修士論文

話し言葉音声認識における
韻律的特徴を用いた
フィラー検出

平成20年2月4日

指導教員 広瀬 啓吉 教授

東京大学大学院情報理工学系研究科
電子情報学専攻 66405

稲垣貴彦

目 次

第1章　はじめに

従来の音声認識では、音声の持つ音韻的特徴のみが注目され、韻律的特徴は排除されることが多かった。しかし近年は音声認識に韻律を利用する研究がされており、朗読調音声の大語彙連続音声認識について韻律を利用し、認識率の向上を見た研究はいくつかある [1, 2]。しかし朗読調音声はそもそもそれなりに高い認識率が得られるため韻律利用の効果は限定的である。

これを解決する試みとして, 阿部ら [3] は話し言葉の大語彙連続音声認識に韻律的特徴を利用し、特にはっきりした韻律的特徴を持つフィラー [4] を検出することで、認識率を向上させうることを示したが, サンプル数は 100 と少なかった。そこで本論文では、サンプル数を増やし、また韻律モデルと音韻モデルを分離するような手法を検討する。

第 2 章では韻律と音韻の違いと持っている情報について述べる。第 3 章では音声認識の分類と大語彙連続音声認識の理論的背景を述べる。第 4 章では、韻律的特徴を本研究とは異なる側面から音声認識に利用した例を取り上げる。第 5 章では、本研究で使用したコーパス、日本語話し言葉コーパスについて述べる。第 6 章では、実際に韻律的特徴によるフィラー検出器を構成し、音声認識器と組み合わせて音声認識に利用する実験を報告する。

第2章　韻律と音韻

音声の持つ特徴は音韻的特徴と韻律的特徴の二つに分けられる。以下に、特に韻律的特徴に注目して二つの特徴の概要を述べる。

2.1　音韻的特徴

音韻とは言葉を構成する基本単位のことで、音素 (おおよそローマ字表記の個々のローマ字に相当する) のことを指す。そして音韻的特徴とは、音声がどの音素らしいかを示す特徴のことである。音韻的特徴は主に声道 (喉頭より上の部分で、咽頭、鼻腔、口腔などからなり、全体として一つの連続した管を成しているもの。成人では約 15〜17cm の長さを有する) の形を変化させることによって生み出される。

現在の音声認識システムでは、音声を周波数分析したときのスペクトル包絡が音韻的特徴としての用いられている。図 2.1 にその様子の概略を示す。上のグラフは音声波形、下のグラフはその中央部分の対数スペクトルである。太線はスペクトル包絡であり、この形が音韻的特徴として用いられる。

2.2　韻律的特徴

韻律的特徴は主に以下の三つで構成される。

- 基本周波数。声の高さ。声帯の振動周期によって生じる。F_0 とも呼ばれる。
- 音素・休止の継続時間長 (リズム、テンポなどと関連)。
- パワー (音の強さ)。

韻律的特徴は、音韻的特徴に比べて相対的に広い時間領域にわたって生じる現象である。

韻律的特徴のうち F_0 とパワーの変化を図 2.2 に示した。上のグラフがパワーの変化、下のグラフが F_0 の変化である (元の音声波形は図 2.1 と同じである)。F_0 は周期的でない波形には推定できないため、抽出されないことがある。

韻律的特徴によって伝えられる情報は、主に表 2.1 に示したものがあげられる [5]。言語情報とは、辞書・統語・意味・談話のレベルで文字言語によって陽に表現されるか、あるいは文字言語による表記およびその前後の文脈から容易に一義的

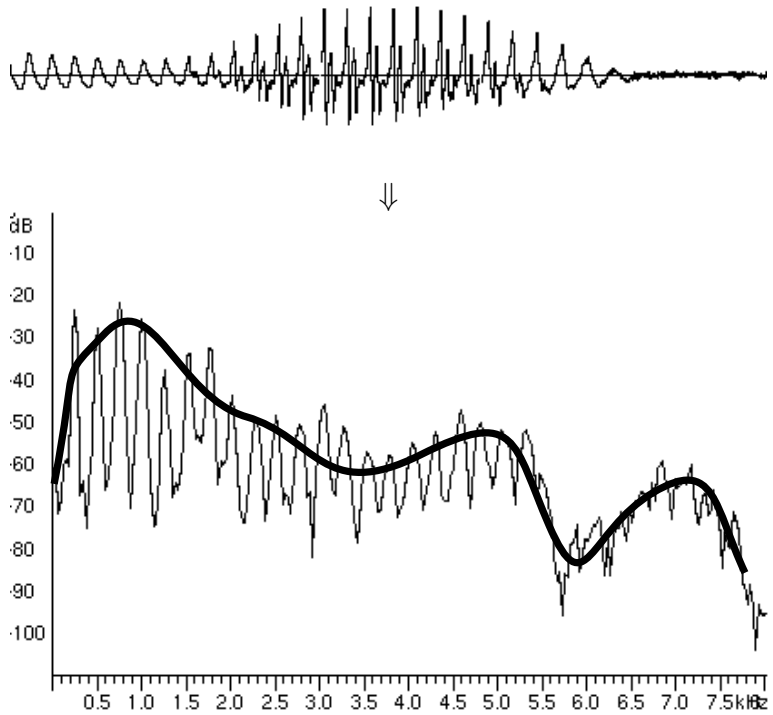

図 2.1: スペクトル包絡の抽出

に導出し得るものを指す。言語情報は基本的に音韻的特徴によって伝達されるが、韻律的特徴もアクセント型などの言語情報を伝達していると考えられる。また言語情報以外の情報も音声は伝達するが、そのうち話者が意識的に制御して伝えることができるものをパラ言語情報という。例えば、断定・疑問・勧誘などの種々の意図、丁寧・くだけた、などの種々の態度、職業などによる特有の発話スタイルなどがあげられる。逆に話者が意図的に制御できないものを非言語情報という。例えば、話者の性別・年齢・身体的状態など個人的な特徴や、感情などの心理状態に関するものなどである。これらパラ言語情報、非言語情報は主に韻律的特徴によって伝達されると考えられる [6]。

