音声言語情報処理 44－37
(2002. 12. 20)

討論音声を対象とした話者モデル選択による話者インデキシングと自動書き起こし

西田 昌史† 秋田 祐哉†,†† 河原 達也†,††

† 科学技術振興事業団 さきがけ研究21 「協調と制御」領域
†† 京都大学 情報学研究科 知能情報学専攻
〒606-8501 京都市左京区吉田本町
E-mail: †{nishida, akita, kawahara}@kuis.kyoto-u.ac.jp

あらまし 討論音声を対象とした教師なし話者インデキシングとそれを用いた音声認識について報告する．討論音声では，話者の交替が頻繁に発生し，継続時間の短い発話が多く，発話時間のばらつきが大きいため，画一的なモデルで話者インデキシングを行うのが困難である．そこで，BICに基づいて最適な話者モデル(GMMまたはVQ)を選択する方式を提案する．本方式では，発話時間の短い音声に対してVQモデル，長い音声に対してはGMMモデルが選択される枠組みを実現する．実際の討論音声に対して，従来法に比べて高いインデキシング精度を得ることができた．次に，討論音声認識のための音響・言語モデルについて検討を行う．話者インデキシング結果に基づいて音響モデルを適応することにより，音声認識精度の改善を得ることができた．

キーワード 音声認識，話者認識，討論音声，教師なし話者インデキシング，モデル選択，BIC

Speaker Indexing based on Speaker Model Selection and Automatic Speech Recognition in Discussions

Masafumi NISHIDA†, Yuya AKITA†,††, and Tatsuya KAWAHARA†,††

† PRESTO, Japan Science and Technology Corporation (JST)
†† School of Informatics, Kyoto University, Kyoto, 606-8501 Japan
E-mail: †{nishida, akita, kawahara}@kuis.kyoto-u.ac.jp

Abstract This paper addresses unsupervised speaker indexing for discussion audio archives. In discussions, the speaker changes frequently, thus the duration of utterances is very short and its variation is large, which causes significant problems in applying conventional methods such as model adaptation and Variance-BIC (Bayesian Information Criterion) methods. We propose a flexible framework that selects an optimal speaker model (GMM or VQ) based on the BIC according to the duration of utterances. When the speech segment is short, the simple and robust VQ-based method is expected to be chosen, while GMM will be reliably trained for long segments. For a discussion archive, it is demonstrated that the proposed method achieves higher indexing performance than that of conventional methods. The speaker index is useful for speaker adaptation of the acoustic model, which improves the performance of automatic speech recognition.

Key words Speech recognition, Speaker recognition, Discussions, Unsupervised speaker indexing, Model selection, Bayesian information criterion

1. はじめに

本研究では，討論音声を対象とした話者インデキシングと自動書き起こし（音声認識）の実現を目指している．討論音声に対して話者のインデキシングを行うことで，特定の話者の発話を検索したり，話者毎に賛成や反対といった立場をタグ付けすることができると考えられる．また，話者適応により音声認識精度を改善することも考えられる．

話者インデキシングは，あらかじめ話者モデルを学習すれば容易に実現できる．しかし，同じ話者が常に討論に参加しているとは限らず，毎回話者毎に学習用のデータを収集することは多大な労力が必要であり，様々な条件に適用することが困難である．したがって，本研究では，事前に話者モデルを用意せず，話者数が未知である教師なしの話者インデキシングについて検討を行う．

近年，話者インデキシングの研究が主に，ボイスメール [1]，ニュース音声，Switchboard コーパス [2] を対象として行われている．ボイスメールのタスクでは，10 秒以上のメッセージを対象に話者のインデキシングが行われている．また，Switchboard コーパスは，電話での会話音声を対象としており，発話の平均時間は 31 秒，最小時間は 14 秒である [2]．これらのタスクでは，話者モデルは背景話者モデルから適応することで得られ，背景話者モデルと適応して得られた話者モデルとの尤度比に基づいて，話者のクラスタリングを行っている [3]．これらのタスクに対して，本稿で扱う討論音声は，継続時間が短い発話が非常に多く，MLLR のような適応技術を用いて話者モデルを得ることができない．また，発話の継続時間のばらつきが非常に大きいので，同一話者のセグメントでも分散が大きく異なることが予想される．

