音声言語情報処理 21−5
(1998. 5. 29)

自動抽出されたアナウンサー発話に対するニュースディクテーションと記事分類

緒方 淳　　西田 昌史　　有木 康雄

龍谷大学 理工学部

〒 520-2194 大津市瀬田大江町横谷 1-5
Tel: 077-543-7427
E-mail: {ogata,nishida}@arikilab.elec.ryukoku.ac.jp, ariki@rins.ryukoku.ac.jp

あらまし　ビデオ・オン・デマンドを目指したニュースデータベースを構築するには、ニュース記事を分類しておく必要がある。本研究では、ニュース音声に対してディクテーションを行ない、キーワードを抽出することにより、自動的に記事の分類を行なう。記事を分類する上では、アナウンサーの発話区間のみをディクテーションすれば十分であり、処理の短縮にもつながる。しかし、人手でニュース音声中からアナウンサーの発話区間を切り出すのは現実的ではない。そこで、本研究では、アナウンサーの発話区間のみを自動的に抽出した場合と、レポータなどを含めた場合に対するディクテーションを行ない、記事の分類精度の比較を行なう。

キーワード：話者照合、部分空間法、話者セグメンテーション、音声ディクテーション、記事分類

News Dictation and Article Classification for Automatically Extracted Announcer Utterance

Jun Ogata　　Masafumi Nishida　　Yasuo Ariki

Faculty of Science and Technology, Ryukoku University

1-5 Yokotani, Oe-cho, Seta, Otsu-shi, 520-2194 Japan
Tel: +81-77-543-7427
E-mail: {ogata,nishida}@arikilab.elec.ryukoku.ac.jp, ariki@rins.ryukoku.ac.jp

Abstract　In order to construct a news database with a function of video on demand (VOD), it is required to classify a news articles into topics. In this study, we describe a system which can dictate news speech, extract keywords and classify news articles based on the extracted keywords. We propose that it is sufficient to dictate only the announcer utterance for classifying the news articles and it contributes to reduce the processing time and increases the classification accuracy. As an experiment, we compared the classification performance of news articles between in two cases; in the case of dictating only the announcer utterances which are automatically extracted and in a case of dictating a whole speech which includes reporter or interviewer utterances.

Key words : speaker verification, subspace method, speaker segmentation, speech dictation, article classification

1 はじめに

近年、放送の多チャンネル化により、多くのニュース番組が放映されるようになった。これを受けて視聴者には、知りたいニュースだけを見たいという要求が生じている。この要求に対応するには、ニュース記事を分類してデータベースを構築しておく必要がある。このとき、人手でニュース記事を分類することは不可能であり、機械によるニュース記事の自動分類が望まれる。

本研究では、ニュース音声に対してディクテーションを行ない、キーワードを抽出することにより、自動的に記事分類を行なうことを目的としている。ニュース音声は、レポータあるいはインタビューの発話部分に比較的雑音が重畳している場合が多いので、レポータあるいはインタビューの発話をディクテーションすると、単語を正しく認識できないと考えられる。このことから記事を分類する上では、比較的雑音が少ないアナウンサーの発話区間のみをディクテーションすることにより、雑音の影響を抑えることができ記事の分類に必要なキーワードを抽出できると考えられる。また、アナウンサーの発話のみをディクテーションすることにより、処理の短縮にもつながる。しかし、人手でニュース音声中からアナウンサーの発話区間を切り出すのは現実的ではない。そこで本研究では、アナウンサーの発話区間のみを自動的に抽出した場合と、レポータなどを含めた場合に対するディクテーションを行ない、記事の分類精度の比較を行なう。

