1C1-06

音響情報を利用した言い直しの検出

Detection of Japanese Self-Collections using Acoustic Information

船越 孝太郎[*1] 鈴木 泰裕[*2] 徳永 健伸[*1] 田中 穂積[*1]
Kotaro FUNAKOSHI Yasuhiro SUZUKI Takenobu TOKUNAGA Hozumi TANAKA

[*1]東京工業大学大学院 情報理工学研究科
Graduate School of Information Science and Engineering, Tokyo Institute of Technology

[*2]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

Detecting self-corrections or speech-repairs is an important task in spoken dialog systems. This paper reports experimental result to detect self-correction by using only acoustic information, such as pause length, pitch and energy. Our detection method is based on pause detection, because 99% of self-corrections entail a pause at their interruption sites. The CART and SVM method are used to decide if a detected pause is a self-correction or not. We achieved a recall of 77.9% and a precision of 2.58%, showing the limitation of self-correction detection with only acoustic information.

1. はじめに

音声対話を実現しようとするときに問題となる言語現象の一つに，自己修復，いわゆる言い直し (self-correction, speech-repair) がある．言い直しは文法の適格性を破壊し，計算機による言語理解を妨げる要因となるため，適切な処理が必要であり，Hindle の研究を始めとして多くの研究が行なわれてきた [Hindle 83, Bear 92, O'Shaughnessy 92, Nakatani 93, Kikui 94, 佐川 94, Heeman 97, 伝 97, 中野 98, Core 99, Spilker 00, Funakoshi 02].

言い直しは，どの研究でも概ね，Nakatani らの Repair Interval Model(RIM)[Nakatani 93] に類するモデルで定式化される (cf.[Funakoshi 02])．RIM では，言い直しの発生点とみなす Interruption Site(IS: 図 1 参照) を中心として，REPARANDUM，DISFLUENCY，REPAIR の連続する 3 つの区間 (それぞれ，図 1 において $[\ldots]_{rpd}$，$[\ldots]_{df}$，$[\ldots]_{rp}$ で表された区間) に分けて言い直しを捉える．IS は REPARANDUM の終端であり，DISFLUENCY の始端である．

じゃ，$[3\text{時}]_{rpd}$ IS $[\text{えっと ごめん}]_{df}$ $[4\text{時の}]_{rp}$ 電車で

図 1: Repair Interval Model [Nakatani 93]

言い直しの処理は，(1) 言い直しの検出 (IS の位置の特定)，(2) 表層上での言い直しの範囲の特定 (REPARANDUM 区間の始端，および REPAIR 区間の終端の特定)，(3) 言い直しの修正 (REPAIR 区間，DISFLUENCY 区間の削除)，の 3 段階で行なわれる．本稿では，最初の段階の言い直しの検出を，音響情報のみを用いて日本語音声対話コーパス上で行なった実験の結果について報告する．音響情報としてはポーズ，パワー，ピッチの 3 種類を利用した．

2. 言い直しの検出

言い直しは，単語列上でのパターンマッチング [Bear 92, Kikui 94] や構文的な情報 [Hindle 83, 佐川 94, 伝 97, 中野 98, Funakoshi 02]，n-gram 言語モデルへの組み込み [Heeman 97] などを用いることである程度の検出が可能だが，言い直しが発声した位置によって複数の解釈が可能な場合 (例 (1)) や，省略あるいは音声認識の誤りによる助詞の脱落が起きた場合 (例 (2)) などは，言語情報だけで正確に言い直しの場所を特定することはできない．

(1) . . . arriving Fort Worth twenty two twenty one forty . . .
(21:40, 22:21, 22:40 の 3 つの解釈が可能 [Nakatani 93])

(2) 机，台に載せて
(机を台に載せるのか，何かを机ではなく台に載せるのか)

