バギングCAN2とギブス分布を用いる多段話者認識の性能向上

九州工業大学　○溝部祐太　植木琢也　黒木秀一

Performance improement of multistep speaker recognition using bagging CAN2 and Gibbs-distribution-based-extended Bayesian inference

Yuta Mizobe, Takuya Ueki and Shuichi Kurogi
Kyushu Institute of technology

Abstract: This paper presents a method of multistep speaker identification using Gibbs-distribution-based extended Bayesian inference (GEBI) for rejecting unregistered speaker. The method is developed for our speaker recognition system which utilizes competitive associative nets (CAN2s) for learning piecewise linear approximation of nonlinear speech signal to extract feature vectors of pole distribution from piecewise linear coefficients reflecting nonlinear and time-varying vocal tract of the speaker. In this paper, we focus on the problem of Bayesian inference (BI) in multistep identification for rejecting unregistered speaker and introduce GEBI to solve the problem. The effectiveness of the present method is shown by means of experiments using real speech signals.

1. まえがき

従来の話者認識の研究手法としてHMM法などが実用的であるため，一般的に用いられているが，それらは線形モデルを枠組みとして理論が作られている．しかし，近年，音声信号が非線形であるという多くの報告があり，非線形性を扱う手法により，より高い認識性能を実現できるのではないかと考えられる．

そこで，非線形関数を区分的に線形近似する能力を持つニューラルネットの1つとして提案されている競合連想ネット(CAN2:Competitive Associative Net 2) による手法の開発が進められた[1-3]．先行研究では，このCAN2にアンサンブル学習の1つであるバギング法を適用したバギングCAN2を用いて，音声時系列を性能よく再現した[4-6]．また，再現する際に抽出される音声時系列の極の分布の配置パターンを特徴パターンとして用いた話者認識は，高い認識率を得ている[7]．しかし，これまでの実験では，ベイズ推定を用いて行っていたために，識別すべき音声データの話者s が学習した話者集合S に含まれない場合に適応していなかった[8, 9]．そこで，ギブス分布を用いて多段認識を行うことで認識率の信頼性が向上するか検討する必要があると考えられる．

本稿では，この研究方法を話者認識に応用し，ギブス分布に基づく拡張ベイズ推定を用いた多段認識実験を行い，学習していない音声を入力したときにそれを判断できるかを検討する．

2. CAN2と単純ベイズ推定を用いる多段話者認識

2.1 話者認識の概要

CAN2 を用いた話者認識のブロック図をFig.1. に示す．一般に話者認識は音声をアナログ信号からディジタル信号へ変換する前処理，特徴量抽出，パターンマッチング，話者の決定の4工程で構成される．更に，話者認識には話者照合と話者識別に分類される．前者は，ある人物が本人の主張している通りの個人であるかを判断し，認証または棄却するもので，これは2クラスの分類問題と見なせる．一方，後者は，予め登録された話者の中で，入力された声が誰によるものかを特定するもので，こちらは多クラスの分類問題と見なすことができる．さらに，話者認識はテキスト依存型とテキスト独立型の2つに分類される．前者は，話者認識をする際に発声する単語や文章が予め登録されてあるものと同じでなくてはならないが，後者は発声する単語や文章が予め登録されてあるものと異なるものを用いる．

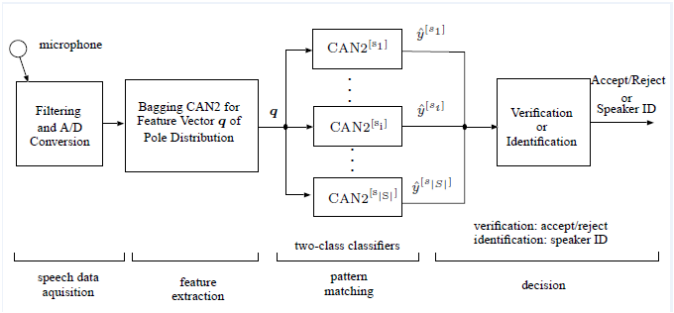

Fig.1. Speaker recognition system using the CAN2s

2.2 単一話者認識

本実験では，音声信号から得た極の分布を特徴量として用いることで話者認識を行う．話者 $s \in S\{s_i | i \in I_S\}$ の極分布を表す特徴ベクトル $\boldsymbol{q} = (q_1, q_2, \cdots, q_k)^T$ の集合を $Q^{[s]}$ とする．ただし，$I_S = \{1,2,\cdots,|S|\}$ である．非線形関数を区分的に線形近似する能力を持つニューラルネットの 1 つである CAN2 を用いる．

