TR-I-0265

話者選択手法を用いた音声認識の基礎検討

A study of speech recognition using speaker selection approach

大倉計美　　杉山雅英
Kazumi OHKURA　Masahide SUGIYAMA

1992.6.10

概要

近年、不特定の話者が発声した音声を高精度に認識するために、話者適応手法、不特定話者認識、話者選択手法等の研究が活発になされている。本報告では、これらの研究を行なうための基礎実験として、ATR 音声データベースに属する話者 16 名を用いて、話者間の関係を連続分布型 HMM を用いた 6 音素認識実験により評価した。また、話者選択手法の可能性を不特定話者モデルとの比較により検討した。

ATR 自動翻訳電話研究所
ATR Interpreting Telephony Reserch Laboratories

目次

1 はじめに

近年、不特定の話者が発声した音声を高精度に認識するために、話者適応手法、不特定話者認識、話者選択手法等の研究が活発になされている。本報告では、これらの研究を行なうための基礎実験として、ATR 音声データベースに属する話者 16 名を用いて、話者間の関係を連続分布型 HMM を用いた 6 音素 (/b,d,g,m,n,N/) 認識実験により評価した。また、話者選択手法の可能性を不特定話者モデルとの比較により検討する。以下の節は、次の様に構成される。第 2 節では実験条件を述べる。第 3 節では不特定話者モデルと話者選択を用いた認識手法との基礎検討として、標準話者と入力話者の音素認識率の関係を調べる。第 4 節では、不特定話者モデルの性能を評価する。第 5 節では不特定話者モデルと話者選択を用いた認識手法との比較を行なう。

2 実験条件

分析条件は、標本化周波数 12kHz、高域強調 $1-0.97z^{-1}$、フレーム長 21.3ms、フレームシフト幅 3ms、LPC 分析 14 次とした。特徴量は、ケプストラム (CEP)、ΔCEP および Δ 対数パワーを用いた。CEP および ΔCEP 計算打ち切り次数は 16 次とした。ΔCEP および Δ 対数パワーは、前後 8 フレームより、回帰係数を計算した。HMM 学習資料には、重要語 5240 単語の偶数番目の単語中より切り出した音素 /b,d,g,m,n,N/、認識資料には奇数番目の単語中より切り出した 6 音素 /b,d,g,m,n,N/ を用いた。/b,d,g/ は語頭と語中の 2 種類のモデルを用いた。HMM を作成した話者 (標準話者) と入力話者の詳細を表 1 に示す。HMM は連続分布型 (無相関正規分布)、4 状態 3 ループであり、各状態のアークはタイドアークとした。

表 1: 標準話者と入力話者

標準話者	男性 : MSH,MHT,MMS,MXM 女性 : FKN,FKS,FFS,FTK
入力話者	男性 : MAU,MMY,MNM,MTK 女性 : FSU,FMS,FYM,FYN

3 各標準話者と入力話者間の認識率の検討

本節では、不特定話者モデルと話者選択手法の比較を行なうための基礎検討として、各標準話者と入力話者間の認識率の関係を調べる。表 1 に示した標準話者の音声資料を用いて各話者の HMM(混合分布数は 3) を作成し、入力話者の音声を認識する。認識結果を表 2 に示す。表 2 より、同性間の認識率は良いことと、同性間の認識においても標準話者の選び方によりかなり認識率にバラツキのあることが分かる。このように入力話者に対して認識率の低い標準話者を含めて学習した不特定話者モデルがどの程度の認識性能を示すのか、また、音声特徴の近い話者を選択した場合のモデルはどの程度の認識性能を示すのかを、次節以降検討していく。

表 2: 特定話者の認識率および入力話者の認識率

		MSH	MHT	MMS	MXM
	特定話者	92.91%	95.30%	94.13%	93.25%
入	MAU	70.99%	70.65%	66.92%	53.86%
	MNM	61.77%	57.60%	61.51%	50.81%
力	MMY	75.56%	73.08%	66.07%	58.46%
	MTK	66.50%	46.07%	60.51%	58.80%
話	FSU	42.56%	40.17%	32.99%	23.59%
	FMS	44.36%	47.09%	35.04%	31.79%
者	FYM	33.16%	32.22%	30.68%	22.39%
	FYN	37.12%	24.74%	39.42%	17.75%

		FKN	FKS	FFS	FTK
	特定話者	95.30%	94.45%	96.92%	94.62%
入	MAU	36.64%	33.16%	21.63%	34.52%
	MNM	29.31%	36.36%	27.87%	35.51%
力	MMY	42.22%	40.00%	34.10%	39.49%
	MTK	34.10%	38.72%	30.68%	38.12%
話	FSU	75.38%	72.31%	59.57%	77.09%
	FMS	52.99%	63.33%	54.02%	65.04%
者	FYM	66.07%	64.70%	44.27%	54.87%
	FYN	64.42%	65.87%	45.90%	74.91%

