音声認識と話者認識を統合した話者の人名付与システム

西田昌史[†] 緒方淳[†] 有木康雄[†]

本研究では,「クリントン大統領が，情報スーパーハイウェイについて話しているシーンを見たい」といった，特定の話者がある話題について話しているシーンの検索を目指している．このような話者と発話内容を同時検索するには，話者の交替を検出し，発話区間に対して話者の名前を付与し，重要語を検出する必要がある．そこで，本研究では，まず話者セグメンテーションにより話者の交替を検出し，話者モデルを自動構築する．次に，大語彙連続音声認識とワードスポッティングにより，ニュース音声中から人名および話者の交替を促すフレーズ（キーフレーズ）を抽出する．抽出された人名およびキーフレーズを利用して，自動構築された話者モデルに話者の名前を付与する．この人名インデキシングと，大語彙連続音声認識による重要語検出により，話者と発話内容を同時検索することが可能となる．

Speaker Name Indexing System by Integrating Speech Recognition and Speaker Recognition

Masafumi Nishida,[†] Jun Ogata[†] and Yasuo Ariki[†]

The purpose of this study is to retrieve a video clip where a specific speaker talks about some topics, for example, “I would like to watch a video clip where President Clinton talks about information super highway”. In order to retrieve the speaker name and the spoken contents simultaneously, it is required to detect speaker changes, index the speaker name to the obtained speaker section and extract important words. In this study, the speaker changes are detected by performing the speaker segmentation and a speaker model is automatically constructed. A phrase suggesting the speaker change as well as the speaker name in a news speech data is extracted by large vocabulary continuous speech recognition and word spotting technique. Thus, the extracted speaker names are automatically indexed to the speaker section obtained by the speaker segmentation. Therefore, we can simultaneously retrieve the speaker name and the spoken contents based on the speaker name indexing and the important words extracted by the large vocabulary continuous speech recognition.

1. はじめに

近年，放送のディジタル化により，多くのテレビ番組が放映されるようになった．放送がディジタル化されると，放映された番組をハードディスクに録画しておくことができるホームサーバが，各家庭に設置されるようになる．そうなると，ユーザは，ホームサーバから見たい番組や，その中の関心のある場面を検索したり，番組の内容を要約して見たいといった要求を持ったりするようになる．この要求に対応するには，番組の内容に対して，音声や映像といったメディアの解析を通して，索引情報を自動付与しておく必要がある．

我々は，これまでニュース番組を対象に，音声や映像メディアを解析して，ニュース番組の構造化ならびに検索に関する研究を行ってきた[1)~3)]．ニュース番組の構造化に関しては，話者セグメンテーションによるアナウンサー発話の自動抽出と大語彙連続音声認識を統合化したニュース記事の自動分類手法を提案した[1)]．これをふまえて，本研究では,「クリントン大統領が，情報スーパーハイウェイについて話しているシーンを見たい」といった，特定の話者がある話題について話しているシーンの検索を目指している．この処理を実現するためには，話者の交替を検出して，各話者の名前を付与し，大語彙連続音声認識により重要語を抽出しておく必要がある．ここでいう重要語とは，特定の話者が話している話題を表すキーワード，たとえば情報スーパーハイウェイなどを示している．

ニュース音声や対話における話者セグメンテーションおよびクラスタリングの研究がさかんに行われている[4)~8)]．これまで，我々は，ニュース音声中のアナウンサー発話の自動抽出[9)]，座談会やドラマに対する話

† 龍谷大学理工学部
Faculty of Science and Technology, Ryukoku University

者セグメンテーション[9),10)]を行ってきた．これらの研究では，特定の話者をあらかじめ学習することなく，入力音声中から話者モデルを自動学習し，部分空間法による話者照合[11)]に基づいて自動的に話者のセグメンテーションを行っている．本研究では，この話者セグメンテーション法をベースとして，話者の学習データ量に応じて話者判定の閾値を設定する方法を提案する．

人名の付与に関しては，ニュース映像中の顔の名前付けを自動的に行う Name-It という手法が提案されている[12)]．Name-It は，ニュース映像中から顔を検出するとともに，クローズドキャプションならびにテロップから名前を抽出し，両者を対応付ける手法である．しかし，ニュースでは顔が写っているからといって，その話者がしゃべっているとは限らない．そこで，本研究では，大語彙連続音声認識とワード スポッティングによりニュース音声中から人名を抽出し，対応する話者の名前を付与する方法を提案する．これは，ニュース音声に対して大語彙連続音声認識を行い，人名および話者の交替を促すフレーズを抽出する方法である．本研究では，話者の交替を促すフレーズをキーフレーズと呼ぶ．この際，大語彙連続音声認識では，辞書にない未知語を認識することができないため，あらかじめ用意された人名辞典により，人名スポッティングを新たに行う[13)]．したがって，本研究では，話者セグメンテーションにより自動構築された話者モデルに対して，大語彙連続音声認識とワード スポッティングにより抽出した話者の名前を付与する．この人名インデキシングと，大語彙連続音声認識により得られた重要語を基に，話者と発話内容を同時検索することが可能となる．本手法により，98 年 8 月の NHK15 分間のニュースに対して，話者セグメンテーション，大語彙連続音声認識，人名スポッティングを行い，人名インデキシング実験を行った．

