

高品質な感情音声を合成する手法と 顔型ロボットへの実装例

阪本 正治 斉藤 隆

日本アイ・ビー・エム（株）東京基礎研究所

あらまし 本論文では、ターゲットキャラクタの音声表出を模擬した音声合成器を目指して、ターゲットキャラクタが、感情が昂ぶったときに多用する言い回しや擬態語などを含んだ音声を高品質に合成する手法を提案する。また、実装例として、感情音声合成機能と音声認識機能および対話機能を顔型ロボットに搭載したデモンストレーションプロトタイプを紹介する。

A High Quality Emotional Speech Synthesis Method and A Face Robot

Masaharu Sakamoto, Takashi Saito

IBM Research, Tokyo Research Laboratory, IBM Japan

Abstract: In this paper, we propose an emotional speech synthesis method in which we aim to simulate the speech expression of a target character. Our method enables to synthesize high quality emotional speech by seamlessly combining phrases and onomatopoeia peculiar to the character. As an example of a practical use, we mention here a face robot that is equipped with our high quality speech synthesizer, a facial expression controller, a speech recognizer and a simple dialog manager.

1. はじめに

近年、テキスト音声合成でテキストを読み上げるだけでなく、感情を込めた音声（以下、感情音声）など多様なスタイルの音声を合成する需要が生まれつつある。たとえば、ロボットの発話機能においては、ロボットとの会話をより魅力的にし、人とのコミュニケーションを円滑にするために、より豊かな表現力を持つ音声合成が求められる。

これまで、感情とその音声での表出の研究がなされ、感情を伴った音声とそうでない音声、異なる感

情を伴った音声などの韻律的特徴や音響的特徴の違いが分析されている[1,4]。また、分析にとどまらず、実際に感情音声を合成する試みも報告されている[1-7]。

例えば、文献[2]では、感情音声の韻律的な特徴と音響的特徴を規則化し、フォルマント合成によって感情音声の合成を行っている。文献[3]では、感情音声の韻律的特徴や音響的特徴を、ユーザインターフェースを使って設計する方法を提案している。また、文献[1]では、大規模な感情音声コーパスを用いた波形重畳方式による感情音声の合成法を提案してい

る。これらの研究は、感情音声の韻律的特徴と音響的特徴を明らかにし、それらに基づいて、任意文に対して指定した感情を伴った音声を合成することをねらっている。しかし、任意文に対して、指定した感情を伴った音声を合成することは困難性が高い。また、合成音声を聴いて感情を判定する実験からは、高い正解率は得られていない[1]ことから、改善の余地がありそうである。

一方、文献[7]では、任意文を対象としたテキスト音声合成出力と感情音声や会話口調用の韻律テンプレートを用いた合成音を組み合わせた方式が提案されている。

本論文で述べる内容は、ターゲットキャラクタの音声表出を模擬する音声合成器をつくることをねらっており、それには、必ずしも任意文に対して、指定した感情で音声を合成しなくても、ターゲットキャラクタが、感情が昂ぶったときに多用する言い回しや擬態語などを含んだ音声を、高品質に合成することをねらえばよい。例えば、ロボットの発話機能に応用する場合、ロボットに、個性（キャラクタ）を与えることを考える。すると、語彙や話し方も、そのキャラクタを反映したものが求められる。このような応用では、そのキャラクタ特有の感情表現を幅広く取り揃えて、それらを高品質に合成するほうがより効果的であろう。

本論文では、感情音声の韻律的特徴と音響的特徴を登録し、合成時にそれらを用いて高品質な感情音声を合成する手法を提案する。一方、筆者らは、平静な感情における音声に対しては、ターゲット話者の音声から、その声の特徴を表す音声辞書（ボイスフォントと呼ぶ）を生成し、合成音声へ反映できるような音声合成システムを構築している [8-10]。このボイスフォントと感情音声合成手法を組み合わせることによって、文の一部が感情音声として登録されていない場合に、ボイスフォントを使ったテキスト音声合成出力で補間し、合成した感情音声と継ぎ目なく接続する（シームレス音声合成と呼ぶ）ことも可能となっている。

