Vol.2010-SLP-80 No.9
2010/2/13

音素情報を利用した BIC に基づくオンライン話者識別

奥　貴裕[†]　佐藤　庄衛[†]　小林　彰夫[†]
本間　真一[†]　今井　亨[†]

字幕放送の拡充やメタデータの効率的な制作を目的とした音声認識では，発話の内容だけでなく，誰がいつ話したのかを検出する「話者識別」の併用が有効である．本報告では，音素情報を利用した，ベイズ情報量基準（BIC）に基づくオンライン処理向けの話者識別手法について述べる．提案する話者識別手法では，音素認識から得られる音素情報を用い，特徴量を音素クラスに分類することで，精度向上を図る．報道系情報番組の対談部分を対象とした識別実験を行った結果，話者交替点毎の判定手法において，従来の全音素を用いる場合に比べ「母音＋鼻音」クラスの場合で 1.2 ポイント識別率が改善することを確認した．また，話者照合などで用いられている混合ガウス分布（GMM）による手法との比較実験も行い，提案手法の有効性を確認した．

Online Speaker Diarization with Phonetic Information Based on BIC

Takahiro Oku,[†] Shoei Sato,[†] Akio Kobayashi,[†]
Shinichi Homma[†] and Toru Imai[†]

In speech recognition for closed-captioning and efficient production of metadata, not only the contents of the speech but also “speaker diarization” detecting ‘who spoke when’ is effective in combination. In this paper, we describe a new online speaker diarization method with phonetic information based on Bayesian Information Criterion (BIC). To improve the diarization accuracy, we classify speech features according to phonetic information obtained by phoneme recognition. In a speaker diarization task of conversational TV news programs, our new online method determining a speaker with a class of vowels and nasals at a each speaker change point reduced the diarization error rate (DER) by 1.2 point. We also show that our method yields better performance compared to the conventional method using a Gaussian mixture model (GMM).

1. はじめに

NHK では，ニュースなど生放送番組を対象とした，字幕制作のための音声認識の研究を行っている．NHK における音声認識の現在の課題は，対談など自由発話を含む番組の認識精度の改善である[1]．音声の認識時に，話者識別によって音声から「誰が，いつ」発話したかが検出できれば，音響モデルの話者適応などにより，認識率の改善が期待できる[2]．また，放送番組の発話内容の書き起こしだけでなく，話者名や話者の交替点を抽出できれば，番組の検索やメタデータの制作を効率よく行うことができる[3]．

本稿では，報道系情報番組の対談部分を対象とした，オンライン処理向けの話者識別を検討する．従来の話者識別システムでは，発話区間検出によって一定区間の無音で切り出された発話を単位として，話者識別をするものが多い[4][5]．しかし，これを対談番組に適用しようとした場合，無音を挟まずに話者交替があるような箇所では，１発話内に複数の話者が含まれてしまうという状況が発生する．よって，対談番組では，１発話毎の話者判定ではなく，発話内の話者交替点をオンラインで探索しながら話者判定をする必要がある．Liu らは，ニュース音声を対象とし，音素境界を話者交替点の候補として，話者の交替点をオンラインで逐次検出しつつ，交替点に挟まれた発話区間の話者を判定する話者識別システムを提案している[6][7]．

本報告では，報道系情報番組の対談部分を対象として，音素情報を利用したベイズ情報量基準（BIC）[8][9]に基づくオンライン処理向けの話者識別手法を提案する[10]．提案手法は，音素認識[11]によって得られる音素情報を利用し，個人性をより多く含むと考えられる「母音＋鼻音」とそれ以外の「子音」のクラスに音響特徴量を分類することで精度向上を図る．識別実験では，話者交替点毎および発話区間検出による発話末毎の判定や，一定の窓幅以前の話者の逐次確定など，オンライン性を考慮した判定手法を用いて提案手法の性能を評価し，提案手法の有効性を確認する．更に，提案する BIC による話者判定手法を，話者照合などで用いられる混合ガウス分布（GMM）に基づく従来手法[12][4]と比較した実験も行い，対談音声における提案手法の有効性を示す．

