ヒューマンインタフェース 65－18
音声言語情報処理 10－18
（1996. 3. 1）

ニュース音声中の語彙反復による情報検索
—部分空間射影に基づく話者正規化の応用—

田頭茂明　　有木康雄
龍谷大学 理工学部
〒 520-21 大津市瀬田大江町横谷 1-5

あらまし 音声を入力インターフェースとした情報検索システムの構築を試みた。具体的には、マルチメディア情報の一つとして、ニュース音声を対象としている。「日本はＰＫＯに参加することになりました」と言うようなニュース音声を聞いていて、＜ＰＫＯ＞と言う用語を知らない場合、その場で「そのＰＫＯってどういう意味？」と知らない用語を音声で反復して尋ねることができるシステムである。これにより、情報を得ているメディアと同じメディアを使って情報検索することができる。このシステムでは検索対象のキーワードをニュース音声とユーザー発話との共通区間と設定している。共通区間を切り出す場合に問題となる話者性の違いに対しては、部分空間射影に基づく話者正規化を用いている。

キーワード：共通区間、自由発話、話者正規化、テレビニュース、未知語、情報検索

An Enquiring System of Unknown words In TV News by Spontaneous Repetition
- Application of Speaker Normalization by Speaker Subspace Projection -

Shigeaki Tagashira and Yasuo Ariki
Faculty of Science and Technology, Ryukoku University
1-5, Yokotani Ooe-Cho Seta Otsu-Shi 520-21, Japan

Abstract We tried to construct a system of enquiring unknown words, by spontaneous repetition, appearing in spoken sentences of TV. For example, we hear "Japan would join PKO." from TV news and if "PKO" is an unknown word, then we can enquire it by saying "What's the PKO?" The system recognizes the word "PKO" and explains its meaning. The system estimates a common section between news speech and user speech and recognizes the word corresponding to the common section. We solved a problem of speaker difference in extracting common sections by speaker subspace projection.

英文 Key words：Common Sections, Spontaneous Speech, Speaker Normalization, TV news, Unknown Words, Information Retrieval.

1 はじめに

マルチメディア情報では、情報を得ているメディアは、文字だけでなく音声であり映像である。このようなメディアに対しては、情報を得ているメディアそのものを使って、インタラクティブに情報検索することが望ましい。例えば、音声で情報提供を受けている場合には、その内容について音声で質問するというのが自然である。

このことから、音声を入力インターフェースとした情報検索システムの構築を試みた。具体的には図1に示すように、マルチメディア情報の一つとして、ニュース音声を対象としている。「日本はＰＫＯに参加することになりました」と言うようなニュース音声を聞いていて、＜ＰＫＯ＞と言う用語を知らない場合、その場で「そのＰＫＯってどういう意味？」と音声で尋ねることができるシステムである。その結果としてＰＫＯの意味が検索され、音声で説明を聞くことができる。

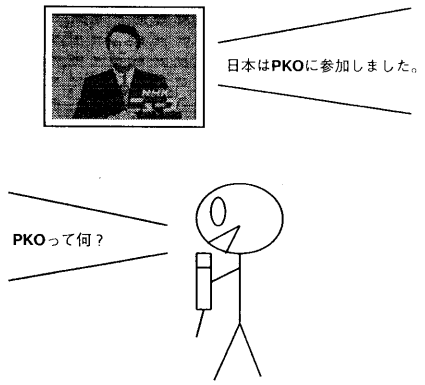

図 1: システムのイメージ図

また入力音声としては、人間のごく自然な発話である自由発話を対象としている。自由発話では「あのー」「えーと」などに代表される間投詞や、言い淀み、言い誤りおよび言い直しなどが頻繁に出現する。このことから構文駆動の認識システムで、自由発話の音声を認識することは非常に難しいと考えられる [1]。そこで、我々はワードスポッティングに着目し、連続音声中の重要な単語のみを認識する手法をシステムに採り入れた。

