. E90-D, No. 5, pp. 816–824，2007.
[5] H. Kawahara *et al.*, Speech Communication, Vol. 27，No. 3-4，pp. 187–207，1999.
[6] H. Kameoka *et al.*, IEEE Trans. on Audio, Speech and Language Processing，Vol. 18，No. 6，pp. 1507-1516，2010.
[7] 全他，情報処理学会研究報告，No. 2007-12-20, pp. 301–306,，2007.
[8] 吉里他，音講論（春），in press，2013.
[9] 北条他，音講論（秋），no. 2-2-7，2012.
[10] http://hts.sp.nitech.ac.jp/