単語アクセント型ごとのF0波形に着目した単語音声の感情変換*

☆劉丹, 堂元健太郎, 宇津呂武仁 (筑波大)

1 はじめに

音声は，我々の日常的なコミュニケーシ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 [1–4] 等の研究が行われており，例えば，文献 [1] 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声変換手法の開発が必要である．

そこで，本研究では，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 [5]．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，本研究では，

- 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．

Table 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 /naname/	いのる /inoru/
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/

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 [4] 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．

本稿でも，文献 [4] の枠組みを参考にして，モーラ数2~6の計20種類の単語アクセント型を評価対象とし，各アクセント型について，文献 [4] のデータベースに収録されている単語を教師語とする．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から選定する．ここで，2~6の各モーラ数について，教師語と評価語の例をTable 1 に挙げる．

本研究で取り扱う感情の対象は，平静，怒り，悲しみとした．一方，喜びの感情については，文献 [4] において，感情伝達において重要な物理量が韻律成分で

*Emotional Voice Conversion based on F0 Contour and Duration of Word Accent Type, by LIU, Dan, DOMOTO, Kentaro, UTSURO, Takehito (University of Tsukuba)

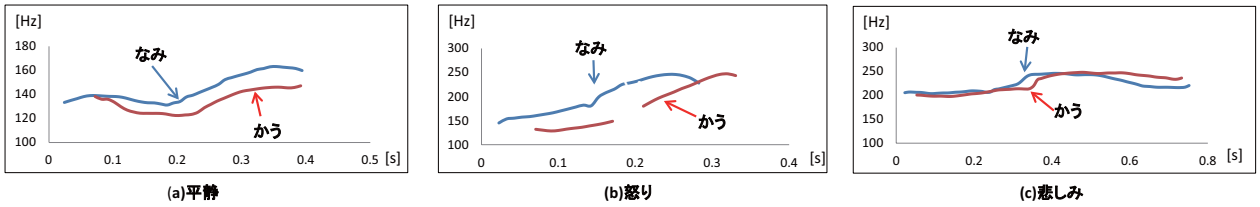

Fig. 1　平静および感情音声の F0 波形 (モーラ数: 2，アクセント型: LH，話者 A)

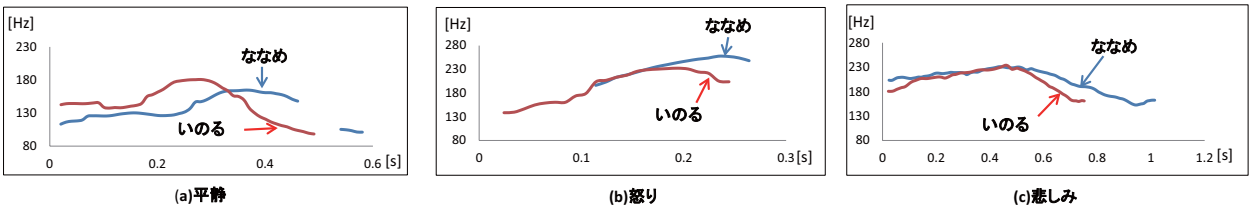

Fig. 2　平静および感情音声の F0 波形 (モーラ数: 3，アクセント型: LHL，話者 A)

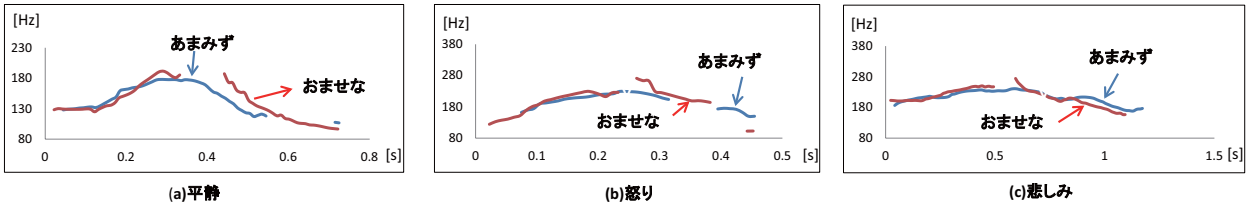

Fig. 3　平静および感情音声の F0 波形 (モーラ数: 4，アクセント型: LHLL，話者 A)

