音声言語情報処理 33－6
（2000．10．27）

統計的言語特徴を利用したテキスト音声合成の韻律制御

石川 泰　　中島邦男

三菱電機株式会社　情報技術総合研究所

{yasushi,kunio}@isl.melco.co.jp

あらまし　テキスト音声合成においては，合成音声の自然性の観点から，韻律制御が最も重要な課題のひとつである．韻律の決定要因のうち，文のイントネーションを決める要因が文の統語情報である．従来，この統語情報としては，文の統語解析から得られる木構造に基づくパラメータが利用されることが多かったが，木構造の抽出は必ずしも容易ではない．そこで，文の局所的な言語情報として，文節の言語的なカテゴリーの連鎖に着目し，その統計的な特徴を利用して韻律制御のパラメータを抽出する方式を提案する．本稿では，提案方式について述べ，ポーズ位置の抽出についての評価実験について報告するとともに，韻律パラメータの予測についての可能性を議論する．

Prosodic Control for Japanese Text-to-Speech Systems Using Statistical Language Models

Yasushi Ishikawa and Kunio Nakajima

Information Technology R&D Center, Mitsubishi Electric Corporation

Abstract　A prosodic control method is one of the most important issues for naturalness of synthesized speech in Japanese text-to-speech systems. In prosodic control, a syntactical feature of a sentence is an important factor for generating intonation. In general, parameters based on tree structure of a sentence are used as a syntactical parameter, However, it is not easy to extract tree structure from sentences. Thus, we proposed a prosodic control methods using statistical language models which represent local syntactical features of sentences. We also report results of experiments on estimation of pause position of utterances.

1. はじめに

テキスト音声合成の品質については，近年の技術の進歩により著しく向上している．そのひとつの要因がコーパスベースの研究手法にある．すなわち，以前の音声合成の研究が音声学・言語学の知見に基づく少数の音声の分析・観察から，テキストを音声に変換するための「規則」を発見的に求めることを目的とした研究手法であったのに対し，1980 年代以降，音声データベースの整備が行われ，統計的な手法を中心とした数理モデルと，評価尺度を導入し，コーパスに対して，最適なモデルを求めるとともに，オープンデータについての評価によりその問題点を把握する手法，すなわちコーパスベースの音声合成が主流になったのである[1]．しかし，この手法が音声合成分野で主に利用されて

いるのは，合成単位の作成と選択，音韻継続時間長制御の課題といえる．前者は合成単位で表現したときのスペクトル特徴の歪最小化問題であり，後者も種々の問題はあるものの予測時間と実際の継続時間との歪を最小化するという問題として捉えられ，その評価尺度が明確であることに加え，コーパスが整備されている音素ラベル付の音声データにより研究が進められることが大きい．

一方，韻律特徴のうち，イントネーション，アクセントの制御に対応する基本周波数の制御では，データベースの整備が遅れていること，評価尺度が本来は複雑な知覚上の尺度を考慮しなくてはならないことが問題となり，従来の発見的な方法や極めて少数のデータからのモデルの検討にとどまっている．

本稿では，韻律制御における我々の基本的な考え方を示し，つぎに少量の音声データからモデルを効率的に学習する一方法について提案し，さらに，提案手法をポーズ位置の推定に利用した場合の評価実験と，基本周波数パラメータの予測に利用する可能性について述べる．

2. ネットワークモデルによる韻律制御

2.1 ネットワークモデル

韻律モデルは，入力テキストに対して，

1) 基本周波数を決定し，その時間パタンとして聴取されるアクセント，イントネーション，パラ言語特徴を正確に再現する．
2) 各音韻の継続時間長を制御し，その文における特徴から，自然なリズムや，文節特徴を再現する．

ものであり，通常，入力テキストから「読み」，「統語情報」，「意味情報」などを抽出するとともに，テキストに表れない，「発話スタイル」，「感情」などの特徴を指定することにより，パラメータを生成する．従来は，このうち最も重要な統語的な特徴として，文解析の結果得られる木構造を基本とするパラメータが用いられることが多かった．例えば，隣接するフレーズの共通のノードまでのアーク数により数量化される分離度が典型的なパラメータであり，一般的に統語情報を大まかなクラスに分類し，韻律との関係を抽出しようとする方式であった[2]．しかし，木構造の抽出は容易な問題ではなく，大きな誤りが生じる場合がある．さらに，大まかなクラス分類に基づく方法では，十分な精度のモデルが得られない可能性が高い．我々は，韻律モデルとして，

