音楽情報科学　29－14
（1999. 2. 19）

局所的話速変化検出を指向した簡易セグメンテーション手法と実環境音声への適用について

佐々木 淳一　広重 真人　宮永 喜一　栃内 香次
北海道大学工学部
〒060-8628　札幌市北区北13条西8丁目
Tel.　(011)706-6823
FAX　(011)709-6277
E-mail:　ssk@media.eng.hokudai.ac.jp

あらまし

本報告では、自発音声の局所的話速変化自動検出を目的とした簡易セグメンテーション手法について述べる。一般に話速変化に乏しいと言われる日本語音声では、まれに現れる話速変化箇所が強い表現力を持つ可能性があると考えられる。本手法は話速変化情報を音素の認識以前に得ることを目的とし、将来的に連続音声認識や言語的復元・意味理解等の各処理に有益な情報を提供することを想定している。本報告では、話者の意図と対応をとりやすい意味的な「まとまり区間」ごとに音素変化回数を算出することにより話速変化を表現することを試みる。また、話速変化を豊富に含むと考えられる自発音声を実環境において収録し、話速変化検出実験を行なった結果について述べる。

キーワード　話速変化、韻律情報、音素セグメンテーション、実環境音声

A simple method of phoneme segmentation for detection of local speech rate variation and its application to speech data recorded in real environment

Junichi SASAKI, Makoto HIROSHIGE, Yoshikazu MIYANAGA and Koji TOCHINAI
Department of Engineering
Hokkaido University
13-8, Kita-ku, Sapporo, 060-8628, Japan
Tel.　+81-11-706-6823
FAX　+81-11-709-6277
E-mail:　ssk@media.eng.hokudai.ac.jp

Abstract

In this report, a simple method of phoneme segmentation for detection of local speech rate variation is described. It is said that Japanese speech contains less variation of speech rate than other languages. Thus if there is a large variation of speech rate, such portion of speech may contain speaker's strong intention. The speech rate obtained with this method is expected to supply useful information to speech recognition system, recognition error correction system, or speech understanding system. In this report, we express speech rate variation by calculating the number of phoneme changes within a particular time duration that is selected to easily express speaker's intention. Spontaneous speech data which may contain large amount of rate variation are recorded in real environment. The results of rate detection experiments for the recorded speech data are described.

key words　Speech rate variation, Prosodic information, Phoneme segmentation, Speech data in real environment

1 はじめに

近年音声研究において、発声された音韻そのものが何であるかという言語情報だけではなく、音声として表現された話者の意図情報などもある程度とらえられるような、韻律情報処理に関する研究を行なうことは重要な課題となってきている[1][2]。

音声波形に含まれる韻律情報は、大きく分けてピッチ、パワー、話速の変化で表されるが、そのうちピッチとパワーについてはすでに各所で研究がなされており、島駆動的音韻認識、係り受け解析[3][4]、強調箇所選定、重要語選定等への応用が考えられている。しかし、局所的な話速変化に含まれる情報に関する研究例[5][6][7]は多いとはいえない。筆者らは音韻認識から意味理解にわたるまで利用可能な「意図的な話速変化情報」を検出することを目標としている。一般に日本語は話速変化に乏しいと言われている。したがって意図的な話速変化は日本語にはまれであるが、まれであるゆえ強い表現力を持つと考えられる。そのような話速の変化のある個所を捉える能力を持つことは、感性情報処理の観点からも重要であろう。

話速に関する研究では、発話速度の変化パターンを自然な合成音作成に利用することを目的としているものが多い[6][8]。これは自然な発話中に現れる無意識な音素時間長の規則を生成することにあたり、これらは筆者らが目指す意図的な話速変化を認識するものとは方向性が異なる。本では将来的に、得られた話速変化情報をピッチ・パワー等の情報と合わせて連続音声認識システムに提供し、重要な部分から順に音韻認識を行うことや、音韻認識後の言語的復元・意味理解等の各処理に有益な情報を提供することを想定しているため、音素認識以前に簡潔な手法によっておおまかな話速変化を捉えることを目的としている。

意図的な話速変化を検出する際には、得られた話速変化が人間の知覚と一致していることが望ましい。しかし現在のところ、話速をどのように表現すれば人間の知覚に最も近くなるかは充分には明らかになっていないと考えられる。したがって本報告では、当面の方法として以下の2点に基づいて話速を表現することとした。

- 話速の単位について
  話速の単位では、現時点で利用可能で一般的なものとしてモーラ速度 [mora/sec] がある。しかしモーラ速度を音響的特徴のみから直接算出することは困難である。したがって本報告では、より音響的特徴に対応している音素速度 [phoneme/sec] を用いることとした。