本論文では、ニュース音声中からアナウンサーの発話区間を自動的に抽出するために、一般的にセキュリティ応用を目的として研究されている話者照合 [1][2][3] の技術を用いる方法を提案する。現在、話者照合技術は、技術的には高い照合精度が得られているが、時期差に対するロバストネスが問題となっている [4] [5]。一方、音声メディア処理としても応用可能な技術である。例えば、座談会などで特定の人の発言を拾い出して聞くことも可能となるであろう [6]。また、討論会における話者のインデキシングを話者識別の技術を用いて行う研究が報告されている [7]。これは、あらかじめ討論会の参加者の音声を学習しておき、その話者モデルにより話者識別を行って話者のラベル付けを行なう方法である。これに対して、本研究では、ニュース音声を対象としているため、アナウンサーの話者モデルをあらかじめ学習しておくことは現実的でない。なぜならニュース番組では、アナウンサーが日によって替ることもあり、またあらかじめアナウンサーの話者モデルを学習しておくと、時期差に対処しなくてはならなくなるからである。

そこで、本研究では、特定の話者をあらかじめ学習することなく、入力音声中から各話者の話者モデルを自動学習し、話者照合に基づいて自動的に話者区間を切り出す方法を提案する。話者照合の方法としては、リアルタイム処理のために、部分空間法を用いている。この手法を用いて、NHK5 分間のニュース 45 日分に対して、アナウンサーの発話区間の切り出し実験を行ない、本手法の有効性を評価した。

次に、記事の分類手法としては、単語 bigram と不特定話者 HMM を用いて、ニュース音声のディクテーションを行ない、その結果得られるキーワードをもとに、「朝日新聞データベース分類表索引」を用いて、記事を 10 の分野に分類する。

本研究では、NHK5 分間のニュース 48 記事に対して、アナウンサーの発話区間のみを自動的に抽出した場合とレポータなどを含めた場合に対するディクテーションを行ない、記事の分類精度の比較を行なった。

2 アナウンサーの発話区間の抽出

2.1 話者照合

話者照合 [8] とは、入力音声と同時に自分が誰であるかの ID を入力して、その音声が本当にその ID に対応する人の発話であるかどうかを判定するものである。図 1 に示すように、入力音声と本人の標準パターンとの距離が、閾値よりも小さければ本人の発話であると判定し、そうでなければ他人の発話であると判定するものである。

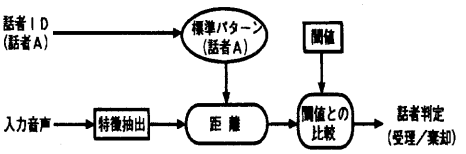

図 1: 話者照合

2.2 主成分分析法

本研究では、リアルタイム処理のために、従来法に比べて計算量の少ない部分空間法 (主成分分析法) [9] を話者照合のベースにしている。図 2 に主成分分析法に基づく話者照合の概念図を示す。

主成分分析法による話者照合では、話者ごとに入力音声データ $x_k(1 \leq k \leq N)$ から平均ベクトル $\mu^{(i)}$ を求め、式 (1) により共分散行列 $R^{(i)}$ を求める。ここで、(i) は話者の識別番号を表す。

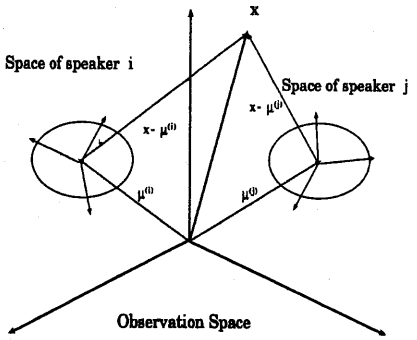

図 2: 主成分分析法

$$R^{(i)} = \frac{1}{N}\sum_{k=1}^{N}(x_k - \mu^{(i)})(x_k - \mu^{(i)})^T \qquad (1)$$