言い直しが発生するとき，IS とその周辺には，発話の中断による単語の断片化，短めのポーズ，パワー (エネルギー) やピッチの急激な変化，発声速度の変化などの音響的・音韻的な特徴が観察できる [Bear 92, O'Shaughnessy 92, Nakatani 93, 神田 96]．Bear らは，パターンマッチングによって検出された箇所の音響的特徴を手動で検査し，音響・音韻情報がパターンマッチングによる誤検出の識別に有効であることを示している [Bear 92]．また Nakatani らは，人手で書き起こされたコーパスに，ポーズ長，F0(ピッチに相当)，エネルギーなどの情報を同じく人手で付与し，そこから得られる素性の集合に対して CART(Classification and Regression Trees) 法 [Breiman 84] 適用することで書き起こしコーパス上で言い直しを検出する分類木を生成し，83.4%の再現率と 93.9%の適合率を得ている [Nakatani 93]．Spilker ら [Spilker 00] は言い直しの検出に，Batliner ら [Batliner 00] の "Prosody Module" を用い 49%の再現率と 70%の適合率を得ている[*1]．Batliner らの Prosody Module は，ポーズ，エネルギー，F0，発話速度，および品詞情報をニューラルネットワークと n-gram を使ってモデル化し，発話の区切りや承認や確認といった発話行為を予測する．

しかしながら，単語認識を基本とする現在の自動音声認識では，言い直しの強い手がかりとなる単語断片の検出は難しい．また，言い直しが発生すると，フィラーや単語断片のために言い直し近辺の単語認識に失敗する可能性が高い．従って現状で

連絡先: 船越 孝太郎, 〒 152-8552 東京都目黒区大岡山 2-12-1 東京工業大学 計算工学専攻 西 8 号館 E607, 03-5734-3086, koh@cl.cs.titech.ac.jp

*1 単語断片が完全に検出できると仮定した場合は，71%の再現率と 85%の適合率．

は，実際の音声対話システムでの利用に際して，品詞情報などに依存した言い直しの検出がどれだけ有効に働くかは未知である．

もし単語や品詞の情報を利用せずにある程度の言い直しの検出が可能ならば，良好な認識結果を得られない場合でも言い直しの発生を認識し，誤った理解による対話状態の致命的な遷移[*2]を抑制することが期待できる．あるいは，言い直し近辺の情報を用いずに聞き返しを行なえば，言い直しそのものを正しく解釈はできなくても，わずかでも対話の進行状態を先に進めることができる可能性がある．

以上のことを踏まえ，本稿では，音響情報のみを用いた言い直しの検出を試みる．

3. 検出に用いる音響情報

本稿では，音声データから自動的に抽出可能な音響情報のみを利用して，言い直しの検出，すなわち IS の検出を行なう．

まず，全ての言い直しは，IS の位置にポーズを伴うと仮定する．今回使用したデータでは約 99%の言い直しがポーズを伴っていた．すなわち，本稿での言い直しの検出とは，音声データ中のポーズを検出し，そのポーズの始端が IS であるかどうかを判別することである．判別には，音声対話コーパスから自動抽出したポーズの周辺の音響情報を素性とする分類器を用いる．以下で，分類器が用いる素性について説明する．

パワー

ある時点の音声のパワーを，振幅値に Hamming 窓関数を適用した値を 2 乗して前後 15msec の区間内で平均した値の常用対数を 10 倍したものと定義する．パワーに関する素性として，ポーズ前後の平均，最大値，パワー曲線の直線近似後の傾きの変化を抽出する．

ポーズ前後の平均，最大値に関しては，実測値と音声ファイル内で正規化した値の両方を素性として用いる．また，ポーズ前後での大小関係も素性に含める．

パワーのポーズ前後での変化の様子を素性として取り込むために，佐々木ら [佐々木 96] が発話末の状態 (継続か終了か) の推定のために用いた，抑揚情報パターンの直線近似とコード化の手法を利用する．この手法では，ポーズ区間の前後を一定間隔毎に最小 2 乗法によって直線近似し，求めた傾きに対して予め定めた範囲毎にラベルを付与してコード化する (表 1 参照)．表 1 は予め調査した傾きの分布から定めた．上記のコード以外に，傾きの値そのもの，上下の変動を表す符合も別個の素性として使用する．

ピッチ (F0)

音声データから F0 を抽出し，無声音領域など F0 が抽出できない区間に関してはその区間の両端を直線で結び，ピッチ曲線とする (図 2 参照)．F0 の抽出には，井本和範氏の作成したプログラムを借用した[*3]．パワーと同様に，ポーズ区間の前後の平均，最大値，ピッチ曲線の直線近似の傾きの変化などを素性として用いる．