本稿では，データ長に応じて最適な話者モデルを選択することで，話者のインデキシングを行う手法を提案する．従来，話者認識においては，主に GMM [4] と VQ に基づく手法が用いられている．学習データが十分にある場合は，VQ よりも GMM の方が認識精度が高いことが知られている [5]．しかし，学習データが十分得られない場合，GMM のパラメータを頑健に推定することができない．そこで，BIC に基づいてデータ長に応じた最適な話者モデル (GMM または VQ) を選択し，適応することなく学習データから直接話者モデルを構築する．

また，このような討論音声を対象とした音声認識についても検討する．討論は基本的に話し言葉音声であるため，話し言葉の特徴を反映したモデルが必要である．また，言語モデルにおいて討論の話題を十分にカバーすることも必要である．そのため，日本語話し言葉コーパス (CSJ) と新聞記事コーパスを混合することにより言語モデルを構成する．

2. データベースとタスク

討論音声としてのデータは，NHK の『日曜討論』という 1 時間番組を用いる．本番組は，経済や政治に関する問題を司会者の進行のもと，ジャーナリストや政治家が討論しているものである．今回の話者インデキシングの実験では，2001 年 6 月から 12 月までに放送された 9 回分の音声を用いた．

表 1　討論音声のテストセット
Table 1　Test set of discussion speech

	A	B	C	D	E
#Speaker	5	5	5	8	6
#Utterance	534	665	609	541	612
	F	G	H	I	–
#Speaker	8	5	5	5	–
#Utterance	474	371	613	559	–

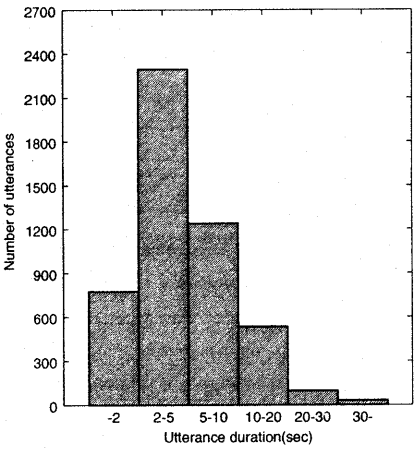

図 1　発話区間の継続時間分布
Fig. 1　Distribution of utterance duration

話者インデキシングは，発話区間の分割とクラスタリング処理により行う．発話区間の分割は，音声のパワーと零交差数に基づいて行い，分割されたセグメントを一発話とする．表 1 は，各討論音声の話者と発話区間数を示している．図 1 は，発話区間の継続時間の分布を示している．図 1 中の "5 – 10" は，5 秒から 10 秒までの発話区間数を表している．

全発話区間の平均継続時間は 6 秒，最小時間は 1 秒，最大時間は 71 秒である．また，10 秒未満の発話区間が，全体の 87%を占めている．このように，短い発話区間が非常に多く，発話時間のばらつきが大きいため，発話毎に頑健なモデルを構成することが困難であると考えられる．

3. Variance-BIC に基づく話者インデキシング

BIC(Bayesian Information Criterion) は，モデルの複雑度をペナルティとした尤度基準である．データの集合を $X = \{x_j \in \Re^d : j = 1, \cdots, N\}$，モデルのパラメータを $\lambda = \{\lambda_i : i = 1, \cdots, K\}$，モデル λ_i のパラメータ数を β_i とすると，BIC は，次式のように定義される．

$$BIC_i = \log P(X|\lambda_i) - \alpha \frac{1}{2} \beta_i \log N \qquad (1)$$

ここで，α は重み係数である．

話者交替を検出して話者インデキシングを行う方法として，各セグメントに対して単一ガウス分布を仮定し，セグメントの

分散比に基づいてクラスタリングを行う方法が提案されている[6]．本稿では，このBICに基づく手法が尤度を分散で近似していることから，“Variance-BIC”と呼ぶことにする．

Variance-BICに基づく話者インデキシングは，話者交替の検出とクラスタリング処理からなる．話者交替の有無，すなわち連続する二つのセグメントが同一話者の発話であるかどうかの判断は，二つのセグメントの分散に基づいて，次式により行われる．

$$\Delta BIC^{i}{}_{variance} = -\frac{n_1+n_2}{2}\log|\Sigma_0| + \frac{n_1}{2}\log|\Sigma_1| + \frac{n_2}{2}\log|\Sigma_2| + \alpha\frac{1}{2}(d+\frac{1}{2}d(d+1))\log(n_1+n_2) \quad (2)$$

