Vol.2014-MUS-103 No.33
2014/5/24

単語アクセント型ごとのF0波形と時間長を利用した単語音声の感情変換

劉 丹[1]　堂元 健太郎[1]　井上 祐輔[1]　宇津呂 武仁[2]

概要：本論文では，単語アクセント型ごとに，各感情のもとでの単語音声のF0波形と時間長を分析し，単語アクセント型・各感情の組の単位で，F0波形と時間長が似通っていることを示す．そして，教師用単語(教師語)の感情音声のF0波形・時間長を流用することにより，評価用単語(評価語)の感情音声変換が可能であることを実験的に示す．

キーワード：感情音声変換，単語アクセント型，基本周波数，時間長

Emotional Voice Conversion utilizing F0 Contour and Duration of Word Accent Type

Dan Liu[1]　Kentaro Domoto[1]　Yusuke Inoue[1]　Takehito Utsuro[2]

Abstract: This paper analyzes the F0 contour and duration of emotional voice of a word for each word accent type. Then, for each pair of a word accent type and an emotion type, we show that words belonging to the corresponding word accent type with the corresponding emotion type share almost similar F0 contour and duration. Finally, we show experimental results of emotional voice conversion of words for evaluation, by replacing their F0 contour and duration with those of words for training within each word accent type.

Keywords: emotional voice conversion, word accent type, F0 contour, duration

1. はじめに

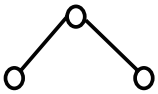
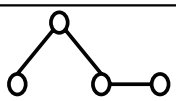
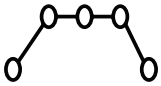
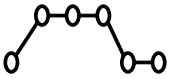
音声は，我々の日常的なコミュニケーヒョン手段の一つである．通常音声中には，発話者の意志を伝達する言語情報，年齢・性別等の個人性情報，感情・気分などを伝達する感情情報等を含め様々な情報が含まれている．このうち，感情情報は，人間関係を改善するための重要な役割を担うこともあれば，逆に，人間関係や社会関係に悪い影響を与える場合もある．よい人間関係と社会関係を構築するため，感情音声変換技術を備えた音声情報伝達システムの開発は非常に重要であると考えられる．このようなシステムを開発することができれば，音声上の感情情報の取り扱いが容易になり，人間関係における感情の取り扱いを容易にすることができる．そして，産業面，教育面，医療面等，多様な局面における様々な応用が期待できる．

感情音声合成の研究においては，文献 [2, 5, 7, 9] 等の研究が行われており，例えば，文献 [2] では，感情音声コーパスを作成し，音素ごとに，基本周波数，パワー，時間長の三つの韻律パラメータを分析している．その結果として，基本周波数の大きさは，悲しみ，怒り，喜びの感情との間の相関が大きく，悲しみ，怒り，喜びの順に高くなる，としている．一方，時間長に関しては，悲しみの感情における時間長は，怒り，および，喜びよりも長い，としている．また，パワーに関しては，三つの感情の間で有意な差はないとしている．

一方，感情音声変換の研究は多くは行われておらず，現状において，実用レベルの感情音声変換技術は実現されていない．今後の実用化のためには，より精度が高い感情音

[1] 筑波大学大学院システム情報工学研究科
Graduate School of Systems and Information Engineering, University of Tsukuba, Tsukuba, 305–8573, Japan
[2] 筑波大学 システム情報系
Faculty of Engineering, Information and Systems, University of Tsukuba, Tsukuba, 305–8573, Japan

表 1　分析対象アクセント型 (モーラ数ごと)

モーラ数	アクセント型	教師語	評価語
2	LH 	なみ /nami/	かう /kau/
3	LHL 	ななめ /midori/	はいる /hairu/
4	LHLL 	あまみず /amamizu/	おませな /omasena/
5	LHHHL 	やわらげる /yawarageru/	いれかわる /irekawaru/
6	LHHHLL 	いわずもがな /iwazumogana/	うちのめした /uchinomeshita/