2. 音素情報を利用した話者識別

話者識別は，個人性をより多く含む音声区間の特徴量を用いることで，その精度向上が期待される．聴取による話者識別実験では，母音や鼻音が識別に有効であるという報告があり[13]，話者識別で用いる特徴量を音素ごとに分類することによる効果が

[†] NHK 放送技術研究所
NHK Science and Technology Research Laboratories

表 1　音素クラス

母音＋鼻音	a, a:, i, i:, u, u:, e, e:, o, o:, n, ny, m, my, N
子音	b, by, ch, d, dy, f, g, gy, h, hy, j, k, ky, p, py, r, ry, s, sh, t, ts, w, y, z

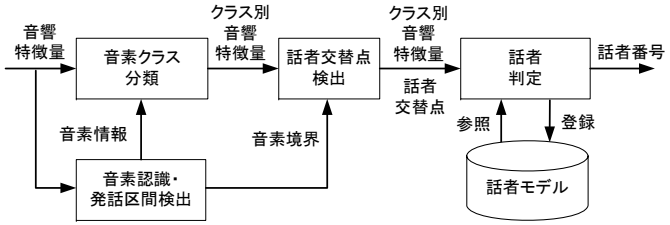

図 1　話者識別の流れ

期待できる．提案する話者識別手法では，音素認識から得られる音素情報を用い，特徴量を音素クラスに分類することで精度向上を図る．聴取実験の知見に基づき，表 1 に示すように，個人性情報をより多く含むと考えられる「母音＋鼻音」と，それ以外の「子音」のクラスに特徴量を分類した．識別実験では，「母音＋鼻音」と「子音」それぞれの音素クラスの特徴量のみによるモデルを用いた場合と，それらのモデルを統合した「混合モデル」について評価し，表 1 に示す音素（無音を示す音素以外）すべてに対応する特徴量を用いた「全音素」の場合と比較する．

3. オンライン話者識別

本提案手法の話者識別の流れを図 1 に示す．まず，前章で述べたように，音素認識によって得られる音素情報により，表 1 に示す音素クラスに音響特徴量を分類し，クラス別音響特徴量を得る．ここで，音素認識と同時に発話区間検出も行い，この発話区間情報を後述する話者の判定で利用する．次に，クラス別音響特徴量を用い，音素境界を候補として話者の交替点を逐次検出しつつ，登録された話者モデルを用いて話者の判定を行う．また，今回の識別実験では，識別開始時の話者モデルの登録は 0 名とし，過去に発話した話者以外の新規話者と判定された場合に，オンラインで新規話者モデルを作成し登録していくタスクを想定している．ただし，ニュース番組のキャスターなど，あらかじめ出演することが分かっている話者については，前もって話者モデルを作成しておくという方法も考えられる[3]．話者の判定は，音声の入力からの遅れ時間が少ないほど話者識別のオンライン性が高いと考えられ，後述する各種条件の判定タイミングにおいて，実験的に提案手法を検証する．

3.1 ベイズ情報量基準(BIC)

話者交替点の検出，および話者の判定には，共に BIC に基づく(1) 式の ΔBIC を用いる[8][9]．ΔBIC は 2 つの発話の特徴ベクトル列 x , y に対して，それらが同一話者によるものかどうかを判定する基準である．

$$\Delta BIC(x,y)=\log\left(\frac{p(x|\lambda_x)p(y|\lambda_y)}{p(x\,y|\lambda_{xy})}\right)-\alpha P$$
$$=\frac{1}{2}\left[N_{x\,y}\log\left|\Sigma_{x\,y}\right|-N_x\log\left|\Sigma_x\right|-N_y\log\left|\Sigma_y\right|\right]-\alpha\left(\frac{d(d+3)}{4}\right)\log\left(N_{x\,y}\right) \quad (1)$$

