音声言語情報処理 34－8
(2000. 12. 21)

音声に含まれる感情の判別に関する検討

門谷 信愛希†, 阿曽 弘具†, 鈴木 基之‡, 牧野 正三‡

† 東北大学大学院工学研究科 ‡ 東北大学大型計算機センター

〒980-8579 仙台市青葉区荒巻字青葉05
TEL : (022)217-7088
E-mail : nob@aso.ecei.tohoku.ac.jp

あらまし　本論文は, 音声に含まれる感情の判別を目的としている. 最初に感情 (怒り, 悲しみ, 喜び) を含んだ音声が連続音声認識システムに与える影響を13名の話者によって発話された1040文章を用いて調査した. その結果, これらの感情は平均で10～20%, 最大で50%程度の認識率の低下をもたらすことが分かった. 次に, 感情の判別にはどのようなパラメータが有効であるかを調べた. 正準判別分析の結果, 文中の最大基本周波数, 最大振幅, 基本周波数の変動範囲などの特徴量が有効であることが分かった. 全話者の混合データに対する判別分析では, 怒り (61.4%), 悲しみ (53.1%), 喜び (45.8%) の順で判別率が低下することが分かった. 一方で, 各パラメータの判別に対する寄与度は話者によって異なっていることが分かった.

キーワード　基本周波数, 感情の判別, 判別分析

An investigation on discrimination among emotion expressions contained in speech

Nobuaki Kadotani†, Hirotomo Aso†, Motoyuki Suzuki‡, Shozo Makino‡

† Graduate School of Engineering, Tohoku University
‡ Computer Center, Tohoku University

05 Aramaki-aza-aoba, Aoba, Sendai, Miyagi 980-8579

Abstract　This paper describes the discrimination among emotion expressions contained in speech. At first, we investigated an influence of these emotion expressions(anger,pleasure and sadness) on performance of continuous speech recognition system using 1040 sentences uttered by 13 speakers. We found that those three emotions gave 10% to 20% down on the performance. Next, we investigated effective parameters for discrimination among emotions. Based on canonical discriminant analysis, the following parameter are effective for all speakers: maximum fundamental frequency,maximum power and variation range of fundamental frequency. Discriminant rates for all speakers decrease in order of anger(61.4%),pleasure(53.1%) and sadness(45.8%). On the other hand, we also found that contribution rate of these parameters is different dependent on speaker characteristics.

key words　fundamental frequency, emotion discrimination, discriminant analysis

1 はじめに

近年の音声認識研究の発展はめざましいものがあり, 一般家庭などの低騒音環境下においては,90%以上の認識率を得る認識エンジンなども存在している [1]. また, 切符購入システムや秘書ロボットなどのエージェントシステムにおいても, 音声をそのインターフェースとして用いるための研究が進んでいる. しかし, ディクテーションシステムは基本的に, 平静時の音声を対象として設計されており, 話者の心理状態の変化に起因する音声の様々な変動がどのような影響を及ぼすかの調査はこれまであまり行われていない. もしこのような変動により影響があるとすれば, 感情ごとに音響モデルを切り替えるなどの対応が必要になると思われる.

また, 対話システムにおいては従来, 合成音声による応答がメインであって, 話者の心理状態に関わらず, 一定調のレスポンスをするだけであった. このため, ごく限定されたタスクから, より汎化性を増したシステムにそのまま適用すると, ユーザーに違和感を与えることが予想される. このような違和感は, 装置と人間との外観の差異よりもむしろ, 人間が行っているような相手の表情, 声の調子, しぐさなどを判断して応答を柔軟に変化させる機構の欠如がもたらすものと考えられる. 従って, 今後様々な社会システムに音声認識・理解・応答システムを導入する際には, 人間の能力に近い機能を持つようにさせる必要がある.

本研究においては以上のような観点から, まず感情を含んだ音声を従来の連続音声認識システムで認識し, どのような影響があるかを調べた. また, 感情の判別にはどのようなパラメータが有効であるか, 判別分析を行って検証した.

2 使用した音声データ

音声データは, 特定の感情の影響を比較的受けにくいと思われる会話文を 20 文選択した. 文章内容を表 1 に示す.