声変換手法の開発が必要である．

そこで，本論文では，韻律パラメータとして基本周波数および時間長を用いて，単語音声の感情変換を行う方式を提案する．まず，単語アクセント型ごとに，各感情のもとでの単語音声の F0 波形と時間長を分析し，単語アクセント型が同一である異なる二単語の間で，感情音声の F0 波形と時間長を比較し，各感情のもとでの F0 波形と時間長が似通っていることを示す．そして，教師語の感情音声の F0 波形・時間長を流用することにより，評価語の感情音声変換が可能であることを実験的に示す．特に，本論文では，

- 読み上げ調の平静音声から，怒り，および，悲しみの感情音声への変換，
- 悲しみの感情音声から怒りの感情音声への変換，
- 怒りの感情音声から悲しみの感情音声への変換，

を対象として評価実験を行った結果を報告する．

2. 単語アクセント型ごとの感情音声データベースの作成と分析

2.1 教師語・評価語の選定

日本語の単語が持つアクセント型は，モーラ (拍) を単位として，ピッチの高低変化で表現する．ピッチが高いモーラの記号を H(high) とし，ピッチが低いモーラの記号を L(low) とする．個々の単語のアクセント型はピッチの高 (H) と低 (L) の記号の組み合わせで表記する．モーラは，日本語における仮名一文字の音の長さの音韻単位である．また，ピッチが下がる直前の位置を決定するモーラを，アクセント核と呼ぶ．

ここで，文献 [7] においては，任意の単語の感情音声を合成するために，モーラ数 2~6 の計 20 種類の単語アクセント型の各々に対して，日本人男性声優 1 人が，平静を含む様々な感情のもとで発声した感情音声を収録した感情音声データベースを作成している．本論文においても，文献 [7] の枠組みを参考にして，2~6 の各モーラ数について，表 1 に示すアクセント型一つを選び，合計で 5 種類のアクセント型を評価対象とする．そして，各アクセント型について，文献 [7] のデータベースに収録されている単語を教師語 (表 1) とする．一方，評価語としては，各アクセント型について，オンライン日本語アクセント辞書 OJAD [6][*1] に収録されている単語として，表 1 に挙げたものを選定する．

本論文では，表 1 に示す 5 種類のアクセント型について，教師語・評価語の計 10 単語を用意した後，感情音声の収録を行った．収録対象の感情は，平静，怒り，悲しみとした．一方，喜びの感情については，文献 [7] において，感情伝達において重要な物理量が韻律成分ではない可能性があると報告されていることから，本論文においても分析・評価の対象としなかった．実際に，喜びの感情については，感情変換についての予備実験の結果において，提案手法の有効性が十分に確認できなかったことから，MFCC 等の声質パラメータ等，感情音声変換についての関連方式，および，感情音声合成方式において有効性が報告されている他の特徴量を併用する方式の検討が必要であると考えられる．

2.2 感情音声データベースの作成手順

表 1 の計 10 単語について，平静，怒り，悲しみの各感

[*1] http://www.gavo.t.u-tokyo.ac.jp/ojad/

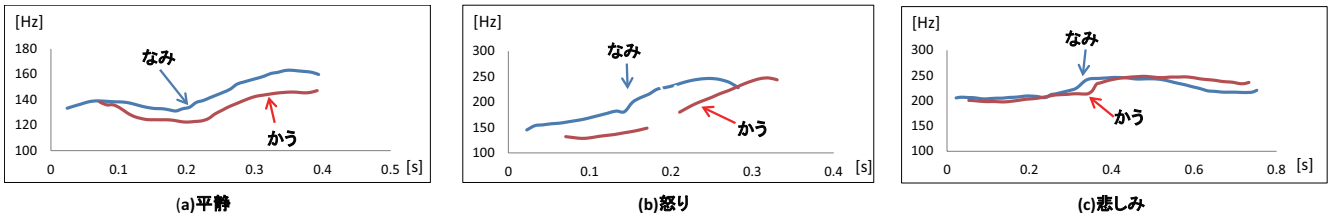


図 1 平静および感情音声の F0 波形 (モーラ数: 2，アクセント型: LH，話者 A)