- 局所的話速の表現について
  連続的に発声された音声の局所的話速の表現には、その時間分解能の大小により瞬時速度に相当するものから文章1つにわたる平均速度など様々なものが考えられる。本報告では、検出された話速変化を人間の知覚とよりよく対応させるために、適切な長さ（時間分解能）の「まとまり区間」で平均音素速度を算出する。まとまり区間が長すぎると、意図的変化が埋もれてしまい、短すぎると、意図的ではない(無意識の)音素長変化の影響を受ける。これらの中間の長さで、かつ話者の意図との対応がとり易いような、単語や文節、アクセント句程度の、意味的なまとまりを持った長さが望ましい。この「まとまり区間」への自動区分けは、将来的にはアクセント句等への自動区切りの研究成果[3][9]を利用することを想定し、現時点では手動で区切ることとする。

本報告の目標は局所的な話速変化の検出であるため、音韻認識を想定した従来のセグメンテーションに比べて正確な区切り位置は必要なく、区切り回数のみ検出できればよい。さらに、話者の意図情報という観点では、絶対的な話速値よりも相対的話速変化に意味があると思われる。これらをふまえて、本報告では簡潔なセグメンテーション手法により音素変化回数に基づく話速変化の記述を試みる。

また、話速変化について論じるには朗読音声と自発的に発声された音声との違いも考慮しなければならない。本研究ではより自然な自発的音声へシステムを適用するため、実環境において新たに音声データの収録を行った。

本報告では、話速を求めたい音声データから有声部と無声部でそれぞれ別のスペクトル変化パラメータを抽出して、スペクトル変化量の極大点を可変閾値[11]を用いて検出し、それを音素境界とみなして話速を算出する。まとまり区間ごとに得られた話速値の差分をとり、その正負によってシステムの評価を行う。

以下、第2章で話速変化検出の手法について、有声部と無声部それぞれに対して説明する。次に第3章で実環境音声データの収録について説明し、第4章で話速変化検出結果を示し、第5章で考察を述べる。

2 話速変化検出手法

本報告では音素境界検出のためにスペクトル変化パラメータを使用する。有声部と無声部の性質の違いに基づき、異なるパラメータを用いている。

2.1 有声部でのスペクトル変化パラメータ

有声部のスペクトル変化パラメータとしては、パワースペクトル全体の形状から変化量を算出するLPC ケプストラム距離などがあるが、パワースペクトル全体の変化情報を用いたパラメータでは、あまり知覚されないスペクトルの谷の部分や微細構造部分の変化も、フォルマント部分の変化と同様にパラメータに等しく影響してしまう。このため、音韻が変化しない部分でも小さなピークが検出されたり、

母音連続部など、スペクトルのゆっくりとした変化に基づく特徴が、微細構造部分による変化によって検出しにくくなる。そこで有声部ではスペクトルの微細構造にはこだわらず、フォルマント部分の情報のみを利用したスペクトル変化パラメータを用いることにする。

入力音声に対して有声／無声判定を行ない、有声と判定された区間内でフレームごとに12次の線形予測分析によるフォルマント周波数を推定する。フレーム切り出しにはハニング窓を用い、フレーム長30msec、フレーム周期5msecとした。このとき無声と判定された箇所はフォルマント周波数を0としておく。

隣接フレームごとに得られたフォルマント周波数の差分をとり、これを有声部におけるスペクトル変化パラメータとして音素境界検出に利用する。フォルマント移動軌跡から音素境界を検出する手法については、2.3節で述べる。一般的に、有声音の知覚には第1から第3の3つのフォルマントが重要だといわれているが、複数フォルマントの組合せによるシステムの複雑化をさけ、最も安定に推定可能な第1フォルマントのみを用いることにした。

2.2 無声部でのスペクトル変化パラメータ

日本語において、表記上は無声子音の連鎖は存在しないため、本システムのプロトタイプにおいては無声と判定された箇所については一つの音素とみなしていた[10]。しかし実際の音声では母音の無声化、脱落、無声子音との融合化などが頻繁に起こり、ひと続きの無声区間に複数の音素が含まれる場合がある。したがって、局所的話速変化検出を議論するにあたって無声子音の連鎖も考慮する必要がある。本研究ではフォルマントによる処理ができない無声区間において、LPCケプストラムを用いた音素境界検出を行なっている。

