動的特徴量を考慮したピッチの高精度モデル化手法 *

◎全 炳河　徳田 恵一　益子 貴史 †　小林 隆夫 †　北村 正　(名工大, † 東工大)

1. はじめに

ピッチパターンのモデル化手法として, 多空間上で定義される確率分布に基づく HMM (multi-space probability distribution HMM: MSD-HMM) を用いた手法を提案した [1]. この手法では, 有声/無声境界においての動的特徴量を考慮せずピッチをモデル化していたため, 尤度最大化基準に基づく HMM からのパラメータ生成手法 [2] を用いてピッチパターンを生成 [3] した際, 有声/無声境界において不連続なピッチパターンが生ずる場合があった. そこで本研究では, 有声/無声境界での動的特徴量を計算し, MSD-HMM によりピッチパターンを動的特徴量を含めてモデル化する. また, ピッチパターン生成例及び受聴試験より, 提案する手法の有効性を示す.

2. 有声/無声境界での動的特徴量

従来は図 1(a) のように, 当該フレームとその前後それぞれ 1 フレームずつで計算した 1 次及び 2 次の回帰係数から微分係数に対応する値を求め, 動的特徴量としてモデル化した. 時刻 t における 1 次と 2 次の微分係数 $\delta p_t, \delta^2 p_t$ は以下のようになる.

$$\delta p_t = \frac{1}{2}(p_{t+1} - p_{t-1}) \tag{1}$$

$$\delta^2 p_t = \frac{1}{4}(p_{t+1} - 2p_t + p_{t-1}) \tag{2}$$

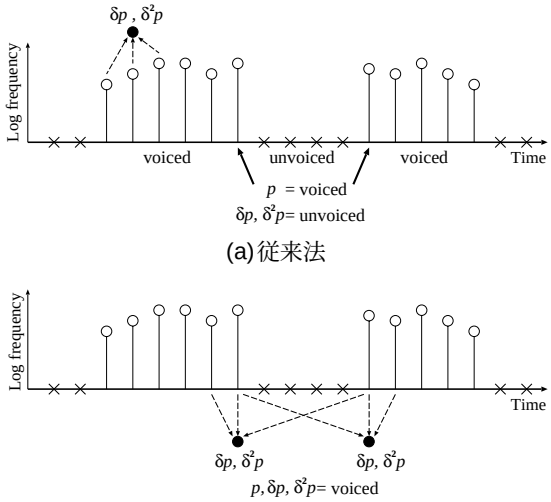
$\delta p_t, \delta^2 p_t$ はそれぞれ式 (1),(2) において, 計算に使われる静的特徴量が全て有声の場合のみ計算され, 無声区間及び有声/無声の境界の有声フレームについては, 動的特徴量を計算していない.

そこで, 本研究では図 1(b) に示すように, 当該フレーム及び前後最近傍の有声フレームから回帰係数を計算する. 得られた回帰係数から微分係数に対応する値を求め, これを動的特徴量とする. m フレーム前の有声フレーム p_{t-m}, 当該フレーム p_t 及び n フレーム先の有声フレーム p_{t+n} より当該フレームの微分係数を求める式は以下のようになる.

$$\delta p_t = \frac{m^2 p_{t+n} + (n^2 - m^2)p_t - n^2 p_{t-m}}{m^2 n + m n^2} \tag{3}$$

$$\delta^2 p_t = \frac{m p_{t+n} - (m+n)p_t + n p_{t-m}}{2(m^2 n + m n^2)} \tag{4}$$

また, 音声の最初の有声/無声境界では, 当該フレームより前の有声フレーム, 最後の有声/無声境界では当該フレームより後の有声フレームが得られないため, 最初の有声/無声境界では式 (3),(4) の m , 最後の有声/無声境界では n を無限大として微分係数を計算する. この場合は 1 次の微分係数は得られるが, 2 次の微分係数は 0 となる.



(a) 従来法

(b) 提案法

図 1．動的特徴量の計算

3. MSD-HMM によるピッチのモデル化

ピッチパターンは連続値をとる有声区間と値を持たない無声区間の時系列として表されるため, 通常の連続 HMM や離散 HMM では直接モデル化することができない. そこで, ピッチパターンを有声区間での静的特徴量と動的特徴量を出力する m 次元空間 Ω_1 と, 無声区間に対応する 0 次元の空間 Ω_2 の二つの空間から出される観測事象と考え, MSD-HMM によりモデル化する.

