論文 / 著書情報
Article / Book Information

論題(和文)	ＨＭＭ音声合成における数量化Ｉ類を用いた発話速度制御法
Title(English)	
著者(和文)	外川 太郎, 山田 真裕, 岩野 公司, 古井 貞熙
Authors(English)	Koji Iwano, SADAOKI FURUI
出典(和文)	日本音響学会 2002年秋季講演論文集, Vol. , No. 3-10-9, pp. 345-346
Citation(English)	, Vol. , No. 3-10-9, pp. 345-346
発行日 / Pub. date	2002, 9

HMM 音声合成における数量化Ｉ類を用いた発話速度制御法 *

◎外川 太郎　山田 真裕　岩野 公司　古井 貞熙　(東工大)

1　はじめに

柔軟性のある音声合成を実現するために，分析合成系を用いた HMM に基づく音声合成 [1, 2] において，様々な発話速度での音声合成を実現する手法を提案する．本手法は，3 種類の発話速度に対応した継続時間長のモデルを数量化 I 類を用いて作成し，それらを補間することによって任意の発話速度の継続時間長モデルを構築する．本手法を用いて合成した様々な発話速度の音声について，主観評価実験により自然性の評価を行う．

2　数量化Ｉ類による音素継続時間長制御

高精度な音素継続時間長の制御手法として，統計的手法である数量化 I 類 [3] を利用した音素継続時間長制御法 [4, 5] が提案されている．本研究では，これらの手法と同様に数量化 I 類を用いた音素継続時間長制御を行う．

数量化 I 類 [3] とは，質的説明変数（制御要因）と目的とする量的変数を線形重回帰分析に基づいてモデル化する手法であり，以下の式で定式化される．

$$\hat{y}_i = \bar{y} + \sum_f \sum_c x_{fc}\delta_{fc}(i) \quad (i = 1, \cdots, N) \quad (1)$$

ここで，$\hat{y}_i$ は i 番目のサンプルの推定値，$\bar{y}$ は全データの平均値，N はサンプル数である．$\delta_{fc}(i)$ は i 番目のデータの制御要因 f がカテゴリ c に属する場合に 1，それ以外のときに 0 を与える関数である．x_{fc} は制御要因 f のカテゴリ c の数量であり，推定二乗誤差 $E = \sum_i (\hat{y}_i - y_i)^2$ を最小化するように求められる．制御要因 f のカテゴリ c に属するデータの総数を n_{fc} とするとき，数量 x_{fc} は以下の条件を満たしている．

$$\sum_f \sum_c x_{fc} n_{fc} = 0 \quad (2)$$

数量化 I 類は，調音様式にしたがって分類した 13 の音素クラスごとに行い，各クラスごとに音素継続時間長モデルを作成する．制御要因としては，推定対象音素とその前後 2 音素の計 5 つを用いた．

3　任意の発話速度での音声合成手法

任意の発話速度での音声合成を実現するためには，与えられた発話速度に対応した音素継続時間長モデルを用意し，それを用いて音素継続時間長を決定する必要がある．そこで本節では，1) 普通（normal）・速い（fast）・遅い（slow）発話速度で読み上げた音声データを使って 3 種類の音素継続時間長モデルを作成し，2) そのモデル間を補間することで目標速度での継続時間長モデルを作成し，3) そのモデルを利用して音声を合成する，という発話速度制御法を提案する．

3.1　各発話速度の音素継続時間長モデル作成

男性話者 1 名による，3 種類の発話速度（普通・速い・遅い）での音声データを収録した．読み上げる文章として，ATR 日本語連続音声データベース [6] の 503 文章を使用した．まず,「普通」速度で 503 文章を読み上げ，それらを用いて合成フィルタ作成用の triphone HMM の学習を行う．そのうち前半の 300 文について，学習した HMM を用いて強制切り出しを行って音素境界を定め，数量化 I 類によって音素継続時間長モデルを作成する.「速い・遅い」に関しても前半の 300 文を，発声者が不自然さを感じない範囲でできる限り速く，あるいは遅く，読み上げてもらった．音素切り出しは，普通速度の音声で学習した HMM で行い，同様にそれぞれの発話速度での音素継続時間長モデルを作成した．このように，数量化Ｉ類の目標値となる音素継続時間長を合成用 HMM を用いた強制切り出しによって得ることで，継続時間長制御と音声合成で用いる音素単位の定義基準が一致し，基準の不一致による合成音声の品質劣化を避けることができる．

