情報処理学会論文誌 Vol.56 No.3 993–1002 (Mar. 2015)

合成音声への自然なポーズ挿入のための音声の自然性に影響を与えるポーズ位置に関する要因の分析と評価

武藤 博子[1] 井島 勇祐[1,a)] 宮崎 昇[1] 水野 秀之[1] 阪内 澄宇[1]

受付日 2014年6月30日, 採録日 2014年11月10日

概要：本研究では，テキスト音声合成における自然なポーズ挿入の実現のために，テキストから抽出可能なポーズ位置に関する素性と音声の自然性との関係の分析と評価を行った．まず，ポーズ位置およびテキストが異なる合成音声を用いた主観評価実験を実施し，ポーズ位置の自然性の観点で主観評価値を収集した．次に，ポーズ位置に関する素性をポーズ句の長さと係り受け構造の観点から複数設計し，重回帰分析および判別分析により主観評価値との関係について分析を行った．分析の結果，長さが極端に短いポーズ句の存在に関する素性等 13 種類の素性が，自然性に対する影響が大きいことが確認できた．最後に，提案する素性のポーズ位置決定における有効性を評価するため，従来の素性に基づくポーズ位置決定手法と，主観評価値と素性との関係に基づくポーズ位置の自然性の評価器とを組み合わせた枠組みでポーズ位置の主観評価実験を行い，主観評価値の向上が確認できた．

キーワード：テキスト音声合成，ポーズ位置，素性，自然性，韻律

Analysis and Evaluation of Factors Relating Pause Location for Natural Text-to-speech Synthesis

Hiroko Muto[1] Yusuke Ijima[1,a)] Noboru Miyazaki[1] Hideyuki Mizuno[1] Sumitaka Sakauchi[1]

Received: June 30, 2014, Accepted: November 10, 2014

Abstract: This paper reports a study in which the relationship between various pause location-related features and speech naturalness was analyzed to achieve natural pause insertion for text-to-speech synthesis. First, a subjective experiment was conducted using speech samples with different pause locations and text contents to collect naturalness scores regarding pause location. Next, multiple regression and discriminant analysis were carried out. The analysis results confirmed that 13 features have a significant impact on speech naturalness. To confirm the features' effectiveness for pause location prediction, a speech naturalness evaluator was constructed that uses the relationship between the obtained naturalness scores and features. The results of a subjective evaluation experiment performed with the evaluator confirmed that its use resulted in improved evaluation scores.

Keywords: text-to-speech synthesis, pause location, feature, naturalness, prosody

1. はじめに

近年，自然な合成音声の実現に向けて，様々なテキスト音声合成システムの開発が行われている．テキスト音声合成システムは，入力されたテキストを解析し言語的に推定が必要な読み，アクセント，ポーズ位置を付与するテキスト解析部と，テキスト解析部から得られた情報に基づき音声生成を行う音声生成部から構成される．このうち，音声生成部は，音声波形素片を接続して音声を生成する波形接続型音声合成 [1] や，音響特徴量を統計的にモデル化し音

[1] 日本電信電話株式会社 NTT メディアインテリジェンス研究所
NTT Media Intelligence Laboratories, NTT Corporation, Yokosuka, Kanagawa 239–0847, Japan
[a)] ijima.yusuke@lab.ntt.co.jp

声パラメータの生成を行う HMM 音声合成 [2] 等が提案されており，自然な合成音声が実現されつつある．しかし，音声生成部は，テキスト解析部から正しい情報が与えられることを前提とした処理を行うため，テキスト解析部で付与される情報の精度も合成音声の自然性に大きな影響を与える．そのため，音声合成に必要となる読み，アクセント，ポーズ位置の付与についても，従来から多くの手法が提案されており，新聞やニュース等のテキストにおいて高い精度が得られている．本研究では，これらの情報の中でも，音声の自然性や理解しやすさに大きな影響を与えるポーズ位置に焦点をあてる．前述のとおり，従来研究によりポーズ位置の付与精度は向上しつつあるが，ポーズ位置が不自然であることが原因で合成音声の自然性が低下することも多く，適切な位置へのポーズ挿入は自然で理解しやすい合成音声を実現するうえで重要な課題となっている．

合成音声のポーズ位置を決定するうえで，人間の発話におけるポーズ挿入は参考とするべき重要な指標であり，従来より様々な観点で調査が行われている．比企 [3] は，呼気や呼気段落長がポーズ挿入における重要な要因であることを示し，Fujisaki ら [4] は，日本語の読み上げ音声を対象として，文中の各位置におけるポーズの発生確率および継続時間長を調査し，句の長さに応じてポーズ挿入確率が線形に増加する傾向があることを確認している．一方で，杉藤ら [5] は，意味的な区切りとなる句境界には呼気にかかわらずポーズが挿入されることを確認し，文の統語構造もポーズ挿入に与える影響が大きいことを示している．また，自然な合成音声を実現するためのポーズ位置の決定に関する研究も報告されている．Mandal ら [6] は Bangla 語の読み上げ音声を対象として Fujisaki らと同様の分析を実施し，句の長さがポーズ位置の決定に有効であることを示している．その他，係り受け構造の重要性を示す研究結果も報告されている．海木ら [7] は，ナレータの発話の分析により，文の係り受け構造によってポーズの入りやすさが大きく異なることを示している．藤尾ら [8] は，海木らの研究により得られた素性に基づくポーズ位置決定手法を提案し，係り受け構造を考慮した素性がポーズ位置の決定に有用であることを示唆している．また，鈴木ら [9] は，隣り合った N 文節の品詞情報を用いて推定した係り受け関係に基づくポーズ位置決定手法を提案し，その有効性を確認している．これらの研究により，句の長さおよび係り受け構造がポーズ位置を決定するうえで重要であることが確認されているが，これら 2 つの要因がポーズ位置の自然性に与える影響の差について，十分な分析や検討はされていない．

このように，自然発話におけるポーズ位置と様々な素性との関係について検討が行われている一方で，ポーズが音声の理解しやすさに与える影響についても検討が行われている．杉藤 [10], [11]，河野 [12] らは，ポーズ句の長さや回数，意味的な境界におけるポーズの有無が聞き手の理解度に影響を与えることを示しており，ポーズの位置に関わる要因が，音声の評価に影響を与えることが定性的に示唆されている．合成音声の主観評価では，ポーズ位置を含め人間が不自然に感じる箇所が音声中に 1 カ所でも存在すると，評価が大きく低下することがよく知られており，合成音声の品質向上の観点からもポーズ位置の自然性に関わる要因を定量的に明らかにすることは有用と考えられる．

