論文 / 著書情報
Article / Book Information

論題(和文)	自由発声中の連続数字音声認識
Title(English)	Dijit string recognition in spontaneous speech
著者(和文)	Etienne Bauche, 南泰浩, Bojana Gajic, 松岡達雄, 古井貞熙
Authors(English)	SADAOKI FURUI
出典(和文)	日本音響学会 1997年春季講演論文集, Vol. , No. 2-Q-15, pp. 169-170
Citation(English)	, Vol. , No. 2-Q-15, pp. 169-170
発行日 / Pub. date	1997, 3

2—Q—15　自由発声中の連続数字音声認識*

△Étienne Bauche ○南　泰浩 △Bojana Gajic 松岡達雄　古井貞熙
(NTTヒューマンインタフェース研究所)

1．はじめに

現在の音声認識の性能は、小語彙のタスクであれば、かなり高水準に達して来ている。しかし、実際にシステムを作成すると、期待していたほどの性能が達成されないということがしばしばある。これは、音声の切り出しの誤りや、不用語や言いよどみなどによって音声認識システムの性能が劣化するためであると考えられる。そこで、ここでは、自然に発声された音声を使って、音声認識結果を悪化させる様々な現象を分析する。それらの原因に対処する様々な手法を導入し、音声認識性能の改善を図る。

2．認識タスク

対象としたタスクは桁数既知の4桁の数字認識である。ここでは、この認識を行うために数字HMM、background HMM、alldigit HMM、garbage HMMを作成した。background HMMは発声のされていない区間から作成したHMMであり、alldigit HMMは全ての数字の発声から1つのHMMを作成したものであり、garbage HMMは長時間の音声から1つのHMMを作成したものである。数字HMMとbackground HMMの作成には電子協日本語共通音声データベース中の4桁数字を9704個使用した。garbage HMMの作成には音響学会研究用連続音声データベース中の男性30人が発声した音声を使用した。使用したパラメータはケプストラム、デルタケプストラムとデルタパワーを用いた。これらのHMMではCMN[1]を使っている。

評価用のデータベースとしては、音声始端終端検出を評価するために1784個の4桁数字を用意した。これとは別に音声認識システムの総合評価をするために、27人の発声した1013個の4桁数字を用いた。評価用のデータは「あなたの車の番号は何ですか」などのような質問や数字当てゲームに答えさせるような質問を使って収録した。このため、「あの」、「えっとー」などの不用語、番号を思い出す間のポーズ、コンピュータの予期せぬ答えによる混乱した発声などが数多く録音できた。

3．音声始終端検出

音声の切り出しは(1)式の値が閾値より大きいか小さいかに基づいて判定を行う。Nは尤度を平均化するフレーム数である。ここでは5フレーム（40msec）に設定している。このフレーム数をここではブロックと呼ぶことにする。

$$diff = \sum_{i=1}^{N} [\log P(o_i \mid alldigit) - \log P(o_i \mid background)] \quad (1)$$

閾値を越えたブロックが2ブロック以上続く領域の始端、あるいは閾値以上のブロックの間に閾値以下の連続8ブロック以内の領域（発声と発声の間のポーズに対応する）が存在する領域の始端を音声の仮の始端とする。始端はこの仮の始端より時間的に2ブロック戻った場所に設定する。終端は、音声の後に閾値以下の領域が8ブロック以上続いた場所とする。以上で説明したブロック数や閾値は4桁の数字認識の予備実験により最適化した。この予備実験から、パワーだけを用いた音声始終端検出法では誤認識率が5.2%であるのに対し、本手法を用いると誤認識率が3.1%まで改善されることが確認された。このとき、人の視察による音声始終端検出では誤認識率が2.4%であった。

4．ベースラインの認識結果と誤認識の分析

ここでは文法のネットワークとして、次のものを使った。

bak — d — d — d — d — bak

図1 4桁数字認識のためのネットワーク

ここで、bakはbackground HMM、dは数字HMMを表す。このネットワークでの誤認識率は24.4%であった。

誤認識であった247個の数字列を分析した結果、以下のようにグループ化できた。

(1) 数字間の長いポーズ：121個

これは、発声を考えたり躊躇したりしたため、8ブロックを越える長いポーズが数字と数字の間に挿入されたためと考えられる。

(2) out-of-vocabularyによる影響：69個

「ええ」や「んん」などによる誤認識が31個、「ええと」による誤認識が13個確認された。

(3)数字の置換：45個

「はち」と「いち」を間違えるなどの数字間の置換による誤認識である。

(4) 数字の挿入や脱落：12個

このネットワークでは数字を4桁に限定しているため、パターンの一致しないところに数字を挿入したり削除したりする現象がみられた。

5．音声認識システムの改善

(A) garbage HMMの導入

out-of-vocabularyに対応するために図1のネットワークにgarbage HMMを以下のように導入する。

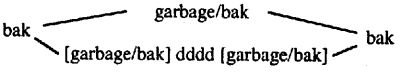

図2 garbage HMMを導入したネットワーク

*Digit string recognition in spontaneous speech.
By Étienne Bauche, Yasuhiro Minami, Bojana Gajic, Tatsuo Matsuoka and Sadaoki Furui
(NTT Human Interface Laboratories)