4 不特定話者モデルの評価

本節では、不特定話者 HMM の評価実験について述べる。不特定話者モデルは、8 名の標準話者 (MSH,MHT,MMS,MXM,FKN,FKS,FFS,FTK) から作成した。混合分布数は 9 とした。表 3に、不特定話者 HMM を用いて入力話者の音声を認識した場合の認識率と、比較のため 6 音素の認識率が最も高かった標準話者の認識率を示す。表 3より不特定話者モデルは、6 音素の認識率が最も高かった標準話者の音素認識率よりも高い認識率を示すことが分かる。この結果より、話者 1 名を選択するよりも、入力話者に対して認識率の低い話者が含まれてはいるが、多数の話者から学習した不特定話者モデルを用いる方が認識性能が良いことが分かる。

5 不特定話者モデルと話者選択を用いたモデルとの比較

本節では不特定話者 HMM との比較により、話者選択により複数話者を選択した場合の認識性能の検討を行なう。

話者選択方法を以下に述べる。ここでは、2 方法の選択手法を用いた。第 1 の方法は、従来から用いられている実際的な方法である “VQ 歪みによる話者選択 ” つまり、話者選択用の単語を各標準話者のコードブックにより量子化した場合の量子化歪みにより話者を選択する方法である。第 2 の方法は、本実験の認識対象である /b,d,g,m,n,N/ の各話者の認識結果より認

識率の高かった話者を選択する方法である。この第2の方法は、実際的ではないが最も良い認識性能を実現できると考えられ、認識性能の上限を調べるために行なった。本2方法の詳細を以下に示す。

1. VQ歪みによる話者選択とモデルの作成方法

   8名の標準話者各々のバランス単語(先頭から100単語)を用いて作成したケプストラムのコードブック(サイズ256)を用いて、入力話者(MAU,FSU)の発声した5単語(バランス単語の101~105番目の単語を使用)を量子化した場合の量子化歪み(ケプストラムのユークリッド距離)により話者選択を行った。話者は量子化歪みの小さかった1名および3名を選択した。認識には選択された3名のHMMを1つに合成したHMMを用いた。合成方法は、遷移確率については量子化歪みの最も小さかった話者のものを使用し、分布(平均・分散)は各話者のHMMをそのまま用いる。分岐確率は話者選択時に使用した歪みより計算されるファジイ級関数値にしたがい重みをつけた。つまり、話者A、B、CのHMMにおいて、状態iから状態jへの遷移によってシンボルベクトル(y)が出力される確率がそれぞれ、$b_{ij}^{A}(\mathrm{y})$、$b_{ij}^{B}(\mathrm{y})$、$b_{ij}^{C}(\mathrm{y})$である場合、合成したHMMにおける出現確率($b_{ij}(\mathrm{y})$)は次式で表される。

$$b_{ij}(\mathrm{y}) = \mu_A b_{ij}^{A}(\mathrm{y}) + \mu_B b_{ij}^{B}(\mathrm{y}) + \mu_C b_{ij}^{C}(\mathrm{y})$$

   ここで、μ_A, μ_B, μ_Cは話者選択時に使用した歪みより計算されるファジイ級関数値を表す。本実験ではファジネスを1.3とした。

2. 認識率による話者選択手法とモデルの作成方法

   第3節で行なった6音素認識実験において認識率の高かった話者を選択する。この場合のHMMの作成方法は、遷移確率については認識率の最も高かった話者のものを使用し、分布(平均・分散)は各話者のHMMをそのまま用いる。各話者の分岐確率は次式に従い、合計が1になるように正規化した。

$$b_{ij}(\mathrm{y}) = \mu_A b_{ij}^{A}(\mathrm{y}) + \mu_B b_{ij}^{B}(\mathrm{y}) + \mu_C b_{ij}^{C}(\mathrm{y})$$

$$\mu_A = \mu_B = \mu_C$$

$$\mu_A + \mu_B + \mu_C = 1$$

結果を表4に示す。表4には、話者選択を行なわずに標準話者8名を全て用いてHMMを合成した(合成方法には、上記の”認識率による話者選択手法とモデル”を使用)場合の認識率も合わせて示した。表4の結果より、以下のことが分かる。

1. 話者の静的な特徴のみを用いたVQ歪みによる話者選択を用いた場合の認識率は、不特定話者モデルを用いた場合の認識率よりも低い。

2. 入力話者に対して /b,d,g,m,n,N/ の認識率が高かった標準話者を 3 名選択した場合は、不特定話者モデルを用いた場合の認識率よりも高い認識率を示す。

