

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*

☆相原龍, 高島遼一, 滝口哲也, 有木康雄 (神戸大)

1 はじめに

従来, 感情音声合成の手法で一般的なのはコーパスを用いて韻律を制御する方法であったが, 文章合成には膨大な手間がかかるという問題があった. そこで岩見らは, 話者変換に用いられていた「GMM による声質変換法」を用いて, 文章単位の感情声質変換を行った [2]. また Veaux らは同様の手法を用いて韻律の変換を行っている [3]. これらは声質・韻律どちらか一方のみの変換であり, 完全な感情音声変換とはいえない.

本論文では, GMM を用いて音声の特徴量であるスペクトルと韻律を, 平静音声から感情音声のものへと変換する. 従来の感情音声変換はスペクトル, 韻律どちらか一方のみの変換にとどまっていたが, 本研究ではその両方を特徴量として変換する新しい手法を提案する.

2 GMM に基づく声質変換法

2.1 変換関数の学習

入力特徴量系列を $x = [x_0, x_1, \dots, x_T]^T$, 出力特徴量系列を $y = [y_0, y_1, \dots, y_T]^T$ とする. 静的な特徴量, x_t, y_t に対して, その前後のフレームとの関係も考慮した動的特徴量 $\Delta x_t, \Delta y_t$ を考える. 以下, 静的特徴量と動的特徴量を統合した $X_t = [x_t^T, \Delta x_t^T]^T$, $Y_t = [y_t^T, \Delta y_t^T]^T$ を用いる.

入力・出力特徴量を統合した統合特徴量ベクトルを $Z = [X^T, Y^T]^T$ とする. 統合確率モデル関数は以下のように表せる.

$$p(Z|\lambda^{(Z)}) = \sum_{m=1}^M \alpha_m N(Z; \mu_m^{(Z)}, \Sigma_m^{(Z)}) \quad (1)$$

変換関数のパラメータ $\lambda^{(Z)}$ は m 番目の正規分布において $\alpha_m, \mu_m^{(x)}, \mu_m^{(y)}, \Sigma_m^{(xx)}, \Sigma_m^{(yx)}$ である. ここで, 統合特徴量ベクトルの m 番目の正規分布における共分散行列 $\Sigma_m^{(Z)}$, 平均ベクトル $\mu_m^{(Z)}$ は, 以下のように表せる.

$$\Sigma_m^{(Z)} = \begin{bmatrix} \Sigma_m^{(XX)} & \Sigma_m^{(XY)} \\ \Sigma_m^{(YX)} & \Sigma_m^{(YY)} \end{bmatrix} \quad \mu_m^{(Z)} = \begin{bmatrix} \mu_m^{(X)} \\ \mu_m^{(Y)} \end{bmatrix}$$

パラメータは, EM アルゴリズムを用いて最尤推定で学習する.

2.2 最尤変換

入力特徴量系列を $X = [X_0, X_1, \dots, X_T]^T$, 出力特徴量系列を $Y = [Y_0, Y_1, \dots, Y_T]^T$ とする. 静的・動的特徴量系列に対する尤度は以下ようになる.

$$p(Y|X, \lambda^{(Z)}) = \sum_m p(m|X, \lambda^{(Z)}) p(Y|X, m, \lambda^{(Z)}) \quad (2)$$

ここで, m は分布系列 $m_1, m_2, \dots, m_t, \dots, m_T$ である. t フレーム目における m 番目の条件付き確率分布は以下のように表せる.

$$p(Y|X, m, \lambda^{(Z)}) = N(Y_t; E_{m,t}^{(Y)}, D_{m,t}^{(Y)}) \quad (3)$$

$$E_{m,t}^{(Y)} = [\mu_m^{(Y)} + \Sigma_m^{(YX)}(\Sigma_m^{(XX)})^{-1}(X_t - \mu_m^{(X)})] \quad (4)$$

$$D_{m,t}^{(Y)} = \Sigma_m^{(YY)} - \Sigma_m^{(YX)}(\Sigma_m^{(XX)})^{-1}\Sigma_m^{(XY)} \quad (5)$$

ここで, 式 (2) を最大にする変換された静的特徴量系列を $\hat{y}$ とすると,

$$\hat{y} = \arg \max p(Y|X, \lambda) \text{ subject to } Y = Wy \quad (6)$$

と表せる. W は静的特徴量系列を, 静的・動的特徴量系列に拡張するための変換行列 [4] である. $\hat{y}$ の推定には, 単一分布系列を近似的に用いる.

$$p(Y|X, \lambda^{(Z)}) \approx p(m|X, \lambda^{(Z)}) p(Y|X, m, \lambda^{(Z)}) \quad (7)$$

入力特徴量系列に対する尤度が最大になる分布系列 $\hat{m}$ を決定し, その後, 上の式の尤度を最大化する変換特徴量系列を決定する.

$$\hat{m} = \arg \max p(m|X, \lambda^{(Z)}) \quad (8)$$

$$\hat{y} = (W^t D_{\hat{m}}^{-1} W)^{-1} W^T D_{\hat{m}}^{-1} E_{\hat{m}} \quad (9)$$

ここで,

$$E_{\hat{m}}^{(Y)} = [E_{\hat{m}1,1}^{(Y)}, E_{\hat{m}2,1}^{(Y)}, \dots, E_{\hat{m}t,t}^{(Y)}, \dots, E_{\hat{m}T,T}^{(Y)}] \quad (10)$$

$$D_{\hat{m}}^{(Y)} = \text{diag}[D_{\hat{m}1}, D_{\hat{m}2}, \dots, D_{\hat{m}t}, \dots, D_{\hat{m}T}] \quad (11)$$

となる [4].

