発話文のモデル化と感情豊かな音声合成

山本 隼佑 †　　中林 知子 †　　鈴木 朋子 ††　　田村 直良 †††

† 横浜国立大学大学院 環境情報学府
†† 跡見学園女子大学 短期大学部
††† 横浜国立大学大学院 環境情報研究院
{yamamoto,tam}@tamlab.ynu.ac.jp

1　はじめに

本稿では文の発話をモデル化し, 機械の合成音声の発話を感情豊かにする手法に付いて述べる.

情報化社会の発展に伴い音声合成の技術が発展し, 様々な場所で機械の合成音声が使われるようになった. しかし, 介護ロボットやエンターテインメントロボットなど人と深いコミュニケーションをとる場合, 無機質な音声では人間同士のような感情豊かなやり取りは難しい. また, 機械の無機質な対応は, 聞き手の人間に対しストレスを与える事も多い. また, 名前の呼び掛けなど意味的に中立的な発話の場合, 発話のニュアンスが話し手の感情を表している事がある.

そこで, 発話文の音響パラメータを操作することで無感情な機械の合成音声の発話を感情豊かにすることを目的とする.

まず感情付与の対象を限定する. 次に感情を込めた発話の音声波形をそのまま用いる [5] のは膨大な音声データを必要とするため, 対象の発話文を「品詞」, 「モーラ数」, 「アクセント位置」による分類でモデル化し記憶するデータを軽量化する. モデルに基づいて感情を込めた音声を採取し, 音響分析をして音響パラメータの変化率を算出してデータベース化する. 音声合成時には入力文を形態素解析し「品詞」,「モーラ数」,「アクセント位置」を算出し, データベースから音響パラメータの変化率を求め, オリジナルの「平静」の音声の音響パラメータを調整することにより, 「感情豊かな音声合成」を実現する.

2　発話文のモデル化

2.1　感情付与の対象文の選定

表現したい感情が適切に伝わるということを主眼におくと, 文の中でも特に感情が表れている文に焦点を当て, その文の感情を表す音声物理量を修正することで全体として有効に感情を伝えることが出来ると考えられる. そこで特に感情が現れる文に何か傾向があるかを日常的な発話文として漫画「ちびまる子ちゃん」 [4] を用いて調査を行った. 過半数の被験者が感情が表れているとした文 91 文中, 感動詞だけの文が 44 文, 名詞一語からなる文が 28 文を占め, こうした文に感情が表れているという結果を得た. 感動詞は表記方法や発話した時のモーラ数が曖昧であるといった問題点があるのに対し, 名詞は語形の変化が殆んど無く, 扱いやすいといった点から名詞を中心とした文に感情を付与する事で, 文章全体を感情豊かに発話させる手法を考える. 本研究では感情付与の対象を以下の様に定義し, これを名詞一語文と定義する.

- 名詞が 1 語含まれる文
- 述語を伴わない文
- 名詞およびそれを修飾する語句が一つつけられた文

形態素の分類手法は形態素解析システム茶筌 [6] の定義に基づく.

2.2　基本感情の選定

付与する感情を選定するため, 日本語語彙大系」 [1] の中で属性が「感情」である 53 語を項目に用いて調査を行い, 因子分析（主因子法, promax 回転）を行うことで結果を整理する. いずれの因子にも負荷量が低い項目, 複数の因子に負荷量が高い項目を除き繰り返し分析を行った結果, 4 因子 12 項目にて安定した因子が抽出され, この 4 因子をそれぞれ「平静」,「怒り」,「楽しさ」,「悲しみ」とし, 基本感情と定義する.

2.3　発話文のモデル化

名詞一語文に出現するすべての単語の感情ごとの音響パラメータの変化率を, 作成し記憶させる事は非常に困難である. そこで, 出現する単語を品詞の種類, モーラ数, アクセントの位置が同じ場合, 感情ごとに同じ韻律の変化をすると仮定し, 以下の手順

でグループ化し, 記憶させる変化率のデータの軽量化を図る.

まず出現単語を品詞の種類に分類する. 名詞一語文に出現する品詞は名詞, 接頭辞, 接尾辞, 形容詞, 連体詞, 格助詞, 副助詞, 終助詞である. また名詞に関しては名詞 1 語のみの文と前後に他の品詞が接続した場合で韻律の変化が異なると考え, 接続情報についても分類する.

次に品詞ごとにモーラ数で分類する. モーラ数の多い単語のある名詞, 形容詞は出現頻度を考慮して上限を定め, これを越える単語は対象外とする.

