感情音声の合成

小池 和仁† 斎藤 博昭†† 中西 正和††
† 慶應義塾大学大学院 理工学研究科 計算機科学専攻
†† 慶應義塾大学 理工学部 情報工学科
〒223-8522 横浜市港北区日吉 3-14-1
045-563-1141(内線 3765)
kazuhito@nak.ics.keio.ac.jp

あらまし よりよいインターフェースの実現を考えると、感情の伝達は重要である。本研究では、『怒り』『驚き』『悲しみ』『嫌悪』『喜び』の感情について、その表現を制御するパラメータとして話すテンポ、声の高さ、声の大きさを用い、それにより決定されるパラメータを用いて合成された単語音声から、各感情を知覚できるか検証した。結果として、『怒り』『嫌悪』『悲しみ』については 85%以上の高い正反応率が得られ、合成に用いたパラメータ値が妥当なものであることが裏付けられた。一方『喜び』と『驚き』については、韻律だけからその 2 つを区別することが困難であるとの仮説を得ることができた。4 感情について、フランス人にとって自然な音声を新たに作成し、同様の実験を試みたが、言語によって感情パターンが必ずしも同一でないことが示された。

キーワード
音声合成、感情、韻律

Synthesis of emotional speech

Kazuhito Koike Hiroaki Saito Masakazu Nakanishi
Department of Computer Science, Keio University
3-14-1 Hiyoshi, Kohoku-ku, Yokohama, 223-8522, Japan
+81-45-563-1141(ex3765)
kazuhito@nak.ics.keio.ac.jp

Abstract Emotional aspect plays an important role in user interface. We deal with five emotions that are “anger”, “surprise”, “sadness”, “dislike” and “joy”. We verify whether one can feel each emotion from speech of a word that is synthesized with characteristics of each emotion. Those characteristics are represented by three parameters of tempo, pitch and volume of speech. As a result, we get a correct answer with a high rate of 85% in “anger”, “dislike” and “sadness”. This result validates the value of the parameters for those emotions. On the other hand, in “joy” and “surprise”, we hypothesize that it is difficult to distinguish these two emotions only by its prosody. Preliminary experiments in which four emotions are attached to a word by a native French show that emotion patterns are somewhat language dependent.

key words speech synthesis, emotion, prosody

1. はじめに

人間は、感情を伝えるだけで相手とコミュニケーションをとることができる。また、声の調子や態度から相手の感情を読み取ることができれば、言葉では逆のことを言っていてもその真意を知ることができる。このようにコミュニケーションをとるにあたって、感情の伝達が重要な役割を担っていることは明らかである。

現在、人間とコンピュータのインターフェースとして、キーボードやマウスなどがあるがそれらは誰もが容易に扱うことのできるものではない。そこで人間の最も基本的なコミュニケーションの手段である音声を、コンピュータと人間のインターフェースとして用いることが現在試みられている。

コンピュータと、音声によってコミュニケーションをとることを考えた場合、ただ単に発声内容を伝える言語的情報の伝達だけを扱うのではなく、人間同士においても重要な役割を担っている感情の伝達が行われれば、より円滑なコミュニケーションの実現が期待される。よって、よりよいインターフェースを考えるにあたって、コンピュータによる発声者の感情理解は、必要不可欠であると考える。

2. 目的

人間はその感情を、表情や態度、声によって表すが、例えば電話で話していても相手の感情が分かるように、音声だけからもある程度の感情理解は可能である。そこで本研究では、音声に含まれる情緒性情報[1]に注目し、平静音声(何の感情も込められていない音声)との比較により、感情を表すために必要とされる特徴パラメータ値を決定した。さらに、それを用いて感情の込められた音声を再現し、合成された音声から実際にその感情を読み取ることができるかを知覚的に検証した。

3. 感情パラメータ

本研究で対象とした感情は、『怒り』『嫌悪』『喜び』『驚き』『悲しみ』の5つである。これは基本6感情の内、『恐怖』については[3]において、十分な認識結果を得ることができていなかったので、それを除いた5感情を用いた。音声において感情は、話すテンポ、声の高さ、声の大きさによって表される。よって、各々に対応するパラメータとして発話速度、基本周波数(ピッチ)、振幅値を感情を表す特徴量として用いた。ピッチについては、その時系列パターンを各感情間の比較に用いた。

