2004－NL－159　(16)
2004／1／14

自然発話中の冗長語区間検出に関する研究

丹羽　啓二　　佐川　雄二　　杉江　昇
名城大学大学院理工学研究科
〒468-8502 名古屋市天白区塩釜口１丁目 501 番地
E-mail：c302j016@ccmailg.meijo-u.ac.jp,sagawa@ccmfs.meijo-u.ac.jp,sugie@ccmfs.meijo-u.ac.jp

あらまし　自然な発話には，話者の意図とするものではない，文法的・意味的・文脈的な不適格性が生じる．発話中に不適格性が不可避なものとして現れる理由として，思考プロセスが発話プロセスに追いつかない場合が現れ，次の発話内容が発話プロセスに届く間，話者は時間を稼ぐため言い淀みを用いると考えられており，この不適格性を含む音声は音声認識システムによって音声認識を行う際の誤認識を起こす原因となる事が報告されている．本研究は，不適格性を含む音声を正しく認識する為に，前処理として不適格性の一つである「あー，えー」といった冗長語区間の延伸母音を検出・除去する手法についての提案を行い，その検出方法について評価実験を行ったので報告する．

キーワード　　冗長語，延伸母音，言い淀み，有声休止

DETECTING PROLONGED VOWEL IN SPONTANEOUS SPEECH

Keiji NIWA　　Yuji SAGAWA　　Noboru SUGIE

Graduate School of Science & Technology, Meijo University
1-501 Shiogamaguchi , Tempaku-ku, Nagoya, 468-8502, Japan
E-mail：c302j016@ccmailg.meijo-u.ac.jp , sagawa@ccmfs.meijo-u.ac.jp ,sugie@ccmfs.meijo-u.ac.jp

Abstract　Spontaneous speech includes grammatical, semantical and contextual disfluency which speaker does not intend. One of the reason why disfluency is inevitable in spontaneous speech, is that the speaker uses ill-formed-ness while the case where the idea process does not catch up with the utterance process, so that the content of the following utterableness contents reaches the utterance process to get time. It was reported that voice including disfluency cause recognition error in speech recognition system. To recognize the voice including disfluency correctly, this paper proposes the technique for detecting and removing one kind of disfluency (filled pause) including the prolonged vowel, such as “a-, e-” in Japanese, and reports result of evaluation experiment of the method.

KEY WORDS　prolonged vowel, disfluency, filled pause

1.はじめに

本研究では音声認識システムによる自然な発話の書き起こしを行う事を最終目的としている．その為には話者がその場で内容を考えながら発話した音声を，計算機が理解できる事が必要となる．しかし，自然な発話には，話者の意図とするものではない，発話内での文法的・意味的・文脈的誤りである不適格性が頻繁に発生する．この不適格性を含む音声は，音声認識システムによって書き起こしを行う際の誤認識を起こす原因となる事が報告されている[1][2]．これは通常の音声認識システムは言い淀みのない，書き言葉のような音声文章を基に設計されている為で，この不適格性によって普段我々が発話する音声は通常の音声認識システムでは正確に認識する事は困難とされている．またこれら不適格性は発言権の保持や話者の心理状態・思考状態を表す大切な役割を果たしていることもあり，計算機との対話を行う際の重要な因子である．このような事から不適格性を音声認識システムで取り扱う方法として，音声・言語モデルに頻度の多い言い淀みを含ませ認識精度の向上をさせる方法や比較的長時間の言い淀み部分をリアルタイムに検出し，音素列のアライメントを調整する手法などがある．しかし，音声を書き起こした場合，研究・調査に利用する場合等の例外はあるが，通常では不適格性は書き起こしても不必要な言葉であり，文字情報としては不要なものとみなすことができる．そこで本研究では，この自然な発話内で不可避な不適格性が生成される理由として，音声言語の生成プロセスも含めて不適格性の考察を行い，音声を正しく認識する為の前処理の一つとして，音声認識を行う前に不適格性の一つである長音を含む言い淀みを冗長語と定義し，冗長語を除去する事によって音声認識システムの精度が向上する可能性がある事，また冗長語の特長である，母音が安定して伸びている区間を延伸母音とし，延伸母音の部分を検出する手法を提案し，提案した手法が実際にどの程度適用する事ができるのか，音声コーパスを用いた評価実験を行ったので，その結果を報告する．

