1J4-OS-22b-3

マルチモーダル情報を用いた情緒的な発話検出と議論分析

Emotional Speech Detection from Multimodal Data for Discussion Analysis

坂原誠 *1　岡田将吾 *1　新田克己 *1
Makoto Sakahara　Shogo Okada　Katsumi NITTA

*1東京工業大学大学院 総合理工学研究科
Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology

Much research about the discussion is being done in terms of logic, however, hardly give the consideration to emotion. To detection emotional parts and analyze how those influence the logical structure is essential in the cooperative discussion. We define heat-up and complication as emotion in the discussion and annotate those in a real-life discussion data. To discriminate emotion, pitch features are extracted from the data. The pitch statistics is more emotionally features than the pitch shapes. Therefore we calculate pitch statistics and select the best features to distinguish emotional versus neutral speech. Finally, we compare conventional classification schemes with advanced scheme using neutral model. Now we challenge more difficult task than previous work on using a real-life discussion data. The results show that the recognition accuracy is over 56% in conventional classification schemes (baseline 50%).

1. はじめに

長時間に及ぶ議論や, 複雑な議題を扱う場合, 論点がどこにあるのか, 議論がどのように進行したのか, 何が結論となったのかなどを, 発言記録から正確に判断するのは困難であり, 負担の大きい作業である. 議論解析では, それらの問題点を工学的に扱う.

議論は, 競合的な議論と協調的な議論の 2 種類に大きく分けられる. 競合的な議論では, 相手の意見を破ることを目的とするが, 協調的な議論では, お互いの意見を一致させるという合意形成を目的とする. 多くの議論解析の研究では, 議論の論理的構造を中心に解析が行われてきたが, 協調的な議論においては, 論理的な側面だけの解析では不十分である. 議論における情緒的な部分を検出し, それが論理展開の進行にどのような影響を与えたのかを解析することも必要になる. しかし, 情緒を考慮した議論解析はほとんど行われていない.

われわれは, 動画による議論解析の 1 つとして「身振り」に着目した情緒場面の検出を試みた [Sugimoto 12]. その結果, 身振りは情緒場面の検出精度を上げるのに一定の効果はあった. しかし, 身振りが表れる場面は限られており, 身振りが出現しない情緒場面は検出できなかった.

そこで, 本稿では議論の音声情報を利用して情緒場面の検出精度を高めることを目的として研究を行った. これは, プロの読み上げた感情音声を用いた先行研究 [Busso 09] とは異なり, 実際の議論の音声データを用いたよりチャレンジャブルなタスクである.

2. 議論解析の概要

本研究の議論解析は, 論理分析と情緒識別の 2 つからなる. 議論解析の流れを図 1 に示す.

論理分析は, 議論の発言記録を入力としてとる. 議論の発言記録に対して, 論証の一般的な記述法である Toulmin ダイアグラムにおける根拠, 主張, 反証などを各発言に対してアノテーションする. そしてダイアグラムベースで論理的構造を可視化し, 発言の論証を出力する [Kubosawa 12].

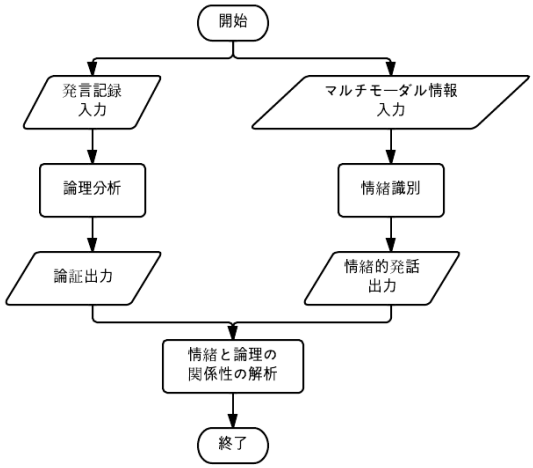

図 1: 議論解析の流れ

情緒識別では, 議論のマルチモーダル情報を入力としてとる. 入力された情報から議論の情緒的な部分を検出し, 検出部分の発言を出力する. 実際の議論において, 合意形成を目的とする際の情緒とは, 対立しているか, あるいは和やかであるかの 2 種類であると考えられる. そこで, 本研究では, お互いの意見が対立して言い争いになっている場面を『紛糾』, お互いが談笑していたり, 声が大きくなっている場面を『盛上がり』として定義した.

