アクセント成分を用いた講演の強調検出

Prominence detection in a presentation using an accent component

小島淳嗣
Atsushi Kojima
法政大学情報科学研究科
情報科学専攻
E-mail:atsushi.kojima.2v@stu.hosei.ac.jp

Abstract

We propose a method for detecting a prominence in Japanese presentations. The prominence is not clearly defined enough to detect the prominence quantitatively in Japanese, because it is covered only in phonetics and Japanese language education qualitatively. In order to quantify the prominence and propose features for detecting the prominence, we reviewed the literature and we understood how words are emphasized in Japanese sentences. In addition, we analyzed acoustic features (e.g., F0, energy, accent component, pause and speech rate) in a data of utterances including an emphasized word based on the knowledge. As a result, we propose using the accent component and Δ accent as a feature for detecting prominence. In an evaluation experiment to detect prominence, we used the intensity of the accent component and its delta features. The experimental results show that a detection accuracy of 0.82 was obtained, which is higher than that achieved in an experiment using features proposed in a method for prominence detection in stress accent language. In an evaluation experiment to detect prominence using every features, Δ accent was most effective. This result dovetailed with a knowledge that Japanese accent is pitch and a word is emphasized by suppressing pitch accent of words before/after the prominence. Therefore, it was suggested that the proposed method is a one of efficient methods for detecting the prominence.

1 はじめに

音声認識の発展に伴い, 言語情報だけでなく, パラ言語情報の認識が課題となっている. パラ言語とは, 発話に含まれる情報の内, 意図をはじめとする話者が制御できる情報である [1]. 例えば, 否定, 肯定, 戸惑い, 不満といったものがある [1]. この情報は, 円滑な対話の実現のために, 重要である.

パラ言語情報の内, 最も重要なものの一つは強調である. 強調は, 話者が聞き手にもっとも訴えたい部分で行われるため, 重要な情報を含んでいる [2]. 例えば, 語の重要性 [3] や 2 語間の対照性 (e.g. 晴れと雨)[4] や新情報 [5] などの情報を含む. この情報は, 講演の要約 [6] やスキミング [7], 対話における重要な箇所へのアノテーション [8], セグメンテーション [9] 等様々な場面で応用できる.

本稿では, 講演で効果的な強調を習得するために, 講演音声から重要性強調を検出する. 講演には, 論文のキーワードのように重要な語 (名詞) が含まれる. 発表者がそのような語を講演で強調できれば, 聴衆にキーワードを認識させることによって, 講演内容の理解補助となる. このような強調習得までに, 話し方は改善までの過程が視覚的に残らないため, どこが悪いか把握しにくく, 改善が難しい. そのため, 提案法によって, 発表者の講演の発話に含まれる全ての単語の強調/非強調を判定して, 発表者に提示する. 発表者は, 提示された結果から, 強調したい単語が強調されているか, あるいは強調したくない単語が強調されていないかを確認する. 発表者は, 強調したい語が強調に判定されるまでこれを繰り返す. このようにして, 強調習得の支援方法の 1 つとして, 強調検出を用いることを想定する.

2 関連研究

強調検出の近年の手法として, 情報の権限性の高い語の抽出を意図して, フランス語の音声から重要な語の強調を検出する手法が提案されている [10]. 具体的には,F0 とパワーの軌跡の振幅を 1-Q 段階に量子化し, マルコフ過程に基づく統計的な手法を用いて, 強調されている音節を精度 0.78 で検出できる. さらに, この手法では, 音節のセグメンテーションは, 音響特徴量 (波形の包絡) によってなされるため,transcription を作成するコスト削減に至っている. 本研究は, この枠組を拡張し, 日本語の強調検出に有効な特徴量を明らかにし, マルコフ過程に基づき, 強調された音節を検出する.

従来の強調語検出の研究では, 強調検出のための特徴量に関して, 有効な特徴量がアクセントによって変わることが示唆されている. 例えば, 英語/オランダ語/フランス語の対話から強調語を検出した文献 [3, 10, 11] によると, 最も有効な特徴量はパワー最大値である. この知見は, 英語のアクセントが, ストレス (強弱) によって決まる [12] ことと合致する (「morning」では/m/を強く読む). 一方で, 日本語のアクセントは, ピッチ (高低) によって決まる (「オリンピック」では,「ピ」の直後にピッチが下降する). そのため, 従来のストレスアクセントの言語の強調検出で有効な特徴量を用いるのは適切でない. それゆえ, 日本語のアクセントに適した特徴量を用いる必要がある.