感情の選択は, 最も基本的な感情であると考えられる「怒り」「悲しみ」「喜び」に, 平静時を表す「標準」を加え, 計 4 感情とした. これらの文章を番号順に被験者に提示し, それぞれの話者が最もよくその感情を表現していると思われる方法で発話してもらった. また, 表 2 に示すような, 感情ごとに想定される場面の一例を示し, 話者が感情移入しやすいようにした. また, 録音前の練習は自由に行ってよいこととした. 録音はいずれも晴天の日に一般木造家屋 (絨毯張り) で行い,DAT とヘッドセット型マイクロフォン (SENHEIZER 製 HMD410) を使用した. サンプリングは周波数 48KHz, 16bit/sample の精度で行い,DAT-Link を用いて 16KHz にダウンサンプリングしながら計算機に取り込みを行った.

表 1: 発話文章リスト

No	内容
1	電車, いつ来るのかなあ
2	山田君, そこのドア閉めてくれる?
3	今日は,8 時に会社にいかないと
4	机の上を, 片づけなさい
5	私は別に, 問題はありませんよ
6	明日の予定, あいてます?
7	そんな自分勝手なこと, ダメだよ
8	マイクの音量を, 上げて下さい
9	まな板を, 洗っておいて
10	車の調子が, 悪いんだ
11	今日は, 雨が降ったね
12	勉強もっとしなさい
13	この色で, 塗っていいのかい
14	あの時計,3 分進んでいるよ
15	昨日, みんなでカラオケに行ったの?
16	そんなに心配しなくても, 大丈夫だよ
17	後で, 手紙を書いておきます
18	辞書で調べれば, わかるよ
19	電話がくるのを, 待っています
20	最近, 天気がよく変わるね

表 2: 「山田君, そこのドア閉めてくれる?」の場面例

感情	想定される場面
怒り	「虫の居所の悪い上司のつもりで」
悲しみ	「憂鬱な会社に 8 時に行くなんて … 」
喜び	「仲の良い友達の家での会話」

3 音声認識システムに与える影響

3.1 実験条件

使用した連続音声認識エンジンは JULIUS[2] である. 実験時のパラメータを表 3 に示す. 音響モデルは男性話者と女性話者で切り替えた. これ以外のオプションはす

表 3: 連続音声認識の実験条件

バージョン	JULIUS-3.1p2
データ形式	MFCC(12 次元)
音響モデル	triphone-HMM(mix:16)
言語モデル	20K 単語辞書+3gram
話者数	13 名 (男性 7 名, 女性 6 名)
文章数	1040 文 (20 文 ×4 感情 ×13 名)

べてデフォルト値を採用した. なお, 使用した言語モデルの実験対象文に対する未知語率は 1.32%であった.

3.2 評価基準

評価は入力文章の音素表記と JULIUS の第 2 パスから出力される音素系列を DP マッチングすることで行った. 評価基準としては, 挿入および脱落を考慮した (1) 式で表される音素正解精度を採用した.

$$Aqurracy = \frac{N - R - D - I}{N} \cdot 100 \ [\%] \qquad (1)$$

但し,N は全音素数,R,D,I は順に置換, 脱落, 挿入誤りを示すものとする.

3.3 実験結果

男性話者の認識率を総合した結果を図 1 に, 女性話者の認識率を図 2 に示す. 値はいずれも 20 文の平均値である.「全体平均」は全話者の平均認識率を表す.

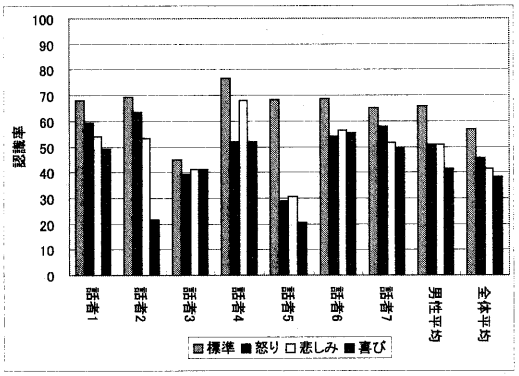

図 1: 平均音素正解精度 (男性話者)