そこで本研究では，句の長さおよび係り受け構造の観点でポーズ位置に関する素性を検討し，主観評価値との関係を分析することで，ポーズ位置の自然性の評価に影響を与える素性を明らかにする．まず，ポーズ位置およびテキストの内容が異なる 500 種類の合成音声を用いた主観評価実験を実施し，ポーズ位置の自然性に関する主観評価値を収集する．次に，収集した主観評価値とポーズ位置に関する素性との関係を重回帰分析および判別分析により分析し，主観評価値と関連が高い素性を同定する．最後に，同定した素性のポーズ位置決定における有効性を評価するため，機械学習の枠組みに基づいて，従来研究で用いられている素性および本研究で同定した素性を用いたポーズ位置決定実験を行い，それぞれのポーズ位置の自然性を主観評価実験により比較評価する．

2. 自然な発話におけるポーズ位置に関する素性

2.1 素性の設計方針

1 章で述べたとおり，従来研究によりポーズ句の長さと統語構造がポーズ位置の決定の重要な要因であることが確認されている．杉藤 [11] は，複数人のナレータの音声を用いた聴取実験において，理解しやすいと聴取者に判断されたナレータの発話は，ポーズ句の長さが平均 1.6 文節程度と，他のナレータの半分程度の長さであることを確認している．また，別宮 [13] は，日本人は，日本語の発話において，およそ 4 拍子のリズムで等時的に発音する傾向があると述べている．これらの報告から，ポーズ句の平均的な長さやばらつきが音声の自然性に影響を与えていると考えられる．一方，統語構造，特に日本語においては係り受け構造が，音声の理解しやすさに影響を与えることが知られている．海木ら [7] は，規則ベースのポーズ位置決定手法を提案するにあたり，複数人の発話の分析を行い，ポーズ位置の決定に関わる主要因を特定している．特に，当該句境界（アクセント句）の係り受けの深さ，（直前の句からの係り受け関係が存在するかどうかを示す）先行句境界の右・左枝分かれの相違，当該句境界の読点の有無等，数文節程度の局所的な句構造に着目し，これらの構造を考慮してポーズ位置を決定することが自然な合成音声を実現するうえで重要であると示唆している．これにより，聞き手にとって理解しやすく話すためには，係り受け構造で定義できる意

味上の区切りにおいて，適切にポーズを挿入することが重要と考えられる．したがって，本研究ではポーズ位置に関する素性として，句の長さに関する素性と係り受け構造に関する素性の 2 種類を検討した．

2.2 分析に用いる素性

本研究では，上述の分類に基づき，句の長さに関する素性を 12 種類，係り受け構造に関する素性を 13 種類，計 25 種類のポーズ位置に関する素性を設計した．**表 1** に素性の一覧を示す．まず，句の長さを表す素性として，ポーズ句の長さの平均値および分散値（a–d）を用いた．また，自然性に与える影響が大きいと考えられる素性として，ポーズ句の長さの外れ値の有無（e–l）を用いた．これは，平均値から逸脱した長さのポーズ句の存在が，音声のリズムを損ない，自然性を低下させる要因となると考えたためである．本素性は，基準値より長いもしくは短いポーズ句の存在有無を 2 値で示しており，該当する長さのポーズ句が 1 つでも存在した場合は 1 が，それ以外の場合は 0 が設定される．基準値は，3 章で述べる主観評価実験で用いる音声試料のポーズ句の長さの平均値と標準偏差に基づき算出した．ポーズ句の長さの単位には，日本語の発音のリズムを構成する単位であるモーラと，文法的なまとまりを構成する単位である文節の 2 種類を用いた．

次に，係り受け構造に関する素性について述べる．本研究では，基本的な係り受け構造として，従来より自然性に対する影響が大きいことが確認されている左枝分かれ境界（境界直前の句が直後の句を直接修飾する境界）および右枝分かれ境界（境界直前の句が直後の句を直接修飾しない境界），そして，比較的出現数が多くポーズをともなって出現することが多い並列関係の 3 種類について，文節境界におけるポーズ有無の観点で素性を設計した（m–p，u–v）．これは，海木ら [7] が，複数人の発話の分析において，左枝分かれ境界はポーズが挿入されにくく，右枝分かれ境界はポーズが挿入されやすい傾向にあることを定量的に示していることから，左枝分かれ境界にポーズが挿入される（m），または右枝分かれ境界にポーズが挿入されない（p）は，一般的に起こりにくいポーズ挿入および脱落であり，自然性に影響を与えると考えたためである．また，（m，p）と対になる素性として，一般的に起こりやすいポーズ挿入および脱落を表す同様の素性（n：左枝分かれ境界におけるポーズなし，o：右枝分かれ境界におけるポーズあり）を用いた．さらに，個数以外の観点の素性として，上記の素性に対応する境界数またはポーズ有無の数に占める割合（q–t）を導入した．これは，係り受け構造におけるポーズの有無と自然性との関係が，音声全体の長さや係り受け構造，ポーズの総数との兼ね合いによって変わる可能性があるためである．**図 1** に，上述した 3 種類の係り受け構造（左枝分かれ，右枝分かれ，並列関係）における素性の具体例を示す．各構造の素性を代表して，各構造間の文節境界にポーズが存在する場合の例（それぞれ，素性 m，o，u に対応）を示す．各構造間にポーズが存在しない場合は，記載したそれぞれの例において，ポーズを示す記号が存在しない場合に対応する．これらの文節境界におけるポーズ有無を示す素性に加えて，文節境界以外の箇所にポーズが

表 1 分析に用いるポーズ位置に関する素性
Table 1 Features related to pause location.