Supervisor: Prof. Katunobu Itou

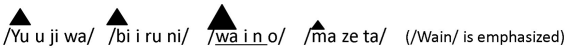

図 1. 強調語を含む発話の単語の強調の程度の変化. 三角形の大きさは, 強調の程度を示す (大：アクセントの山が高められる語, 中：特に高められも抑えられもしない語, 小：アクセントの山が抑えられる語).[13]

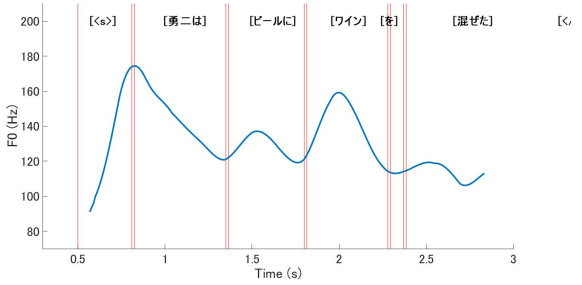

図 2. 強調語を含む F0 軌跡.「ワイン」が強調

3 音声学・日本語教育学の知見に基づく強調の定量化

3.1 音声学・日本語教育学の日本語の強調方法

我々は, 日本語の文中での単語の強調方法に関する知見を得るため, 音声学・日本語教育学の文献 [2, 13, 14] を調査した. その結果, 強調がアクセントの高さ, 語の強さ, ポーズ, 話速, 文中での位置と関連があるとの知見を得た. 具体的には, アクセントの高さに関しては,「強調された語の音調の盛り上がりが増大する」[13]. 強さに関しては,「強調の置かれた語は強く発声される」[14]. ポーズの挿入に関しては,「強調する語の直前, 直後あるいは両方にポーズを置く」[2]. 話速に関しては,「強調したいところをゆっくりいったりする」[13], 発話中での語の位置に関しては,「重要な語を文中で先に示す」[15], といったことである.

また, 発話中の語を強調する際に, 強調する語以外の語も影響を受ける, との知見を得た. 具体的には,「フォーカスのある語は, 語アクセントによる音調の山が高くなり, 以後の語群はアクセントによる音調の山が抑えられる」[13]. 「フォーカスが感じられるのは強調部分に比べて後部要素のピッチが相対的に高く現れているからである」[14]. 「フォーカス語を大げさに際立たせるような場合には, フォーカス前の語群のアクセントも抑えられることがある」[13] といったことである. これらの知見を表した, 音声学の図 の例を図 1 に示す. この図は,「勇二がビールにワインを混ぜた」という発話において, 強調された語 (ワイン) とその他の語の強調の程度を示す. 下線の引かれた語が強調されており, 語の上の三角形は, 語の強調度を示す (大：アクセントの山が高められる語, 中：特に高められも抑えられもしない語, 小：アクセントの山が抑えられる語). この図より, 強調された語に隣接するアクセント句が抑えられていることが分かる.

3.2 強調の定量化

3.1 で得た文中の語の強調方法の知見を元に, アクセントの高さ, 語の強さ, 語の直前のポーズ長, 語の直後のポーズ長, 話速, 語の文中の位置を定量化する. さらに, 強調語を含む発話では

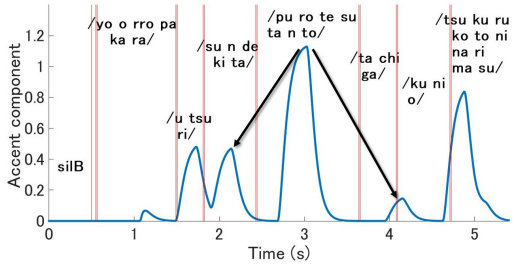

図 3. 強調語を含む発話のアクセント成分.「プロテスタント」が強調

隣接する語の強調が抑えられる, という知見に基づいた強調判定手法を提案する.

アクセントは F0 に対し, 藤崎モデル [17] を仮定して, 各単語のアクセント成分の最大値を推定して求める. アクセント成分を推定する理由は, 日本語の発話のピッチの特徴に起因する.

図 2 に強調語を含む F0 の例を示す. この図は, 図 1 の上から 3 段目の図のように,「勇二がビールにワインを混ぜた」という発話中で,「ワイン」を強調した発話の F0 である. この図より, 強調された「ワイン」に対応する F0 の最大値が, 発話中で最大になっていないことが分かる. これは, 日本語の発話の F0 が, 文頭から文末に向けてなだらかに下降する特徴 [17] に起因する. これによって, アクセントの起伏が不明瞭になっている. そのため, 強調検出のために,F0 の値を直接使うのは不適切である. よって,F0 が下降する影響を除去し, アクセントの起伏を表すアクセント成分のみを推定する.