ワードスポッティングの手法として様々なものが提案されている [2] [3] [4] [5]。しかし、HMMを使用した手法は種々の制約と認識率からみて、良好な結果が得られていないと考えられる。本研究では、ニュース音声を聞きながら、知らない用語を問い合わせるというタスクに限定しているので、図2に示すように、まず、ニュース音声とユーザー発話との共通区間をキーワードとして切り出し、その後切り出した共通区間に対してHMMを用いて認識するという手法を用いた。

共通区間を抽出する手法としては伊藤らが提案している Reference Interval-free 連続 DP(RIFCDP) を用いた [6]。これは連続 DP を拡張したものであり、テンプレートと入力波形（ここではニュース音声とユーザ発話）との整合度を求め、整合度の高い区間を共通区間（検索したい区間）として抽出してくる手法である。しかし RIFCDP の基本はＤＰマッチングであるため、アナウンサーとユーザという異なる話者間では特徴量が異なり、切り出し精度が劣化すると予測される。

この問題に対して、本研究では部分空間射影に基づく話者正規化で対処している。この方法は、ますアナウンサーの音声データをよく表現する低次元の部分空間を設定する。この部分空間をアナウンサーの話者性と考え、この部分空間内で表された音声データを話者正規化された音声データと考える。ユーザの音声データが与えられると、この音声データをよく表現する低次元の部分空間を設定し、部分空間内で音声データを表して話者正規化する。このとき二人の部分空間の軸の相関を最大にすることにより、話者の違いを吸収する方法である。

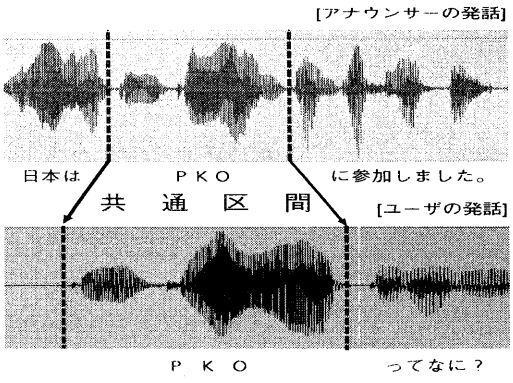

図 2: 共通区間の切り出し

2 システムの構成

システムの構成を図3に示す。システムは、音声入力部、話者正規化部、共通区間切り出し部、単語認識部、情報検索部の5つからなる。まず、音声入力部でニュース音声とユーザ発話を録音し、話者正規化部で音声を正規化した後、共通区間切り出し部でニュース音声とユーザ発話の共通区間を切り出す。切り出した共通区間に対して、単語認識部で知りたい単語を認識し、情報検索部で認識した単語の情報を検索する。

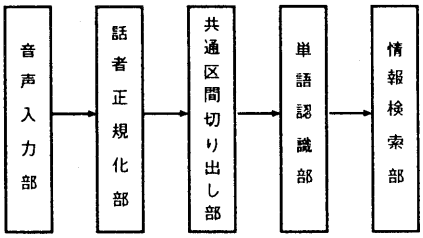

図 3: システムの構成図

2.1 音声入力部

ニュース音声とユーザの発話を入力する。ニュース音声としては、ユーザの発話が行なわれた時刻より5秒前を入力開始時刻とし、ユーザー発話の終了をもって終了時刻とする。サンプリング周波数は12 KHzである。ここで音響分析も同時に行なう。分析条件を表1に示す。ユーザ発話中に混入するニュース音声は、実際上、ほとんど問題とならない程小さい。

表 1: 情報検索システムの分析条件

標本化周波数	12KHz
高域強調	$1-0.97z^{-1}$
時間窓	ハミング窓
分析区間長	20ms
分析周期	5ms
特徴量	16次LPCケプストラム

2.2 話者正規化部

話者正規化部は、次の共通区間切り出し部での前処理的意味をもっている。DPマッチングはHMMとは違い確率構造としてモデルを保持していないため、話者の違いに対して弱いという問題がある。この問題に対して、話者正規化を用いて話者の違いを吸収することにより、DPの照合を精度よく行なえることが期待できる。本研究では、我々が既に提案したCLAFIC正準相関分析法を用いて話者正規化を行なっている[7]。