句の長さに関する素性	
a	ポーズ句のモーラ数の平均値
b	ポーズ句のモーラ数の分散値
c	ポーズ句の文節数の平均値
d	ポーズ句の文節数の分散値
e	ポーズ句のモーラ数の外れ値の有無（$\mu_m + \sigma_m$）
f	ポーズ句のモーラ数の外れ値の有無（$\mu_m + 2\sigma_m$）
g	ポーズ句のモーラ数の外れ値の有無（$\mu_m - \sigma_m$）
h	ポーズ句のモーラ数の外れ値の有無（$\mu_m - 2\sigma_m$）
i	ポーズ句の文節数の外れ値の有無（$\mu_p + \sigma_p$）
j	ポーズ句の文節数の外れ値の有無（$\mu_p + 2\sigma_p$）
k	ポーズ句の文節数の外れ値の有無（$\mu_p - \sigma_p$）
l	ポーズ句の文節数の外れ値の有無（$\mu_p - 2\sigma_p$）
係り受け構造に関する素性	
m	左枝分かれ+ポーズありの文節境界数
n	左枝分かれ+ポーズなしの文節境界数
o	右枝分かれ+ポーズありの文節境界数
p	右枝分かれ+ポーズなしの文節境界数
q	左枝分かれ+ポーズありの文節境界数/左枝分かれの文節境界数
r	左枝分かれ+ポーズありの文節境界数/ポーズありの文節境界数
s	右枝分かれ+ポーズなしの文節境界数/右枝分かれの文節境界数
t	右枝分かれ+ポーズなしの文節境界数/ポーズなしの文節境界数
u	並列関係+ポーズありの文節境界数
v	並列関係+ポーズなしの文節境界数
w	並列関係+ポーズなし/並列関係の文節境界数
x	並列関係+ポーズなし/ポーズなしの文節境界数
y	文節境界以外にポーズが挿入された数

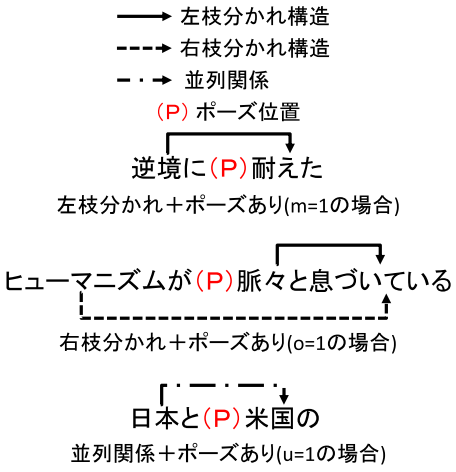

図 1 左枝分かれ，右枝分かれ，並列関係におけるポーズの例
Fig. 1 Examples of dependency structure features.

挿入された数（y）も素性として導入した．文節境界以外の箇所へのポーズ挿入は読み上げ音声における発生数がきわめて少なく，自然性に対する影響が大きいと考えたためである．また，右枝分かれ境界をさらに細分化した素性である係り受けの深さ（境界直前の句が直接修飾する句までの句数）もポーズ決定の素性として用いられることがあるが，本研究では導入しないこととした．これは，右枝分かれ境界を細分化することで，個々の事象数が少なくなり，分析に十分なサンプル数が得られないと考えたことと，まずは基本的な 3 種類の係り受け構造について，句の長さに関する素性と組み合わせたときの自然性との関係を明らかにすることが重要であると考えたためである．従来より，係り受け構造に関する素性の句構造の単位として，アクセント句，文節等，複数の基準が用いられているが，本研究では日本語の係り受け構造の基本的な単位である文節に統一し，素性を設計した．

3. 音声中のポーズ位置の自然性に関する主観評価値の収集

2 章で設計した素性と自然性との関係について分析するため，ポーズ位置のみが異なる合成音声を用いた主観評価実験を実施し，自然性の主観評価値の収集を行った．

3.1 音声試料の作成

音声試料として，テキストの内容およびポーズ位置が異なる 500 種類（テキスト 50 種類 × ポーズ位置の組合せ 10 種類）の合成音声を作成した．テキストには，ATR 音声データベースセット B に含まれる読点のない音素バランス文 50 文を用いた．ポーズ位置の組合せは，あらゆる文節境界におけるポーズ有無の組合せの網羅が主観評価実験のコストの観点から困難であるため，以下の観点に基づき選定した．

- 従来研究ですでに明らかにされているようなポーズの決定要因が含まれている．
- 自然性に関する評価が大きく異なる．

本研究では，人間の発話におけるポーズ位置は，従来研究ですでに明らかにされているようなポーズの決定に関わる要因に影響されていると考え，人間の実際の発話から抽出した多数のポーズ位置の組合せから，ポーズ位置の偏りが少なくなる組合せを選択した．発話の収録は，テキストをプロナレータ 1 名およびナレーションを生業としない一般話者 100 名の計 101 名に発声させることで行った．発声にあたり，事前トレーニングは実施しなかった．また，ポーズ位置に関する教示は与えず，各文を感情をこめず淡々と発話し，文ごとに口調を変えないように指示した．発話に含まれる多数のポーズ位置の組合せから，以下の手順でポーズ位置の偏りが少ない組合せを選択した．以下，ポーズ位置候補集合を一般話者 100 名の発話から抽出したポーズ位置の組合せの集合とする．

(1) 初期値としてプロナレータのポーズ位置の組合せを評価用ポーズ位置として選択する．
(2) ポーズ位置候補集合に含まれるすべての組合せについて，全評価用ポーズ位置との編集距離を算出する．評価用ポーズ位置が複数存在する場合は，各評価用ポーズ位置に対する編集距離の平均値を算出する．
(3) ポーズ位置候補集合から，(2) で得られた編集距離が最も大きい組合せを選択し，評価用ポーズ位置に追加する．追加された組合せはポーズ位置候補集合から除く．
(4) 評価用ポーズ位置の総数が 10 に達した場合は処理を終了する．そうでなければ，(2) に戻る．

ポーズ位置の組合せは文ごとに選択した．また，各話者の音声に対するポーズ位置の決定は，ポーズ位置の付与に関する作業に十分な経験を有する作業者が音声を聴取し，聴感上ポーズが認められる箇所にポーズを付与することで行った．判断に迷った場合は，音声波形を視察し，明確な無音区間（60 ms 以上）が存在した場合に，ポーズと決定するよう作業者に指示した．プロナレータには，ATR 音声データベースセット B に含まれる女性 FKN を用いた．声質，韻律の違いによる自然性の評価への影響を避けるため，合成音声には，波形接続型音声合成システム Cralinet [14] で作成した 1 種類の女性合成音声を用いた．ポーズの長さは，評価に影響を与えることがないように 0.5 秒に統一した．韻律の生成には ATR 音声データベースセット B に含まれる男性プロナレータ MHT の音声に付与されたアクセント句境界，アクセント型を用いた．音声信号のサンプリング周波数は 22.05 kHz とした．