3.1 特徴量の抽出

DAT-Link[7]を用いて、サンプリング周波数を24kHz、16bit量子化によりディジタル化した。発話時間は、音声波形の目視により感情ごとの長さを比較した。振幅値は、512個分のデータの2乗和の平方根をその区間の振幅値とした。窓関数にはハミング窓を用い、フレーム長を512個分のデータ、フレーム周期を64個分のデータとし、ピッチ抽出には、FFTケプストラム法を用いた。

3.2 特徴パラメータの決定

本研究では、音声サンプルとして人名を用いた。人名であれば、どの感情を込めても発声しやすいし、その語自体に感情は含まれないと判断したからである[3]。人名以外にも語自体に感情が含まれているような“あーあ”や“うまい”などについても、平静の音声と、その単語に対応する感情を込めて発声してもらった音声を収録し、それを他の人にどの感情に聞こえるか、解答してもらった。ところが、これらの音声は感情の込められた音声を正確に言い当てることが出来るが、平静の音声についてもその感情が込めれた音声として聞こえてしまった。

実際に特徴パラメータを決定するにあたっては、収録した音声の中で、各感情の認識率が高かったある発話者による音声サンプルを用いた。

4. 音声合成システム

4.1 合成法

本研究で用いた音声合成システムは、音素を合成単位とした、波形合成方式により音声合成を行うものである。保存データは、音素の波形データそのものではなく、波形の特徴となる局所的なピークを制御点としたデータを、各音素について作成し、合成に用いる。合成の際には、ピッチ周期にあわせて、制御点を時間軸上に配置し、それを余弦関数による補間を行うことにより、音声波形を生成する。これは、音声波形が時間軸上において局所的に見れば、非常に滑らかな形をしていることに基づくものである。データとして持っているのは各音素のデータだけであり、音素間のわたりの部分の波形には、その両端の音声データから制御点の移動により得られる中間波形を用いる[6]。

4.2 ピッチパターンの制御

ピッチパターンの生成には、藤崎モデル[4]を用いた。このモデルは、対数基本周波数の時間パターンが、ベースラインと句頭から句末に向かう緩やかな

下降のフレーズ成分、さらに局所的な起伏に対応するアクセント成分の和によって表されるとしたものであり、(1) 式のように表される。

$$\ln F_0(t) = \ln F_{\min} + \sum_{i=1}^{I} A_{\mathrm{p}_i} G_{\mathrm{p}_i}(t - T_{0_i}) + \sum_{j=1}^{J} A_{\mathrm{a}_j} \{G_{\mathrm{a}_j}(t - T_{1_j}) - G_{\mathrm{a}_j}(t - T_{2_j})\} \quad (1)$$

また、フレーズ成分 $G_{\mathrm{p}_i}(t)$、アクセント成分 $G_{\mathrm{a}_j(t)}$ は、各々式 (2)、(3) のように表される。

$$G_{\mathrm{p}_i}(t) = \begin{cases} \alpha_i^2 t \exp(-\alpha_i t) & (t \geq 0) \\ 0 & (t < 0) \end{cases} \quad (2)$$

$$G_{\mathrm{a}_j}(t) = \begin{cases} \min[1 - (1 + \beta_j t)\exp(-\beta_j t), \theta_j] & (t \geq 0) \\ 0 & (t < 0) \end{cases} \quad (3)$$

$F_{\min}$: 最低周波数,
I: フレーズコマンド回数,
J: アクセントコマンド回数,
A_{p_i}: i 番目のフレーズ成分の大きさ,
A_{a_j}: j 番目のアクセント成分の大きさ,
T_{0_i}: i 番目のフレーズコマンドの開始時刻,
T_{1_j}: j 番目のアクセントコマンドの開始時刻,
T_{2_j}: j 番目のアクセントコマンドの終了時刻,
α_i: i 番目のフレーズ成分の固有角振動数,
β_j: j 番目のアクセント成分の固有角振動数,
θ_j: j 番目のアクセント成分の閾値.