表 2.1: 韻律的特徴によって伝えられる情報

種類		内容
言語情報	辞書情報	アクセント型
	統語情報	係り受け
	意味情報	疑問文
	談話情報	話題・焦点、段落
パラ言語情報		意図、態度
非言語情報		個人性、感情

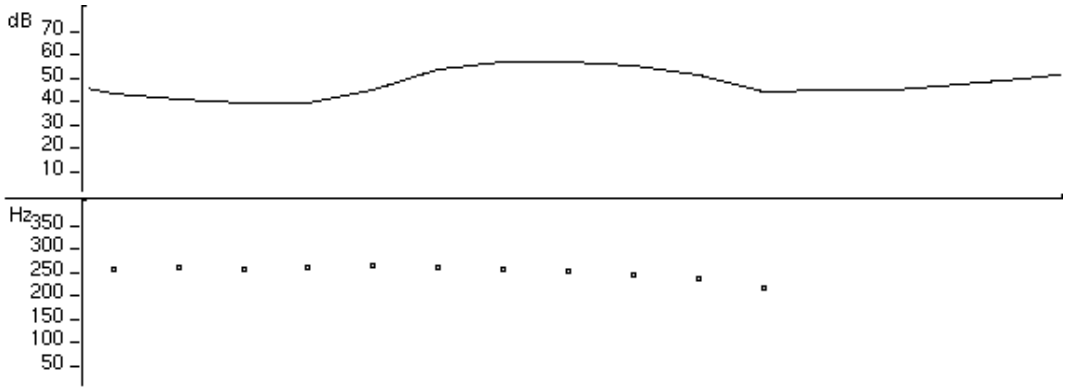

図 2.2: 抽出されたパワーと F_0

2.3 音声分析と特徴パラメータの抽出

現在の音声認識では音韻的特徴すなわち音声のスペクトル包絡を効率よく表現できる特徴パラメータが用いられている。また分析にあたっては、入力音声全体を一度に扱うのではなく、数十 ms 程度のフレームと呼ばれる区間を切り出し、その区間は定常状態にあるとみなして時間窓を掛けて周波数分析する。そして 10ms 程度ごとにフレームをずらして周波数分析を繰り返すことで音声全体を処理する。

スペクトル包絡を表わすのによく用いられているのがケプストラム (spectrum の spec を逆さにした造語) である。ケプストラムは対数スペクトルの逆フーリエ変換として定義される。例えば、図 2.1 の上のグラフの中心を 30ms ほど切り取り、窓をかけ、フーリエ変換して絶対値の対数を取ると下のグラフの細線部分すなわち対数スペクトルが得られる。しかし対数スペクトルは太線で示した包絡成分だけでなく微細構造 (周期的な繰り返し) を含んでいる。ここから音韻的特徴すなわち包絡成分のみを抽出するのがケプストラム分析である。ケプストラムは対数スペクトルを逆フーリエ変換、すなわち周波数軸を時間に見立てて再度周波数分析したものであり、ケプストラムの低次の項はスペクトルの包絡成分に、ケプストラムの高次の項はスペクトルの音源波形に対応する。

そこで音声の特徴パラメータとしては、ケプストラムの低次の項 (第 10 次程度まで) とパワー、また時間差分である Δ ケプストラムや Δ パワーがよく用いられる。

以上のようなフレーム単位の分析を、シフトしながら波形全体に対して施すことで、特徴パラメータ列 $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_t)$ が得られ、これが認識に用いられる。

2.4 まとめ

言語音声の持つ特徴には音韻的特徴と韻律的特徴がある。音韻的特徴は声道の形によって生じる。また数十ミリセカンドという比較的短時間に存在し、通常の音声認識で音素を区別するために使われる。言語情報を主に伝えているのは音韻

的特徴である。韻律的特徴は比較的長い時間に渡って生じる。韻律的特徴はアクセントや係り受けなどの一部の言語情報と、意図や態度などのパラ言語情報、個人性などの非言語情報を伝える。

第3章　大語彙連続音声認識の概要

本研究では最終的に大語彙連続音声認識器に韻律モジュールを組み込んで駆動する予定である。以下に大語彙連続音声認識の概要を述べる。

3.1　音声認識技術の分類

音声認識技術は対象とする発話の種類によって幾つかに分類される。ここでは話者による分類と発声による分類を取り上げる。

3.1.1　話者による分類

特定話者音声認識

特定の話者の音声を認識するための技術で、事前に話者の音声を学習させる必要がある。

不特定話者音声認識

事前の学習を必要とせず、不特定の話者の音声を認識する技術である。

話者適応

上記二つの中間的なもので、最初は不特定話者認識だが話者に合わせてパラメータを修正して話者に適応していく技術である。

3.1.2　発声による分類

孤立単語音声認識

発声の前後にポーズを設けて発声された単語を認識する技術である。入力された音声データを予め登録してある単語モデルと比較して最も似ているものを認識結果とする。

連続音声認識

文章として発話された音声を認識する技術である。単語の境界が明確ではなく、音が隣接した単語の影響を受けることを考慮しなければならない。

3.2 認識の基本的な枠組み

大語彙連続音声認識とは、特徴パラメータの時系列 $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_t)$ が与えられたときに、単語列 $\boldsymbol{W} = (\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_i)$ が発声された確率 $P(\boldsymbol{W}|\boldsymbol{X})$ が最も高くなるような $\hat{\boldsymbol{W}}$ を探すことである。式で表わすと次のようになる。

$$\hat{\boldsymbol{W}} = \underset{\boldsymbol{W}}{\operatorname{argmax}} P(\boldsymbol{W}|\boldsymbol{X})$$