話者ごとの共分散行列 $R^{(i)}$ を固有値分解して正規直交ベクトル $v_j^{(i)}, j = 1,2,\cdots,p^{(i)}$ を求め、式(2)により射影行列 $P^{(i)}$ を求める。

$$P^{(i)} = \sum_{j=1}^{p^{(i)}} v_j^{(i)} v_j^{(i)T} \qquad (2)$$

本人であると申告された話者の部分空間に対して、入力音声データの系列 $x_k(1 \le k \le N)$ との距離を式(3)により求める。この距離が閾値より小さければ、本人の音声であると判定する。

$$\begin{aligned} y^{(i)} &= \frac{1}{N}\sum_k \|x_k - \{P^{(i)}(x_k - \mu^{(i)}) + \mu^{(i)}\}\|^2 \\ &= \frac{1}{N}\sum_k \|(I - P^{(i)})(x_k - \mu^{(i)})\|^2 \end{aligned} \qquad (3)$$

2.3 アナウンサーの発話区間抽出法

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、同一話者の区間とみなす必要がある。また、1つの音声区間を発話区間と呼ぶ。

話者区間の切り出し方法[6]を以下に示す。

1. 入力音声の1秒毎に、1秒間の平均パワーを求め、閾値処理により音声か無音かを判定する。これにより抽出された無音から無音までの各音声区間に対して、以下の処理を行う。
2. 最初に抽出した音声区間で話者モデルを学習するとともに、話者照合のための閾値を設定する。この音声区間が最初の話者区間となる。
3. 次の音声区間に対して、直前に発話した話者と照合する。照合した結果、
   (a) 直前の話者と判断されれば、直前までの話者区間と今抽出した音声区間を合わせて、その話者の新しい話者区間とする。この話者区間に対して話者モデルの再学習、閾値の再設定を行なう。
   (b) 直前の話者でないと判断されれば、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断する。(a)と同様に、話者モデルの再学習、閾値の再設定を行なう。一方、直前の話者以外と照合した結果、閾値を下回った話者がいなければ、新しい話者と判断し、今抽出した音声区間をその話者の話者区間とする。この話者区間に対して話者モデルの学習、閾値の設定を行なう。

話者区間の切り出し方法を図3に示す。

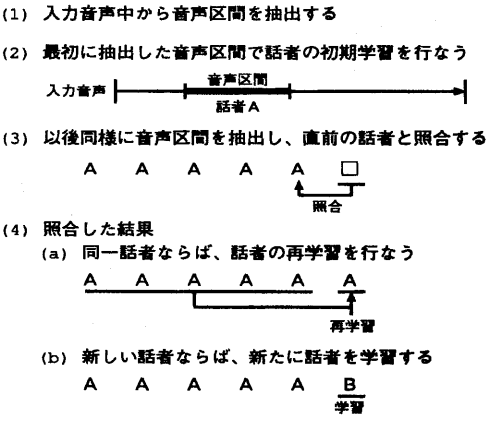

図 3: 話者区間の切り出し法

既知話者と未知話者を判定する閾値 θ は、学習時に作成された部分空間と、学習に用いた音声との距離の平均 μ と標準偏差 σ を求めて、次式のように設定している。

$$\theta = \mu + \frac{\sigma}{3} \qquad (4)$$

2.4 アナウンサーの発話区間抽出実験

2.4.1 実験条件

NHKの5分間のニュース45日分を用いて、アナウンサーの発話区間を切り出す実験を行なった。部分

空間の次元数は、最も切り出し率が高かった7次元とした。実験条件を表1に示す。

表 1: 実験条件

データ	NHK5 分間のニュース 45 日分
サンプリング周波数	12kHz
フレーム長	20ms
フレーム周期	5ms
窓タイプ	ハミング窓
特徴抽出	LPC ケプストラム (16 次)
部分空間の次元数	7
閾値 θ	$\theta = \mu + \frac{\sigma}{3}$

実験の評価は、アナウンサーの発話区間切り出し率・適合率で評価した。これらは、次式で定義される。

$$\text{切り出し率} = \frac{\text{アナウンサーと正しく認識した発話区間数}}{\text{全ニュース中のアナウンサーの発話区間数}} \quad (5)$$

$$\text{適合率} = \frac{\text{アナウンサーと正しく認識した発話区間数}}{\text{アナウンサーとして切り出された発話区間数}} \quad (6)$$