２.音声言語の生成モデルと冗長語

普段我々が発話を行う際，発話内容に関する事柄を心的データベースから引き出し，その中の当面の対話に必要な情報は心的バッファで心的操作が行われ，更にその心的バッファの領域内で，概念化，構文化，音韻変換というプロセスを行う事で発話を行うとされている[3][4]．このモデルを図１に示す．

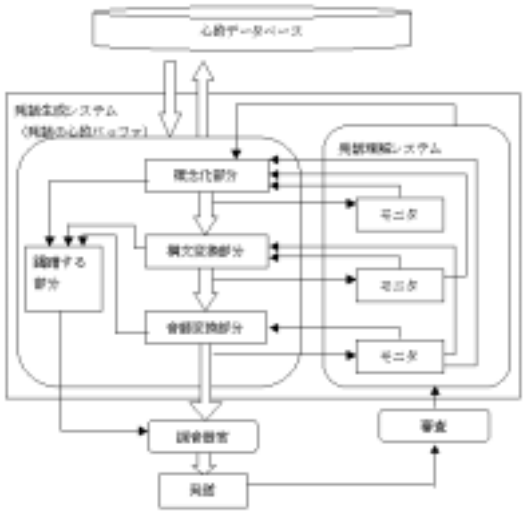

図１．発話の生成モデル

発話中に不適格性が不可避なものとして現

れる理由として，思考プロセスが発話プロセスに追いつかない場合が現れ，次の発話内容が発話プロセスに届く間，話者は時間を稼ぐため言い淀みを用いると考えられる．この事から，言い淀みには以下の様な大切な役割を持っていると考えられる．

・ 発言権の保持
音声対話では，その進行に従い，話者間で発話権が移動していく．話者の立場からは，次の発話が心的バッファに準備できていないにも関わらず発言権を保ち続けたい時，あるいはとりあえず何か発言しなければならない時，発話を準備しながら言い淀む事で，聴取者に次の発言権を待って欲しいと暗に伝える事ができる．

・ 話者の心理状態・思考状態を表す機能
円滑な対話を進める為に，話者は自分の心的状態・思考状態を，無意識に意識的にも聴取者と共有する行動をとる．話者の立場からは，有声休止の方法（音韻やイントネーション，発声法等）によって，発話内容に対する自信の無さ，不安，躊躇，謙遜といった心的状態を表現できる（あるいは露呈してしまう）．また，その間投詞・冗長語の種類によって，何かを思い出そうとしているのか，あるいは聴取者にとって適切な表現を探しているのかといった，異なる思考状態を表現する事ができる．

従って，話者が心的バッファでの発話内容の処理を待っている間，冗長語が発声されるのであれば，調音器官をどのように変化させて良いか分からず，喉頭を含む調音器官の位置・状態を変化させる事ができない．従って冗長語の区間では調音器官がほぼ一定のまま，声道形状が殆ど変化しない状態で声帯が振動していると考えられるので，声帯の振動周期である音声の基本周波数を求め基本周波数の時間的な変化を利用することは有効な手法であると考えられる．

３．発話セグメントを用いた冗長語検出[5]

２章で述べた冗長語の基本周波数の特徴を利用する為，ケプストラムによって基本周波数を推定する方法を用いた．図２に処理手法を示す．

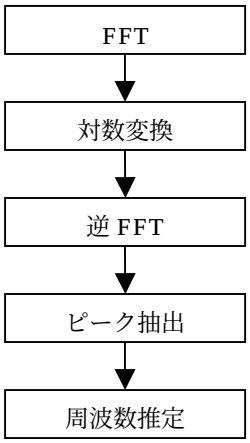

図 2．基本周波数推定処理の流れ

この手法によって得られた基本周波数の特徴と図１の発話モデルと調査により，冗長語前後に無声区間が現れやすいという特徴を利用して，無声区間に挟まれた有声区間を発話セグメントとし，発話セグメント中の基本周波数の変動量が少ない区間との時間比率を利用する事で，冗長語の検出し，除去を行うシステムを実装した．図 3 に処理の流れを示す．処理で用いている基本周波数の安定時間は対話音声コーパスの調査

