論文 / 著書情報
Article / Book Information

論題(和文)	英語不特定話者連続音声認識の試作
Title(English)	
著者(和文)	磯谷亮輔, 渡辺隆夫, 畑崎香一郎, 永野敬子, 篠田浩一, 田海真一, M. Chong
Authors(English)	Koichi Shinoda
出典(和文)	日本音響学会平成4年度春季研究発表会講演論文集, Vol. 1-P-10, No. , pp.
Citation(English)	, Vol. 1-P-10, No. , pp.
発行日 / Pub. date	1992,

1—P—10

英語不特定話者連続音声認識の試作 *

⊙ 磯谷 亮輔　渡辺 隆夫　畑崎 香一郎　永野 敬子　篠田 浩一
田海 真一　Michael Chong (日本電気(株) C&C 情報研究所)

1. まえがき

筆者らはこれまでに、半音節を認識単位とした混合ガウス分布HMMによる日本語不特定話者連続音声認識の検討を行ない、その有効性を示してきた[1, 2]。

今回これらの結果をふまえ、diphoneを認識単位とする英語不特定話者連続音声認識システムを試作した。発声変形の取り扱い、単語間接続などに関し検討を加え、評価を行なった結果について報告する。本方式は自動通訳システムINTERTALKER[3]の英語認識部として用いられた。

2. 認識単位

認識単位としては、一般に音素の他 generalized triphone [4]など環境に依存した音素などが用いられている。今回は認識単位として、音素と diphone について試みた。音素としては無音を含め50音素を定義した。diphoneは音声学的知識に基づき各音素を後続音素に応じて2～11個に細分化したもので、総数は354種類である。

各音素または diphone を left-to-right、スキップなしのHMMで表す。各モデルの状態数は1～3である。HMMの各状態ごとの特徴ベクトル出力確率を混合ガウス分布で表現する。

3. 認識方式

3.1. ネットワーク制御連続音声認識

認識方式として、ネットワーク制御連続音声認識[5]を用いる。文法を構文ネットワークとして記述し、各単語の発音を音素または diphone の列として単語辞書に記述する。構文ネットワーク、単語辞書とHMMとをあらかじめ単一のネットワーク上にコンパイルしておく。認識時には、入力フレームに同期したViterbiアルゴリズムによりネットワーク上で最も尤度の高い経路を求める。認識処理の高速化のため、「バンドルサーチ」[6]を行なう。

3.2. 発声変形の扱い

文発声では、同じ単語でも話者あるいは文脈などにより発音はさまざまに変形する。この変形を、単語辞書における発音表記のレベルで記述する。すなわち、各単語に対し複数の発音表記を対応させる(マルチエントリ辞書)。

3.3. 単語間接続

認識単位として diphone を用いる場合には、単語の接続部に前後の単語に応じた diphone モデルを挿入する。処理量削減のため、先行単語の語尾に後続の音素環境に依存しない音素モデルを用いる方法についても比較検討した。ここで用いる音素モデルとして、

(1) 音素を認識単位として学習した音素モデル
(2) 中心音素が同じで後続環境の異なる diphone モデルを「平均化」したもの

の2通りを検討した。(1)の方式では、独立に学習した音素モデルと diphone モデルを認識時に接続することになり、この接続部で時間的な不連続が生じる可能性がある。(2) はこの問題に対処するものである。「平均化」にはHMMの学習におけるパラメータの「結び」の手法を用いる。すなわち、対応する状態、遷移に関するすべてのパラメータを「結び」の関係として、1ループだけforward-backward アルゴリズムで学習する。

4. 評価実験

4.1. 音声データ

音声データとして、HMM の学習用に

(a) 音素バランスを考慮した273単語36名(男性14名、女性22名)各1回発声
(b) 観光案内・チケット予約タスク単語305単語、観光案内タスク文35文18名 (男女各9名) 各1回発声

を用い、評価用に

(c) 観光案内タスク文35文18名(男女各9名)各1回発声

を用いた。(a)、(b)、(c) の話者はすべて異なり、また (b) と (c) のタスク文には同じ文は含まれない。

これらのデータをサンプリング周波数16kHz、帯域 0.1～7.2kHz、分析窓長 32msec、フレー

* An Experimental English Speaker-Independent Continuous Speech Recognition System, by R. Isotani, T. Watanabe, K. Hatazaki, K. Nagano, K. Shinoda, S. Taumi, and M. Chong (NEC Corporation)