$P(\boldsymbol{W}|\boldsymbol{X})$ はベイズの定理を用いて次のように書き直せる。

$$P(\boldsymbol{W}|\boldsymbol{X}) = \frac{P(\boldsymbol{X}|\boldsymbol{W}) \cdot P(\boldsymbol{W})}{P(\boldsymbol{X})}$$

ここで $P(\boldsymbol{X})$ は入力パターン自体の生起確率であり、同じ入力波形については一定であるから、統計的音声認識は次の式を解く問題となる。

$$\hat{\boldsymbol{W}} = \underset{\boldsymbol{W}}{\operatorname{argmax}} P(\boldsymbol{X}|\boldsymbol{W}) \cdot P(\boldsymbol{W})$$

$P(\boldsymbol{X}|\boldsymbol{W})$ はある単語列 $\boldsymbol{W}$ が発声されたきに特徴パラメータの時系列として $\boldsymbol{X}$ が観測される確率を表わす。このようにある単語列 $\boldsymbol{W}$ の音響的特徴を記述する統計的モデルを音響モデルという。実際の処理の例として、モノフォンないしトライフォンによって確率を計算する場合、並んだ音素モデルの確率の積として観測確率を定義する。

また $P(\boldsymbol{W})$ は音声が何も観測されていない時点での単語列 $\boldsymbol{W}$ の出現確率である。これは言語的な性質だけによる統計的モデルであり言語モデルと呼ばれる。実際の処理では、N グラムが使用されることが多い。

このように、モデルは音響的な性質と言語的な性質に分割してはいるが、確率という共通の尺度で 2 つのモデルが結合されているため、認識においてはこれらのモデルを統一的に扱うことができる。つまり、特徴パラメータ列をまず音素系列に変換して、それから単語列に変換するというような処理ではなく、これらを同じ枠組みで同時に実行できる。このような音声認識システムの流れを図 3.1 に図示した。図中の Search において、両モデルを用いて様々な単語列仮説の確率値が計算され、最も確からしい単語列が最終的に出力される。

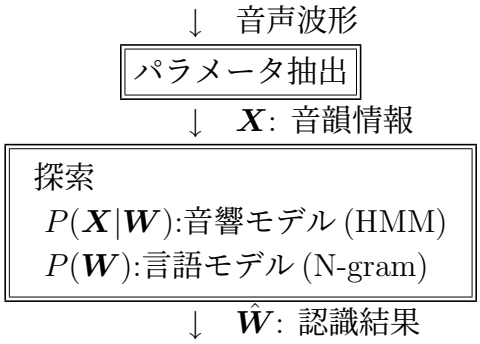

図 3.1: 音声認識システム

第4章　韻律的特徴の音声認識への利用

この章では韻律的特徴を音声認識に利用した事例として、韻律的特徴からアクセント句境界を推定して探索を高速化した例を取り上げる。これは表2.1で言うところの統語情報を抽出していると考えることができる。

4.1　探索空間の軽減と音響モデルの高精度化

[7] では、アクセント句 (F_0 パターンの谷間に相当する韻律イベント) 境界を抽出して、それを認識における探索空間の軽減と音響モデルの高精度化に利用する手法を提案している。

大語彙連続音声認識では、一般的に探索空間の増大を防ぐために各フレームごとにスコアが探索ビーム幅以内に収まった仮説だけを対象に探索を進める。[7] では、アクセント句境界情報に基づいてビーム幅を動的に制御することで探索空間の軽減を実現している。図4.1は、「このところずっと考えていた」という文を認識するときの正解経路のスコアとビーム幅以内に残った仮説の最低スコアの変動を示したものである。横軸は時刻、縦軸はそのフレームにおけるスコアである。また、点が正解経路のスコア、実線が最低スコアを示している。これを見ると、正解経路のスコアは単語の境界にあたる時刻で一時的に落ち込んでいるのが分かる。これは、単語が遷移するときに言語モデルの尤度が加えられるためである。したがってこのときに正解経路がビーム幅の外に出る危険性が高くなる。かといって十分なビーム幅を用意しておき、認識過程の間ずっと固定しておくのでは無駄も多い。

そこで [7] では、あらかじめ求めておいたアクセント句境界情報を元に、句境界をまたぐ場合にはビーム幅を大きくし、句内ではビーム幅を徐々に小さくしていくというビーム幅の動的な制御を行っている。これによって効率的な枝刈りができる。10名の話者がそれぞれ5文ずつ発声した計50文の新聞記事読み上げ音声の認識において、ビーム幅の動的制御によって単語正解精度86%を維持したまま、探索空間を約51%、探索時間を約30%軽減できたと報告されている。