語の強さは, 語の対数パワー最大値を推定することで定量化する. ポーズは, 語の直前・直後のそれぞれのポーズの長さを推定することで定量化する. 話速は,2 つの方法で定量化する.1 つめの方法は, 単語の継続長によって定量化する. これは, 話す話速が遅くなれば, 継続長が長くなるからである. 単語継続長は, 単語の終了時刻と開始時刻継続の差を推定することで定量化する. 2 つめの方法は, 単語内の 1 音節あたりの継続長によって定量化する (i.e., 平均話速). 語の文中での位置は, 発話長を 1 とする単語の開始位置, 終了位置を求める.

さらに, 強調語の直前・直後のアクセントの抑制に関しては, 強調語直前のアクセント句内のアクセント成分最大値と強調語のアクセント成分の振幅最大値の変化量, 強調語直後のアクセント句内のアクセント成分の振幅最大値と強調語のアクセント成分最大値の変化量に着目する. 図 3 に強調語を含む発話のアクセント成分の例を示す. この図は,「ヨーロッパから移り住んできたプロテスタント達が国を作ることになります」という発話中で「プロテスタント」を強調した時のアクセント成分の軌跡である. 直前のアクセント句である「住んできた」と直後のアクセント句である「国を」は, 強調語「プロテスタント」に比べ, アクセントの高さが抑えられていることが分かる. そのため, 語が強調されていれば, 強調語のアクセント成分の振幅最大値と直前/直後のアクセント句内のそれぞれのそれとの差分 (i.e., 変化量) が大きくなる.

3.3 強調検出のための特徴量推定

アクセント成分の推定について説明する. アクセント成分を推定するには, 矩形波で表されるアクセント指令を推定する必

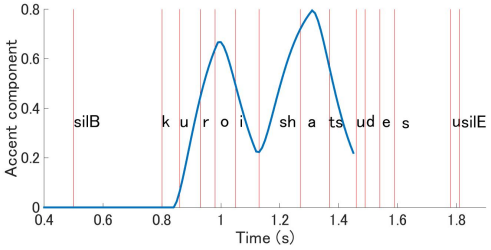

図 4. アクセント成分推定例

要がある [18]. アクセント指令は, 矩形波であり, 指令開始時刻, 指令終了時刻, 振幅の 3 つのパラメータによって決まる. そのため, これらのパラメータを推定する必要がある. これらのパラメータ推定のために,[19] の手法より求めた F0 を線形補間し, メディアンフィルタで平滑化したあと, 区分的に 3 次曲線でフィッティングして, 時間方向に 1 次微分して得られた極値の値を振幅, 極値の得られた時刻を指令開始時刻, 指令終了時刻とする. アクセント指令 $u_{\mathrm{a}}(t)$ からアクセント成分 $y_a(t)$ は式 (1) で得る.

$$y_a(t) = G_{\mathrm{a}}(t) * u_{\mathrm{a}}(t), \tag{1}$$

ただし,$G_a(t)$ は以下の式で表される.

$$G_a(t) = \begin{cases} \beta^2 t \exp(-\beta t) & (t \geq 0) \\ 0 & (t < 0). \end{cases} \tag{2}$$

ここで,t は時間を表す. また,β は最大値に到達するまでの傾きを表すが, 個人性が小さいことが示されているため, 定数 β=20.0 rad/s とできる [20]. 推定されたアクセント成分から, 単語開始時刻から終了時刻までのアクセント成分の最大値を特徴量とする.

事前実験として, この手法を用いたアクセント成分推定の性能を評価した. これまで, 評価データには, 日本語教育学の文献の CD[23, 24] に収録されている強調語をそれぞれ 1 単語だけ含む 11 発話 (女性話者 2 人, 男性話者 3 人) を用いた. これらの発話は, すべて肯定形である.

これらの語の開始時刻から終了時刻の範囲において, アクセント成分が, 推定されているかどうかを評価する. 具体的には, アクセント成分のピークが, 単語の開始時刻から終了時刻までの範囲に位置していれば, 推定できているとした. これは, 単語のアクセント核の位置でアクセント成分が最大になるからである. 評価尺度は精度を用いる. 算出方法を式 (3) に示す.

$$\text{検出精度} = \frac{\text{アクセント成分が検出された語}}{\text{発話の全ての単語}}. \tag{3}$$

実験の結果, 提案法は検出精度 1.00 となった.

図 4 に推定例を示す. この発話は,「/kuroi shatsu desu/」である. この発話は,「/kuroi/」と「/shatsu/」の 2 単語を持つ. また,「/shatu/」が強調されている.

