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　 ６ 月　 １ 日現在

機関番号：３４３１５
研究種目：基盤研究(C)
研究期間：2012～2014
課題番号：２４５００２２３

研究課題名（和文）話者性再現度の自動評価に基づいた個性豊かな音声合成に関する研究

研究課題名（英文）A Study on Speech Synthesis with Rich Personality Based on Automatic Scoring of Reproduction of Speaker Identity

研究代表者
山下　洋一（YAMASHITA, Yoichi）

立命館大学・情報理工学部・教授

研究者番号：８０１７４６８９

交付決定額（研究期間全体）：（直接経費）　　3,700,000円

研究成果の概要（和文）：個性豊かな音声合成を実現するために，音声における個人性や多様性に関する研究を行った。
音声における声質の違いを物理パラメータに基づいて自動的に予測する手法を提案した。音声の特徴量としてMFCCパラメータを用い，その重み付きユークリッド距離で声質類似度を算出する。音声の音韻性の変化を取り除き，イントネーションなどの韻律情報を保存した合成音声を用いた聴取実験を行い，韻律情報と個人性知覚の関係を分析した。方言音声，アニメの声優の音声，アナウンサーの音声，感情を含む音声など，様々な音声を分析し，音声の物理パラメータとの関係を分析した。

研究成果の概要（英文）：This research addresses measurement of personality and analysis of diversity in speech aiming at realizing speech synthesis with rich personalization.
I proposed a new method for measuring the difference of voice quality based on feature parameters of speech. The similarity of voice quality is calculated by weighted Euclidean distance of MFCC parameters which represent spectrum features of speech. I analyzed the relationship between prosodic information and personality perception using synthetic speech in which phonemic information is removed but prosodic information, such as intonation, is preserved. I also analyzed various types of speech which include dialect, character voices in ‘Anime’, announcer voices, emotional voices, and so on.

研究分野：音声情報処理

キーワード：声質　個人性　多様性　韻律　音声合成　音声分析

１．研究開始当初の背景

誰でも簡単に使えるヒューマンインタフェースの実現を目指して，音声を用いた対話システムの研究が広く行われている。計算機の画面上に特定の人の顔を生成し，その人と音声で対話を行う擬人化音声対話エージェントの研究も進められている。このような対話システムにおける音声合成では，画面に表示されたその人の声で合成されることが好ましい。また，音声合成システムがあらかじめ用意している複数の話者から一人の話者を単に選択して利用するだけでなく，「自分の声で合成音を生成したい」など，多様な話者性での音声合成に対するニーズもある。一人の声を多数収録し，「その人の声でしゃべる合成音声合成システム」を構築する Polluxstar のような商用サービスも始まっているが，必ずしも安価なサービスとは言えないのが現状である。

本研究が対象とする HMM に基づいた音声合成 (HMM 音声合成) では，音声を合成するために必要な声の特徴（話者モデル）を「音素モデルのセット」としてあらかじめ学習しておく。一般には１名の話者のデータを学習音声データとして用いる。合成音において，どのくらい「その人らしい声」に聞こえるかという程度（話者性再現度）は，作成した話者モデルによって異なる。話者性再現度は，学習音声データとして用いる文の量や内容によって異なり，さらに，同じ量の同じ文章を用いて話者モデルを学習しても話者によって異なることが経験的に知られている。「その人らしい」声での音声合成をどの話者に対しても実現するには，高い話者性再現度の得られる話者モデル開発手法を確立することが重要である。

２．研究の目的

音声合成システムでは，ある話者（一般には１名）の発声した音声データを用いて声の特徴（話者モデル）をあらかじめ学習しておき，その話者の音声を合成する。合成音において「その人らしい声」に聞こえるかという程度（話者性再現度）が対象話者によって異なることが経験的に知られている。本研究では，話者性再現度を自動評価する手法を確立し，見込まれる話者性再現度が低い場合には，話者モデルの学習を工夫することにより個性豊かな音声合成を実現する手法を開発することを目的とした。また，多様な声質の音声合成を実現するために，様々な音声の特徴を分析することも目的とした。