入力音声データから2.1節で無声と判定された区間を切り出し、その区間内でフレームごとに20次のLPCケプストラム係数を算出する。フレーム長は30msec、フレーム周期は2.5msecとした。隣接フレーム同士でLPCケプストラム距離を計算し、これを無声部におけるスペクトル変化パラメータとして音素境界検出に利用する。

また、スペクトルによる音素境界検出では、無声破裂音に先行する閉鎖区間（closure：以降、clと表記）も1音素として検出されてしまう。無声子音が連鎖して間の母音が無声化する例では、後半の無声子音が破裂音となる頻度が比較的高いため、clの扱いには留意する必要がある。本システムでは、各無声区間内でパワー閾値を用いてcl区間を検出しておき、clの「開始点」にあたる境界を、ケプストラム距離によって検出された音素境界候補列から取り除いて音素数の補正を行なっている。cl区間を前の音素と統合するのは、スペクトルによる音素境界候補検出ではcl終了点よりも開始点の方が検出されにくく、また音素変化回数を求めるという本報告の目的においては各音素境界の位置がずれることは問題にならないと判断したからである。

2.3 スペクトル変化パラメータからの音素境界検出手法

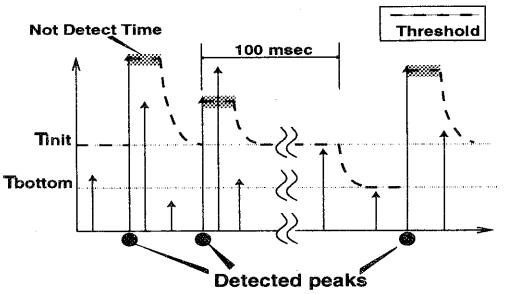

図1. 指数減衰閾値法の説明図

有声部での第1フォルマント移動軌跡および無声部でのLPCケプストラム距離から音素境界を検出するには、スペクトル変化量の極大値を捉えるための閾値を設定する必要がある。本システムでは、適切な間隔で音素境界を検出するために、指数的に減衰する可変閾値$y(t)$を用い、ピーク検出直後には不検出区間を設定する[11]。

検出操作開始時、$y(t)$の初期値はT_{init}に設定される。パラメータが$y(t)$を越えれば、その時刻にピーク検出フラグを出力する。このとき、検出時点でのパラメータ値を$y(t)$とし、一定期間τだけ保持して、保持期間内はピークの検出を行なわない（不検出区間）。

不検出区間が終了した時刻から、閾値は式（1）に従って降下し始める。

$$y(t) = Ae^{at} \tag{1}$$

A：（直前の不検出区間での閾値）
t：（減衰開始点からの経過フレーム数）
a：（減衰定数）

$y(t)$の下限はT_{init}とするが、母音が連続する箇所などはスペクトルが少しずつゆっくりと変化するため、長い時間にわたって距離値が低い値に抑えられてしまい、音韻変化を検出できない場合がある。その対策として、100msec以上ピークが検出されなければ、さらに$y(t)$を下げる。そのときの減衰式を式（2）、下限をT_{bottom}とする。

$$y(t) = Ae^{bt} \tag{2}$$

b：（減衰定数）

ただし、無声区間は有声部に比べて継続時間長が短いので、可変閾値に第2の下限T_{bottom}を使用するメリットが少ない。したがって、無声区間では

	有声部	無声部
τ	20msec	10msec
a	-0.39	-0.6
b	-0.02	—
T_{init}	4.8Hz	0.0054
T_{bottom}	2.4Hz	—

表 1. 指数減衰閾値法の各パラメータ

T_{bottom}を設定しないでT_{init}のみを用いる。図1は指数減衰閾値法の説明図である。各パラメータの決定には実験データと同じ環境で収録された学習用データを用意し、式3で表される音素検出誤り個数Eが最小になるような値を与えた[12]。表1に各パラメータ値を示す。

$$E = \sum_{s=1}^{S} \sum_{n=1}^{N} |Pr_{s,n} - Pd_{s,n}| \qquad (3)$$

S :(学習データの総文数)
N :(各学習データ内のまとまり区間数)
Pr :(リファレンス側のまとまり区間内音素数)
Pd :(本システムによって検出されたまとまり区間内音素数)

3 実環境音声データの収録

本研究において従来より用いてきた、独自に録音した実験データは、予め発話内容を指定し、無響室内で録音した読み上げ音声であった[10]。これらは読み上げるテキストの内容について無声子音・連続母音等のバランスを考慮し、さらに話速が単調にならないように、特定の箇所を特にゆっくり発話するよう依頼することによって部分強調を施すなどの工夫がされていた。しかし本研究で興味の対象としている話速変化は、本来実生活の中での会話に含まれているものである。この点を考慮すると、なるべく録音を意識しないような方法で収録した、雑音や残響を含む実生活での音声から、充分な精度で話速変化を抽出できるようにシステムを構築し調整する必要がある。したがって本報告では、新たに実環境で自発的に発声された音声データを収録した。