最後にモーラごとにアクセント位置で分類する. アクセントのの分類は音声合成 API（Application Program Interface）「FineVoice」[1]内部のアクセント辞書に基づく.

3 感情音声の音響分析手法

3.1 音声採取実験

本研究では合成音声の自然性を重視するため, 実際に人間が感情を込めて発声した音声を音響分析するという手法を取る. 名詞一語文の一般化と基本感情に基づいて, すべてのグループをモーラするよう発話文を作成し, 演劇経験がある男性 2 名に発声させる. 録音機材には DAT(Sony TCD D-100) を用いた. 48kHz サンプリング,16bit 量子化の設定にて録音を行う.

3.2 音声物理量の抽出

音声採取実験にて採取した音声から音声物理量を抽出し, 音響分析を行う. 音声物理量は基本周波数, パワー, 発話時間の 3 種類を抽出した. 基本周波数抽出は Praat [2] を使用し, パワーの抽出には wavesurfer-164-win [3] を使用する. 発話時間は人手によってモーラ単位でラベル付を行う.

3.3 音声物理量の補正

音声物理量の中でも得に基本周波数は発話内に無声音（調音するときに声帯の振動を伴わない音. 日本語では「p」,「t」,「k」,「s」の子音）が含まれていたり, 発話中に急激に音の高さが変化したり, 録音中に雑音が含まれたりすると周波数がうまく採取できない場合があるので, これを補正しなくてはならない. 以下の手順を用いて, 欠落した基本周波数, 音の大きさの値の補正を行う.

[1]NTT-IT の音声合成編集インターフェイスである. 任意の漢字仮名混じり文及び仮名アクセント文を自動的に合成音声に変換するものである. 作成した合成音を聞くための,DA 機能も有している.

1. **直線による補正**

   まず, 欠落したデータを直線によって補完する. 時間 x_i から x_j にかけて測定値が欠落していた時, その前後のデータ $(x_{i-1}, y_{i-1}),(x_{j+1}, y_{j+1})$ から欠落した間の傾き $a = \frac{y_{j+1}-y_{i-1}}{x_{j+1}-x_{i-1}}$ を算出出来る. 測定値の欠落した時間 x_k ($x_i \leq x_k \leq x_j$) における補完された値は $y_k = y_{i-1} + a(x_k - x_i)$ となる.

2. **n 次関数への近似よる補正**

   直線による補正のみでは曲線であった波形が不自然になる. また採取したデータには値のばらつきがある場合があるため, これを音声合成に採用すると不自然な音声になる可能性がある. そこで n 次関数の曲線に近似することで直線の不自然さや値のばらつきを補正する. 近似には最小自乗法を用いる.

   ある時間 x_i における測定値 y_i の組 (x_i, y_i) が M 個あるとする. n 次関数の係数を $\vec{b} = (b_0, b_1, \cdots, b_{n-1}, b_n)$ とすると n 次関数 $f_{\vec{b}}(x)$ は次の様に表す事が出来る. $f_{\vec{b}}(x) = b_n x^n + b_{n-1}x^{n-1} + \cdots + b_1 x + b_0$

   最小自乗法により, 測定値と n 次関数の残差二乗和 $U = \sum_{i=0}^{M}(y_i - f_{\vec{b}}(x_i))^2$ が最小になるような $\vec{b}$ を求める.

   近似の次元数は, 3 次, 5 次, 7 次, 10 次で試行し, ここでは基本周波数の変化をもっとも良く近似できた 10 次を用いた. 直線と 10 次関数の近似による補正を行った結果を図 1に示す.

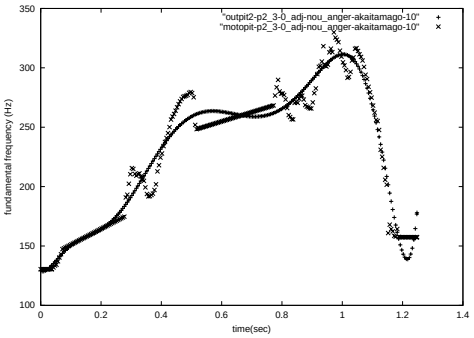

図 1: 直線と 10 次関数の近似による補正を行った基本周波数データ

3.4 音声物理量変化率のデータベースの構築

音声物理量の補正後,「平静」に対する他の 3 感情の音声物理量の変化率を求め, 名詞一語文に出現

するすべての品詞, モーラ数, アクセントに対応する音声物理量変化率のデータベースを構築する. 発話時間の変化率は「平静」に対する他の 3 感情の発話時間の変化を 1 モーラごとに百分率（%）で求め, 基本周波数, パワーの変化率は 1 モーラの時間を 10 分割し, それぞれの時間の基本周波数, パワーの変化を百分率（%）で求める.