ここで，$\lambda(N,\Sigma)$は話者モデルを示し，Σ は特徴ベクトルの共分散行列，N はフレーム数である．$\lambda_{x\,y}$ は x と y が同一話者による発話と仮定した場合のモデルを示す．P , α , d は，それぞれペナルティ項とその重み係数，および特徴ベクトルの次元数である．ΔBIC の値が正のとき，x と y は別話者による発話であると判定される．

また，音素クラスの混合モデルを考えた場合，ΔBIC は(1)式の拡張として(2)式のように表現できる[5].

$$\Delta BIC(x,y)=\log\left(\frac{\prod_{m=1}^{M}p(x_m|\lambda_x^m)p(y_m|\lambda_y^m)}{\prod_{m=1}^{M}p(x_m\,y_m|\lambda_{xy}^m)}\right)-\alpha P$$
$$=\frac{1}{2}\left[\sum_{m=1}^{M}N_{x\,y}^m\log\left|\Sigma_{x\,y}^m\right|-\sum_{m=1}^{M}N_x^m\log\left|\Sigma_x^m\right|-\sum_{m=1}^{M}N_y^m\log\left|\Sigma_y^m\right|\right]-\alpha M\left(\frac{d(d+3)}{4}\right)\log\left(N_{x\,y}\right) \quad (2)$$

ここで，M は混合する音素クラスの数を示し，$\lambda_x^m\ (m=1,\cdots,M)$ は，音素認識結果から音素クラス m に属すると判定された音声区間の統計量である．[5]では，混合分布の尤度を分布内の最大尤度で置き換えることにより，近似的に BIC の混合モデルへの拡張を行っている．一方，本提案手法では，音素認識によって得られる音素情報により，特徴量を音素クラスに分類し，各音素クラス毎のモデル $\lambda_x^m\ (m=1,\cdots,M)$ を作成することで，混合モデルへの拡張を実現していることになる．よって，提案手法の話者モデルは，各音素クラスのフレーム数 N_m と共分散行列 Σ_m で表現される．

3.2 話者交替点検出

話者交替点検出の動作例を図 2 に示す．話者交替点検出では，候補となる交替点 $T_{\text{hyp}}=\{t_{\text{last}},\cdots,t_{\text{curr}}\}$ の前後での話者交替の有無を判定する．T_{hyp} は音素認識から得られる音素境界の集合であり，t_{last} は最後に確定された話者交替点，t_{curr} は現時刻を示す．話者交替点の候補を音素境界に制限[6]することで，効率的な交替点検出が可能

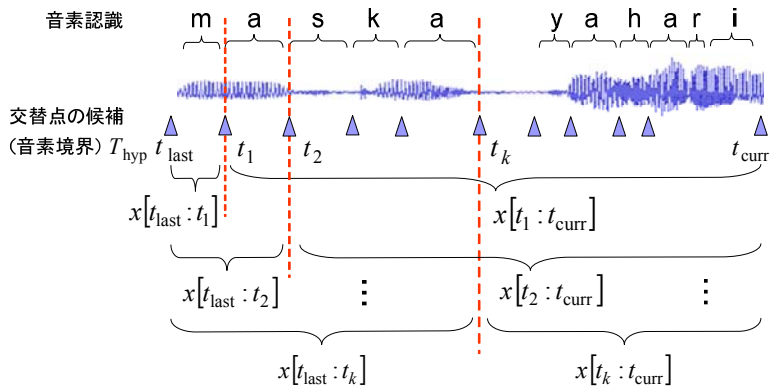

図 2　話者交替点検出

であり，次の(3)，(4)式を満たす t_{h} を話者交替点とする．

$$t_{\mathrm{h}} = \arg\max_{t_{\mathrm{k}} \in T_{\mathrm{hyp}}} \Delta BIC\left(x[t_{\mathrm{last}}:t_{\mathrm{k}}], x[t_{\mathrm{k}}:t_{\mathrm{curr}}]\right) \quad (3)$$

$$\Delta BIC\left(x[t_{\mathrm{last}}:t_{\mathrm{h}}], x[t_{\mathrm{h}}:t_{\mathrm{curr}}]\right) \geq 0 \quad (4)$$