３．研究の方法

数名の大学院生に研究協力者として本研究に参加してもらい，データ整備，聴取実験，ツール開発などを行ってもらいながら，研究代表者が全体の計画を立てながら研究を遂行していく。まず，多数の話者の音声データを収録し，声質およびイントネーションに関して話者の類似度を評価する聴取実験を行い，研究の基礎資料とする。聴取実験結果に基づき，話者性の違いを測る尺度の検討，話者性を表現する音声の特徴の分析を行う。様々な音声の分析を行うために，具体的には，感情を含む音声，アニメの声優の音声，アナウンサーの音声などを収録し，研究に用いる。

４．研究成果

(1) 音声の声質は，スペクトルにおける周波数成分の違いによって主に表現され，音声の個人性に大きく関わっている。個人性による音声の違いを定量的に算出するために，瞬時的なスペクトルによって表現される声質の違いに焦点をあて，声質の類似性を自動的に予測する手法について検討した。音声の特徴量として MFCC パラメータを用い，その重み付けユークリッド距離で声質類似度を予測する。まず，文音声を用いた声質類似度の予測を試みた。30 名の話者が発声した文音声を収録し，声質類似度を予測するモデル（重み付きユークリッド距離における重み係数）の学習と評価で話者が異なるように，30 名のうち 20 名の話者が発声した学習用データセット L と残りの 10 名が発声した評価用データセット T に分割した。L の発声内容は１文のみであり，T の発声内容は 2 文でうち１文は L の文と同じである。18 名の被験者がデータセットごとに総当りで声質類似性の判定を行った。予測結果と聴取実験結果との相関値が最大で 0.49 であった。聴取実験で得られた評価値の分散が大きいことから，文音声を用いた場合には，声質の類似性を安定して算出することは困難であった。次に，文音声に含まれる韻律や特徴量の時間変化等の様々な要因を取り除くた

め，孤立発声された単母音を用いて声質類似度の推定を試みた。20 名の話者が 140Hz のトーン信号を聴取しながら同じ声の高さで発声した単母音/a/ と/i/ を収録し，10 名ずつをモデルの学習と評価に用いた。話者の同一性を判定する聴取実験を行い，孤立発声された単母音を聴取することで話者間の声質類似度をある程度判定できることを確認した。単母音/a/および/i/を用いた声質類似度の予測において，聴取実験との相関値で，それぞれ 0.81, 0.78 を得た。

(2) スペクトルによって表現される声質だけでなく，韻律情報と呼ばれる声の高さ（基本周波数）・声の大きさ・話すスピードの時間変化によっても音声における個人性の違いがもたらされる。個性豊かな音声合成を実現するために，音声における韻律情報に関する個人性知覚の分析を行った。まず，音声の音韻性の変化を取り除き，韻律情報を保存した合成音声を作成して，この合成音声を聴取し個人を同定する聴取実験を行った。合成音声の聴取によって韻律情報から個人を同定することは困難であったが，よく知っている身近な話者については同定できることがわかった。また，アナウンサーのニュース読み上げ音声に関しては，多くの被験者がアナウンサーであることを同定できていた。次に，アナウンサーと一般人話者の合成音声を対比較し識別する聴取実験を行った。多くの被験者がアナウンサーと一般人話者を識別できており，アナウンサーのニュース読み上げ音声には，なんらかの特徴があることが示唆された。音声を分析した結果から，アナウンサーの韻律情報は一般人話者の読み上げ音声に比べて抑揚が大きくなっており，アナウンサー音声の特徴の一つであることがわかった。しかし，アナウンサーの音声を真似て一般人話者が発声した物真似音声に関しても，一部の話者においてはアナウンサーと同程度の抑揚の変化が見られるものの，アナウンサー音声と物真似音声の判別ができていることから，抑揚以外にもアナウンサーらしさを特徴づける特徴があると考えられる。さらに，聴取実験の結果からアナウンサーらしさを求め，抑揚と話速を変数としてアナウンサーらしさを予測する重回帰モデルを作成し，評価を行った。特異な傾向を示した 1 名の話者を除外してモデルを学習することによって，実験で得られたアナウンサーらしさと予測値の間で 0.36 の相関係数を得た。