ポーズ

各ポーズの長さを素性として用いる．ポーズの検出には上で定義したパワーを用いた．録音室のような，定常的で小さな雑音しかない環境で録音された音声データの場合，パワーの分布をグラフ化すると，音声の平均パワーと背景雑音の平均パワーとを極大とする二つの山が現れる (図 3 参照)．

*2 話者が対話の終了を宣言したと誤認識してしまう場合など

*3 http://vision.kuee.kyoto-u.ac.jp/lecture/dsp/pitch/pitchprog.html

表 1: パワーとピッチの傾きに対応するコード

パワー		ピッチ	
傾き	コード	傾き	コード
-0.91 未満	1	-42 未満	1
-0.91〜-0.65	2	-42〜-30	2
-0.65〜-0.39	3	-30〜-18	3
-0.39〜-0.13	4	18〜-6	4
-0.13〜0.13	5	-6〜6	5
0.13〜0.39	6	6〜18	6
0.39〜0.65	7	18〜30	7
0.65〜0.91	8	30〜42	8
0.91 以上	9	42 以上	9

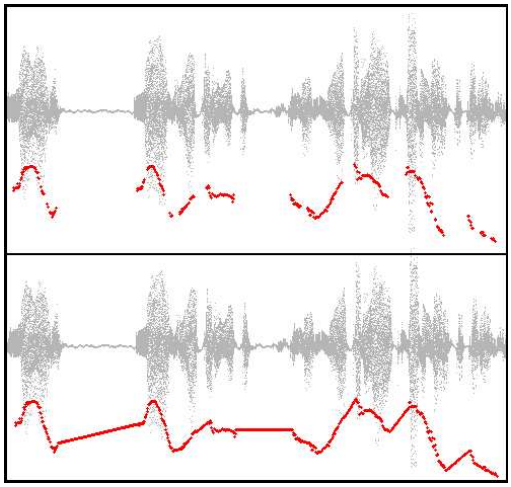

図 2: ピッチの補完

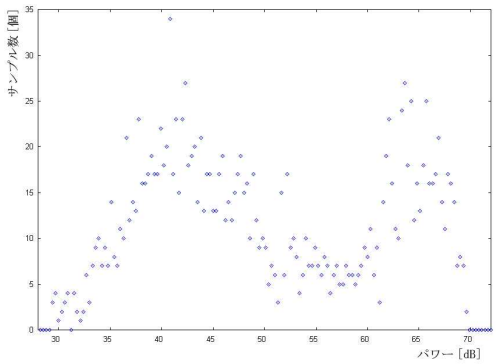

図 3: パワーの分布 (横軸がパワー [dB]，縦軸がサンプル数)

ポーズを検出する為に，音声データ毎に，二つの山の間に挟まれた谷間の領域の中で最小のサンプル数を持つ値を計算し，その値を閾値をとする．そして，その閾値よりもパワーが小さい区間をポーズとする．短い発話や，ポーズがほとんど含まれていない場合には山が二つできないことがあるが，この場合には予め設定したデフォルト値を用いた．いくつかの音声データで試した結果，ポーズ，発声区間ともに，最短長を 25msec として，それより短い区間はノイズとみなすと良い結果が得られた．

4. 検出実験

4.1 学習・テストデータの作成

コーパスとして ATR 研究用自然発話音声データベース (SDB) [塚田 97] の模擬対話データを用い，この中から，言い直しを含む 872 発話と言い直しを含まない 242 発話，全 1114 発話を取り出した．この 1114 発話の音声データから，自動的にポーズ，パワー，ピッチを計算し，これらに関連する素性を抽出した．素性の数は，全部で 144 個である．そして，検出されたポーズ毎にそこが IS であるかどうかを人手でチェックし，949 個の正例 (言い直し) と 17164 個の負例 (非言い直し) を得た．

ピッチやパワーの最大値や平均値を求め素性として使用するには，ポーズからどこまで離れた情報をどのような間隔で計算するかを予め定めておく必要がある．本稿では，音響情報のみの使用を考え単語の境界情報などの音韻的な情報も手に入らないという前提なので，単語内で平均値・最大値を計算するといったことはできない．今回は，平均値・最大値を求めるための区間 (フレーム) の幅，素性として扱う区間の個数などは，事前の調査によって求めた値を用いた．