3.2 主観評価

作成した全音声試料に対して，MOS（Mean Opinion Score）による主観評価を行った．評価尺度は，「5：非常に自然」～「1：非常に不自然」とした．評価は正常聴力を有する成人 24 名が行い，全評価者の評点の平均値を各音声の主観評価値とした．

図 2 に全音声試料の主観評価値のヒストグラムを示す．主観評価値は 1.5 から 4.5 まで幅広く分布しており，ポーズ位置の違いが自然性に非常に大きな影響を与えていることが分かる．**図 3** に，自然性が平均的に高いまたは低いと評価されたポーズ位置の例と，評価者によって評点がばらついたポーズ位置の例を示す．主観評価値が高い例（c）は，意味の区切りが良い箇所に，ほぼ一定のポーズ句の長さでポーズが挿入されている．一方，主観評価値が低い例は，極端に長いポーズ句（a），短いポーズ句（b）が存在する，または意味の区切りが悪い箇所にポーズが挿入される等の事象が確認できる．また，評価者間の評点にばらつきがある例（d）は，「見上げる」と「フジもいいが」の間

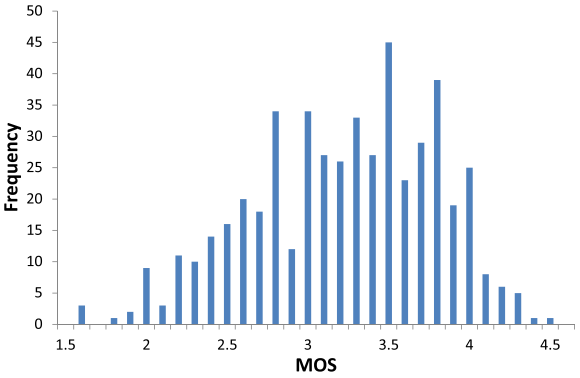

図 2 全音声試料の主観評価値のヒストグラム
Fig. 2 The MOS histogram for speech stimuli.

このプロジェクトの入札者の中では弊社がもっとも
ふさわしい資格を有していると(P)自負いたしております
(a) 主観評価値が低いポーズ位置の例(主観評価値:2.1)
(極端に長いポーズ句が存在)

自動車や精密機械などで(P)
技術系の採用を抑えるところが(P)目立ち(P)
売り手市場の技術系にもかげりが見え始めた
(b) 主観評価値が低いポーズ位置の例(主観評価値:2.7)
(極端に短いポーズ句が存在)

インタビューは午後十時から始まり(P)
途中で夕食をはさみ(P)延々四時間に及んだ(P)
(c) 主観評価値が高いポーズ位置の例(主観評価値:4.3)

見上げる(P)フジもいいが(P)路地植え(P)
また鉢植えの花もきれいです
(d) 評点にばらつきのあるポーズ位置の例
(主観評価値:3.1 [内訳, 5:4名, 4:5名, 3:7名, 2:5名, 1:3名])

図 3 音声試料におけるポーズ位置の例
Fig. 3 Examples of pause location in speech stimuli.

や,「また」と「鉢植えの」の間等,ポーズが入っていることに不自然さを感じるか否かが人によって異なることが評価者間の評点のばらつきの原因になったと考えられる.

4. 主観評価値とポーズ位置に関する素性との関係の分析

4.1 音声試料における素性の分布

3 章で作成した音声試料について,素性の分布を調査した.まず,連続値で表される素性(a–d, m–y)について,各素性の値を試料ごとに算出し,平均値,分散値を求めた.結果を**表 2** に示す.次に,外れ値に関する素性(e–k)について,素性値が 1 となる数を調査した.**表 3** に素性値が 1 をとる数および素性値の決定の基準となる基準値を示す.素性(e, f, i, j)においては,基準値以上の場合素性値は 1 となり,素性(g, k)においては基準値以下の場合 1 となる.基準値の算出に必要となるポーズ句のモーラ数,文節数の平均値および標準偏差は全データから算出し,それぞれ $\mu_m = 9.6$, $\sigma_m = 6.5$, $\mu_p = 2.3$, $\sigma_p = 1.2$ となった.なお,表 1 の素性(h)および(l)は,基準値が負の値となったため,以降の分析から除いた.表 2 および表 3 に示すとおり,連続値で表される素性は,(q–t, w, x)の比率に関する素性以外は分散が大きいことから,音声試料中に様々なポーズ位置の組合せが含まれていることが分かる.また,素性値が 2 値で表される素性のうち,素性値が 1 となる素性は,該当数がすべて 100 を超えており,分析に用いるうえで十分な数が得られていることが分かる.

表 2 連続値で表される素性の分布
Table 2 Distribution of continuous value features.

	素性	平均	分散
a	モーラ数の平均値	11.5	35.9
b	モーラ数の分散値	28.6	1,781
c	文節数の平均値	2.27	1.48
d	文節数の分散値	1.17	3.61
m	左枝分かれ+ポーズあり	1.31	1.55
n	左枝分かれ+ポーズなし	3.11	3.04
o	右枝分かれ+ポーズあり	1.91	1.50
p	右枝分かれ+ポーズなし	0.83	1.06
q	左枝分かれ+ポーズあり/左枝分かれ	0.30	0.08
r	左枝分かれ+ポーズあり/ポーズあり	0.37	0.09
s	右枝分かれ+ポーズなし/右枝分かれ	0.28	0.10
t	右枝分かれ+ポーズなし/ポーズなし	0.17	0.04
u	並列関係+ポーズあり	0.34	0.37
v	並列関係+ポーズなし	0.14	0.16
w	並列関係+ポーズなし/並列関係	0.10	0.07
x	並列関係+ポーズなし/ポーズなし	0.04	0.01
y	文節境界以外にポーズが挿入された数	0.06	0.06

表 3 外れ値に関する素性の該当数および基準値
Table 3 Number of features regarding outliers.

