様式 C-19

科学研究費補助金研究成果報告書

平成２３年　６月　１日現在

機関番号：１２６０８
研究種目：研究活動スタート支援
研究期間：２００９～２０１０
課題番号：２１８０００２０
研究課題名（和文）ヒューマノイド音声対話システムのための話し言葉音声合成に関する研究

研究課題名（英文）Study on speech synthesis for humanoid spoken dialog system

研究代表者
能勢　隆（NOSE TAKASHI）
東京工業大学・大学院総合理工学研究科・助教
研究者番号：９０５５０５９１

研究成果の概要（和文）：ヒューマノイド音声対話システムの実現に向けた話し言葉音声合成のための基盤技術として、(1)統計モデルに基づく話し言葉音声合成の検討、(2)統計モデルに基づく不特定話者声質変換の検討、(3)音声合成における音韻・韻律コンテキストの詳細な評価、を行った。

研究成果の概要（英文）：Two novel techniques and an investigation were presented that is key technologies of speech synthesis for the development of humanoid spoken dialog system as follows. (1) Spontaneous speech synthesis based on statistical parametric modeling (2) Speaker-independent voice conversion based on statistical parametric modeling. (3) Investigation of phonetic and prosodic contextual factors in speech synthesis.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
２００９年度	1,060,000	318,000	1,378,000
２０１０年度	970,000	291,000	1,261,000
年度			
年度			
年度			
総　計	2,030,000	609,000	2,639,000

研究分野：工学
科研費の分科・細目：知覚情報処理・知能ロボティクス
キーワード：テキスト音声合成、隠れマルコフモデル、話し言葉音声、話者適応、HMM 音声合成、ヒューマノイドロボット、音声対話システム、声質変換

１．研究開始当初の背景

(1) 音声合成は人間をサポートするロボットなどの実現のための重要な基盤要素技術の一つである。しかし、現在実用化されているシステムは読み上げ調の音声のみしか出力できないものが大半である．ユーザにとってより自然で違和感のない音声対話インターフェースを実現するためには，感情などの多様な表現を含んだ合成音声の生成が必須である。

(2) このような背景を受け、研究代表者らはこれまでに，感情や発話様式を柔軟に制御可能な統計モデルに基づいた音声合成手法を提案し，プロのナレータによる模擬感情音声を対象としてその有効性の高さを示してきた。

２．研究の目的

本研究では、新たにより自発性の高い話し言葉音声についても対象とし、より多様な合成音声の生成を目的とし、以下の３点について検討を行う。

(a) 話し言葉音声の自然性の向上：話し言葉音声は従来整備されてきたナレータの音声に比べて統計モデルの学習・合成音声の生成に必要なラベル情報の付与が難しく特定話者の大量の音声データを準備することが容易ではない。このような場合に少量の学習データのみで自然性の高い手法を検討する。

(b) 話者性の多様化：従来の波形接続型の合成方式では大量の音声データを用いるため明瞭で自然な音声を生成できるが、一方で話者の数を増やすことが困難である。この問題に声質変換のアプローチで取り組む。

(c) 韻律コンテキストの分析：統計モデルに基づく合成方式ではアクセントや形態素などの様々な韻律情報をコンテキストとして考慮する。これらが合成音声の品質に寄与しているかを調べ、より適切なコンテキストを検討する。

３．研究の方法

(a) 話し言葉音声の自然性の向上：限られた音声データのみから自然性の高い合成音声を生成するために、平均声と話者適応に基づく手法を導入する。この手法ではあらかじめ整備されているプロのナレータによる大量の音声データにより学習した統計モデルを事前知識とし、それに対して話者適応によるパラメータ変換を行うことで目標話者の音声データがごく少量の場合でも自然性の高い合成音声を得ることができる。

(b) 話者性の多様化：目標話者の少量の音声データのみを用いて精度良く声質変換を行うために、隠れマルコフモデル（HMM）に基づく手法について検討を行う。HMM では音声のスペクトル特徴だけでなく、時間的な変化も同時にモデル化を行うことができ、従来の混合ガウス分布に基づく手法に比べ、より話者性の変換性能が向上することが期待できる。

(c) 韻律コンテキストの分析：具体的な韻律コンテキストとして形態素、アクセント、呼気段落情報、文長のそれぞれに対し、モデルの学習に用いた場合とそうでない場合について客観および主観評価実験を行う。

４．研究成果

(a) 話し言葉音声の自然性の向上：整備されているプロのナレータ音声と一般話者による少量の話し言葉音声の両方を最大限に活用できる手法として二段階モデル適応を用いた手法を提案した。これによりごく限られた目標話者の音声のみでその話者の話し言葉音声を合成することができ、様々な話者の声を容易に作成することが可能となった。

(b) 話者性の多様化：元話者音声に対して不特定話者モデルを導入したことで、従来必須であった元話者音声を用いることなく、任意の話者への声質変換を実現した。また発話毎に適応的に量子化された基本周波数情報をモデルの学習および変換に利用することで、声の高さおよびその変化に現れる個人性についても適切に変換を行えるようになった。

(c) 韻律コンテキストの分析：客観評価により考えうるすべての要因の組み合わせをそれぞれコンテキストとしてモデルを学習し、各要因の寄与度を調べた。その結果に基づき従来に比べてよりコンパクトなコンテキストセットを提案し、それを用いた場合でも従来と同程度の品質の合成音声が生成可能であることを主観評価により示した。

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１２件）

(1) Takashi Nose, Yuhei Ota, Takao Kobayashi, “HMM-based voice conversion using quantized F0 context,” IEICE Trans. on Information and Systems, vol.E93-D, 9, pp.2483-2490, 2010, 査読有り