ここで，Σ_0は二つのセグメントをマージしたときの共分散行列，Σ_1は一つ目のセグメントの共分散行列，Σ_2は二つ目のセグメントの共分散行列を表している．また，n_iは各セグメントのデータサイズ(フレーム数)，dは特徴ベクトルの次元数を表している．式(2)のΔBICの値が正であれば，二つのセグメントはマージされる．

討論音声では，発話の継続時間の分散が大きく，極端に短い発話に対して特に分散の推定ならびに比較が困難である．また，本手法は，重み係数がタスクに依存し，タスク毎に新たに設定し直す必要がある[7]．

4. 話者モデル選択に基づく話者インデキシング

4.1 統計的話者モデル選択

本稿では，データに応じて最適な話者モデルを選択する手法を提案する．GMMは，近年最も使われている統計的手法であるが，モデルのパラメータを頑健に推定するには，多くの学習データが必要になる．学習データが十分に得られない場合は，VQ歪みに基づく手法がGMMより認識精度が高いことが知られている[5]．従来の話者認識のタスクでは，各話者の学習データ量はほぼ同じであり，話者モデルは学習データ量やタスクに応じて，手動で設定するのが一般的であった．

これに対して，BICに基づいて学習データ量に応じた最適な話者モデル(VQまたはGMM)を自動的に選択する方式を提案する．しかし，GMMとVQでは，モデルの構造と識別尺度が異なっているので，そのままではBICを求めて比較することができない．そこで，VQを拡張した“CVGMM (Common Variance GMM)”というモデルを導入する．CVGMMは，通常のGMMにおけるすべての混合分布の重みと共分散を共通化したもので，GMMの尤度と比較することが可能になる．また，CVGMMは，共分散行列を単位行列に置き換えることで，VQモデルになる．

最初に，各セグメント毎にGMMの混合分布を推定する．クラスタsのGMMに対するBIC値は，次式により求められる．

$$BIC^{(s)}_{GMM} = \log P(X|\lambda^{(s)}_{GMM}) - \alpha M d \log N \quad (3)$$

ここで，$\log P(X|\lambda^{(s)}_{GMM})$はGMMに対する学習データ$X$の対数尤度である．$M$は混合分布数，$d$は特徴ベクトルの次元数，$N$は学習データのフレーム数を表しており，重み係数$\alpha$は1に設定している．

次に，CVGMMを生成する．ここで，CVGMMの混合分布の重みは，$\bar{w}^{(s)} = 1/M$により与える．CVGMMの共分散は，次式により各クラスタで学習したGMMの共分散の平均を用いる．

$$\bar{\Sigma}^{(s)}_{CVGMM} = \frac{1}{M\cdot S}\sum_{i=1}^{S}\sum_{j=1}^{M}\Sigma^{(i)}_{GMM_j} \quad (4)$$

ここで，Sはクラスタ数である．CVGMMの平均ベクトルには，GMMの平均ベクトルを用いてもよいが，過学習によりCVGMMの平均ベクトルを頑健に推定することが困難であるので，GMMのパラメータ推定に用いられているEMアルゴリズムの繰り返し回数を制御し，1回のみ学習した平均ベクトルを用いる．これは，VQのセントロイドを推定するのにほぼ対応する．なお，本研究では，GMMの学習におけるEMアルゴリズムの学習回数は，10回に設定している．

CVGMMに対するBIC値は，次式のようになる．

$$BIC^{(s)}_{CVGMM} = \log P(X|\lambda^{(s)}_{CVGMM}) - \alpha\frac{1}{2}d(M+1)\log N \quad (5)$$

学習データ量が少ない場合は，GMMとCVGMMに対する尤度の差が小さく，CVGMMにおけるモデルの複雑度が小さいため，CVGMMが選択されることが期待される．このように，本手法は，データ量に応じてモデルの構造と識別尺度を動的に制御することができ，任意の発話長に対して柔軟に話者のインデキシングが可能になる．

4.2 話者インデキシング

話者インデキシングは，BICに基づいてGMMまたはCVGMMを選択することにより行われる．以降，本手法を“SMS (Speaker Model Selection)”と呼ぶことにする．