それぞれの出力結果を用いて, 情緒が議論の論理展開にどのような影響を与えたのか, 相関 · 回帰関係を分析する. また, 円滑な議論と難渋した議論では, 情緒の生起関係にどのような違いがみられるのかを解析する.

連絡先: 坂原誠, 東京工業大学大学院 総合理工学研究科 知能システム科学専攻, 〒 226-8502 神奈川県横浜市緑区長津田町 4259 J2-53, Tel:045-924-5205, Fax:045-924-5218, E-mail:sakahara@ntt.dis.titech.ac.jp

3. 手法

3.1 韻律的特徴

文字情報では, 文脈などによって言語情報以外の情報が伝達されるが, 音声言語では, 文字言語に比べて態度や感情といった情報の比重が増す. そして, 感性情報は主に韻律的特徴によって伝えられる. 韻律的特徴は音源に関連した特徴であり, 具体的には, 音源振動の基本周波数 (以下 F_0), 音素継続時間長, 音源強度の 3 つからなる. その中でも特に, 声帯振動の F_0 が最も直接的に韻律を表現する特徴量であるとされている.[広瀬 08][Busso 09]

そこで, 本手法でも F_0 を特徴量として扱う. しかし, 感情音声と韻律的特徴の相関関係については, 多くの研究がなされているが, その規則は非常に複雑でまだ十分に解明されていない.

3.2 認識方法

感情識別の研究では, 抽出した特徴量を用いて, 何らかの識別器による認識を行うことが一般的である.(図 2) 本稿では,LDA(Linear Discriminant Analysis) を用いてインプットデータが情緒タグのない発言 (以下ニュートラスピーチ) か情緒タグのある発言 (以下エモーショナルスピーチ) であるかを識別する. また, 先行研究 [Busso 09] において, よりロバストで正解率の高いとされるニュートラルモデルを用いた手法 (3.2.1 項) が提案されている. そこで, この手法との比較も行った.

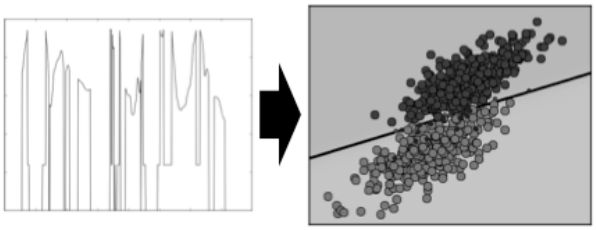

図 2: 一般的な認識手法

3.2.1 ニュートラルモデルを用いた手法

エモーショナルスピーチは話者に依存するところが大きく, さらに感情自体にもはっきりとした定義は確立されていない. また, エモーショナルスピーチのコーパス化は質・量ともに困難であり, それぞれの感情についてモデルを生成することは, 過学習となりロバスト性が失われる危険性がある. そこで, まず別のニュートラルコーパスを用いてニュートラルモデルを訓練する. そして, インプットデータのニュートラルモデルに対する尤度を識別器により分類する.(図 3) 仮に, ニュートラルモデルにニュートラルスピーチが入力されると高い尤度が, エモーショナルスピーチが入力されると低い尤度が出力される. ニュートラルモデルを用いることで, 話者性に対してよりロバストな識別が期待される.

ニュートラルモデルは各特徴量 f に対して GMM(Gaussian Mixture Model) を用いて生成する.

$$F_f(X_f = \chi_f|\Theta) = \sum_{j=1}^{K} \alpha_j \frac{1}{\sigma_j\sqrt{2\pi}} \exp\Big(\frac{(\chi_f - \mu_j)^2}{-2\sigma_j^2}\Big) \quad (1)$$

with

$$\Theta = \{\alpha_j, \mu_j, \sigma_j\}_{j=1}^{K} \alpha_j > 0 j = 1, \ldots, K, \sum_{j=1}^{K} \alpha_j = 1.$$

GMM の Θ は EM アルゴリズムにより求められる. 本稿では, インプットデータに対して求められる尤度 ($F_f(X_f = \chi_f|\Theta)$) を LDA を用いてエモーショナルスピーチかニュートラスピーチであるかに識別する. 以下, この手法を NM-LDA と表記する.