*GMM-Based Emotional Voice Conversion of Spectrum and Prosody. by Ryo AIHARA, Ryoichi TAKASHIMA, Tetsuya TAKIGUCHI, Yasuo ARIKI (Kobe University)

3 提案手法

GMMに基づく声質変換法を用いて、感情音声変換を行う。同一人物・同一発話内容の元感情音声とターゲット感情音声データを用いる。この2つのパラレルデータが学習音声となる。音声分析変換合成技術STRAIGHTを用いて基本周波数、スペクトル包絡、非周期成分に分離・合成する。学習時、GMMは基本周波数、スペクトル包絡についてそれぞれ別のGMMを構築する。非周期成分については使用せず、変換の際には平静音声のものをそのまま用いる。

声質を表すスペクトル包絡は、STRAIGHTで抽出する。学習時は、抽出された特徴量を、平静発話と感情発話データセットに関して、ユークリッド距離に基づくDPマッチングを行い、時間アライメントをとり、学習する。韻律の一部である基本周波数は、STRAIGHTで抽出したものを手動で音節に分割する。それぞれの音節ごと、離散コサイン変換をかけ低次5次元を使用する。特徴量は $5 \times (\text{音節数})$ の行列となる。この特徴量抽出を平静音声、感情音声の両方について行い、入力・出力特徴量を統合した統合特徴量ベクトルを学習する。

4 評価実験

4.1 実験条件

データベースとして、文献[1]で作製された感情音声データベースを用いる。日本人男性声優1人が、20単語に対して、平静を含む様々な感情をこめて録音したものである。本研究では、平静、怒り、悲しみ、喜びの4感情を用い、平静からそれぞれの感情に独立して変換する。混合数はスペクトルは64、基本周波数は32である。学習に用いるデータ数がGMMを構築するのに十分でないため、テスト時は、学習に用いた20単語平静音声から1つ選んで行った。

4.2 聴取実験

成人男女20名による聴取実験を行った。変換した単語に対して、平静、怒り、悲しみ、喜びのうち、どの感情が含まれているかを判定した。

4.3 実験結果

Table 1は聴取実験の結果を示したものである。(a)はスペクトル包絡のみを変換し、基本周波数は平静のものを合成、(b)は基本周波数のみを変換しスペクトル包絡は平静のものを合成、(c)はその両方を変換し聴取した結果である。スペクトル包絡のみの変換では、全ての感情がほとんど表現できていない。基本周波数のみの変換では、悲しみ以外の感情表現は正確とは言えない。両方を変換すると、基本周波数

Table 1 Results of Classification [%].

(a) Spectrum conversion

Tar. \ Percept.	Angry	Sadness	Joy	Neutral
Angry	5	5	5	85
Sadness	15	5	0	80
Joy	10	0	5	85

(b) F0 conversion

Tar. \ Percept.	Angry	Sadness	Joy	Neutral
Angry	0	10	25	65
Sadness	0	80	10	10
Joy	10	25	30	35

(c) Spectrum and F0 conversion

Tar. \ Percept.	Angry	Sadness	Joy	Neutral
Angry	65	0	10	25
Sadness	5	80	5	10
Joy	10	25	40	25

で感情を表現できている悲しみ以外の感情において、感情の認知率が上昇していることが分かる。

5 おわりに

実験結果から、GMMによる声質変換法で韻律とスペクトルの感情音声変換を行うとどちらか一方のみの変換にくらべて、より感情が表現できることがわかった。しかし、全ての感情を完全に表現できているとはいえない。今後の課題として、変換による音声の劣化の改善が必要である。また、今回は単語単位の変換を行ったが、データベースを作れば文章単位での変換も可能である。

謝辞 本研究には、慶應義塾大学研究用感情音声データベース[1]を用いた。

参考文献

- [1] 森山剛 ほか, “ 韻律の部分空間を用いた感情音声合成,” 情報処理学会論文誌, 50(3):1181-1191, 2009.
- [2] 岩見洋平 ほか, “ GMMに基づく声質変換を用いた感情音声合成,” 電子情報通信学会技術研究報告, SP2002-171, pp.11-16, 2003.
- [3] C. Veaux, X. Robet, “ Intonation conversion from neutral to expressive speech,” INTERSPEECH 2011.
- [4] T. Toda, *et al.*, “ Voice Conversion Based on Maximum-Likelihood Estimation of Spectral Parameter Trajectory,” IEEE Trans. ASLP, Vol.15, No.8, pp.2222-2235, 2007.