収録は大学のゼミ室（奥行 8m ×幅 4.6m ×高さ 2.8m）で行なった。この部屋には特別な雑音対策は施されていない。話者自身の声の残響がある程度存在するが、接話形ヘッドセットマイクロホンを用いることによって残響の影響や他話者の音声を抑制した。収録データはゼミでの論文紹介という状況で、内容を制限されない自発的な発話である。論文紹介の開始から終了まで、およそ 60 分にわたってすべての発話を収録した。DAT にて 48kHz で録音した後、パーソナルコンピュータ上で数文を切り出し、ワークステーション上で 10kHz にダウンサンプリングした。切り出す文を選択する際には、以下の点に注意した。

- 意味の通るひとまとまりの文で、前後がポーズによって他の発声と明確に分けられているもの。
- 他話者の発声や体を動かすことなどによって発生するノイズ、咳払いなどの非言語音が少ないもの。
- 1 文が長過ぎたり短か過ぎたりしないもの。（2.5～8sec 程度を目安とする）
- 話者が録音を意識しない自然なデータを得るために、収録開始直後のものを避ける。
- 文頭において出現頻度が高かった単一モーラの無意味語（「で、」「や、」「ま、」など）は文に先行するポーズ区間の関係上、含めた状態で切りだすが、話速変化に与える影響がが少ないと思われるため、以降の処理では考慮しない。

切り出した音声データのラベリングは、ワークステーション上で行なった。ラベリングに用いたソフトウェアは Entropic 社の「ESPS/waves+」である。ラベリングは時間波形とスペクトルの目視および聴取によって音素ごとに行なった。

このようにしてラベル付与した音声データの総数は 19 文である。一般に自発音声に多く含まれる次のような特徴が、本データにおいても観測された。

- 「えー」「えーその」といった無意味語が多く含まれている
- 句の最終モーラを伸ばして発話するサンプルがあった
（例:「あつかっていく」→「あつかっていくー」）
- 表記上は 2 モーラとみなされるべき長母音が短縮しているため、発音上は 1 モーラとなっているサンプルがあった
（例：「とゆーことで」→「とゆことで」）

話速値を算出していく「まとまり区間」は本来は自動化するべきものであるが、本研究では話者の意図との対応をとり易い区間するため、以下の手順に従って手動で決定する。

(1) テキストに書き起こした発話内容、時間波形の目視、聴取によっておおまかな区切り点を定める。ここでの「まとまり区間」は、話者の意図と対応をとりやすい単位とするため、文節相当の長さとする。

(2) ピッチ情報も考慮するため、ESPS の get_f0 コマンドにより得られたピッチ変化を目視しながら、分割･統合を再度検討する。ピッチ変化を用いた「まとまり区間」は、アクセント句に相当する「山型」を 1 区間の基本とする。したがって、上述（1）の処理で 1 区間と判断された箇所

においても、2つ以上の「山型」が含まれている場合には、その谷の部分で「まとまり区間」を分割する。さらに本データにおいては、話者がピッチを平坦（あるいは下降気味）に発話した後「しりあがり」になるものをひとまとまりの単位としながら発話する例が多く見られた。したがって、上述（1）の処理で「まとまり区間」が「平坦（あるいは下降）」＋「しりあがり」の並びになっている箇所は統合した。また、「山型」の途中で区切れていた場合も統合した。

「まとまり区間」への分割作業に大きなかたよりがないことを確認するため、全音声データ19文にわたって、まとまり区間の継続時間長と含有音素数の統計をとった。結果を表2に示す。また、全音声サンプルにわたる「まとまり区間」の時間長と含まれる音素数のヒストグラムを図2に示す。ヒストグラムから「まとまり区間」の継続時間長、含有音素数ともに多くが平均のまわりに分布し、極端に長い／短いものは少ないことがわかる。時間長が最も長い区間は、「デルタケプストラム（を）用いた」という発声であり、含有音素数も19と比較的多いため分割することも考えられるが、この間は助詞「を」が脱落するほど強い結び付きでひと息に発声されたものであるため、区切ることは不可能と考える。同様に、時間長が2番目に長いのは「セグメンテーションとの間に」と発声された区間であるが、これも聴取してみた結果明らかにひと息で発声されたものであったため、分割はしなかった。含有音素数が1となっている区間は、すべて「えー」という無意味語であったため、これも統合の必要はないと判断した。