ここで，$x[t:t']$は時刻 $t+1$ から t' までの音響特徴量系列を示す．また，十分な統計量を得るため，評価する発話長は 2 秒以上と設定した．

3.3　話者判定（クラスタリング）

提案する話者クラスタリング手法の概要を図 3 に示す．話者クラスタリングでは，登録された話者モデルの集合 C を考え，入力音声が C のいずれかの話者か，新規話者であるかを判定する．話者判定は，話者交替点毎および発話区間検出による発話末毎や，一定の窓幅以前の話者の逐次確定など，オンライン性を考慮して行うものとする．y_i を話者 i の発話，t_{d} を話者判定する時刻としたとき，

$$\Delta BIC\left(x[t_{\mathrm{last}}:t_{\mathrm{d}}], y_i\right) \geq 0 \quad \forall i \in C \quad (5)$$

であれば，発話 $x[t_{\mathrm{last}}:t_{\mathrm{d}}]$ は新規話者と判定する．(5)式が満たされなければ，

$$j = \arg\min_{i \in C} \Delta BIC\left(x[t_{\mathrm{last}}:t_{\mathrm{d}}], y_i\right) \quad (6)$$

を発話者と判定する．話者の判定後，$x[t_{\mathrm{last}}:t_{\mathrm{d}}]$ の統計量を y_j に追加して，当該話者モデルを更新する．新規話者と判定された場合には，新たに話者モデルの作成と登録を行う．

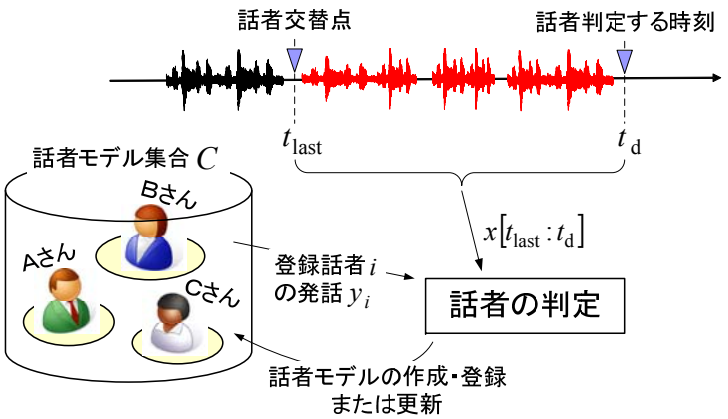

図 3　話者クラスタリング

4.　話者識別実験

4.1　実験条件

以上の提案手法により，話者識別実験を行った．識別の評価指標には，NIST が提案する diarization error rate(DER)を用いた[9]．DER は以下の(7)式で定義される．

$$DER = \frac{FS + MS + SE}{\text{総発話時間}} \quad (7)$$

ここで FS(False alarm speech)は発話者なしの区間で発話と誤判定した時間，MS(Missed speech)は発話者ありの区間で発話なしと誤判定した時間，SE(Speaker error)は話者を誤った時間を示す．

評価データには 2008 年 5 月の NHK の報道系情報番組「クローズアップ現代」の対談部分（総発話時間 2000 sec，話者 7 名，話者交替数 70）を用いた．開発データには評価データの前週の同番組を使用し，ΔBIC のペナルティ項の重み α を決定した．特徴ベクトルは 12 次元 MFCC＋対数パワー＋ Δ ＋ ΔΔ の計 39 次元とした．音素認識には[11]で提案された男女並列の性別依存音響モデルによる発話区間検出手法を用い，これにより得られる発話末 t_{e} を後述の話者判定手法で用いた．音素認識率は 58%であり，表 1 に示した音素クラスの認識率は 72%であった．また，上記 MS，FS は音素認識による発話区間検出で決定され，それぞれ総発話時間の 1.0%，0.5%であった．