さらに、本論文では、以上の手法を適用したデモンストレーション・プロトタイプを紹介する。それ

は、感情音声合成機能を有したテキスト音声合成と音声認識機能および対話機能を搭載した顔型ロボットであり、高度な対話機能は有していないが、高品質な任意文の音声合成、感情音声の合成、顔の表情制御、音声認識機能、およびダイアログ・マネージャーを組み合わせることで、効果的なデモンストレーションとなった。

以下の章では、まず、本研究で使用した音声資料を紹介する。続いて、ボイスフォント技術と感情音声の合成手法について述べ、最後に、実装例としてデモンストレーション・プロトタイプを紹介する。

2. 音声資料

本研究では、表現力が豊かなプロの声優を選び、“男の子”のキャラクタの声で以下のテキストを読み上げてもらい録音した。

(A) ATR 連続音声スクリプト 503 文

(B) 音素バランス単語セット 1576 語

(C) 感情音声を収録するための 1600 文から構成されるモノローグ

(A)、(B) は、平静な感情における韻律的特徴と音響的特徴を抽出し、ボイスフォントを構築するために用いる。(C) は、そのキャラクターの多様な感情が表出する、さまざまなストーリーから成り立っている。その音声は、朗読調ではなく、感情表現など多様な発声様式の音声が多量に含まれているので、平静な感情で発声されているものと感情音声とに分類し、平静な感情で発声されている音声はボイスフォント構築用に用い、感情音声は感情音声の合成に用いる。

これらの発声は、すべて録音スタジオにおいて 16 ビット、48kHz のサンプリング周波数で DAT テープに記録し、その後、22.05kHz にダウンサンプリングし PC へのハードディスクに蓄積した。

3. ボイスフォント

筆者らは、ターゲット話者の音声からその声の特徴を表す音声辞書（ボイスフォントと呼ぶ）を生成し、合成音声へ反映できるような音声合成システム

を構築している。それは、話者再現性を向上させるために、音声合成に必要となるさまざまな特徴をターゲット話者の音声から抽出し、その抽出した特徴を音声合成器が扱える仕組みになっている。さらに、簡便でありながらも精度のよい特徴抽出を行うことが可能な音声合成器自身を利用した音声ラベリングシステムを採用している[9]。一方、ラベリングの不具合や不良発声は避け難いので、それらに対処するために、構築プロセスの中で検査と編集（修正）が行えるようになっている。その作業をできるだけ容易にし、効率よくボイスフォントを作成するために、“ボイスフォントビルダ”とよぶ GUI プログラムを提供している[10]。ボイスフォントビルダは、エラーの修正だけでなく、新しい発声や語彙の追加登録も容易に行なえるよう設計されている。

図1にボイスフォント構築フローを示す。ボイスフォント構築システムへの入力、新しい話者の音声波形と対応する正書テキストであり、出力は、波形ユニット、音素セグメント情報、アクセント句情報、各アクセント句の F0 概形、継続長制御情報などを含んだボイスフォントである。ボイスフォントビルダは、(1) 自動ラベリング、(2) 検査・編集、(3) 辞書生成、の機能を持っている。

