有声休止，無声休止の自動検出を考慮したデコーディングによる自由発話音声認識*

○緒方 淳† 後藤 真孝† 伊藤 克亘‡ (†産総研，‡法政大)

1 はじめに

自由発話音声認識においては，不明瞭な発声や口語表現，言い淀み，発話速度の変動など，様々な要因により認識性能が劣化してしまう．本研究では，その中でも特に，現在の音声認識では扱うことが困難な，有声休止，無声休止の２つの非言語情報に着目する．本報告では，自然発話中の有声休止，無声休止の音響的特徴をボトムアップな信号処理にて検出し，それらを認識時に考慮することで，両休止に対する頑健な音声認識手法を提案し，その評価を行う．

2 有声・無声休止の自動検出に基づく音声認識手法

まず，本研究の提案手法の概要について述べ，有声・無声休止区間の検出手法，休止区間スキップを用いたデコーディング手法について説明する．

2.1 システムの概要

提案するシステムの概要を図１に示す．まず，入力音声に対し，後述する有声休止区間検出，無声休止区間検出をそれぞれ実行し，有声休止，無声休止の区間情報 (始端時刻，終端時刻) を算出する．次に，得られた休止区間情報を音声認識の探索過程に考慮することで，有声・無声休止に頑健なデコーディングを実行し，認識結果を出力する．

また，更なる高精度化のために，音響モデルの学習の段階にも，有声・無声休止区間検出を適用することを考える．音響モデルの学習データの全発話に対し，同様の各休止検出を実行し，区間情報を得る．それらの時刻の情報をもとに，各発話の音響特徴量データ (MFCC) の中から，休止区間に相当する箇所を除去する．そして，休止区間が除去された特徴量データを用いて，音響モデルの再学習を行う．

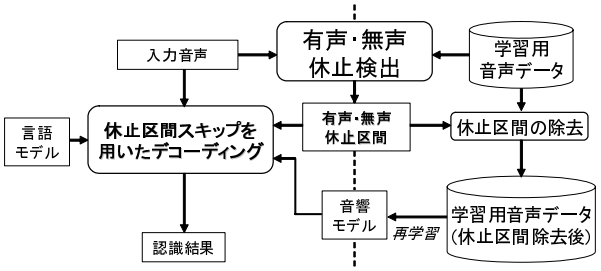

図 1: 提案システムの概要

2.2 有声休止，無声休止の検出手法

有声休止の検出には，後藤らによって提案されたリアルタイム有声休止検出手法 [1] を採用した．本手法では，有声休止が自然な発話において不可避なのは，それが思考プロセスが発話プロセスに追い付かない場合に表れる現象であるからだという仮説に基づく．すると有声休止は，調音器官がほぼ一定のまま声帯が振動し続けるときの音声，すなわち，音韻的に変化が少ない持続した母音の引き延ばしを伴っていると仮定できる．そこで，そうした有声休止が持つ２つの音響的特徴 (基本周波数の変動が小さい，スペクトル包絡の変形が小さい) をボトムアップな信号処理によってリアルタイムに検出する．そのため，任意の母音の引き延ばしの開始点を終了点を，言語非依存に検出できるという特長を持っている．

無声休止 (無音) の検出法としては，従来から様々な手法が存在するが，本実験では，音声信号中のパワー情報に基づく検出手法を検討する．すなわち，振幅スペクトルのパワー値に対して閾値処理を行うことにより，無声休止区間の決定を行う．

2.3 休止区間スキップを用いたデコーディング手法

以上の手法にて検出された有声・無声休止区間を用いたデコーディング手法について述べる．発話中に有声休止あるいは無声休止が検出され，その区間情報 (始端時間，終端時間) が与えられると，以下の処理が実行される．フレーム同期ビーム探索において，認識処理が検出された休止区間の開始時刻に到着すると，音声認識器の動作を一時停止し，現時点の認識処理過程 (それまでの仮説情報，探索空間での現在の位置情報等) を保存する．休止区間のフレームは，認識処理の対象とならず，スキップされる．認識処理が休止区間の終端時刻に到着すると，保存された各情報をもとに認識処理が再開される．