ム周期 10msec でメルケプストラム分析し、メルケプストラム係数 10 次元、各メルケプストラム係数の変化量、正規化パワー差分の計 21 次元のベクトルを特徴ベクトルとした。

4.2. 実験条件および結果

観光案内のタスクを用いた連続音声認識実験を行なった。タスク規模を表1に示す。評価に用いた文の平均文長は 6.9 単語である。

表 1: タスク規模

異なり単語数	267
単語パープレキシティ	6.8

はじめに、音素を認識単位としてマルチエントリ辞書をシングルエントリ辞書 (各単語に 1 通りの発音表記が対応) と比較した。マルチエントリ辞書の 1 単語あたりの平均エントリ数は 5.4 である。HMM モデルのガウス分布混合数は 4 とした。

結果を表2に示す。辞書をマルチエントリにすることにより認識率が大きく改善された。

表 2: 辞書の比較 (音素モデル)

辞書	文認識率
シングルエントリ	62.4%
マルチエントリ	70.3%

つぎに、diphone を単位として学習・認識実験を行なった。まず、学習データとしてタスクに依存しないデータ (a) だけを用いた場合と、タスクのデータ (b) も用いた場合との比較を行なった。各 diphone の HMM モデルのガウス分布混合数は 4 とし、辞書はマルチエントリ辞書を用いた。単語間の接続には前後の単語に応じた diphone モデルを用いた。結果を表3に示す。タスク独立な学習では認識率が大きく低下した。

表 3: 学習データの比較 (diphone モデル)

学習データ	文認識率
タスク独立 ((a) のみ)	58.2%
タスク依存 ((a)+(b))	79.4%

最後に diphone を単位とした場合の単語間接続法について比較した。学習にはタスクのデータ (b) も用いた。ガウス分布混合数は 4 で、マルチエントリ辞書を用いた。結果を表4に示す。表中「文認識率 +」は冠詞や単数複数の誤りなどささいな誤りを正解としたときの文認識率である。

表 4: diphone モデルでの単語間接続法の比較

単語間接続	文認識率	文認識率 +
音素モデル	75.9%	84.9%
diphone 平均化	74.9%	85.4%
diphone モデル	79.4%	87.9%

認識単位として diphone を用いることにより認識率が向上し、単語間を diphone モデルで接続した場合ささいな誤りを除くと 87.9% の文認識率が得られた。

また、単語間の接続に音素モデルあるいは diphone を平均化したモデルを用いると性能は低下するが、低下の度合は比較的小さかった。認識処理に要する時間は、diphone を用いた場合の約 1/4 強であった。今回の実験では、diphone の平均化により単語間を接続する方法と diphone とは独立に学習した音素モデルで接続する方法とで性能上の差は確認できなかったが、前者の方が diphone モデルとの接続部での連続性が保たれる、すでにできている diphone から容易に作成できる、などの点で有利な方法であると考えている。

5. むすび

混合ガウス分布 HMM にもとづく英語不特定話者連続音声認識システムを試作した。認識単位として diphone を用い、単語間接続部にも後続単語に応じた diphone モデルを挿入した。267 単語のタスクで評価実験を行ない、本方式の有効性を確認した。

今回はタスクのデータを用いて学習を行なったが、タスクに依存しない学習でも同等以上の性能を得るためにさらに認識単位の設定や学習法を含めた基本方式の改良が必要である。

謝辞

日頃ご指導いただく亘理部長をはじめとするメディアテクノロジ研究部の諸氏に感謝します。

参考文献

[1] 磯谷他、音学講論、1-8-19 (1990.9).
[2] 磯谷他、音学講論、3-5-13 (1991.3).
[3] T.Watanabe, et al., "An Experimental Automatic Interpretation System INTERTALKER" (本予稿集)
[4] K.F.Lee, et al., ICASSP89, S9-3 (1989.5).
[5] 服部他、信学技報、SP89-15 (1989.6).
[6] 渡辺他、音学講論、3-5-19 (1991.3).