比較手法は,F0 の平滑化において, 移動平均フィルタを用いる手法 [18] とした. 評価尺度は, 平滑化後の F0 軌跡の有声区間の値と, 合成 F0 軌跡 (i.e., アクセント成分 + フレーズ成分 + ベースライン成分) のそれとの RMSE (Root Mean Square Error) とする. RMSE の値は,F0 へのフィッティングの良さを表す. つまり, この値が小さいほど, 実際の F0 にフィッティングした成分が推定できている. RMSE は, 以下の式で計算する.

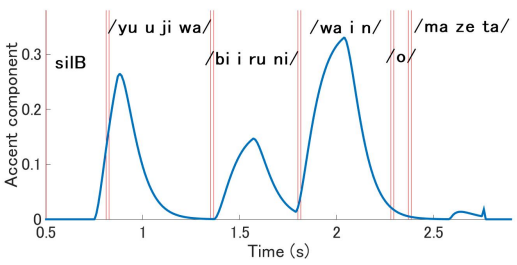

図 5. 強調語を含むアクセント成分軌跡.「ワイン」が強調

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{k=1}^{N}(\log F0 - \log F0_{\mathrm{synth}})^2}. \tag{4}$$

ただし,N は対数 F0 の有声区間のフレーム数,$\log F0$ は対数 F0 軌跡の有声区間の値,$\log F0_{synth}$ は, 対数合成 F0 軌跡の有声区間の値を表す. 実験の結果, 提案法における RMSE の平均は,0.19 となった. また, 比較手法における RMSE の平均は,0.39 となった. さらに, 平均の差を検定した結果, 有意差が確認された (危険率 0.05). よって, 提案法は, 比較手法に比べ, 高い性能で成分を推定できている.

図 5 に図 1 に示した F0 からアクセント成分を推定した結果を示す. アクセント成分は, 語のアクセントに起因する成分のみを表しているため, 強調語である「ワイン」で最大値となっている事がわかる.

パワー E は, 式 (5) で計算する.

$$E = \sum_{m=0}^{L-1}(s[(rR+m]\cdot w[m])^2, r = 0, 1, 2, ..., \tag{5}$$

ここで,r は各フレームの番号,L,R をそれぞれ窓長とシフト幅のサンプル数を表す. ただし, 音声の収録された環境による違いの影響を除去するために, 式 (6) で発話中の最大値が 1 になるように正規化する. さらに, 対数を取る.

$$E_{\mathrm{norm}} = \log_{10}\frac{E}{\max(E)}. \tag{6}$$

図 6 に図 1 の発話からパワーを推定した結果を示す. 日本語は, ストレスアクセントの言語とは異なり, ピッチアクセントであるため, パワーでは強調部分 (「ワイン」) とその他の部分で大きく差が出ていない.

ポーズ長の推定には, 背景雑音のみ含むフレームから得たゼロクロスと対数パワーのしきい値による VAD (Voice Activity Detection) を行い, 音声以外の区間をポーズとみなし, ポーズが挿入された開始時刻と終了時刻を求める [25]. そして, 単語の開始時刻, 終了時刻のそれぞれ直前と直後に位置するポーズがあれば, それを直前, 直後に挿入されたポーズとみなし, ポーズ終了時刻から開始時刻を引くことでポーズ長を計算する.

単語の継続長は, 単語の終了時刻から開始時刻を引くことで求める. 発話中における単語の開始位置と終了位置は, 発話長を 1 に正規化したときの相対位置として推定する.

提案する静的特徴量を強調語から推定した結果を図 7 に示す (経済学という学問です, という発話の分析結果). この図において,F は対数 F0 最大値,A はアクセント成分,Po は対数パワー最

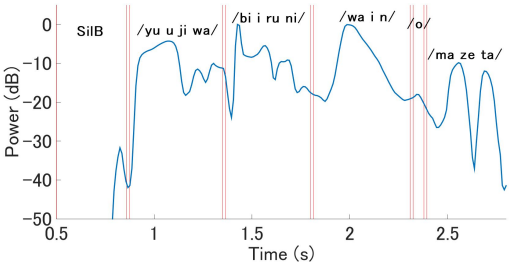

図 6. 強調語を含むパワー軌跡.「ワイン」が強調

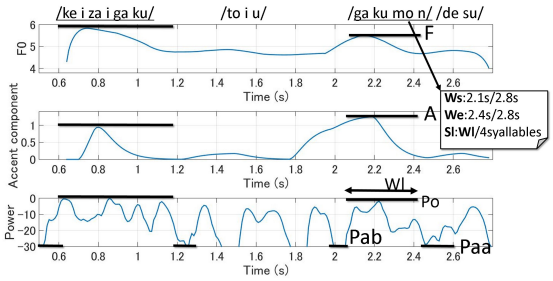

図 7. 音響特徴量の例 (F:対数 F0 最大値 A:アクセント成分の振幅最大値 Po:対数パワー最大値 Pab:単語の直前のポーズ長 Paa:単語の直後のポーズ長 Wl:単語の継続長 Ws:発話長を 1 とした時の単語の開始位置 We:発話長を 1 とした時の単語終了位置 Sl:1 音節あたりの継続長)