・複雑な統語解析、意味解析を行わない
・音声コーパスからの学習が可能

である方式を検討し、ネットワーク遷移による韻律制御モデルを提案した[3][4]。これは、文を、文節を入力とする状態遷移で表し，状態は，文の統語的な構造のみでなく，韻律的な状態も表すものとし，状態の遷移，すなわち，統語的韻律的状態の遷移により，その韻律的な特徴量が決まるとするものである．

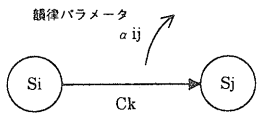

図 1. ネットワーク遷移による韻律パラメータの生成

従って，文全体の構造をネットワークで考えるのであれば，文頭から文末までの状態遷移ネットワークを、局所的な文節コンテキストのみを要因として考えるのであれば[5]、N 文節系列を表現するネットワークを考えればよい。この方法によれば、種々の韻律パラメータを状態遷移時に生成するモデルの学習が可能である。しかし、このようなモデルでは、

1) 利用する言語カテゴリをどのように選択するか、
2) 小規模な韻律データベースからどのようにモデルの学習を行うか、

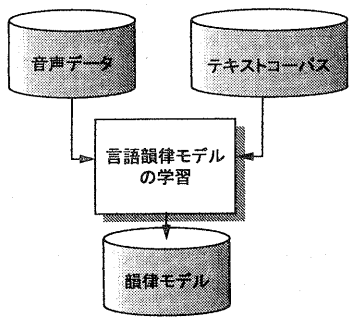

図 2 テキストの言語特徴を利用する韻律モデル

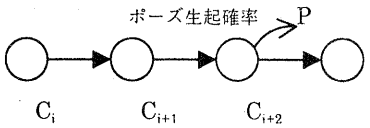

図 3. 3 文節連鎖に対するポーズの正義

が問題となる．特に，後者は学習に韻律特徴を付与した大量の音声データが必要となるコーパスベースの手法では，避けられない重要な課題である．そこで，大規模なコーパスを活用できるテキストデータを用いて言語特徴を効率的に活用する方式を提案した[4]（図 2）．すなわち、文に明示的に示されている言語的、韻律的境界である読点と文節列との関係を学習し、テキスト合成における韻律パラメータの予測の要因として活用するとともに、ポーズ予測のためのパラメータを韻律パラメータ推定の学習に利用しようというものである．読点の付与には，日本語の場合，決まった規則があるわけではないが，音節系列に対して読点がどの程度生起するかを統計的に求めることで，文の意味的な境界である確率，あるいは，その程度をモデル化できるものと予想される．

3. 統計的言語情報による韻律制御

3.1 読点予測

新聞記事[6],Email などのテキストコーパスを収集し，連続する N 文節の言語カテゴリ系列、C_i, C_{i+1}, C_{i+2},.. C_{i+N-1} について C_{i+N-2}, C_{i+N-1} 間，すなわち最後の 2 文節間に読点が生起する平均確率を求める（図 3）．

読点の位置の予測については，次に示す 2 つの方法が考えられる．

1) 確率最大の位置に読点を予測：与えられた入力文節系列について，確率の高い方から J 個の読点を予測する．あるいは，閾値判定を行い，確率が閾値以上であれば，読点を予測する．
2) 累積の確率を用いる予測：入力の i-L 番目の文節の前から I 番目の文節の前までに 1 回も読点が生起しない確率は，生起確率を p_i とすれば$(1-p_{i-L})(1-p_{i-L+1})\ldots(1-p_i)$となる．これが閾値を下回ったとき，すくなくともその間に 1 つ以上のポーズが生起したものとみなし，その間の確率値が最大の位置にポーズを予測する．