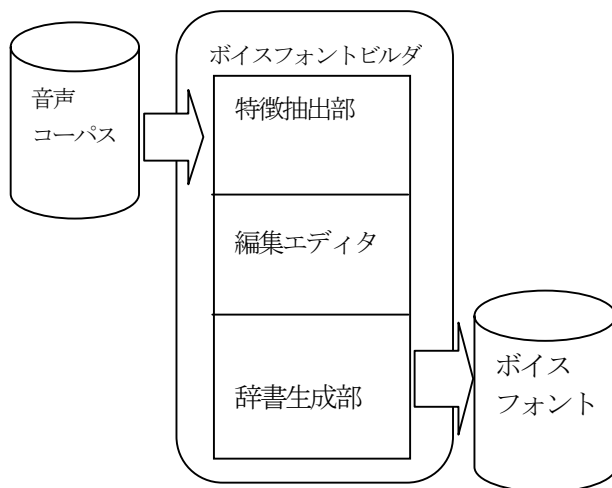


図1 ボイスフォント構築の流れ

4. 感情音声の合成

ボイスフォントによって、平静な感情では、任意文に対して話者性を有した音声の合成をできるようになったが、それだけでは、ターゲットキャラクターの音声表出のしかたを十分にカバーしているとは言えない。ターゲットキャラクターが、感情を伴って発声する音声も高品質に合成することが必要である。しかし、感情音声を対象とする場合、韻律的特徴や音響的特徴が多様であり、通常のテキスト音声合成技術を適用すると、平静な感情での音声に比べそれらの変形が大きくなり、品質が劣化することが予想される。しかし、ターゲットキャラクターの音声表出を模擬する目的では、任意文に対して指定した感情で音声を合成しなくても、ターゲットキャラクターが、感情が昂ぶったときに多用する言い回しや擬態語などの音声表現を利用し、合成する対象を絞り込むことによって、韻律的特徴や音響的特徴の変形を小さくし、高品質な感情音声を得ることが可能となる。

本手法では、ターゲットキャラクターの感情を伴った音声を録音し、個々の音声の韻律情報と波形素片をテキストラベルと結びつけ、フレーズ程度の長さに区切ってユーザ辞書に登録する。合成時には、入力テキスト中に登録したテキストラベルが出現した場合、登録した韻律情報と波形素片を用いて感情音声を合成する。さらに文の一部が感情音声として登録されていない場合に、ボイスフォントを使ったテキスト音声合成出力で補間し、合成した感情音声と継ぎ目なく接続することも可能となっている。

4.1 感情音声の特徴抽出と感情ラベリング

感情音声も、自動ラベリングと検査・修正までは、ボイスフォントビルダを使って、平静な音声と同じ処理を行い、音素ラベリングや特徴抽出を行う。しかし、辞書生成の段階では異なる手続きをとる。まず、音声資料に感情の観点からラベリングを施す。本研究では、アクセント句単位で“平静”、“喜び”、“悲しみ”、“怒り”の4つの感情ラベルを付与した。さらに、平静以外の感情を伴った音声には、感情

の強さをあらわすラベルも付与した。感情ラベリングの結果、平静な感情に分類された発声は、ボイスフォントの構築用としボイスフォントビルダの辞書生成部に入力する。一方、喜び”、“悲しみ”、“怒り”の感情を伴った音声は、感情音声としての次節に述べる登録手続きを行なう。

4.2 感情音声の登録

感情音声の韻律情報と波形情報は、アクセント句単位で、正書のテキストと関連付けてユーザ辞書に登録する。登録する項目は、(1) 正書テキストと読み、(2) 音素波形ユニットのインデックス、(3) ピッチ軌跡、(4) 音素毎の継続長、(5) 前後のアクセント句のモーラ数、(6) 前後のアクセント句のアクセント型、(7) 前後のアクセント句の平均ピッチ、である。ここで、感情音声の波形素片は便宜上ボイスフォントの波形ユニットに加えている。(2)の音素波形ユニットのインデックスとはそのユニットインデックスを指している。しかし、その音響的特徴は、平静な発声とは著しく異なることがあるので、登録した感情音声専用とし、通常の音声合成時に使われないように設定している。

4.3 感情音声の合成とシームレス音声合成

入力テキスト中に、ユーザ辞書に登録した感情音声のテキストラベルが出現した場合、登録した韻律情報と波形素片を用いて、高品質な感情音声合成される。一方、文の一部が感情音声として登録されていない場合は、それをテキスト音声合成出力で補間し、合成した感情音声と継ぎ目なく接続する。我々は、この手法をシームレス音声合成と呼ぶ。