2. 話者セグメンテーション

2.1 部分空間法による話者照合

観測空間で観測される話者 s の学習音声データを n 次元特徴ベクトルの集合 $\{x_t^{(s)}\}$ $(t = 1, 2, \cdots, N)$ とする．この学習データから，平均ベクトル $\mu^{(s)}$ および分散共分散行列 $R^{(s)}$ を次式で求める．

$$\mu^{(s)} = \frac{1}{N}\sum_{t=1}^{N} x_t^{(s)} \tag{1}$$

$$R^{(s)} = \frac{1}{N}\sum_{t=1}^{N}(x_t^{(s)} - \mu^{(s)})(x_t^{(s)} - \mu^{(s)})^T \tag{2}$$

この分散共分散行列 $R^{(s)}$ を固有値分解すると，次式のようになる．

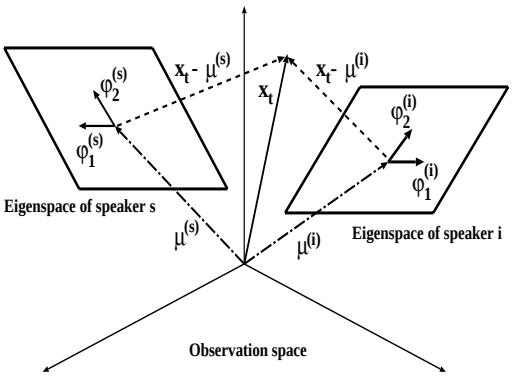

図 1 話者固有空間
Fig. 1 Speaker eigenspace.

$$R^{(s)} = \Phi^{(s)}\Sigma^{(s)}\Phi^{(s)T} \tag{3}$$

ここで，$\Sigma^{(s)}$ は，分散共分散行列 $R^{(s)}$ の固有値 $\lambda_i^{(s)}(i = 1, \cdots, k, \cdots, n)$ を対角成分に持つ対角行列である．また，$\Phi^{(s)}$ は，分散共分散行列 $R^{(s)}$ の固有ベクトル $\varphi_i^{(s)}(i = 1, \cdots, k, \cdots, n)$ を列ベクトルとする行列である．

図 1 に示すように，固有ベクトル $\varphi_i^{(s)}$ は正規直交基底ベクトルとして，話者 s の音声データを表現する固有の空間（話者固有空間）を形成している．この意味で，$\Phi^{(s)}$ は，話者性を表していると考えられる．

話者固有空間を構成する場合，固有値 $\lambda_i^{(s)}$ が大きい上位 k 個の固有ベクトル $\varphi_i^{(s)}$ を正規直交基底ベクトルとして，k 次元話者固有空間を構成する．

照合は，本人 s であると申告された話者の固有空間と，入力特徴ベクトルの集合 $\{x_t\}$ との平均距離を式 (4) の投影距離により求める．

$$PD^{(s)} = \frac{1}{N}\sum_{t=1}^{N}\{||x_t - \mu^{(s)}||^2 - \sum_{i=1}^{k}(x_t - \mu^{(s)}, \varphi_i{}^{(s)})^2\} \tag{4}$$

ここで，N は入力音声の総フレーム数である．

3 次元の観測空間における投影距離の概念図を**図 2** に示す．図 2 で話者 s の固有空間は，正規直交基底ベクトル $\varphi_1{}^{(s)}$ と $\varphi_2{}^{(s)}$ で張られる空間として表されている．式 (4) で表される話者 s の投影距離とは，話者 s の学習データの平均ベクトル $\mu^{(s)}$ と入力ベクトル x_t との差ベクトルの長さ $||x_t - \mu^{(s)}||^2$ から，差ベクトル $x_t - \mu^{(s)}$ を話者 s の固有空間へ射影した射影ベクトルの長さ $\sum_{i=1}^{k}(x_t - \mu^{(s)}, \varphi_i{}^{(s)})^2$ を差し引いたものとして定義される．

式 (4) により求めた投影距離が閾値より小さければ，

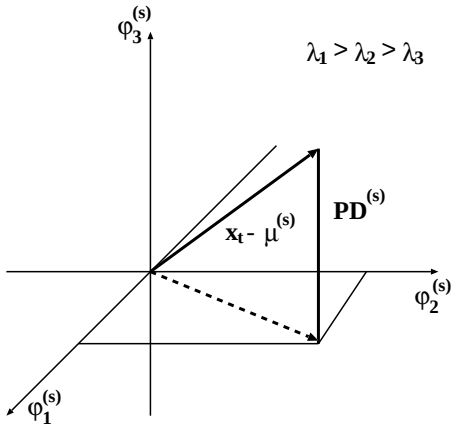

図 2 投影距離
Fig. 2 Projection distance.