本手法は，基本的に，入力音声中の検出された休止区間を無視してデコーディングを行う方法である．ただし，無声休止の場合，検出された全ての休止区間を完全に無視すると，単語内において発生した無声休止に対しては有効に働くと考えられるが，単語間に発生した無声休止の場合，実際の認識の際に単語の区切りを同定することが困難になり，認識性能を劣化させる可能性がある．そこで，検出された休止が無声休止の場合は，休止区間全てを無視するのではなく，ある一定の長さだけ休止区間を残すようにする．ここでは，単語内での無声休止に対する効果も考慮し，検出された無声休止を比較的短い時間長に統一する．

以上のデコーディング手法は，先に述べた音響モデルの再学習時と同様に音響特徴量データから休止区間をあらかじめ除去し，除去したデータに対して通常のデコーディングを直接行う場合と基本的には同等の効果をもたらす手法といえる．ただし，提案手法のようにデコーディングの過程において，休止区間の情報を直接組み込むことによって，様々な発展，応用が考えられる．例えば，休止の前後での文脈に関する統計情報を学習しておき，実際の認識時の休止区間においてそれらをデコーディングの過程に組み込む方法や，休止中やその前後では話速が変化する傾向があることから，休止区間に基づいてデコーディングパラメータを動的に決定する手法，などが考えられる．このようなデコーディングの高精度化については今後の課題とする．

3 評価実験

提案手法の効果を調べるため，実環境の自由発話音声データを用いて評価実験を行った．

3.1 実験条件

本実験では，使用するデータベースとして，CIAIR 車内コーパス [2] を用いた．文献 [3] によると，CIAIR の対話音声は，自由発話の特徴が顕著に現れている音声データであり，発話速度の変動も他のコーパス (CSJ,

*Spontaneous Speech Recognition Based on Automatic Filled and Silent Pause Detection.
by Jun OGATA, Masataka GOTO (AIST), Katsunobu ITOU (Hosei University)

JNAS) に比べて大きいことが報告されている．また，本研究で対象としている有声休止，無声休止が頻繁に発生した音声データとなっている．これは，CIAIR はカーナビゲーションを想定した対話であり，発話者は運転中のため，発話内容をその場で考えなければならないことが多いためである．

音響モデルには，長母音化を考慮した日本語音節モデル [4] を用いた．本モデルは，自由発話の特徴である長母音化を個々のサブワード単位内に考慮したもので，日本語自由発話音声認識において，一般的な triphone モデルと同等以上の性能を持つことが示されている．サブワード単位数は 245，そのうちの 3 つは無音モデル (発話開始の無音，発話終端の無音，単語間ショートポーズ) である．サブワード間のコンテキスト依存はなく，mono 音節モデルとなっている．学習データには，CIAIR の音声データのうち，401 名のドライバにより発声された音声データ 79093 発話を用いた．言語モデルは，CIAIR の対話音声の書き起こしテキスト 94306 文を用いて学習した単語 bigram(語彙数 4305) である．

評価用データとしては，まず上記の学習データとは別のデータ 11021 発話に対して，有声，無声検出器を実行し，その中から，より多くの休止が含まれている発話 (発話中の休止区間長の合計が長い発話) を順に 600 発話選択した．

3.2 実験結果

有声・無声休止区間検出に基づく音声認識の評価実験を行った．2.3 節で述べた，無声休止検出後の区間長は予備実験の結果より，0.1sec とした．実験結果を図 2 に示す．ここで,「ベースライン」はベースライン音声認識の結果,「有声」は有声休止検出のみを用いた場合,「無声」はパワー情報に基づく無声休止検出のみを用いた場合,「有声・無声」は両方を用いた場合の結果である．提案手法の 4 種類では，検出結果に基づいて音響モデルを再学習したときの結果も示している．

まず，有声休止区間検出を考慮した音声認識結果とベースラインの結果を比較すると，音響モデルの再学習も併用することで約 3%の性能向上がみられた．認識結果を調べたところ，改善されたパターンとしては，単語末尾の有声休止，単語内部の有声休止に対するものが特に多かった．CIAIR のような対話タスクの場合，ある特定の名詞 (店名や地名，製品名等) において，有声休止が発生することが頻繁にあり (例: コンビニ ⇒ コンーービニ)，これを言語モデルや発音モデルなどの事前学習の枠組みのみで解決することは困難と考えられる．また，極端に長く発声されたフィラーに関しても，本手法によって湧き出し誤りが抑えられ，多くの改善が得られた．