図2において継続時間長が1100～1200 msecの間となったまとまり区間は6つあったが、それらに対応する含有音素数を調べてみると、9～21個と比較的広い範囲にわたっていた。図中の矢印が各サンプルの対応を表す。同程度の継続時間長にもかかわらず含有音素数が異なるということは、収録した音声データが話速の変化を豊富に含んでいることを示すと言える。

	継続時間長	含有音素数
平均	655msec	8.6
標準偏差	240msec	3.6
最大値	1325msec	21
最小値	215msec	1
総区間数	113	

表 2. まとまり区間の継続時間長と含有音素数の統計

4 局所的話速変化値検出実験

第3章で収録した実環境音声データに第2章で説明した音素境界検出システムを適用すると、入力音声データの音素境界列が得られる。これによって、手動で決定した意味的まとまり区間ごとに単位時間

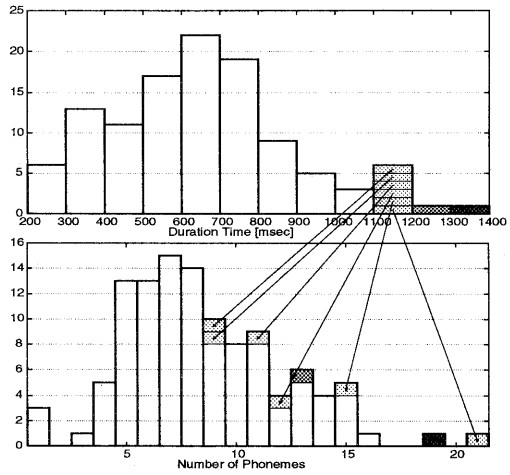

図 2. 上：まとまり区間の時間長のヒストグラム、下：含有音素数のヒストグラム

あたりの音素数[phoneme/sec]を計算すれば、局所的話速変化値が得られる。

4.1 音素境界検出の結果

表3は各「まとまり区間」ごとに検出された音素の個数である。

各音声データは「まとまり区間」の数が異なっている。表中上段の数値は本システムによって検出された音素数で、下段がラベルに基づくリファレンスの音素数である。音素境界検出の精度を評価するために、音素境界正解率C_{phone}を式（4）のように定義した。

$$C_{phone} = \left(1 - \frac{\sum_{n=1}^{N} |Pr_n - Pd_n|}{\sum_{n=1}^{N} Pr_n}\right) \times 100 \quad (4)$$

N :(各学習データ内のまとまり区間数)
Pr :(リファレンス側のまとまり区間内音素数)
Pd :(本システムによって検出されたまとまり区間内音素数)

図3は各文のC_{phone}である。19文全体（まとまり区間の数113区間）に対するC_{phone}は81%であった。個々の「まとまり区間」に注目すれば検出精度の悪い箇所もあるが、本報告の目的は厳密な音素セグメンテーションではなく局所的な話速変化を検出することであるため、ここで得られた音素境界検出結果をさらに話速の変化と関連づけた形で表現する必要がある。次節では、ここで得られた結果を音素速度[phoneme/sec]の形と、その差分の形での表現を試みる。

No.		まとまり区間 1	2	3	4	5	6	7	8	9	10
1	P	7	11	8	8	6	5	9	5	12	
	P_{ref}	9	8	6	7	5	5	7	4	9	
2	P	21	21	4	4	9					
	P_{ref}	19	21	5	4	6					
3	P	4	6	14	8	4	12	2	13	15	12
	P_{ref}	5	6	15	7	5	10	1	11	14	12
4	P	13	5	9	15	6					
	P_{ref}	12	6	7	13	6					
5	P	9	7	16	7	1					
	P_{ref}	13	7	15	6	3					
6	P	6	11	9	7						
	P_{ref}	7	10	11	8						
7	P	7	16	13	4	11					
	P_{ref}	6	10	13	4	9					
8	P	10	7	11	7	10	8				
	P_{ref}	11	7	9	8	8	8				
9	P	3	8	6	13	12	10				
	P_{ref}	5	5	6	9	11	7				
10	P	5	11	10	9	4	13				
	P_{ref}	5	16	7	9	7	9				