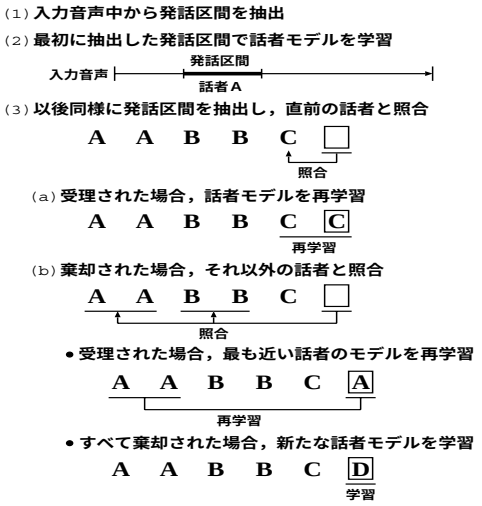

図 3 話者セグメンテーション法
Fig. 3 Speaker segmentation method.

本人の音声であると判定する．

2.2 話者セグメンテーション法

ここでは，異なる話者が発話したそれぞれの区間のことを話者区間と呼ぶ．同一話者がポーズをおいてしゃべっても，同一話者の区間と見なす必要がある．

部分空間法による話者照合に基づいた話者セグメンテーション法を**図 3** に示し，処理内容を以下に述べる．ここで，本手法では，話者交替には無音をともなうと仮定して処理を行っている．また，話者を判定する閾値は，話者モデルごとに設定している．

(1) 入力音声中のパワーを求め，閾値処理により音声区間を検出する．これにより抽出された無音から無音までの区間を発話区間と呼び，この発話区間に対して，以下の処理を行う．

(2) 最初に抽出した発話区間で話者モデル（話者固有空間）を学習するとともに，話者照合のための閾値を式 (5) により設定する．この発話区間が最初の話者区間となる．

(3) 次の発話区間に対して，直前の発話区間の話者（以後，直前の話者と呼ぶ）と照合する．

(a) 受理された場合，直前までの話者区間と今抽出した発話区間を合わせて，その話者の新しい話者区間とし，話者モデルの再学習，閾値の再設定を行う．

(b) 棄却された場合，直前の話者以外と照合し，閾値を下回った話者の中で最も距離が小さい話者に判定する．この場合，(a) と同様に，話者モデルの再学習，閾値の再設定を行う．一方，直前の話者以外と照合した結果，閾値を下回る話者がいなければ，新しい話者と判断する．今抽出した発話区間をその話者の話者区間とし，話者モデルの学習，閾値の設定を行う．

ここで，話者の再学習の際，学習データ量が 30 秒を超えた場合，再学習は行わない．これは，話者の発話時間に偏りができ，学習データ量が多い特定の話者モデルに照合されやすくなることを避けるためである．予備実験として，学習データ量の上限を変化させて話者のセグメンテーションを行った．その結果，学習データ量の上限を 30 秒より大きくすると，学習データ量が多い話者に受理される誤りが増加したため，本研究では，学習データ量の上限を 30 秒に設定した．

このように最初に抽出した発話区間で話者のモデルを学習し，以後抽出した発話区間に対して，直前の話者モデルとの照合を繰り返すことで，事前に話者モデルを作成しておくことなく，話者交替の検出と話者の識別が可能になる．

同一話者か新しい話者かを判定する話者照合の閾値 θ は，学習に用いた発話区間中の特徴ベクトル集合と，学習時に作成された話者固有空間との投影距離の平均を m，標準偏差を σ として次式のように設定している．

$$\theta = m + w\sigma \tag{5}$$

$$w = \frac{1}{\alpha \log T} \tag{6}$$

式 (6) 中の T は，話者の学習データ量（秒）を表している．従来法では，学習データ量を考慮せずに閾値を設定していたため，学習データ量が少ない場合，学習話者のモデルが話者性を十分表現できていないため，閾値が小さくなり本人を誤って棄却しやすくなってし

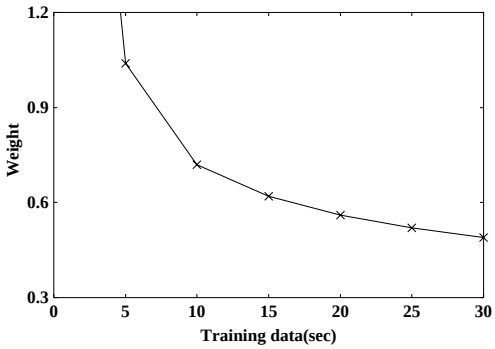

図 4 学習データ量と重みの関係
Fig. 4 Relation between training data and weight.

表 1 分析条件
Table 1 Analytical condition.