	素性	該当数	基準値
e	モーラ数の外れ値の有無($\mu_m + \sigma_m$)	219	16.1
f	モーラ数の外れ値の有無($\mu_m + 2\sigma_m$)	106	22.6
g	モーラ数の外れ値の有無($\mu_m - \sigma_m$)	146	3.0
i	文節数の外れ値の有無($\mu_p + \sigma_p$)	386	3.5
j	文節数の外れ値の有無($\mu_p + 2\sigma_p$)	183	4.7
k	文節数の外れ値の有無($\mu_p - \sigma_p$)	105	1.1

4.2 重回帰分析

各素性と自然性との関係を調査するため,3 章で収集した主観評価値と各素性との関係を重回帰分析により分析した.重回帰分析の目的変数には主観評価値を,説明変数には表 2 および表 3 に示す素性の値を用いた.まず多重共線性を回避するため,素性の分散拡大要因(VIF)が一定値以上の場合,または素性を用いることで他の素性の VIF が一定値以上となる場合は,該当する素性を説明変数から除いた.本分析では,VIF を 10 に設定した.この処理により,文節数の平均値(c),右枝分かれ境界にポーズがない数とポーズなしの文節境界数の割合(t),文節数の外れ値に関する素性(i–k)が説明変数から除かれた.残った 18 種類の素性と主観評価値との重相関係数を算出したところ

表 4 各素性と主観評価値の偏相関係数
Table 4 Partial correlation coefficients for each feature.

	素性	偏相関係数
a	モーラ数の平均値	−0.22
b	モーラ数の分散値	0.05
d	文節数の分散値	−0.17
e	モーラ数の外れ値の有無（$\mu_m + \sigma_m$）	0.13
f	モーラ数の外れ値の有無（$\mu_m + 2\sigma_m$）	−0.09
g	モーラ数の外れ値の有無（$\mu_m - \sigma_m$）	−0.22
m	左枝分かれ+ポーズあり	−0.38
n	左枝分かれ+ポーズなし	0.04
o	右枝分かれ+ポーズあり	−0.12
p	右枝分かれ+ポーズなし	−0.11
q	左枝分かれ+ポーズあり/左枝分かれ	−0.09
r	左枝分かれ+ポーズあり/ポーズあり	0.02
s	右枝分かれ+ポーズなし/右枝分かれ	−0.07
u	並列関係+ポーズあり	−0.02
v	並列関係+ポーズなし	−0.17
w	並列関係+ポーズなし/並列関係	0.03
x	並列関係+ポーズなし/ポーズなし	0.19
y	文節境界以外にポーズが挿入された数	−0.27

0.79 となり，高い相関が確認できた．

表 4 に主観評価値と各素性の偏相関係数を示す．偏相関係数が負となる素性の中では，左枝分かれ境界におけるポーズ数（m），文節境界以外へのポーズ数（y）が，比較的大きい偏相関係数が得られた．次に偏相関係数が大きい素性は，ポーズ句のモーラ数の平均値（a）およびポーズ句のモーラ数の外れ値の有無（$\mu_m - \sigma_m$）（g）であった．これらの結果から，係り受け関係にある句境界および文節境界以外の箇所にポーズが存在する場合，平均より極端に短いモーラ数のポーズ句が存在する場合，全体的にポーズ句が長い場合に，自然性の評価に負の影響を与えることが分かった．一方，偏相関係数が正になる素性の中では，並列関係かつポーズが挿入されていない境界数とポーズが挿入されていない境界の総数の割合（x）の偏相関係数が比較的大きく，並列関係におけるポーズの脱落が自然性の評価に正の影響があることが示唆されている．しかし，本素性は，偏相関係数が負となる（a，g，m，y）の 4 種類の素性と比べると偏相関係数の絶対値が小さく，主観評価値に与える影響は比較的低いと考えられる．

4.3 判別分析

重回帰分析により，複数の素性の組合せが主観評価値と高い相関を示すことが確認できたが，4.2 節で述べた 6 種類以外の素性は偏相関係数が全体的に小さく，自然性との関係は明らかではない．そこで，判別分析に基づく素性選択により，自然性と強い関係を有する素性の同定を行った．また，素性選択で得られた素性の妥当性を評価するため，判別分析によるクラス分類の精度を調査した．

4.3.1 クラスタ分析による主観評価値のクラス分類

まず，3 章で得られた 500 個の主観評価値を，1）自然性高，2）自然性中，3）自然性低の 3 クラスに分類した．これは，樹形図に基づく分析において，4 以上のクラスに分類した場合，各クラスに含まれるデータ数が最大で 4 倍以上の大きな偏りが生じたためである．また，定性的にも自然性を高，中，低の 3 段階で分類するのは妥当であると考えられる．各クラスの境界値はクラスタ分析により決定した．クラスタ分析は，評価者 24 名の評点および全評価者の平均値（主観評価値）をケースとして実施し，同一の主観評価値を持つデータが複数のクラスに分類された場合は，主観評価値が同じデータを用いてクラスの多数決を行い，最も多くのデータが含まれるクラスに分類した．非類似度計算法にはユークリッド距離を用い，クラスタ結合手法には Ward 法を用いた．分析の結果，クラス間の閾値は，2.9，3.7 となった．各クラスのデータ数の内訳は，自然性高が 104，自然性中が 244，自然性低が 152 となった．

4.3.2 変数増減法による素性選択

次に，前項で得られた主観評価値のクラスに分類するうえで有効な素性を決定するため，クラスと表 4 に示す素性との関係を判別分析により分析した．判別分析における素性選択には変数増減法 [15] を採用し，判別に有効な説明変数の組合せを決定した．変数増減法は，用意した変数の中から，選択する変数を事前に固定せず，変数の導入および除去を繰り返しながら，統計的に最も妥当性の高いモデルを作成する手法である．変数増減法における変数の導入および除去水準には偏 F 値を用いた．偏 F 値は，判別式を構成する各説明変数の係数の有意性を判断する検定統計量であり，変数増減法において，変数を順次導入する際に偏 F 値が基準値以上となった場合その変数を導入し，一方偏 F 値が基準値以下となった変数を除去する基準として用いられる．またその高低により，各説明変数の判別式への寄与度合を判断することができる．今回は基準値として 2.0 を採用し，表 4 で示した素性に対して偏 F 値を算出し，基準値を超える素性を順次選択する．ただし，素性を追加した際にはすでに選択した素性に対しても再度偏 F 値を算出し，基準値以下になる素性は削除した．

表 5 に，素性選択により得られた 13 種類の素性および各素性の偏 F 値を示す．選択された素性の中で最も偏 F 値が高いのは左枝分かれ境界のポーズ数（m）であり，自然性との関係が他の素性と比べて特に強いことが確認できた．また，文節境界以外にポーズが挿入された数（y），ポーズ句のモーラ数の外れ値の有無（$\mu_m - \sigma_m$）（g）の偏 F 値も比較的高いことが確認できた．

4.3.3 クラス分類による評価

素性選択により得られた 13 種類の素性の妥当性を評価するため，判別分析によるクラスの分類の精度を確認した．クラス分類には 4.3.1 項で分類したクラスを用い，判

表 5 変数増減法により選択された素性の偏 F 値
Table 5 Partial F values for each selected feature.