4.3 振幅の制御

本音声合成システムで用いた振幅制御モデルを以下に示す。このモデルにより求まる値を基本として、各音節ごとに振幅値の比率を決定する。

- 母音
  母音の振幅には標準正規分布を用いる。そして、有声子音に続く母音 (式 (4)) の場合と、それ以外の場合 (式 (5)) の 2 つに分けて考える。これは実際の音声波形を観察した結果、母音の振幅は有声子音に続く場合とそうでない場合とで特に顕著な違いが見られたからである。

$$y = 0.1 + \frac{1.0}{\sqrt{2.0\pi}} \exp\left(\frac{-(\frac{4.0}{\mathrm{T}}t - 1.0)^2}{2.0}\right) \quad (4)$$

$$y = 0.1 + \frac{1.0}{\sqrt{2.0\pi}} \exp\left(\frac{-(\frac{6.0}{\mathrm{T}}t - 3.0)^2}{2.0}\right) \quad (5)$$

  ここで T は、その母音の発話時間長である。
- 有声子音
  母音の場合と同様に標準正規分布を用いる。有声子音の振幅は式 (6) で定義する。

$$y = 0.1 + \frac{0.8}{\sqrt{2.0\pi}} \exp\left(\frac{-(\frac{3.5}{\mathrm{T}}t - 3.0)^2}{2.0}\right) \quad (6)$$

  ここで T は、その有声子音の発話時間長である。
- 無声子音
  無声子音の振幅は 0.05 から 0.11 までの間の値を任意に選択する。
- 破裂音
  破裂音には振幅が急激に変化する部分が存在する。音素のデータとして切り出す部分には、それが含まれないので、振幅が急激に変化する部分を作る必要がある。破裂音の発声される時間を始めの 40%と残りの 60%とに分け、残りの 60%は無声子音と同様に 0.05 から 0.11 までの間の任意の値を選択し、始めの 40%の部分は式 (7) で振幅を定義する。

$$a = \begin{cases} 0.05 + \frac{0.6}{0.2\mathrm{T}}t & (0 < t < 0.2\mathrm{T}) \\ 0.65 - \frac{0.6}{0.2\mathrm{T}}t & (0.2\mathrm{T} \leq t < 0.4\mathrm{T}) \end{cases} \quad (7)$$

  ここで T は、その破裂音の発話時間長である。

4.4 時間長の制御

匂坂らは、視察による音韻境界を決定する手法を用い、自然音声における音韻長の平均的性質、隣接音韻間の音韻長依存関係等を分析した [5]。ただし、感情により同じ言葉を発声していてもその発声時間が異なるので、各音節ごとの発声時間長を音節の母音部分を変化させることにより調節する。音節ごとの発声時間長の変化は、主として母音部で担われているからである。本研究では、匂坂らの分析により得られた有意味単語中における音韻長を基本の音素長として用いる。

4.5 音声合成システムへの入力

以下に、本音声合成システムに対する入力パラメータを示す。

1. フレーズコマンドの数 (I)
2. フレーズごとの音素列
3. 各音節の発話時間長の比率
4. フレーズ間の無音区間の長さの比率
5. アクセントコマンドの数 (J)
6. 各アクセントコマンドの開始時刻 (T_{1_j})

7. 各アクセントコマンドの終了時刻 (T_{2_j})
8. 各フレーズコマンドの開始時刻 (T_{0_i})
9. 各フレーズ成分の大きさ (A_{p_i})
10. 各アクセント成分の大きさ (A_{a_j})
11. 各音節の振幅値の比率
12. 最低周波数 (F_{min})

5. 知覚実験及び考察

22歳から30歳の男女20人を対象とし、合成音声から先に挙げた5つの感情を読み取ることが出来るかを調べた。本研究では、"やまもと"という4音節の人名を認識、合成のサンプルとした。