ここで，発話速度 s $(s = fast, normal, slow)$ の音声データの無音区間を除いた総時間長を $t(s)$ とする．この値から 1 モーラあたりの平均モーラ時間長 $ML(s)$ を求めたところ，

$$\begin{aligned} ML(fast) &= 104.5 \ [\text{ms}] \\ ML(normal) &= 149.6 \ [\text{ms}] \\ ML(slow) &= 307.6 \ [\text{ms}] \end{aligned} \quad (3)$$

であった．

3.2　補間による音素継続時間長モデル作成

3 つのモデル間を補間することで，任意の発話速度での音素継続時間長モデルを生成する．そのため，普通速度での平均モーラ時間長 $ML(normal)$ を基準にして，発話速度 s における時間長伸縮率 $R(s)$ を以下のように定義する．

$$R(s) = \frac{\log ML(s) - \log ML(normal)}{\log 2} \quad (4)$$

各発話速度データの伸縮率は，

$$\begin{aligned} R(fast) &= -0.52 \\ R(normal) &= 0 \\ R(slow) &= 1.04 \end{aligned} \quad (5)$$

となる．

モデルの補間は，この伸縮率に基づいて行われる．発話速度 s における数量化 I 類の数量を $x_{fc}(s)$ とし

∗ A method of speech rate control using Quantification Theory (Type 1) for HMM-based TTS

By Taro Togawa, Masahiro Yamada, Koji Iwano and Sadaoki Furui (Tokyo Institute of Technology)

たとき，目標とする発話速度 t における数量 $x_{fc}(t)$ を以下のように求める．

$$x_{fc}(t) = \begin{cases} x_{fc}(normal)(1 - \frac{R(t)}{R(slow)}) \\ \quad + x_{fc}(slow)\frac{R(t)}{R(slow)} & (R(t) \geq 0) \\ x_{fc}(normal)(1 - \frac{R(t)}{R(fast)}) \\ \quad + x_{fc}(fast)\frac{R(t)}{R(fast)} & (R(t) < 0) \end{cases} \tag{6}$$

なお，補間によって作成した新しい数量 $x_{fc}(t)$ についても，式 (2) で示される条件は満たされる．

このような数量の直線補間を 13 の音素クラスごとに行うことで，継続時間長モデルの補間を行う．

4　対比較実験

「速い」と「普通」の中間にあたる「やや速い（$R(s) = -0.26$）」,「遅い」と「普通」の中間にあたる「やや遅い（$R(s) = 0.52$）」を目標の発話速度と想定して，それぞれの音素継続時間長モデルを作成し，音声を合成した．また，収録データから直接得られた「速い」「遅い」発話の継続時間長モデルを用いた場合の音声も合成し，併せて 4 種類の発話速度について合成音声の自然性を評価した．

比較対象として，簡単な発話速度の制御を行う方法（ルールベース法）を用意した．この手法は，まず，普通の発話速度で作成した数量化 I 類モデルを利用して，評価文の音素継続時間長を求める．各音素の継続時間長について，1) 母音・長母音ならば r 倍，2) 撥音 /N/ ならば $r/2$ 倍，3) それ以外の音素ならば変化させない，といったルールで時間長制御を行う．これは，発声速度による音素の伸縮は，子音よりも母音・撥音の方が大きいという知見による [7, 8]．係数 r は，提案法で合成された文音声の時間長と同じになるように，文章ごとに設定した．

