論文 / 著書情報
Article / Book Information

論題(和文)	連続数字音声における音響モデル学習法の検討
Title(English)	
著者(和文)	植本 尚子, 松岡 達雄, 松井 知子, 古井 貞熙
Authors(English)	SADAOKI FURUI
出典(和文)	1995年日本音響学会春季講演論文集, , No. 1-Q-13, pp. 121-122
Citation(English)	, , No. 1-Q-13, pp. 121-122
発行日 / Pub. date	1995, 3

1—Q—13

連続数字音声における音響モデル学習法の検討 *

⊙ 植本尚子[1]　松岡達雄[2]　松井知子[2]　古井貞熙[1,2] †
（東京工業大学[1]、NTT ヒューマンインタフェース研究所[2]）

1　はじめに

連続数字音声認識はさまざまな応用の考えられる技術であるが、応用においては数字を1桁でも誤れば意味をなさないので極めて高い認識率が必要である。連続数字の場合、通常、任意の数字の連続が許されるため、連続音声認識において重要な役割を果たす文法的拘束が有効に使えない。したがって、精度の高い音響モデルが重要である。本報告では、Tied-mixture HMM をベースとした不特定話者連続数字認識のための高精度な音響モデルの学習法について報告する。電子協音声データベースを用いた4桁の連続数字認識において、音素文脈を考慮したサブワードモデルを用い、さらに、“誤り最小化原理”を適用した学習を行なうことにより高い認識率を達成することができた。

2　音響モデル学習法

2.1　サブワードモデル

連続発声された音声では、調音結合の影響により孤立発声の音声より認識率が低下する。これに対して、音素文脈を考慮した音響モデルを用いるのが有効であることが知られている。本報告では連続数字認識への音素文脈依存モデルの導入を提案する。提案するモデルは、単語(数字)を語頭(head)、語中(body)、語尾(tail)の3つの部分に分割したものである[1, 2]。以降、単語(数字)ごとに音響モデルを設定したものを単語モデル、一数字を3つの部分に分割した音響モデルをサブワードモデルと呼ぶ。

本報告で提案するサブワードモデルについて説明する。サブワードモデルでは、図1にあるように各数字の body を各一つづつ、head と tail を複数設定する。詳細なモデル化としては、各数字に対してそれぞれ先行、後続する数字に応じて10ずつ head と tail を設定することが考えられるが、10数字に対して音響モデルの数が210となり、その学習のためには1000種類(コンテキスト)の数字列が必要となる。各コンテキストについて数十の学習サンプルが必要としても数万サンプルという多量の学習データが必要となる。また、モデル数が多いと学習時間のみならず認識時の処理時間も増加する。ここでは、学習データが限られていることから調音結合の様子が近いものについては同じ音響モデルを用いるようにグループ化した。

グループ化は以下のように行なった。母音については /u/ と /o/ を同じグループに、子音については調音様式にしたがって破裂音、摩擦音、弾音、鼻音に分類した。破裂音や摩擦音が連続した場合や同じ母音の連続した場合、鼻音「ん」と鼻音が連続した場合については、調音結合が著しく起こるため、このような音素文脈に対してはそれぞれ固有のモデルを設けた。つまり、“roku” - “kyu” という文脈に対して、“kyu” にだけに先行する “roku” の tail、“roku” にだけ後続する “kyu” の head を設ける(図1参照)。この他ヒューリスティックなルールも加えサブワードモデルを設定した結果、数字あたりの平均の head と tail の数はどちらも約3となり、head の数は最小で1、最大で5、tail の数は最小で1、最大で6となった。サブワードモデルの数は全部で72となった(head は29、tail は33)。

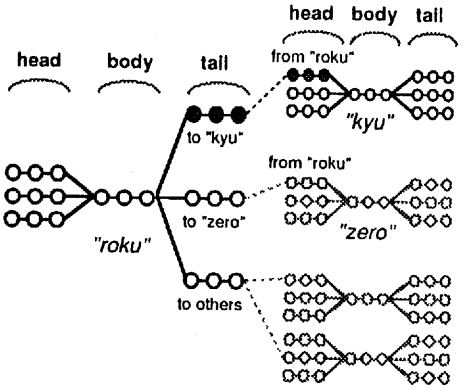

図1　サブワードモデル

2.2　誤り最小化原理

ここまで述べてきた音響モデルは、最尤推定に基づいて学習することを前提としてきた。つまり、音響モデルは学習データに対して尤度最大になるように学習されている。これに対して、識別的な学習によりクラス間の識別力がより高くなるように音響モデルを学習する方法が提案されている[3, 4]。

観測ベクトル X_t に対するクラス i のモデルの対数尤度を識別関数として定義する。

$$g_i(X_t, \Lambda) = \log(\mathcal{L}_i(X_t)) \qquad (1)$$

ただし、Λ は音響モデルのパラメータセットである。この識別関数に基づき、識別誤り関数を以下のよう

*Acoustic modeling for connected digit speech recognition

† By Naoko Uemoto[1], Tatsuo Matsuoka[2], Tomoko Matsui[2] and Sadaoki Furui[1,2]
(Tokyo Institute of Technology[1], NTT Human Interface Laboratories[2])

に設定する。

$$d(X_t,\Lambda) = -g_k(X_t,\Lambda) + G_k(X_t,\Lambda) \quad (2)$$

$$G_k(X_t,\Lambda) = \log\left\{\frac{1}{N-1}\sum_{j,j\neq k}^{N} exp(\eta g_j(X_t,\Lambda))\right\}^{\frac{1}{\eta}} \quad (3)$$