3.2 読点予測実験

このようなモデルにおいては，文の意味的な境界を精度よく予測する言語カテゴリの検討が重要となる．そこで，まず，読点の予測実験を行った．学習データは 43,019 文 615,132 文節、評価データは 13,228 文、188,728 文節である。読点の予測方法は，簡単のため，上記 1)を用いた．結果を表 2 に示す．なお，正解率は正解数/生成数×100%とした。比較した言語カテゴリは以下の 3 種類である．

A. 品詞カテゴリ：名詞句、動詞句など品詞による分類(9 カテゴリ)
B. 用法的カテゴリ：用言句を連用・連体に分類、体言句は主格性、所有性（「の」など)、その他に分類(14 カテゴリ)
C. 意味的・用法的カテゴリ：用言に接続の助詞により意味的な分類を行い，副詞，接続詞についても意味的 2 カテゴリに分類（17 カテゴリ）

表 1 読点予測精度(%)

カテゴリ	A		B		C	
N	2	3	2	3	2	3
精度	26.3	27.0	33.7	35.0	37.8	39.3

最も高い性能を示したカテゴリ C を用いた場合，さらに N=4 とした解きの精度を求めたが，39.7%と精度向上はわずかであった．表 2 にカテゴリ C の分類を示す．

3.3 ポーズ位置の予測

読点は，文の作成者が意図して意味的境界を示したものである．発話においては，ポーズの生起に強い関連があることが予測されるが，テキスト合成においては，読点のみをポーズ位置とみなすと，ききとりにくい合成音声となる場合がある．ポーズの位置を正しく予測することは，統語的な情報を合成システムが韻律特徴により伝送するという観点から重要な課題である．

表 2 カテゴリCの分類

#	意味
1	一般的名詞句
2	助詞「は」「が」などを伴う主格性の名詞句
3	助詞「の」などを伴う連体性の名詞句
4	「今日」など時間を表す名詞句
5	時間を表す名詞句で，助詞「の」など連体性の句
6	用言の連体形
7	「ながら」「から」など接続性の付属語を伴う用言
8	「けれど」「が」など反語的付属語を伴う用言
9	その他の用言
10	副詞句
11	「一方」など発言性の意味の副詞句
12	接続詞
13	反語的意味の接続詞
14	連体詞
15	感動詞
16	その他のカテゴリの文節
17	NULL(文頭，文末)

そこで，前節で学習したテキストに含まれない，種々の文章を用い，発話におけるポーズの特徴を調べるとともに，読点予測モデルによるポーズ予測の性能を調査した．

2名のプロの女性ナレータに、計6種類の文章を10セット読ませた。与えたリストには、原文どおりの句読点がふってあり、発声においては、ポーズの挿入については、一切の指示をせず、単に、「音声分析用のデータとして、明瞭に抑揚を抑えて発声する」こと，すなわち，ポーズをあまり意識せずに発声することを要求するか（以下平板読みと呼ぶ），あるいは「小説の朗読として感情を込めて発声する」こと，すなわち，文意が正確に伝わるよう考慮する（以下朗読と呼ぶ）ことのいずれかを指示した。表3に、音声データを示す。

表 3 音声データ(評;評論、新;新聞、平;平板、朗;朗読)

Set	1	2	3	4	5	6	7	8	9	10
話者	FKK				FKN					
内容	小説 1		小説 2		小説 3		随筆 1		評	新
文数	123		432		111		59		33	41
発声	平	朗	平	朗	平	朗	平	朗	平	平

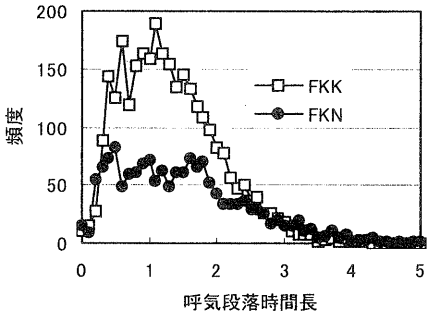

図 4 話者別の呼気段落時間長分布(sec)

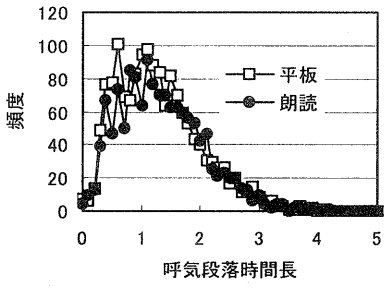

図 5 話者 FKK の呼気段落時間長分布(sec)

