様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成 ２７ 年　 ５ 月 ２６ 日現在

機関番号：１１３０１
研究種目：若手研究(B)
研究期間：2013～2014
課題番号：２５７３０１０６

研究課題名（和文）多様で肉声感の高い音声生成のための素片正規化に基づくハイブリッド音声合成の研究

研究課題名（英文）A study of speech synthesis for achieving synthetic speech with high quality and variability based on hybrid approach

研究代表者
能勢　隆（Nose, Takashi）

東北大学・工学（系）研究科（研究院）・講師

研究者番号：９０５５０５９１

交付決定額（研究期間全体）：（直接経費）　　3,200,000円

研究成果の概要（和文）：本研究は，任意の話者の限られた音声データのみで「人間に近い肉声感」をもち，かつ様々な感情や発話様式を表現可能なハイブリッド型の音声合成方式を確立することを目的として研究を行い，以下の6つの項目について成果が得られた．(1)非言語情報やパラ言語情報を柔軟に再現・制御可能とする，(2)韻律の多様性を自動学習する，(3)多言語の音声合成への拡張を行う，(4)音声だけでなく歌声への応用についても検討する，(5)このような音声コーパスを効率的に構築する方法を確立する，(6)従来のパラメータ生成法を改善し主観品質を向上する，

研究成果の概要（英文）：The purpose of this research is to establish hybrid speech synthesis framework that can synthesize human-like speech with various emotional expressions and/or speaking styles using only a limited amount of speech data. We achieved the following six issues in this research. (1) Flexible control of non- or para-linguistic information appearing in synthetic speech. (2) Automatic training of prosodic variations, (3)Expansion to the multi-lingual or cross-lingual speech synthesis, (4)Application to singing voice synthesis, (5) Efficient designing of speech corpus for synthesis, and (6) Improving subjective quality of synthetic speech by modifying the conventional parameter generation method .

研究分野：音声情報処理

キーワード：統計的音声合成　非言語情報　パラ言語情報　韻律　多言語　歌声合成　パラメータ生成

１．研究開始当初の背景

任意の文章から音声を生成する「テキスト音声合成技術」は，目の不自由な人々や高齢者などの生活支援は当然のこと，ヒューマノイドロボットの実現，各種ナビゲーションシステム，音声対話システムなどにおける最重要基盤技術の一つである．さらに最近では，スマートフォンなどの携帯機器やゲーム機などにも音声合成機能が搭載されるなど今後も需要は急速に拡大することが予想される．しかしその一方で，現在実用化されているほとんどのテキスト音声合成は読み上げ調の音声を出力することしかできず，今後の「高品質で，かつ人間に優しい音声インターフェース」の普及に十分な機能をもっているとはいえない．

従来，感情などを含む多様な音声の合成は，声の高さやリズムなどを規則により変化させる規則合成方式や，大量の音声を使用する素片選択方式が主流であったが，品質やコストの面で問題があった．これに対し，近年は研究代表者らが提案している統計モデルに基づく多様な音声合成がその柔軟性，コストパフォーマンスの面から大きな注目を集めている．事実，音声情報処理分野のトップカンファレンスにおいて関連セッションが増加しており，学術雑誌の特集号の出版，様々なコミュニケーションロボットやインテリジェントシステムの報道発表などからも関心の高さが伺える．

２．研究の目的

本研究は，研究代表者らがこれまでに提案・確立した「統計モデルに基づく表情豊かな音声合成」の枠組を応用し任意の話者の限られた音声データのみで「人間に近い肉声感」をもち，かつ様々な感情や発話様式を表現可能なハイブリッド型の音声合成方式を確立することを目的として研究を行い，音声インターフェース基盤技術の普及と向上に資する．

具体的には，本研究では従来の言語情報のみを用いた音声性の枠組みを拡張し，

(1)非言語情報やパラ言語情報を柔軟に再現・制御可能とする
(2)韻律の多様性を自動学習する
(3)多言語の音声合成への拡張を行う
(4)音声だけでなく歌声への応用についても検討する
(5)このような音声コーパスを効率的に構築する方法を確立する
(6)従来のパラメータ生成法を改善し主観品質を向上する

ことを目的とする．

３．研究の方法

上記の(1)については従来のスタイル音声合成において，ハイブリッドな枠組みとしてモデル適応の枠組みを導入し，読み上げ音声のモデルから目標スタイルの音声のモデルに変換する不特定話者スタイル変換法を提案する．

上記の(2)については，声の高さを表す基本周波数(F0)が，強調表現などの韻律の大きな変化を伴う際に，従来の生成 F0 との差分が大きくなることに着目し，差分 F0 を量子化して学習の際のコンテキストとして利用する手法を提案する．提案法の概要を図 1 に示す．