5.1 用いたパラメータについて

- ピッチパターンを制御するためのパラメータについては、自然音声の分析結果から、$\alpha_i = 0.003, \beta_i = 0.02, \theta_j = 0.9$ に固定し得ることが明らかにされている [4]。
- アクセント成分、フレーズ成分の大きさ、アクセントコマンドの開始時刻(アクセントのある音節の母音開始時点の70ms前)についても、文献[4]の中で挙げられている値を元に、まず、平静の音声を作成し、感情による違いを再現するように各値を変更した。
- フレーズコマンドの開始時刻については文献[2]より、フレーズ成分の極大点が、1番目のアクセントコマンドの開始時刻になるように設定した。
- 各アクセントコマンドの終了時刻は、次のアクセントコマンドの開始時刻までとし、フレーズの最後のアクセントコマンドであれば、そのフレーズの終わりまでとした。
- 各感情の特徴が最後の音節に特に現れていると考え、それらの音声については、最後の音節の時間長をそれ以前の音節よりも長く設定した。

以下と表1で、本研究で用いた各パラメータ値を示す。

平静 :アクセントコマンドの開始時刻を、"ま"の母音開始時点の70ms前に設定した。

怒り :他のどの感情音声よりも発話時間長を短くし、振幅値とピッチパターンの変化量を大きくした。アクセントコマンドの開始時刻は、"ま"の母音開始時点の70ms前に設定した。

嫌悪 :他のどの感情音声よりも最後の音節の時間長を長くし、語尾のピッチを上げるためのアクセントコマンドの開始時刻をアクセントのある音節の母音開始時点よりも後ろにした。1つ目のアクセントコマンドの開始時刻を"や"の母音開始時点の70ms前にし、ピッチを下げるために大きさを負の値とした。2つ目を1つ目の開始時刻の200ms後、3つ目を"と"の母音開始時点の120ms後とした。

悲しみ :振幅値とピッチパターンの変化量を小さくした。アクセントコマンドの開始時刻を"ま"の母音開始時点の70ms前とした。

驚き :アクセントコマンドの開始時刻は、"ま"の母音開始時点の70ms前とした。

喜び :振幅値を平坦ながら全体的に大きな値をとるようにした。ピッチパターンを語尾で下げるために、2つ目のアクセント成分の大きさを負の値にした。1つ目のアクセントコマンドの開始時刻を"ま"の母音開始時点の70ms前とし、2つ目を"と"の母音開始時点の30ms後とした。

表 1: 5感情に用いたパラメータ値

	平静	怒り	嫌悪	悲しみ	驚き	喜び
発話時間比率	1.0	0.3	0.8	0.8	0.5	0.5
	1.0	0.3	0.8	0.8	0.5	0.5
	1.0	0.3	0.8	0.8	0.5	0.5
	1.0	1.0	1.7	1.0	0.7	1.0
振幅比率	0.75	0.7	0.7	0.5	0.75	1.0
	1.0	2.5	0.75	0.2	1.5	1.0
	0.7	2.0	1.0	0.08	1.3	1.0
	1.3	5.0	3.0	0.08	3.0	2.5
A_{p_i}	0.43	0.43	0.43	0.43	0.43	0.43
J	1	1	3	1	1	2
A_{a_j}	0.26	0.4	−0.1 0.4 0.9	0.1	0.8	0.8 −0.1
F_{min}	120	100	100	100	135	135

$(\alpha_i = 0.003, \beta_i = 0.02, \theta_j = 0.9, I = 1)$

5.2 実験方法

実験は2種類行った。実験1としては、まず、合成した平静音声を聞いてもらい、その後ランダムに選択した3つの音声について、『怒り』『嫌悪』『悲しみ』『驚き』『喜び』の5感情のどれに感じるかを解答してもらった。実験2でも、まず平静音声を聞いてもらい、次に5感情全ての音声を聞いてもらう(順番はランダム)。そして、相対的に5つの音声がどの感情に当てはまるかを解答してもらった。

表 2: 実験 1 の結果 (%)

	怒り	嫌悪	悲しみ	驚き	喜び
怒り	75.0	0.0	0.0	16.7	8.3
嫌悪	0.0	58.3	0.0	41.7	0.0
悲しみ	0.0	0.0	100.0	0.0	0.0
驚き	0.0	0.0	0.0	50.0	50.0
喜び	16.7	0.0	0.0	25.0	58.3

表 3: 実験 2 の結果 (%)

	怒り	嫌悪	悲しみ	驚き	喜び
怒り	95.0	5.0	0.0	0.0	0.0
嫌悪	5.0	85.0	0.0	10.0	0.0
悲しみ	0.0	5.0	95.0	0.0	0.0
驚き	0.0	0.0	0.0	50.0	50.0
喜び	0.0	5.0	5.0	40.0	50.0

