5S-10

ベイジアンネット混合モデルを用いた感性ロボットのための対話者感情の推定法

杉野 良樹　加藤 昇平　伊藤 英則

名古屋工業大学 大学院 工学研究科 情報工学専攻

1　はじめに

近年，多くのエンタテイメントロボットが開発されており，人間との円滑なコミュニケーションを目的としたロボットの研究が盛んに行なわれている [1]．ロボットと人間とのより豊かなコミュニケーションのためには，お互いの感情や情動を把握する必要がある．その実現には，人間が後天的に学習し獲得している「対話者感情を理解する知能」の推論モデルをロボットに持たせることが必要であると考える．そこで本研究では，感性会話ロボット Ifbot（図 1）の対話者感情推定手法として発話音声と言語情報を用いたベイジアンネット混合モデルを提案する．

図 1: Ifbot

2　Bayesian Network

ベイジアンネット (BN) は，複数の確率変数の間の定性的な依存関係を非循環有向グラフ (DAG) により表現し，個々の変数の間の定量的な関係を条件付確率で表した確率モデルである [2]．確率変数をノードとし，変数間に確率的依存関係が強いと判断される場合に対応するノード間に有向リンクを付ける． 依存関係を確率的相関と同一視した場合，N 個の確率変数 $(X_1, ..., X_n)$ の同時確率分布 P は次式で表現される．

$$P(X_1, ..., X_n) = \prod_{i=n}^{n} P(X_i|Pa(X_i)). \qquad (1)$$

ここで，$Pa(X_i)$ は確率変数 X_i の親ノードを表す．式 (1) は，各ノード X_i が $Pa(X_i)$ にのみ依存し，X_i から辿って到達できるノードを除いた他のノードとは条件付独立となることを表している．

親ノードがある状態 $Pa(X_i) = x(x$ は親ノード群の各値で構成したベクトル) のもとでの n 通りの離散状態 $(y_i, ..., y_n)$ を持つ変数 X_i の条件付確率分布は $p(X_i = y_1|x), ..., p(X_i = y_n|x)$ となる．これを各行として， 親ノードがとりうる全ての状態 $Pa(X_i) = x_1, ..., x_m$ のそれぞれについて列を構成した表の各項目に確率値を定めたものが X_i についての条件付確率表 (CPT) である．これにより，確率変数間の確率的な依存関係がモデル化できる．ベイジアンネットを用いて知識をモデル化することで，知識の記述量・計算量を大幅に削減される．また部分的な証拠からでも確率的に推論できる長所を持つ．このため本研究では，ロボットに搭載する感情推定のための知識モデルとして効率とロバストを得ることが可能なベイジアンネットを応用する．

3　感情推論器の学習

3.1　音声関連ベイジアンネットワーク

本研究では，対話者が発話した音声から対話者感情を推定する為に音声に対する感情推定知識をベイジアンネットワークとしてモデル化した．本節では，モデル構築の流れについて概説する．

3.1.1　音声資料

使用する音声資料は，感情表現がなされている必要がある．本研究では TV ドラマ，映画，それに準ずるものより女優，俳優が感情を込めて発話したフレーズを抽出し「怒り」「嫌悪」「悲しみ」「恐怖」「驚き」「喜び」の 6 種類 (以降，6 感情) に分類した．それらの中から，聴取実験により感情が適切に表現されていると判断された音声資料をサンプルデータとする．

3.1.2　特徴量の抽出

音声は，3 つの要素 (韻律，音質，音韻) から成り立っている． この中で，韻律的特徴が人間の感情表現に最も関連することが過去の様々な研究から明らかになっている (例えば文献 [3])．そこで本研究では，音声資料から振幅構造を反映する「短時間パワー」(PW)，ピッチ構造を反映する「基本周波数」($F0$) 及び時間構造を反映する「1 モーラあたりの発話継続時間」(Tm) をそれぞれ計測する．PW 及び $F0$ に関しては，平均，最大，最小，標準偏差を抽出した．このとき，短時間分析におけるフレーム長を 23ms（250 samples），フレーム周期を 11ms とし，窓関数として Hamming 窓を使用した．以上により 9 個の特徴量を音声関連ベイジアンネットワークの確率変数とする．

