論文 / 著書情報
Article / Book Information

論題(和文)	平均声に基づく音声合成における合成音声の品質評価
Title(English)	Objective evaluation of synthesized speech in average-voice-based speech synthesis
著者(和文)	大川高志, 緒方克海, 山岸順一, 小林隆夫
Authors(English)	Takashi Okawa, Katsumi Ogata, Junichi Yamagishi, Takao Kobayashi
出典(和文)	日本音響学会2007年春季研究発表会講演論文集, Vol. , No. , pp. 193-194
Citation(English)	Proceedings of the ASJ 2007 Spring Meeting, Vol. , No. , pp. 193-194
発行日 / Pub. date	2007, 3

平均声に基づく音声合成における合成音声の品質評価*

◎大川高志, 緒方克海, 山岸順一, 小林隆夫 (東工大)

1 はじめに

我々は，所望の話者性を持った音声を容易に合成可能な音声合成システムの実現を目指し，平均声と話者適応に基づく音声合成方式を提案した [1]．この方式では，平均声モデルの学習データ量を増加させることで合成音声の品質が改善され，目標話者の音声のみで学習を行う特定話者 (SD) 方式と同等かそれ以上の品質を得られることが示されている [2, 3]．本論文では，平均声方式における学習データ増加に伴う合成音声の品質向上の要因を SD 方式との比較を通してより詳細に調べた結果を報告する．

2 平均声方式

平均声と話者適応に基づく音声合成方式 [1] は，複数の話者の平均的な音響的特徴を HMM によりモデル化した平均声モデルに対し，目標話者の音声データをもとに話者適応を行うことで，平均声モデルを目標話者モデルへと変換し，目標話者の音声に近い合成音声を生成する手法である．

この手法において，目標話者の学習データ量が一定の場合には，平均声の学習データ量を増加させることにより，合成音声の品質が向上する [3]．その際，平均声の学習文章数の増加に伴って，平均声モデルのパラメータ共有に利用される決定木のリーフノード数が増加していく．この事実より，モデルパラメータの共有構造が，合成音声の品質に影響を与える一因になっていると考えられる．

3 SD 方式におけるモデルサイズの調整

SD 方式においても，学習に用いる話者のデータ量は一定のままで，リーフノード数を増加させることが可能 [4] である．

決定木のリーフノードに HMM の分布が 1 つ対応付けられているとする．λ を決定木のリーフノードの集合 $\{S_1, S_2, \ldots, S_M\}$ で定義されるモデルとすると，λ の記述長 $\mathcal{D}(\lambda)$ は

$$\mathcal{D}(\lambda) \equiv -\mathcal{L}(\lambda) + cLM \log W + C \qquad (1)$$

と表せる．ここで $\mathcal{L}(\lambda)$ はモデル λ の対数尤度を表し，M はリーフノード数を，L は観測ベクトルの次元を表す．C はモデル選択に関する項であるが，ここでは定数項であると仮定する．また，$W = \sum_{m=1}^{M} \Gamma_m$ であり，Γ_m は学習データ中にノード S_m が出現する頻度である．$LM \log W$ はペナルティ項であり，c は構築する決定木のサイズを調整するための重み係数を表し，c が大きくなるにつれて木のサイズが小さくなる．なお，$c = 1$ のとき最小記述長 (MDL) 基準 [5] となる．

4 実験

本研究では平均声方式における合成音声の品質向上が，単にリーフノード数の増加によるものであるのか，平均声モデルの学習データ量の増加に伴うコンテキスト情報の増加によるものかを調べている．そのために，文献 [3] の目標話者に加えて，他の話者についても SD 方式において決定木の大きさを決めるパラメータを変化させ，パラメータ共有構造を変化させた場合について，SD 方式と平均声方式による合成音声の品質を客観評価により比較検討する．

Table 1　平均声の学習文章数とリーフノード数の関係

平均声の学習文章数	450	1350	2250	4050
コンテキスト数	26445	59932	85321	119646
メルケプストラム部	429	999	1464	2247
対数基本周波数部	891	2164	3228	5227