サンプリング周波数	12 kHz
高域強調	$1 - 0.97z^{-1}$
フレーム長	20 ms
フレーム周期	5 ms
窓タイプ	ハミング窓
特徴パラメータ	24 次メルフィルタバンク 18 次 MFCC

まう．また，学習データ量が多くなってくると，学習話者のモデルが話者性を十分表現することができ，閾値が大きくなり他人を誤って受理しやすくなってしまう．それに対して，本手法は，学習データ量に対応して動的に閾値を設定するために，式 (6) に示す関数を用いた．学習データ量と重み w の関係を**図 4** に示す．図のように，この関数は，学習データ量が少ない場合に重みが大きくなり，学習データ量が多くなるにつれて重みが小さくなる滑らかに減衰する関数である．つまり，閾値 θ は重み w によって，学習データ量が少ない場合に大きくなり，本人が受理されやすくなる．また，学習データ量が多い場合には閾値 θ が小さくなり，他人が棄却されやすくなる．このように，式 (6) に示す関数を用いることで，学習データ量に応じた照合が可能となる．

2.3 話者セグメンテーション実験

2.3.1 実験条件

98 年 8 月の NHK15 分間のニュース 4 日分を用いて，話者セグメンテーション実験を行った．以後，このニュース音声に対して，大語彙連続音声認識，ワードスポッティング，人名インデキシングの実験を行った．音声の分析条件を**表 1** に示す．

本研究では，評価用データとは異なるニュースデータを用いた予備実験の結果，話者固有空間の次元数を 6 次元，式 (6) の係数 α を $\alpha = 0.55$ に設定した．

表 2 話者セグメンテーション結果
Table 2 Speaker segmentation result.

	話者境界数	%
話者交替 検出率	50 / 71	70.4
話者交替 適合率	50 / 63	79.4
	発話区間数	%
話者識別率	402 / 493	81.5

実験の評価は，話者交替検出率・適合率，話者識別率で評価した．ここでいう話者識別率は，本手法により直前の話者あるいはそれ以外の話者と照合を行い，正しく話者を識別できた割合を表している．これらは，次式で定義される．

$$\text{話者交替検出率} = \frac{\text{正しく検出した話者間の境界数}}{\text{全ニュース中の話者間の境界数}} \quad (7)$$

$$\text{話者交替適合率} = \frac{\text{正しく検出した話者間の境界数}}{\text{検出した話者間の境界数}} \quad (8)$$

$$\text{話者識別率} = \frac{\text{正しく話者を識別した発話区間数}}{\text{全ニュース中の発話区間数}} \quad (9)$$

2.3.2 実験結果と考察

話者セグメンテーション実験を行った結果を**表 2** に示す．また，提案手法による学習データ量に対応した閾値設定の有効性を示すために，閾値を固定した場合に対する話者セグメンテーション実験を行った．式 (6) に示す閾値の重み w を変化させて実験を行った結果，重み w が 0.7 のときに最も精度が高く，話者交替検出率 59.2%，適合率 71.2%，話者識別率 66.7%という結果が得られた．したがって，閾値を固定した手法に比べて，本手法が有効であることが分かった．

本研究では，話者交替の検出において，1 発話区間分でも話者の境界がずれて検出された場合，検出誤りと判断している．話者の照合においては，新たな話者の学習データが少なすぎると，以後棄却されてしまったり，話者が交替しているにもかかわらず，発話区間が短いために受理されてしまうといった誤りがあった．

3. ニュース音声の自動書き起こし（大語彙連続音声認識）

3.1 認識システムの構成

大語彙連続音声認識システムとしては，ワードグラフを中間結果とする 2-pass 構成のシステムを用いている (**図 5**)[14]．まず，1st-pass で単語 bigram を用いた lexical tree search を行い，認識結果をもとにワードグラフを作成する．このとき，最もスコアの高い単語にのみ back-off 接続を行う，最尤単語 back-off 接続を用いることにより，認識精度を落とすことなく処理時間を大幅に削減している[15),16]．2nd-pass では，ワー

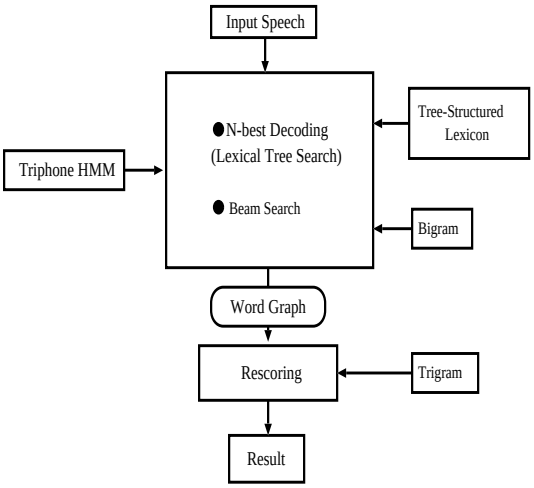

図 5 認識システム
Fig. 5 Recognition system.

表 3 音響分析と HMM
Table 3 Acoustic analysis and HMM.

