修　士　論　文

題　　目

ニュース番組音声を用いたアンカーの発話区間抽出精度向上による音声認識への効果

指導教員

松永　昭一 教授

平成30年度

長崎大学大学院 工学研究科

総合工学専攻

野崎　大智 (52117321)

目 次

図目次

表目次

第1章

研究背景

1.1 研究背景

近年、通信・放送業界では地上デジタル放送の開始や、新たな高速通信規格の誕生など、通信ネットワークの急速な発達が見られる。それに伴い、誰もがテレビやパソコンだけでなくスマートフォン・タブレットなど様々なデバイスを通して手軽に膨大な量の音声・映像データを入手し、好きな時に好きな場所で視聴することが容易な時代となった。しかし、入手できる情報量が増えた分、それら全てが必要であるとは限らず、自分に必要な情報のみを手軽に取捨選択できれば便利である。映像・音声データに、話者や内容のインデックスの情報が付与されていれば、その部分だけを選択して視聴できる。しかし、世の中には膨大な量の映像・音声データが存在するため、それら全てに人手でインデックスを付与することは事実上不可能である。そこで、自動的にインデクシングすることが望まれる。

自動でインデクシングを行うためには、映像・音声データ内の発話区間、発話者、発話内容の特定が必要である。これらを推定する技術のことをダイアライゼーションと呼び、本研究はこの技術の実現を目指す。

本研究では、世の中に存在する映像・音声データの中である特定の人物に情報が集中する形式で行われるニュース番組に着目した。ニュース番組は主にアンカー (司会役のアナウンサー) を中心として、アンカーがレポーターなどに話を振りながら、ニュースが進行していく。また、ニュース番組は収録環境が良いため、研究対象としても適しており、ニュース番組で高精度にダイアライゼーションができると、同じスタイルのその他の映像・音声データにも用いることができると考えられる。そこで、本研究ではニュース番組を対象として研究を行う。

1.2 研究目的

ダイアライゼーションで推定する情報の中で、本研究では発話者の特定に着目した。本研究で対象とするニュース番組には以下の特徴がある。

ニュース番組の特徴

- 30 分程度のニュース番組の中で複数の多様な話題がある
- 1 人または複数のアンカーおよび天気予報士など複数の話者が存在する
- 話者情報（話者数、性別、話者の声質など）および発話区間が未知である

- 同一話者が連続で発話することが多い

このようなニュース番組において、ニュースの話題にインデクシングが行われていることは必要な話題の検索に重要である。ニュース番組のアンカーには以下のような特徴があり、インデクシングに重要な情報を持つと考えられる。

アンカーの特徴

1. 発話数が多い
2. ニュース番組の司会および話題の切り替えを行う
3. ニュース番組の全体にわたって発話している

このため、アンカーの発話区間を検出、音声認識を行うことによって、より高精度なインデクシングが実現可能であると考えた。

先行研究では特定話者の声の特徴に着目し、話者特徴量 (i-vector) を用いることでアンカーの発話区間を検出し、検出精度の皇后が確認された。また、音声認識においても音声認識精度の向上が確認されている。i-vector は、話者識別、話者特定の分野では高い精度を示しているが、短い発話データや雑音を含む発話データからは話者の特徴を十分に抽出できない問題がある。この問題に対して、短い発話の話者識別の検討 [1] やスペクトラルクラスタリングを用いた話者クラスタの作成手法 [2] が提案されているが、問題の解決には至っていない。そこで、ニュース番組では同一話者が連続で発話することが多いことに着目した。

本研究では、前後の発話区間が同一話者の発話である可能性が高いとき発話区間を結合、長い発話を擬似的に作成して話者特徴量の抽出を行い、i-vector 抽出精度の向上を目指す。また、以上の手法で抽出された i-vector を用いてアンカーの発話区間を検出と音声認識を行い、発話区間を結合による効果を検証する。

1.3 論文構成

次章以降における本論文の構成は、まず 2 章で音声認識システムの概要について説明を行う。次に 3 章では話者識別システムの概要として i-vector、コサイン類似度の理論的背景の説明、および使用する音声データの概要の説明を行い、ニュース番組音声における発話間の i-vector のコサイン類似度を用いることによる効果を検証する。4 章では i-vector を用いた単純なアンカーの発話区間抽出アルゴリズムによる発話区間抽出実験を行い、問題提起を行う。5 章では本研究で提案するアルゴリズムの説明を行う。6 章では提案手法を用いた発話区間抽出実験を行い、本研究における提案手法を用いることで話者情報と発話区間が未知の場合におけるアンカーの発話区間抽出への効果を検証する。7 章では 6 章で抽出したアンカーの発話区間の音声認識を行い、どれほど単語が正確に認識されているかを検証する。 8 章では本研究において検証された実験の結果を元に結論を述べる。

第2章

基礎知識

2.1 音源識別

音響データ（ニュース音声、会話音声等）の中にはさまざまな音源種別（声、音楽、雑音等）の音が混在している。音源識別とは、音響データ中に含まれる音源種別を自動的に識別することである。ここでの処理は、音響データのスペクトル解析を行い、音響特徴パラメータを求め、あらかじめ用意した各音源種別の音響特徴パラメータの分布と比較することで音源種別を識別する。