4.1 実験条件

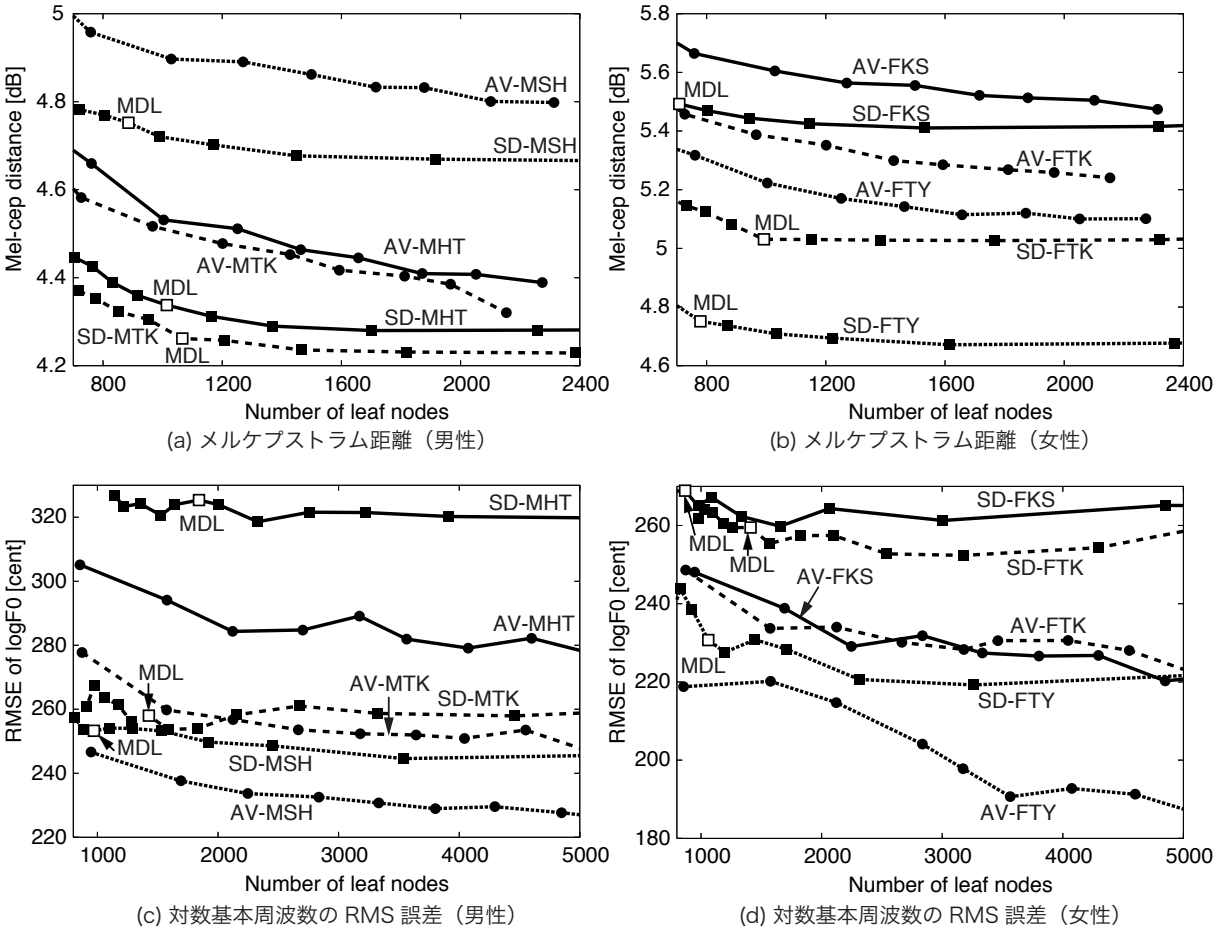
本研究では，隠れセミマルコフモデル (HSMM) に基づく音声合成システム [6] を用いて実験を行った．学習データには，ATR 日本語音声データベースセット B に含まれている男性話者 6 名と女性話者 4 名，および文献 [7] で用いた女性話者 FTY をあわせた男女 11 名を用いた．目標話者には文献 [3] で用いた話者セットを含む，(FTK,MTK),(FTY,MHT),(FKS,MSH) の男女各 1 名の組合せを 3 通り選び，残りの話者 9 名を平均声モデルの学習話者とした．平均声の学習では，学習データの偏りを軽減するため，各学習話者ごとに 50 文章ずつのサブセットを重複しないように，かつ話者ごとに異なるように 450 文章まで増加させた．平均声モデルのコンテキストクラスタリング手法には STC[8] を用い，分割停止基準には MDL 基準を用いた．また，平均声モデルの学習には SAT を利用した手法 [1] を用いた．このときの学習文章数と平均声モデルのリーフノード数との関係を Table 1 に示す．ここでは目標話者の組合せごとの 3 種類の平均声モデルのリーフノード数の平均値を示している．話者適応アルゴリズムとしては，文献 [3] と同様に，SMAPLR に MAP 推定を組合せた手法を採用し，適応データには目標話者の 450 文章を用いた．