次に，無声休止区間検出を考慮した音声認識結果とベースラインの結果を比較する．音響モデルの再学習も併用することで，最高で 2.6%の性能向上がみられた．本手法にて改善されたパターンとしては，まず単語内部の無声休止に対するものが挙げられる．以下に実際に本手法で改善された，単語内無声休止の発声例を示す (発声文中の単語のみを表示，[sp] は無声休止を表す)．

- ホテル ⇒ ほ [sp] てる
- 今池支店 ⇒ いまい [sp] けしてん

実環境での対話タスクにおいては，発話者は発話内容を思考しながら発声することが多いため，単語間だけでなく，単語の内部においてもこのような無声休止が多くみられる．また，単語間の無声休止については，今回の実験では，ベースラインのシステムにおいて，単語のエントリとしてショートポーズを学習しているため，無声休止が発生したことによる直接的な誤認識はみられなかった．しかし，比較的長い無声休止中に，発話者のリップノイズや小さな舌打ちなどの雑音が発生し，これによる湧き出し誤りが発生していた．パワー情報に基づく無声休止検出では，これらの比較的小さな雑音が閾値処理によって棄却されており，認識改善に寄与していた．

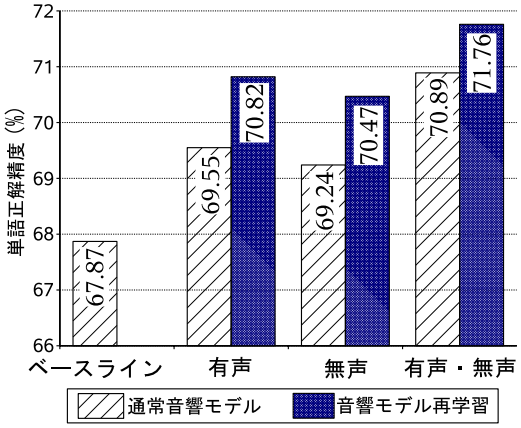

図 2: 各認識手法の認識精度

最後に，有声・無声休止の両方の検出結果を考慮した音声認識結果について述べる．ベースラインと比べて，音響モデルの再学習も併用することで，有声休止，無声休止を単独で行うよりも更に改善が得られ，最終的に約 4%の性能向上がみられた．本実験で扱ったような自然発話，特に対話音声においては，有声休止，無声休止が 1 発話中に同時に発生することも頻繁にあると考えられる．また，1 発話中の，ある 1 単語においてこれら 2 つの休止が同時に発生することも考えられる．本実験データにおいて実際に存在した例を以下に示す．

- 正解：**和食** の お店 に …
- 発声：**わしょ** [sp] **くー** の おみせ に …
- 従来：**場所 くう** の お店 に …
- 提案：**和食** の お店 に …

上から順に，正解単語列，実際の発声，従来手法での認識結果，提案手法での認識結果をそれぞれ示している．この例では,「和食」という単語の内部に無声休止が発生し，末尾に有声休止が発声している．本手法を用いることで，このような，従来の音声認識手法では困難な発声に対しても改善が得られることを確認した．

4 まとめ

本報告では，自由発話音声認識の性能改善を目的とし，有声・無声休止区間の自動検出に基づく音声認識手法を提案し，実験によりその効果を確認した．従来検討されてきた，音響，言語，発音などの各モデルの事前学習において，休止情報をモデル化する手法では，タスクやデータベースに依存することが避けられず，あらゆる自由発話に対処することは困難であった．本研究は，言語やタスクに非依存な休止区間検出手法を用いることで，あらゆる自由発話中の休止に対して頑健な手法の実現を目指したものである．

今後の課題としては，デコーディングアルゴリズムの高精度化，他のタスク，データベースによる評価などが挙げられる．

参 考 文 献

[1] 後藤 他: 信学論, Vol.J83-D-II, No.11, pp.2330-2340, 2000.
[2] K.Takeda, et al.: IEICE Transactions on Information and Systems, Vol. E88-D, No. 3, pp.553-561, 2005.
[3] 山田 他: 情処研報, 2005-SLP-58, pp.1-6, 2005.
[4] 緒方 他: 信学論, Vol.J86-D-II, No.11, pp.1523-1530, 2003.