

論文 / 著書情報
Article / Book Information

論題(和文)	N-bestに基づく話者適応化法の検討
Title(English)	A study of N-best based speaker adaptation method
著者(和文)	松井知子, 橋本直己, 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. 2-6-18, pp. 79-80
Citation(English)	, Vol. , No. 2-6-18, pp. 79-80
発行日 / Pub. date	1997, 3

○ 松井知子[†] △ 橋本直己[‡] 松岡達雄[†] 古井貞照^{††}
 († NTT ヒューマンインタフェース研究所 ‡ 東京工業大学)

1 まえがき

連続 HMM に基づく音声認識システムにおいて、不特定話者用音韻 HMM を即時に認識対象話者に適応化することで、認識性能を改善しようとする試みがなされてきた。しかし、従来の即時型の話者適応化では、不特定話者用音韻 HMM から推定した単語系列を用いる必要があるために、不特定話者用音韻 HMM に対する性能が低い話者セットに対しては、単語系列が正しく推定できず、話者適応化の効果が必ずしも得られなかった。そのために、筆者らは不特定話者用音韻 HMM に対する性能が低い話者についても有効な、N-best に基づく即時型話者適応化法を提案し、連続数字音声認識実験においてその有効性を確かめた [1]。本論文では、不特定話者用音韻 HMM に対する性能が低い話者セットに対するこの方法の効果を一層向上させるために、N-best による平滑化推定法について検討する。さらに、この方法で必要とする計算量を、認識性能を劣化させることなく削減するために、発声照合を行うことを考える。また、不特定話者用音韻 HMM のパラメータを認識対象話者のパラメータへ写像するモデル変換関数として、各混合ガウス分布の平均値ベクトルのバイアスモデル、線形回帰モデルを用いた場合を比較し、話者適応化に有効なモデル変換関数について検討する。

2 N-best による平滑化推定法

前回の N-best に基づく話者適応化法 [1] では、適応化を階層的に繰り返し行う。各繰り返しにおいて、N 個の各単語系列について適応化を行い、最も高い尤度を示す系列を選択し、その系列に従って適応化した音韻 HMM を次の繰り返しで順次用いていく。その時、はじめの方の繰り返しでは適応化がまだ十分ではないために、正解の単語系列が必ずしも最も高い尤度を示すとは限らない。前回の方法には、不特定話者用音韻 HMM を用いた時に正解の単語系列が下位に認識される場合には、適応化しても正解が 1 位にならないことが多いという問題があった。そのために、ここでは適応化後に最も高い尤度を示す系列だけではなく、下位の系列に関しても考慮する平滑化推定法について検討する。

この平滑化推定法では、前回の方法と同じく、まず入力音声の N-best デコーディングによって得られた N 個の単語系列 $\{W_1, W_2, \dots, W_N\}$ のそれぞれについて、不特定話者用音韻 HMM のパラメータ θ を認識対象話者のパラメータへ写像するモデル変換関数 $G_{\eta}(\theta)$ の中の変換パラメータ η_n を次式に従って求める。

$$\eta_n = \arg \max_{\eta} f(X|W_n, \eta, \theta)g(\theta)$$

ここで、 X は入力音声、 $f(\cdot)$ は尤度関数、 $g(\cdot)$ は事前確率密度関数を表す。次に認識対象話者に適応化した音

韻 HMM のパラメータ θ_n を、この η_n を用いて次式のように計算する。

$$\theta_n \leftarrow G_{\eta_n}(\theta)$$

そして、新たに信頼尺度 C_n を例えば、次のように定義する。

$$C_n = \left(\frac{L_n}{L_{\max}} \right)^{\alpha}$$

ここで、 L_n は適応化後に単語系列 W_n が示す尤度、 $L_{\max}$ は適応化後に各単語系列が示す尤度の中で最大のもの、 α は実験的なパラメータである。この信頼尺度を各系列について計算し、音韻 HMM のパラメータを次式に従って求める。

$$\hat{\theta} \leftarrow \frac{\sum_{n=1}^N C_n \theta_n}{\sum_{n=1}^N C_n}$$

(この音韻 HMM は次の繰り返しで順次用いていく。)

3 実験

3.1 条件

ここでは、連続数字音声認識実験において本方法を評価する。学習では、NTT が作成した電話音声データベースから、男性話者 177 名が発声した計 24,194 個の連続数字 (1、2、または 4 桁数字) を用いて、不特定話者用 HMM を計 113 個のサブワード HMM (54 語頭、13 語中、46 語尾の 4 混合ガウス分布 HMM) として文脈依存に生成した [2]。

適応化及び認識では、NTT データ通信 (株) が作成した電話音声データベース [3] から、男性話者 50 名が発声した 4 桁数字 (6 発声/話者) を用いた。本方法は、評価話者を不特定話者用 HMM を用いた場合に、誤って認識される 4 桁数字が 2 個以上の話者 (難しい話者) と、1 個もしくは 0 個の話者 (易しい話者) に分けた場合の 4 桁数字の単語系列認識率によって評価した。本実験ではデフォルトとして 10-best の単語系列を用いたが、10-best 内に正解が含まれる確率は 97.3 % であった。また、モデル変換関数 $G_{\eta}(\cdot)$ としては、各混合ガウス分布の平均値ベクトルのバイアスモデルを用い、16 個の平均値バイアスを 5 回の繰り返しで階層的に推定した。(はじめの繰り返しで 1 個のバイアスを、次の繰り返しで 2 個、3 回目で 4 個、4 回目で 8 個、5 回目で 16 個のバイアスを推定した。) さらに、信頼尺度のパラメータ α は 3.0 に設定した。

特徴パラメータとして、フレーム長 32ms、フレーム周期 8ms、LPC 分析次数 12 でケプストラム、 d ケプストラムを抽出した。各発声ごとに、ケプストラム平均値正規化を行った。

*A study of N-best based speaker adaptation method.