(3) 合成音声の音質を改善する手法について検討した。HMM（隠れマルコフモデル）を用いたパラメータ音声合成手法においては，パラメータの時間変化が過度に平滑化され音声が劣化する問題がある。この問題を解決するために，パラメータの時間変化が自然音声における変化と近づくようにモデル化を行う新しいパラメータ生成手法を提案した。

(4) 音声の多様性を構成する要因として方言が挙げられ，方言を表現する重要な要因として，方言ごとのアクセントの違いがある。アクセントの生成において重要な役割を果たす物理量が，声の高さとして知覚される基本周波数の時間変化 (基本周波数パターン) である。具体的な方言として大阪方言を取り上げ，基本周波数パターンと方言の自然性の関係を分析するために，不自然な大阪方言の基本周波数パターンにどのような調整を加える事で，自然な大阪方言となるのかを検討した。基本周波数パターンの分析には藤崎モデルと呼ばれる代表的なモデル化手法を用いた。基本周波数パターンは，基本周波数の上がり下がりの位置を表すアクセント位置と上がり下がりの大きさを表すアクセント指令で表現される。実験では，不自然な合成音声の基本周波数パターンを大阪方言のアクセント型に対応したものに変更し，それを用いて大阪方言発話の合成を行った。アクセント型の変更は，アクセント指令の大きさはそのままにして位置だけを調整した。聴取実験によって，調整した合成音声がもとの合成音声よりも自然であるかどうかを 5 段階で判定した。聴取実験の結果，もとの音声のスコア 2.2 に対して，調整後の音声では 3.76 というスコアが得られ，大阪方言においては，アクセントの位置を適切に設定することにより，その他の成分を変更しなくても自然に近い発話を音声合成することが可能であることが示された。

(5) 多様な声質を持つ音声の例としてアニメにおける声優の音声（アニメボイス）を取り上げ，分析を行った。アニメボイスは独特

の声質・口調であり，その特性は一般人の声の特性とどう異なるのか，アニメボイスであると判断する要素とは何かという疑問を明らかにすることを目指して，一般の声とアニメボイスの比較を行った。声優の発話訓練を受けた女性の音声を収録し分析に用いた。収録した音声は６０発話である。これは，単母音５種類と，システムボイスやブラウザゲームのキャラクターの台詞を参考にした発話１５種類を，３種類の声質「平常」「キュート」「クール」で発声してもらった音声である。「キュート」と「クール」をアニメボイスらしい音声とした。声の分析の結果，単母音のスペクトル包絡において殆どの台詞について,「キュート」の包絡の形状が「平常」,「クール」と比べ大きく異なるものとなった。また，F0 の変移は「キュート」の高低差が最も目立っていた。これは分散の値と深く対応する。平均や分散においても，殆どの音声について数値の大きさが「平常」,「クール」と比べ非常に大きなものとなった。

(6) 音声の明瞭性を自動予測する手法を提案し評価した。音声スペクトルにおけるダイナミックレンジを用いて明瞭性を予測する。アナウンサー，セミプロ話者，一般話者の 3 種類に音声を分類し，この順に明瞭性が高いとの予測結果を得た。スペクトルダイナミックレンジを算出する音声区間や周波数帯域について検討を行い，それらを変更しても大きな差がないことを確認した。