例として図 2に名詞のデータベースの構造を示す.

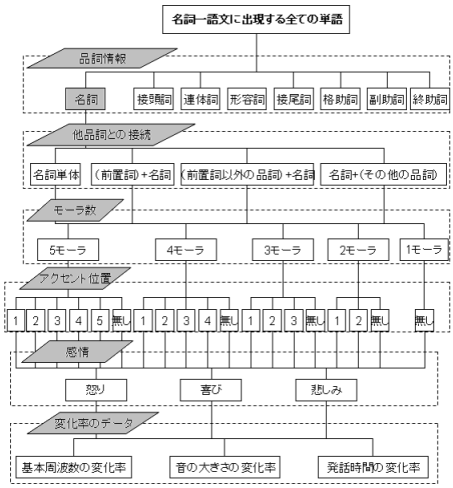

図 2: 名詞のデータベース構造

4 音声合成器を用いた感情の音声合成

4.1 入力文のテキスト解析

入力テキストを解析して品詞ごとに分割し, 品詞, モーラ数, アクセントの位置の情報を検索する.

テキストの形態素へ分割, および品詞情報の取得には形態素解析システム茶筌を用いる. アクセントの位置の取得は, 音声合成器 FineVoice 内部に組み込まれているアクセント辞書を使用する.

4.2 音響パラメータ変化率の合成

付与する感情と品詞, モーラ数, アクセントの位置の情報を音響パラメータ変化率データベースに伝え, 基本周波数, 音の大きさ, 発話時間の変化率を検索する.

最初のテキスト入力文から作成された無感情の音声の音響パラメータに, 指示された変化率を付与し, 感情音声を作成する.

入力文のテキスト解析から感情の音声合成までの処理の流れを図 3に示す.

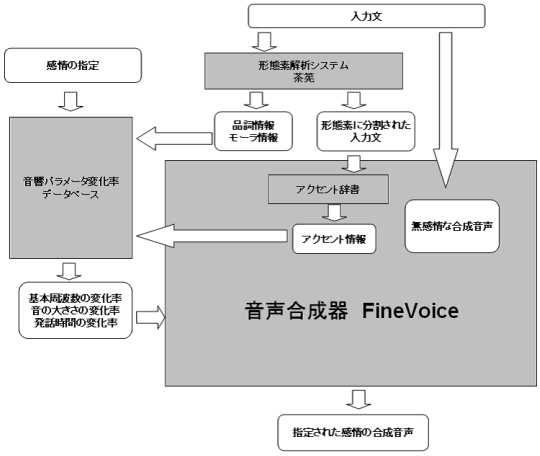

図 3: 音声合成システムの流れ

5 音声合成システムの評価と考察

5.1 合成音声の聴取実験

構築した音声合成システムに名詞一語文と任意の感情（「喜び」,「怒り」,「悲しみ」,「平静」の 4 感情のいずれか）の指定を行い, 出力された感情合成音声の評価を行う. 採取時に用いていない中立な意味の名詞一語文 80 文にランダムに感情を振り分け（各 20 文）,8 名の聴取者に 4 感情の中でもっとも近と感じる感情を判定させる.

結果を表 1に示す. 例えば「怒り」では,「怒り」を正しく知覚した確率が 51.9%,「怒り」を「平静」と誤認識した確率が 8.1%,「喜び」と誤認識した確率が 33.8%,「悲しみ」と誤認識した確率が 6.3% である.

5.2 結果の評価と考察

- 平静

  音声合成器の音声をそのまま使用しているので, 正しく知覚した確率は高くなった.

- 怒り, 喜び「怒り」と「喜び」については,「怒り」を「喜び」に,「喜び」を「怒り」に誤認識した割合が高い. この 2 つの感情は音声

表 1: 聴取実験結果（%）

感情	平静	怒り	喜び	悲しみ
平静	**91.9**	3.8	0	4.4
怒り	8.1	**51.9**	33.8	6.3
喜び	4.4	37.5	**55.0**	3.1
悲しみ	20.6	5.6	8.1	**65.6**

物理量の変化が非常に良く類似しており, 音声物理量のばらつきや欠落の補完時に若干パラメータが変化した事が感情の認識に影響してしまったと考えられる. しかし実験で採取した音声を聞いても「怒り」か「喜び」か判別することは非常に難しい. 人間は文意から相手の感情を読み取ることが多く, 単発的な名詞一語文のみから「怒り」と「喜び」の感情を聞き分ける事は困難であり, このことも誤認識の原因の 1 つであると考えられる. また, ある被験者 A は「怒り」を「喜び」と誤認識した文が 12 文,「喜び」を「怒り」と誤認識した文が 2 文であったのに対し, 被験者 B は「怒り」を「喜び」と誤認識した文が 4 文,「喜び」を「怒り」と誤認識した文が 11 文であった. これは人によって感情の認識が異なるか, または被験時の心境が影響したのではないかと考えられる.