	変数	偏 F 値
a	モーラ数の平均値	4.9
b	モーラ数の分散値	3.3
e	モーラ数の外れ値の有無（$\mu_m + \sigma_m$）	6.8
f	モーラ数の外れ値の有無（$\mu_m + 2\sigma_m$）	2.7
g	モーラ数の外れ値の有無（$\mu_m - \sigma_m$）	12.4
m	左枝分かれ+ポーズあり	84.6
n	左枝分かれ+ポーズなし	7.3
o	右枝分かれ+ポーズあり	7.2
p	右枝分かれ+ポーズなし	7.6
u	並列関係+ポーズあり	2.3
v	並列関係+ポーズなし	2.9
x	並列関係+ポーズなし/ポーズなし	5.3
y	文節境界以外にポーズが挿入された数	12.4

表 6 自然性のクラスの推定結果
Table 6 Results of naturalness class estimation.

入力	推定クラス			正判別率
クラス	自然性高	自然性中	自然性低	(%)
自然性高	78	25	1	75.0
自然性中	77	132	35	54.1
自然性低	6	46	100	65.2
総合				62.0

別分析および主観評価値のクラスの推定の評価には同一のデータを用いているため，クローズな条件での評価となっている．

表 6 にクラスの分類結果を示す．自然性高，自然性低のクラスに関しては 6〜7 割の推定精度が得られており，特に，自然性高のクラスが自然性低のクラスに，自然性低のクラスが自然性高のクラスに誤推定された割合は，それぞれ 1 割以下に抑えられた．これより，表 5 で示す素性の組合せを自然性の判別に用いた場合，自然性が低い音声を高いと誤判断する可能性を下げることができる．前述のとおり，音声の主観評価では不自然に感じる箇所が 1 カ所でも存在すると評価が大きく低下する傾向があるため，本素性を用いて自然性の低いポーズ位置による合成音声を識別し除外することは，自然な合成音声を実現するうえで有用であるといえる．

4.4 選択された素性に関する議論

これまでの分析結果より，表 5 に示す素性の組合せでポーズ位置の自然性を説明できることが確認できたが，各素性自体は従来研究でもポーズ位置の決定に有効な要因としてあげられているものも多い．左枝分かれ＋ポーズあり（m），左枝分かれ＋ポーズなし（n），右枝分かれ＋ポーズあり（o），右枝分かれ＋ポーズなし（p），文節境界以外にポーズが挿入された数（y）の 5 種類の係り受け構造に関する素性の偏 F 値が比較的高いことは，係り受け構造がポーズ位置の決定に影響を与えることを確認した従来研究の結果と一致する．また，ポーズ句の長さに関する素性であるモーラ数の平均値（a）および分散値（b）も選択されているが，係り受け構造に関する素性と比べると偏 F 値が小さく，自然性に与える影響が副次的なものであることを示している．この点も杉藤ら [5] の研究でも示唆されているとおりである．一方，モーラ数の外れ値の有無（e，g）が，素性（n–p）と同等の偏 F 値を示しており，自然性に与える影響がこれらの素性と同程度の大きさであることが新たに確認できた．また，左枝分かれ＋ポーズあり（m）が他の素性の中でも突出して偏 F 値が大きく，係り受け関係にある句境界におけるポーズの存在が自然性に特に大きな影響を与えることも確認できた．本素性は表 4 の重回帰分析での偏相関係数が示すとおり，主観評価値とは逆相関であることから，自然性を下げる方向で働くことが分かる．これらの結果から，従来研究で示されていた係り受け構造とポーズ句の長さに関する素性が自然性に影響を与えることを，自然性の評価の観点から再確認しただけでなく，以下に示す新たな知見を得た．

- 句の長さが平均から大きく外れたポーズ句の存在は，自然性に与える影響が係り受け構造と同程度である．
- 係り受け関係にある句境界におけるポーズ挿入は，自然性を引き下げる最大の要因である．

5. ポーズ位置決定実験

前章の分析により，自然性に対する影響が大きいことが確認された 13 種類の素性が，音声合成におけるポーズ位置の決定に有効であることを確認するため，本素性を用いたポーズ位置の決定実験を実施した．

5.1 ポーズ位置決定の枠組み

従来より，音声合成において機械学習に基づくポーズ位置決定手法が提案されている [16], [17], [18]．また，識別モデルに基づく機械学習法である CRF [19] は，様々な分野で有効性が確認されており，ポーズ位置の決定においても有効であることが報告されている [20]．そのため，本研究で提案する素性のポーズ位置決定における有効性を確認するためには，これらの素性を直接用いて CRF を学習することが望ましい．CRF を用いた場合，動的計画法によりテキストの先頭から順次ポーズ位置が決定される．しかし，提案する素性の中にはテキスト中のポーズの位置関係に基づくものが含まれており，すべてのポーズ位置が確定していないと求められないものがある．そのため提案する素性をそのまま CRF で利用することは困難である．そこで，本研究では，主観評価値と提案する素性との関係に基づくポーズ位置の自然性の評価器と，従来の素性に基づくポーズ位置決定法を組み合わせた枠組みによりポーズ位置の決定を行う．

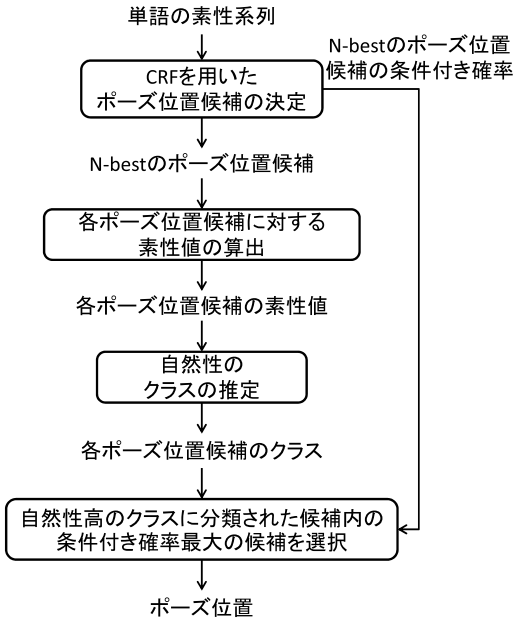

図 4 ポーズ位置決定の枠組み

Fig. 4 Overview of pause location prediction.