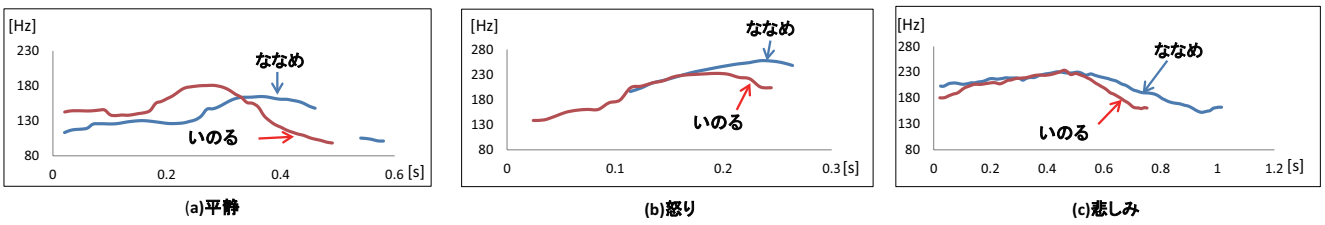


図 2 平静および感情音声の F0 波形 (モーラ数: 3，アクセント型: LHL，話者 A)

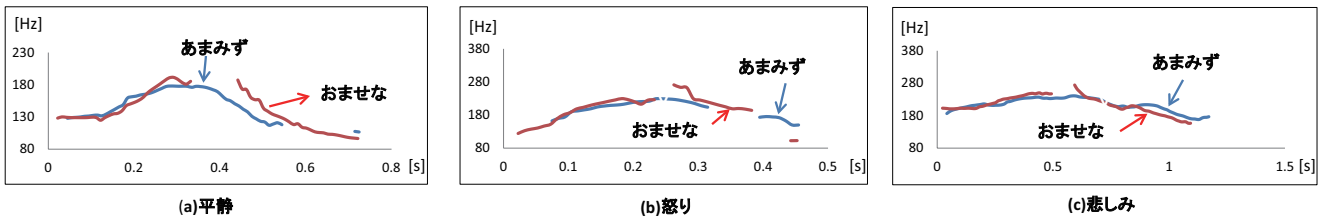


図 3 平静および感情音声の F0 波形 (モーラ数: 4，アクセント型: LHLL，話者 A)

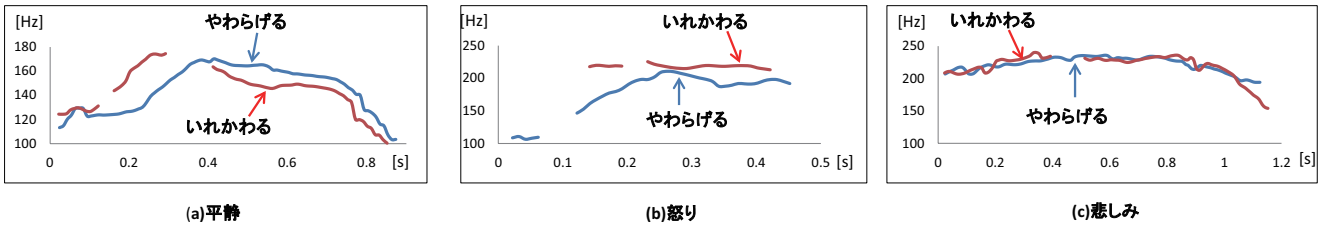


図 4 平静および感情音声の F0 波形 (モーラ数: 5，アクセント型: LHHHL，話者 A)