また、アクセント句境界を利用した音響モデルの高精度化も検討されている。調音結合に対処するため、精密な音響モデルは音素文脈を考慮して構築される。し

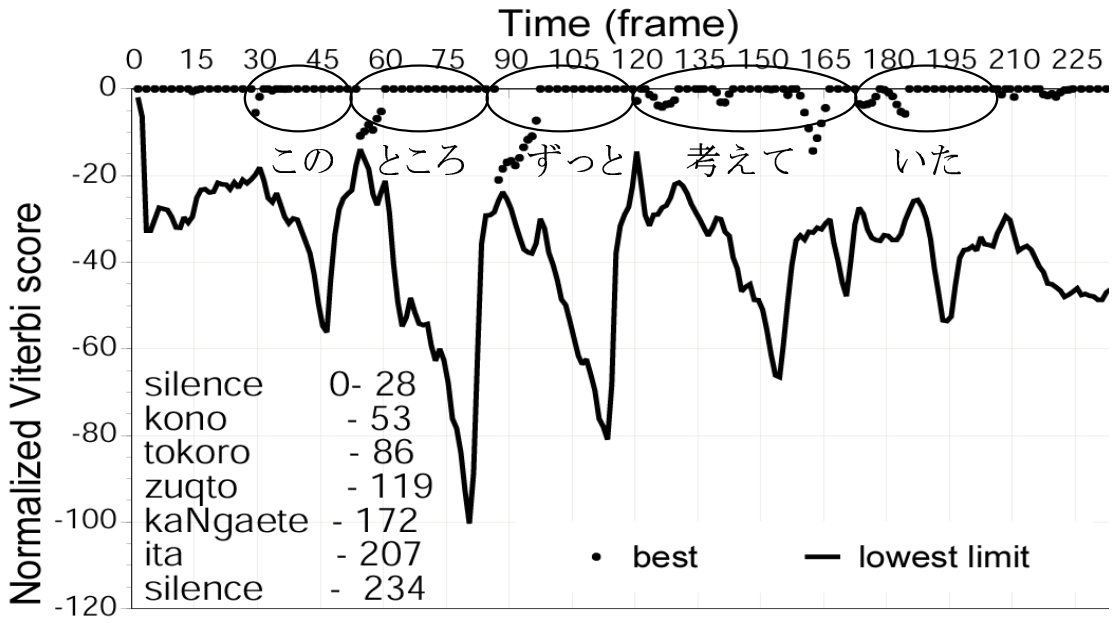

図 4.1: 探索ビーム内における仮説の最低スコアと正解仮説のスコアの変動

かし、アクセント句境界が存在する場合、そこで発声の一区切りがあると考えられ、句内部での発声ほど調音結合の影響を受けないことが予想される。そこで、音素文脈を考慮した音響モデルは句内部で使用し、句境界では音素文脈を考慮しないモデルを用いる手法を提案している。これによって、単語正解精度が 90.1%から 91.3%へと約 1%改善されている。

第5章　日本語話し言葉コーパス

本研究で使用したコーパスは日本語話し言葉コーパス (the Corpus of Spontaneous Japanese) である [8]。CSJ は、現代日本語の自発音声 (主に学会講演など) を大量に集めて多くの研究用情報を付加した話し言葉研究用のデータベースであり、語数にして約 750 万語、時間にして約 660 時間の音声が含まれている。CSJ は以下のような従来の音声データベースにない多くの特徴を有している。

- 多数の話者による多少とも自発的な音声を対象としていること
- 豊富な研究用付加情報を提供していること
- 発話スタイルないし自発性に対する評価を与えていること
- XML 文書化されたデータも公開していること

このコーパスは独立行政法人国立国語研究所と独立行政法人通信総合研究所が推進してきている文科省科学技術振興調整費開放的融合研究制度研究課題「話し言葉の言語的・パラ言語的構造の解析に基づく『話し言葉工学』の構築」プロジェクトの一環として構築されている。このプロジェクトの目標は自然な話し言葉を工学的に処理するための基盤技術を開拓することにおかれているが、CSJ はそのために必要不可欠なデータベースとして位置づけられており、その構築作業は主として国立国語研究所が分担している。CSJ は 2004 年 6 月に公開された。

5.1　音響モデルと言語モデル

CSJ の中には、CSJ で学習された音声認識用の音響モデルと言語モデルが含まれている。音響モデルは時間にして 486 時間分の 2496 講演から、言語モデルは語彙にして 6.6M を含む 2592 講演から学習されている。音素体系は表 5.1 に示す 42 種類である。ここで、q は促音に伴う無音、sp は音声中の短い無音である。N は撥音、a:〜o:は長母音を表す。silB は発話の先頭の無音、silE は発話の終端の無音であり、発話は基本的に 500ms 以上の無音区間で区切ったものと定義している。形態素は国立国語研究所で定義された短単位に基づいている。

表 5.1: CSJ で定義されている音素

母音	a i u e o a: i: u: e: o:
子音	N w y j my ky by gy ny hy ry py
	p t k ts ch b d g z m n s sh h f r
	q sp silB silE

5.2 F タグ

CSJ では、付加的な情報を示すため様々なタグが書き起こしテキストに付与されている。これらのタグは多人数で作業をするためにマニュアルに従って一定の基準で付けられた。ここでは本研究に関係する F タグ (フィラータグ) について述べる [9]。

F タグは感動詞のうち、以下のものに付与された。ただし応答表現は対話においてのみ付与され、CSJ の大半を占める独話には付与されてない[1]。

フィラー 「あの」「そのー」「えっとー」「うんとー」「あー」「んー」「あのですね」など (全講演対象)

感情表出系感動詞 「あ (っ)」「あら (っ)」「うわわ (っ)」「おー」「わー」など (全講演対象)

応答表現 「はい」「ええ」「うん」「ああ」「おお」「ううん」「いえ」「いいえ」「いや」など (対話のみ)

フィラーや感情表出系感動詞に F タグを付与するのは、本来、転記テキストの可読性の問題と、自動形態素解析の精度の問題からである。これらの言葉には「あ」「え」「ん」など短い表現が多く、特にフィラーは「法則式イーを」や「多次元項目エー反応モデル」のように文節や単語の途中に発話されることも少なくない。そこで転記テキストの可読性と自動解析の精度を改善するために F タグを付与されたのである。

フィラーと感情表出系感動詞に同じタグを付与するのは、実際の作業において、両者を区別することのできない場合が少なくないためである。例えば「これがアー解答なんだと思って」といった場合の「アー」は、音調によっては、場を繋ぐためのフィラーなのか、簡単を表現する感動詞なのかを区別することが難しいことがある。