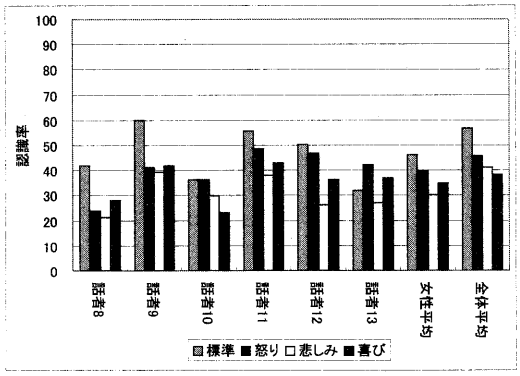

図 2: 平均音素正解精度 (女性話者)

まず, 図 1,2 のデータについて, 感情間に平均認識率の有意差を仮定した検定を話者ごとに行った. その結果, 有意水準 5%では話者 3 で仮定が棄却されたが, それ以外の 12 名の話者については仮定が採択された. このことから, 平静音声に感情が含まれることで, 有意差を伴う認識率の低下を起こすことが確認された.

男性話者では 60〜70%程度の「標準」の認識率が得られているのに対し, その他の感情ではいずれも 10%から 20%, 最大で 50%近い認識率の低下が見られた. 特に「喜び」の影響が大きいことがわかる. 男性話者の平均では標準 → 怒り → 悲しみ → 喜びの順で影響度が増している.

一方, 女性話者では 50〜60%程度の精度が得られているが, 感情ごとの影響度合では,「悲しみ」の方が「喜び」よりも大きいことが判明した. 平均でも標準 → 怒り → 喜び → 悲しみの順に精度が低下している. これらの結果は, 男女の感情表現に差があることを示していると考えられるが, 一方で影響度合いは話者によって著しく異なっている. このため, もし感情ごとに音響モデルを切り替えるシステムを構築する場合は, 認識の前段階で, 感情を推定することが不可欠であるといえる.

4 感情の判別

4.1 特徴量の選択

音声から抽出する特徴量には, 感情間の相違をよく含んだものである必要があり, これまでにも様々な特徴量が検討されてきた. 音声の感情表現は韻律や振幅, および文長などの時間構造, 等に現れると考えられ, 従来の研究 [3][4][5] においてもその有効性がある程度確認されている. 本研究ではこれらを考慮して, 特に重永の研究 [5] に着目し, 韻律的特徴としては抽出が容易な基本周波数 (F0) を選択した. 重永の研究 [5] においては,

$F0_{max}$: 文中の F0 の最高値 [Hz]
A_{max} : 文中の最大振幅
T : 文の長さ [sec]
$F0_{init}$: 文頭の F0 値 [Hz]
$F0_{range}$: $F0_{max} - F0_{min}$ [Hz]

の 5 種の特徴量が基礎量として用いられていたが, 本研究ではこれらに加えて,

T_{zero} : 文中の無音区間の累積長 [sec]
$F0_{avg}$: F0 の平均値 [Hz]
$F0_{diff}$: F0 のフレーム間差分平均 [Hz]

の 3 種の統計量を加え, 全部で 8 次元のパラメータを構成することとした. 基本周波数軌跡上でこのうち 5 つを図示したものが図 3 である.

T_{zero} を採用した理由は, 特に悲しみの場合に, 単語間のポーズ (無音区間) が長くなる傾向が見られ, 判別に寄与すると考えたためである. また, $F0_{diff}$ については, 怒りや喜びなどの場合で基本周波数の変化が他にくらべて急

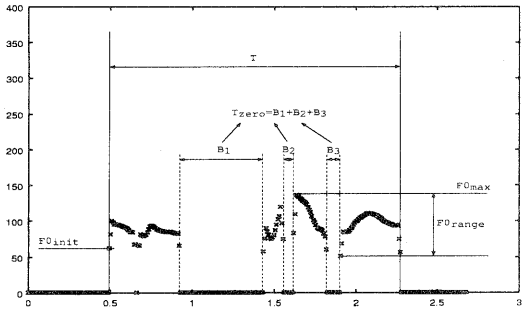

図 3: 基本周波数軌跡と各パラメータの関係

峻になっている点に注目した. $F0_{avg}$ については, $F0_{max}$ や $F0_{range}$ と相関を持つことが予想されるが, 補助的に導入した.

4.2 基本周波数の抽出

基本周波数の抽出には, 発話全体の連続性を考慮した DP マッチングによる方法 [6] を使用した. 基本周波数抽出時に使用したパラメータは表 4 の通りである.