本手法の処理手順を以下に示す．

（1） 学習：各クラスタに対して，GMMとCVGMMが話者モデルとして学習される．初期学習では，各発話が一つのクラスタになる．

（2） モデル選択：各クラスタにおいて，GMMとCVGMMに対するBIC値を求め，これに基づきいずれかのモデル（GMMまたはCVGMM）が選択される．

（3） 距離計算：クラスタ間の距離をCross Likelihood Ratio (CLR)[8]に基づいて計算する．CLRは次式により求められる．

$$d_{ij} = \log\frac{P(X_i|\lambda_i)}{P(X_i|\lambda_j)} + \log\frac{P(X_j|\lambda_j)}{P(X_j|\lambda_i)} \quad (6)$$

ここで，X_iはクラスタiに含まれている発話，λ_iはクラスタiで選択されたモデル(GMMまたはCVGMM)を表す．

（4） 識別処理によるクラスタリング：各クラスタに対して，距離が最小になるクラスタを探索し，最小距離のクラスタが一致するクラスタどうしをマージする．

上記の処理(1)から(4)までを繰り返し実行する．マージすべきクラスタがなくなった場合，モデルの学習と選択処理を終了し，処理(5)を実行する．

（5） 照合処理によるクラスタリング：クラスタ間の最小距離を計算し，距離が閾値θより小さければ，それらのクラスタをマージする．

すべてのクラスタ間で距離が閾値θより大きくなるまで，上記の処理 (3) と (5) を実行する．

本手法は，クラスタリング処理を 2 段階に分けて実行している．処理 (4) の最初のクラスタリングでは，短いセグメントからなるクラスタに対して，尤度を安定して得られないために，識別処理によりクラスタリングを行う．各クラスタの学習データが十分に得られた後，処理 (5) で尤度に基づく閾値処理により話者のクラスタリングを行う．

5. 話者インデキシング実験

5.1 実験条件と評価法

2 章で述べた 9 回分の討論音声を対象に話者インデキシング実験を行う．提案手法の有効性を示すために，Variance-BIC に基づく手法，VQ ならびに GMM に基づく手法との比較実験を行った．VQ ならびに GMM に基づく手法は，提案手法において，VQ ならびに GMM をすべてのクラスタのモデルとして選択した場合に相当する．

音声データのサンプリング周波数は 16 kHz で，特徴量は 12MFCC，12ΔMFCC，logPow，ΔlogPow の計 26 次元である．

話者インデキシング実験の評価には，BBN 尺度 [9] を用いる．BBN 尺度は，次式により求められる．

$$I_{BBN} = \sum_{i=1}^{N_c} n_i p_i - QN_c, \tag{7}$$

ここで，n_iはクラスタ i に含まれている発話数，N_cはクラスタ数，p_iはクラスタ i の純度を表している．クラスタの純度は，$p_i = \sum_{j=1}^{N_s} (n_{ij}/n_i)^2$と定義される．$N_s$は話者数，$n_{ij}$はクラスタ i に含まれている話者 j の発話数を表している．式 (7) の Q は，クラスタの純度とクラスタ数を制御するパラメータである．本稿では，Q を 0.5 に設定した．

インデキシング精度は，正しく話者のインデキシングを行った場合と，各手法で自動的にインデキシングを行った場合の BBN 尺度の比により，評価を行った．また，話者数の推定精度を正解率 (REcall rate) と適合率 (PRecision rate)，F 値 (F-value) を次式により求め，評価を行った．

$$RE = \frac{\text{インデキシングが正しくできた話者数}}{\text{正解話者数}} \times 100 \tag{8}$$

$$PR = \frac{\text{インデキシングが正しくできた話者数}}{\text{インデキシングされた話者数}} \times 100 \tag{9}$$

$$F\text{-}value = \frac{2 \cdot RE \cdot PR}{RE + PR} \tag{10}$$

5.2 実験結果

各手法により得られたインデキシング精度の平均を表 2 に示す．表 2 中の “Index” は BBN 尺度の比，“Spk num” は話者数の推定精度を表している．また，Variance-BIC に基づく手法では，重み係数αを 2.5 に設定したときの結果，VQ に基づく手法では，コードブックサイズを 4 から 32，GMM ならびに SMS に基づく手法では，混合分布数を 4 から 32 に変えて実験を行った結果を示している．