5.2.1 フィラー

フィラーとは、言い淀み時などに出現する場繋ぎ的な表現のことである。

[1] よってこの発表では触れない

表 5.2: フィラーの語彙

基本表現
あ (ー)、い (ー)、う (ー)、え (ー)、お (ー)、 ん (ー)、と (ー)*、ま (ー)*、う (ー) ん、 あ (ー)(ん)(ー) の (ー)*、そ (ー)(ん)(ー) の (ー)*、 う (ー) ん (ー)(っ) と (ー)*、あ (ー)(っ) と (ー)*、 え (ー)(っ) と (ー)*、ん (ー)(っ) と (ー)*
組み合わせ
基本表現 +「ですね (ー)」「っすね (ー)」 *印の基本表現 +「ね (ー)」「さ (ー)」

例:これは そのー 重要な問題なので あのー 今回の おー 議論でも ん 大きく取り上げたいと思います。

CSJ ではフィラーの語彙を表 5.2 のように限定し、しかも語が場繋ぎ機能を有する場合のみ F タグを付与している。

フィラーか否かで迷う場合については以下のように定められた。

フィラーか連体詞かで迷う場合 「あの (ー)」「その (ー)」についてはフィラーか連体詞かで迷うことが多い。文脈や音調から判断が付かない場合には、迷った旨をコメントに記した上で、F タグを付ける。

フィラーが語断片かで迷う場合 「あ」「え」「ん」などの短い音は、フィラーか語断片かで迷うことが多い。そこで以下の操作的な判断基準に従って決定された。

1. 母音の引き伸ばしがあればフィラーである。
2. フィラーが後続し、かつそのフィラーの冒頭の音と同じであればフィラーである。
3. 後続する内容語あるいは語断片の冒頭の音と同じであれば語断片である。
4. 問題の表現の直後にポーズがあればフィラーである。
5. 以上の条件に該当しなければ語断片である。

フィラーが接続詞・接続助詞・格助詞かで迷う場合 「と」については、フィラーか、接続詞・接続助詞・格助詞かで迷うことが多い。「X(前文脈) と Y(後文脈)」とあった場合に、X と Y の関係が以下の条件のいずれも満たさない場合に限ってフィラーとみなす。

- X と Y が並立関係にある場合 (「それと」に置換可能な場合)

- X と Y に因果関係がある場合 (「すると」に置換可能な場合)
- X を引用的に受けて Y に続くと考えられる場合

フィラーが副詞かで迷う場合 実際の発話において、音調や文脈などからフィラーの「まー/まあ」と、副詞の「まあ」とを区別することは極めて難しい。そこで全て F タグを付与し、表記は「まー」と統一した。しかし「まあまあ」は以下のいずれかの条件を満たす場合は副詞とした。

- 韻律条件: 後者の「マー」のみにアクセント核がある場合
- 文法条件: 助詞・助動詞が後続するか、用言・副詞に係り程度を表わす場合

5.2.2 感情表出系感動詞

感情表出系感動詞とは、驚いたときや落胆したときなどに発する感動詞である。

例:あ、あっ、あー、あーあ、あら、ありゃ、うっ、うーむ、うわ、え、えー、お、おー、おや、げっ、へっ、ほー、わっ

フィラーの場合とは異なり、語は限定されていない。

5.3 まとめ

CSJ で定められたフィラーの認定基準について述べた。この基準ではフィラーの語彙と文脈が限定されており、韻律的にはフィラーであるような語であっても必ずしも F タグが付いているとは限らないことに注意が必要である。

5.4 予備実験:F タグの検査

本研究は CSJ の質に依存している。書き起しテキストとそのタグ付けに誤りが多ければ研究は根底から覆る。CSJ がどの程度信頼できるものなのか、あるいはどの程度の誤りが含まれているかを把握しておかなければならない。そこで予備実験として書き起こしテキストに付けられた F タグの正確さを検証した。

表 5.3: F タグの検証結果

	誤り		フィラーらしい語			総数 (参考)		
講演 ID	F タグ	書起し	コノ	コー	デ	短単位	フィラー	発話
A01F0122	0	1	0	0	0	5278	41	268
A01F0132	0	1	0	0	0	5010	218	134
A01M0007	0	0	3	9	13	11476	341	792
A01M0048	0	0	0	0	0	5621	117	301
A03M0004	0	0	0	0	0	6558	220	372
A04M0047	0	0	0	0	1	7856	206	382

5.4.1 方法

層別標本法によって標本調査を行なった。

コアに含まれる A 講演[2]は 70 講演あり、そのうち 46 講演が男性によるもので、24 講演が女性によるものである。ここから男性 4 人、女性 2 人の講演を無作為に抽出して調査した。

本研究では発話 (転記) 単位以外のコーパスの時間情報を使用しないので、F タグの時間情報がずれているといったことは問題にならない。それで調査方法と内容は以下のようにした。

1. F タグが付いた部分の書き起こしテキストとその位置を表示する。
2. 音声を再生し、フィラーがあること、他には無いことを確認する。問題があれば記録する。

5.4.2 調査結果

調査の結果は表 5.3 のようになった。F タグに誤りは見られなかった。ただし、書き起こし誤りと見られる部分が 2 箇所あった。また、誤りではないが、韻律的にはフィラーであるが語彙条件 (表 5.2) によって F タグが付いていないと思われる短単位[3]が見られた。内訳は「コノ」が 4 回、「コー」が 9 回、「デ」が 14 回の全 27 回であり、「デ」1 回を除いて全て一人の話者に集中していた。

5.4.3 まとめ

F タグの誤りは見られず、基準どおりに付与されていることが確認できた。また一部の話者が CSJ の条件にない語彙をフィラーに使っていることが確認された。

[2]講演の分野ごとに記号が振られている
[3]複合語を一つの単語ではなく複数の単語とみなすときの、単語

しかしそうした語はフィラー全体に対して相対的に少なく、CSJ においては、F タグの付いた短単位がフィラーでありそれ以外のフィラーはないとみなしても害はないと考えられる。