ここで、アナウンサーの判定方法は、ニュース一日分毎に話者を切り出し、最も長時間発話している話者をアナウンサーと判定している。

今回用いたNHK5分間のニュース45日分すべてに、レポータあるいはインタビューが含まれている。

2.4.2 実験結果と考察

アナウンサーの発話区間切り出し実験を行なった結果を表2に示す。

表 2: アナウンサーの発話区間切り出し結果 (%)

アナウンサー切り出し率	92.1
アナウンサー適合率	91.1

発話区間の切り出しでは、各話者の初期学習に用いる音声が短かすぎると、閾値が低く設定され、以後その話者の音声は棄却されやすくなる。また、話者モデルの作成後、各話者の発話が短い場合にも、同一話者の発話を棄却されたり、異なる話者の発話を同一話者として受理されやすくなる。さらに、発話区間に雑音が重畳していると、誤って話者を認識する場合がある。

3 ディクテーション

3.1 実験条件

今回、用いた言語モデルは、毎日新聞 CD-ROM 版の 45ヶ月分 (91 年 1 月～ 94 年 9 月) の記事から学習したものである。形態素解析結果は、毎日新聞記事の解析結果である RWC テキストデータベースによるものである。語彙数 5K の back-off bigram で、cut-off は 2 とした。

音響モデルは、男性不特定話者音素 HMM で、1 状態あたりの混合数は 8、音素数は 41 種類の monophone モデルである。学習には、日本音響学会新聞記事読み上げ音声コーパスのうち、男性話者 137 人分の 21782 発話を用いた。音響特徴量には 39 次元の特徴パラメータ (12 次元のメルケプストラム係数とパワー、およびそれぞれの Δ、ΔΔ係数) を用いた。

評価用音声データは、NHK5 分間のニュース 45 日分から選び出した 48 記事を用いた。各記事は、アナウンサーとレポータあるいはインタビュアーを含んでいる。この 48 記事に対するアナウンサーの発話区間切り出し率は 92.6%、適合率 82.9% であった。これにより切り出された発話に対してディクテーションを行なう。

3.2 連続音声認識実験

ディクテーションは、アナウンサーの発話区間抽出実験により、切り出された各々の発話に対して行なった。認識評価は、切り出されたアナウンサーのみの発話区間 (Anchor)、アナウンサー以外のインタビューやレポーター等の発話区間 (Other)、両方ともに含んだ発話区間 (All)、の 3 セットについて求めた。

評価用データ 3 セットそれぞれのテストセットパープレキシティを表3に示す。

前節までに述べた音響モデルと言語モデルを用いて、beam-search を用いたビタビデコーディングを行なった。デコーダーには HTK(HMM Toolkit)[10] を用いた。実験結果を表4に示す。

今回用いた音声データは比較的、背景雑音等が多く、全体的に認識精度は低い値となった。特に Other セットのインタビュー等の区間では特に背景雑音が多く、またテストセットパープレキシティも高い値を示し、認識精度はかなり低い結果となった。

表 3: 評価用ニュース音声データ

	発話数	5K 未知語率	perp
Anchor	247 発話	13.9%	153.7
Other	116 発話	29.3%	285.2
All	363 発話	20.6%	177.6

perp: test-set perplexity

表 4: 連続音声認識実験結果

	単語誤り率	単語正解率	単語正解精度
Anchor	39.7%	66.5%	60.3%
Other	79.3%	23.5%	20.7%
All	54.6%	48.7%	45.4%