音響分析	サンプリング周波数	16 kHz
	特徴パラメータ	MFCC（39 次元）
	フレーム長	20 ms
	フレーム周期	10 ms
	窓タイプ	ハミング窓
HMM	状態数	5 状態 3 ループ
	タイプ	triphone
	混合数	12
	学習方法	連結学習

ドグラフに登録された 1st-pass の音響尤度と trigram を用いてリスコアリングを行う．

3.2 大語彙連続音声認識実験

3.2.1 実験条件

音響モデルには，前後の音素環境を考慮した triphone HMM を用いた．音響モデルの学習には，まず ATR 連続音声データベース a〜j セットの 6 名分のデータの視察ラベルを用いて初期モデルを作成し，次に日本音響学会新聞記事読み上げコーパス（JNAS）のうち，男性話者 137 名分の 21782 発話を用いて連結学習を行った．音響分析条件と HMM のトポロジを**表 3** に示す．

言語モデルには，IPA モデル 98 年度版のうち，語彙サイズが 20 K であり，bigram と trigram の cut-off の閾値がそれぞれ 4 のモデルを用いている．言語モデルの学習データは，毎日新聞記事 75ヵ月分である[17]．

評価用データは，98 年 8 月の NHK15 分間のニュース 4 日分（2.3.1 項と同様）である．

3.2.2 実験結果

大語彙連続音声認識の実験結果を**表 4** に示す．参考として，アナウンサー発話のみを認識した結果も示す．ここで，“Corr” は単語正解率，“Acc” は単語正解精度を示しており，“Anchor” はアナウンサー発話のみを認識した結果，“All” はアナウンサーのほかに，インタビュアーやレポータ発話なども含めた結果を表している．

表 4 大語彙連続音声認識結果（%）
Table 4 LVCSR result (%).

	Corr	Acc
Anchor	80.1	78.4
All	38.7	35.4

全体的に認識率は低くなっているが，まず Anchor 部分に関しては，言語モデルの学習データの時期的な差異などが原因と考えられる．また，全体の結果（All）においては，インタビュアーなどの完全な自由発話や外国語発話なども多く含まれており，そのような発話はほとんど認識不能であったため，さらに認識率を低下させる原因となった．

3.2.3 大語彙連続音声認識結果からの人名，キーフレーズ検出

本研究における話者交替検出，人名インデキシングシステムにおいては，音声認識により得られた人名や話者交替を促すフレーズ（キーフレーズ）が重要となる．そこで，本研究での大語彙連続音声認識システムにおいて，人名，キーフレーズがどの程度検出されたかを調べた．なお，人名やキーフレーズは，アナウンサーだけでなく，レポータも発話しているので，アナウンサーのみではなくレポータの発話も含めた音声認識結果から，人名とキーフレーズの検出を行った．

検出対象の人名としては，比較的著名な 50 人分の名前リスト（たとえば，クリントン大統領，野中官房長官...）をあらかじめ用意した．実際に用いたニュースデータ中で出現した人名は，50 名のうち 23 名であった．また，23 名分の人名が出現した回数は，全部で 82 回であった．

キーフレーズとしては，予備調査として評価データ以外のニュース中において，実際に話者交替時に検出されたもの 12 種類（**表 5**）を用意した．実際の評価データ中に出現したキーフレーズの種類数は，12 種類（あらかじめ用意したものすべて）であった．また，12 種類のキーフレーズが出現した回数は，全部で 24 回であった．

人名，キーフレーズの検出結果を**表 6** に示す．実験結果より，キーフレーズは比較的高い検出率を得たが，人名の検出率は非常に低い結果となった．この原因としては，言語モデルの時期差の関係が考えられる．す

表 5 キーフレーズリスト
Table 5 Key phrase list.

次のように述べています
次のように述べました
次のように発言しています
次のように発言しました
お聞きください
考えを示しました
非難しました
語りました
反論しました
考えを明らかにしました
述べました
見通しを示しました

表 6 人名，キーフレーズの検出結果
Table 6 Speaker name and key phrase detection result.

	検出率
人名	41.5%(34/82)
キーフレーズ	79.2%(19/24)

なわち，言語モデルの学習において，人名は言語モデル学習データの時期に強く依存するのに対し，キーフレーズはどの年代のニュースにも独立に出現する．また，人名，キーフレーズとも，レポータなどの背景雑音が比較的多い区間に関しては，誤認識が多かった．

4. 人名スポッティング

ここでは，人名検出の改善法について述べる．先に述べたように，N-gram 言語モデルを用いた大語彙連続音声認識では，人名においては言語モデルの学習データの時期差に強く依存するといった問題（未知語の問題）がある．そこで，本研究においては，あらかじめ用意した人名リストを用いて人名スポッティング（キーワードスポッティング）を行い，それと 3.2 節の大語彙連続音声認識結果を併用することで人名検出の改善を行う．