実験では，オンライン性を考慮して以下の３通りの判定手法で評価した．判定手法の概要を図 4 に示す．

Vol.2010-SLP-80 No.9
2010/2/13

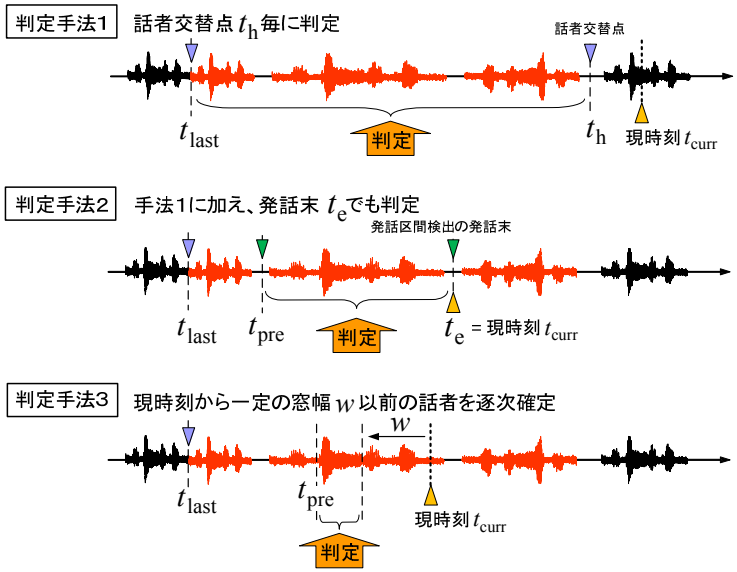

図 4　オンライン話者識別の判定手法

判定手法 1：話者交替点 t_h が検出されるたびに，$x[t_{last}:t_h]$ の話者を判定する．

判定手法 2：手法 1 に加え，発話区間検出における発話末 t_e にて $x[t_{pre}:t_e]$ の話者を判定する．t_{pre} は，話者の確定が終了している最終時刻を示す．この手法では，時刻 t_{last} 以降は話者交替は発生していないと考えられるので，判定精度向上のため，$x[t_{last}:t_e]$ の統計量を用いて $x[t_{pre}:t_e]$ の話者を判定する．ただし，すでに確定している t_{pre} 以前の話者判定結果は変更しないものとする．

判定手法 3：現時刻から一定の窓幅 w 以前の発話者を逐次確定する．判定手法 2 と同様に，判定精度向上のため，$x[t_{last}:t_{curr}]$ の統計量を用いて，$x[t_{pre}:t_{curr}-w]$ の話者を判定する．また，確定している t_{pre} 以前の話者判定結果は変更しない．

話者の判定は，音声の入力からの遅れ時間が少ないほど話者識別のオンライン性が高いと考えられる．上述の判定手法の遅れ時間は，判定手法 1 では話者交替点間の発話時間，判定手法 2 では発話区間検出における発話時間，判定手法 3 では一定窓幅 w である．判定手法 3 では，一定窓幅 w を小さくすれば，遅れ時間を小さくすることができるので，判定手法 1 よりも判定手法 2，判定手法 2 よりも判定手法 3 の方が話者判定の遅れ時間は少なく，よりオンライン性の高い判定手法であると言える．

表 2　話者識別結果（判定手法 1,2）

識別手法	話者交替点で判定（判定手法 1）		発話末でも判定（判定手法 2）	
	DER[%]	SE[%]	DER[%]	SE[%]
全音素（従来法）	4.0	2.4	4.7	3.1
母音＋鼻音	2.8	1.2	4.2	2.6
子音	4.1	2.5	6.8	5.2
混合モデル	3.1	1.5	4.3	2.7

