F-008

音声の音響的特徴を考慮した感情を付与した音声合成手法
Speech Synthesis Method with Acoustic Features of Emotion Voice

平井 秀人† 芋野 美紗子‡ 土屋 誠司‡ 渡部 広一‡
Shuto Hirai Misako Imono Seiji Tsuchiya Hirokazu Watabe

1. はじめに

近年，ロボットは日々進化しており，主に産業分野において様々な用途で利用されている．しかし，今後ロボットは産業分野のみならず，日常生活の中で人のパートナーとして活動することが求められている．そのため，ロボットはユーザーフレンドリーであることが求められている．人間は感情表現をしながら他の者とコミュニケーションを行うことが多い．この感情を含めたコミュニケーションは特に会話時に見受けられる．そこで，ロボットが人間とコミュニケーションを行う際に，ロボットが出力する音声に感情の特徴を付与させることで円滑に違和感なくコミュニケーションがとれると考えられる．

本研究では，エクマンの基本感情[1]を基に分けられた感情の内，喜び，怒り，悲しみ，驚きの 4 感情に着目し，それらの基本周波数，発話速度，音の大きさの 3 つのパラメータ(以下 3 つまとめて音響的特徴)の解析を行い，一般化する．それをロボット発話に適用することで，感情音声の実現した．

2. 解析対象のデータ

本研究では音声コーパス[2]の音声データを使用する．本コーパスは，自発対話音声と演技音声が含まれている．自発対話音声のデータ群はオンラインゲーム中のプレイヤーに音声チャットを利用させて，自然に感情が表出した音声を計 9114 発話収録したものである．演技音声のデータ群は自発対話音声から，いくつかの文を選択し，喜び，悲しみ，驚きなど 8 感情でプロの声優が発声した音声を計 2656 発話収録したものである．本研究では自発対話音声には環境音，ブレスなどの雑音が多く入っているため，演技音声(以下音声データ)のみ使用する．

3. 研究内容

本研究は，まず Praat[3]で音響的特徴を抽出し，その特徴を一般化した．また，AITalk[4]で音声合成を行う際に一般化した特徴を付与することで感情音声を作成した．

3.1 音声の音響的特徴の抽出

Praat に音声を入力すると音響的特徴とフォルマントが描画される．フォルマントとは，声道の共鳴により増幅された振動数のことで，主に母音推定に用いられるデータである．本研究ではフォルマント周波数を基に音声をモーラごとに区切り，音響的特徴を抽出する．モーラとは一定の時間的長さを持った音の単位のことである．分析する音声データは各感情 80 個である．表 1 に平静の感情の 3 モーラの周波数の抽出データ，表 2 に喜びの感情の 3 モーラの周波数の抽出データの具体例を示す．

表 1. 平静の周波数[Hz]の抽出例

単語	1 モーラ	2 モーラ	3 モーラ
リンゴ	300	200	200
バナナ	350	300	200
イチゴ	250	400	200

表 2. 喜びの周波数[Hz]の抽出例

単語	1 モーラ	2 モーラ	3 モーラ
リンゴ	500	300	400
バナナ	550	400	400
イチゴ	450	500	400

「リンゴ」の平静の周波数の，1 モーラ目を 300[Hz]，2 モーラ目を 200[Hz]，3 モーラ目を 200[Hz]と抽出したとし，以下同じモーラ数ごとに分類していき，表に格納する．同様の手順で各感情，モーラ数ごとに表に格納していく．

3.2 AITalk による感情を付与した音声合成

3.1 節で抽出した特徴をもとに各感情と平静時の音声を除算して倍率を求め，AITalk が出力する音声の音響的特徴に付与することで感情音声を出力する．以下の数式(1)で音響的特徴を一般化する．

$$X_i = \frac{1}{M} * \sum_{n=1}^{M} A_{in} \quad \cdots式(1)$$

式(1)ではモーラ数 N とそのときのデータ数が M の時，音声の i モーラ目の音響的特徴を A_{in} とし，X_i は平均を計算した結果を示す． $(1 \leqq i \leqq N)$．表 1 を用いて表すと，モーラ数 N は 3，データ数 M は 3 であり，i=1(1 モーラ目)のときは(300+350+250)/3=300[Hz]となる．同じ手順で i=2 のとき(200+300+400)/3=300[Hz]，i=3 のとき(200+200+200)/3=200[Hz]となる．同様に各感情，モーラ数ごとに平均を計算していく．最後にモーラごとに倍率を求める．表 1 と表 2 の平均を計算したものと各モーラの倍率を表 3 に示す．

表 3. 一般化結果

	1 モーラ	2 モーラ	3 モーラ
平静	300	300	200
喜び	500	400	400
喜び倍率	1.67	1.33	2.00

1 モーラ目の平静に対する喜びの倍率は 500/300≒1.67 となり，同様に各感情，モーラごとに倍率を求めていく．この倍率を AITalk が出力する音響的特徴に掛ける．具体例を表 4 に示す．

表 4. 感情音声の作成

単語「テレビ」	1 モーラ	2 モーラ	3 モーラ
AITalk	500	600	400
喜び倍率	1.67	1.33	2.00
喜び音声	835	798	800

†同志社大学大学院理工学研究科
Graduate School of Science and Engineering, Doshisha University
‡同志社大学理工学部
Faculty of Science and Engineering, Doshisha University

AITalk が「テレビ」という単語を出力した時の周波数を[500，600，400][Hz]としたとき，表 3 で求めた喜びの倍率を掛けると[835，798，800][Hz]となり，これが喜びの感情が含まれた音声となる．この倍率を AITalk が出力する音声の音響的特徴に掛けることで，感情が含まれる音声が表現できると考えられる．