2.2.1 CLAFIC正準相関分析法

この話者正規化法は、ある話者の音声データに対して、このデータをよく表現する少数の軸を設定し、この軸で形成される部分空間を話者性と考える。またこの部分空間内での音声データの相対位置を音韻性と考える。話者性の違いを部分空間の違いと考え、各話者の部分空間を形成する軸の相関を話者間で最大にするという基準で、それぞれの軸を設定し話者の違いを吸収する。具体的には、図4に示すようにある話者A（モデル話者）が発話した音声データX_AからCLAFIC法[8]により、部分空間（モデル空間）を求めておく。

CLAFIC法は、観測空間内に存在するある話者の音声データから相関行列を求め、それを固有値分解して得られる固有ベクトルを、部分空間の軸とする方法である。次に、異なる認識話者Bが発話した音声データX_Bに対して、モデル空間の軸と相関が高くなるように正準相関分析をおこなって、認識話者Bの部分空間（認識空間）を設定する方法である。

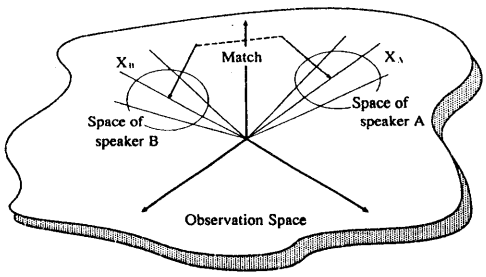

図 4: 話者正規化のイメージ図

それぞれの空間の軸相関を最大にすることにより、モデル空間と認識空間での相対位置が一致すると考えられる。処理の手順は次の通りである。

STEP(1) 話者 A の音声データ X_A を基に CLAFIC 法により正規直交基底 V_A を求める。

STEP(2) 話者 A の音声データ 1 文と話者 B の音声データ 1 文を DP マッチングして特徴ベクトル間の対応付を行う。これを複数の文に適用して X_A, X_B を求める。

STEP(3) 話者 A の軸 v_A を固定し、話者 B の音声データ X_B から話者 A の軸に相関の高い軸 v_B を次式により求める。

$$v_B = \frac{\sqrt{C}\Sigma_{22}^{-1}\Sigma_{21}v_A}{\sqrt{v_A^T\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}v_A}} \qquad (1)$$

ここで C は軸 v_A の分散、Σ_{12}、Σ_{21} はそれぞれ話者 A と話者 B の相互相関行列であり、Σ_{22} は話者 B の自己相関行列である。話者 A と話者 B の音声データを、こうして得られた各々の話者空間に射影して正規化する。

2.2.2 予備実験

この話者正規化が、ＤＰのようなデータ照合に対しても有効であることを確かめるための実験を行なった。テンプレートを作成した標準話者とは異なる話者を入力話者として、話者正規化を行い、ＤＰによる単語認識実験を行なった。実験条件としては、東北大・松下単語音声データベース 212 単語を用いた。212 単語のうち 162 単語を話者正規化に用いて、残り 50 単語を認識対象とした。テンプレートを作成した標準話者は男性 (sp210) であり、認識話者は男 (sp301) 女 (sp606) 各 1 名で、テンプレートの話者とは異なる話者である。

2.2.3 結果

結果を表 2に示す。男性 (sp301) の場合で、正規化しない場合より正規化した場合で１０％の認識率の向上が見られ、女性 (sp606) の場合についても３０％の向上が得られた。認識に用いた部分空間の次元数は 16 次元である。

2.2.4 考察

話者正規化しない場合で、男性である sp301 は女性の sp606 より認識率が高いが、正規化すると男性、女性に関係なく良好な認識結果が得られている。男女間での差がなくなっていることから話者性を吸収でき、本研究で提案している話者正規化が、ＤＰのようなデータ照合に対しても有効であることがわかる。

表 2: DP による５０単語の認識結果 (%)