4 記事の分類方法

4.1 ニュース記事の分類手順

本研究でのニュース音声記事の分類は、図4に示すように2段階に分かれている [11]。

(1) まず、ニュース音声に対してディクテーションを行ない、キーワード列とキーワードの存在確率 $Ps(w)$ を求める。

(2) 次に、キーワードが記事の分類（トピック）に寄与する割合と、求めたキーワードの存在確率をもとにニュース記事を分類する。

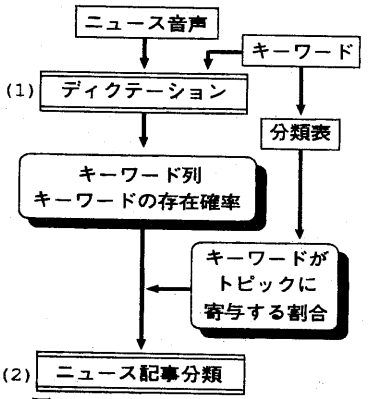

図 4: ニュース記事分類の手順

4.2 寄与率の計算

単語列を基に記事を分類する場合には、まず単語列から分類に有効なキーワードを抽出する。各キーワードと分類の分野（トピック）との関係は、トピック n に対するキーワード w の寄与率 $P(n|w)$ として式(7)で求めておく。

$$P(n|w) = \frac{\text{トピック n に含まれるキーワード w の数}}{\text{全トピックに含まれるキーワード w の総数}} \quad (7)$$

これは、キーワード w を含むトピック n の事後確率である。寄与率を求めるデータベースとして、「朝日新聞記事データベース 分類表索引」 [12] を用いた。これには約1万2千個のキーワードが載っており、キーワード毎に分類付けがなされている。大分類、中分類、小分類とあるが、今回は大分類（10分類）を行なった。例えば、表5の例では、「日米」というキーワードを含む複合語の数は、トピック「政治」では3語、「経済」では2語、「社会」では0語あり、全体で5語あることを表している。この場合、寄与率は以下のように計算される。

表 5: 分類表とキーワードの関係の例

トピック	政治	経済	社会
日米	日米安全保障条約 日米行政協定 日米関係	日米経済摩擦 日米貿易摩擦	
合計	3	2	0

$$P(\text{政治}|\text{日米}) = \frac{3}{5} = 0.6 \quad P(\text{経済}|\text{日米}) = \frac{2}{5} = 0.4$$

$$P(\text{社会}|\text{日米}) = \frac{0}{5} = 0.0$$

記事分類に用いる全てのキーワードについて、この方法により事前に寄与率を求めておく。

4.3 分類の確率計算

抽出したキーワード列が $w_1, \cdots, w_k$ のときの、記事がトピック n に分類される確率 $P(n|w_1, \cdots, w_k)$ は式(8)で求められる。

$$P(n|w_1, \cdots, w_k) = \sum_{i=1,\cdots,k} P(w_i) \times P(n|w_i) \quad (8)$$

ここで、$P(w_i)$ はディクテーションにより求められた単語尤度 $Ps(w_i)$ を用いて、式(9)のようにして求められる。

$$P(w_i) = \frac{Ps(w_i)}{\sum_{j=1,\cdots,k} Ps(w_j)} \quad (9)$$

このようにしてキーワードの音響的な確率と、キーワードの分類に関する知識的な確率を統合する。そして、トピック n に分類される確率 $P(n|w_1, \cdots, w_k)$ が最大値をとるトピックを、その記事のトピックであると判定し分類する。

5 記事分類実験

5.1 実験条件

自動抽出されたアナウンサー発話区間に対するディクテーション結果を基に、記事分類を行なった。また比較として、インタビュー、レポーター等も含んだ場合についても記事分類を行なった。登録したキーワードは、評価用音声データに含まれており、かつ分類表データベースに適合したもので単語である。キーワー

ドの寄与率は、「朝日新聞記事データベース 分類表索引」から求めた。分類数は大分類（総類、政治、経済、労働、文化、科学、社会、事件、スポーツ、国際）の10分類である。