4. 評価

大学生 5 人を被験者とし，平静時の音声と感情音声をランダムに聴取してどの感情が含まれているかアンケート形式で回答させた．被験者に「平静」，「喜び」，「悲しみ」，「怒り」，「驚き」の 5 つより 1 つ感情を選択してもらった．その中で聴取させた音声と被験者が選んだ感情が合致するか否かで評価した．このときのテストセットは会話コーパス[5]に含まれている文の中からランダムに選択した 24 文である．

5. 評価結果

「喜び」は 30.2%，「怒り」は 53.7%，「悲しみ」は 47.5%，「驚き」は 41.8%，「平静」は 83.7%の精度を得られた．また，付与した感情に対しての回答率を表 5 に示す．

表 5. 各感情の回答率(%)

		回答した感情				
		平静	喜び	怒り	悲しみ	驚き
付与した感情	平静	83.7	0.0	14.6	1.7	0.0
	喜び	2.6	30.2	7.8	33.6	25.9
	怒り	29.8	3.3	53.7	9.9	3.3
	悲しみ	6.6	18.9	9.0	47.5	18.0
	驚き	1.6	42.6	4.1	9.8	41.8

表 5 は例えば「平静」の感情を付与したとき 14.6%の割合で「怒り」と認識し，1.7%の割合で「悲しみ」と認識したことを示す．また音声聴取実験後にテストセットを被験者に見てもらい，どの感情のときにそのテキストの内容を発するのが適切かを評価するアンケートを各感情複数回答が出来る形式で行った．その具体例として「意地悪だよね」という文章の評価を表 6 に示す．

表 6. 目視実験回答例

平静	喜び	悲しみ	怒り	驚き
0	0	5	5	1

この例では大学生 5 人中平静のときに発するのが適切だと考えた人は 0 人，同様に喜びが 0 人，悲しみが 5 人，怒りが 5 人，驚きが 1 人選択したことを示す．

6. 考察

各感情の精度にばらつきがあった．そこで，式(1)を用いた一般化が正しかったか否かについて考察する．一番精度が低かった「喜び」に関して表 5 を参照すると「悲しみ」や「驚き」に認識されやすいことがわかる．ここで表 7 に各感情 3 モーラの発話時間の倍率の一部を示す．

表 7. 各感情 3 モーラの周波数の倍率

	1 モーラ	2 モーラ	3 モーラ
喜び	1.17	1.23	1.36
怒り	0.94	1.10	1.06
悲しみ	1.50	1.22	1.13
驚き	1.69	1.85	1.84

各感情の倍率を確認すると，「怒り」以外の特徴が類似しており被験者は両者を区別することが困難だったと考えられる．

また「喜び」の 3 モーラの音響的特徴を分析したデータの標準偏差を表 8 に示す．解析したデータにばらつきがあり，表 8 を参照すると，発話時間に関して周波数や声の大きさの標準偏差より大きいことがわかる．そのため，3.2 節で述べた式を用いるとデータの丸め込みが起こり，特徴が失われてしまうため不適切だと考えられる．

表 8. 「喜び」3 モーラの標準偏差

	1 モーラ	2 モーラ	3 モーラ
周波数	0.26	0.34	0.37
発話時間	0.89	0.56	0.66
声の大きさ	0.05	0.09	0.07

以上のことから，各感情，モーラごとのデータの標準偏差の小さい値に関しては 3.2 で示した数式(1)を用いるのは適切だと考えられるが，そうでないものは異なる手法を用いるべきだと考えられる．

次に，テキストの評価実験について考察する．表 6 で示した例では，「悲しみ」と「怒り」のときにそのテキストの内容が発せられるべきだということを示している．このテキストの音声を聴取したとき「悲しみ」と「怒り」の正答率が高く，そのほかの感情の正答率は高くなかった．この結果より被験者はテキストの内容に作用されて回答したと考えられた．今回の実験では会話コーパスからランダムに 24 文のテキストを選択した．しかし，それでは内容に作用されるテキストが含まれるため，会話コーパスに含まれる大量のテキストの中からあらかじめ人に作用されないテキストを選択させて，それをテストセットとすると精度が高くなったのではないかと考えられる．

7. おわりに

本研究では，音声データを解析し，抽出した音響的特徴を一般化し音声合成した．その際に音響的特徴を付与して感情音声を表現することができた．今回，使用したデータを解析したときに，その解析データに偏りが見受けられ，またテストセットの文章の内容に誘導されて精度が下がったと考えられた．音声データを増やす，一般化する際の用いた数式，使用するテストセットの精錬を行うことが今後の課題である．

謝辞

本研究の一部は，科学研究費補助金(若手研究(B)24700215)の補助を受けて行った．

参考文献

[1] ポール・エクマン，W.フリーセン，”表情分析入門-表情に隠された意味をさぐる”，工藤力訳，第 7 版，誠信書房，2000.
[2] 国立情報学研究所，”感情評定値付きオンラインゲーム音声チャットコーパス”.
[3] Paul Boersma，David Weenink，http://www.fon.hum.uva.nl/praat/，2015/1/12
[4] 株式会社 AI，AITalkⅡSDK，http:www.ai-j.jp/sdk2，2015/1/12.
[5] 名古屋大学，”名大会話コーパス”，http://tell.cla.purdue.edu/chakoshi/meidai-chuui.html，2008/6/30.