人名スポッティングは，サブワードデコーダの併用によるリジェクションを用いることで行う[18),19)]．具体的には，**図 6** に示す (a)，(b) 2 種類の言語モデルを用いる．(a) はキーワード（ここでは人名に相当）とフィラーモデルにより構成されており，キーワードモデルと呼ぶ．フィラーモデルは，キーワード以外の区間を近似するもので，ここでは任意の音素を用いている．(b) は音素モデルの出現のみを許したもので，ここではバックグラウンドモデルと呼ぶ．(b) を用いたデコーディングは，サブワードデコーダに相当する．

入力音声に対して，2 種類の言語モデルを利用してそれぞれビタビアルゴリズムに基づき，式 (10)，(11) で示される尤度を算出する．

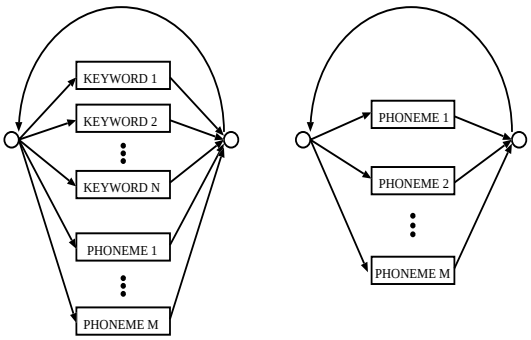

図 6 キーワードスポッティング用言語モデル
Fig. 6 Language model of keyword spotting.

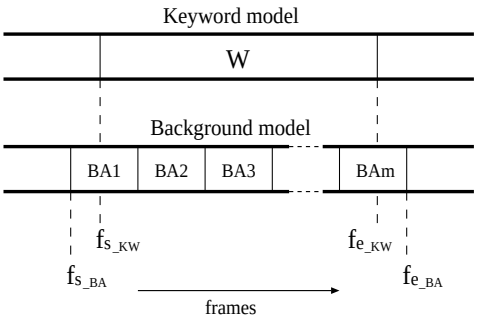

図 7 2 つのモデルでの尤度区間の対応
Fig. 7 Correspondence of likelihood section for two models.

$$S_{KW}(W)=\frac{\text{単語 } W \text{ が認識されたときの対数尤度}}{\text{フレーム数}} \quad (10)$$

$$S_{BA}(W)=\frac{\text{単語 } W \text{ に対応する，音素列の対数尤度}}{\text{フレーム数}} \quad (11)$$

ここで $S_{KW}(W)$ は，キーワードモデル中の単語 W が認識されたときの対数尤度をそのときのフレーム数（**図 7** で $f_{e-KW}-f_{s-KW}$）で割ったものであり，$S_{BA}(W)$ はバックグラウンドモデルでの音素列の対数尤度を，単語 W に対応するフレーム数（図 7 で $f_{e-BA}-f_{s-BA}$）で割ったものである．図 7 に単語 W が認識されたときの各尤度とフレーム数の対応を示す．

湧き出し単語のリジェクションは，式 (12) により求めた対数尤度差 $S(W)$ が，ある一定の閾値以下の値のものに対して行った．

$$\text{対数尤度差 } S(W)=S_{KW}(W)-S_{BA}(W) \quad (12)$$

4.1 人名スポッティング実験

4.1.1 実験条件

音響モデルには，音素環境独立な HMM（monophone HMM）を用いた．音素数は，無音を含む 41 種

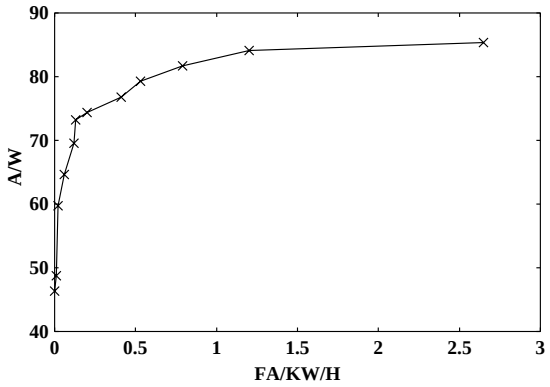

図 8 人名スポッティング結果 1
Fig. 8 Speaker name spotting result 1.

表 7 人名スポッティング結果 2
Table 7 Speaker name spotting result 2.

検出率（%）	湧き出し誤り率（%）
85.4	2.7
79.3	0.5
46.3	0.0

類である．学習データは，日本音響学会新聞記事読み上げコーパス（JNAS）のうちの 21782 発話（男性話者 137 名分）である．音響分析条件等は，3.2.1 項で述べた音響モデルと同様である．

評価用データは，98 年 8 月の NHK15 分間のニュース 4 日分（2.3.1 項と同様）で，継続時間長は 1.05 時間である．スポッティング対象単語（人名）の出現回数は 82 回，語彙サイズは 50 である．

4.1.2 実験結果

人名スポッティングの結果を**図 8**，**表 7** に示す．ここで，図 8 中の A/W は検出率，$FA/KW/H$ は単位時間あたりの湧き出し誤り率を表しており，湧き出し単語のリジェクションの閾値を変化させたときの，検出率と湧き出し誤り率を示している．人名スポッティング実験では，レポータやインタビュアーなどの背景雑音が比較的多い発話に関しては精度が低く，特に湧き出し誤りが多くなっていた．今回は，ワードスポッティングのフィラーモデルとして単純な音素連接モデルを用いたが，今後は，さらに頑健かつ汎用的なフィラーモデルを検討する必要がある．

