

日中聴取者による音声から受ける印象を用いた感情認識モデルの構築と比較

川村 剛[†] 政倉 祐子^{††} 大野 澄雄^{††}[†] 東京工科大学大学院 バイオ・情報メディア研究科 ^{††} 東京工科大学 コンピュータサイエンス学部

1 はじめに

一般的に異なる文化圏では音声による感情表現とその認識には文化的習慣により違いがある。そのため、文化的要因による影響により特定の言語の感情認識モデルは他の言語では認識精度が低下する。既存の重回帰分析を用いてモデルを構築する方法では、文化的要因による影響を切り分けるには難しい。感情認識に影響する文化的要因を明確にできれば他言語に適応可能な感情認識モデルの構築が可能になると考える。本報告では、音声から受ける印象を利用し、2段階に分けてモデル構築を行う。印象(値)から感情(値)を推定するモデルを構築する。このモデルへの入力となる印象値は音響的特徴量からそれを推定するモデルを別に構築して用いる。日本人と中国人それぞれについて同様に分析し比較した結果を報告する。

2 音声資料収集と音響的特徴量

本研究では、自然会話音声収集のために大学生及び大学院生の男女2名ずつ計4名が、ポータブルゲーム機を無線LANで相互に接続しゲームする際に発せられた音声を収録した。その際、Skypeを通して会話しTapurで音声を収録した。音声は日本語音声、中国語音声それぞれ収録した。日本語音声は東京工科大学、中国語音声は北京理工大学の学生に協力をしてもらい、学生は1年以上それぞれの言語の標準語を使い続けているという条件を満たしている。

3時間程度収録した会話音声データをPraatを用いて、100ms以上の無音区間を境界として発話を切り出した。話者以外の音、笑い声を含むものは除いた後、男女混合で560個をランダムに選出し評価実験に利用した。前後30個ずつはダミーとしているため分析に使用するデータは500個である。日本語と中国語とを合わせて計1000個の音声コーパスを構築した。

各々の音声に対して表1に示した特定の言語に依存しない特徴量をMATLABを用いて自動的に抽出した。なお、 F_{0min} 、 F_{0mean} 、 F_{0max} については話者ごとに標準化した値を用いている。

表 1: 韻律的特徴量

特徴量	説明
F_{0min}	F_0 値の発話内最低値
F_{0mean}	F_0 値の発話内平均値
F_{0max}	F_0 値の発話内最高値
F_{0stdv}	F_0 値の発話内標準偏差
P_{mean}	短時間平均パワーの発話内平均値
P_{max}	短時間平均パワーの発話内最大値
P_{stdv}	短時間平均パワーの発話内標準偏差
P_{magn}	短時間平均パワーの変動量
C_{1min}	第1次ケプストラム係数の発話内最小値
C_{1mean}	第1次ケプストラム係数の発話内平均値
C_{1max}	第1次ケプストラム係数の発話内最大値
C_{1stdv}	第1次ケプストラム係数の発話内標準偏差

3 評価実験

評価実験では聴取者が音声から受ける印象と感情を文献[1]と同様の手法で評価した。以下のように、聴取者は日本人と中国人の2グループで、日本語コーパスと中国語コーパスそれぞれについて評価を行った。印象評価は森らが文献[2]で提案した6つの軸を用いた(表2)。印象評価は、それぞれの軸に対して7段階(-3, -2, -1, 0, +1, +2, +3)で程度を評価した。感情評価は喜び、悲しみ、怒り、恐れについて0(全くその感情はない)、1~5の6段階で程度を評価した。評価は各聴取者グループの母語に翻訳した評価システムを用いた。

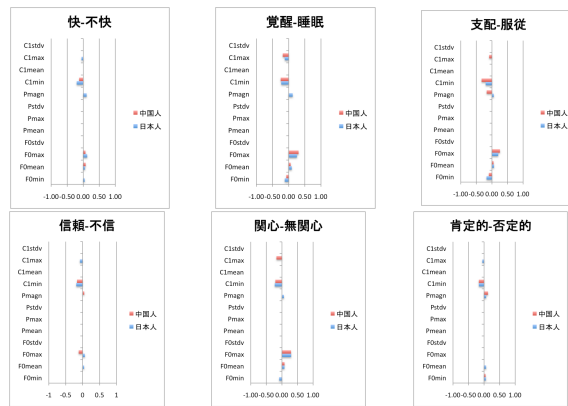
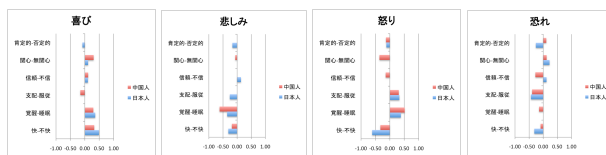
表 2: 森らが提案した感情評価軸 [2]

印象軸	説明
快 - 不快 (I_1)	話者の感情状態
覚醒 - 睡眠 (I_2)	
支配 - 服従 (I_3)	
信頼 - 不信 (I_4)	対人関係
関心 - 無関心 (I_5)	
肯定的 - 否定的 (I_6)	態度