ここで、一例をあげて説明する。ある対話アプリケーションにおいて、名前が“一郎”である人を賞賛するときに、以下の文に対する音声を合成することを考える。

いやあ、やっぱり一郎はいいやつだ。

この文のうち、四角で囲まれたアクセント句は、感情音声として登録され、それぞれ”喜び”に分類

されている。しかし、名前の部分は感情音声として登録されていないので、テキスト音声合成出力によって補う。本方式では同一キャラクタのボイスフォントを用いて合成するので、韻律的特徴や音響的特徴はまったく異なるものではなく、テキスト音声合成出力に韻律の変更を加えることとする。本手法を模式的に図2に示す。まず、補間する部分を含めた文全体のテキストをテキスト音声合成エンジンに与え、ボイスフォントを使って合成音声出力を得る。そして、補間したい部分を切り出し、感情音声と接続するが、通常、感情音声とピッチレンジが合わないことがあるのでピッチレベルを変更する。この処理を行うため、感情音声のアクセント句の情報とし

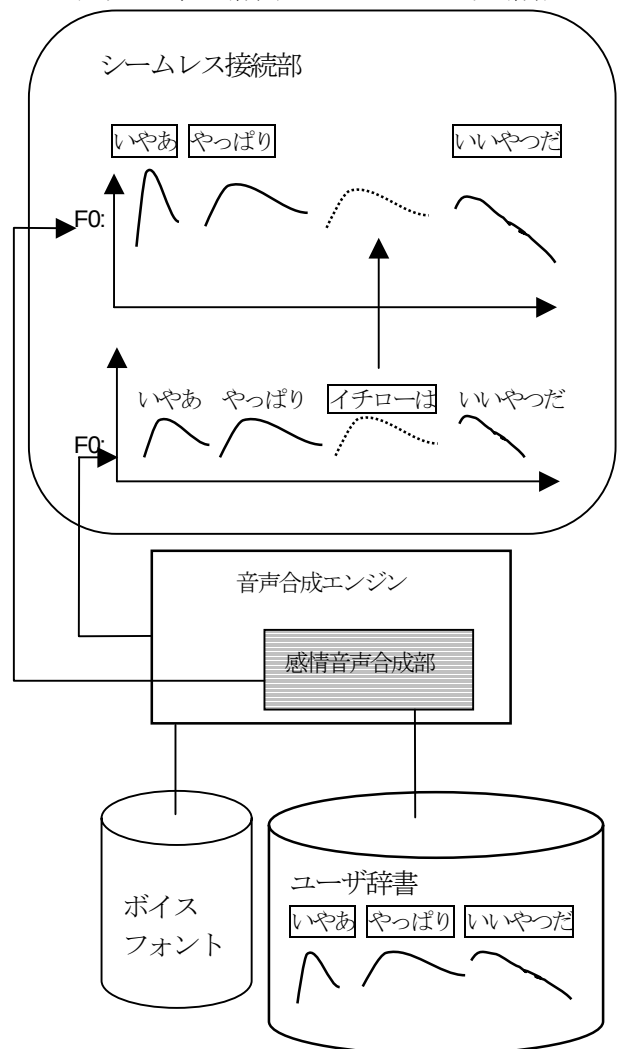


図2 シームレス音声合成手法の模式図

て、前後のアクセント句の平均ピッチが登録されており、それに合うように、F0 パターンをシフトした後、感情音声の合成出力と接続する。

この例では、ピッチのシフトについて説明したが、場合によっては、音素継続長やアクセントの強さなどの変更が必要となることも予想される。

5. F0 軌跡の再現性

ここでは、感情音声の F0 軌跡の再現性について考察する。感情音声のピッチパターンの登録する場合、各音素中心でのピッチで代表させて登録する方法と、原波形のピッチマークをそのまま利用する方法が考えられる。これら 2 つの方法で合成した感情音声と原音声と比較する主観評価テストを予備的に行ったところ、代表ピッチを使った場合、元の音声のニュアンスとは異なって聞こえることがあるとい