結果を利用した．発話比率や基本周波数の変動幅は実験によって調整を行った．

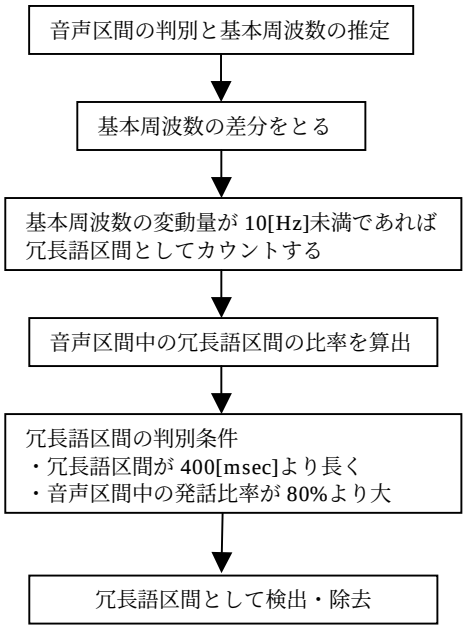

図 3．冗長語区間の判別処理の流れ

４.音声認識の予備実験

３章で示した手法によって検出し，除去された冗長語を，市販の音声認識システムによって改善されるか調査する予備実験を行った．今回使用した音声認識システムは ScanSoft 社が開発した Dragon Speech セレクト Ver.6 で，話者本人によるシステムへの学習を行った後，自動除去された音声ファイルを音声認識させた．その結果を以下に示す．

＜冗長語除去前＞
小泉総理は広島をいやAA長崎を訪問した
＜冗長語除去後＞
小泉総理は広島をいや長崎を訪問した

結果より“えー”という冗長語が“AA”と認識され，誤認識が発生していたが，システムによって除去され，誤認識が発生しなくなっていることがわかる．

しかし通常の発話であっても基本周波数の変動量が少ない時間が比較的長く継続している箇所や，冗長語であっても発話時間が短いものや，途中で基本周波数の変動量が大きくなっている箇所があり，継続時間がリセットされてしまい検出しない場合が起きていた．また，無音区間に挟まれた発話セグメントに依存している為，通常の発話と連続して発話されている冗長語の区間が検出できないといった問題があった．

５.フォルマント周波数の推定

基本周波数だけでは冗長語の区間を推定するのは困難であるという問題と，発話セグメントに依存してしまうと連続した発話での検出ができないという問題に対し，母音区間のスペクトルの特徴が強く現れる，第１・第２フォルマント周波数を利用する事を試みた[6]．フォルマントを抽出しその周波数を求める処理方法を図４に示す．

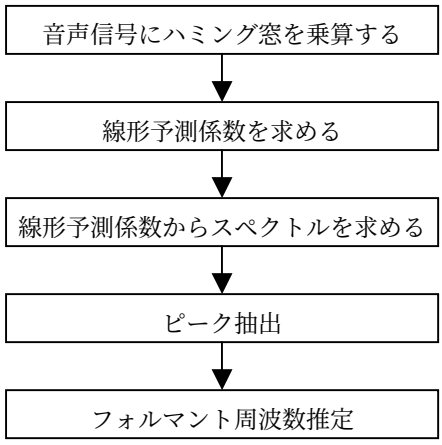

図 4．フォルマント周波数推定処理の流れ

６.抑制フィルタの作成

線形予測を利用した方法のみではフォル

表１．設定したフィルタのパラメータ

パラメータ	フィルタ 1	フィルタ 2	フィルタ 3	フィルタ 4	フィルタ 5
周波数[Hz]	187.50	406.25	1593.75	2000.00	2500.00
ゲイン[dB]	- 60.00	- 60.00	- 59.00	- 59.00	- 59.00
尖度	全て 400.00				