表 2 話者インデキシング結果
Table 2 Speaker indexing result

	Index	Spk num		
		RE	PR	F-value
Variance-BIC	85.6	100.0	81.6	89.6
VQ				
(4 cb)	70.7	96.0	80.0	86.0
(8 cb)	89.6	97.8	81.8	88.3
(16 cb)	91.7	100.0	86.8	92.4
(32 cb)	94.7	98.1	91.3	92.7
GMM				
(4 mix)	73.0	95.6	76.0	82.2
(8 mix)	94.9	100.0	84.9	90.9
(16 mix)	94.4	98.1	91.4	94.2
(32 mix)	93.2	100.0	89.6	94.2
SMS				
(4 mix)	73.0	95.6	76.0	82.2
(8 mix)	94.8	100.0	81.8	88.8
(16 mix)	96.3	100.0	90.1	94.3
(32 mix)	96.8	100.0	90.7	94.3

提案手法である SMS は，混合分布数が 32 のときに，インデキシング精度が 96.8%，話者数の推定精度が F 値で 94.3 という Variance-BIC や GMM といった従来法に比べて最も高い結果を得ている．したがって，このインデキシング結果を次章の討論音声認識に用いることにする．

GMM に基づく手法では，短い発話が多い場合，その発話で推定した混合分布の分散が非常に小さくなり，誤った識別を引き起こす．したがって，同じ話者を正しくマージすることができず，話者のインデキシング精度が低くなっている．これに対して SMS では，VQ の拡張として GMM の共分散を共通化した CVGMM を導入しているため，混合分布数が 16 以上でも安定して学習することができる．実際に混合分布数が大きくなると，CVGMM が選択されることが多くなっている．混合分布数が 4 のとき，GMM と SMS で同じ結果が得られた．これは，混合分布数が少ない場合，GMM のパラメータは頑健に推定することができ，CVGMM よりも BIC 値が大きくなり，各クラスタに対して GMM が選択されるためである．

VQ に基づく手法では，短い発話に対しても頑健なモデル化が可能になるが，継続時間が長い発話に対しては，GMM の方がよいモデルを与える．実際に，GMM の結果と比べると，コードブックと混合分布数が同じ場合，GMM の方がインデキシング精度が高くなっている．

Variance-BIC に基づく手法は，話者数の推定精度が高いが，インデキシング精度は低くなっている．これは，重み係数αの値が固定である一方，発話の継続時間の分散が大きく，短い発

話に対するクラスタリングを正しく行うことができないためである．

6. 討論音声の自動書き起こし

次に，この討論の音声認識について検討する．

6.1 言語モデル

『日曜討論』においては，(1) 政治・経済分野の専門用語や，時事に関する語句，(2) 話し言葉特有のフィラーや文末表現などが観測される．しかし，これらの言語的特徴を十分に含んだ，一般に利用可能なテキストコーパスは存在せず，タスクにマッチした言語モデルを直接構築することができない．『日曜討論』を対象とした先行研究 [10] では WWW 上の講演録から構築したモデル [11] を用いているが，未知語は辞書に別途登録する必要があるほか，講演録は人手により整形済みであるために話し言葉表現が十分に含まれず，討論音声とのミスマッチが大きいと考えられる．実際，このモデルによる日曜討論音声のテストセット・パープレキシティは 292.2 と報告されている．

そこで，本研究では上記 (1) と (2) を表現するモデルを混合することで，討論音声に対応したモデルを作成することとした．(1) には，政治・経済分野の話題に親和性の高い新聞記事コーパスを利用した単語 trigram の新聞モデルを構築した．(2) については，学会等の講演からなる『日本語話し言葉コーパス』(CSJ) [12] を用いて構築した講演モデル [13] を用いた．形態素解析は (1)(2) ともに ChaSen 2.02 [14] を用いている．これらのモデルの仕様を表 3 に示す．表 3 中のテストセットパープレキシティ(PP) および未知語率 (OOV) の算出には，『日曜討論』9 回分の書き起こしを用いている．また，語彙選択の際のカットオフの値は 1 にしている．

これらのモデルを重み付け混合し，討論音声認識用のモデルを作成した．語彙サイズは 42,471 語となった．新聞モデルと講演モデルの混合比は事後的に 0.7:0.3 と定めた．

『日曜討論』の各討論に対するパープレキシティと未知語率を表 4 に示す．9 回分 (1 回あたり平均 14,503 語) の平均のテストセットパープレキシティは 212.42，未知語率は 2.41%であり，語彙サイズが約 1/3 の講演モデルと同等のパープレキシティの値を維持したまま未知語率を削減することができた．