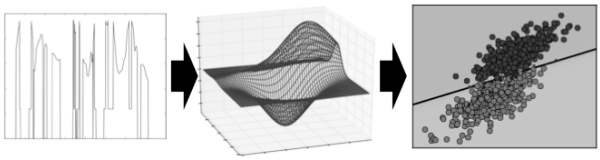

図 3: ニュートラルモデルを用いた認識手法

4. 実験

4.1 情緒タグのアノテーション

感情を含んだ大規模な日本語音声コーパスというものは現状では存在しない. また, 自由発話型の対話音声コーパスというものも一般には公開されていない. そこで, 本稿ではまず, テレビの討論番組の音声データ (表 1) に対して 2. 節で定義した情緒にもとづいて発言単位での情緒タグのアノテーションを複数人で行った. 情緒タグの内訳を表 2 に示す.

表 1: テレビの討論番組の音声データ詳細

サンプリング周波数	16KHz(48KHz をダウンサンプリング)
量子化ビット数	16bit
チャンネル数	モノラル (ステレオをミックス)
話者数	13 名 (男性 9 名, 女性 4 名)
発言数	559

表 2: 情緒タグ内訳

	発言数
紛糾	69
盛上がり	119
紛糾 $\wedge$ 盛上がり	32
紛糾 $\vee$ 盛上がり	156

4.2 特徴選択

Speech Signal Processing Toolkit[SPTK] の pitch コマンドを用いて発言ごとに F_0 を抽出した. pitch コマンドのパラメータを表 3 に示す.

表 3: pitch コマンドパラメータ

パラメータ名	値
ピッチ抽出アルゴリズム	SWIPE
フレームシフト	80ms
SWIPE 閾値	0.3
最小 F_0	60.0(Hz)
最大 F_0	240.0(Hz)
出力形式	F_0(Hz)

F_0 の形状そのものよりも, 発言単位での F_0 の中央値, 四分位数のような統計量の方がより有意な特徴量であることがわかっている. そこで, 発言ごとに抽出した F_0,∇F_0 の値から統計量を算出した.(表 4) ここで,∇F_0 は F_0 の近似微分の値である.

特徴量の性能は, 対象ドメインに敏感であるため, 実際の議論の音声データに適したものを選択する必要がある. そこで, 分散分析 (ANOVA) により各特徴量の p 値を求めた. テレビの討論番組の音声データ (表 1) の全 559 発言のうち, 情緒タグがアノテーションされている発言は 156 発言である. インプットデータのニュートラスピーチとエモーショナルスピーチの数を等しくするために, 各実験毎にニュートラスピーチからランダムに 156 発言を選び, 選ばれたニュートラスピーチとエモーショナルスピーチを合わせた 312 発言をインプットデータとした.

100 回の実験における各特徴量の p 値の平均値を図 4 に示す. $(-\log(p_{value}) > 0.4)$ である Sdstd, Siqr, SQ75, Sstd, Sdiqr の 5 つの特徴量を選択した.

表 4: F_0 特徴量

	F_0	∇F_0
平均値	Smean	Sdmean
標準偏差	Sstd	Sdstd
レンジ	Srange	Sdrange
最小値	Smin	Sdmin
最大値	Smax	Sdmax
中央値	Smedian	Sdmedian
下側四分位数	SQ25	SdQ25
上側四分位数	SQ75	SdQ75
四分位数範囲	Siqr	Sdiqr

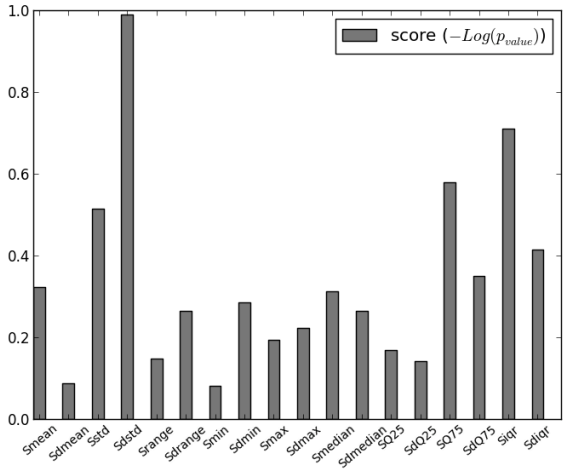

図 4: 各特徴量のスコア

4.3 実験設定

4.2 節と同様に, 各実験毎に議論音声データからニュートラスピーチをランダムに 156 発言選び, 選ばれたニュートラスピーチとエモーショナルスピーチを合わせた 312 発言をインプットデータとした. したがって, 正解率のベースラインは 50.0%である.