マントの状態が大きく変動してしまう為，変動を抑制する２次の IIR フィルタを５段作成した．表１にフィルタ作成に用いたパラメータを，図５にその周波数応答を示す．このパラメータは参考文献[6][7]と調査した結果を基に設計したものである．

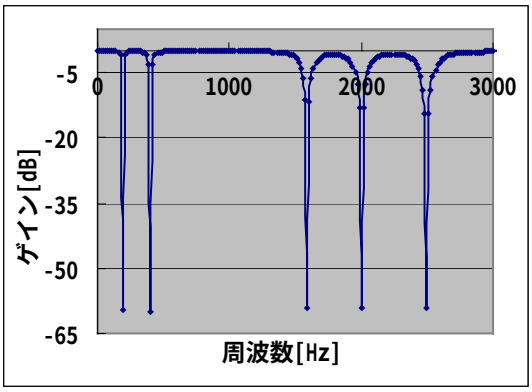

図５．フィルタの周波数応答

７.システムの実装

実装した線形予測法による第１，第２フォルマントを推定し，分散を求める方法と，2 次の IIR フィルタを 5 段組み合わせたものを，これまでの基本周波数から冗長語区間を求めるシステムに組み込んだ．また発話セグメントを利用せずに，基本周波数とフォルマントが安定している区間のカウントを行い，閾値を超えると，その区間を長音区間として判別する事ができるようにした．図 6 にシステムの処理の流れを示す．現在の実装ではフレーム幅は 512point，フレームシフトは 160point で，音声はサンプリング周波数 16kHz，量子化ビット数 16bit での処理を行っている．検出された区間は除去できる様になっている．

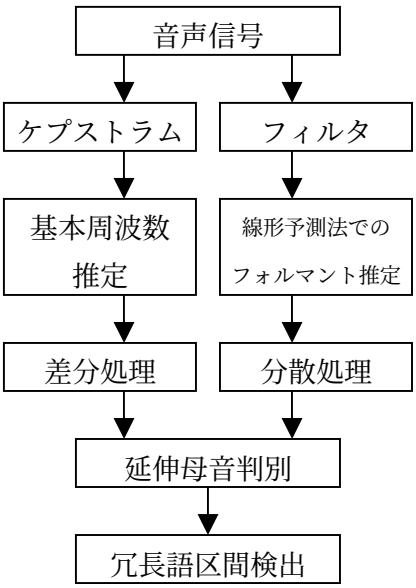

図 6．システムの処理の流れ

このシステムによって検出された延伸母音区間を図 7 に示す．

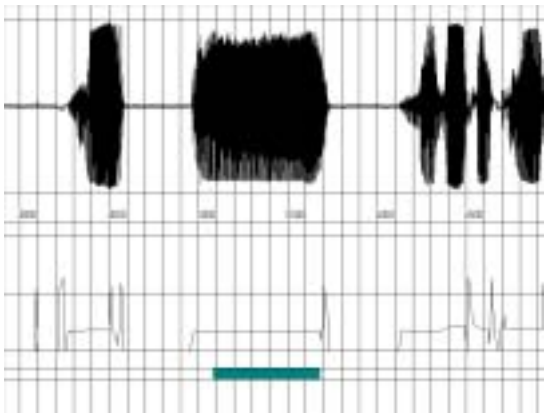

図７.検出された延伸母音区間

処理結果の確認の為，出力形式として上段はシステムに入力された音声信号波形，中段は音声信号の変化に対応した変化を確認できるように基本周波数の時間変化のグラフを表示し，下段では音声区間と基本周波数に対応した延伸母音区間を示すインジケータになっている. 図 7 に示す音声は「小泉総理は広島を，いや，えー，長崎を訪問した」という発話内の「いや，えー，長崎を」という部分である．「えー」の延伸母音部分が検出されている．システムは CPU が Pentium4　2GHz の PC 上で動作しており, 処理速度は 10 秒の音声を約 12 秒で処理が可能となっている．