4.2 実験

分類器として，CART 法によって生成される分類木と，Support Vector Machine(SVM)[Vapnik 98] を用いた．CART の実装として，Edinbugh Speech Tools [*4]に含まれる wagon を，SVM の実装として，工藤拓氏の TinySVM [*5]を用いた．

正例の数は負例の 5%強と極端に少ないので，負例の数を正例の数に合わせて学習を行なった．すなわち，正例 949 個の内，800 個を学習用，149 個をテスト用とするのに合わせ，17164 個の負例から 800 個を学習用に抜き出した．CART の学習には 144 個の素性全てを用いた．SVM の学習には，パラメータを変えて 35 回 CART に学習させた時に，出力された分類木の中で一度以上利用されていた 64 個の素性のみを用いた．カーネル関数には線形関数を用いた．

表 2 に 700 個の正例と負例併せて 1600 個のサンプルデータから学習したモデルを，149 個の正例と負例併せて 298 個のテストデータに適用した結果を示す．再現率は，テストデータ内の言い直しの中で，言い直しとして判定されたものの割合を示す．適合率は，分類器が言い直しとして判定したポーズの中で正しかったものの割合を示す．正解率は，テストデータのなかで言い直しあるいは非言い直しとして正しく判定されたポーズの割合を示す．表 3 には同じモデルを，149 個の正例と残り全ての負例 16364 個，併せて 16513 個のテストデータに適用した結果を示す．

表 3 の正解率のベースラインは，全てのポーズを言い直しでないとした場合の 99.1%となり，今回の結果ははるかにそれを下回ってしまった．表 3 で適合率が著しく低いのは，例えば CART の場合，言い直しのポーズが 149 個中 116 個正しく認識されているものの，言い直しでないポーズが 163364 個中 4373 個言い直しとして誤認識されてしまったためである．

*4 http://festvox.org/packed/festival/1.4.3/

*5 http://cl.aist-nara.ac.jp/ taku-ku/software/TinySVM/

表 2: 結果 1

分類器	データセット	再現率	適合率	正解率
CART	学習用	80.1	76.1	77.5
	テスト用	77.9	74.8	75.8
SVM	学習用	80.0	75.1	76.7
	テスト用	77.2	70.6	72.5

表 3: 結果 2

分類器	再現率	適合率	正解率
CART	77.9	2.58	73.3
SVM	77.2	2.35	70.8

Heeman ら [Heeman 99] は，同じ ATR で作成された別の日本語音声対話コーパスでの実験で，再現率 78.6%，適合率 74.9%の結果を得ている (正解率は 99.8%)．彼らは単語断片や編集表現を含めたほぼ完全な音声認識を想定している上，使用したコーパスも方法も異なるので一概には比較できないが，彼らが 304,000 語の中の 659 個の言い直しを上記のような再現率と適合率で検出していることを考えると，音響情報だけを使用する方法に対する望みは小さい．Spilker ら [Spilker 00] の用いた手法は，ポーズやピッチなど本稿と同じような情報によって言い直しを検出しているものの，同時に品詞レベルの言語モデルも利用しており，この言語モデルが単語列上のパターンマッチングのように作用し，検出に強く影響を与えている可能性がある．

また，CART が生成した分類木をみると，ピッチ・パワーの傾きコードに関する素性はほとんど使われておらず，判別に利用されている素性はポーズ長とピッチ・パワーの最大値，平均値に関するものがほとんどであった．ピッチやパワーの傾きが適切にコード化されているのか，そもそも傾きをコード化しても言い直しの検出には有効ではないのかを検討する必要がある．

5. まとめ

本稿では，ポーズに注目して，言い直しの音響情報のみによる全自動検出を試みた．音響情報として，ポーズ，ピッチ，パワーを利用し，ピッチとパワーに関しては，最大値，平均，傾きの変化，ポーズの前後での落差などを特徴として抽出し，それらの特徴量を素性として，CART 法と SVM によって分類器を学習させた．

しかしながら，再現率に関してはかなり良い値を得られたものの，適合率が非常に悪く，全体としては良好な結果は得られなかった．今回利用した音響情報のみから言い直しを正確に検出することは難しいと結論する．適合率を改善する為には，言い淀みの音響的な検出や言語的な情報が必要である．

今回は，単語境界なども分からないものとして，予め定めた画一的な間隔や幅で平均や最大値などの計算を行なったが，より音韻的な情報を利用して話者速度を抽出したり，最大値などを単語の長さに合わせて計算することで一定の向上は望める