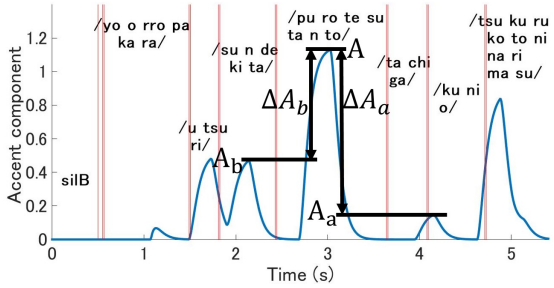

図 8. Δ アクセントの推定例

大値,Pab は単語の直前のポーズ長,Paa は単語の直後のポーズ長, Wl は単語の継続長,Ws は発話長を 1 とした時の単語の開始時刻,We は発話長を 1 とした時の単語の終了位置,Sl は 1 音節あたりの継続長を示す.

強調語のアクセント成分の振幅最大値と前後のアクセント句内のアクセント成分の差分を計算する. 強調語と直前/直後のアクセント成分の振幅最大値の差分は, 直前/直後の差分を ΔA_b, ΔA_a とすると, 式 7,8 で計算される.

$$\Delta A_b = A - A_b. \tag{7}$$

$$\Delta A_a = A - A_a. \tag{8}$$

ただし,A は, 強調語のアクセント成分の振幅最大値, A_b は強調語直前のアクセント句内のアクセント成分の振幅最大値,A_a は強調語直後のアクセント句内のアクセント成分の振幅最大値を示す. これらを計算した例を 8 に示す.

これらの動的特徴量を加えると, 提案する特徴量は, 計 11 次元となる.

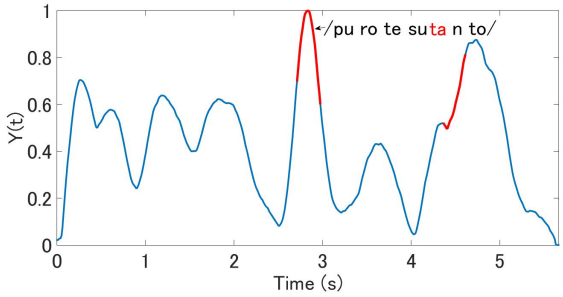

図 9. 強調検出結果 (赤線: 検出された強調音節)

4 実験

4.1 データセット

我々は, 提案法による強調語検出手法の精度を評価するために, 強調語を含む発話をテレビの講演より収集した. これは, 従来の講演データベースを調査した結果, 本研究で扱う効果的な強調がなされていなかったためである.

収集した講演は,1 回あたり 15-20 分程度のもので, 経済学や時事問題等をテーマにしており, 話者 5 名 (男性 3 名, 女性 2 名) で行われている. これらの講演を男性 1 名が聴取し, 強調語を含む発話を切り出した.

強調/非強調のラベルが単語レベルで付いた 20 発話 (強調 32 単語, 非強調 32 単語) に対し, 正しく強調検出できるか実験する. これらの発話の強調ラベルは 1 名によって付与された.

4.2 実験条件

音声は 48 kHz で記録され, 分析時には 16 kHz にダウンサンプリングする.F0 とアクセント成分とパワーはフレーム長 25 ms, フレームシフトは 10 ms で計算する. 窓関数は hann 窓を用いる. 特徴選択の事前実験 [26] の結果，提案法では，強調判定のためにアクセント成分の振幅最大値，Δ アクセントのみ用いる.

比較手法は,3 PRO のオリジナルの特徴量 (F0, パワー,spectrum tilt) を用いる方法 [10] とした.3PRO のパラメータは, 論文で最も良い精度となった $Q = 8$,$L = 0.2$ とした. 評価尺度は, 論文に記載されている精度とした. 具体的には, データに単語レベルで強調/非強調のラベルが付いており, 強調としてラベルされた単語において, 音節の位置は問わず, どこかの音節が強調と判定されていれば, 検出できたとみなす.

4.3 結果と考察

実験の結果, 提案法では, 精度が 0.82, 比較手法では 0.42 となった. 提案法による強調音節を推定した結果を図 9 に示す. この例は, 図 8 の発話から, 強調音節を検出した結果である. 検出された音節を聴取した結果,「プロテスタント」の「タ」の区間が強調として検出されていた.