(B) ポーズアルゴリズムの導入

このアルゴリズムは数字間の長いポーズに対処するために導入した。まず、数字の信頼度を調べるために以下の3つの尺度を用いる。この信頼度から数字を3段階に評価し、この評価結果に基づいて、入力された数字列が受理できるかどうかを確認する。受理できない場合は数字間にポーズがあると仮定して処理を行う。

・継続時間長

ある継続時間長より短い数字は「不可」とラベルづける。

・background HMM との信頼度

background HMM と数字HMM間の(2)式の確信度が0以下であれば、その数字を「不可」とする。ここでTはビタービ探索によって得られる数字の長さであり、Pはビタービ探索で得られた尤度を表している。

$$D=\frac{\sum_{i=1}^{T}[\log P(o_i \mid digit)-\log P(o_i \mid background)]}{T} \quad (2)$$

・garbage HMM との信頼度

以上の信頼度でラベル付けされなかった音声に対しては(3)の信頼度を計算する。この値により数字を「不可」、「可」、「良」とラベル付けする。

$$D=\frac{\sum_{i=1}^{T}[\log P(o_i \mid digit)-\log P(o_i \mid garbage)]}{T} \quad (3)$$

以下にこれらの信頼度を使ったポーズ処理のアルゴリズムを示す。

(1)最大のポーズの長さを8ブロックに設定し、音声始端検出を行う。

(2)音声終端の検出し、認識を行う。

(3)もし、認識された数字が全て「可」以上であれば、結果を出力する。もし、数字列の中に一つも「良」の数字がなければ(1)へ（音声と仮定した入力の中に数字が存在しないので、入力を捨て去り新たな始端の検出を行う）。そうでなければ、(4)へ。

(4)最大のポーズの長さを+8して(2)を実行。このとき最大のポーズが64ブロックを越えていて、かつ数字の3つ以上が" 可" 以上であれば結果を出力し終了。そうでなければ(1)に戻る。

(C) filler HMM の導入

「ええと」などのような発声に対処するためにfiller HMMを数字の発声の前に挿入した。このfiller HMMを作成するために、10人の発声した300数字列を使った。このデータは評価用の音声を収録した場所と同じ場所で収録した。

(D) HMMの精密化

置換誤りに対処するために、HMMの混合数を4から8に増加させた。

(E) 環境雑音への適応

background HMMを作成した環境と認識を行っている環境では、雑音が異なるため雑音への適応をおこなう必要がある。雑音適応としてはHMM合成法[2]などの手法も考えられるが、すでにCMNを使っているためHMM合成法は使えない。そこで、ここでは、MAP推定[3]を雑音適応として用いた。この適応用のデータとしては(D)で使ったデータを利用した。

6．実験結果

実験結果を表1に示す。garbage HMMの導入により誤認識率は21.5%改善された。しかし、このモデルを入れてもまだ「ええと」や「ええ」などによる誤認識は改善されなかった。ポーズアルゴリズムの導入により、誤認識率はさらに33%改善された。このアルゴリズムによって数字間のポーズによる数字の挿入誤りが230から8へと大幅に改善された。「ええと」や「ええ」などに対処するようなfiller HMMを導入すると、若干の改善が見られた。しかし、このときのfiller HMMの学習話者には評価話者が6人含まれているので、正当な評価とはなっていない（この後の実験のためこの6人を除いた誤認識率も（）内に記述しておく）。分布の混合数を8に増やすと若干の誤認識率の改善が確認された。しかし期待したほど，置換誤りを改善するには至らなかった．MAPによる適応化によりさらに誤り率が30%改善された。

最終的には、全ての処理を行った結果、誤認識率が24%から5.5%まで改善された。

表1　認識結果

手法	誤認識率	誤認識数字列の数
ベースライン	24.4%	247
(A)garbage HMMの導入	19.2%	194
(B)ポーズアルゴリズムの導入	11.4%	115
(C)filler HMM の導入	9.5% (8.7%*)	96 (68*)
(D)HMMの精密化（8混合分布）	8.3%*	65*
(E)環境雑音への適応	5.5%	43*

＊は784数字列での評価を示す．その他は1013個の数字列での評価である．

7．まとめ

本論文では、自由に発声された数字列を対象に、認識率を向上させるための各種の手法を提案した。この結果、誤認識率が24%から5.5%まで改善された。特に今回提案したポーズアルゴリズムは単語と単語の間の長いポーズに対して有効であることが確認された。

文献

[1] B. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," Journal of the Acoustic Society of America, vol. 55, pp. 1304-1312, June 1974.

[2] Lee, C.-H. and J.-L. Gauvain, "Speaker adaptation based on map estimation of hmm parameters," Proc. ICASSP, 558-561, April 1993.

[3] F. Martin, K. Shikano, and Y. Minami, "Recognition of noisy speech by composition of hidden Markov models," Proc. Eurospeech, pp. 1031-1034, September 1993.