第6章　韻律によるフィラー検出を組み込んだ音声認識

韻律的特徴によるフィラー検出器を学習ベクトル量子化 (以下 LVQ、自己組織化マップ=SOM の仲間) によって構成し、その性能を実験で評価した。さらに大語彙連続音声認識器にそのフィラー検出器を組み込んで性能を評価した。

6.1　韻律によるフィラー検出実験

韻律によるフィラー検出器を LVQ によって構成し、その単体での検出性能を評価した。以下に使用した韻律的特徴や各種条件を述べる。

6.1.1　フィラー検出器の構成

韻律によるフィラー検出器の構成を図 6.1 に図示した。フィラー検出器の入力は、音声の基本周波数とパワー、および単語アライメント情報である。コーパスに含まれる手作業で付けられたアライメントではなく、書き起こしテキストから Julian によって機械的に作成したアライメントを用いた。これは、未知の音声入力からアライメントをを取る際には Julius を用いてアライメントを取得するからである。出力は各単語がフィラーであるか否かを表わすカテゴリのベクトルである。

今回、各単語の韻律パラメータとして用いた情報は表 6.1 のとおりである。 F_0 およびパワーの推定には Snack[10] の pitch 関数 (いわゆる `get_f0` 関数) を用いた。

検出器の本体には二次元の LVQ を使用した。LVQ はニューラルネットの一種であり、教師ありの競合学習によって入力データをパターン分類することができる。LVQ が学習する様子を図 6.2 に図示した。LVQ では隣り合ったノードが “接続” されている。そして、入力ベクトル v に最も近い重みベクトル w を持つノード (図の黒いノード) とその “近隣” の (点線の円の中の) ノードが w を修正して v に近づける ($w_{\mathrm{new}} = (1-\alpha)w + \alpha v$) ことで学習を進める。学習が進むにつれて、修正される “近隣” の半径は狭く、修正量 (α) は 0 に近くされる。今回は入力ベクトル 300 個ごとにこれらのパラメータの修正を行なった。

今回用意した LVQ は、ノード数が 10×10 から 16×16 まで、およびそれぞれについて学習回数が 100×300 から 900×300 回まで (200×300 回刻み) のもの、

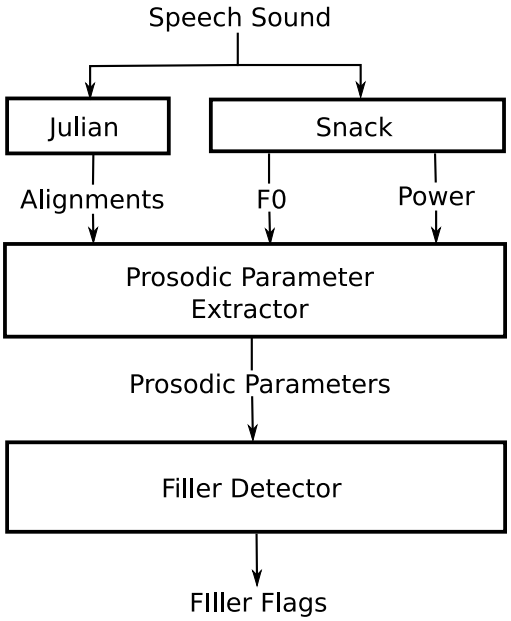

図 6.1: 韻律によるフィラー検出器

表 6.1: フィラー検出に使用した韻律パラメータ

音素数
F_0 の最大・最小の差
F_0 を一次直線で近似したときの傾き
(母音の平均 F_0)/(発話全体の母音の平均 F_0)
最終母音と後続単語の最初の母音の F_0 の差
パワーを一次直線で近似したときの傾き
母音のパワーの平均
最後の母音音素の持続時間
先行するパワーの小さい区間の長さ
後続するパワーの小さい区間の長さ

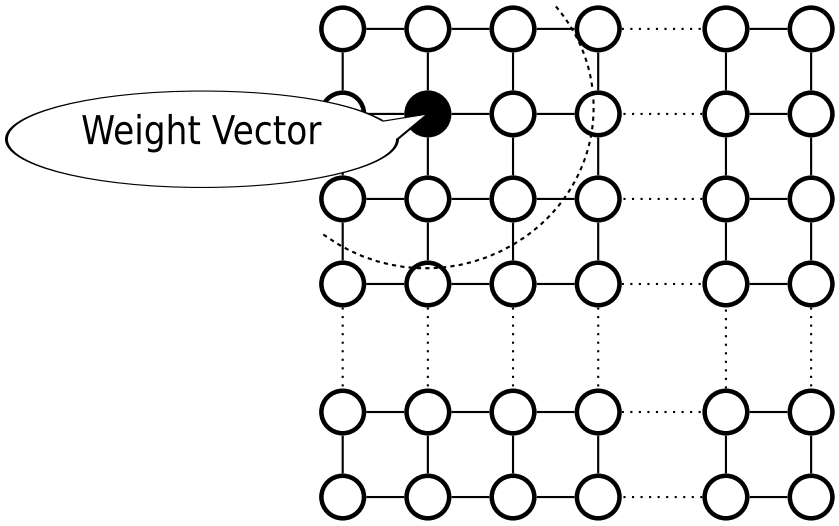

図 6.2: LVQ の学習

表 6.2: 学習と評価に使用した講演の内容

	講演	発話	フィラー	単語
学習用	1	334	167	2474
評価用	176	53450	30011	440039

合計 35 種類である。

6.1.2 評価実験に使用したデータ

検出器の学習には CSJ のコアに含まれる A01F0055 講演を使用した。 評価には CSJ のコアに含まれる A 講演と S 講演のうち、合計 176 講演を使用した。ただし、プライバシー保護のためノイズで音声を消されている発話、および F_0 が推定できないなどの理由で韻律パラメータを抽出できなかった発話は学習には使用していない。表 6.2 に学習用・評価用データの講演の数とそこに含まれる発話・単語およびフィラー単語の数を示す。 学習に使う教師データとしては、単語に F タグが付与されていれば 1、なければ 0 を与えた。