強調語判定の精度が従来法に比べ向上したことに関して, 正しく判定できた強調語の ΔA_b の平均と, 正しく判定できた非強調語の ΔA_b の平均の差を検定した. その結果, 前者の平均は 0.18, 後者の平均は 0.01 となり, 有意差があった (危険率 0.05). さらに, 正しく判定できた強調語の ΔA_a と, 正しく判定できた非強調語の ΔA_a の平均の差を検定した. その結果, 前者の平均は 0.18, 後者の平均は -0.02 となり, 有意差があった (危険率

0.05). これは, 強調時には, 強調語前後のアクセント成分の振幅最大値が, 強調語のそれに比べ, 低くなることを示している. これは, 発話中の語を強調する際に, 前後の語のアクセントを抑える, といった音声学・日本語教育学の知見 [13, 14] と合致する.

また, 判定が改善しなかった 7 単語に関して, これらの語の ΔA_b の平均と正しく判定された強調語の ΔA_b の平均の差を検定した. その結果, 前者は -0.06 となっており, 有意差があった (危険率 0.05). さらに, 判定が改善しなかった語に関して, これらの語の ΔA_a の平均と正しく判定された強調語の ΔA_a の平均の差を検定した. その結果, 前者は 0.01 となっており, 有意差があった (危険率 0.05). これは, 強調語の直前/直後のアクセント成分の振幅最大値が抑えられていなかった語の判定が改善しなかったことを示している. 実際にこれらの発話を聴取したところ, 強調語の直前/直後の語も強調されて聞こえた. このような語を判定するためには, 強調語を中心として, 周囲数単語分の特徴量を判定に用いるか検討する必要がある. これは, 音声学において, 強調語前の語群, 強調語以後の語群のアクセントが抑えられる, との知見があるからである. 実際に聴取した結果, 強調語を含め, 周囲 3 単語必要であった.

5 演技音声における感情表現解析への応用

5.1 Δ アクセントによる演技音声の感情分析

Δ アクセントによるアクセント, 強調表現の解析手法の枠組みを応用し, ドラマ/アニメーション (アニメ) における声優や俳優の感情表現解析を検討した. 従来研究において, 感情とアクセントの関連が示唆されている. 具体的には,F0 パターン [1] やアクセントによる F0 下降のタイミング [1] やアクセント指令の振幅 [21] を分析した結果, 感情のカテゴリごとに異なる結果を得ている. しかし,Δ アクセントに相当する特徴量を解析した研究はない. また,Δ アクセントは, 日本語の強調を捉えるのに有効なモデルの 1 つであるが, 強調と感情の関連もまた示唆されている. 具体的には, 怒りと喜びは, 悲しみに比べ, 強調に聞こえやすい [22] といったことである. この知見は,Δ アクセントが感情表現の解析にも応用可能である可能性を示唆している.

感情音声の解析手法の応用として, 感情の識別や, 声/俳優向けの感情表現の演技練習の支援や, 感情ごとにアクセントのパラメータを保存しておき, 従来の音声合成のための F0 生成モデルにこれを組み込むことで, ドラマ/アニメーション向け音声合成器に応用すること等が考えられる.

5.2 データセット

我々は,「のだめカンタービレ」のアニメ/ドラマから, 声/俳優の発話データをそれぞれ収集した. この作品は, アニメ/ドラマともに, 原作の漫画に忠実に制作されており, 同一の状況・コンテキスト下での, ほとんど同一の声/俳優の発話セットが得られる. これにより, 声/俳優の強調表現の比較が可能となる.

収集の結果, ドラマ (1 回約 50 分程度) とアニメーション (1 回約 20 分程度) の第 1-2 話から, 発話データを 25 セット収集できた. これらの発話セット以外に, 話者のオーバーラップがあるものや効果音を含むものが 41 セットあったが, これらの発話セットは音響特徴が正しく計算できなくなるため, 除外した. ただし,Back Ground Music (BGM) に関しては,REPET-SIM [27] を適用して, 音声と BGM が畳み込まれた素材から音声のみを抽出することで, 分析データとして採用する. 収集された発話は, 男性話者 5 人, 女性話者に 4 人によって発話された. アニメから収集された発話データにおいて, 最も短い発話継続長が 2.83 s, 最も長い発話継続長が 5.74 s, 全発話の平均継続長が 4.15 s となった. ドラマから収集された発話データにおいて, 最も短い発話継続長が 2.49 s, 最も長い発話継続長が 6.98 s, 全発話の平均継続長が 4.00 s となった.