No.		まとまり区間 1	2	3	4	5	6	7	8	9	10
11	P	6	12	12	7	14	8	5	10	4	13
	P_{ref}	6	13	9	4	14	7	7	11	1	12
12	P	7	15	13	8						
	P_{ref}	7	15	12	10						
13	P	10	8	12	9	8	10				
	P_{ref}	8	8	8	8	5	8				
14	P	7	10	9	5	7	7	6			
	P_{ref}	10	9	8	8	5	10	8			
15	P	2	6	1	8	4	13				
	P_{ref}	4	7	1	6	6	13				
16	P	13	10	8	8						
	P_{ref}	15	14	11	5						
17	P	8	15	11	7						
	P_{ref}	11	14	10	8						
18	P	4	17	18	9	13	6				
	P_{ref}	5	11	15	11	13	6				
19	P	9	10	7	6	4					
	P_{ref}	10	9	7	5	6					

P_i: 検出された音素の個数
P_{ref_i}: リファレンスの音素数

表 3. 各まとまり区間ごとの検出音素数

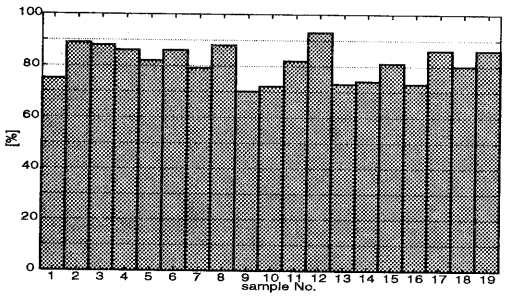

図 3. 音素境界検出率 C_{phone}

4.2 局所的話速値変化の算出

前節で得られた音素境界をもとに、単位時間あたりの音素数を計算した。図4～7に特徴的な例を示す。

図4、5は、話速変化が比較的良く検出されている例である。図中、上段の実線がシステムによる検出結果で、破線がラベルによるリファレンスの話速値である。下段は前フレームとの差分をとったもので、1文全体の絶対速度の大小を除いて相対的な話速の変化だけを表すものと言える。図中の縦線がまとまり区間である。まとまり区間でラベルが付与されていないものはポーズ区間である。ポーズ区間の含み方も聴覚的な話速変化に影響を与える重要な要素であるが、本研究では今のところポーズ区間を処理の対象外としている。図4、5ともリファレンスのグラフと良く似た形をしている。特に図5から話速値が「ゆっくり」であると予想される「／ゆーいな／」の区間では、実際に聴取してみてもゆっくり発話されていることが確認できる。

図6、7は、話速変化がうまく検出されなかった例である。図6の例では全体的に検出過多であり、差分をとっても変化の増減方向がほとんど逆になってしまっていることがわかる。図7の例では、スペクトル変化では境界を捉えにくい音素（例えば/r/など）が多く含まれていることにより、結果として話速値が実際のものと大きくずれてしまっている。