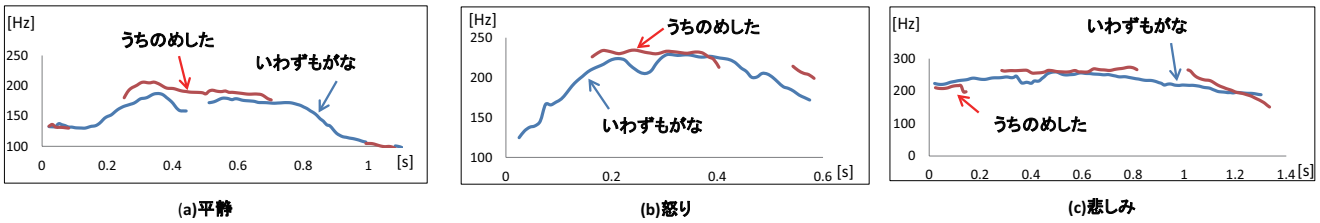


図 5 平静および感情音声の F0 波形 (モーラ数: 6，アクセント型: LHHHLL，話者 A)

情音声計 30 音声の収録を行う．話者としては以下の 2 名を選んだ[*2]

話者 A 声優のための専門学校の教育を 1 年間受けた日本人男性 1 名.

話者 B 日本語能力試験 N1 の資格を持ち，日本に 2 年 6 ヶ月在住しており，日本において日本語学校を卒業後，総合大学大学院に入学・在籍している中国人男性 1 名.

音声の収録は，Mac 上の Praat[*3] を用い，サンプリング周波数は 48kHz とした．感情表現の仕方およびその程度においては，個人差や録音する時の状況依存性等があるため，ある程度それらを統一するために，文献 [7] の感情音声データベースに収録されている教師語の各感情音声 (平静，怒り，悲しみ) を話者に聴取させ，できる限りそれらの感情音声と同じ基準で感情音声を発声するように指示をした．

[*2] 感情の表現の仕方は話者によって個人差があるが，感情変換は話者ごとに行なったため，3.2 節の主観評価実験においては，話者 A と話者 B の間では，ほぼ同等の評価結果となった.

[*3] http://www.fon.hum.uva.nl/praat/

2.3 分析結果

単語アクセント型ごとに，教師語・評価語の各感情音声の基本周波数の時間軸方向の推移をプロットした結果を図 1~5 に示す．プロットの際には，話者 A から収録した感情音声に対して，Praat を用いて基本周波数および時間長を 10ms 間隔で抽出し，教師語のプロットを青線で，評価語のプロットを赤線で，それぞれ示す．この結果から分かるように，各アクセント型・各感情においてプロットした教師語・評価語の基本周波数の推移の形はかなり類似していると言える．したがって，教師語の感情音声の基本周波数・時間長を流用することによって評価語音声の感情変換を行う提案方式が，ある程度適切であることが期待できると言える．

3. 単語アクセント型ごとの F0 波形と時間長を利用した感情音声変換

3.1 変換手順

まず，教師語および評価語の感情音声を表すために表記 X を用い，単語 w，基本周波数 F0 波形 f，時間長 d，感情 e，話者 SP を用いることにより，感情音声 X を次式で表す．

$$X = \langle w,\ f,\ d,\ e,\ SP \rangle$$

すると，教師語 w_{tr} について，感情変換前後の感情を，それぞれ e_s および e_t とすると，話者 SP_Z による教師語 w_{tr} の変換前感情音声 X_s^{tr}，および，変換後感情音声 X_t^{tr} は，それぞれ次式で表される．

$$X_s^{tr} = \langle w_{tr},\ f_s^{tr},\ d_s^{tr},\ e_s,\ SP_Z \rangle$$
$$X_t^{tr} = \langle w_{tr},\ f_t^{tr},\ d_t^{tr},\ e_t,\ SP_Z \rangle$$