A study on difference of emotional perception between Japanese and Chinese listeners using impressions

[†] Tsuyoshi KAWAMURA (Graduate School of Bionics, Computer and Media Sciences, Tokyo University of Technology)

^{††} Yuko MASAKURA, Sumio OHNO (School of Computer Science, Tokyo University of Technology)

図 1: 印象軸の β 値図 2: 感情軸の β 値

4 重回帰分析によるモデル構築

印象認識モデル (*model1*) と感情認識モデル (*model2*) の 2 つのモデルを構築し検定した。以下にその結果について述べる。

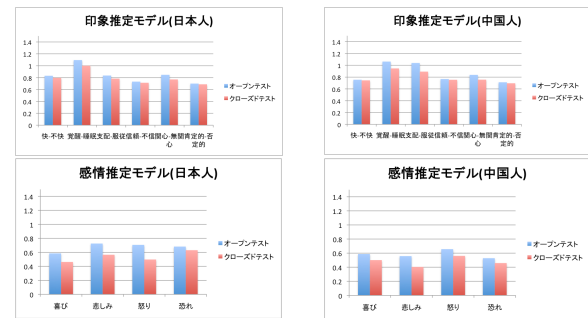
4.1 印象認識モデル (*model1*) と考察

各印象ごとに、音響的特徴量 (1) を説明変数、印象値を目的変数として変数増減法による重回帰分析を用いてモデルを構築した (*model1*)。モデルに対してクローズドテスト、オープンテストを行った。その結果、各印象モデルの決定係数は、日本人 ($I_1 = 0.08, I_2 = 0.17, I_3 = 0.12, I_4 = 0.05, I_5 = 0.18, I_6 = 0.03$)、中国人 ($I_1 = 0.03, I_2 = 0.21, I_3 = 0.26, I_4 = 0.03, I_5 = 0.19, I_6 = 0.04$) であった。また、得られたモデルの各変数の標準化偏回帰係数 (β) を図 1 に示す。クローズドテスト、オープンテストにおける残差の RMS を図 3 に示す。

決定係数は日本人、中国人ともに 0.3 以下とモデルの精度は良くない。 I_1, I_4 を除き比較的中国人の方がよい。図 1 から全体的に β の値は小さいが、各変数の β の傾向が日本人と中国人間で共通している部分は多い。例えば C_{1min} は I_3 を除く全ての軸において β の値がほぼ同等と言える。一方、 C_{1max} のようにどちらかにしかない、 I_4 における F_{0max} は逆方向を示している変数もある。また、図 3 から全体的にクローズドテストが良いが、 I_4, I_6 それぞれ β 値が 0.015 前後の差しかなくほぼ同等の結果となった。

4.2 感情認識モデル (*model2*) と考察

印象値を説明変数、感情値を目的変数として 4.1 と同様にモデルを構築し検定を行った (*model2*)。得ら

図 3: 検証における残差の RMS(左: 日本人聴取者、右: 中国人聴取者、上段: *model1*、下段: *model2*)

れたモデルの β を図 2、検定結果を図 3 に示す。

各感情モデルの決定係数は、日本人 (喜び = 0.74, 悲しみ = 0.59, 怒り = 0.60, 恐れ = 0.27)、中国人 (喜び = 0.59, 悲しみ = 0.64, 怒り = 0.33, 恐れ = 0.31) であった。決定係数は日本人の恐れが 0.27 が一番低く、日本人の喜びの 0.74 が最も高い。これは印象と感情間に強い関係性があることを示唆していると言える。図 2 から β は日本人と中国人でかなり異なると言える。また、オープンテストはクローズドテストに比べて、*model1* の日本人は 2 ~ 10%、中国人は 1 ~ 16%、*model2* の日本人は 8 ~ 42%、中国人は 15 ~ 38% の範囲で低い。図 3 から *model2* おいて、特に怒りについて中国人に比べて日本人の方が劣化していることが分かる。

5 まとめ

感情認識モデルに影響する文化的要因を明確にするため、文化圏別聴取者グループごとに印象を用いた感情認識モデルを構築し比較検討した。*model1* は精度は良くないが β 値の傾向においては日本人と中国人とで共通する部分が多いことが分かった。これは声から受ける印象については文化に影響が少ないことを示唆している。一方で *model2* の β 値では日本人、中国人の差が明確に現れていると言える。以上のことから文化の差は、声から受けた印象を感情へ変換する過程において生じていると考えられる。今後、文化差をより明確にするためには *model1* の精度向上、モデル間の検定を行う必要があると考える。

参考文献

- [1] Kawamura, T., Ohno, S., " A study on the difference of emotional perception between Japanese and Chinese ", PACLING 2011 in Malaysia, 2011.
- [2] Mori, H., Satake, T., Nakamura, M., and Kasuya, H. "Constructing a spoken dialogue corpus for studying paralinguistic information in expressive conversation and analyzing its statistical/acoustic characteristics", Speech Communication, **53**, 1, pp.36-50, 2011.