6.2 音響モデルの話者適応

音響モデルも，言語モデルと同様の理由により，討論音声から直接構築することができない．そのため本稿では CSJ の学会講演音声を用いて作成された PTM triphone HMM の音響モデル [13] を利用する．音響モデルの仕様を表 5 に示す．討論話者の発話には言い淀みや発音の怠けといった講演音声と共通の現象が観測され，実際に予備実験において，講演音声モデルの方が読み上げ音声によるモデルよりも高い性能が得られた．

この音響モデル (ベースライン) に対して，話者インデキシング結果を用いた教師なし MLLR 話者適応を行う．各話者の適応に用いる音声はインデキシングによりその話者のラベルが付与された発話とし，各発話の音素トランスクリプションとして，ベースラインモデルによる音声認識結果の音素表記を与える．MLLR 適応におけるクラスタ数は 32 とした．

なお，話者適応による精度の改善に関する上限値を調べるため，あらかじめ人手により付与された話者ラベルと書き起こしを用いて，ベースラインモデルに教師あり話者適応を行った音響モデルを用いた評価も行う．

6.3 評価実験

これらのモデルを用いて，『日曜討論』の音声認識実験を行った．デコーダは Julius 3.3 を利用し，1 分を超えるような長い発話に対処するために逐次デコーディング [15] を適用している．テストセットは話者インデキシングと同じ 9 討論である．認識結果を図 2 に示す．

ベースラインの音響モデルを用いた場合では 47.4%，話者インデキシング結果を用いた教師なし話者適応を行った場合では 51.3%，教師あり話者適応を行った場合では 52.9%の認識精度が得られた．したがって，話者インデキシングを行い，その結果を用いて話者適応を行うことで，音声認識精度を改善することができた．

特に，討論音声データ D, E, F に対しては，他の討論データに比べて，認識精度が低い．これは，他の討論音声データに比べて話者数が多く，さらに複数の話者が同時に発話している区間が比較的多かったためである．また，認識結果を分析したところ，言い淀みやフィラーが多く，これらの単語の前後で連続して認識に失敗していた．

講演音声を対象とした音声認識に比べて，認識精度が低い結果となった．これは，学会講演音声を用いて音響モデルを作成し，また言語モデルを新聞記事と学会講演を混合して作成しているため，討論音声に対して音響モデルと言語モデルがマッチしていないためであると考えられる．今後，さらに音響モデルと言語モデルの作成について検討していく必要がある．

7. おわりに

本稿では，話者交替が頻繁に発生し，発話が短く継続時間のばらつきが大きい討論音声を対象とした教師なし話者インデキシング法について検討を行い，BIC に基づいて発話の継続時間に応じた最適な話者モデル (VQ または GMM) を選択する手法を提案した．NHK の『日曜討論』9 回分を対象として，話者インデキシング実験を行った結果，本手法は従来法に比べて高いインデキシング精度が得られた．また，本手法では，事前に話者数が未知で，各話者の話者モデルを用意しなくても，話者インデキシングが可能である．

また，この討論音声データの自動書き起こしについても検討を行った．特に言語モデルについては，新聞記事モデルと講演モデルを混合することにより効果的なモデルを作成することができた．話者インデキシング結果を利用した話者適応により，認識精度の改善を得ることができた．

今後は，複数の話者がオーバーラップして発話している場合を対象とした話者のインデキシング，話者適応による音声認識精度の改善について検討を行う予定である．

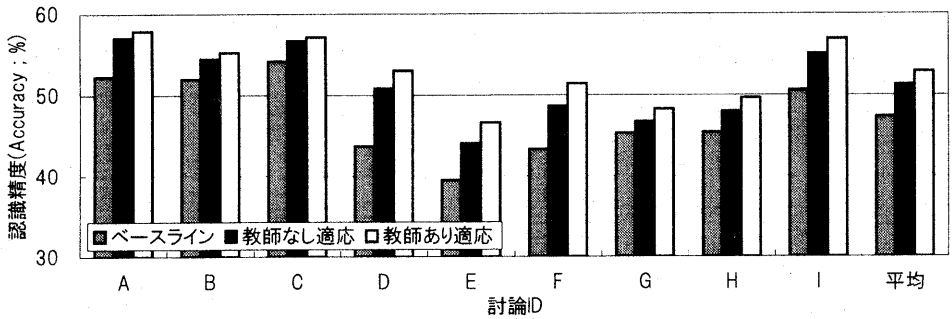

図 2　討論音声の認識結果

Fig. 2　Automatic speech recognition result

文　献

[1] D. Charlet, "Speaker Indexing for Retrieval of Voicemail Messages," Proc. ICASSP, vol. 1, pp. 121-124, 2002.