一方，評価語 w_{ts} について，同様に，感情変換前の感情 e_s での，話者 SP_Z による感情音声 X_s^{ts} は次式で表される．

$$X_s^{ts} = \langle w_{ts},\ f_s^{ts},\ d_s^{ts},\ e_s,\ SP_Z \rangle$$

ここで，感情音声変換を表す関数を EmoConv とすると，本論文の感情音声変換においては，評価語の変換前感情音声 X_s^{ts} 中の基本周波数 F0 波形 f_s^{ts} および時間長 d_s^{ts} を，教師語の変換後感情音声 X_t^{tr} 中の基本周波数 F0 波形 f_t^{tr} および時間長 d_t^{tr} に差し換えることにより，関数 EmoConv を実現する．ただし，基本周波数には個人差があるため，本論文の感情音声変換において，教師語の変換後感情音声を用いて評価語音声の感情変換を行う際には，同一話者の範囲での感情音声変換のみを行う．以上をまとめると，感情音声変換関数 EmoConv は次式で表される．

$$\mathrm{EmoConv}\left(X_s^{ts}\right) = \langle w_{ts},\ f_t^{tr},\ d_t^{tr},\ e_{t'},\ SP_Z \rangle$$

ただし，感情変換後の感情については，$e_{t'}$ と表記している．

以上の変換手順のうち，時間長の変換においては，音声パラメータ操作ツール STRAIGHT[*4]の「音声のピッチなどを変換せずに時間長だけを一定に変換させる」の機能を用いる．ただし，この際，評価語の変換前感情音声 X_s^{ts} と教師語の変換後感情音声 X_t^{tr} の間で，STRAIGHT を用いて手動で音節のアライメントをとることにより，音節の時間長単位で変換を行う．また，基本周波数 F0 波形の変換においても，STRAIGHT を用いて抽出したピッチ軌跡を差し換えることにより，変換操作を行う．

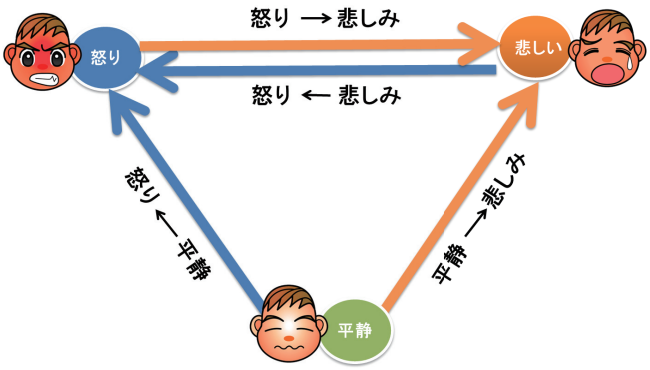


図 6 教師語音声を利用した評価語の感情音声変換実験の流れ (4 通りの変換)

また，本論文では，変換前感情 e_s および変換後感情 e_t の組として，図 6 および以下に示す 4 通りを対象とする．

$$\begin{aligned} e_s = \text{平静} &\longrightarrow e_t = \text{怒り} \\ e_s = \text{平静} &\longrightarrow e_t = \text{悲しみ} \\ e_s = \text{悲しみ} &\longrightarrow e_t = \text{怒り} \\ e_s = \text{怒り} &\longrightarrow e_t = \text{悲しみ} \end{aligned} \quad (1)$$

3.2 主観評価実験

3.2.1 評価手順

主観評価実験によって提案手法の有効性を検証するために，評価語 w_{ts} について，感情変換後の感情 e_t での，話者 SP_Z による感情音声 X_t^{ts}

$$X_t^{ts} = \langle w_{ts},\ f_t^{ts},\ d_t^{ts},\ e_t,\ SP_Z \rangle$$

を参照用目標感情音声として，5 段階 DMOS 主観評価実験を行う．

具体的には，提案手法による感情変換音声 $\mathrm{EmoConv}\left(X_s^{ts}\right)$ を含む以下の 3 種類の音声を被験者に聞かせた上で，主観評価対象の感情変換音声に 1~5 点の 5 段階の点数を付与させる．