$$f^{[s]}(\boldsymbol{q}) = \begin{cases} 1, \text{ if } \boldsymbol{q} \in \mathrm{Q}^{[s]} \\ -1, \text{ otherwise} \end{cases} \tag{1}$$

話者 s を受理または棄却するための 2 クラスの分類器を $\mathrm{CAN2}^{[s]}$ とする．

まず訓練データ $\left(\boldsymbol{q}, f^{[s]}(\boldsymbol{q})\right), \boldsymbol{q} \in Q^{[s]}$ を用いて連続関数 $\hat{y}^{[s]} = \hat{f}^{[s]}(\boldsymbol{q})$ により $\mathrm{CAN2}^{[s]}$ を式(1)の関数のように近似させる．それにより単一話者照合の出力を次式のように 2 値化する．

$$v^{[s]} = \begin{cases} 1, & \text{if } \hat{y}^{[s]} = \hat{f}^{[s]}(\boldsymbol{q}) \geq y_\theta \\ -1, & \text{otherwise.} \end{cases} \tag{2}$$

すなわち，もし $v^{[s]} = 1$ ならば，話者を s_i とみなしてもよいということである．ここで，閾値 y_θ は後述する認識の性能を調節するためのものである．次式により与えられる最大検出よって単一話者識別を行う．

$$\mathrm{ID} = \underset{i \in I_S}{\operatorname{argmax}} \left\{ \hat{y}^{[s_i]} = \hat{f}^{[s_i]}(\boldsymbol{q}) \right\} \tag{3}$$

上記の話者識別値 ID を得ることで第 $|S|$ 番目の話者の音声信号を見分ける．

2.3 単純ベイズ推定を用いる多段話者識別

まず，訓練データセットより得られる 2 クラス分類器 $\mathrm{CAN2}^{[s_i]}$ の照合確率 $p(v^{[s_i]} | s)$ は次式のようになる．

$$p(v^{[s_i]} = 1 | s) = \frac{n(v^{[s_i]} = 1 | s)}{n(v^{[s_i]} = 1 | s) + n(v^{[s_i]} = -1 | s)} \tag{4}$$

$$p(v^{[s_i]} = -1 | s) = 1 - p(v^{[s_i]} = 1 | s) \tag{5}$$

$n(v^{[s_i]} = 1 | s)$ は話者 $s \in S$ の音声に対して $\mathrm{CAN2}^{[s_i]}$ が $v^{[s_i]} = 1$ と出力した数を示す．話者識別では，話者 $s \in S$ から得た特徴ベクトル $\boldsymbol{q}$ を入力した $\mathrm{CAN2}^{[s_i]}\left(s_i \in S\right)$ の出力ベクトル $\boldsymbol{v}^{[S]} = (v^{[s_1]}, \cdots, v^{[s_{|S|}]})$ の照合確率を次式のように定義する．

$$p(\boldsymbol{v}^{[S]} | s) = \prod_{s_i \in S} p(v^{[s_i]} | s) \tag{6}$$

なぜならば，2 つの確率 $p(v^{[s_i]} | s)$ と $p(v^{[s_j]} | s)$ は $i \neq j$ で独立していると考えられるからである．話者 s から得た $\boldsymbol{v}^{[s]}$ を $\boldsymbol{v}_{1:t}^{[s]} = \boldsymbol{v}_1^{[s]}, \cdots, \boldsymbol{v}_t^{[s]}$ として，次式よりベイズ推定を行う．

$$p_{\mathrm{I}}^{[\mathrm{Bys}]}(s | \boldsymbol{v}_{1:t}^{[s]}) = \frac{p_{\mathrm{I}}^{[\mathrm{Bys}]}(s | \boldsymbol{v}_{1:t-1}^{[s]}) p(\boldsymbol{v}_t^{[s]} | s)}{\Sigma_{s_i \in S} \, p_{\mathrm{I}}^{[\mathrm{Bys}]}(s_i | \boldsymbol{v}_{1:t-1}^{[s]}) \; p(\boldsymbol{v}_t^{[s]} | s_i)} \tag{7}$$