3.1.3　モデルの構造決定

本研究では，音声資料より抽出した各音声特徴量を適当な量子化数で量子化し学習データとする．学習データに含まれる目標属性 (6 感情) と属性 (音声韻律特徴) との間の依存関係を表現するために属性間の結合とその強さ (CPT) を学習することでベイジアンネットワークの構造を決定する．学習方法として，本研究では情報理論的妥当性がありデータへの過度なフィットを回避することで予測精度の高いモデルが学習可能できる BIC(Bayesian Information Criterion) に基づくモデル選択手法を採用する．M をモデルとし，$\hat{\theta_M}$ を M を表すパラメータ，d をパラメータ数とすると M の評価値 $BIC(\hat{\theta_M}, d)$ は次のように定義される．

$$BIC(\hat{\theta_M}, d) = \log_{\hat{\theta_M}}^{N} P(D) - \frac{d \log N}{2} \qquad (2)$$

*Bayesian-Mixture-Based Inference of Dialogist's Emotion for Sensitivity Robots, Yoshiki Sugino, Shohei Kato, and Hidenori Itoh, Department of Computer Science and Engineering, Graduate School of Engineering Nagoya Institute of Technology, Gokiso-cho, Showa-ku, Nagoya 466-8555, Japan.

表 1: 言語関連 BN 学習データの一例

$WORD$	$INFO1$	$INFO2$	$GRAM$	$EMOT$
楽しい	快	興奮	肯定	喜び
ひどい	不快	興奮	疑問	怒り
ごめんなさい	不快	鎮静	肯定	悲しみ
⋮	⋮	⋮	⋮	⋮

ここで D は学習データ，N は D のデータ数を表す．$\hat{\theta_M}$ は最尤法により求めた．D が部分観測の場合には EM アルゴリズムを用いて推定し CPT を補間する．本研究では，BIC が最大となるモデルを求めこれを対話者の感情推定のための知識としてロボットに与える．BIC を最大にするモデルの探索には K2 アルゴリズムを用いた．K2 アルゴリズムではノード間の親子順序を事前知識として与えることで探索空間を制限することが可能だが，ここでは音声の振幅 (PW)，ピッチ ($F0$) ，及び時間 (Tm) の 3 グループに分けグループ内のノードのオーダリングを固定することにより探索空間を軽減させ準最適な構造を決定する．

3.2 言語関連ベイジアンネットワーク

本研究では，対話者の発話に含まれる音声特徴に加えて発話文の言語情報を感情推定に利用する手法を提案する．本節では，音声とは独立した言語情報のみからなる言語関連ベイジアンネットワークの構築について概説する．

3.2.1 言語情報

本研究では，まず蓄積した Ifbot-対話者間の会話履歴から感情を含む対話者の発話文を抽出し，その時点で対話者が抱いた感情 ($EMOT$) を 6 感情から選び記録する．次に対話者が各文中で同感情を最も強く表す単語 ($WORD$) を抽出し各語に対して話者が感じる 2 つの感性ラベル ($INFO1$：快，不快 と $INFO2$：興奮，鎮静) を付ける．続いて各文の文型情報 ($GRAM$ ：肯定，否定，疑問) を抽出し上記の 5 属性から成る事例集合を作成する．そして，事例集合を Ifbot が知覚できる語彙でフィルタリングしたものを学習データとする．表 2 に学習データの一例を示す．学習データを基に音声関連ベイジアンネットワーク同様にモデルの構造を決定する．