特徴量は,4.2 節で選択した Sdstd, Siqr, SQ75, Sstd, Sdiqr の 5 つを用いた.

ニュートラルモデルを訓練するためのニュートラルコーパスには, ニュース放送用の原稿をプロのアナウンサが読み上げたもの [RWCP-SP99](表 5) を用いた. GMM の混合数は先行研究 [Busso 09] と同じ $K = 2$ に設定した.

表 5: RWCP-SP99 詳細

サンプリング周波数	16KHz
量子化ビット数	16bit
チャンネル数	モノラル
話者数	6 名 (男性 3 名, 女性 3 名)
発言数	(全話者共通原稿 10 件＋個別原稿 30 件)

4.4 実験結果

各実験には,5-fold Cross Validation を用いた. 100 回の実験における正解率 (Acc), 適合率 (Pre), 再現率 (Rec),F 値 (F) の平均値を表 6,7 に示す. また, 各手法の Confusion Matrix を表 8,9 に示す.

表 6: 正解率

	Acc
LDA	56.19%
NM-LDA	47.47%

表 7: 識別性能

		Pre	Rec	F
LDA	ニュートラル	0.5744	0.6510	0.5826
	エモーショナル	0.6054	0.5268	0.5332
NM-LDA	ニュートラル	0.5058	0.5986	0.5081
	エモーショナル	0.5168	0.4247	0.4172

表 8: Confusion Matrix(LDA)

	ニュートラル	エモーショナル
ニュートラル	97	59
エモーショナル	78	78

表 9: Confusion Matrix(NM-LDA)

	ニュートラル	エモーショナル
ニュートラル	88	68
エモーショナル	96	60

5. 考察

先行研究に比べて厳しいデータセットを用いたが,LDA を用いた一般的な判別分析手法では, 約 56%の正解率を達成し, ベースラインの 50%を上回った. 先行研究では,NM-LDA で約 77%の正解率を達成していたが, 本稿ではこちらの手法の正解率は約 47%にとどまり, ベースラインの 50%を下回った. LDA を用いた一般的な判別分析手法では, 本実験においても約 56%の正解率を達成しているので, これはニュートラルモデルの訓練方法に問題があったのではないかと考えられる. 今回

の実験では, 音声データの正規化はデータセットごとに一律に行なっており, 話者ごとに行なっていないことが原因として挙げられる.

6. おわりに

議論における検出すべき情緒を定義し, 実際の議論である討論番組の音声データにアノテーションを行った. また, このデータセットにおける有意な特徴量を選択し, それらの特徴量を用いて判別分析を行った. 先行研究よりも厳しい条件のデータセットであったが, 約 56%の正解率を達成しベースラインを約 6%上回った.

今回の実験では, データセットの量が十分であったとは言い難い. また, 認識された情緒のうち, 紛糾と盛上がりのそれぞれの識別性能がどれほどなのかも検証する必要がある.

今後は, データセットをより充実させるとともに, 話者ごとの正規化を行った NM-LDA の識別性能の再実験を行う. さらに, 議論中の発言者の表情や身振りを認識することで, マルチモーダルな特徴量として利用することを考えている.

参考文献

[Busso 09] C. Busso, Sungbok Lee, S. Narayanan: Analysis of Emotionally Salient Aspects of Fundamental Frequency for Emotion Detection, IEEE Transactions on Audio, Speech & Language Processing, pp.582-596, (2009).

[Sugimoto 12] Takahiro Ueda, Masaki Sugimoto, Shogo Okada, Yukio Ohsawa, Yoshiharu Maeno and Katsumi Nitta: Discussion Analysis Using Temporal Data Crystallization, The 6th International Workshop on Jurisinformatics, (2012).

[RWCP-SP99] 技術研究組合 新情報処理開発機構 RWCP (Real World Computing Partnership) 知的資源 WG: 検索・要約用ニュース音声データベース (RWCP-SP99)

[SPTK] a suite of speech signal processing tools, http://sptk.sourceforge.net/

[広瀬 08] 広瀬 啓吉: 韻律と音声言語処理, 電子情報通信学会技術研究報告.SP, 音声 108(265), pp.25-30, (2008).

[Kubosawa 12] Shumpei Kubosawa, Youwei Lu, Shogo Okada, Katsumi Nitta: Argument Analysis with Factor Annotation Tool, The 25th International Conference on Legal Knowledge and Information Systems, JURIX, (2012).