5.3 実験結果

実験結果を表 2 と表 3 に示す。表の見方としては、表 2 を例にとると『嫌悪』の音声を『嫌悪』と正しく知覚した確率が 58.3%,『嫌悪』の音声を『驚き』と誤認識した確率が 41.7%となる。

5.4 考察

『怒り』と『悲しみ』の感情については、実験 1 においても実験 2 においても正反応率が高く、合成音声の感情表現に用いたパラメータ値が各々の感情を表すのに必要かつ有用なものであることが確認できる。実験 1 において『嫌悪』の間違いが『驚き』に集中し、『驚き』の間違いが『喜び』に集中しているのはそれぞれに似た特徴が含まれているからであると考えられる。『嫌悪』と『驚き』については、共に語尾でピッチパターンが上がるという特徴を持ち、『驚き』と『喜び』についても、共に最低周波数が高く、全体的に高い音声であるという特徴を共有している。そのため、平静の音声との比較によってのみ感情を評価する実験 1 においては、共通した特徴を持つ正解ではない感情を解答してしまうことがあると考えられる。実験 1 と比較して、実験 2 における『怒り』と『嫌悪』の感情の正反応率は大幅に上昇した。これは、他の感情の音声と比較しながら聞くことにより、その音声が持つ特徴がよりクローズアップされたことによるものと考えられる。このことから、付加した特徴は間違ったものではなく、それをより顕著に表現することにより正反応率がよくなることが予想される。

両実験において、『驚き』と『喜び』の正反応率は低く、それぞれの感情を表現するために有効な特徴が抽出できたとは言えない。この結果については、以下のような推測をすることが出来る。

『喜び』であるか『驚き』であるかを決めかねるどの解答者も、2 つの音声の違いは分かるのだが、それぞれがどちらの感情に当てはまるかを迷ってしまうと言っていた。このことから、感情による音声の性質の違いはあるものの、日常、人間がこの 2 つの感情を感じるときは話者の態度や表情とともに感じるときが多いため、音声だけによって判断することには不慣れであり、そのために間違った解答をしてしまったことが 1 つの理由として考えられる。また、2 つの感情が同居していたために、どちらか一方の感情には決めかねるということも考えられる。一方で、この 2 つの感情を言い当てた解答者の中には、自分がその感情を込めて発声するとしたらどうなるかを考えたら、正解であったという人もいた。このことを考慮すると、この 2 つの感情を表現するために用いた特徴は間違ってはいないものの、それがあまり顕著に現れていなかったことが考えられる。

6. 結論

本研究で使用した、音素を合成単位とする波形合成方式による音声合成モジュールを用いて、ピッチパターン、振幅の時間的変化、発話時間を変化させることにより、音声に含まれる感情を表現できることが確かめられた。本研究で扱った 5 つの感情のうち、『怒り』『悲しみ』については、抽出した特徴を再現することにより各々の感情を表現することができた。『嫌悪』は、『驚き』の感情との違いをよりはっきりさせることで、正反応率の向上が見込まれ、それをもとに音声に含まれる『嫌悪』の特徴が抽出できることが考えられる。『喜び』『驚き』については今回、その特徴を抽出できたとは言えない結果となった。この 2 つの感情については、さらに分析を重ねて、各々の感情を表現するために有効な特徴を発見することが必要である。ただし、感情とは音声のみによってではなく、話者の態度や表情によっても表現されるものであるから、本研究のような実験によって表現しきれない感情も存在することが考えられる。

7. 今後の展望

今回、『喜び』『驚き』『嫌悪』の感情について (とくに『驚き』『喜び』について) 有効な特徴が抽出できたとは言えない。これらの感情を表現するために、本研究で変化させたパラメータのなかから、とくにその

感情を表すのに必要とされるパラメータを発見できれば、それぞれの感情を音声によって表現した場合、その正反応率も向上すると思われる。本研究では、1つの感情を表現するために複数のパラメータを変化させている。そこで、ある1つのパラメータを平静の音声のパラメータに変換した音声データについて聴取実験をし、表現しようとする感情が著しく失われた場合、そのパラメータはその感情を表現するために重要な役割を担っていることが考えられる。これをあまり結果の良くなかった『喜び』『驚き』『嫌悪』だけでなく、『怒り』『悲しみ』の感情についても行うことによって、より正反応率の高い音声が合成できると考える。