- 評価語 w_{ts} の平静音声 $X_{\text{平静}}^{ts}$ (1 点に固定)
- 参照用目標感情音声 X_t^{ts} (5 点に固定)
- 評価対象の感情変換音声 $\mathrm{EmoConv}\left(X_s^{ts}\right)$

被験者は成人男女 10 名とし，

[*4] http://www.wakayama-u.ac.jp/~kawahara/STRAIGHTadv/index_j.html

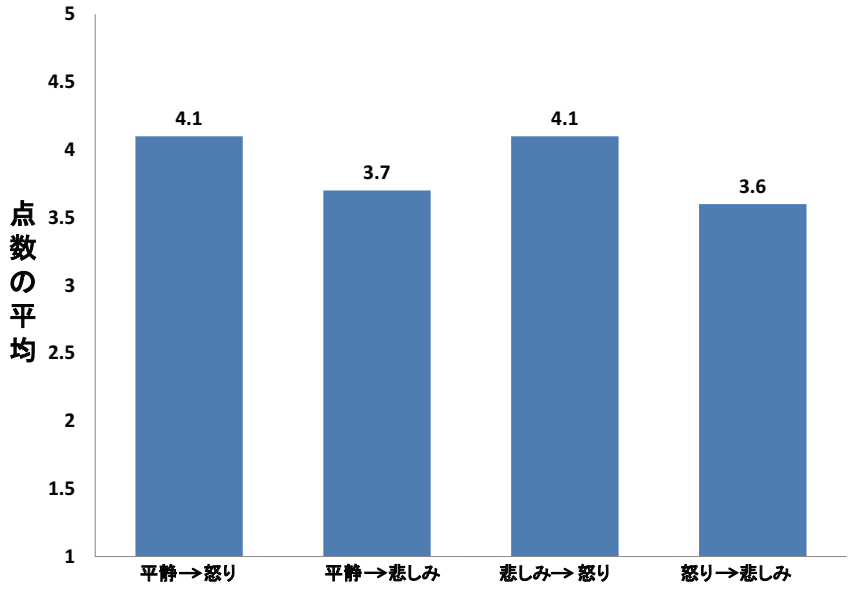


図 7 感情変換結果音声に対する主観評価結果 (感情組ごと)

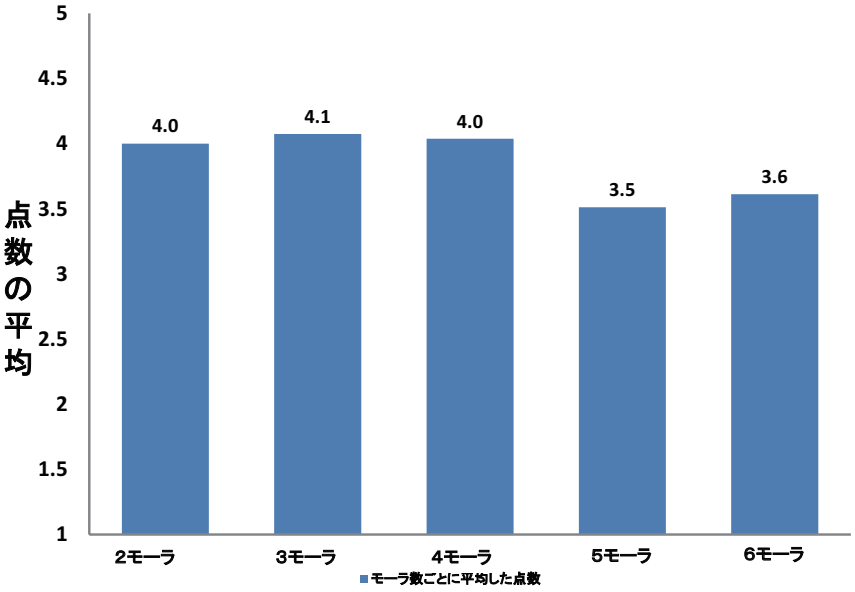


図 8 感情変換結果音声に対する主観評価結果 (モーラ数ごと)