う意見が多かった。図 3 にその例を示す。この発声は“怒り”に分類されている。図 3 の (a) は原音声波形で、(b) はその F0 軌跡である。F0 軌跡は、音声波形のピッチマーク間隔の逆数をプロットしたものである。図中の○印は、代表ピッチである。(c) は、その代表ピッチを使って合成した音声の F0 軌跡である。0.4sec 付近までは、よく再現されているが、0.5sec 以降はかなり異なっている。原音声と聴き比べてみると、語尾のニュアンスが違って聞こえる。一方、(d) は、原音声のピッチマークを用いて合成した感情音声の F0 軌跡である。このようにマイクロプロソディーも含めて再現される。以上のような結果から、感情音声のピッチパターンを再現するには、音素毎に 1 点で代表させるのでは不十分であり、原音声のピッチパターンを、そのまま登録するのがよいと考えられる。

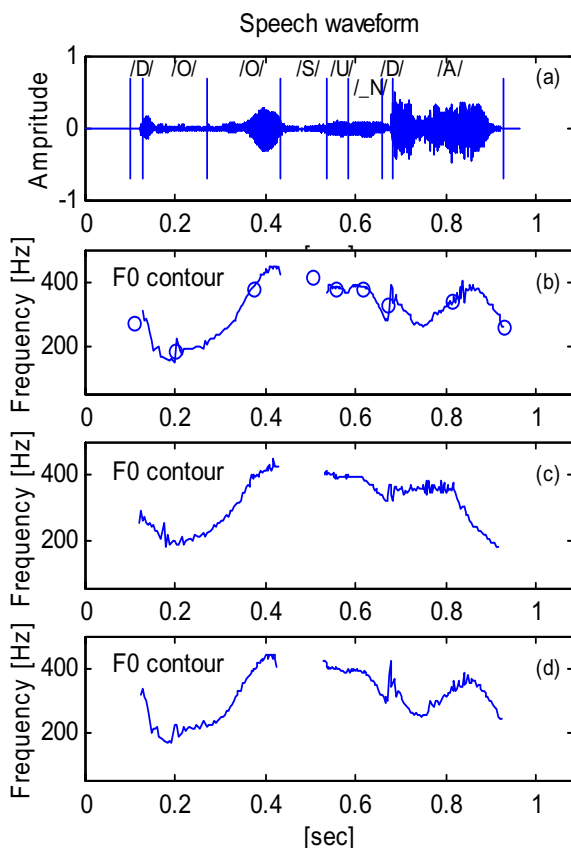


図3 F0 軌跡の再現性。怒りの感情で「どーすんだ」と叫んでいる発声。(a) 音声波形 (b) 原音声の F0 軌跡 (c) 点ピッチを使った合成音の F0 軌跡 (d) 原音声のピッチマークを使った合成音の F0 軌跡

6. デモンストレーション・プロトタイプ

以上の手法を適用したデモンストレーション・プロトタイプを紹介する。それは、感情音声合成機能と音声認識機能および対話機能を搭載した顔型ロボットである。その顔型ロボットは PC の RS-232 ポートに接続され、PC には、感情音声合成できる発話機能、音声認識機能、対話制御機能、顔制御機