6.1.3 実験結果

学習回数-ノード数-フィラー再現率のグラフを図 6.3 に示す。フィラー再現率は、フィラー単語の韻律パラメータ入力に対して、フィラー検出器が 1 を出力した率である。再現率は 300×300 回ないし 500×300 回の学習で再現率は最も高くなった。ノード数は再現率には大きく影響しなかった。最も再現率の高かったのは 500×300

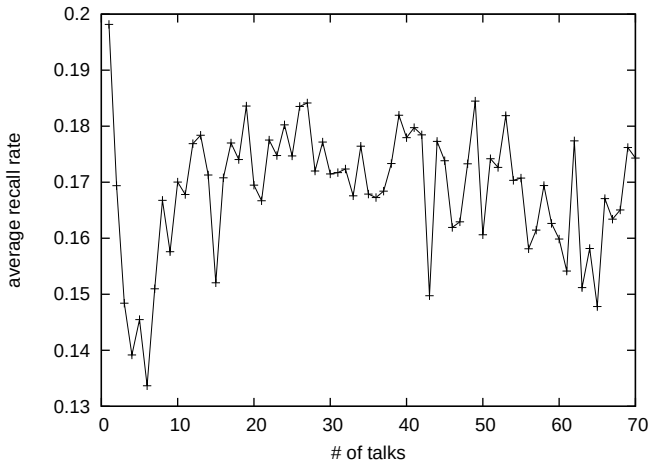

図 6.3: 学習回数-ノード数-再現率

回学習した 10×10 ノードの LVQ であり、フィラーの再現率は 53.5%、非フィラーの再現率は 96.3%であった。

6.1.4 考察

フィラーの正解率が非フィラーに比べて低いのは、学習データにフィラーでない単語が多過ぎるからであろう。さらなる性能向上を求めるなら、学習データに含まれるフィラー/非フィラー単語の比率を変えて実験をするべきである。それと合わせて、必要な学習データの数についても実験的に検討する必要がある。 しかしながら実際に音声認識器に組み込むときは、誤った単語アライメントを与えられる可能性があるため、正しい単語アライメントのみでのテストには限界があると考えてそこまでの追求はしないことにした。

6.2 韻律を利用したフィラー検出器を組み込んだ音声認識

6.2.1 フィラー検出と利用の仕組み

製作したフィラー検出器を音声認識器と組み合わせて音声認識を行なった。その構成を図 6.4 に図示した。フィラー検出器に入力されるアライメントは、6 節では Julian を用いて書き起こしテキストから推定されたが、ここでは Julius(1) によって推定される。認識対象の音声の書き起こしテキストは事前に得られるものではないからである。フィラー検出器の出力は、単語ごとのカテゴリではなく、フィラーである (らしい) 単語候補の存在する時刻 (フィラー時刻) である。Julius(2) は、得

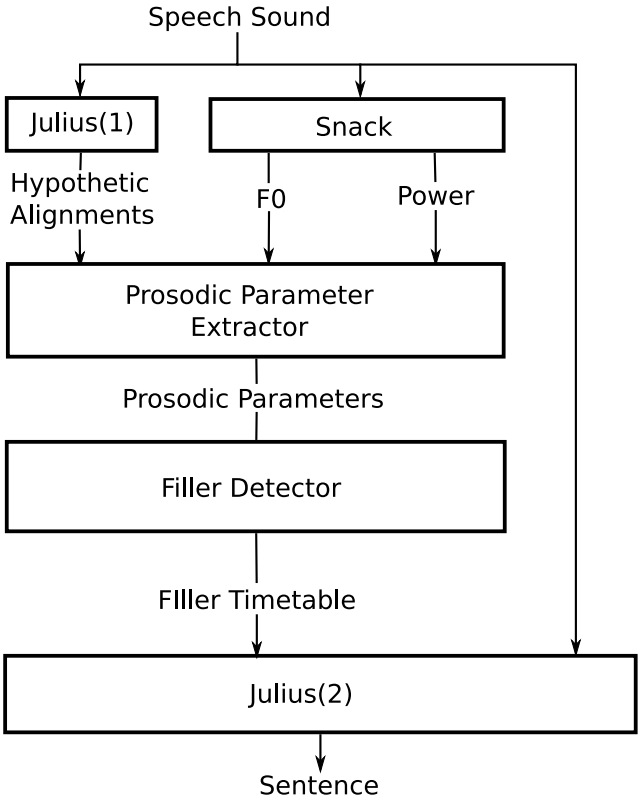

図 6.4: フィラー検出器と組み合わせた音声認識器

表 6.3: 正解した文の数とその割合

	baseline	proposed	total
w/ filler	4265	4358	22984
ratio	0.1856	0.1896	
w/o filler	7933	8258	30466
ratio	0.2604	0.2711	

られたフィラー時刻をまたいで存在する仮説単語のスコアを適宜修正しながら探索を進める。

使用したフィラー検出器は 6.1.3 で作成した最もフィラー再現率の高いものである。再現率はフィラー単語について 53.54%, 非フィラー単語について 96.38%であった。適合率はフィラー単語について 53.36%, 非フィラー単語について 96.40%であった。

6.2.2 認識実験

Julius v3.5.3(fast) とフィラー検出器を組み合わせて認識実験を行なった。認識実験には CSJ のコアに含まれる A 講演と S 講演を使用した。使用したデータは 53450 発話, 440576 単語である。

検出器 (SOM) の出力と単語仮説の持つ属性 (フィラーであるか否か) との組み合わせによって, フィラー時刻およびスコア修正にはいくつかの方法が考えられる。今回は, 仮説においてフィラー単語であり, かつ検出器もフィラーであると判定した単語のある時刻をフィラー時刻とした。また Julius の第一パスで生成された単語仮説が, フィラー時刻にまたがって存在し, かつフィラーであるときその単語仮説に 5 ポイントのスコアを加算した。