- 悲しみ

  「悲しみ」については,「怒り」や「喜び」と比べて認識率は高かったが「平静」に誤認識されることが多かった.「怒り」や「喜び」に比べて音声物理量の変化に特徴が少ないことが, 誤認識につながったと考えられる. 原因として実際の会話での「悲しみ」を込めた発話は, 時に小さくなったり擦れたりする声であるが, 音声採取時には音声物理量を欠落無く採取するため, ある程度の音量を発話者に求めたことが, 現実の発話の特徴と異なったのではないかと推測される. また,「喜び」に誤認識された文は, 文の語尾が採取した音声とは異なってやや上がっていた. 原因は「悲しみ」の発声は語尾の基本周波数のデータが欠落することが多く, 欠落した基本周波数の関数近似による補完時に, 関数が文末を右上がりに近似をしてしまったと考えられる.

全体の結果に対する考察として, 平静を除いた 3 感情の認識率は 57.5% と高い数字では無いが,「怒り」と「喜び」の混同は上記の考察によりある程度仕方の無いものと考えると,「怒り」,「喜び」の音声物理量の付与を行った結果, 共に 90% 前後, 被験者は近い感情を知覚しており, 機械の合成音声に一定の感情付与が成功したと言える. むしろ「悲しみ」の感情音声が「平静」と誤認識された点が,「悲しみ」の音声物理量の特徴があと一つとらえられなかったと言える. 認識率の向上には音声採取の環境をより整えデータの欠落を少なくする事, 欠落したデータの補完方法を改善し, より元の波形の特徴を正確に掴む事, より多くの音声データを用い, 安定した変化率のデータを算出するといった改善が有効であると考えられる.

6　まとめと今後の課題

本研究では, 機械の合成音声に感情を付与することを目的とし, 会話中で良く感情の込められる文として名詞一語文を対象に,「怒り」,「喜び」,「悲しみ」の 3 感情の付与を行った. 名詞一語文に出現する単語を「品詞」,「モーラ数」,「アクセント」に分類することで出現単語を一般化し, 採取した音声から「基本周波数」,「音の大きさ」,「発話時間」の音声物理量を抽出した. データの欠落は直線と n 次関数による近似で補い,「平静」に対する 3 感情の音声物理量の変化率を求め, 音声合成器の無感情な音声に付与した.

合成した感情音声の聴取実験の結果, 音声物理量の変化率の付与は一定の感情付与の効果をもたらした. しかし,「怒り」と「喜び」の区別が付きにくい事や,「悲しみ」の感情付与が聴取者に知覚されにくい問題点があった.

今後の課題は認識率の向上と, 会話文の全ての単語を対象として音響パラメータを変化させることである. 自然な感情豊かな会話文の発話には, 感情に抑揚を付けることが必要で, ターゲットの文または単語の文章中における位置や重要度により, 音声物理量の変化率を変えなくてはならない. また, 現在は話者の感情をユーザーが指定しているが, 機械が文意から話者の感情を自動的に推測出来るようにする事も必要である.

参考文献

[1] NTT コミュニケーション科学基礎研究所監修, 池原悟, 宮崎正弘, 白井諭, 横尾昭男, 中岩浩巳, 小倉健太郎, 大山芳史, 林良彦. 日本語語彙大系 ＣＤ－ＲＯＭ版, 1999.

[2] http://fonsg3.let.uva.nl/praat/. Praat.

[3] http://www.speech.kth.se/wavesurfer/. wavesurfer.

[4] さくらももこ. ちびまる子ちゃん（1）. 集英社文庫, 1987.

[5] ニック・キャンベル. 音声合成 chatr のしくみ. 信学技法, pp. 47–54. SP98-84, NOV.1998.

[6] 松本裕治, 北内啓, 山下達雄, 平野善隆, 松田寛, 浅原正幸. 日本語形態素解析システム「茶筌」version2.0 使用説明書 第二版. NAIST Technical Report, 奈良先端科学技術大学院大学 松本研究室, 1999.