評価データには，数量化に用いなかった 24 文章を用い，提案法とルールベース法での自然性に関する対比較実験を行った．各発話速度の合成音声について，被験者一人あたり 3 文章の比較実験を行う．この 3 文章は評価データ 24 文章中から無作為に選択した．被験者は 15 名であり，各発話速度について 2 手法間の比較実験が 45 文章，90 合成音声で行われている．

なお，無音区間については両手法とも同じ制御を行っており，普通の発話速度で収録したデータから強制切り出しによって求めたポーズ長を，発話速度に併せて伸縮して用いた．

各発話速度における対比較実験のプリファレンススコアを図 1 に示す．全ての発話速度で提案法がルールベース法より高いスコアとなっていることがわかる．全発話速度の結果をまとめると，提案法のプリファレンススコアは 65% となる．この結果について，2 つの手法間に性能の優劣が見いだせるかどうかを，二項分布に基づいて有意水準 5% で検定したところ，提案法の有意性が確認され，提案法が有効であることが示された．

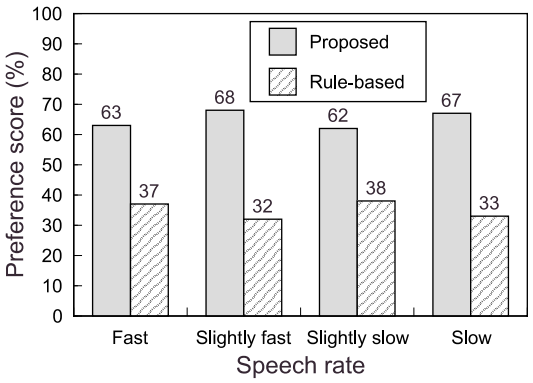

図 1. 様々な発話速度による提案法とルールベース法による合成音声の対比較実験結果

5　まとめ

HMM 分析合成系を用いた，様々な発話速度での音声を合成する手法を提案した．提案法は,「速い・普通・遅い」の 3 段階の発話速度の音声データから，数量化 I 類を用いて音素継続時間長の制御モデルを作成し，そのモデル間を補間することで，任意の時間伸縮率での継続時間長モデルを作成するものである．

提案法と，ルールベースの制御方式で合成した音声とを複数被験者による対比較実験を行った結果，提案法の有効性が確認された．

今後の課題としては，さらなる自然性向上のために，発話速度に応じたスペクトル特徴や F_0 パターンの制御があげられる．

参 考 文 献

[1] K.Tokuda, T.Kobayashi, and S.Imai, “Speech parameter generation from HMM using dynamic features,” Proc. ICASSP 95, vol.1, pp.660–663 (1995-5).

[2] T.Masuko, K.Tokuda, T.Kobayashi, and S.Imai, “Speech synthesis using HMMs with dynamic features,” Proc. ICASSP 96, vol.1, pp.389–392 (1996-5).

[3] C.Hayashi, “On the prediction of phenomena from qualitative data and the quantification of qualitative data from the mathematico-statistical point of view,” Ann. Inst. Statist. Math., vol.3, no.2, pp.69–98 (1952).

[4] 海木延佳，匂坂芳典，“文音声における子音継続長の設定，” 秋季音講論，vol.1, pp.259–260 (1990-9).

[5] 海木延佳，武田一哉，匂坂芳典，“言語情報を利用した母音継続時間長の制御，” 信学論，vol.J75-A, no.3, pp.467–473 (1992-3).

[6] 阿部匡伸，匂坂芳典，梅田哲夫，桑原尚夫，“研究用日本語データベース利用解説書（連続音声データ編）,” TR-I-0166, ATR 自動翻訳電話研究所 (1990-9).

[7] 舛田剛志，戸田智基，川波弘道，猿渡　洋，鹿野清宏，“発話速度の異なるデータベースを用いた音声合成手法の検討，” 信学技報，vol.101, no.603, pp.61–68 (2002-1).

[8] 古井貞熙，音声情報処理，森北出版株式会社，東京 (1998).