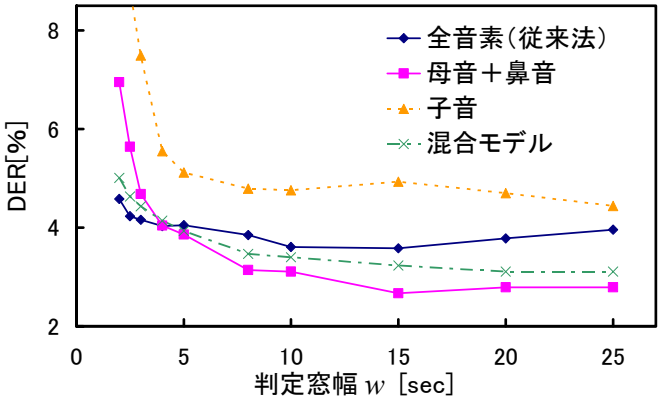

図 5　話者識別結果（判定手法 3）

4.2　実験結果

表 2 に判定手法 1，判定手法 2 の識別結果を示す．全音素を用いる従来法に比べ，「母音＋鼻音」や「混合モデル」は DER が低かった．また，「子音」クラスの特徴量が比較的低い値であったのは，子音にも個人性の情報が存在する可能性や，音素認識において母音を子音と誤認識している部分が識別に寄与しているためと考えられる．

判定手法 3 において，判定の窓幅を変化させたときの DER を図 5 に示す．ただし，図の横軸は判定窓幅 w であり，音素認識で判定された無音区間は除かれている．w を無限に大きくした場合の値は，判定手法 1 の場合に相当する．w が 10～25 秒では，「母音＋鼻音」と「混合モデル」ともに，従来の全音素に比べて DER が低く，識別性能は向上した．しかし，w が 2～3 秒では，DER は従来法が最も低かった．また，w が 2～3 秒の区間では，DER が「母音＋鼻音」の場合で 4.7%から 7.0%へ，「子音」の場合

で 7.5%から 10.2%へと急激に悪化した．これは，特に話者交替点の検出直後において話者クラスタリングをする際，従来法に比べて各音素クラスの統計量が十分に得られなかったことが原因と考えられる．

5. GMM に基づく手法との比較検討

5.1 GMM に基づく話者識別手法

従来の話者照合や話者識別は，話者モデルを GMM で表現することが多い[4][5][12]．GMM による話者識別では，判定する発話を X としたとき，(8)式で表される尤度比 $L_1(X)$ を閾値 θ と比較することによって判定を実施する．

$$L_1(X) = \frac{P_{\mathrm{sp}}(X)}{P_{\mathrm{gen}}(X)} \qquad (8)$$

ここで，$P_{\mathrm{gen}}(X)$ はあらかじめ作成した男女別の Universal Back ground Mo del (UB M) のうち，発話 X が生成される尤度の高い方の値，$P_{\mathrm{sp}}(X)$ は登録された話者モデルのうち，発話 X が生成される尤度が最大となる話者モデルの尤度の値を示す．$L_1(X)$ が閾値 θ 以下であれば新規話者と判定し，UBM と発話 X から新規話者モデルを作成してシステムに登録する．そうでなければ登録話者と判定し，該当話者モデル（GMM）を発話 X により学習する．話者モデルの作成および学習は，オンラインで MAP 推定[12]により実施する．

また，文献[4]では尤度が最大となる話者以外の話者の尤度平均 $P_{\mathrm{ave}}(X)$ を用いた，以下の(9)式で表される尤度比 $L_2(X)$ を採用している．

$$L_2(X) = \frac{P_{\mathrm{sp}}^2(X)}{P_{\mathrm{gen}}(X)P_{\mathrm{ave}}(X)} \qquad (9)$$

以下，(8)式による話者判定を GMM 手法 1 とし，(9)式による話者判定を GMM 手法 2 とする．

一般に GMM の混合分布は，対角共分散行列で表すことが多い．一方，BIC による提案手法は，全共分散行列を用いる．したがって，GMM は特徴ベクトルの次元間を独立に扱うため，過学習が生じやすいのに対し，全共分散で表現する BIC は次元間の相関も考慮するため，過学習は比較的生じにくいと考えられる．さらに，GMM による手法は，(8)式や(9)式で表されるように，あらかじめ作成しておいた UBM が必要であり，収録環境の違いなどを適切に考慮しておく必要もあるといった点が，BIC とは異なる．