$$\left(\begin{array}{c}\text{式 (1) の}\\ \text{4 通りの}\\ \text{感情組}\end{array}\right) \times \left(\begin{array}{c}\text{全 5 種類の}\\ \text{評価対象}\\ \text{アクセント}\\ \text{型の評価語}\end{array}\right) \times \left(\begin{array}{c}\text{話者 AB}\\ \text{の 2 名}\end{array}\right) = 40 \text{ 音声}$$

の評価音声を無作為な順序で選び，以下の手順に沿って被験者に主観評価を行わせる．

(1) まず，評価語 w_{ts} の平静音声 $X^{ts}_{\text{平静}}$ を聞かせ，この点数を 1 点とする．
(2) 次に，参照用目標感情音声 X^{ts}_{t} を聞かせ，この点数を 5 点とする．
(3) 最後に，評価対象の感情変換音声 $\text{EmoConv}\left(X^{ts}_{s}\right)$ を聞かせ，参照用目標感情音声 X^{ts}_{t} に類似している程度に応じて，1~5 点の点数をつけさせる．

ただし，被験者には聞き直すことを許可して主観評価実験を行わせる．

3.2.2 評価結果

主観評価結果の平均値は 3.9 となり，比較的高い結果となった．この結果から，基本周波数および時間長の二つの韻律パラメータは，怒りおよび悲しみの感情音声の表現において重要な働きをすることが分かった．また，教師語の感情音声の基本周波数・時間長を流用することより，評価語音声の感情変換が可能であることが分かった．

次に，式 (1) の 4 通りの感情組ごとに主観評価結果を平均した結果を図 7 に示す．この結果から分かるように，悲しみの感情音声への変換よりも怒りの感情音声への変換

の方が高い評価結果となった．一方，モーラ数ごと (実際にはアクセント型ごと) に主観評価結果を平均した結果を図 8 に示す．この結果から分かるように，2∼4 モーラの評価語における主観評価結果は 4 点以上の高評価となったが，5∼6 モーラの評価語においては，語長が長いことが原因で感情変換音声が劣化する場合があり，主観評価結果が 4 点未満に下がる結果となった．

4. 関連研究

感情音声変換に関する関連研究 [1,3,4,8,10] においては，GMM を用いて基本周波数や声質パラメータの変換を行う方式 [1,4,10] や，回帰木等の手法を用いて基本周波数，時間長等の韻律パラメータの変換を行う方式 [3,8] 等が提案されている．これらの研究の多くにおいては，文を対象として感情変換を行った結果に対して，悲しみ，怒り，喜びといった代表的な感情の間の識別の主観評価タスクを通して提案手法の評価を行っている．

GMM を用いる手法の一つとして，文献 [4] では，声質パラメータである 40 次元の MFCC パラメータを用いて，GMM に基づく声質変換手法により，文単位の感情音声変換を行う手法を提案している．この論文では，声質パラメータだけでは十分な感情変換が実現できないと報告している．同様に，文献 [10] においては，GMM を用いる手法において，音節ごとの基本周波数を用いて感情音声変換を行う手法を提案している．一方，文献 [1] においては，GMM を用いる手法において，声質パラメータであるスペクトル包絡と基本周波数 F0 を用いて，単語単位の感情音声変換を行う手法を提案している．

また，文献 [3] においては，回帰木による時間長変換方式，および，GMM によるスペクトル包絡変換方式に加えて，基本周波数の変換を行う方式を併用することにより感情音声変換を行う手法を提案している．一方，文献 [8] においては，感情音声変換において，回帰木による基本周波数変換方式と GMM による基本周波数変換方式の比較を行っている．