ここで，$p(\boldsymbol{v}_t^{[s]} | s, \boldsymbol{v}_{1:t-1}^{[s]}) = p(\boldsymbol{v}_t^{[s]} | s)$ で与えられる条件付き独立性を用いる．このように単純ベイズ推定を用いることにより計算を簡単に行うことができる．これは実際に多くのアプリケーションで効果を示している[10].

2.4 単純ベイズ推定の問題

ベイズ推定により求めた確率 $p_{\mathrm{I}}^{[\mathrm{Bys}]}(s | \boldsymbol{v}_{1:t}^{[s]})$ が閾値 $p_{\mathrm{I}\theta}$ よりも大きくなるとき，話者識別値 ID が求まり話者が決定する．しかし，この方法には問題があるので解析していく．ベイズ推定より

$$\begin{aligned} p_{\mathrm{I}}^{[\mathrm{Bys}]}(s | \boldsymbol{v}_{1:t}^{[s]}) &= \frac{p_{\mathrm{I}}^{[\mathrm{Bys}]}(s | \boldsymbol{v}_{1:0}^{[s]})}{Z_t} \prod_{k=1}^{t} p\left(\boldsymbol{v}_k^{[s]} | s\right) \\ &= \frac{p_{\mathrm{I}}^{[\mathrm{Bys}]}(s | \boldsymbol{v}_{1:0}^{[s]})}{Z_t} \exp\left(-t \, \tilde{L}_{1:t}^{[s]}\right) \end{aligned} \tag{8}$$

となる．Z_t は正規化定数であり，$\sum_{s \in S} p_{\mathrm{I}}^{[\mathrm{Bys}]}(s | \boldsymbol{v}_{1:t}^{[s]}) = 1$ となる．また，$\tilde{L}_{1:t}^{[s]}$ は次式のような負の正規化対数尤度である．

$$\tilde{L}_{1:t}^{[s]} \equiv -\frac{1}{t} \log L_{1:t}^{[s]} = -\frac{1}{t} \left(\sum_{k=1}^{t} \log p\left(\boldsymbol{v}_k^{[s]} | s\right) \right) \tag{9}$$

ただし，$L_{1:t}^{[s]} \equiv \prod_{k=1}^{t} L_k^{[s]}$ は出力 $\boldsymbol{v}_{1:t}^{[s]}$ の尤度であり，$L_k^{[s]} \equiv p(\boldsymbol{v}_k^{[s]} | s)$ は各 $\boldsymbol{v}_k^{[s]}$ の初期尤度である．$\tilde{L}_{1:t}^{[s]}$ は初期尤度の幾何平均 $\left(\prod_{k=1}^{t} p\left(\boldsymbol{v}_k^{[s]} | s\right) \right)^{1/t}$ の負の対数であるので，t が十分に大きくなると一定値に収束する．これより話者 $s_i \in S$ と話者 $s_m = \underset{s_i \in S}{\operatorname{argmax}} \, p_{\mathrm{I}}^{[\mathrm{Bys}]}\left(s_i | \boldsymbol{v}_{1:t}^{[s]}\right)$ の確率の比率は $t \to \infty$ のとき

$$\begin{aligned} r_i^{[\mathrm{Bys}]} &\equiv \frac{p_{\mathrm{I}}^{[\mathrm{Bys}]}\left(s_i | \boldsymbol{v}_{1:t}^{[s]}\right)}{p_{\mathrm{I}}^{[\mathrm{Bys}]}\left(s_m | \boldsymbol{v}_{1:t}^{[s]}\right)} \\ &= \frac{p_{\mathrm{I}}^{[\mathrm{Bys}]}\left(s_i | \boldsymbol{v}_{1:0}^{[s]}\right)}{p_{\mathrm{I}}^{[\mathrm{Bys}]}\left(s_m | \boldsymbol{v}_{1:0}^{[s]}\right)} \exp\left(-t\left(\tilde{L}_{1:t}^{[s_i]} - \tilde{L}_{1:t}^{[s_m]}\right)\right) \\ &\to \begin{cases} 1, & s_i = s_m \\ 0, & s_i \neq s_m \end{cases} \end{aligned} \tag{10}$$