5.2 比較実験

文献[5]では，ニュース音声を対象とし，1 発話毎に話者交替が発生したかどうかの判定について，各手法による比較検討を行っている．一方，本章では，報道系情報番組の対談部分を対象として，提案するオンライン話者識別の BIC による話者判定手法を，GMM による従来手法と比較検討する．

表 3　真の話者交替点での GMM による話者識別結果 DER(%)

混合数	GMM 手法 1	GMM 手法 2
32	14.1	14.1
64	14.1	8.9
128	9.8	7.6
256	9.9	7.6
512	9.9	8.3

表 4　真の話者交替点での BIC による話者識別結果 DER(%)

音素情報 利用なし	母音＋鼻音
1.8	1.8

表 5　各誤り要因に対する誤り発話時間(sec)　（総発話時間：2000sec）

誤り要因	GMM 手法 2（混合数 128）	BIC（母音＋鼻音）
あいづち	134.5 0.	0
雑音	48.4 0.	0
背景音楽	0.0 6.	9
短い発話（2 秒以下）	3.9 1.	0

話者識別の評価指標には，DER を用いた．評価データには 4 章と同様，報道系情報番組「クローズアップ現代」の対談部分を用いた．GMM による手法では，対数パワーが 16 以上の特徴量のみを話者識別に用いることとした．また，男女別の UBM は，男性 698 分，女性 417 分の NHK のニュース音声から作成した．

まず，真の話者交替点毎に切り出された発話の話者判定について，GMM による手法と BIC による手法の比較実験を実施した．ここでは，話者モデルの作成および学習に，話者識別による判定結果ではなく，真の話者情報を基に行うこととした．これは，GMM と BIC で，話者モデルの学習データを同一の条件にして，話者判定の性能だけを比較するためである．

混合数を変化させたときの GMM 手法 1 および GMM 手法 2 の話者識別結果を表 3 に示し，音素情報を利用しない場合の BIC と「母音＋鼻音」クラスの BIC の話者識別結果を表 4 に示す．ここで，音素情報を利用しない場合の BIC では，GMM による手

表 6　オンライン話者識別のシステム評価 DER(%)

GMM BIC		
手法 2 （混合数 128）	音素情報 利用なし	母音＋鼻音 （提案法）
11.8	4.5	2.8

法と同様に, 対数パワーが 16 以上の特徴量を話者識別に用いることとしている. また, 識別結果は閾値θ，およびペナルティ項の重みαを変化させたときの最良の結果（closed 条件）を示している．GMM による話者識別では，全ての混合数について，GMM 手法 1 よりも GMM 手法 2 の方が DER は低かった．また，最も精度が良かった GMM 手法 2 の混合数 128 に比べて，BIC による話者識別は，音素情報を利用しない場合も「母音＋鼻音」の場合も DER が 5.8%低い.

話者判定を誤った発話を，その誤りの要因となりえるもの（あいづち，雑音，背景音楽, 短い発話であったなど）で分類した結果を表 5 に示す. GMM による手法では，誤り発話の多くが発話内に短いあいづちや雑音を含んでいた．一方，BIC の話者判定誤りは，背景に音楽がある発話や 2 秒以下の短い発話のみであった．このことから，BIC による手法は GMM に比べ，発話内の雑音などの影響を受けにくく，よりロバストな話者判定が行えるのではないかと考えられる.

次に，オンライン話者識別システム全体の評価をするため，BIC によりオンラインで話者交替点を逐次検出しつつ，検出された話者交替点毎に切り出された発話の話者判定を行う比較実験を実施した．開発データには，評価データの前週の同番組を使用し，閾値θ，およびペナルティ項の重みαを決定した（open 条件）.