表 4: 基本周波数の抽出パラメータ

DFT 次元数	1024 次元
フレーム周期	8ms(128 フレーム)
F0 抽出範囲	50〜400Hz
C の重み係数	$w1 = 6000, w2 = 1.5, w3 = 1$

また, 発話区間と非発話区間の区別には, パワーをしきい値で区切って判断するようにした.

4.3 実験結果

4.3.1 話者ごとに閉じたデータの場合

抽出したデータに対して正準判別分析を行い,4 感情の判別を試みた. まず, 話者ごとに閉じたデータで分析を行い, 選択した特徴量の基礎的性能を判定した. 得られた判別率を表 5 に示す.

話者によって判別率に若干のばらつきが見られるものの,80〜95%程度, 平均でも 90%近い判別率が得られた. また, 感情ごとの全話者平均値を見ると, 標準 → 怒り → 悲しみ → 喜びの順で判別率が低下する傾向が見られた.

次に, 感情ごとに, 文章がどの感情に分類されたかを調べた. 結果を図 4 に示す. 判別割合は, 判別によりその感情に分類された文章データの割合を表している.

喜び以外の感情では, ある特定の感情との判別誤りが他と比較して多いのに対し, 喜びの場合は他の 3 感情とまんべんなく誤りが起きていることがわかる.

表 5: 感情ごとの正判別率

No	標準	怒り	悲しみ	喜び	平均
1	90%	85%	80%	90%	86.3%
2	75%	85%	95%	85%	85.0%
3	100%	100%	85%	85%	92.5%
4	95%	95%	95%	65%	87.5%
5	95%	80%	90%	95%	90.0%
6	100%	100%	80%	75%	88.8%
7	75%	85%	85%	75%	80.0%
8	90%	95%	95%	80%	90.0%
9	100%	75%	90%	70%	83.8%
10	100%	85%	100%	95%	95.0%
11	100%	75%	80%	80%	83.8%
12	90%	100%	75%	80%	86.3%
13	70%	95%	75%	95%	83.8%
平均	90.8%	88.9%	86.5%	82.3%	87.1%

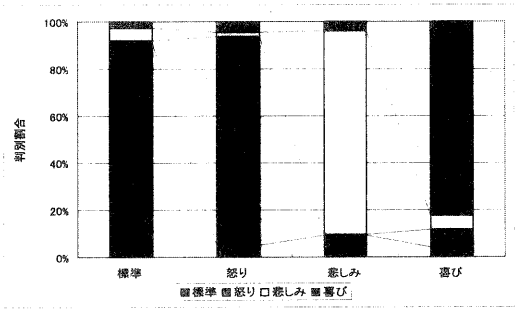

図 4: 感情ごとの判別率と割合 (話者全体)

4.3.2 話者混合データの場合

次に, 同じ分析手法を用いて, 各話者のデータを混合し, その全体に対して判別を試みた. 混合は男性話者全体, 女性話者全体, それに話者全体, の 3 通りについて別々に行った. 結果を図 5,6,7 に示す.

全体的には 60%程度の判別率が得られているが, 特に図 5 の怒りの判別率が他と比較して高い値が得られている. これは, 男性話者が総じて怒りの感情表現を, 他の感情とは明らかに異なる方法で行っていることを示している. また, 対照的に喜びの判別率は低い値であり, 男性の喜びの表現方法が, 他とあまり大差ないためであるとも考えられる.

一方, 女性話者については, 誤って分類される感情が, ある特定の感情に集中する傾向が見られている. 特に, 怒りが喜びに多く誤判別されることや, 悲しみが標準になることが多く, 事前に予想された結果にある程度沿ったものとなっている. 喜びについては, 男性話者同様, まん

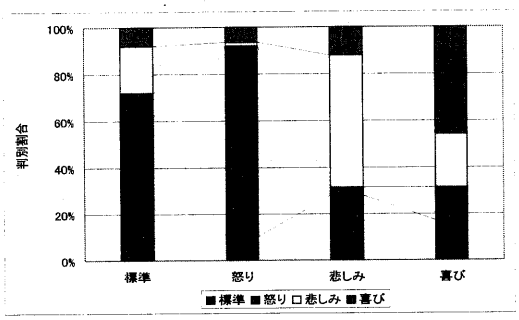

図 5: 感情ごとの判別率と割合 (男性話者)