(2) Shuji Yokomizo, Takashi Nose, Takao Kobayashi, “Evaluation of prosodic contextual factors for HMM-based speech synthesis,” Proc. 11th Annual Conference of the International Speech Communication Association, INTERSPEECH 2010, pp.430-433, 2010, 査読有り

(3) Tomoki Koriyama, Takashi Nose, Takao Kobayashi, “Conversational spontaneous speech synthesis using average voice model,” Proc. 11th Annual Conference of the International Speech Communication Association, INTERSPEECH 2010, pp.853-856, 2010, 査読有り

(4) Takashi Nose, Takao Kobayashi, “Speaker-independent HMM-based voice conversion using quantized fundamental frequency,” Proc. 11th Annual Conference of the International Speech Communication Association, INTERSPEECH 2010, pp.1724-1727, 2010, 査読有り

(5) Takashi Nose, Takao Kobayashi, “HMM-based robust voice conversion using

adaptive F0 quantization," Proc. 7th ISCA Workshop on Speech Synthesis, SSW7-2010, pp.80-85, 2010, 査読有り

(6) Yusuke Ijima, Takashi Nose, Makoto Tachibana, Takao Kobayashi, "A rapid model adaptation technique for emotional speech recognition with style estimation based on multiple-regression HMM," IEICE Trans. on Information and Systems, vol.E93-D, 1, pp.107-115, 2010, 査読有り

(7) Takashi Nose, Takao Kobayashi, "A technique for estimating intensity of emotional expressions and speaking styles in speech based on multiple-regression HSMM," IEICE Trans. on Information and Systems, vol.E93-D, 1, pp.116-124 , 2010, 査読有り

(8) Takashi Nose, Koujirou Ooki, Takao Kobayashi, "HMM-based speech synthesis with unsupervised labeling of accentual context based on F0 quantization and average voice model," Proc. 2010 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2010, pp.4622-4625, 2009, 査読有り

(9) Yusuke Ijima, Takeshi Matsubara, Takashi Nose, Takao Kobayashi, "Speaking style adaptation for spontaneous speech recognition using multiple-regression HMM," Proc. 10th Annual Conference of the International Speech Communication Association, INTERSPEECH 2009, pp.552-555, 2009, 査読有り

(10) Takashi Nose, Junichi Asada, Takao Kobayashi, "HMM-based speaker characteristics emphasis using average voice model," Proc. 10th Annual Conference of the International Speech Communication Association, INTERSPEECH 2009, pp.2631-2634, 2009, 査読有り

(11) Junichi Yamagishi, Takashi Nose, Heiga Zen, Zhenhua Ling, Tomoki Toda, Keiichi Tokuda, Simon King, Steve Renals, "A robust speaker-adaptive HMM-based text-to-speech synthesis," IEEE Trans. on Audio, Speech, and Language Processing, vol.17, 6, pp.1208-1230, 2009, 査読有り

(12) Yusuke Ijima, Makoto Tachibana, Takashi Nose, Takao Kobayashi, "Emotional speech recognition based on style estimation and adaptation with multiple-regression HMM," Proc. 2009 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2009, pp.4157-4160, 2009, 査読有り

〔学会発表〕(計２７件)

(1) Takashi Nose, Takao Kobayashi, "Speaker-independent HMM-based voice conversion using quantized fundamental frequency," Proc. 11th Annual Conference of the International Speech Communication Association, INTERSPEECH 2010, 2010/9/29, Makuhari, Japan.

(2) Tomoki Koriyama, Takashi Nose, Takao Kobayashi, "Conversational spontaneous speech synthesis using average voice model," Proc. 11th Annual Conference of the International Speech Communication Association, INTERSPEECH 2010, 2010/9/28, Makuhari, Japan.

(3) Shuji Yokomizo, Takashi Nose, Takao Kobayashi, "Evaluation of prosodic contextual factors for HMM-based speech synthesis," Proc. 11th Annual Conference of the International Speech Communication Association, INTERSPEECH 2010, 2010/9/27, Makuhari, Japan.

(4) Takashi Nose, Takao Kobayashi, ``HMM-based robust voice conversion using adaptive F0 quantization," Proc. 7th ISCA Workshop on Speech Synthesis, SSW7-2010, 2010/9/22, Kyoto, Japan.

(5) Takashi Nose, Koujirou Ooki, Takao Kobayashi, "HMM-based speech synthesis with unsupervised labeling of accentual context based on F0 quantization and average voice model," 2010 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2010, 2010/3/17, Dallas, USA.

(6) Yusuke Ijima, Takeshi Matsubara, Takashi Nose, Takao Kobayashi, "Speaking style adaptation for spontaneous speech recognition using multiple-regression HMM," Proc. 10th Annual Conference of the International Speech Communication Association, INTERSPEECH 2009, 2009/9/7, Brighton, U.K.

(7) Takashi Nose, Junichi Asada, Takao

Kobayashi, “HMM-based speaker characteristics emphasis using average voice model,” Proc. 10th Annual Conference of the International Speech Communication Association, INTERSPEECH 2009, 2009/9/10, Brighton, U.K.

(8) Yusuke Ijima, Makoto Tachibana, Takashi Nose, Takao Kobayashi, “Emotional speech recognition based on style estimation and adaptation with multiple-regression HMM,” Proc. 2009 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2009, pp.4157-4160, 2009/4/21, Taipei, Taiwan.

６．研究組織

(1)研究代表者

能勢　隆（NOSE　TAKASHI）

東京工業大学・大学院総合理工学研究科・助教

研究者番号：９０５５０５９１