GMM 手法 2（混合数 128）と BIC による話者識別結果を表 6 に示す．音素情報を利用しない場合の BIC でも，GMM に比べて DER は低かった．「母音＋鼻音」の場合には，識別精度は更に改善し，GMM の場合に比べて DER は 9.0%低いことを確認した.

6.　おわりに

本稿では，音素認識による音素情報を利用したオンライン話者識別手法を提案した. 識別実験により, 窓幅以前の話者を逐次確定する場合において, 窓幅が約 10 秒以下では，提案法は従来法と同程度か及ばなかった．しかし，話者の判定を話者交替点毎に行う場合には，提案法により DER が 4.0%から 2.8%へ改善した．また，提案手法と，GMM による従来手法の比較検討も行った．識別実験では，真の話者交替点毎に切り出された発話を話者判定する場合において，closed 条件で，提案手法は GMM による手法よりも DER が 5.8%低いことを確認した．また，話者交替点を BIC により検出しながら話者判定する場合においても，open 条件で，提案手法は，DER が 9.0%低いことを確認した．今後は，提案手法によるオンライン話者識別結果の音声認識への適用法などについて検討を進めていく.

参考文献

1) 本間真一，小林彰夫，奥貴裕，佐藤庄衛，今井亨，都木徹：ダイレクト方式とリスピーク方式の音声認識を併用したリアルタイム字幕制作システム，映像情報メディア学会論文誌, Vol.63, No3, pp.331-338 (2009)
2) Zhang, Z., Furui, S. and Ohtsuki, K.: On-line incremental speaker adaptation with automatic speaker change detection, in Proc. IEEE Int. Conference on Acoustics, Speech and Signal Processing (ICASSP), Vol. Ⅱ, pp961-964 (2000)
3) 小林彰夫，奥貴裕，本間真一，佐藤庄衛，今井亨，都木徹：コンテンツ活用のための報道番組自動書き起こしシステム, 情報処理学会研究報告. SLP, 音声言語情報処理, Vol.2009, No.20 (2009)
4) Markov, K. and Nakamura, S.: Never-Ending Learning System for Online Speaker Diarization, in Proc. ASRU, pp.699-704 (2007)
5) 中川聖一，森一将，：発話間の VQ ひずみを用いた話者交替識別と話者クラスタリング，信学論, J85-DII, 11, pp.1645-1655 (2002)
6) Liu, D., Kubala, F.: Fast Speaker Change Detection for Broadcast News Transcription and Indexing, EUROSPEECH'99, Vol. 3, pp.1031-1034 (1999)
7) Liu, D., Kubala, F.: Online Speaker Clustering, in Proc. IEEE Int. Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.333-336 (2004)
8) Chen, S. and Gopalakrishnam, P.: Speaker, Environment and Channel Change Detection and Clustering via the Bayesian Information Criterion, in Proc. 1998 DARPA Broadcast News Transcription and Understanding Workshop, pp.127–132 (1998)
9) Tranter, S. and Reynolds, D.: An Overview of Automatic Speaker Diarization Systems, IEEE Trans. ASLP, Vol.14, no.5, pp.1557–1565 (2006)
10) 奥貴裕，佐藤庄衛，小林彰夫，本間真一，今井亨：音素情報を利用した対談番組におけるオンライン話者識別，日本音響学会，講演論文集, 3-1-13(2009.9)
11) Imai, T., Sato, S., Homma, S., Onoe, K. and Kobayashi, A.: Online speech detection and dual-gender speech recognition for captioning broadcast news, IEICE Trans. Inf. & Syst., Vol.E90-D, no.8, pp.1286-1291 (2007)
12) Reynolds, D., Quatieri, F, Dunn, R.: Speaker Verification Using Adapted Gaussian Mixture Models, Digital Signal Processing, 10, pp.19-41 (2000)
13) 網野加苗，菅原勉，荒井隆行：聴取による話者識別における音韻間の格差と音響的対応，信学技報, SP2004-164, pp.1-6 (2005)