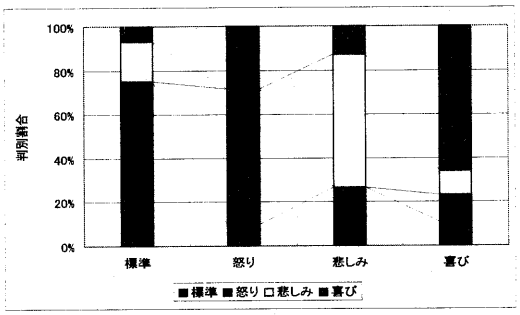

図 6: 感情ごとの判別率と割合 (女性話者)

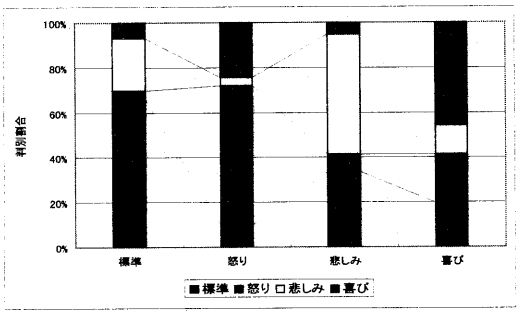

図 7: 感情ごとの判別率と割合 (話者全体)

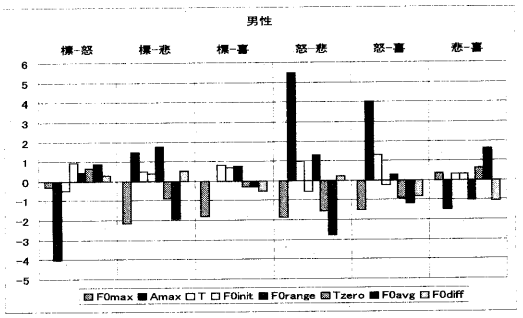

図 8: 標準化判別係数 (男性話者)

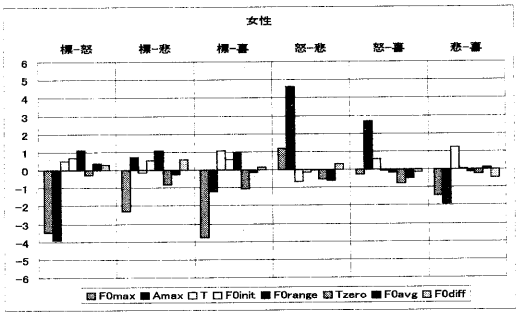

図 9: 標準化判別係数 (女性話者)

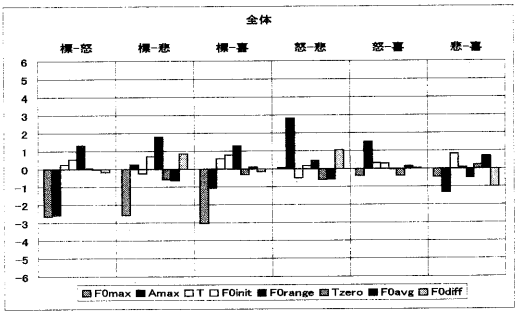

図 10: 標準化判別係数 (話者全体)

べんなく他の感情と誤判別されている.

最後に全話者を統合した場合であるが, 概ね両者の平均をとった値に近い結果が得られた. しかし, 男性話者の怒りにあった特徴的な高い判別率は失われており, データを混合することでこれら男性特有の感情表現が埋もれてしまっていることが見て取れる.

4.4 各パラメータの寄与率

次に, どのような特徴パラメータが判別に寄与しているかを知るため, 各感情の区別に用いる判別関数の係数を調査した. 図 8,9,10 に各パラメータの係数を示す. 判別係数の値は標準化されているので, 絶対値が多きいほど判別への寄与が大きい. まず, 男性話者においては A_{max} と $F0_{max}$, $F0_{avg}$ の絶対値が大きい. 逆に $F0_{init}$ や $F0_{diff}$ は小さく, 判別にほとんど寄与していないことが判明した. 次に女性話者では, やはり A_{max} と $F0_{max}$ が他と比較して大きくなっている. 実質的に, これら 2 つの特徴量で判別が行われているともいえる. しかし, 男女を混合したデータにおいては $F0_{range}$ も判別に寄与していることが分かる. このことから, ある話者に有効なパラメータが, 他の話者に対しても普遍的に有効であるとは考えにくい. そこで, 話者ごとに閉じた環境における判別係数も併せて調査した. 図 11 に 3 名の話者の判別係数を示