本研究では、ニュース番組の音声データに音響特徴パラメータを用いた音源識別 [3] を用い、音声データ中の音源種別を以下の 4 つに分類した。

(1) 音声区間: アナウンサーやインタビューの声

(2) 音楽区間: オープニングやエンディングなどの音楽、BGM

(3) 背景雑音区間: 自動車走行音や鳥の泣き声

(4) 無音区間: 音量が極めて小さい区間

また、音源識別システムは音響データを各種別へ識別するための音響特徴パラメータの分布に混合ガウス分布を用いている。本研究では、混合数 8 のガウス分布を使用している。本研究の音響特徴パラメータを表 2.1、各音源種別の識別性能 F 値を表 2.2 に示す。ただし、表の評価値とは、音源種別間のデータ数を考慮するために、各音源種別の F 値に各音源種別の割合をかけ、合計したものである。

表 2.1: 音源識別のための音響特徴パラメータ

スペクトルの変化	スペクトルの傾き
白色雑音との近さ	ピッチ
パワー	中心周波数
中心周波数のバンド幅	

表 2.2: 音源識別システムの F 値

	F 値 [%]
Speech	94.84
Music	83.32
Noise	63.73
Pause	77.43
評価値	87.19

以下に、音源識別のためのスペクトル解析と音響特徴パラメータについて説明する。

2.1.1 スペクトル解析

音響データのスペクトル解析の手法として最も一般的に利用されている方法は、短時間フーリエスペクトル分析がある。この方法は、音響データから連続する数 10ms 程度の時間長の信号区間を切り出し、切り出された信号が定常性（一定周期で繰り返す）と仮定して、スペクトル解析を行う。

スペクトル解析の流れは以下の通りである。

(1) フレーム化処理:与えられた信号 $s(n)$ に長さ N の分析窓を掛けることで以下のような信号系列 $s_\omega(m;l)$ を取り出す。

$$S_w(m;l) = \sum_{m=0}^{N-1} \omega(m)s(l+m) \qquad (l = 0, T, 2T, \cdots) \tag{2.1}$$

ここで、添え字 l は信号の切り出し位置に対応している。すなわち、l を一定間隔 T で増加させることで定常とみなされる長さ N の信号系列 $s_\omega(n) \quad (n = 0, 1, \cdots, N-1)$ が間隔 T で得られる。この処理をフレーム化処理と呼び、N をフレーム長、T をフレーム間隔と呼ぶ。

(2) 窓関数をかける:ある有限区間以外で 0 となる関数であり、フレーム化されたデータに対して重みをつける関数である。フレーム化処理を行う場合、離散的なデータの繋ぎ目においての信号の急激な変化の影響を和らげるため、原則として窓関数をかけなければならない。代表的なものとして音声信号だけに有効なハニング窓と、音声信号以外にも様々な信号にも有効なハミング窓がある。

$$\text{ハニング窓} : \omega(n) = 0.5 - 0.5\cos(\frac{2\pi n}{N-1}) \qquad (n = 0, 1, \cdots, N-1) \tag{2.2}$$

$$\text{ハミング窓} : \omega(n) = 0.54 - 0.54\cos(\frac{2\pi n}{N-1}) \qquad (n = 0, 1, \cdots, N-1) \tag{2.3}$$

(3) スペクトル分析（離散時間フーリエ変換、高速フーリエ変換):フレーム化処理によって得られた信号系列の短時間フーリエスペクトルは、離散時間フーリエ変換により以下の式で与えられる。

$$S(n) = \sum s_\omega(n)e^{-j2\pi\frac{nk}{N}} \qquad (k = 0, 1, \cdots, N-1) \tag{2.4}$$

離散フーリエ変換 (DFT) は、離散的なデータをフーリエ変換する際に、通常のフーリエ変換の無限区間積分を有限の和で書き換えたもので、時間領域、周波数領域ともに離散化されたフーリエ変換のことであり、時間領域の表現を周波数領域における表現に変換する。また、逆に周波数領域の表現を時間領域の表現に変換する、つまり元の音響データに戻す変換を離散フーリエ逆変換 (IDFT) と呼び以下の式で与えられる。

$$S(n) = \frac{1}{N}\sum S(k)e^{j2\pi\frac{nk}{N}} \qquad (k = 0, 1, \cdots, N-1) \tag{2.5}$$

実際の信号処理過程では、離散的フーリエ変換 (DFT) をその高速算法である高速フーリエ変換 (FFT) を用いて実行し、当該音声区間のスペクトル表現とすることが一般的である。高速フーリエ変換は式 (2.2)、(2.3) の N が 2^n 個であるとき、その処理を高速にできる性質がある。フーリエ変換の式には、

$$S\prime(n) = S(e^{j\frac{2\pi}{n}k}) = \sum s_\omega(n)e^{-j2\pi\frac{2\pi}{N}kn} \qquad (k = 0, 1, \cdots, N-1) \tag{2.6}$$