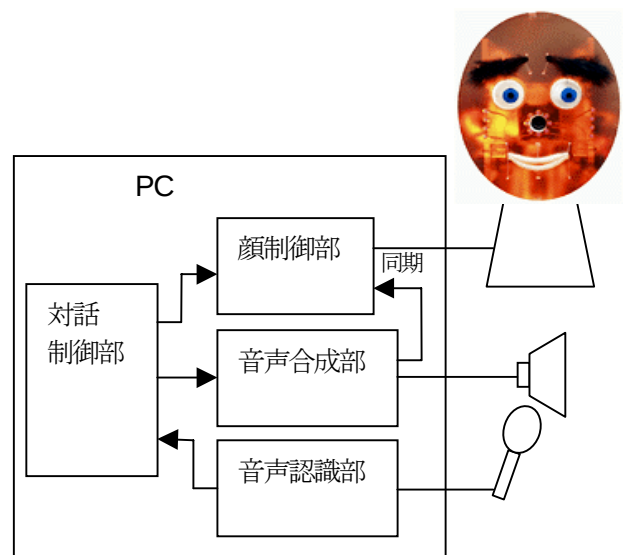


図4 デモンストレーション・プロトタイプの構成

能が搭載されている。その構成は図4のようになっている。ボイスフォントおよび感情音声の合成には2章で述べた音声資料を用いた。感情音声としては、“喜び”と、“悲しみ”または“怒り”、の感情音声を40個ずつ選んで登録した。対話シナリオは、簡単なQ&Aである。顔型ロボットは、ユーザに合成音で出題し、ユーザは、それに音声で回答する。ユーザの音声の認識結果が対話制御部に送られ、もし正解であれば、“喜び”の感情を伴った音声でユーザを誉め称える。もし不正解であれば、“怒り”または“悲しみ”の感情を伴った音声でユーザを叱咤する。感情音声は、登録されているものの中からランダムに選ばれて合成されるが、正解数あるいは不正解数が増えると、感情の強度が強いものが合成されるように作られている。

このプロトタイプでは、ごく少数の感情表現フレーズしか登録していないので、ターゲットキャラクターの音声表出を模擬するというねらいを達成しているとは言えないが、対話を魅力的なものにするという、感情音声の合成の効果を示すものとなった。

7. まとめ

ターゲットキャラクターの音声表出を模擬する音声合成器をねらった、感情音声の合成手法を提案した。本手法によれば、高品質な感情音声の合成が可能である。しかし、適用例では、わずかな感情表現しか登録していないので、今後、感情表現の登録数を増やして、音声表出の模擬の度合いをあげていきたい。

今回は、感情音声を“喜び”、“悲しみ”、“怒り”に分類しており、表層的で大まかな分類にとどまっている。今後は、ターゲットキャラクターの音声表出に適合した分類のしかたを検討したい。また、感情音声を使ったボイスフォントを作成することも試みたいが、それには、感情音声データをどのように分類してボイスフォントビルダに入力すべきかなど、検討すべき事項は多く残されている。

謝辞

顔型ロボットの制御および音声認識と合成機能を使った対話アプリケーションを作成された日本アイ・ビー・エム東京基礎研究所の長尾確研究員と学生研究員の皆様に深謝いたします。顔型インターフェースは、IBM アルマデン研究所センターのBlueEyesプロジェクトのために作られた。

参考文献

- [1] Iida, A., Iga, S., Higuchi, F., Campbell N., and Yasumura, M., “Acoustic nature and perceptual testing of corpora of emotional speech,” *Proc. of ICSLP 98*, Sydney, 1998: 1559-1562.
- [2] Murray, I. R., Arnott J. L., “Implementation and testing of a system for producing emotion-by-rule in synthetic,” *Speech Communication*, 16, 1995: 369-390.
- [3] Cahn, J. E., “Generating expression in synthesized speech,” MIT Media Laboratory Technical Report.
- [4] 小林, 新美, “音声の感情を反映する韻律情報制御方式について”, 日本音響学会秋季講演論文集, pp.233-234, (1993)
- [5] 片江, 木村, “感情音声合成における声質と韻律の制御の効果”, 日本音響学会秋季講演論文集, pp.187-188, (2000)
- [6] 丸本, キャンベレ, “波形接続型音声合成方式における発話様式の制御”, 日本音響学会春季講演論文集, pp.213-214, (2000)
- [7] 篠崎, 阿部, “テンプレート・テキスト音声合成方式とCGキャラクターエージェントWebMessengerを用いたコンテンツ製作”, 電子情報通信学会総合大会, SD-4-3, pp.267-268, (2000)
- [8] Saito, T., Sakamoto, M., “A method of creating a new speaker's VoiceFont in a text-to-speech system,” *Proc. of ICSLP 2000*, Beijing, Vol.2, 2000: 771-774.
- [9] 斉藤, 阪本, “テキスト音声合成を利用した音素・韻律統合ラベリングシステム”, 信学技法, SP99-88, pp.17-24, (1999)
- [10] 斉藤, 阪本, “ボイスフォントビルダー -テキスト音声合成における音声辞書作成ツール-”, 電子情報通信学会総合大会, SD-4-2, pp.265-266, (2000)