と考える．また，本稿では試さなかったが，[Bear 92] のようにパターンマッチングや構文的な情報を用いた言い直しの検出と組み合わせることで，ある程度の有効性を見込める可能性もある．

参考文献

[Batliner 00] Batliner, A., Buckow, J., Niemann, H., Noth, E., and Warnke, V.: The Prosody Module, in Wahlster, W. ed., *Verbmobil: Fundations of Speech-to-Speech Translation*, pp. 107–121, Springer (2000)

[Bear 92] Bear, J., Downing, J., and Shriberg, E.: Integrating multiple knowledge sources for detection and correction of repairs in human-computer dialog, in *Proceedings of 30th Annual Meeting of ACL*, pp. 56–63 (1992)

[Breiman 84] Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J.: *Classification and Regression Trees*, Wadsworth & Brooks (1984)

[Core 99] Core, M. G. and Schubert, L. K.: A syntactic framework for speech repairs and other disruptions, in *Proceedings of 37th Annual Meeting of ACL*, pp. 413–420 (1999)

[伝 97] 伝 康治：統一モデルに基づく話し言葉の解析, 自然言語処理, Vol. 4, No. 1, pp. 23–40 (1997)

[Funakoshi 02] Funakoshi, K., Tokunaga, T., and Tanaka, H.: Processing Japanese Self-correction in Speech Dialog Systems, in *Proceedings of COLING2002* (2002)

[Heeman 97] Heeman, P. A. and Allen, J. F.: Intonational Boundaries, Speech Repairs and Discourse Markers: Modeling Spoken Dialog, in *Proceedings of 35th Annual Meeting of ACL*, pp. 254–261 (1997)

[Heeman 99] Heeman, P. A. and Loken-Kim, K. H.: Detecting and Correcting Speech Repairs in Japanese, in *ICPhS Satellite Meeting on Disfluency in Spontaneous Speech, Berkeley* (1999)

[Hindle 83] Hindle, D.: Deterministic parsing of syntactic non-fluencies, in *Proceedigns of 21st Annual Meeting of ACL*, pp. 123–128 (1983)

[神田 96] 神田 祐和, 堀内 靖雄, 小磯 花絵, 市川 薫：対話における言い直しの分析, 人工知能学会研究会資料 SIG-SLUD-9601, pp. 55–62 (1996)

[Kikui 94] Kikui, G. and Morimoto, T.: Similarity-based Identification of Repairs in Japanese Spoken Language, in *Proceedings of ICSLP-96*, pp. 915–918 (1994)

[中野 98] 中野 幹生, 島津 明：言い直しを含む発話の解析, 情報処理学会論文誌, Vol. 39, No. 6, pp. 1935–1943 (1998)

[Nakatani 93] Nakatani, C. and Hirschberg, J.: A speech-first model for repair identification and correction, in *Proceedings of 31th Annual Meeting of ACL*, pp. 200–207 (1993)

[O'Shaughnessy 92] O'Shaughnessy, D.: Analysis of False Starts in Spontaneous Speech, in *Proceedings of ICSLP-92*, pp. 931–934 (1992)

[佐川 94] 佐川 雄二, 大西 昇, 杉江 昇：自己修復を含む日本語不適格文の分析とその計算機による理解手法に関する考察, 情報処理学会論文誌, Vol. 35, No. 1, pp. 46–52 (1994)

[佐々木 96] 佐々木 聡, 神田 祐和, 堀内 靖雄, 市川 薫：発話末の基本周波数とパワーのパターン分類とその分析, 人工知能学会研究会資料 SIG-SLUD-9602, pp. 1–6 (1996)

[Spilker 00] Spilker, J., Klarner, M., and Gorz, G.: Processing Self-Corrections in a Speech-to-Speech System, in Wahlster, W. ed., *Verbmobil: Fundations of Speech-to-Speech Translation*, pp. 131–140, Springer (2000)

[塚田 97] 塚田 元, 中村 篤, 竹澤 寿幸, 匂坂 芳典：研究用自然発話音声データベース解説書, Technical Report TR-IT-0222, ATR 音声翻訳通信研究所 (1997)

[Vapnik 98] Vapnik, V.: *Statistical Learning Theory*, Wiley, Chichester (1998)