	男性 (sp301)	女性 (sp606)
話者正規化なし	84.0	62.0
話者正規化あり	94.0	92.0

2.3 共通区間切り出し部

入力したニュース音声とユーザの発話間で共通区間を切り出す。２つの音声データ間で任意の共通区間を取り出す手法として、Reference Interval-free 連続 DP(RIFCDP) を用いている [6]。

RIFCDP は標準パターン中の任意の区間と、入力音声中の任意の区間との間で整合度を計算し、共通区間の切り出しを可能にした手法である。この概念図を図 5に示す。標準パターン中の τ 番目のフレームを Γ_τ、あるいはフレーム τ とする。

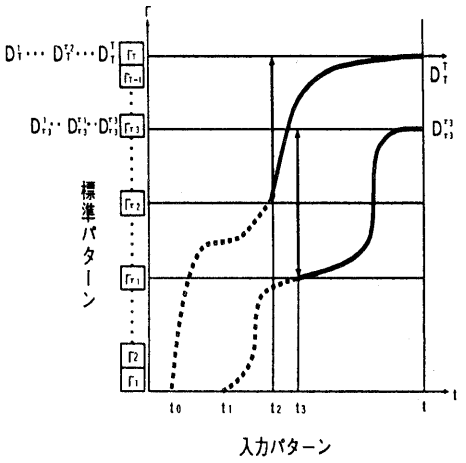

図 5: RIFCDP の概念図 [文献 [6] より引用]

標準パターン中の任意の区間、例えば図 5で、時刻 t におけるフレーム τ_1 からフレーム τ_3 まで

の整合度($D_{\tau_3}^{\tau_1}$)を求めるには、フレームτ_3に至るまでの累積距離と、その最適パス上でのフレームτ_1通過時の累積距離（時刻t_3でフレームτ_1の累積距離）が必要である。そこで、標準パターン側の各フレームにそのフレームに至る最適パス上の累積距離の履歴を与えることにする。例えば図のように、フレームτ_3では、フレームτ_3に至る最適パス上で、フレームτまでの累積距離$D_{\tau_3}^{\tau}$ ($1 \leq \tau \leq \tau_3$)の履歴(τ_3)個を与える。（累積距離履歴と同様に重み係数履歴、始端時刻履歴も必要）。これにより、フレームτ_1からフレームτ_3までの整合度は、これらの累積距離の差を累積重み係数で正規化すれば求められる。この整合度を全フレーム、全時刻に対して求め、閾値またはローカルピークから共通区間を抽出する。

2.4 単語認識部と情報検索部

共通区間切り出し部の結果である切り出された区間に対して、不特定話者のHMMにより単語を認識する。区間を限定しているので連続音声認識ではなく、単語認識であり、認識率の精度は向上する。また、共通区間切り出しにおいては、波形的に照合しているだけなので、目的としているキーワードの一部分や、キーワード以外の部分など複数の共通区間を切り出す場合がある。そこで、複数の共通区間に対して単語認識を行なった後、HMMの出力確率が最大のものを認識結果として出力している。

また、単語認識部での認識結果を、情報検索部で時事用語事典により検索し、検索結果を音声とテキストで出力する。

3 システムの評価実験

システムの性能評価実験を行なった。実験は音声入力部、話者正規化部、共通区間切り出し部、単語認識部、情報検索部を通して行ない、検索結果で評価した。

3.1 実験条件

音声データはNHKの5分間のニュース30日分から取得した。ニュース音声とユーザの発話、それぞれ20文を用いて2.2で述べた話者正規化を行ない、それぞれの部分空間を設定した。部分空間の次元数は16次元である。音声の分析条件は表1に示したとおりである。キーワードとしては表3に示す25単語を設定した。評価では、キーワードを含んだニュース文とユーザーの発話をそれぞれ25文用意し、目的のキーワードを検索できれば正解とした。

表3: キーワードの一覧

建設省	日米首脳会談	エリツィン大統領
サハ共和国	ハバロフスク地方	モスクワ
CIS	通信衛星	郵政省
光ファイバー	北朝鮮	河野外務大臣
サミット	OECD	オウム真理教
薬事法違反	有権者	連立政権
中小企業対策	ベトナム	オイルショック
OBサミット	青酸ガス	硫酸
気象庁		