3. 話者選択を行なわずに標準話者 8 名から HMM を合成した場合は、不特定話者モデルよりも低い認識率を示す。

話者選択手法としては、認識率で話者を選択する方法が VQ 歪みによる選択手法よりも良い結果を示しているが、音素認識率による話者選択手法は現実的ではないことと、VQ 歪みによる話者選択では選択用の 5 単語中 6 音素全ては含まれていない (/d/ が 1 回、/N/ が 4 回出現するだけで、その他の音素 /b,g,m,n/ が 1 回も出現しない) ことを考慮すると、/b,d,g,m,n,N/ 認識の枠組では、どちらが優れているとは言えない。しかし以上の結果より、複数話者を選択することの有効性と、うまく話者を選択すれば不特定話者モデルよりも認識性能の良いモデルを作成できることが明らかになった。また、選択された複数の話者のモデルを 1 つに合成する方法により、標準話者のモデルを全て使用するよりも認識性能が良く、混合数の少ない効率の良いモデルを作れることが明らかになった。今後は、少数単語から単語内に含まれない音素に対する特徴をも推定でき、かつ話者の動的な特徴をも考慮できる話者選択手法の研究が必要である。

表 3: 不特定話者モデルの認識率

	不特定話者モデル	認識率が最も良かった 標準話者の HMM を用いた場合
MAU	74.98%	70.99% (MSH)
MMY	76.41%	75.56% (MSH)
MNM	66.02%	61.77% (MSH)
MTK	70.00%	66.50% (MSH)
FSU	81.79%	77.09% (FTK)
FMS	73.16%	65.04% (FTK)
FYM	66.58%	66.07% (FKN)
FYN	77.13%	74.91% (FYN)

() 内は認識率が最も良かった標準話者名

表 4: 不特定話者モデルと話者選択手法の認識率

	不特定話者モデル	量子化歪みによる選択1名	量子化歪みによる選択3名	認識率による選択1名	認識率による選択3名	標準話者8名全てを使用
MAU	74.98%	66.92% (MMS)	70.99% (MMS) (MXM) (MSH)	70.99% (MSH)	76.59% (MSH) (MHT) (MMS)	67.35%
FSU	81.79%	77.09% (FTK)	75.98% (FTK) (FFS) (MHT)	77.09% (FTK)	83.33% (FTK) (FKN) (FKS)	80.60%

() は、選択された話者名を示す

6 あとがき

話者適応手法、不特定話者認識、話者選択手法等の研究を行なうための基礎実験として、ATR 音声データベースに属する話者 1 6名を用いて、話者間の関係を連続分布型 HMM を用いた 6 音素 (/b,d,g,m,n,N/) 認識実験により評価した。結果は、同性間の認識率は高いことと、標準話者の選び方により認識率がかなり変化することが明らかになった。話者選択手法を用いた認識手法と不特定話者モデルとの比較により、うまく話者を選択すれば不特定話者モデルよりも高い認識率を実現できるモデルを作成できることを示した。また、選択された複数の話者のモデルを 1 つに合成する方法により、標準話者のモデルを全て使用するよりも認識性能が良く、効率の良いモデルを作れることを示した。今後は、少数単語から単語内に含まれない音素に対する特徴をも推定でき、かつ話者の動的な特徴をも考慮できる話者選択手法の研究が必要である。

謝辞

研究の機会を与えて頂いた、ATR 自動翻訳電話研究所樽松明社長に感謝いたします。また、熱心に討論頂く嵯峨山茂樹室長をはじめ、ATR の皆様に感謝いたします。

参考文献

[1] 花沢, 川端, 鹿野 : "Hidden Markov Model による音韻認識実験の結果," ATR テクニカルレポート, TR-I-0147, (Feb. 1990).

[2] 中村, 花沢, 鹿野 : " ベクトル量子化話者適応アルゴリズムの HMM 音韻認識による評価," 信学技報, SP88-106, pp.1-8 (1988).

[3] 杉山, 鹿野 : " 量子化歪み最小の原理に基づく母音標準パターンの教師なし学習法," 研究実用化報告, Vol.34, No.12, pp.1717-1725 (1985)

[4] 杉山, 好田：“認識結果を利用した母音標準パターンの教師なし話者適応化法,” 信学論 (D), Vol.J69-D, No.8, pp.1197-1204 (1986-08).

[5] 杉山：“母音の教師なし話者適応における各種の方法の比較,” 信学論 (D), Vol.J 70-D, No.5, pp.958-963 (1987-05).