[2] S. Meignier, J.F. Bonastre, and I.M. Chagnolleau, "Speaker Utterances Tying Among Speaker Segmented Audio Documents Using Hierarchical Classification: Towards Speaker Indexing of Audio Databases," Proc. ICSLP, pp. 577-580, 2002.

[3] D.A. Reynolds, "Comparison of Background Normalization Methods for Text-Independent Speaker Verification Systems," Proc. EUROSPEECH, pp. 963-966, 1997.

[4] D.A. Reynolds and R.C. Rose, "Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models," IEEE Trans.SAP, vol. 3, no. 1, pp. 72-83, 1995.

[5] T. Matsui and S. Furui, "Comparison of Text Independent Speaker Recognition Methods Using VQ Distortion and Discrete/Continuous HMMs," Proc. ICASSP, vol. 2, pp. 157-160, 1992.

[6] S. Chen and P. Gopalakrishnan, "Speaker, Environment and Channel Change Detection and Clustering via the Bayesian Information Criterion," Proc. DARPA Broadcast News Transcription and Understanding Workshop, 1998.

[7] A. Tritschler and R. Gopinath, "Improved Speaker Segmentation and Segments Clustering Using the Bayesian Information Criterion," Proc. EUROSPEECH, vol. 2, pp. 679-682, 1999.

[8] D.A. Reynolds, E. Singer, B.A. Carlson, G.C. O'Leary, J.J. McLaughlin, and M.A. Zissman, "Blind Clustering of Speech Utterances based on Speaker and Language Characteristics," Proc. ICSLP, pp. 3193-3196, 1998.

[9] A. Solomonoff, A. Mielke, M. Schmidt, and H. Gish, "Clustering Speakers by Their Voices," Proc. ICASSP, pp. 757-760, 1998.

[10] 田熊 竜太, 岩野 公司, 古井 貞熙, "逐次話者適応を用いた並列処理型会議音声認識システムの検討," 音講論集, 2-5-16, Mar. 2002.

[11] 篠崎 隆宏, 細川 貴生, 古井 貞熙, "話し言葉コーパスを用いた音声認識の検討," 音講論集, 1-3-14, Mar. 2001.

[12] 小磯 花絵, 前川 喜久雄, "『日本語話し言葉コーパス』の設計の概要と書き起こし基準について," 情処学研報, SLP-36-1, 2001.

[13] 南條 浩輝, 河原 達也, "講演音声認識のための種々の形態素解析及び音響モデルの評価," 話し言葉の科学と工学ワークショップ講演予稿集, pp.47-52, Feb. 2002.

[14] 松本 裕治, 北内 啓, 山下 達雄, 平野 善隆, 松田 寛, 浅原 正幸, "日本語形態素解析システム『茶筌』version 2.0," Dec. 1999.

[15] 河原 達也, 加藤 一臣, 南條 浩輝, 李 晃伸, "話し言葉音声認識のための言語モデルとデコーダの改善," 情処学研報, SLP-36-3, 2001.

表 3　言語モデルの仕様

Table 3　Language model

	新聞モデル	講演モデル
学習データ	CD-毎日新聞 (2001 年版)	日本語話し言葉コーパス (CSJ)
総単語数	29.5M	2.7M
異なり単語数	146K	40K
語彙サイズ	40K	13K
平均 PP	354.17	208.37
平均 OOV	4.72%	7.74%

表 4　各討論のパープレキシティと未知語率

Table 4　Perplexity and Out-Of-Vocabulary rate

	A	B	C	D	E
PP	185.45	206.92	204.53	231.26	275.29
OOV	1.83	1.97	2.23	2.48	2.94
	F	G	H	I	–
PP	211.82	168.53	241.85	186.15	–
OOV	2.52	2.68	3.08	1.98	–

表 5　音響モデルの仕様

Table 5　Acoustic model

学習データ	日本語話し言葉コーパス (CSJ) 60 時間 (男性)
特徴量	MFCC(12) ΔMFCC(12) ΔPow(1) 計 25 次元
音素数	43
状態数	3,000
コードブック数	129
混合数	128