図 4 にポーズ位置決定のフローを示す．まず，従来の素性を用いて CRF によりポーズ位置を決定し，ポーズ位置の候補（N-best）を得る．次に，各候補に対して，表 5 に示す 13 種類の素性の値を算出し，主観評価値と提案する素性との関係に基づくポーズ位置の自然性の評価器を用いて自然性のクラスを推定する．そして，自然性が最も高いクラスに推定された候補の中から，CRF により算出される条件付き確率が最大となる候補を，ポーズ位置の決定結果として出力する．次節以降の実験では，従来の素性および提案する素性に基づいて決定したポーズ位置の自然性を主観評価実験により比較することで提案する素性の有効性を確認する．

5.2 実験条件

評価実験には，ニュース文と 1 名の男性プロナレータが読み上げ口調で各文を発話した音声を用いた．各文には，日本語形態素解析器 JTAG [21] および日本語係り受け解析器 JDEP [21] を用いて，単語境界，品詞，読み，文節境界，係り受け構造を付与した．また各発話には，3.1 節と同様に作業者の視察および聴取により，ポーズ位置，アクセント句境界，アクセント型を付与した．作業者が付与したポーズ位置と，自動解析により付与された文節境界が一致しない箇所が存在する場合は，該当箇所を含む文をデータから除外した．その結果，実験には 4,159 文を用いている．1 文あたりの単語数，文節数，ポーズ数の平均値は 23.5，15.5，3.3 である．

CRF の学習およびポーズ位置の決定には，CRF++ [22] を用い，モデルパラメータの学習には準ニュートン法である L-BFGS を用いた．素性には，文献 [8], [17], [18] で用いられた素性を参考として，単語の読み，品詞，文節境界の有無，係り受け構造（右左枝分かれ構造の種別）を用いた．学習データとして 3,743 文を用い，それ以外の 416 文は評価データとした．ポーズ位置の自然性の評価器には，4.3 節で述べた判別分析によるクラス分類を使用し，判別分析の係数は 4.3 節で求めたものを用いた．また，評価器への入力として，CRF で決定したポーズ位置候補の 10-best を用いた．

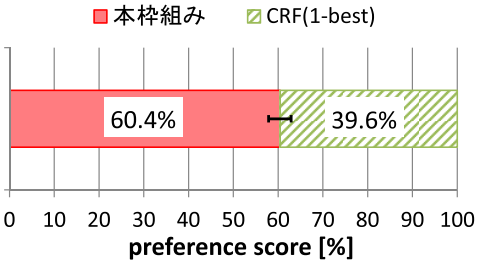

図 5 主観評価実験におけるプリファレンススコア

Fig. 5 Preference score from the subjective evaluation. (The error bar shows 95% confidence intervals.)

5.3 主観評価実験

本枠組みで決定したポーズ位置の合成音声に対する効果を確認するため，対比較による主観評価実験を実施した．評価音声には，本枠組みおよび CRF により得られた 1-best のポーズ位置に基づきポーズ挿入を行った合成音声を用いた．評価者として，3 章の主観評価実験とは異なる，正常聴力を有する成人 24 名を用いた．提示順序による影響を防ぐため，合成音声を入れ替えて，同一の組合せの音声を 2 回提示した．合成音声の生成には，3.1 節と同様に，波形接続型音声合成システム Cralinet を用いた．アクセント句境界，アクセント型は，評価データに付与されている情報を用いた．挿入するポーズの長さは一律で 0.5 秒，サンプリング周波数は 22.05 kHz とした．評価文として，CRF の学習に用いていない 416 文のうち，本枠組みと CRF のみの結果との間で差分が存在した 89 文から，さらにランダムに選択した 50 文を用いた．

図 5 に実験結果を示す．本枠組みで決定したポーズ位置に基づく合成音声が，60.4%の割合で，従来の素性を用いた CRF のみで決定したポーズ位置に基づく合成音声よりも自然性が高いと評価された．本実験結果により，本研究で提案した素性を用いることで，合成音声の自然性の向上が可能であることが確認できた．

6. おわりに

本研究では，合成音声の自然性の向上を目的として，音声の自然性に影響を与えるポーズ位置に関する要因の分析と評価を行った．まず，ポーズ位置およびテキストの内容が異なる 500 種類の音声試料を用いた主観評価実験により

ポーズ位置の自然性に関する主観評価値を収集し，ポーズ位置に関する素性と主観評価値との関係を重回帰分析および判別分析により分析した．分析の結果，ポーズ句の長さと係り受け構造の観点で設計した 13 種類の素性を自然性に対する影響が大きい要因として同定できた．また，句の長さが平均から大きく外れたポーズ句の存在は，係り受け構造と同程度に自然性に対する影響が大きいことを明らかにした．最後に，従来の素性に基づく CRF によるポーズ位置決定手法と，主観評価値と素性との関係に基づくポーズ位置の自然性の評価器とを組み合わせた枠組みでポーズ位置決定を行い，主観評価により同定した素性の有効性を確認した．なお，本研究では，ポーズ位置のみに着目した分析と評価を行ったが，ポーズの長さも音声の自然性や聞き手の理解度に対する影響が大きいことが従来研究からも示されている．今後，合成音声において，より自然なポーズ挿入を実現するためには，音声の自然性や聞き手の理解度と関係の強いポーズの長さに関する要因の分析と特定が必要と考えられる．

参考文献

[1] Hunt, A. and Black, A.: Unit selection in a concatenative speech synthesis system using a large speech database, *Proc. ICASSP*, pp.369–372 (1996).