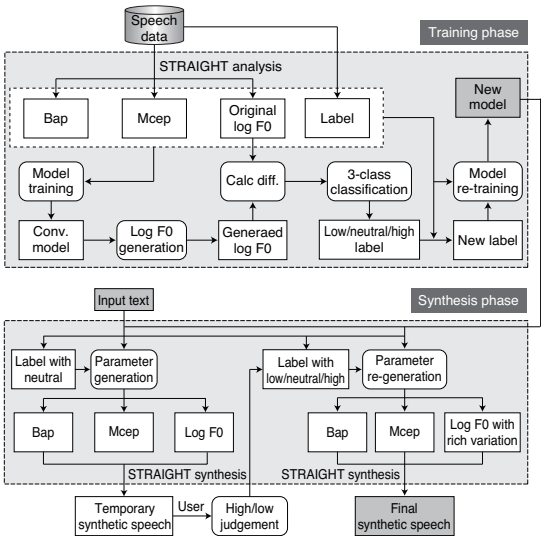

図 1 韻律の多様性を自動学習・再現するシステム

上記の(3)については，目標話者の母国語の音声のみを用いてその話者の外国語の音声を合成することができるクロスリンガル話者適応法を提案する．この手法ではあらかじめ他の複数の話者の両言語の音声で言語依存の平均声モデルを学習しておき，母国語のモデルと目標話者の音声を用いて話者適応の枠組みで変換行列を求め，これを目標言語のモデルに適応する手法として，共有決定木コンテキストクラスタリングに基づく 2 段階クラスタリングを用いた新たな手法を提案した．提案法における 2 段階クラスタリングの仕組みを図 2 に示す．

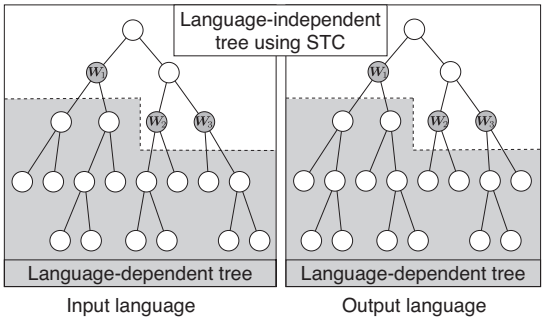

図 2 2 段階クラスタリングを用いた状態マッピング

上記の(4)については，従来のスタイル制御法を歌声にも応用するために，F0 のモデル化として音高正規化学習に基づく重回帰隠れマルコフモデルを新たに提案する．また，ビブラートが顕著でない場合についてもモデル化可能なセグメント単位でのモデル化手法を提案する．

上記の(5)については，複数の音韻・韻律に関するコンテキストのバランスを最大化するエントロピーに基づく文選択手法を提案する．

上記の(6)については，従来の音響モデルの尤度を最大化する際に系列内分散を最適化する高速でハイブリッドなアルゴリズムを提案する．

４．研究成果

上記の(1)については，数名の話者の複数のスタイルについて，合成音声の再現性と自然性について評価実験を行った．結果として従来の韻律を単純な規則で変化させる場合よりも，若干自然性は劣るが，スタイルによってはその再現性が向上することを示した．

上記の(2)については差分 F0 の平均をアクセント句毎に粗く量子化して学習時のコンテキストラベルとして使用することで，従来困難であった強調表現の自動学習が適切に行われることを客観および主観評価実験により確認した．

上記の(3)については，従来の状態マッピングに基づく手法に比べて，提案法である共有決定木コンテキストクラスタリングを利用した手法のほうが客観および主観評価において上回ることを示した．

上記の(4)については，音高正規化学習によりスタイルの制御を実現しつつ，F0 のモデル化を高精度で行えることを示し，特に学習データにない音高を再現する際に，その効果が顕著であることを示した．また，ビブラートについても，従来のビブラート表現が顕著であることを前提としたものではなく，表現が曖昧であるような場合についても，歌声独特の揺らぎを再現し，自然性が改善することを示した．

上記の(5)については，提案法により文選択を行うことで，ランダムに選んだ場合に比べ，客観および主観評価のスコアが向上し，ランダムに選ぶ場合に生じる潜在的リスクを回避できることを示し，既存のデータベースではなく，話し言葉などの多様なスタイルや口調を含む新しくデータベースの構築の際に有効であることを示した．

上記の(6)については，パラメータのアフィン変換により，系列内分散の誤差を学習時に最小化するような変換を求め，それを生成時に利用することにより，従来の系列内変動をパラメータ生成時に考慮する手法に比べ，高速でかつ歪の少ないスペクトルパラメータを生成できることを示した．