これらの発話に感情のラベルを付けた. ラベル情報は, 感情認識の研究 [22] を参考にして平静, 怒り, 悲しみ, 喜びとした. ラベル情報は, 男性 1 名によって付けられた.Table 1 に付与されたラベル情報とラベルを付与された発話数を示す.

表 1. 感情ラベルと発話数

ラベル	平静	怒り	悲しみ	喜び
発話数	11	6	3	5

5.3 実験条件

感情表現を含む発話における, 声/俳優のアクセント成分,Δ アクセントを分析する. 音声は 48 kHz で記録され, 分析時には 16 kHz にダウンサンプリングする.F0 とアクセント成分はフレーム長 30 ms, フレームシフトは 10 ms で計算する.

5.4 結果

各感情ごとの, 声/俳優の発話の特徴量 (i.g., 発話に含まれる全てのアクセント句のアクセント成分 (A) の平均と最大値, それらの句の ΔA_b と ΔA_a の最大値, 及び基底周波数 B) を分析し, 平均の差を t 検定により, 比較した. その結果, 有意差があったのは, 喜びにおける ΔA_b 最大のみとなり, 他の感情における声優と俳優間の特徴量では, 平均の差に違いはなかった (危険率 0.05). 喜びにおける声/俳優の特徴量の比較結果を 表 2 に示す. 表 2 より, 声優の発話の ΔA_b 最大値は, 俳優のものより

表 2. 喜びの発話の声/俳優の特徴量の比較

特徴量	声優	俳優
A 平均	0.44	0.35
B 平均	4.93	4.89
A 最大	0.83	0.73
ΔA_a 最大	0.26	0.34
ΔA_b 最大	0.36	-0.20

0.56 も大きく, 声優が, アクセントを抑える, といった制御をしている事がわかる. また, 実際に音声を聴取した結果, 俳優に比べ, 声優の発話のほうが強調の度合いが強く, 感情の度合いも強いことを主観的に確認した.

カテゴリ「喜び」における, 声/俳優の発話の分析結果の比較例を図 10 に示す. この発話は, 女性話者によるものである. 発話内容は, 俳優は「でも胸がキューンて苦しい」, 声優は「胸がキューンて苦しいです」である. これらの発話では,「キューン」が強調されている. 上段は対数 F0, 中段はアクセント成分, 下段はフレーズ成分を示す. F0 の 2 つの成分の内, フレーズ成分は, 短時間に急激に上下する成分ではないため,F0 を急激に変化させるには, アクセント成分を制御する必要がある. アクセント成分を比較すると, 俳優は,「キューン」に対応するアクセント成分の高さが発話中で最大になっていない. 一方, 声優は,「キューン」の前後の語のアクセント成分の高さを抑えること

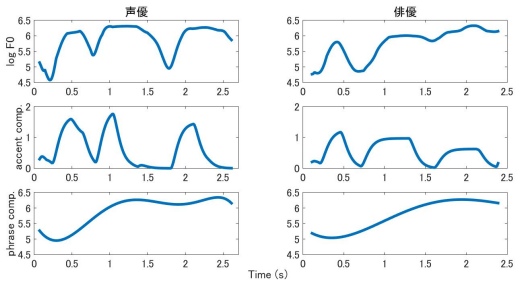

図 10. 声/俳優の強調表現分析結果の例

で強調していることが分かる. さらに, 俳優のフレーズ成分を見ると, 文頭から文末にかけて上昇している. つまり, 声優は語のアクセントを制御することで F0 を上昇させているが, 俳優はフレーズ成分を制御することで,F0 を上昇させている.

また, すべての感情カテゴリにおいて, 声優/俳優間で A に違いが見られなかったことに関しては, 「同一の感情の発話であっても, アクセント成分の振幅は感情の強さに影響されない」[21] といった知見と合致する.

一方で,「喜び」において, 俳優に比べて声優の方が強く強調しており感情の度合いもまた強く聞こえたものの, 声/俳優間で B に有意差がみられなかったことは,「喜びでは, 感情の強さが強くなるに従って基底周波数が単調に増加する」 [21] といった知見とは異なる結果となった. これに関して,[21] では, 話者が演劇経験のある男女それぞれ 2 名が感情の度合いを意図して 3 段階に使い分けて録音していた. その結果, 基底周波数の平均は, 弱で 1.05, 中で 1.10, 強で 1.20 と変化している. 一方, 表 2 の基底周波数 B をみると, 声優で 4.93, 俳優で 4.89 となっており, はるかに高い値となっている. よって, 声優や俳優の発話における強い感情表現では, 基底周波数は十分に高いため, 基底周波数以外の要因である ΔA_b によって, 感情の度合いを制御している可能性がある. 今後は, 声/俳優の感情の度合いに影響を与えるパラメータを明らかにするため, 発話を感情の度合いごとに分類し, 特徴量を分析する.