８.コーパスを用いた調査

製作したシステムによって，対話音声コーパス[8]（発話ドメインは講演, 話者は男性 5 人）を用い, 調査を行った. 音声ファイルは，文字化された対話音声コーパスに話者別の公演毎に収められていた物の中から延伸母音が含まれていた物を使用した．調査方法はシステムで検出（ラベル付けを）された音声区間が実際に聞いた場合，延伸母音であるかという主観評価によって比較を行った．表 2 にその結果を示す．調査に用いたパラメータとして，フィルタ以外のパラメータは以下のもの使用した.

- 最低発話継続時間：200[msec]
- 基本周波数許容変動幅：20[Hz]
- 第 1 フォルマント許容分散：500
- 第 2 フォルマント許容分散：750

これらのパラメータについてはこれまでの調査結果を考慮し設定したものであるが，フィルタ同様に話者によっての依存性があり，最適な値は不定である．これらのパラメータの決定法は今後の課題と考えている.

表 2 .検出実験の結果

時間長(msec)	検出数（個）
200～300	38
300～400	14
400～500	7
500～600	2
600～700	5
700～800	0
800～900	1
900～1000	0
1000 以上	0
合計	67
合計時間	17 分 47.801 秒

この検出結果と，音声を実際に聞いた主観評価によって比較した適合率を求めると以下のような結果となった.

$$\text{適合率} = \frac{\text{実際の結果}}{\text{システムの出力結果}} = \frac{59}{67} \cong 0.8806$$

９.まとめ

音声認識システムの誤認識の原因となる冗長語を，発話モデルを用いてその役割を示し，その役割から特徴を考え，冗長語を除去する手法を構築し，音声認識システムによる実験を行った.

また冗長語を検出するアルゴリズムを基にして，連続した発話の中から，比較的短時間な延伸母音を検出のできる手法を提案しシステムの実装を行い，評価実験を行っ

た．その結果約 88%の適合率を得る事ができた．

今後の課題として，システムに設定するパラメータの決定方法，検出された延伸母音区間から冗長語を除去する方法の考案がある．

＜参考文献＞

[1]後藤　真考，伊藤　克亘，　速水　悟：“ 自然発話中の言い淀み箇所のリアルタイム検出システム ”, 情報処理学会　音声言語情報処理研究会　SLP-27-2（1999）.

[2]中川　聖一, 小林　聡：“ 自然な音声対話における間投詞・ポーズ・言い直しの出現パターンと音響的性質 ”，日本音響学会誌 51 号第 3 号，pp.202-210(1995).

[3] Yuji Sagawa，Masahiro Ito，Noboru Ohnishi，Noboru Sugie：“ A Model For Generating Self － Repairs ”：Proc. of International Conference on Spoken Language Processing，(1994).

[4]定延　利之, 田窪　行則：“ 談話における心的操作モニター機構－心的操作標識「ええと」と「あのー」－ ”，統語構造と韻律的特長との対応関係に関する研究，pp15-33(1993).

[5]Keiji Niwa, Yuji Sagawa, Noboru Sugie：“ Detecting Prolonged Vowel In Spontaneous Speech ”: Proc. of the IASTED International Conference on Artificial Intelligence and Applications，pp.29-33，(2003).

[6]古井貞熙：“ 音響・音声工学 ”, pp.115-120, pp.124-127，近代科学社(1998).

[7]Robert Bristow-Johnson：”Cookbook formulae for audio EQ biquad filter coefficients”, <http://www.harmony-central.com/Computer/Programming/Audio-EQ-Cookbook.txt. >

[8]堂下 修司：“ 音声・言語・概念の統合的処理による対話の理解と生成に関する研究 ”, 人工知能学会誌 Vol.12 No.1, (1997).