有声/無声を表す空間インデックスの集合を X , 有声区間におけるピッチの値とその動的特徴量を含んだ特徴ベクトルを $\boldsymbol{x}$, ピッチに関する観測事象を $\boldsymbol{o} = (X, \boldsymbol{x})$ とする. $X = \{1\}$ のときには, 有声区間を表し, $\boldsymbol{x}$ はピッチの静的特徴量と動的特徴量を含んだ m 次元のベクトルである. また, $X = \{2\}$ のときには, 無声区間を表し, $\boldsymbol{x}$ は 0 次元 ($\boldsymbol{x}$ は値を持たない) となる. このとき, MSD-HMM の状態 i における観測事象 $\boldsymbol{o}$ に対する出力確率は, 次のように表される.

$$b_i(\boldsymbol{o}) = \sum_{g \in S(\boldsymbol{O})} w_{i_g} \mathcal{N}_{i_g}^{n_g}(V(\boldsymbol{o})) \tag{5}$$

但し, $V(\boldsymbol{o}) = \boldsymbol{x}$, $S(\boldsymbol{o}) = X$ であり, w_{i_g} は各空間に対する重み, $\mathcal{N}_{i_g}^{n_g}$ は各空間の分布で, $n_1 = m$, $n_2 = 0$ であり, $\mathcal{N}_{i_2}^{0}(V(\boldsymbol{o})) = 1$ とする.

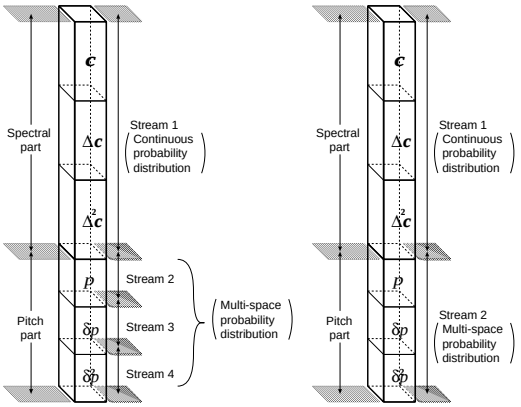
各状態の出力確率分布を式 (5) で定義することにより, HMM の枠組みでピッチパターンを直接モデル化することができる.

4. 特徴ベクトル

スペクトルとピッチをそれぞれ別々の HMM でモデル化した場合, スペクトルとピッチで音素境界を

∗ An accurate modeling method of pitch pattern considering dynamic features.

By Heiga Zen, Keiichi Tokuda, Takashi Masuko†, Takao Kobayashi† and Tadashi Kitamura (Nagoya Institute of Technology, †Tokyo Institute of Technology)



(a)従来法　　(b) 提案法

図 2．特徴ベクトル

合わせるためには何らかの工夫が必要となる．また, ピッチのみを特徴ベクトルとした場合では, 有声区間, 無声区間ともに音素に関する情報が不足するため, 音素境界を適切に学習することができない．そこで, メルケプストラムと対数基本周波数を合わせて一つのベクトルとし, これを特徴ベクトルとする．

また, 従来は図 2(a) に示すように, ピッチの静的特徴量, 1 次動的特徴量, 2 次動的特徴量をそれぞれ別のストリームでモデル化していた．これに対して提案法では, 図 2(b) のように一つのストリームでピッチをモデル化する．

HMM の各状態の有声/無声を決定する際, 従来は各状態におけるピッチの静的特徴量のストリームにおける空間の重み (式 (5) 中の w_g) のうち, $w_1 > w_2$ となる状態を有声区間, それ以外を無声区間としていた．提案法では単一のストリームでピッチの静的特徴量及び動的特徴量をモデル化している．このため, 提案法では有声/無声の判定を動的特徴量も含んでいるストリームの空間重みで決定している．