5.2 実験結果

記事分類結果を表6に示す。実験結果の分類率は、記事の総数に対する正解トピックの記事数の割合である。正解トピックはテキストデータを基に正解キーワード列を求め、抽出したキーワードは全て正解なので、単語の存在確率 $Ps(w_i) = 1$ として分類した結果である。

表より、アナウンサー発話区間のみで記事分類を行なった結果、単語正解精度60.3%のとき記事分類率63.6%を得た。比較として行なったAllセットに対する結果と比べると、0.6%程高い結果であった。これは、Allセットのインタビュアー等の区間における単語認識誤りによって、誤分類が生じたものと考えられる。また2つのデータセットそれぞれに対する分類率はほぼ変わらないことから、記事分類をする上ではアナウンサーの発話区間のみでも十分であり、ニュース記事の話題に関する情報は、比較的アナウンサーの発話区間に多く含まれていることがわかる。

表 6: 記事分類結果

	単語正解精度	記事分類率
Anchor	60.3%	63.6%
All	45.4%	63.0%

6 おわりに

部分空間法（主成分分析法）をベースとする話者照合に基づいた方法により、NHK5分間のニュース45日分に対して、アナウンサーの発話区間抽出実験を行なった。その結果、アナウンサーの発話区間切り出し率92.1 %、適合率91.1 %となった。

また、自動抽出されたアナウンサーの発話区間に対して、ディクテーションを行ない記事分類を行なった結果、インタビュアー等を含めた場合よりも分類率を落とすことなく、また処理の短縮にもつなげることができた。今回実験で用いたニュースデータには、アナウンサーの発話区間にも背景雑音が重なっているものもあった。そのようなデータについては、比較的明瞭に、また文法的にも正しいアナウンサー発話区間に対しても、低い認識結果となった。今後、そのような雑音下でもロバストなディクテーションができるよう考慮する必要がある。

参考文献

[1] 松井知子, 古井貞熙: “音韻・話者独立モデルによる話者照合尤度の正規化”, 信学技報, SP94-22, pp.61-66, (1994-06).

[2] 松井知子, 古井貞熙: “テキスト指定型話者認識”, 信学論, Vol.J79-D-II, No.5, pp.647-656, (1996-05).

[3] Konstantin P.Markov, Seiichi Nakagawa: “EVALUATION OF FRAME-BASED LIKELIHOOD NORMALIZATION FOR SPEAKER VERIFICATION”, 日本音響学会 平成9年度春季研究発表会, 2-6-13, pp.69-70, (1997-03).

[4] 松井知子, 西谷隆, 古井貞熙: “話者照合におけるモデルとしきい値の更新法”, 信学論, Vol.J81-D-II, No.2, pp.268-276, (1998-02).

[5] 松井知子, 相川清明: “時期差による発声変動を考慮した話者モデルの生成法”, 日本音響学会 平成9年度秋季研究発表会, 1-1-23, pp.45-46, (1997-09).

[6] 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, (1997-11).

[7] 三村正人, 河原達也, 堂下修司: “パネル討論音声の話者と話題に関する自動インデキシング”, 音声言語情報処理, 11-3, pp.13-18, (1996-05).

[8] 松井知子, 古井貞熙: “VQひずみ、離散/連続HMMによるテキスト独立型話者認識法の比較検討”, 信学論(A), J77-A, 4, pp.601-606, (1994).

[9] エルッキ・オヤ, 小川英光, 佐藤誠訳: “パターン認識と部分空間法”, 産業図書 (1986).

[10] Cambridge University Engineering Department Speech Group and Entropic Research Laboratory Inc.:“HTK Hidden Markov Model Toolkit V2.0”

[11] 緒方淳, 森晴, 有木康雄: “単語bigramを用いた日本語ニュースディクテーションによる記事分類”, 音響学会平成10年度春季大会, 2-Q-16, pp.151-152, (1998-03).

[12] “朝日新聞記事データベース分類表索引”, 朝日新聞社ニューメディア本部, (1992).