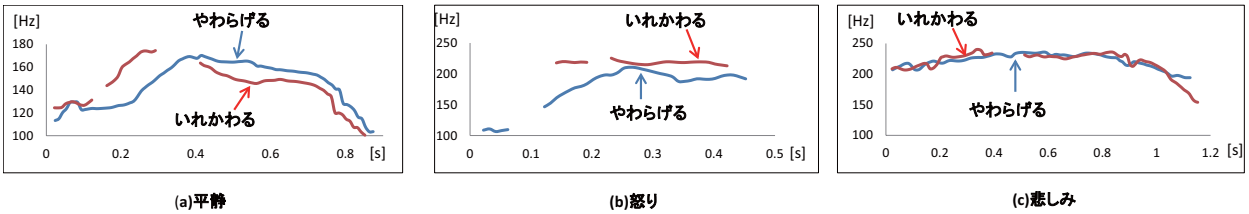

Fig. 4　平静および感情音声の F0 波形 (モーラ数: 5，アクセント型: LHHHL，話者 A)

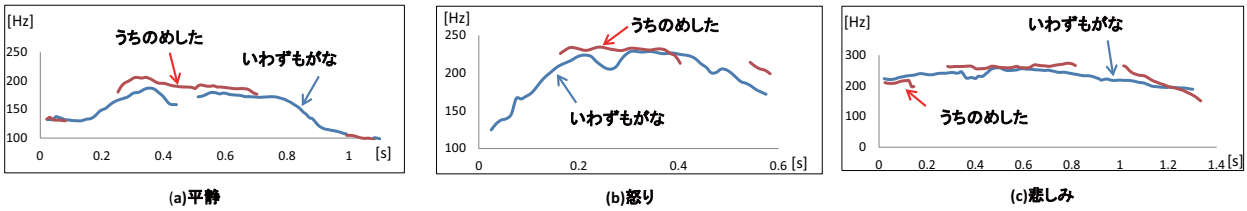

Fig. 5　平静および感情音声の F0 波形 (モーラ数: 6，アクセント型: LHHHLL，話者 A)

はない可能性があると報告されていることから，本研究においても分析・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．

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

2.2　感情音声データベース

本研究では，20 種類のアクセント型の教師語・評価語の計 40 単語について，平静，怒り，悲しみの各感情音声計 120 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.

[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.

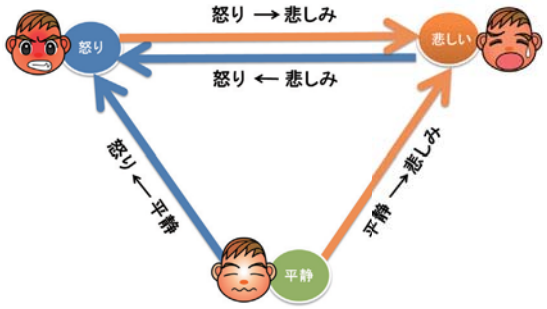

Fig. 6　教師語音声を利用した評価語の感情音声変換の流れ (4 通りの変換)

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 [4] 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.

2.3　分析結果

20 種類の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の抜粋として，Table 1 に挙げた単語アクセント型についてのプロットを Fig. 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 波形と時間長を用いた感情音声変換

3.1　変換手順

変換前感情 e_s および変換後感情 e_t の組としては，Fig. 6 に示す読み上げ調の平静音声から悲しみと怒り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，の 4 種類の感情変換を行う.

基本周波数には個人差があるため，本研究の感情音声変換において，変換後感情 e_t の教師語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．変換手順として，まず，時間長の変換を行う．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．評価語音声の時間長を，変換後感情 e_t の教師語音声の時間長と同じ長さに変換させる．ただし，この際，変換前感情 e_s の評価語音声と変換後感情 e_t の教師語音声の間で，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.

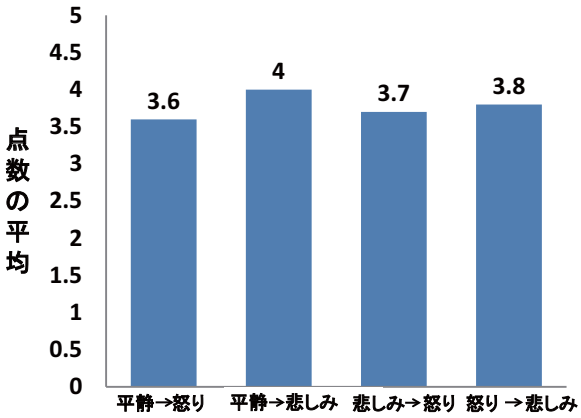

Fig. 7　感情変換結果音声に対する主観評価結果 (感情組ごと)