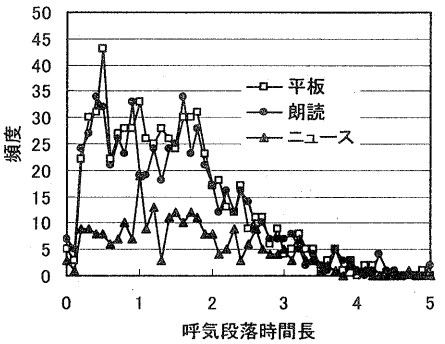

図 6 話者 FKK の呼気段落時間長分布(sec)

3.4 呼気段落長

ポーズの予測では、ポーズの挿入時点からのくの長さにより生起確率が上昇するとの報告がある[7]。そこで、ポーズで区切られた呼気段落の時間長について、分析を行った。話者別の呼気段落継続時間長の分布を図4,5,6に示す．図からは，今回の実験では、話者や発話スタイル、テキストと呼気段落時間長の間には，明白

な関連は見られず，0.3秒から2秒の間に多く分布すること，および，話者により若干の相違があるものの，時間長の分布は広いものであり，特定の時間長に集中する傾向は見られなかった．このことは，「息継ぎ」という観点で発話継続時間長が主たる要因となっているものではないことを示すものである．

3.5 読点とポーズの一致

ポーズ位置と、読点の一致の程度を分析した。表4に、文末以外の文中の全ポーズ位置に対しての、読点位置でのポーズの割合を再現率として、全読点に対するポーズ生起の割合を一致率として示す。

表からは、ポーズ位置と読点位置の一致については，テキストに大きく依存することがわかる。特に，話し言葉ではない新聞などでは，高い一致率と低い再現率が確認された．このことは，発声者が，専門用語や助詞の省略などが多発する新聞の読上げにおいて，誤解が生じないようポーズを挿入することで，意味的なまとまりを表出しているのではないかと推測できる．

表4 読点とポーズの一致率、再現率(%)

Set	1	2	3	4	5	6	7	8	9	10
一致	88	93	60	52	94	82	94	94	100	81
再現	51	52	57	62	58	65	42	45	30	40

3.6 読点位置予測

テキストに対する読点の位置予測を行った。読点の生成手法は，前述の方式2）を用いた．閾値は，正解の読点数とほぼ同じ数の読点を予測するよう、事後的に決定した．その結果，正解数1241に対して、1219の読点を生成し、うち、482個、39.5%が正解位置と一致した。

3.7 ポーズ位置予測

同様の方法で、ポーズ位置に対する予測実験を行った。結果を表5に示す。なお、読点位置予測と同様、話者別に事後的に生起ポーズ数が、発声ポーズ数とほぼ同じ数となるよう、閾値を調整した。

表5 ポーズ位置推定精度

話者	発声数	生成数	正解数	予測精度
FKK	2138	2283	771	33.8%
FKN	1165	1071	478	44.6%

4. F0パラメータと読点生起確率

5.1 実験

読点の生起確率を表すモデルによりポーズ位置の予測が比較的精度よく行えることが，確認できた．次に，このモデルと，韻律パラメータとの間の関係を分析する．プロの男性ナレータ1名が話者したATR音韻バランス文503文章について，ケプストラム法によりピッチパタンを求た．ピッチパタンとスペクトルの視察により，アクセントフレーズを抽出し，アクセントフレーズをフレーズ成分とアクセント成分の重畳で表現した場合のアクセント成分の高さ（図7）を視察により求めた．このパラメータは，テキスト音声合成における韻律パラメータである．このパラメータについては，統語的意味的情報以外の多くの要因が影響していることが確認できているため，その影響を低減するため、モーラ数（6カテゴリ）、アクセント型（2カテゴリ）を要因とする数量化I類分析を行い、得られた係数によるパラメータ補正を行った．前章で高い精度を示した3音節組みのカテゴリーCを用いて，3音節目のアクセント成分の高さと，読点生起確率の関係を調査した．発話におけるポーズが生起する確率が高いことは、その間の意味的な分離性が高いことを意味するため、3文節目のアクセント成分の高さは，先行文節との関係が弱いため，高くなることが予想される．