By Tomoko Matsui[†], Naoki Hashimoto[‡], Tatsuo Matsuoka[†], and Sadaaki Furui^{††} (†NTT Labs., ‡Tokyo Inst. of Tech.)

表 1: 単語系列認識率% ([]: 誤り削減率%)

方法	難しい話者	易しい話者	平均
ベースライン	59.3	93.9	87.7
前回の方法	70.4 [27.3]	92.7	88.7
平滑化推定法	74.1 [36.4]	92.7	89.3

3.2 結果

表1に、不特定話者用 HMM をそのまま用いた場合（ベースライン）、前回の方法で適応化した場合、この平滑化推定法によって適応化した場合の単語系列認識率と誤り削減率を示す。この結果から、平滑化推定法は難しい話者に対して有効であることがわかる。

4 尤度比に基づく発声照合

本方法では、繰り返しの度に N 個の単語系列に対して適応化を行うので、比較的大きな計算量を必要とする。ここでは、本方法で必要とする計算量を認識性能を劣化させることなく、平均的に削減するために、発声照合を行うことを考える。予備実験から、不特定話者用 HMM を用いた時には誤って認識されるが、本方法を適用することによって正しく認識される4桁数字は、不特定話者用 HMM を用いた時に、1 位に認識される単語系列の尤度 l_1 と 2 位の尤度 l_2 との対数尤度比 $\log(l_1) - \log(l_2)$ が小さいことがわかった。つまり、対数尤度比が比較的大きな4桁数字は、もともと不特定話者用 HMM によって正しく認識されているか、本方法を適用しても効果がない。そのために、この対数尤度比に予め適当なしきい値を設定し、そのしきい値に基づいて発声照合を行えば、本方法を適用する有効性を判定でき、本方法に必要な計算量を平均的に大きく削減できる。

表2に平滑化推定法において、上記の発声照合（しきい値は事後的に設定）を行った場合の、適応化に必要な計算時間（実時間）の平均削減率、平均単語系列認識率を示す。発声照合を行うことにより、計算時間が削減されるだけでなく、認識性能も改善した。平均認識率は 89.3 % から 90.0 % に（ベースラインと比較した時の誤り削減率は 13.0 % から 18.7 % に）向上した。なお、しきい値の事前決定は今後の課題である。

5 バイアス vs. 線形回帰モデル

話者適応化に有効なモデル変換関数について検討するために、各混合ガウス分布の平均値ベクトルの線形回帰モデル（maximum likelihood linear regression; MLLR）[4] による N -best に基づく適応化実験を行った。

表3に、発声照合を行う／行わない場合の線形回帰モデルによる単語系列認識率を示す。なお、回帰行列（対角行列）は各発声について1個、1回の推定で求めた。線形回帰モデルによる方法は、バイアスモデルによる平滑化推定法と比べた場合、全話者の平均では同等の性能を示したが、本論文では難しい話者について性能を向上させることが重要であると考えられる。本論文の実験では、難しい話者についてはバイアスモデルによる平滑

表 2: 発声照合を用いた場合の計算時間の削減率%と単語系列認識率%

削減率 (計算)	単語系列認識率		
	難しい話者	易しい話者	平均
80.1	74.1	93.5	90.0

表 3: 線形回帰モデルによる単語系列認識率% ([]: 誤り削減率%)

発声照合	難しい話者	易しい話者	平均
なし	70.4 [27.3]	93.9	89.7
あり	72.2 [31.7]	93.9	90.0

化推定法が最も有効であった。ただし、線形回帰モデルの方が推定すべきパラメータ数は少ない。

6 むすび

本論文では、不特定話者用音韻 HMM に基づく音声認識システムのための、 N -best に基づく即時型話者適応化法について検討した。不特定話者用 HMM に対する性能が低い話者に対する効果を一層向上させるために、 N -best に基づく平滑化推定法を提案した。さらに、発声照合を行うことにより、本方法で必要とする計算量を大幅に削減できることを示した。また、平均値バイアスモデル、線形回帰モデルについての比較実験を行い、線形回帰モデルの方が少ないパラメータ数で同等の性能が得られるが、不特定話者用 HMM に対する性能が低い話者に対しては、平均値バイアスモデルを平滑化推定法とともに用いる方が有効であることを示した。

謝辞

本実験において、データベースを使用することを許可して頂いた、NTT データ通信株式会社に感謝致します。また、本実験を支援して下さい、東京工業大学の波多野健氏に感謝致します。

参考文献

- [1] 松井知子、古井貞照、音声認識のための N -best に基づく話者適応化、日本音響学会秋季研究発表会、岡山、3-3-16、pp. 117-118、1996.
- [2] T. Matsuo, N. Uemoto, T. Matsui and S. Furui, *Elaborate acoustic modeling for Japanese connected digit recognition*, Proc. IEEE Automatic Speech Recognition Workshop, Snowbird, pp. 169-170, 1995.
- [3] M. Morishima, T. Isobe and K. Murakami, *Telephone speech database and the experiment using CDHMM*, Proc. Acoustical Society of Japan, Fall Meeting, 2-8-8, 1994.
- [4] C.J. Leggetter and P.C. Woodland, *Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models*, Computer Speech and Language, Vol. 9, pp. 171-185, 1995.