

音声の構造的表象を用いた感情・発話スタイルの推定*
☆青木史朗, 峯松信明, 広瀬啓吉 (東大)

1 はじめに

音声は人間にとって最も身近なコミュニケーションツールの一つである。音声を用いたコミュニケーションにおいて、人は音声の音響的特徴から言語情報のみでなく、様々な感情や発話スタイルなどのパラ言語情報を抽出することで円滑なコミュニケーションを行っている。

近年提案されている音声の構造的表象は、非言語情報による音声の音響的変形を写像として捉え、写像不変量のみを用いて音声を表象する方法である。即ち、音声から非言語情報を凡そ消失することができる [1][2]。より具体的には、複数の音響事象に対して全ての二事象間距離を測定し、距離行列を求める。一般に距離行列は一つの幾何学的構造を規定するため、これを音声の構造的表象と呼ぶ。この音声の構造的表象は非言語情報の影響を受けないが、発声者の方言、訛、意図、感情などの言語・パラ言語情報には容易に影響を受ける。既に音声の構造的表象に着眼したパラ言語情報に対する分析として、浜野らは感情を含む音声に対して、その構造に様々な変化が現れることを示している [3]。また、藤野らは音声の構造的表象を用いた感情・発話スタイル推定の初期的検討を行い、構造情報が有効であることを示している [4]。本研究では、音声の構造的表象に基づく単語音声認識などで適用された技術を感じ情・発話スタイル推定で用いることで、構造情報が感情推定に与える効果を実験的に検討した。

2 音声の構造的表象

音声事象をケプストラム空間内のガウス分布とし、事象間の距離をバタチャリヤ距離を用いて計測する。二つのガウス分布をそれぞれ $p_1(\mathbf{x})$, $p_2(\mathbf{x})$ とすると、バタチャリヤ距離は以下の式で表される。

$$BD(p_1, p_2) = -\ln \int_{-\infty}^{\infty} \sqrt{p_1(\mathbf{x})p_2(\mathbf{x})} d\mathbf{x}$$
$$= \frac{1}{8} \mu_{12}^T \left(\frac{\Sigma_1 + \Sigma_2}{2} \right)^{-1} \mu_{12} + \frac{1}{2} \ln \frac{|(\Sigma_1 + \Sigma_2)/2|}{|\Sigma_1|^{\frac{1}{2}} |\Sigma_2|^{\frac{1}{2}}}$$

T は転置を、 $|\cdot|$ は行列式を表す。 μ_i , Σ_i は $p_i(\mathbf{x})$ の平均ベクトルおよび分散共分散行列であり、 μ_{ij} は $\mu_i - \mu_j$ である。このとき、二つの分布に対して共通のアフィン変換 $\mathbf{Ax} + \mathbf{b}$ をかけた場合、バタチャリヤ距離はその前後で不変である。その結果、距離行列である構造的表象はアフィン変換に対して不変であり、およそ非言語情報に依存しない不変的な構造となる。音声の構造的表象において、 $\mathbf{x}$ に $\mathbf{A}$ を掛ける演算は構造の回転(声道長の差異など)として近似され、 $\mathbf{b}$ を加える演算は構造のシフト(音響機器特性の差異など)として近似

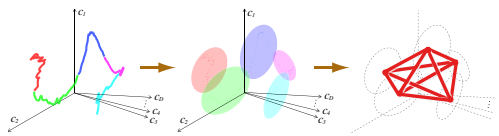


図 1 一発声の音声の構造化

表 1 音響分析条件

サンプリング	16bit / 16kHz
窓	窓長 25msec, シフト長 10msec
パラメータ	MCEP(1~12 次元)
音声事象分布	単一ガウス分布
分布推定方法	MAP with Variational Bayes
状態数	15

されることとなる¹。

一発声の音声の構造化は、発声から音声事象の分布系列を得た後、任意の二分布間距離(即ち距離行列)を求めることで行われる(図 1)。この距離行列のうち、上三角成分を「構造ベクトル」と定義する。