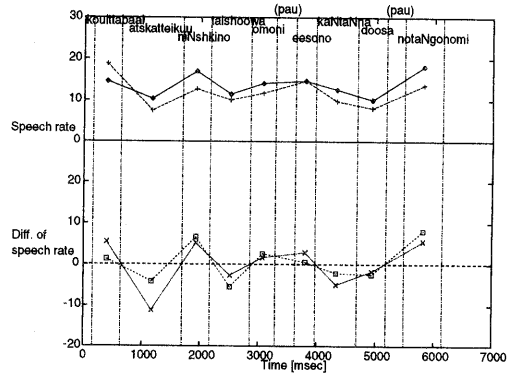

図 4. 局所的話速変化検出結果の例 1

／こういったばあい／あつかっていくー／にんしきの／たいしょーは／おもに／えーその／かんたんな／どーさ／のたんごのみ／

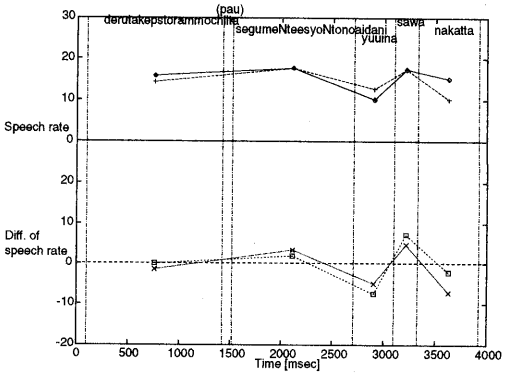

図 5. 局所的話速変化検出結果の例 2

／デルタケプストラムもちいた／セグメンテーションとのあいだに／ゆーいな／さは／なかった／

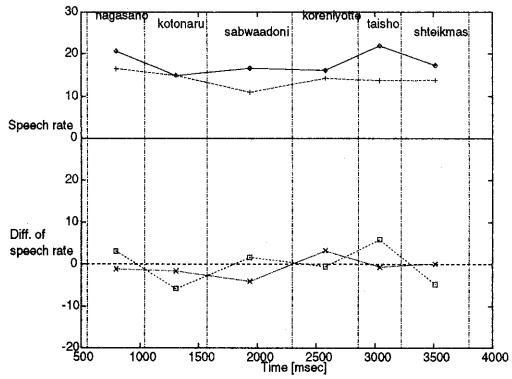

図 6. 局所的話速変化検出結果の例 3

／ながさの／ことなる／サブワードに／これによって／たいしょ／していきます／

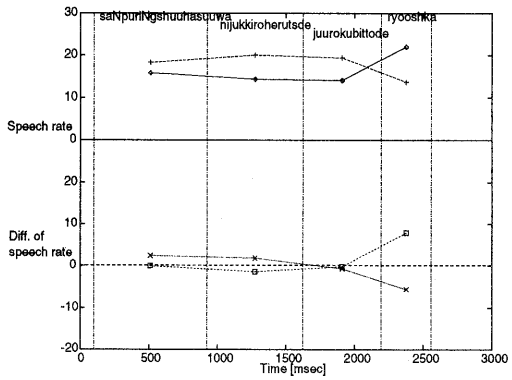

図 7. 局所的話速変化検出結果の例 4

／サンプリングしゅーはすうは／にじゅっキロヘルツで／じゅーろくビットで／りょーしか／

4.3 実験結果の評価

本システムでは、音素の認識以前に捉えたおおまかな話速変化情報を連続音声認識システムに提供し、重要な部分から順に認識を行うことや、音韻認識後の言語的復元・意味理解等の各処理に有益な情報を提供することを目的としている。そのためには、得られた話速変化値を評価するには、絶対的な量でどれだけの話速値かというよりも、局所的・相対的な話速の増減方向が正しく捉えられているかという点を重視すべきである。したがって本節では、4.2節で得られた話速値の差分について、その正負がどの程度リファレンスと一致していたかを評価するため、話速変化検出正解率 C を式（8）で定義する。C は、ひとつ前の「まとまり区間」に比べて話速が速くなったか遅くなったかという増減の方向をどの程度捉えられていることを示す。

$$Dd_i = Rd_i - Rd_{i-1} \tag{5}$$

$$Dr_i = Rr_i - Rr_{i-1} \tag{6}$$

$$c_i = \begin{cases} 1, & (Dd_i \cdot Dr_i \geq 0) \\ 0, & (Dd_i \cdot Dr_i < 0) \end{cases} \tag{7}$$

$$C = \frac{\sum_{i=1}^{N} c_i}{N} \times 100 \tag{8}$$

i：（まとまり区間番号）
N：（まとまり区間総数）
Rd：（検出された話速値）
Rr：（リファレンス側の話速値）

全 19 文の話速変化検出正解率 C を図 8に示す。

4.2節で良好な検出結果が得られた例として挙げた図 4、5はグラフ中の $No.1$、$No.2$ のサンプルに対応しており、話速変化検出正解率 Cで評価しても 100%、80%と高い。他にも $No.3$、$No.4$、$No.5$、$No.7$、$No.12$、$No.18$ で 80%を越える高い正解率が得られた。

逆に 4.2節で検出結果の悪い例として挙げた図 6、7はグラフ中の $No.13$、$No.16$ に対応しており、検出正解率が低くなっている。

5 考察

局所的話速検出実験の結果のうち、式（8）による評価基準で高い値を得た音声データは、話速値変化グラフの目視においてもリファレンス側のグラフとよく似た形をしている。話速の極端に速い（あるいは遅い）箇所を捉えて後のシステムに有益な情報を与えるという本システムの目的からは、得られた話速値の絶対的な量よりも話速変化の概形が正しく捉えられることが重要である。図 8の実験結果では正解率の低いデータも含まれておりシステムの精度という点では改善の余地があるが、正解率の高いサン

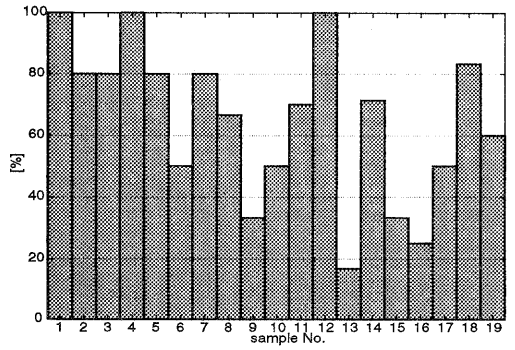

図 8. 話速変化検出正解率 C

プルでは話速変化の増減方向が正しく捉えられている。さらに図5のように、本システムによって検出された話速変化値から予測される「ゆっくり」発話された箇所と実際の聴取とが一致している例もあった。これは本システムが話速変化の特徴を捉える能力を持つ可能性があることを示すと言えよう。今回実験に用いた音声データ全体に対する話速変化検出正解率 Cは、66.4%であった。