となる．これは識別すべき音声データの話者 s が学習した話者集合 S に含まれる場合は最小の損失 $\tilde{L}_{1:t}^{[s]}$ をもつ $p_{\mathrm{I}}^{[\mathrm{Bys}]}(s_m | \boldsymbol{v}_{1:t}^{[s]})$ が十分大きくなり望まし

い識別ができるが, 話者 $s \notin S$ の場合も話者 $s_m \in S$ として誤識別される可能性があるという問題が生じることを示唆する.

3. ギブス分布を用いる多段話者識別

式(8)の代わりに, 式(9)の損失関数 $\tilde{L}_{1:t}^{[s]}$ をエネルギーとするギブス分布

$$p_{\mathrm{I}}^{[\mathrm{Gbs}]}\left(s \mid \boldsymbol{v}_{1:t}^{[S]}\right) \equiv \frac{1}{Z_t} \exp\left(-\beta \left(\tilde{L}_{1:t}^{[s]} + \frac{1}{t} \log p_{\mathrm{I}}^{[\mathrm{Gbs}]}\left(s \mid v_{1:0}^{[S]}\right) \right) \right) \tag{11}$$

を用いることを考える. β は逆温度と呼ばれるパラメータである. t の増加に対して, 話者 $s = s_i$ と話者 $s_m = \underset{s_i \in S}{\operatorname{argmax}}\, p_{\mathrm{I}}^{[\mathrm{Gbs}]}\left(s_i \mid \boldsymbol{v}_{1:t}^{[S]}\right)$ の確率の比率は次式のように 1 より小さい値に収束する.

$$r_i^{[\mathrm{Gbs}]} \equiv \frac{p_{\mathrm{I}}^{[\mathrm{Gbs}]}\left(s_i \mid \boldsymbol{v}_{1:t}^{[S]}\right)}{p_{\mathrm{I}}^{[\mathrm{Gbs}]}\left(s_m \mid \boldsymbol{v}_{1:t}^{[S]}\right)} \to \exp\left(-\beta \left(\tilde{L}_{1:t}^{[s_i]} - \tilde{L}_{1:t}^{[s_m]} \right) \right) \to c_i^{\beta} < 1 \tag{12}$$

このことは上記の単純ベイズ推定の問題が回避できる可能性を示唆する. さらに, c_i は初期尤度比率の幾何平均の収束値である.

$$c_i = \lim_{t \to \infty} \left(\prod_{k=1}^{t} \frac{L_k^{[s_i]}}{L_k^{[s_m]}} \right)^{1/t} = \lim_{t \to \infty} \left(\prod_{k=1}^{t} \frac{p\left(\boldsymbol{v}_k^{[S]} \mid s_i\right)}{p\left(\boldsymbol{v}_k^{[S]} \mid s_m\right)} \right)^{1/t} \tag{13}$$

ここで, 話者 $s_{m_2} = \underset{s \in S \setminus \{s_m\}}{\operatorname{argmax}}\, p_{\mathrm{I}}^{[\mathrm{Gbs}]}\left(s \mid \boldsymbol{v}_{1:t}^{[S]}\right)$ の収束値 $r_{m_2}^{[\mathrm{Gbs}]} = c_{m_2}^{\beta}$ は未登録者の音声よりも登録者の音声の方が小さくなると仮定する. なぜならば, 登録者の音声を入力した場合, 話者 s_m の尤度は未登録者の音声を入力した場合よりも大きな値となり, 一方話者 s_{m2} の尤度はどちらの場合もほとんど同じ値になると予想したからである.