3 感情・発話スタイル推定

3.1 使用したデータ

音声データとして、劇団経験者の男性 2 名、女性 6 名による様々な発話スタイルの連続発声音声を用いた。発話内容は「あいうえお」、「こんにちは」、「ありがとう」の 3 種類である。また、発話スタイルは以下の 15 種類である。

(A) 震えるほどの悲しみ, (B) 押し殺した怒り, (C) 仕方なく, (D) 恥じらい, (E) ささやき, (F) 恐怖, (G) 感謝, (H) 楽しく・うきうき気分, (I) ごまかし, (J) ため息, (K) 自慢, (L) 明確, (M) 激しい怒り, (N) 思いきり, (O) 大袈裟な感謝。

音声データは各発話スタイルにつき 5 回発声されている。よって、音声データは各発話内容につき合計 600 発声(8 名×15 種類×5 回)あり、全体で 1800 発声がある。また、音声の収録は見本音声を真似ることで行っている。この時、女性話者 1 名の音声を見本音声とし、見本話者も含めた 8 名の話者全員が見本音声を収録前に毎回聞き、可能なかぎり真似てもらった。

3.2 音響分析条件

表 1 に示す音響分析条件で各発声から構造ベクトルを求めた。分布系列化は一発声に対する MAP-based な HMM 学習を通して行っている。

3.3 識別器

感情・発話スタイルの推定方法としてサポートベクターマシン(SVM)を用いた。SVM として R 言語の kernlab パッケージの関数 ksvm() を使い、カーネル関数をガウシアンカーネルとした。特徴ベクトルとして

¹なお、バタチャリヤ距離は非線形変換にも不変である。また、任意の可逆かつ連続な写像に対して不変な事象間距離はバタチャリヤ距離を一般化した f-divergence のみである

* Emotion and speaking style estimation using speech structure. by AOKI, Shiro, MINEMATSU, Nobuaki and HIROSE, Keikichi (The University of Tokyo)

表 2 特徴ベクトル

名称	次元数
構造ベクトル (Structure Vector:SV)	105 次元
構造ベクトル+韻律情報 (SV+韻律)	201 次元
構造ベクトル+ MFCC(SV+MFCC)	393 次元
IS09	384 次元

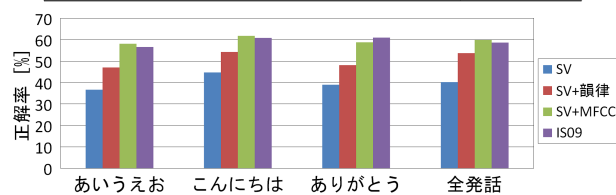


図 2 感情推定正解率

構造ベクトルを用いることにより、感情・発話スタイルの推定を行なった。

3.4 特徴量

評価実験に用いる特徴ベクトルは表 2 に示す通りである。比較する特徴量として Björn Schuller らにより提案された特徴量セットである The INTERSPEECH 2009 Emotion Challenge feature set(以下, IS09)を用いた [5]。IS09 は MFCC(1~12 次元), パワー, F0, フレームベースのゼロ交差率, ハーモニック・ノイズエネルギー比の 16 種類とそのデルタ成分について, 最大値, 最小値, レンジ, 最大値・最小値のあったフレームナンバー, 平均, 線形近似したときの傾き・オフセット, 線形近似と実際の波形の差, 標準偏差, 尖度, 歪度の 12 種類が特徴量として計算される。よって, 一つの音声に対して IS09 の特徴ベクトルは $16 \times 2 \times 12 = 384$ 次元となる。また, 構造ベクトルと韻律情報を併せて用いた場合, 構造ベクトルと MFCC を併せて用いた場合についても実験を行なった。韻律情報としては IS09 のうちの MFCC を除いた特徴量, MFCC は IS09 のものとした。