認識結果

全文一致したもののみを正解としたとき, フィラーのある文 (発話) の認識で正解率は 0.1856 から 0.1896 へ改善した。またフィラーのない文の認識で正解率は 0.2604 から 0.2711 へ改善した (表 6.3)。

また文ごとに単語正解率を計算すると表 6.4 のようになった。単語正解率はフィラーのある文において 0.6399 から 0.6417 に改善した。フィラーのない文では 0.6454 から 0.6511 へ改善した。なお単語正解率は, 正解の書き起こしテキストと認識結果の編集距離を d, 最大の編集距離を m とするとき, $1 - d/m$ と定義した。

認識結果の改善した文を見ると, フィラーが出やすくなった他, 短い単語や助詞がよく挿入されていた (表 6.5)。

表 6.4: 文ごとの単語正解率の平均

	baseline	proposed
w/ filler	0.6399	0.6417
w/o filler	0.6454	0.6511

表 6.5: 認識例

transcr.	えこの周波数差が……
proposed	えーこの小数差が……
baseline	猫の小数差が……
transcr.	音源の方向を振り向く
proposed	音源の方向を振り向く
baseline	音源方向を振り向く

またフィラーとは関係ない部分で認識結果が悪くなっている部分もあった (表 6.6)。

考察

二通りの影響があったと考えられる。まず, 認識できなかったフィラーを期待どおり認識できるようになった部分である。

次に, フィラー単語の仮説が残りやすくなったことで探索から漏れる単語が変化し, 認識結果に影響を与えた部分である。特に認識結果にフィラーのない文であっても, 例えば助詞が出やすくなったなどの認識結果が変わっているものは, この影響を受けたと考えるべきであろう。これを抑えるには探索のビーム幅を広くし枝刈りを減らす必要があるだろう。しかしフィラーを含まない文についても全体として認識率が改善していることから, 認識結果に出ないフィラー単語仮説が探索の中で文節ないし意味的な区切りを示唆していた可能性がある。

表 6.6: 副作用例

transcr.	えー下の
proposed	えー舌の
baseline	えー下の
transcr.	もう凄いマイクを
proposed	もう凄いうまいこう
baseline	もう凄いマイクを

6.2.3 まとめ

韻律的特徴からフィラーを検出するLVQを教師あり学習によって作成した。非フィラーの再現率は96.3%となったが、フィラーの再現率は53.5%にとどまった。

その検出器を大語彙連続音声認識器とJuliusに組み合わせて、韻律によるフィラーモデル付き音声認識を実験した。フィラーを正しく認識する部分は増えたが、一方でフィラーの関係ない部分でも探索に影響を与えていると考えられる。

今後の発展として、使用する韻律パラメータを変える、アライメント情報に依存しないフィラー検出を実現する、フィラー以外の単語属性を検出するという展開が考えられる。

第7章　まとめ

本論文では、まず音声には韻律と音韻の二つの面があることと、それら二面の持つ情報について述べた。音韻的特徴は声道によって生み出され、言語を伝える主な手段となっている。韻律的特徴は一部の言語的特徴と多くのパラ言語・非言語的情報を伝える。次に音声認識、特に大語彙連続音声認識の枠組みを概説した。大語彙音声認識では、音響モデルと言語モデルは確率という一つの基準によって結合されている。次に韻律的特徴の持つ統語情報という側面を大語彙連続音声認識に利用した例を取り上げた。次に本研究で用いるコーパスの概略を述べ、さらに本研究で最も重要なF タグが正しく付与されているかを検証した。最後に、韻律的特徴を利用したフィラー検出器を実際に製作し、話し言葉の大語彙連続音声認識に利用して、認識結果が改善することを確認した。

参考文献

[1] K. Hirose and K. Iwano, “Detection of prosodic word boundaries by statistical modeling of mora transitions of fundamental frequency contours and its use for continuous speech recognition”, *Proc. IEEE ICASSP*, pp. 1763-1766, 2000.

[2] K. Hirose and N. Minematsu, “Use of prosodic features for speech recognition”, *Proc. ICSLP*, pp. 1445-1448, 2004.

[3] 阿部悠, 広瀬啓吉, 峯松信明, “音声認識時の韻律利用によるフィラー検出”, 日本音響学会講演論文集, pp. 1213-1214, 2006.

[4] Felix C. M. Quimbo, Tatsuya Kawahara, Shuji Doshita, “Prosodic analysis of fillers and self-repair in Japapnese speech”, *Proc. ICSLP*, pp. 3313-3316, 1998.

[5] 広瀬啓吉, “韻律情報の処理”, *Journal of Signal Processing*, Vol.2, No.6, pp.415-423, 1998.

[6] 藤崎博也, “韻律研究の諸側面とその課題”, 日本音響学会講演論文集, pp.287-290, 1994.

[7] S. Lee, K. Hirose and N. Minematsu, “Incorporation of prosodic module for large vocabulary continuous speech recognition”, *Proc. ISCA Tutorial and Research Workshop on: Prosody in Speech Recognition and Understanding*, pp.97-101, 2001.

[8] K. Maekawa, “Corpus of Spontaneous Japanese: Its design and evaluation”, *Proc. ISCA and IEEE workshop on Spontaneous Speech Processing and Recognition*, pp.7-12, 2003.
`http://www2.kokken.go.jp/~csj/public/index.html`

[9] 国立国語研究所, “日本語話し言葉コーパスの構築法”, 国立国語研究所, pp. 82-91, 2006.

[10] `http://www.speech.kth.se/snack/`

[11] `http://www-ra.informatik.uni-tuebingen.de/SNNS/`

[12] “大語彙連続音声認識エンジン Julius”, http://julius.sourceforge.jp/