式(11)より漸化式は次式のようになる.

$$p_{\mathrm{I}}^{[\mathrm{Gbs}]}\left(s \mid \boldsymbol{v}_{1:t}^{[S]}\right) \equiv \frac{1}{Z_t} p_{\mathrm{I}}^{[\mathrm{Gbs}]}\left(s \mid \boldsymbol{v}_{1:t-1}^{[S]}\right)^{\beta_t / \beta_{t-1}} p\left(\boldsymbol{v}_t^{[S]} \mid s\right)^{\beta_t} \tag{14}$$

ただし, $\beta_t = \beta / t \ (t \geq 1)$, $\beta_0 = 1$ である. $\beta_t = 1 \ (t \geq 1)$ のとき, 従来の方法であるベイズ推定となるので式(14)をギブス分布に基づく拡張ベイズ推定と呼ぶことにする.

4. 話者認識実験および解析

4.1 実験方法

本研究では, 実際に音声を録音したデータを入力としている. 音声を入力する際には, パソコンにマイクロフォンを接続し, 音声の形式を PCM, サンプリング周期を 8000Hz, 量子ビットを 16bit とし, モノラルで録音した. 録音環境は, ノイズが入らないように, 録音者しか居ない静かな教室で行った. 録音する際のマイクの位置は, 口元から 3cm～5cm の間でセットし, また, 音声データベースは, 5 人の話者の集合を S = {SM, SS, TN, WK, YM}とし, 各話者についてそれぞれ 5 単語 W = {daigaku, fukuokaken, gakusei, kikai, kyuukoudai}を各 10 回録音し“word1”, “word2”,…, “word10” のようにしたもので構成される. これらの音声時系列について, Leave-one-set-out-cross-validation (LOOCV)によって極の分布による話者認識実験を行った.

LOOCV とは, *I* 組の学習データのうち 1 組を取り除いてテストデータセットとし, 残り *I* −1 を使って学習することを全てのデータに対して *I*回繰り返すことを指す.

4.2 実験結果と考察

5 種類の単語それぞれ 10 個ずつのデータを用いて, 話者を 5 人として実験を行い, ベイズ推定とギブス分布のそれぞれを用いた話者識別結果を示す. テキスト依存型話者認識実験では, 5 人の話者を 1 セットとして各単語について 10 セット用意した. 1 つのセットをテストデータセットとし, 残りのセットを訓練データセットとして LOOCV によって話者認識を行った. テキスト独立型話者認識実験では, /daigaku/ をテストデータ, 訓練データを /fukuokaken/, /gakusei/, /kikai/, /kyukoudai/ といったように, ある単語をテストデータ, それ以外を訓練データとして話者照合及び話者識別実験を行った. 未登録者を含む話者識別については, 5 人の話者をそれぞれ学習させた 5 つの CAN2 のうち 1 つを取り除き, 除かれた CAN2 の話者は未登録者となる. そして, 未登録者の音声を入力して識別できるか実験を行った.

Table 1.(a)にテキスト依存型話者識別での登録者のみを対象とした場合と未登録者を含んだ場合の誤差率 E_{I}, $E_{\mathrm{I}}^{[\mathrm{Bys}]}$, $E_{\mathrm{I}}^{[\mathrm{Gbs}]}$ を, Table 1.(b)にテキスト独立型話者識別での実験結果を示す. ここで, E_{I} は式(3)により求めた単一話者識別の誤差率であり, $E_{\mathrm{I}}^{[\mathrm{Bys}]}$ はベイズ推定を用いた場合の誤差率, $E_{\mathrm{I}}^{[\mathrm{Gbs}]}$ はギブス分布を用いた場合の誤差率である.

Table 1. Error rates obtained in the experiments of speaker identification. E_{I} is the singlestep identification error rate. $E_{\mathrm{I}}^{[\mathrm{Bys}]}$ and $E_{\mathrm{I}}^{[\mathrm{Gbs}]}$ indicate the multistep identificationerror rates for BI and GEBI, respectively. The test of "speaker involving unregistered" is executed by LOOCV which leaves each classifier for $s_i \in S$ out as an unregistered speaker. The results of GEBI are obtained with the inverse temperature $\beta = 1$.