3.5 評価実験 1

発話内容を 1 種類に固定し, 評価データを 1 名の発声, 学習データを残し 7 名の発声として推定実験を行い, クロス・バリデーションにより最終的な正解率を求めた。また, 3 種の発話内容を用いた場合(全発声)についても同様の実験を行なった。正解率を図 2 に示す。

図 2 より, SV+韻律と SV+MFCC を見ると構造ベクトルのみである SV と比較して正解率が大きく向上している。特に SV+MFCC は比較手法である IS09 よりも正解率が高くなっている。これより, 構造ベクトルを MFCC や韻律情報などと併せて用いることが効果的であることがわかる。

3.6 評価実験 2

音声の構造的表象は非常に強い不変性により効果的に話者性を取り除くことが可能となるが, それと同時に, 言語的・パラ言語的に異なる二発声を同一と見なすことが容易に起きる。この強すぎる不変性は認識性能の低下を引き起こすため, 構造的表象を用いた音声認識では特徴量を分割することで強すぎる不変性に対する制約を導入している [2]。

今回, 分布系列推定の後に特徴量分割を適用した。

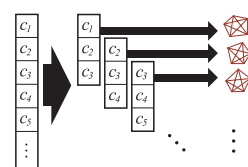


図 3 特徴量分割

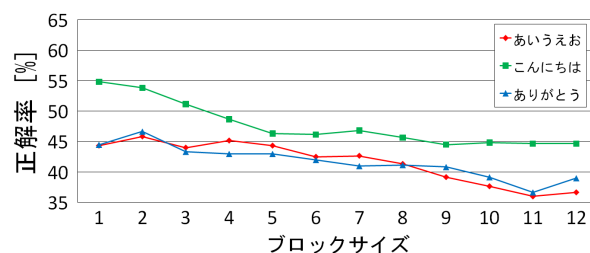


図 4 感情推定正解率 (特徴量分割)

分割前の時点で分布系列は MCEP12 次元ガウス分布として表現されている。MCEP に対してブロックサイズ w で特徴量分割を行い, それぞれの部分特徴量ベクトルを用いて構造化を行う (図 3)。

実験では, ブロックサイズ w を 1~12 と変化させて構造ベクトルを求め, それらの特徴ベクトルとして実験を行なった。この時, それぞれの構造ベクトルの次元数は $105 \times (13 - w)$ 次元となる。即ち, $w = 12$ は図 2 の SV と同等である。正解率を図 4 に示す。

図 4 よりブロックサイズが小さい程, 正解率が高いことがわかる。これより, 特徴量分割が感情推定に対して有効であることがわかる。今後は, 特徴量分割した構造情報とその他の特徴量 (韻律情報, MFCC) を併せて用いた場合についても検討したい。

4 おわりに

音声の構造的表象の情報を MFCC や韻律情報などの他の特徴量と併せて用いることと, 特徴量分割が感情推定に対して有効であることがわかった。

参考文献

- [1] N. Minematsu, "Mathematical evidence of the acoustic universal structure in speech," Proc. ICASSP, pp.889-892, (2005).
- [2] 峯松他, "音声に含まれる言語的情報を非言語的情報から音響的に分離して抽出する手法の提案 ~ 人間らしい音声情報処理の実現に向けた一検討 ~ (招待論文)," 電子情報通信学会論文誌, J94-D, 1, 12-26, 2011.
- [3] 浜野他, "音声の分節的特徴に着眼したパラ・非言語情報推定に関する実験的検討," 電子情報通信学会音声研究会, SP2003, pp.25-30 (2004).
- [4] 藤野他, "音声の音響的普遍構造に着眼したパラ・非言語情報推定に関する実験的検討," 音講論 (春), pp.59-60, 2005.
- [5] B. Schuller *et al.*, "The interspeech 2009 emotion challenge," Interspeech 2009, pp.312-315 (2009).