検出の精度を低下させている原因として、まずスペクトル変化パラメータだけでは捉えにくい音素の存在がある。/r/、/b/、/d/などは音素検出時に脱落してしまうことが多かった。また、正解率が低かったデータの中でも特に、$No.16$ のサンプルはリファレンスの話速値および聴取によって、他のデータに比べて全体的に話速が速く、調音結合の程度が大きかった。そのためスペクトル変化のピークがはっきりと現れずに、音素境界検出時に捉えられない箇所が多かった。話速値変化の特徴を捉えるのが目的である以上、特に速く発話された箇所に対しても同様の精度で処理を行なえることが望ましい。文全体の発話速度と検出精度の関係については今後の課題である。

6 むすび

本報告では局所的話速変化検出システムによる実環境音声の話速値自動検出実験を行なった。その結果、全体で66.4%の話速変化検出正解率が得られた。今後さらに本システムの有効性を向上するためには、人間が捉えることのできる話速変化を十分に捉えているかどうかという観点が必要と考えられる。このためには、聴取実験との比較を行ない、その結果に基づいてより人間の知覚に近い話速の表現を導入することが必要であろう。また、得られた話速変化値から、特に速い（あるいは遅い）箇所を効果的に抽出するための手法も取り入れる必要がある。さらには、実際に話速変化を考慮した音声認識システムと結び付けることにより、どの程度有効な情報を与えることができるかを調べることも重要な課題である。

謝辞

本研究を行なうにあたり、北海道大学大学院工学研究科電子情報工学専攻情報メディア工学講座信号処理工学分野の荒木健治助教授に多くの御助言をいただきましたことを深謝いたします。

参考文献

[1] 藤崎博也: “ 韻律研究の諸側面とその課題”, 音講論 2-5-11, pp.287-290(1994-11).

[2] 中川聖一:“音声言語情報処理研究の動向と研究課題”, 情報処理学会誌, vol.36, No.11, pp.1012-1019 (1995-11).

[3] 小松昭男, 大平栄二, 市川熹: “ 韻律情報を利用した構文推定およびワードスポッティングによる会話音声理解方式”, 信学論 (D) **J71-D**, 7, pp.1218-1228 (1988-07).

[4] 江口徳博, 尾関和彦:” 韻律情報を利用した係り受け解析”, 音響学会誌 **52**, pp.973-978(1996).

[5] 北澤茂良, 小林聡, 市川英哉:“対話音声の振幅に基づく発話速度の測定”, 音講論 2-P-13, pp.327-328 (1995-09).

[6] 川波弘道, 広瀬啓吉:” 韻律構造を考慮した対話音声の発話速度分析と規則化”, 音講論 1-2-13, pp.199-200(1998-09).

[7] 大野澄雄, 藤崎博也, 高橋浩生:” 日本語音声における強調の韻律的特徴に与える影響について”, 音講論 1-2-14, pp.201-202(1998-09).

[8] 匂坂芳典, 東倉洋一:” 規則による音声合成のための音韻時間長制御”, 信学論 (A) **J67-A**, No.7, pp.629-636(1984-07).

[9] 中井満, シンガー ハラルド, 匂坂芳典, 下平博,”F_0 生成モデルを用いたテンプレートに基づく連続音声の句境界検出”, 信学論 (D-II) **J80-D-II**, pp.2605-2614(1997-10).

[10] 稲葉一紀, 広重真人, 宮永喜一, 栃内香次:” フォルマント軌跡を利用した局所話速変化検出手法”, 1996 電子情報通信学会総合大会講演論文集 D-669, p.290(1995-10).

[11] B.Gold,L.R.Rabiner:”Parallel Processing Techniques for Estimating Pitch Periods of Speech in the Time Domain”, J.Acoust.Soc.Am.,46,2, pp.442-448(1969-08).

[12] 佐々木淳一, 広重真人, 宮永喜一, 栃内香次: “ 局所的話速変化検出のための音韻変化検出システムの複数話者適応に関する検討”, 音講論 2-P-15, pp.323-324(1998-03).