7M-06

アクセント成分を用いた講演の強調語検出 *

小島 淳嗣 (法政大学大学院 情報科学研究科), 伊藤 克亘 (法政大学 情報科学部),

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, 書き起こしできない情報の内, 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 [7]. 例えば, 語の重要性 [2] や 2 語間の対照性（e.g. 晴れと雨） [3] や新情報 [2] などの情報を含む. この情報は, 講演の要約 [4] や対話における重要な箇所へのアノテート [5], 講演のスキミング [6] など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 音声学・日本語教育学の知見に基づく強調の定量化

2.1 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 [7, 8, 9] 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」 [8]. 強さに関しては,「強調の置かれた語は強く発声される」 [9]. 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」 [7]. 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」[8], 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」[10], 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」 [8]. 「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」 [9]. 「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」 [8] といったことである.

2.2 強調の定量化

2.1 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.

* : Prominence detection in a presentation using an accent component *Atsushi Kojima (Hosei Univ.) et al.*

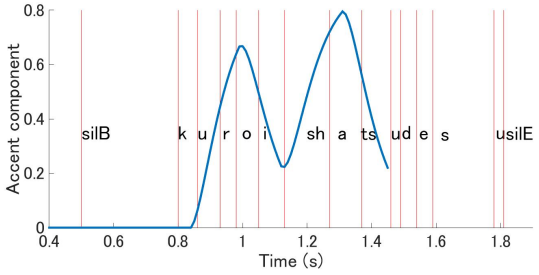

図 1. アクセント成分推定例

アクセントは F0 に対し, 藤崎モデル [11] 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. ただし, 強調語検出には F0 最大値も特徴量に含める. 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. 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.

事前実験として, この手法を用いたアクセント成分の推定精度を評価した. 評価データには, 日本語教育学の 2 つの文献の CD [13, 14] の発話を用いた. これらのデータは, 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) である. これらの発話は, すべて肯定形である. 対象とする語は, 発話に含まれる全ての語 (35 単語) とした. 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. 比較手法は，英語の音声合成のための F0 生成モデルである TILT [15] 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 (1) に示す.

$$\text{検出精度} = \frac{\text{アクセント成分が検出された語}}{\text{発話の全ての単語}}. \qquad (1)$$

実験の結果, 提案法は検出精度 1.00，TILT は 0.86 となった.

Fig. 1 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.

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の振幅の最大値と強調語のアクセント成分の振幅の最大値の変化量, 強調語直後のアクセント句内のアクセント成分の振幅の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.

Δ アクセントは，強調語のアクセント成分の振幅の最大値と前後のアクセント句内のアクセント成分の振幅の最大値との差分によって得る. 強調語と直前/直後のアクセント成分の振幅の最大値の差分は, 直前/直後の差分を ΔA_b, ΔA_a とすると, 式 (2) で計算される.

$$\Delta A_b = A - A_b,\ \Delta A_a = A - A_a. \qquad (2)$$

ただし,A は, 強調語のアクセント成分の振幅の最大値, 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の最大

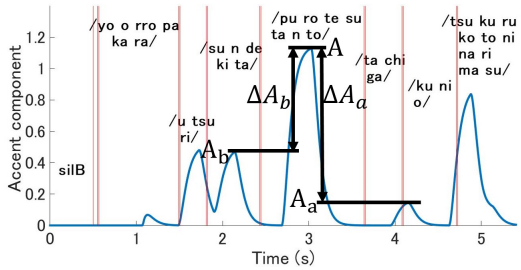

図 2. Δ アクセントの推定例

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の最大値を示す．これらを計算した例を 2 に示す．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の最大値と直前/直後のアクセント句内のそれぞれのそれとの差分 が大きくなる．

これらの動的特徴量を加えると, 発話中の強調語を判定するために用いる特徴量は, 計 11 次元となる．これらの特徴量をまとめたものを表 1 に示す．

表 1. 音響特徴量

特徴量
アクセント成分振幅最大値
F0/パワー最大値
直前/後のポーズ長
単語/音節継続長
単語の開始/終了位置
直前/後の Δ アクセント

3 実験

3.1 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収集した．我々は, これまでに, 強調語含む講演を収集してきた [12].

これらの講演は，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．その結果,

3.2 実験

提案法による強調検出の有効性を検証するため, 収集した発話を利用して評価実験を行う．強調/非強調のラベルが付いた 64 発話（強調 32 発話, 非強調 32 発話）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．とした．評価尺度に用いる適合率・再現率を式 (3)-式 (5) に示す.

$$\text{適合率} = \frac{\text{検出された強調単語数}}{\text{検出された単語数}}. \tag{3}$$

$$\text{再現率} = \frac{\text{検出された強調単語数}}{\text{強調単語数}}. \tag{4}$$

$$\text{精度} = \frac{\text{検出された強調単語数} + \text{検出された非強調単語数}}{\text{強調単語数} + \text{非強調単語数}}. \tag{5}$$

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．パワーはフレーム長 50 ms, フレームシフトは 10 ms で計算する．窓関数はともに hann 窓を用いる．音響特徴量は,Support Vector Machine (SVM) で学習する．カーネルは, ガウス関数を使用する．

さらに，前向き特徴選択 [17] によって，有効な特徴量を選択する．この手法では, まず, 全ての特徴量の集合の中で, 単体で最も検出精度の高い特徴量を選択する．次に, 既に選択された特徴量以外のものの中で, 既に選択した特徴量と組み合わせて精度が最大になるものを選択する．選択する特徴量がなくなるか, 追加前より精度が下がるまで, この処理を繰り返す．これにより，有効な特徴量を優先的に選択される．

3.3 結果と考察

実験の結果, 提案法では, 適合率が 1.00, 再現率が 1.00, 精度が 1.00 となった．また, 特徴選択の結果，選択された特徴量は，それが選択された順番がはやい順に列挙すると，ΔA_b，A，F0 最大値，ΔA_a，単語継続長，直後のポース長，音節継続長となった．

強調語判定の精度が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の最大値が, 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 [8, 9] と合致する．

4 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．まずはアクセント成分の推定精度が十分であるかどうかを評価した．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．まずはアクセント成分の推定精度が十分であるかどうかを評価した．

強調検出の実験では, 検出精度 1.00 となり，提案法が, 日本語講演の強調検出に有効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] V. R. Sridhar et.al. Detecting prominence in conversational speech: pitch accent, givenness and focus, *Proc. SP*, pp. 453-456, 2008.

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

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

[5] Z. Malisz et.al. Acoustic-phonetic realisation of polish syllable prominence: a corpus study of spontaneous speech, *Rhythm, melody and harmony in speech. Studies in honour of Wiktor Jassem*, Vol. 14, no. 15, pp. 105-114, 2012.

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

[7] 中条, 日本語の音韻とアクセント, 勁草書房, 1989.

[8] 郡, 講座日本語と日本語教育, 明治書院, 1989.

[9] 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.

[10] 富山, 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.

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

[12] A. Kojima et.al. Prominence Detection for Presentation Training System, *Proc.SOiCT*, pp. 316-322, 2016.

[13] 田中, 日本語の発音教室, くろしお出版, 1999.

[14] 河野,1 日 10 分の発音練習, くろしお出版, 2004.

[15] P. Taylor, The tilt intonation model, *Proc.ICSLP 98*, pp. 1383-1386,1998.

[16] 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.

[17] P.Pudil et.al. Floating search methods in feature selection, Pattern Recognition Letters, vol.15, pp.1119-1125, 1994.