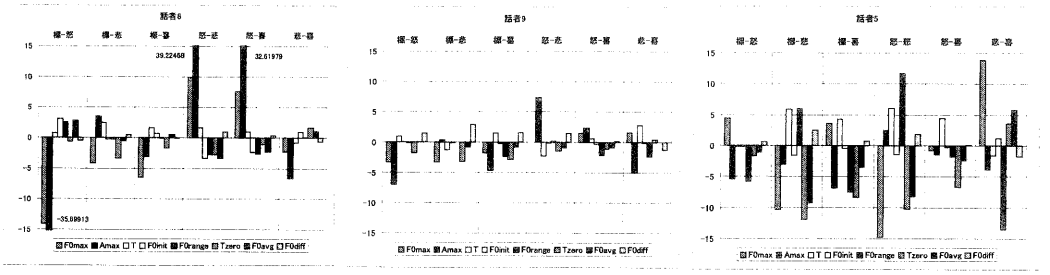

図 11: 標準化判別係数のパターン

す. 今回使用したデータの判別係数の分布は, 大別してこの 3 つのパターンに分類することが出来た. まず, 話者 5 のように $F0_{max}$ や A_{max} が他と比較して大きく, これら少数のパラメータの寄与割合が大きいグループがあった. 話者 3,6,8,12 などがこのパターンであり, 判別率 (表 5) でも, 比較的高い値を得ることができた. 次に, 係数がいずれも同程度の大きさであり, 決め手となるパラメータに欠けるグループがある. 話者 1,2,4,7,9,11,13 がこのパターンである. これらの話者では判別率を見ても, 他の話者と比較して低くなっている. 最後に,$F0_{max}$ や A_{max} 以外で相対的に大きいパラメータが存在するグループがあった. 話者 5,10 が該当する. これらの話者では, 話者 8 のパターンと同等か, それ以上に高い判別率を得ることができた. このことから, 話者の感情表現にはかなり大きい個人差が存在しており, 話者混合データの判別率の低下の原因となっていることが考えられる. 従って, これらの問題を解決するためには, 話者ごとの正規化だけではなく, 個人差を吸収するような別のパラメータも検討する必要があると思われる.

5 まとめ

今回の実験で使用した感情判別パラメータは, 話者単独データではある程度の性能を持っていることが判明したが, 話者によって判別に有効なパラメータにかなりのばらつきがあり, これが話者混合データの場合の認識率低下の原因になることが分かった. 感情表現は話者によってかなり幅があり, 男性と女性との相違も大きい. このような個人差を吸収する頑健な感情判別モデルの構築には, より個人間の変動の小さいパラメータを検討する必要があることが分かった.

謝辞

本研究は, 情報処理振興事業協会 (IPA) が実施した「独創的情報技術育成事業」の援助により開発された「大語彙連続音声認識デコーダ Julius」[2] を利用して行われました. 関係者の方々に深く感謝致します.

参考文献

[1] 西村雅史, 伊東伸泰, 山崎一孝: “単語を認識単位とした日本語の大語彙連続音声認識” , 情処論, Vol.40, No.4, pp.1395-1403 (1999)

[2] 李晃伸, 河原達也, 堂下修司: “単語トレリスインデックスを用いた大語彙連続音声認エンジン JULIUS”, 信学技報, SP98-3 (1998-04)

[3] 上床弘幸, 小林豊, 新美康永: “音声の感情表現の分析とモデル化”, 信学技報, SP92-131 (1993-01)

[4] 川波弘道, 広瀬啓吉: “態度・感情音声における韻律的特徴の考察”, 信学技報, SP97-67 (1997-11)

[5] 重永実: “感情の判別分析からみた感情音声の特性 (III)”, 信学技報,SP97-66 (1997-11)

[6] 城風敏彦, 牧野正三, 城戸健一: “発話全体の連続性を考慮した基本周波数の抽出”, 信学論 (A), Vol.J73-A, No.9, pp.1537-1539 (1990-09)