3.2 結果

結果を表4に示す。話者正規化しない場合において72.0%、話者正規化した場合で80.0%の認識結果を得た。正規化することにより8.0%の認識率の向上が得られた。

表4: システムの評価結果(%)

	認識率
話者正規化なし	72.0
話者正規化あり	80.0

4 考察

間違って認識したものについて分析すると、以下の3つに分類できる。

(1) 共通区間が正確に切り出せていない。
例：「中小企業対策」

(2) 共通区間は切り出せているが、認識を誤る。
例：「郵政省」→誤「建設省」
例：「ベトナム」→「モスクワ」

(3) キーワードを正しく切り出しているが、共通区間が複数あるため、認識の段階で棄却さ

れている。例：「OECD」、「OBサミット」

(1)で切り出した区間は、キーワード以外の区間やキーワードの一部の区間であり、認識部に影響を与えて、認識誤りになったものである。これは、話者正規化しても、まだ切り出せない部分があることを意味しており、今後さらに精度をあげる必要がある。また(2)は、共通区間は切り出せているが、認識において誤ってしまった場合で、紛らわしい単語において誤認識が生じていた。例えば、「郵政省」と「建設省」などの単語で、間違える傾向があった。これは、HMMの認識精度をあげることにより回避できると考えられるので、正規化後不特定話者HMM [9] をこのシステムに採用すれば改善できると考えられる。また、(3)についても(2)と同様なことがいえ、HMMの精度を上げることによって、キーワード以外のわきだしを棄却できるものと考えられる。

5 おわりに

ニュース音声中の知らない用語に対して情報検索するシステムの構築を試みた。このシステムは人間のごく自然なインターフェースである音声を入力とし、自由な発話で未知用語を反復することにより問い合わせできるシステムである。発話の共通区間を検出するためにワードスポティングの手法としてＲＩＦＣＤＰを用いている。ＲＩＦＣＤＰは話者に大きく依存するという問題点をもつが、それに対してCLAFIC正準相関分析法による話者正規化で対処した。

今後の課題としては、認識に用いている不特定話者HMMを話者正規化した不特定話者HMMに変更することによって、さらに認識率の向上を試みる予定である。

参考文献

[1] 北, 川端, 斉藤: “HMM音韻認識と拡張ＬＲ構文解析法を用いた連続音声認識”, 情報処理学会論文誌, Vol. 31, No.3, pp.472-479, 1990-03.

[2] 安藤彰男, 今井亨: “音声認識を用いた放送番組リクエストシステム”, オーディオビジュアル複合情報処理, 10-4, 1995-09.

[3] 村上仁一: “フレーム同期型フルサーチアルゴリズムを用いた連続音声認識と自由発話への応用”, 信学技報, SP95-32, 1995-06.

[4] P.Jeanrenaud,K.Ng,M.Siu,J.R.Rohlicek,H.Gish: “Phonetic-Based Word Spotter:Various Configurations and Application to Event Spotting”, Proc. ESCA EuroSpeech93, pp.1057-1060, 1993.

[5] 河原, 宗続, 堂下: “ヒューリスティックな言語モデルを用いた会話音声中の単語スポッティング”, 信学論 D-II, Vol.J78-DII-No.7 , pp.1013-1020, 1995-06.

[6] 伊藤, 木山, 小島, 関, 岡: “標準パターンの任意区間によるスポッティングのための Reference Interval-free 連続ＤＰ (RIFCDP)”, 信学技報, SP95-34, 1995.

[7] 田頭, 西島, 有木: “話者部分空間への写像による話者認識と話者正規化”, 信学技報, SP95-28, 1995.

[8] オヤ, 小川・佐藤訳: “パターン認識と部分空間法”, 産業図書 , pp.61-89, 1986.

[9] 田頭 茂明, 有木 康雄: “部分空間射影による話者正規化を用いた不特定話者HMM”, 信学技報, SP95-98, 1995.