本研究では、"やまもと"という4音節の単語について分析を行ったが、4音節でない単語や文についても同様の実験を行うことにより、単語や文によらない感情固有の特徴パラメータを発見する。それができればそのパラメータを用いて、ある音声入力に対し、どの感情のもとで発声されたものであるかなどを認識することができる。

フランス語の韻律による感情音声合成

言語により、感情表現の方法が異なることが考えられる。そこで、フランス人にとって、自然な感情パターンを新たに合成し、その音声をフランス人6人(表4のA～F)と日本人5人(表5のA～E)に対し、その感情を正しく知覚できるかを検証した。(この実験は、フランス人留学生によるものである)。本研究で用いた音声合成システムでは、フランス語のデータを用意していないため、本研究で用いた"やまもと"という単語に『Neutral』『Joy』『Surprise』『Anger』『Dislike』の感情を付加した音声を合成した。結果を表4から表6に示す。+は、被験者が感情を正しく認識したことを示し、+/−は、感情を認識はしたがあいまいで、疑わしいことを示し、−は、認識しなかったことを示す。

表4: フランス人に対する実験

	A	B	C	D	E	F
Neutral	+/−	+	+	+	+	+
Joy	+	+/−	+	+	+	−(little disap.)
Surprise	+	+	+	+	+	+
Anger	+	−	−	−	−	−
Dislike	+	−	−(happy)	−	+	−(happy)

(disap: disappointment)

被験者の数が少なかったため、予備的な実験になったが、表4より、本研究で用いた音声合成システムによって、フランス語の韻律をある程度、表現できることが分かり、表5、表6より、言語によって感情の表現法、認識法が異なることが分かった。

表5: 日本人に対する実験(1/2)

	A	B	C
Neutral	+	discouragement	unsatisfied
Joy	+	−(neutral)	+
Surprise	+	doubt	+
Anger	surprise/+	+	−(big joy/neutral)
Dislike	happy	sad	sad

表6: 日本人に対する実験(2/2)

	D	E
Neutral	sad	+
Joy	−(anger)	+
Surprise	doubt	doubt
Anger	−(neutral)	−(happy/sad)
Dislike	doubt/sad	depressed

音声合成システムの公開

本研究で用いた、韻律パラメータの入力による音声合成システムを公開しますので、興味のある方は、メールでご連絡ください*。

謝辞

波形合成法による音声合成の実装をしてくださった鈴木 将貴さんと、合成システムで用いた振幅モデルの実装をしてくださった鈴木 啓高さんに感謝いたします。また、フランス人にとって自然な感情音声合成を行ってくれたSébastien Cagnoli氏に感謝します。

参考文献

[1] 古井貞熙: ディジタル音声処理, 東海大学出版会(1985)

[2] 藤崎 博也, 広瀬 啓吉, 瀬戸重宣: 基本周波数パターンの特徴抽出の自動化, 日本音響学会講演論文集,pp.255-256(1990-9)

[3] 平賀 裕, 斉藤 善行, 森島繁生, 原島 博: 音声に含まれる感情情報抽出の一検討, 信学技報 HC 93-66, pp.1-8(1994)

[4] 広瀬 啓吉, 藤崎 博也, 河井 恒, 山口 幹雄: 基本周波数パターン生成過程モデルに基づく文章音声の合成, 電子情報通信学会論文誌 A Vol.J72-A No.1 pp.32-40(1989-1)

[5] 匂坂 芳典, 東倉 洋一: 規則による音声合成のための音韻時間長制御, 電子通信学会論文誌 Vol.J67-A No.7 pp.629-636(1984-7)

[6] 鈴木 将貴, 中西 正和: 波形の時間的変化に基づいた日本語音声合成, 情報処理学会全国大会講演論文集 Vol.52, No.2, pp.175-176(1996)

[7] DAT-Link http://www.tc.com/

*kazuhito@nak.ics.keio.ac.jp