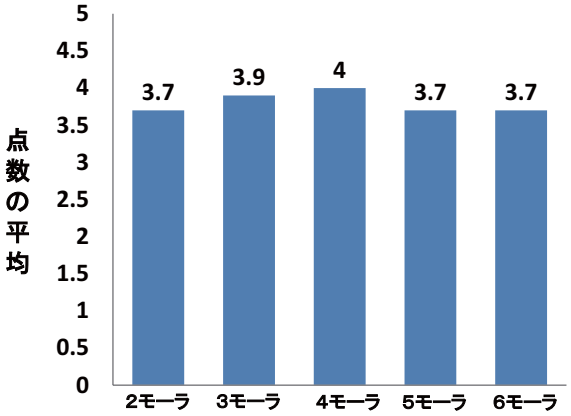

Fig. 8　感情変換結果音声に対する主観評価結果 (モーラ数ごと)

3.2　主観評価実験

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，感情変換後の評価語音声に対して，同一話者

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

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

Table 2　感情変換結果音声に対する識別実験結果

変換後の感情	被験者による判定結果の感情 (%)		
	怒り	悲しみ	平静
怒り	67.3	2.9	29.8
悲しみ	1.3	85.4	13.3
平静 (変換なし)	0.7	5.6	93.7

が発声した評価語の感情音声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．

被験者は成人男女 10 名とし，感情変換後の評価語音声 160 種類 (単語アクセント型ごと 20 単語 × 感情変換 4 通り × 話者 2 人) 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，評価語の平静音声を聞かせ，この点数を 1 点とする．
(2) 次に，参照用目標感情音声を聞かせ，この点数を 5 点とする．
(3) 最後に，感情変換後の評価語音声を聞かせ，参照用目標感情音声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．

DMOS 主観評価結果の平均値は 3.8 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，4 通りの感情組ごとに DMOS 主観評価結果を平均した結果を Fig. 7 に示す．この結果から分かるように，怒りの感情音声への変換よりも悲しみの感情音声への変換の方が高い評価結果となった．一方，モーラ数ごとに DMOS 主観評価結果を平均した結果を Fig. 8 に示す．この結果から分かるように，評価対象アクセント型数の少ない 2 モーラの場合 (2 モーラのアクセント型は 2 種類) を除いて，モーラ数が少ない方が高い評価結果となった．

また，被験者 10 人を対象として，感情変換無しの平静音声 40 通り，および，感情変換された感情音声 160 通りの計 200 音声を対象として，怒り，悲しみ，平静のいずれであるかの識別実験を行った結果を Table 2 に示す．この結果では，平静，悲しみ，怒りの順の識別率となった．怒りの感情変換においては，パワー等の他の特徴量の併用が必要であると考えられる．

4　関連研究

感情音声変換に関する関連研究 [7–11] 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 [7,10,11] 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 [8,9] 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．一方，感情音声合成に関連する研究として，韻律の部分空間を用いて単語の感情音声を合成する手法 [4]，二分回帰木を用いて基本周波数パターン，音素持続時間長の推定を行う手法 [2]，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う手法 [3] 等が挙げられる．

5　おわりに

本研究では，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．今後の課題として，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 [7–11] 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] 飯田他：“感情表現が可能な合成音声の作成と評価”, 情処論誌, **40**, 2, pp. 479–486 (1999).

[2] 桂他：“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”, 信学技報 SP, pp. 31–36 (2003).

[3] 都築他：“HMM 音声合成における感情表現のモデル化”, 信学技報 SP, pp. 25–30 (2003).

[4] 森山他：“韻律の部分空間を用いた感情音声合成”, 情処論誌, **50**, 3, pp. 1181–1191 (2009).

[5] 劉他：“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”, 情報処理学会研究報告, **2014–MUS–103**, (2014).

[6] 峯松：“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, **4**, 3, pp. 174–182 (2014).

[7] 岩見他：“GMM に基づく声質変換を用いた感情音声合成”, 信学技報 SP, pp. 11–16 (2003).

[8] J. Tao, et al.: “Prosody conversion from neutral speech to emotional speech”, IEEE Transactions on Audio, Speech and Language Processing, pp. 1145–1154 (2006).

[9] Z. Inanoglu, et al.: “Emotion conversion using F0 segment selection”, INTERSPEECH, pp. 2122–2125 (2008).

[10] C. Veaux, et al.: “Intonation conversion from neutral to expressive speech”, INTERSPEECH, pp. 27–31 (2011).

[11] 相原他：“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”, 音講論 (春), pp. 503–504 (2012).