4 音声・言語混合ベイジアンネットモデル

本研究では，音声･言語の 2 つのベイジアンネットワークを混合することにより双方の情報から対話者感情を総合的に推定する．$P_{BN_1}(EMOT)$，$P_{BN_2}(EMOT)$ をそれぞれ前節で構築した音声および言語のベイジアンネットワークにおける目標属性 $EMOT$ の確率分布とすると，提案手法による対話者感情の推定値 $P_{mix}(EMOT)$ は以下の確率分布で定義される．

$$P_{mix}(EMOT) = \sum_{i=1}^{2} \beta_i P_{BN_i}(EMOT) \qquad (3)$$

ここで β_i は混合率 ($\sum_i \beta_i = 1, {}^{\forall}\beta_i \geq 0$) である．

5 評価実験

3 節で述べたように音声および言語の資料をそれぞれ 550 事例用意し，そこから任意に 500 の学習事例と 50 のテスト事例を作成した．図 2 にそれぞれの学習事例から生成された音声・言語混合ベイジアンネットモデルを示す．同図のベイジアンネットワークに対しテスト事例を用いて対話者感情の推定実験を行った．表 2 に感情推定の正答率を示す．実験結果から，音声によるモデルでは「悲しみ」「喜び」，言語によるモデルでは「喜び」についての感情推定が困難であることがわかる．これに対して，音声・言語混合ベイジアンネットモデルではそれぞれの特徴量による感情推定の弱点を補うことで 5 感情すべてに概ね高い正答率が得られている．なお，今回の実験では「驚き」に関する音声および言語の学習データが十分に収集できなかったため，同感情の推論に関する知識は獲得できなかった．

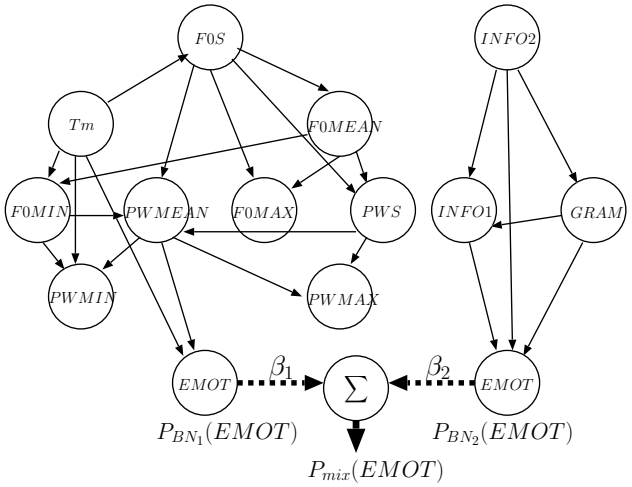

図 2: 音声・言語混合ベイジアンネットモデル

表 2: 感情推定実験結果

感情	正答率		
	音声	言語	混合 ($\beta_1 = 0.7$)
怒り	66.6	66.6	**80.0**
嫌悪	**50.0**	**50.0**	**50.0**
悲しみ	27.2	**72.7**	54.5
恐怖	**80.0**	60.0	**80.0**
喜び	42.8	28.6	**57.1**

6 おわりに

本研究では，音声・言語混合ベイジアンネットワークを用いて感性ロボットのための対話者感情の推定法を提案した．本手法により，音声情報と言語情報の双方から対話者感情を推定することが可能となった．今後の課題としては，学習データを増加させることにより感情推定の性能を改善すること，および，表情を含め年齢，性別等を考慮した総合的な感情推論器を構築し Ifbot への実装を行なう予定である．

参考文献

[1] 竹内将吾，酒井あゆみ，加藤昇平，伊藤英則：対話者への好感度を考慮した感性会話ロボットの感情生成モデル，日本感性工学会第 7 回大会，p. 137，2005.

[2] K. B. Korb and A. E. Nicholson, Bayesian Artificial Intelligence, Chapman & Hall/CRC, 2004.

[3] 重永 寛：感情の判別分析からみた感情音声の特性，電子情報通信学会論文誌， Vol. J83-A No. 6， pp. 726-735，2000.