ここで、N は観測ベクトル X_t に対して考慮する候補の数、その観測に対する正解のクラスは k、k 以外の $(N-1)$ 個の誤りのクラスは j で与えられる。実験では閾値を設けて N が平均 5 になるようにした。損失関数はシグモイド関数を用いて次のように表される。

$$l(d(X_t,\Lambda)) = \frac{1}{1+exp(-a(d(X_t,\Lambda)+b))} \quad (4)$$

正解クラスのモデルの対数尤度がそれ以外のものに対して大きいほど、$l(d(X_t,\Lambda))$ は小さくなる。これは、誤認識を起こしにくいことを表している。学習データ全体に対する損失関数

$$L(\Lambda) = \frac{1}{T}\sum_{X_t}^{T} l(d(X_t,\Lambda)) \quad (5)$$

を最小化することにより、学習データに対して最も誤認識が起こりにくいようなモデルパラメータが求められることになる。音響モデルのパラメータは、次式によって更新される。

$$\Lambda_{n+1} = \Lambda_n - \epsilon_n U_n \delta L(X_t,\Lambda) \quad (6)$$

ここでは、共分散行列として対角成分のみを持つ HMM を用いるので U_n は単位行列とする。

3 実験

3.1 実験方法

HMM の状態数は各音素に 3 状態を割り当て、数字あたり平均 9 状態となった。単語モデルは 4 桁数字を 4 等分した学習データで初期モデルを作成し、そのモデルを連結学習して求めた。サブワードモデルは、上記で求めた単語モデルの状態を 3 等分したものを初期モデルとして、音素文脈依存の制限をつけた連結学習により求めた。学習には、単語モデル、サブワードモデルともに Baum-Welch アルゴリズムを用いた。

識別学習では、始め、全学習データに対する損失を求め、一括してパラメータ更新を行なった。しかし、損失の収束が悪かったため、学習データをいくつかのブロックに分割し、ブロック毎にパラメータを更新した。全学習データに対して一括でパラメータ更新した場合に損失の収束が悪いのは、個々の誤りに対して有効な修正の方向が異なるため、一括してパラメータを更新すると修正量が相殺されてしまうためではないかと考えられる。いわゆるシーケンシャル学習 (学習データ 1 つづつに対してパラメータを更新) を行なってもよいが学習に要する時間の点で、ある程度まとめてパラメータを更新した方が有利であるため 35 種類の 4 桁数字を 7 種類づつ 5 つのブロックに分けて学習を行なった。

サブワードモデルは最尤推定された初期モデルの段階ですでにかなり精度の高いモデルが得られていると考えられるため、誤認識された種類 (コンテキスト) の数字列だけを学習に用いて、さらに、正解モデルの尤度が候補 N 中で最大でない場合、すなわち、誤って認識された場合にのみパラメータを更新した。学習データ中の各音響モデルの出現回数には偏りがある。そこで、式 (5) においては修正量が出現回数で正規化されるようにしている。閾値で枝刈りしているため、候補が挙がらなかった場合や、候補中に正解が含まれなかった場合には式 (5) には計算されない。

アルゴリズムの収束を速めるため ϵ_n の値を複数設け、誤差が最も減少するような ϵ_n を採用しながらパラメータ更新を進めた。更新したモデルパラメータは、重み係数と平均値である。また、アルゴリズムが安定して収束するように、まずケプストラムについて収束するまでパラメータの更新を行ない、後に、デルタケプストラムについてパラメータを更新した。

学習データが少量であるためモデルパラメータの更新の際には修正ベクトルを平滑化した [5]。

3.2 実験結果

本実験において使用したデータは、電子協日本語共通音声データベース中の男性 75 名が各 4 回発声した 35 種類の 4 桁数字である。その内学習用に 70 名分、評価用に 5 名分用いた。20kHz から 12kHz にダウンサンプリングし、32ms 幅のハミング窓を掛けたのち、8ms ごとに 16 次のケプストラムとデルタケプストラムを LPC 分析により求めた。音響モデルは、256 の分布 (平均値、分散) を持つ left-to-right の Tied-mixture HMM である。

音響モデルをサブワードモデルで実現した場合、1.0% の誤り率を得た。これは、単語単位の音響モデルでの誤り率 3.4% と比較すると 70.9% の誤り率の低下である。この結果より、本実験で設計したサブワードモデルは、連続数字音声における音素文脈による調音結合の変化を、単語モデルよりも的確に反映できていると言える。

さらに、上記のサブワードモデルを誤り最小化原理を適用して再学習し音響モデルを作成し、同じ評価データにより認識実験を行なったところ、99.1% の認識率を得た。つまり、もとの最尤推定によって求められたサブワードモデルを使用した場合よりも 14.0% 誤り率が低下した。

4 まとめ

連続数字認識において、数字を head, body, tail に分割した音素文脈を考慮した音響モデルを用いることで、70.9% の認識誤りの削減ができた。さらに、誤り最小化原理を適用して学習を行なうことにより、14.0% の認識誤りの削減ができ、桁数既知での 4 桁連続数字認識において、99.1% の認識率を達成できた。今後は、電話音声に対して高い精度の連続数字認識技術の検討を進めたい。

参考文献

[1] E.Giachin, C.-H.Lee, R.Pieraccini, and L.R.Rabiner, CSELT Technical reports, Vol.XIX, No.1, Feb 1991

[2] E.P.Giachin, C.-H.Lee, L.R.Rabiner, and A.E.Rosenberg, Computer Speech and Language, Vol.6, pp.197-213, 1992

[3] W.Chou, C.-H.Lee, and B.-H.Juang, Proc.ICASSP-94, pp.II-652-655, 1994

[4] W.Chou, B.-H.Juang, and C.H.Lee, Proc.ICASSP-92, pp.I-473-476, 1992

[5] 大倉計美, 杉山雅英, 嵯峨山茂樹, 信学技報, SP92-16, pp.23-28, 1992