[2] Zen, H., Tokuda, K., Masuko, T., Kobayashi, T. and Kitamura, T.: A Hidden Semi-Markov Model-Based Speech Synthesis System, *IEICE Trans. Inf. & Syst.*, Vol.E90-D, No.5, pp.825–834 (2007).

[3] 比企静雄：連続音声中の各種の区分の持続時間の性質，電子通信学会雑誌，Vol.50, No.8, pp.1465–1470 (1967).

[4] Fujisaki, H., Ohno, S. and Yamada, S.: Analysis of occurrence of pauses and their durations in Japanese text reading, *Proc. ICSLP*, pp.1387–1390 (1998).

[5] 杉藤美代子，大山 玄：朗読におけるポーズと呼気—息継ぎのあるポーズと息継ぎのないポーズ，音声言語 IV，近畿音声言語研究会，pp.199–211 (1990).

[6] Mandal, S., Saha, A., Basu, T., Hirose, K. and Fujisaki, H.: Modeling of Sentence-medial Pauses in Bangla Readout Speech: Occurrence and Duration, *Proc. INTERSPEECH*, pp.1764–1767 (2010).

[7] 海木延佳，匂坂芳典：局所的な句構造によるポーズ挿入規則化の検討，電子情報通信学会論文誌 D，Vol.79, No.9, pp.1455–1463 (1996).

[8] 藤尾 茂，匂坂芳典，樋口宜男：確率文脈自由文法を用いた韻律句境界とポーズ位置の予測，電子情報通信学会論文誌 D，Vol.80, No.1, pp.18–25 (1997).

[9] 鈴木和洋，斉藤 隆：日本語テキスト音声合成のための N 文節構造解析とそれに基づく韻律制御，電子情報通信学会論文誌 D，Vol.78, No.2, pp.177–187 (1995).

[10] 杉藤美代子：日本人の声，pp.104–113，和泉書院 (1994).

[11] 杉藤美代子（編集）：講座 日本語と日本語教育 第 2 巻 日本語の音声・音韻（上），明治書院 (1989).

[12] 河野守夫：音声言語の認識と生成のメカニズム：ことばの時間制御機構とその役割，金星堂 (2001).

[13] 別宮貞徳：日本語のリズム 四拍子文化論，ちくま学芸文庫 (2005).

[14] 間野一則，水野秀之，中嶋秀治，宮崎 昇，吉田明弘：顧客へのリアルな音声応答を実現するテキスト音声合成技術「Cralinet」，NTT 技術ジャーナル，Vol.18, No.11, pp.19–22 (2006).

[15] Rencher, A.C. and Christensen, W.F.: *Methods of multivariate analysis*, pp.293–296, John Wiley & Sons, Hoboken (2002).

[16] 布目光生，鈴木 優，森田眞弘：電子書籍の論理構造に基づくポーズ情報の推定と SSML 構造化，情報処理学会研究報告 DD，Vol.80, No.6, pp.1–7 (2011).

[17] 尾関 創，益子貴史，小林隆夫：多空間確率分布に基づくポーズのモデル化，電子情報通信学会技術研究報告 SP，Vol.104, No.30, pp.41–46 (2004).

[18] 海老原充，石川 泰：音声合成におけるネットワークモデルによるポーズ位置予測，電子情報通信学会技術研究報告 SP，Vol.96, No.566, pp.45–50 (1997).

[19] Lafferty, J., McCallum, A. and Pereira, F.: Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data, *Proc. ICML*, pp.282–289 (2001).

[20] Qian, Y., Wu, Z., Ma, X. and Soong, F.: Automatic prosody prediction and detection with Conditional Random Field (CRF) models, *Proc. ISCSLP*, pp.135–138 (2010).

[21] 今村賢治，斉藤邦子，浅野久子：テキストからの知識抽出の基盤となる日本語基本解析技術，NTT 技術ジャーナル，Vol.20, No.6, pp.20–23 (2008).

[22] Kudo, T.: CRF++: Yet another CRF toolkit, available from ⟨http://crfpp.sourceforge.net⟩ (accessed 2014-06-27).

武藤 博子 （正会員）

2009 年東京工業大学情報工学科卒業．2011 年同大学大学院情報理工学研究科計算工学専攻修了．同年日本電信電話株式会社入社．以来，音声合成のテキスト解析処理の研究開発に従事．現在，NTT メディアインテリジェンス研究所音声言語メディアプロジェクト研究員．日本音響学会会員．

井島 勇祐

2007 年八代工業高等専門学校専攻科修了．2009 年東京工業大学大学院総合工学研究科物理情報システム専攻修了．同年日本電信電話株式会社入社．以来，音声合成の研究開発に従事．現在，NTT メディアインテリジェンス研究所音声言語メディアプロジェクト研究員．電子情報通信学会，日本音響学会各会員．

宮崎 昇

1995 年東京工業大学情報工学科卒業．1997 年同大学大学院総合理工学研究科知能科学専攻修了．同年日本電信電話株式会社入社．以来，音声対話システム，音声合成の研究開発に従事．現在，NTT メディアインテリジェンス研究所音声言語メディアプロジェクト主任研究員．電子情報通信学会，人工知能学会，日本音響学会各会員．2000 年度電子情報通信学会猪瀬賞受賞．

水野 秀之

1986 年名古屋大学工学部電気電子学科卒業．1988 年同大学大学院工学研究科情報工学専攻修了．同年日本電信電話株式会社入社．以来，声質変換，音声合成の研究開発に従事．博士（工学）．現在，NTT メディアインテリジェンス研究所音声言語メディアプロジェクト主任研究員．電子情報通信学会，日本音響学会各会員．1993 年日本音響学会技術開発賞受賞．

阪内 澄宇

1993 年山形大学理学部物理学科卒業．1995 年東北大学大学院理学研究科物理学専攻修了．同年日本電信電話株式会社入社．以来，音声および音響信号処理の研究開発に従事．博士（工学）．現在，NTT メディアインテリジェンス研究所音声言語メディアプロジェクト主幹研究員・グループリーダ．電子情報通信学会，日本音響学会各会員．2001 年電子情報通信学会論文賞，2003 年日本音響学会粟屋潔学術奨励賞，2006 年日本音響学会技術開発賞各受賞．