(7) 人間の音声には言語情報の他に話者の発話意図や感情状態の情報が含まれている。これらの情報は円滑なコミュニケーションを実現するために非常に重要である。そこで，感情を含む音声と音声の物理パラメータの関係を明らかにするために，感情音声の印象評価実験を行った。音声は短い文章を読み上げた 22 発話を収録し，被験者にそれぞれの音声に対し「若いー老いた」「明瞭ー不明瞭」「硬いー柔らかい」「上品ー下品」「太いー細い」の５つの印象評価項目を５段階で評価をしてもらった。用いた物理パラメータは,「ダイナミックレンジ」「話速」「スペクトル包絡傾斜」「スペクトル傾斜」の４つである。「ダイナミックレンジ」は音声におけるパワーの強い部分と弱い部分との比率のことで，対数スペクトル包絡において，一次回帰直線とスペクトル包絡との差分を全周波数帯域で加算した値として算出した。印象評価の実験結果から算出した音声ペアの距離と物理パラメータで算出した音声ペアの距離を比較することで印象評定と物理パラメータの関係を分析した。2 種類の距離の間での相関は高々0.36 に留まり，印象評価と物理パラメータの間での明確な関連性は見られなかった。原因として，用いた音声の発話内容が行っていたため，言語情報が印象に影響を与えた可能性があると考えている。

５．主な発表論文等

〔雑誌論文〕（計 2 件）

① Y.Yamashita, A Review of Paralinguistic Information Processing for Natural Speech Communication, Acoustical Science and Technology, 査読有, 34, 2, 2013, pp.73–79
DOI: 10.1250/ast.34.73

② D.Khanh Ninh, M.Morise, Y.Yamashita, A Generation Error Function Considering Dynamic Properties of Speech Parameters for Minimum Generation Error Training for Hidden Markov Model-based Speech Synthesis, Acoustical Science and Technology, 査読有, 34, 2, 2013, pp.123–132
DOI: 10.1250/ast.34.123

〔学会発表〕（計 10 件）

① 小田原一成, 新妻雅弘, 山下洋一, 音声中の検索語検出における共起情報の検討, 日本音響学会 2015 年春季研究発表会, 2015 年 3 月 16 日, 中央大学 (東京都・文京区)

② 銭コウ, 森勢将雅, 新妻雅弘, 山下洋一, 非可聴域の音信号を用いた音信号通信における性能改善の検討, 日本音響学会 2015 年春季研究発表会, 2015 年 3 月 16 日, 中央大学 (東京都・文京区)

③ 島川智行, 山下洋一, 特定話者に対するパラ言語情報の認識, 電子情報通信学会技術研究報告, 2014 年 1 月 24 日, 名城大学 (愛知県)

④ 摺木啓一郎, 森勢将雅, 山下洋一, 韻律情報における個人性知覚の分析, 日本

音響学会 2013 年秋季研究発表会, 2013 年 9 月 27 日, 豊橋技術科学大学 (愛知県)

⑤ 島川智行, 山下洋一, パラ言語情報認識のための個人性の分析, 日本音響学会 2013 年秋季研究発表会, 2013 年 9 月 27 日, 豊橋技術科学大学 (愛知県)

⑥ 摺木啓一郎, 森勢将雅, 山下洋一, 韻律情報の知覚による個人性識別, 電子情報通信学会技術研究報告, 2013 年 6 月 14 日, 新潟大学 (新潟県)

⑦ 島川智行, 森勢将雅, 山下洋一, パラ言語情報処理のための対話音声の収録とラベリング, 電子情報通信学会技術研究報告, 2013 年 1 月 31 日, 同志社大学 (京都府)

⑧ 辻村祥平, 森勢将雅, 山下洋一, 孤立発声母音を用いた声質類似度の評価と自動推定, 電子情報通信学会技術研究報告, 2013 年 1 月 31 日, 同志社大学 (京都府)

⑨ D.Khanh Ninh, M.Morise, Y.Yamashita, Incorporating Dynamic Features into Minimum Generation Error Training for HMM-Based Speech Synthesis, The 8th International Symposium on Chinese Spoken Language Processing, 2012 年 12 月 6 日, Kowloon(中国)

⑩ D.Khanh Ninh, M.Morise, Y.Yamashita, An adaptive weighting approach for minimum generation error training considering dynamic features in HMM-based speech synthesis, 2012 Autumn Meeting of Acoustical Society of Japan, 2012 年 9 月 21 日, 信州大学 (長野県)

６．研究組織

(1) 研究代表者

山下　洋一（YAMASHITA　Yoichi）
立命館大学・情報理工学部・教授
研究者番号：８０１７４６８９