5. ピッチパターンの生成

まず, ピッチパターンを生成するターゲットとなる文章をラベル列に変換する．このラベル列に従って, 学習された HMM を接続し, 一つの文 HMM を作る．次に, 状態継続長分布に基づいて各 HMM の状態継続長を決定し, 各フレームの有声/無声を決定する．有声区間について, パラメータ生成手法を用いてピッチパターンを生成する．従来法では有声/無声境界では動的特徴量を考慮せずにパラメータ生成を行っていたが, 提案法では式 (3),(4) に基づき有声/無声境界における動的特徴量を考慮してピッチパターンを生成する．

6. 実験条件

HMM の学習データとして，ATR 日本語音声データベース B セットの話者 MHT による音韻バランス文 503 文のうち 450 文を用いた．音声データのサンプリング周波数は 16kHz, 分析周期は 5ms で, 長さ 25ms のブラックマン窓を使用した．スペクトル分析には, 24 次のメルケプストラム分析を行った．学習には 0 次から 24 次までのメルケプストラムと対数基本周波数, 及びそれらの動的特徴量を結合したベクトルを用いた．メルケプストラムの分布とピッチの

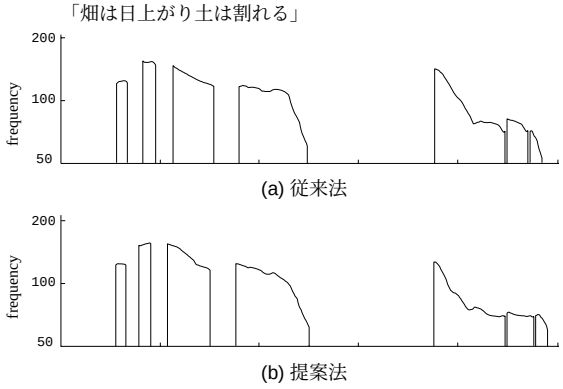


図 3．生成されたピッチパターンの例

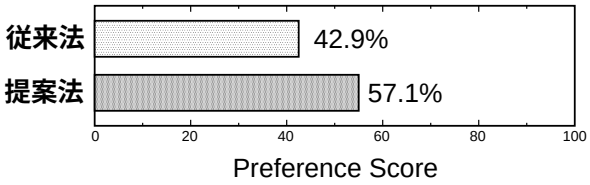


図 4．従来法と提案法の比較

分布それぞれを決定木を用いて MDL 基準でクラスタリング [4] した．クラスタリング後のピッチの総分布数は, 従来法では 1045, 提案法では 1700 となった．

従来法及び提案法で, 学習データに存在しない文章に対して生成されたピッチパターンの例を図 3 に示す．特に, 短い無声区間を挟んだ有声/無声境界におけるピッチパターンの不連続が解消されている．

7. 評価実験

本研究では, 提案法による音質の向上を確認するため, 従来法及び提案法で学習したモデルから音声を合成し, 対比較試験を行った．被験者は男性 8 名で, テスト用の 53 文章の中から, 被験者毎にランダムに 30 文章を選び出して提示し, スペクトル, ピッチ及び状態継続長などを総合的に評価させた．プリファレンススコアを図 4 に示す．図より, 提案法が従来法より優れていることが確認できる．

8. むすび

本研究では, MSD-HMM を用いたピッチパターンのモデル化において, 有声/無声境界での動的特徴量を考慮した手法を提案した．学習したモデルよりピッチパターンを生成し, 有声/無声境界における不連続性が解消されることを示した．また評価試験により, 従来の手法に比べ合成音の品質が向上していることを確認した．

参考文献

[1] 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,” 電子情報通信学会論文誌, J83-D-II, No. 7, pp.1579-1589, 2000.

[2] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, T. Kitamura, “Speech parameter generation algorithms for HMM-based speech synthesis,” Proc. of ICASSP, pp.III-1315-1318, June 2000.

[3] 益子 貴史, 徳田 恵一, 宮崎 昇, 小林 隆夫, “多空間確率分布 HMM によるピッチパタン生成,” 電子情報通信学会論文誌, J83-D-II, No. 7, pp.1600-1609, 2000.

[4] 篠田 浩一, 渡辺 隆夫, “情報量基準を用いた状態クラスタリングによる音響モデルの作成” 信学技報, SP96-79, 1996.