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 9 件）

[1] Takashi Nose, Misa Kanemoto, Tomoki Koriyama, Takao Kobayashi, “HMM-based expressive singing voice synthesis with singing style control and robust pitch modeling,” Computer Speech and Language, 査読有り，2015 年 (in press).
DOI:10.1016/j.csl.2015.04.001

[2] Takashi Nose, Vataya Chunwijitra, Takao Kobayashi, “A parameter generation algorithm using local variance for HMM-based speech synthesis,” IEEE Journal of Selected Topics in Signal Processing, vol.8, no.2, pp.221-228, 査読有り，2014 年.
DOI:10.1109/JSTSP.2013.2283459

[3] Yu Maeno, Takashi Nose, Takao Kobayashi, Tomoki Koriyama, Yusuke Ijima, Hideharu Nakajima, Hideyuki Mizuno, Osamu Yoshioka, “Prosodic variation enhancement using unsupervised context labeling for HMM-based expressive speech synthesis,” Speech Communication, vol.57, pp.144-154, 査読有り，2014 年.
DOI:10.116/j.specom2013.09.014

[4] Takashi Nose, Takao Kobayashi, “Quantized F0 context and its applications to speech synthesis, speech coding and voice conversion,” Proceedings of Tenth International Conference on Intelligent Information Hiding and Multimedia Signal Processing, IIH-MSP 2014, pp.578-581, 査読有り，2014 年.
DOI: 10.1109/IIH-MSP.2014.149

[5] Daiki Nagahama, Takashi Nose, Tomoki Koriyama, Takao Kobayashi, “Transform mapping using shared decision tree context clustering for HMM-based cross-lingual speech synthesis,” Proc. 15th Annual Conference of the International Speech Communication Association, INTERSPEECH 2014, pp.770-774, 査読有り，2014 年.
URL:http://www.isca-speech.org/archive/interspeech_2014/i14_0770.html

[6] 長濱大樹,能勢 隆,郡山知樹,小林隆夫,“共有決定木を利用した話者適応に基づくクロスリンガル音声合成の評価,” 日本音響学会 2014 年春季研究発表会講演論文集, pp. 413-414, 査読無し，2014 年.
DOI/URL:なし

[7] 荒生侑介, 能勢 隆, 小林隆夫, “複数ドメインコーパスからの文選択に基づくキャラクター音声合成の検討,” 日本音響学会 2013 年秋季研究発表会講演論文集, pp. 351-352, 査読無し, 2013 年.
DOI/URL:なし

〔学会発表〕(計 8 件)
[1] Takashi Nose, “Analysis of spectral enhancement using global variance in HMM-based speech synthesis,” 15th Annual Conference of the International Speech Communication Association, INTERSPEECH 2014, Singapore, Singapore, 2014 年 9 月 18 日.

[2] Daiki Nagahama, “Transform mapping using shared decision tree context clustering for HMM-based cross-lingual speech synthesis,” 15th Annual Conference of the International Speech Communication Association, INTERSPEECH 2014, Singapore, Singapore, 2014 年 9 月 15 日.

[3] Takashi Nose, “Quantized F0 context and its applications to speech synthesis, speech coding and voice conversion,” Tenth International Conference on Intelligent Information Hiding and Multimedia Signal Processing, IIH-MSP 2014, Kitakyushu International Conference Center, Kitakyushu, Japan, 2014 年 8 月 28 日.

[4] 荒生侑介, “音声合成のための音韻・韻律コンテキストを考慮した文選択アルゴリズムの評価,” 日本音響学会 2014 年春季研究発表会, 日本大学, 東京都, 2014 年 3 月 10 日.

[5] 長濱大樹“共有決定木を利用した話者適応に基づくクロスリンガル音声合成の評価,” 日本音響学会 2014 年春季研究発表会, 日本大学, 東京都, 2014 年 3 月 10 日.

[6] 荒生侑介, “複数ドメインコーパスからの文選択に基づくキャラクター音声合成の検討,” 日本音響学会 2013 年秋季研究発表会, 豊橋技術科学大学, 豊橋市, 2013 年 9 月 25 日.

[7] Takashi Nose, ”A style control technique for singing voice synthesis based on multiple-regression HSMM,” 14th Annual Conference of the International Speech Communication Association, INTERSPEECH 2013, Lyon, France, 2013 年 8 月 26 日.

６．研究組織
(1)研究代表者
能勢　隆 (NOSE, TAKASHI)
東北大学・大学院工学研究科・講師
研究者番号：９０５５０５９１