(a) text-dependent

		/kyukoudai/	/daigaku/	/kikai/	/fukuokaken/	/gakusei/
single step	E_{I}	0.080	0.060	0.060	0.100	0.040
multi step	登録者のみ					
	$E_{\mathrm{I}}^{[\mathrm{Bys}]}$	0	0	0	0	0
	$E_{\mathrm{I}}^{[\mathrm{Gbs}]}$	0	0	0	0	0
	未登録者を含む					
	$E_{\mathrm{I}}^{[\mathrm{Bys}]}$	0	0	0	0	0
	$E_{\mathrm{I}}^{[\mathrm{Gbs}]}$	0	0	0	0	0

(b) text-independent

		/kyukoudai/	/daigaku/	/kikai/	/fukuokaken/	/gakusei/
single step	E_{I}	0.200	0.200	0.280	0.400	0.420
multi step	登録者のみ					
	$E_{\mathrm{I}}^{[\mathrm{Bys}]}$	0	0	0	0	0
	$E_{\mathrm{I}}^{[\mathrm{Gbs}]}$	0	0	0	0	0
	未登録者を含む					
	$E_{\mathrm{I}}^{[\mathrm{Bys}]}$	0	0	0	0	0
	$E_{\mathrm{I}}^{[\mathrm{Gbs}]}$	0	0	0	0	0

Table 1.(a)より，単一話者識別では誤差が生じているが，多段話者識別だとベイズ推定とギブス分布の両方の場合で誤差をなくすことができた．Table 1(b)の結果も同様に誤差を 0 にすることができた．このことからギブス分布を用いることは未登録者を識別するうえで有効であることが考えられる．また，ギブス分布だけでなく，ベイズ推定を用いた場合でも未登録者を識別できたことから本稿での話者 $s_{m_2} = \underset{s \in S \setminus \{s_{m_1}\}}{\operatorname{argmax}}\, p_{\mathrm{I}}^{[\mathrm{Gbs}]}\left(s \mid \boldsymbol{v}_{1:t}^{[S]}\right)$ の収束値 $r_{m_2}^{[\mathrm{Gbs}]} = c_{m_2}^{\beta}$ は未登録者の音声よりも登録者の音声の方が小さくなるという仮定も正しく，有効であると考えられる．

5. 結論

本稿では，バギング CAN2 を用いる多段話者認識の性能向上のためにギブス分布を導入した．従来の手法ではベイズ推定を用いたことにより識別すべき音声データの話者 s が学習した話者集合 S に含まれない場合に適応していなかったが，本手法では未登録者を対象とした場合も適応させ，性能を向上させることができた．今後は本手法の性能を調べるためにより大きなデータベースやテキスト指定形話者認識での実験を行っていこうと考えている．

参考文献

[1] Ahalt, A.C., Krishnamurthy, A.K., Chen, P., Melton, D.E.: Competitive learning algorithms for vector quantization. *Neural Networks*, Vol. 3, pp.277–290 (1990)

[2] Kohonen, T.: Associative Memory. *Springer Verlag* (1977)

[3] Kurogi, S., Ueno, T. and Sawa, M.: A batch learning method for competitive associative net and its application to function approximation. *Proc. SCI2004*, Vol. V, pp.24–28 (2004)

[4] 根立奈緒子, "競合連想ネットによる母音の解析と認識", 九州工業大学卒業論文,2006.

[5] 船津由起, "競合連想ネットによる母音時系列の解析と認識", 九州工業大学卒業論文,2007.

[6] 嶺石将太, "バギング CAN2 を用いた音声の極分布パターン抽出に基づく話者認識手法の研究", 九州工業大学卒業論文,2010.

[7] Kurogi, S., Mineishi, S. and Sato, S.: An analysis of speaker recognition using bagging CAN2 and pole distribution of speech signals. *Proc. ICONIP2010, Part I, LNCS 6443*, pp.363–370 (2010)

[8] Kurogi, S., Mineishi, S.,Tsukazaki, T. and Nishida, T.: Naive Bayesian multistep speaker recognition using competitive associative nets. *Proc. ICONIP2011, LNCS 7062*, pp.70–78 (2011)

[9] Beigi, H.: Fundamentals of speaker recognition. *Springer-Verlag New York Inc (C)* (2011)

[10] Zhang, H.: The optimality of naive Bayes. *Proc. FLAIRS2004 conference*, (2004)