図8に結果を示す。

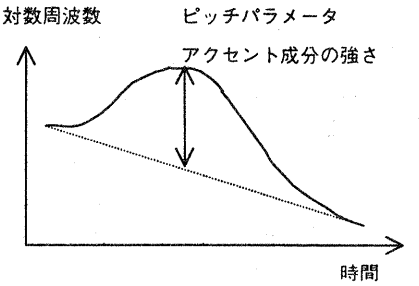

図7 ピッチパラメータ

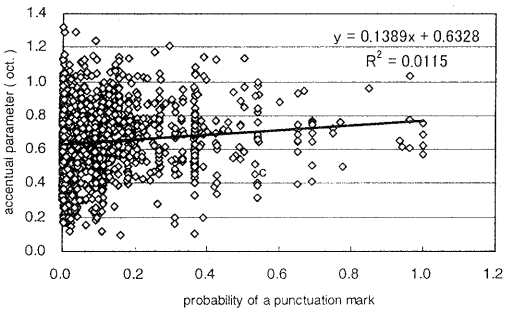

図8 読点生起確率とピッチパラメータとの関係

6.2 考察

ピッチパラメータと読点生起確率との関係を示すグラフ中に，回帰直線と平均誤差を示す．これは，ピッチパラメータを読点生起確率から回帰直線により推定した場合の性能を示している．個々のピッチパラメータと回帰直線の予測誤差は比較的大きいことが分かる．これは，今回の実験では，モーラ数とアクセント型の影響のみを考慮したが，そのほかにも，先行アクセント句のアクセント型，実際のポーズの有無など，ピッチパラメータの変動要因としては，多種多様な要因があるためと考えられる．しかし，予測されたとおり，ピッチパラメータと生起確率の間には相関があることが確認され，読点の生起確率が言語的な情報の表現として，有効であることが確認できた．

5. まとめ

コーパスベースの韻律制御方法を検討した．文の統語的な情報と韻律の関係を明確化し，モデル化するにあたり，

1) 発見的な方法による統語の分類を行わず，なんらかの言語的な特徴量に基づいた統語情報の抽出を行いたい
2) 統語的な状態が，発声という行為の中では，韻律的な状態となっているものと考え，韻律制御と統語解析を統合的に考えたい
3) 少量の韻律コーパスから学習を行いたい

を基本的な考え方として，ネットワーク遷移による韻律制御を検討した．大量のテキストコーパスから得られる意味的な境界，すなわち明白な韻律境界を表す読点に着目し，言語カテゴリーの系列と読点の関連について検討したところ，用法的意味的な言語カテゴリを考えることで，比較的高い精度で読点位置を予測できることが確認された．またこれを直接ポーズの生起位置の予測に適用したところ，約 40%の正解率が得られ，方法の有効性を確認した．さらに，この読点予測のパラメータがピッチパタン生成モデルのパラメータと関連があることを実験から求めた．このことは，韻律モデルの学習における言語情報のパラメータとしての有効性に加え，韻律の学習における初期値，あるいは少数データによる学習時の補間用の値としての利用可能性を示すものである．

今後は，ピッチパラメータの直接予想に適用し，従来の言語情報を少数のクラスに分類して数理モデルによりパラメータを推定する場合との性能の評価を実施する予定である．

参考文献

[1] 匂坂「コーパスベース音声合成」，信号処理，Vol.2, No.6, pp.407-414（1998）
[2] M.Abe et.al, "Two-Stage F0 Control Method using Syllable based F0 Units" ICASSP'92 IIpp.53-56 (1992)
[3] 石川，海老原「ネットワークモデルによる文ピッチ生成法」信学技報告、SP96-41(1997.07)
[4] 石川、中島「テキスト音声合成における統計的言語情報を利用した韻律制御」音講論集、1-7-17(2000.03)
[5] 海木,匂坂，「局所的句構造に基づく F0 制御」，信学技報，SP92-6（1992）
[6] 毎日新聞データベース 95 年
[7] 藤崎ほか「文章朗読における休止挿入の確率的規則とその評価」音講論集,pp.251-252（1999.3）