後述する人名インデキシングにおいては，これまでに述べた話者交替やキーフレーズ，そしてここで述べた人名スポッティング結果を併用するため，人名スポッティング時に起こった湧き出し誤り単語を，ある程度抑制することが可能である．しかしながら，湧き出し誤り単語が大量に起こった場合，そのような方法でも人名インデキシングに悪影響を及ぼすと考えられる．したがって，人名スポッティングにおける湧き出し誤り単語は，ある程度少ない方が望ましい．この点から，検出率 79.3%，湧き出し誤り率 0.5%のときに得られた人名を，以降の実験で用いることにする．

表 8 人名スポッティング結果 3
Table 8 Speaker name spotting result 3.

手法	検出率（%）
(a)	41.5
(b)	79.3
(c)	**81.7**

大語彙連続音声認識結果と人名スポッティング結果との併用による人名検出実験を行った．すなわち，大語彙連続音声認識においては，個々の人名が誤って湧き出すことはほとんどないと考えられることから，大語彙連続音声認識結果と人名スポッティング結果から，どちらか一方に出現していたら，その人名の検出を行うようにする．**表 8** に，大語彙連続音声認識結果のみ (a)，人名スポッティングのみ (b)，そして両方の併用 (c) による人名検出結果をそれぞれ示す．実験結果より，両者の併用により人名検出率が 81.7%と向上していることが分かる．実際，評価データの大語彙連続音声認識結果中には，人名の湧き出し誤りはまったくなかったことから，両者の併用による人名検出は，以降の人名インデキシング実験に有効であると考えられる．

5. 人名インデキシング

4 章で，大語彙連続音声認識とワードスポッティングの併用により抽出された人名と話者の交替を促すフレーズ（キーフレーズ）から，話者セグメンテーションによって得られた話者ラベルに，対応する話者の名前を付与する．

抽出された人名ならびにキーフレーズを発話している話者は，その発話区間の前後で発話しているインタビューイー（インタビューされる人）に関して述べており，アナウンサーあるいはレポータであると見なすことができる．したがって，抽出された人名とキーフレーズを含む発話区間の直前あるいは直後で，その話者と異なる話者ラベルの発話区間をインタビューイーと見なして，話者の名前を付与する．キーフレーズを含む発話区間の前後を判断する方法としては，“次のように” を含むキーフレーズの場合は，直後の発話区間に人名を付与する．それに対して，その他のキーフレーズの場合は，直前と直後の発話区間の話者ラベルを調べて，ラベルが異なる発話区間に人名を付与する．また，前後とも話者ラベルが異なっていれば，発話区

間数が少ない方に人名を付与する．

この方法により，人名のインデキシング実験を行った．実験データは，98年8月のNHK15分間のニュース4日分（2.3.1項と同様）である．今回用いたニュース音声には，インタビューイー区間が18個あり，そのうちキーフレーズが前後に存在していたインタビューイー区間は，10個であった．なお，キーフレーズがなかったインタビューイー区間としては，国会の答弁で複数の話者が次々に質問を行っている場面があった．したがって，この10個のインタビューイー区間に対して，大語彙連続音声認識とワードスポッティングの併用により抽出された人名とキーフレーズによる人名インデキシングを行った結果，7個のインタビューイー区間に対して，話者の名前を付与することができた．残りの区間は，人名あるいはキーフレーズのどちらかが抽出できなかったために，話者の名前を付与することができなかった．したがって，人名とキーフレーズの検出精度，つまり音声認識精度を向上させることで，人名の付与率を改善することができると考えられる．

今後は，話者セグメンテーションにより得られた話者ラベルに対して，抽出された人名を付与し，人名と話者モデルとの共起を利用した人名のインデキシングを行う．また，人名の抽出法として，音声のみではなくテロップの人名を検出することで，音声と画像を統合化した人名のインデキシングについて検討する．さらに，話者と発話内容の同時検索システムの構築について検討する．

6. むすび

本研究では，「クリントン大統領が，情報スーパーハイウェイについて話しているシーンを見たい」といった，特定の話者がある話題について話しているシーンの検索を目指して，話者の交替を検出して，各話者の名前を付与する人名インデキシングについて検討を行った．話者交替の検出においては，部分空間法による話者照合に基づいた話者セグメンテーションを行い，話者の学習データ量に応じた話者判定のための閾値設定法を提案した．また，大語彙連続音声認識により話者の交替を促すフレーズ（キーフレーズ）を抽出し，大語彙連続音声認識とワードスポッティングの併用により人名を抽出することで，話者セグメンテーションによって自動構築された話者モデルに，対応する話者の名前を付与する方法を提案した．