一方 SD 方式の学習話者には平均声方式の目標話者と同じ 6 名を選び，各話者 450 文章でモデルを作成した．評価データには学習及び適応データに含まれない 50 文章を用いた．

4.2 客観評価・考察

SD 方式におけるクラスタリングの分割停止基準を変化させることにより，決定木のサイズを変化させた場合と，平均声において学習データを増加させてパラメータ共有構造を変化させた場合の，各モデルから生成された合成音声と目標話者の分析合成音とのメルケプストラム距離および対数基本周波数の RMS 誤差をまとめたものを Fig. 1 に示す．SD-(目標話者) は SD 方式による結果を，AV-(目標話者) は平均声と話者適応に基づく方式による結果を表している．また図中の MDL は決定木のサイズが MDL 基準の場合の結果である．

*Objective evaluation of synthesized speech in average-voice-based speech synthesis , by OKAWA, Takashi, OGATA, Katsumi, YAMAGISHI, Junichi, and KOBAYASHI, Takao (Tokyo Institute of Technology)



(a) メルケプストラム距離（男性）　(b) メルケプストラム距離（女性）

(c) 対数基本周波数の RMS 誤差（男性）　(d) 対数基本周波数の RMS 誤差（女性）

Fig. 1　客観評価実験結果

平均声方式の学習文章数を増やすことによる合成音声の客観品質の改善は，メルケプストラム距離では 0.1dB〜0.3dB 程度，対数基本周波数の RMS 誤差では 20〜30cent 程度であることが確認できる．一方，SD 方式において分割停止基準の閾値を変化させ，リーフノード数を増加させた場合，メルケプストラム距離で 0.1dB〜0.2dB 程度，対数基本周波数の RMS 誤差では 10〜20cent 程度の改善がみられるが，どちらも平均声の学習文章数を増やすことによる改善には及ばないことがわかる．さらに，SD 方式ではリーフノード数を増やしていくと逆に誤差が大きくなっていく場合もある．これは，データ量が一定の条件下でリーフノード数を増加させているため，それぞれのリーフノードで共有される分布パラメータの推定に利用できる学習データが不足し，適切に推定が行えなくなったためと考えられる．

これに対し，平均声方式では学習文章数を増加させることで誤差が減少する傾向にあり，対数基本周波数の RMS 誤差においては，すべての目標話者において SD 方式を上回る性能を得られることを確認できた．また，メルケプストラム距離においても，さらに学習文章数を増加させることで SD 方式を上回る可能性も考えられる．

以上のことから，HMM 音声合成におけるパラメータ共有構造の構築には，単にリーフノード数を増加させるだけでなく，大量の学習データを用いて豊富なコンテキスト情報を木構造に反映させることが重要であると考えられる．

5　おわりに

本研究では，SD 方式においてモデルサイズを変化させた場合との比較を通して，平均声モデルにおける学習データの増加は，単にリーフノード数を増やすだけでなく，学習データに含まれる豊富なコンテキスト情報をパラメータ共有構造に反映することで，合成音声の品質の改善につながることを示した．今後の課題は平均声方式における合成音声のさらなる品質の改善が挙げられる．

参考文献

[1] J. Yamagishi *et al.*, “A training method of average voice model for HMM-based speech synthesis,” IEICE Trans. Fundamentals, vol. E86-A (8), 1956-1963, 2003.

[2] 大川他, “線形変換と MAP に基づく音響モデル学習法の評価,” 音講論（秋）, 266-267, 2006.

[3] 緒方他, “平均声に基づく音声合成における線形変換と MAP に基づく音響モデル学習法,” 信学技報, SP2006-84, 2006.

[4] 野村他, “HMM 音声合成における決定木の分割停止基準の検討,” 音講論（春）, 291-292, 2005.

[5] K. Shinoda and T. Watanabe, “MDL-based context dependent subword modeling for speech recognition,” J. Acoust. Soc. Jpn. (E), vol. 21, 79-86, 2000.

[6] H. Zen *et al.*, “Hidden semi-Markov model based speech synthesis,” Proc. INTERSPEECH 2004-ICSLP, vol. 2, 1397-1400, 2004.

[7] J. Yamagishi *et al.*, “Acoustic modeling of speaking styles and emotional expression in HMM-based speech synthesis” IEICE Trans. Inf. &Syst.,E88-D (3), 2002

[8] J. Yamagishi *et al.*, “A context clustering technique for average voice models,” IEICE Trans. Inf. &Syst.,E86-D (3), 534-542, 2003