一方，感情音声合成に関連する研究として，文献 [7] においては，韻律パラメータを主成分分析して得られた部分空間を用いて，任意の単語の感情音声を合成する手法を提案している．文献 [5] においては，感情音声合成において，二分回帰木を用いて基本周波数パターン生成過程モデルのパラメータ推定，および，音素持続時間長の推定を行っている．文献 [9] においては，感情を含む音声を訓練事例として HMM 音声合成システムの学習を行った結果について報告している．

5. おわりに

本論文では，韻律パラメータとして基本周波数および時間長を用いて，単語音声の感情変換を行う方式を提案した．まず，単語アクセント型ごとに，各感情のもとでの単語音声の F0 波形と時間長を分析し，単語アクセント型が同一である異なる二単語の間で，感情音声の F0 波形と時間長を比較し，各感情のもとでの F0 波形と時間長が似通っていることを示した．そして，教師語の感情音声の F0 波形・時間長を流用することにより，評価語の感情音声変換が可能であることを実験的に示した．

今後の課題として，文献 [7] において評価対象として用いられた 20 種類のアクセント型全てを対象として，提案手法を評価する予定である．同時に，各アクセント型における評価語の数を増やし，提案手法の評価を行う．また，本論文の単語アクセント型の知識を利用する手法と，関連研究における感情音声変換手法 [1,3,4,8,10] との比較を行い，提案手法の長所・短所について分析を行う必要がある．さらに，評価語話者とは異なる話者が発話した教師語感情音声を情報源として評価語音声の感情変換を行う方式を確立する必要がある．

参考文献

[1] 相原　龍，高島遼一，滝口哲也，有木康雄：スペクトルと韻律を特徴量とした GMM による感情音声変換，日本音響学会 2012 年春季研究発表会講演論文集，pp. 503–504 (2012).
[2] 飯田朱美，ニックキャンベル，安村通晃：感情表現が可能な合成音声の作成と評価，情報処理学会論文誌， Vol. 40, No. 2, pp. 479–486 (1999).
[3] Inanoglu, Z. and Young, S.: Emotion Conversion using F0 Segment Selection, *INTERSPEECH*, pp. 2122–2125 (2008).
[4] 岩見洋平, 戸田智基, 川波弘道, 猿渡　洋, 鹿野清宏：GMM に基づく声質変換を用いた感情音声合成，電子情報通信学会技術研究報告， SP2002-171，pp. 11–16 (2003).
[5] 桂　聡哉，広瀬啓吉，峯松信明：感情音声合成のための生成過程モデルに基づくコーパスベース韻律生成とその評価，電子情報通信学会技術研究報告, SP2002-184，pp. 31–36 (2003).
[6] 峯松信明：オンライン日本語アクセント辞書 OJAD の開発と利用，国語研プロジェクトレビュー， Vol. 4, No. 3, pp. 174–182 (2014).
[7] 森山　剛，森真　也，小沢慎治：韻律の部分空間を用いた感情音声合成，情報処理学会論文誌， Vol. 50, No. 3, pp. 1181–1191 (2009).
[8] Tao, J., Kang, Y. and Li, A.: Prosody Conversion from Neutral Speech to Emotional Speech, *IEEE Transactions on Audio, Speech and Language Processing*, pp. 1145–1154 (2006).
[9] 都築亮介，全　炳河，徳田恵一，北村　正，Bulut, M., Narayanan, N. S.：HMM 音声合成における感情表現のモデル化，電子情報通信学会技術研究報告, SP2003-78，pp. 25–30 (2003).
[10] Veaux, C. and Rodet, X.: Intonation Conversion from Neutral to Expressive Speech, *INTERSPEECH*, pp. 27–31 (2011).