今後の課題としては，人名と話者モデルとの共起や，音声とテロップを統合化した人名インデキシングについて検討し，話者と発話内容の同時検索システムの構築について検討していく予定である．

参考文献

1) 西田昌史，緒方　淳，有木康雄：アナウンサー発話の自動抽出とディクテーションによるニュース記事分類，情報処理学会論文誌，Vol.40, No.4, pp.1482–1490 (1999).
2) 鷹尾誠一，緒方　淳，有木康雄：ニュース音声記事データベースにおける観点の自動抽出と構造化，信学技報，DE2000-12, pp.89–96 (2000).
3) 鷹尾誠一，舩本純一，有木康雄，緒方　淳：ニュース音声データベースに対するクロスメディア検索，*MIRU2000*, pp.457–462 (2000).
4) 森　一将，山本一公，中川聖一：発話間のVQ歪みを用いたオンライン話者交替識別と話者クラスタリング，信学技報，SP2000-18, pp.17–24 (2000).
5) 村井則之，小林哲則：MLLRによる話者適応と統計的発話交代モデルを用いた複数話者対話音声の認識，信学技報，SP2000-14, pp.31–38 (2000).
6) Sonmez, K., Heck, L. and Weintraub, M.: Speaker Tracking and Detection with Multiple Speakers, *Eurospeech99*, Vol.5, pp.2219–2222 (1999).
7) Delacourt, P., Kryze, D. and Wellekens, C.J.: Detection of Speaker Changes in an Audio Document, *Eurospeech99*, Vol.3, pp.1195–1198 (1999).
8) Tritschler, A. and Gopinath, R.: Improved Speaker Segmentation and Segments Clustering using the Bayesian Information Criterion, *Eurospeech99*, Vol.2, pp.679–682 (1999).
9) 西田昌史，有木康雄：自動学習による話者セグメンテーション，信学技報，SP97-57, pp.1–6 (1997).
10) Nishida, M. and Ariki, Y.: Speaker Indexing for News Articles, Debates and Drama in Broadcasted TV Programs, *ICMCS99*, Vol.2, pp.466–471 (1999).
11) 西田昌史，有木康雄：話者固有空間における動的・静的特徴統合による話者照合，信学論，Vol.J83-DII, No.12, pp.2536–2544 (2000).
12) 佐藤真一，中村裕一，金出武雄：Name-It：動画像処理と自然言語処理の統合による映像内容アクセス手法，第3回知能情報メディアシンポジウム，pp.187–194 (1997).
13) 西田昌史，緒方　淳，有木康雄：話者名の自動索引付けと話者・発話内容の同時検索，第15回人工知能学会全国大会，1A4-04 (2001).
14) Ortmanns, S., Ney, H. and Aubert, X.: A Word Graph Algorithm for Large Vocabulary Continuous Speech Recognition, *Computer Speech and Language*, Vol.11, No.1, pp.43–72 (1997).

15) Ogata, J. and Ariki, Y.: An Efficient Lexical Tree Search for Large Vocaburaly Continuous Speech Recognition, *ICSLP2000*, Vol.2, pp.967–970 (2000).
16) 緒方 淳，有木康雄：back-off 接続を考慮した大語彙連続音声認識の高速化，日本音響学会平成12年度春季研究発表会，2-8-8, pp.43–44 (2000).
17) 李 晃伸，河原達也：大語彙連続音声認識エンジン Julius における A*探索法の改善，情報処理学会研究報告，99-SLP-27-5, pp.33–39 (1999).
18) Knill, K.M. and Young, S.J.: Speaker Dependent Keyword Spotting for Accessing Stored Speech, *CUED/F-INFENG/TR 193* (1994).
19) 緒方 淳，有木康雄：ニュース記事分類におけるディクテーションとワードスポッティングの比較，信学技報，SP98-32, pp.67–72 (1998).

(平成13年11月16日受付)
(平成14年 4 月16日採録)

西田 昌史（正会員）

昭和49年生．平成9年龍谷大学理工学部電子情報学科卒業．平成11年同大学大学院修士課程修了．平成14年同大学院博士後期課程修了．工学博士．現在，科学技術振興事業団さきがけ研究21「協調と制御」領域ポスドク研究員．話者認識，音声認識に関する研究に従事．電子情報通信学会，日本音響学会各会員．

緒方 淳

昭和51年生．平成10年龍谷大学理工学部電子情報学科卒業．平成12年同大学大学院修士課程修了．現在，同大学院博士後期課程在学中．音声認識に関する研究に従事．日本音響学会会員．

有木 康雄（正会員）

昭和25年生．昭和51年京都大学大学院修士課程修了．昭和54年同大学院博士課程修了．昭和55年京都大学工学部情報工学科助手．平成2年龍谷大学理工学部電子情報学科助教授，平成4年教授，現在に至る．工学博士．昭和62〜平成2年エディンバラ大学客員研究員．画像処理，音声情報処理に関する研究に従事．電子情報通信学会，日本音響学会，人工知能学会，画像電子学会，映像メディア学会，IEEE 各会員．