6 おわりに

本研究では, 日本語の講演練習システムにおいて, 強調習得を支援するための強調語検出手法を提案した. そのために, 我々は, 音声学や日本語教育学の文献における単語の強調方法に関する知見に基づき, 強調を定量化し, 強調検出に有効な特徴量を提案した. 具体的には, 強調検出のために, アクセント成分とそれのデルタ特徴量を用いることを提案した. 強調検出の実験では, 検出精度 0.82 となり, 提案した特徴量が,「強調語の前後のピッチアクセントの抑制」といった現象をよく捉えたモデルの 1 つであることが示された.

参考文献

[1] K. Maekawa, Production and perception of paralinguistic information, *Proc. SP*, pp. 367-374, 2004.

[2] 中条, 日本語の音韻とアクセント, 勁草書房, 1989.

[3] V. R. Sridhar et.al. Detecting prominence in conversational speech: pitch accent, givenness and focus, *Proc. SP*, pp. 453-456, 2008.

[4] L. Chunrong et.al. Detection and emphatic realization of contrastive word pairs for expressive text-to-speech synthesis, *Proc. ISCSLP*, pp. 93-97, 2012.

[5] S. Calhoun et.al. A framework for annotating information structure in discourse, *Proc. ACL work shop*, pp. 45-52, 2005.

[6] F. R. Chen et.al. The use of emphasis to automatically summarize a spoken discourse, *Proc. ICASSP*, pp. 229-232, 1992.

[7] B. Arons, Pitch-based emphasis detection for segmenting speech recordings, *Proc. ICSLP*, pp. 1931-1934, 1994.

[8] S. Kakouros et.al. Automatic detection of sentence prominence in speech using predictability of word-level acoustic features, *Proc. INTERSPEECH*, pp. 568-572, 2015.

[9] S. A. Moubayed et.al. Prominence detection in Swedish using syllable correlates, *Proc. INTERSPEECH*, pp. 1784-1787, 2010.

[10] S. Kakouros et.al. 3PRO - An unsupervised method for the automatic detection of sentence prominence in speech, *Speech Communication*, Vol.82, pp. 67-84, 2016.

[11] B. B. Jason et.al. The detection of emphatic words using acoustic and lexical features, *Proc. EUROSPEECH*, pp. 3297-3300, 2005.

[12] J. Pierrehumbert et.al. Japanese tone structure, The MIT press, 1988.

[13] 郡, 講座 日本語教室 2, 明治書院, 1989.

[14] 中川, 初級文型でできる日本語発音アクティビティ, アスク出版, 2010.

[15] 外山, 伝わる話し方のための 10 のルール, *Bulletin of aichi shukutoku university*, Vol. 32, pp. 55-64, 2007.

[16] H. Fujisaki, Information, prosody, and modeling-with emphasis on tonal features of speech, *Proc. SP*, pp. 1-10, 2004.

[17] H. Fujisaki et.al. A model for synthesis of pitch contours of connected speech, *Annual report, engg. res. inst., university of Tokyo*, Vol. 28, pp. 53-60, 1969.

[18] S. Narusawa, et.al. Evaluation of a method for automatic extraction of parameters of the fundamental frequency contour generation model, *IPSJ Journal*, Vol. 75, pp. 1-6, 2003.

[19] M. Morise, et.al. Fast and reliable F0 estimation method based on the period extraction of vocal fold vibration of singing voice and speech, *Proc. IES*, CD-ROM, 2009.

[20] H. Fujisaki et.al. Analysis of voice fundamental frequency contours for declarative sentences of Japanese, *J.acount.soc. Jpn*, Vol. 5, No. 4, pp. 233-242, 1984.

[21] 大野, 韻律的特徴の総合的なモデル化と, 感情の表現・伝達過程, 特定領域研究「韻律と音声処理」研究成果報告書, 2005.

[22] K. Jangwon et.al. Relations between prominence and articulatory-prosodic cues in emotional speech, *Proc. SP*, pp. 367-374, 2016.

[23] 田中, 日本語の発音教室, くろしお出版, 1999.

[24] 河野, 1 日 10 分の発音練習, くろしお出版, 2004.

[25] L. R. Rabiner et.al. An algorithm for determining the endpoints of isolated utterances, *The bell system technical journal*, Vol. 54, No. 4, pp. 297-315, 1975.

[26] 小島, アクセント成分を用いた講演の強調語検出, 情報処理学会第 78 回全国大会 Mar. 2017.

[27] Z. Rafii, Music/Voice Separation using the Similarity Matrix, *13th International Society on Music Information Retrieval*, pp. 8-12, 2012.