なる複素系列 $S\prime(k)$ が音声スペクトル表現として最も一般的に用いられる。

(4) パワースペクトルの算出:音響信号の特徴は主として、調音フィルタの振幅伝達特性に含まれる。したがって、音響信号の振幅スペクトル $|S'(k)|$、あるいはその 2 乗のパワースペクトル $|S'(k)|^2$、対数スペクトル $10\log|S'(k)|^2$ が注目すべきスペクトル表現になる。このスペクトルをグラフにすることで、音響データに含まれている周波数成分を解析することができる。音響データに含まれている周波数成分の上限はサンプリング定理により、サンプリング周波数の半分なので、高速フーリエ変換における周波数成分が意味のあるスペクトルとして扱われるのは 0 から π までのスペクトルである。また、音響信号の離散パワースペクトル系列は、離散スペクトル系列から式 (2.7) で表される。

$$|S'(k)|^2 = \frac{1}{N}[\mathrm{Re}\{S'(k)\}^2 + \mathrm{Im}\{S'(k)\}^2] \tag{2.7}$$

この 2 乗値のパワースペクトル $|S'(k)|^2$ を特徴量として扱っている。音響信号に高速フーリエ変換を施すと、時間表現 (縦軸:パワー、横軸:時間) から周波数表現 (縦軸:振幅、横軸:周波数) へと変換できる。しかし、実際には縦軸を周波数、横軸を時間としたグラフがよく使用されており、このようなグラフをスペクトルグラムという。スペクトルグラムは音声を視覚化したものであり、声紋とも呼ばれる。

2.1.2 音響特徴パラメータ

本研究で使用する 7 つの音響特徴パラメータについて述べる [4]

(1) スペクトルの変化

動的特徴量を連続するスペクトルのフレーム間の変化量として取り出す。音響信号のスペクトル分析した連続するフレームにおいて、あるフレームとその一定時間後のフレームとのパワースペクトルの差分によりスペクトルの変化量を得て、そのスペクトルの差分を一定時間足し合わせたものとしている。スペクトルの変化量によって比較する利点は、音声の識別に有利であり音声に比べて背景雑音のほうがスペクトルの変化量が大きく、無音のほうがスペクトルの変化量が小さいということである。

(2) スペクトルの傾き

あらかじめ人手により作成したラベルにより音響データの各区間を各種別（音声、音楽、背景雑音、無音）に振り分け、それぞれに対してスペクトル分析を行い、パワースペクトルを取り出し、各種別内において集められたパワースペクトルの分布を求めることで各種別において傾き値を得る。この傾き値を基に、与えられた音響ファイルから次々に得るパワースペクトルと各種別の学習データとの特徴パラメータの分布の類似度を比較する。この最小単純形は、パワースペクトルにおける一次回帰直線の傾きを比べることと同じである。傾きによって比較する利点は、有色系の音のほうが白色雑音よりも傾きが大きいので、音声と音楽と無音の識別に有利である。

(3) 白色雑音との近さ

パワースペクトルより一次回帰直線からスペクトル波形の切片を求めることで入力信号の白色性の度合を計測する。この白色雑音との近さによって比較する利点は、背景雑音のような定常的に混入した雑音は白色性が高いので、これらの識別ができるこということである。

(4) ピッチ

有声音源の繰り返し周期、いわゆるピッチ（基本周波数）の変化を調べることで、音源の変化を知ることができ、音源の特定のパラメータである。周波数分析によりピッチを求め、学習データと比べることで音源の特定に用いる。

(5) パワー

時間領域の分析だが、音響信号のような非定常的な信号に対して、変化していく信号の大きさにうまく追随するような比較的短い区間に音響データを区切り、その区間の信号 $xl(n)$ に対してエネルギー $E(l)$ を定義する [5]。

$$E(l) = \sum_{n=0}^{N-1} \{x_l(n)\}^2 \tag{2.8}$$

ここでは、整数 N は窓の中に含まれる音響信号の数である。

利点としては、測定が簡単であり、音声認識における有色系の音の区間の抽出にもよく用いられることから、有音と無音の区別に有利である。

(6) 中心周波数

抽出したパワースペクトルにおいて、無音の場合は右下がりに傾斜しているが、有音の場合は傾斜の途中で膨らみまたは突起が発生する。その突起がもっとも大きく発生している周波数帯の中心部分の周波数を中心周波数として定義している。これは有音と無音の識別に効果がある。

(7) 中心周波数のバンド幅

中心周波数を含む膨らみ、あるいは突起の始まりと終わりによる周波数帯の長さをバンド幅として定義する。音声は一定の周波数を含むことが多いためそのバンド幅はある程度の大きさになることが考えられるが、雑音はあまり多くの周波数を含まないものから白色性が高く幅広い周波数を含むものまで様々であり、その違いから音声と雑音の特定に有効である。

2.2 i-vector の概要

近年の話者認識システムの多くは i-vector [3][4] に基づいて構成されおり、この領域における最高水準の技術となっている。i-vector とは、ある発話から得られた音響特徴量を因子分析を用いて、話者固有の特徴を抽出したものである。i-vector の抽出においては、因子分析の入力として、発話毎に GMM(Gaussian Mixture Model) の平均ベクトルを結合した GMM スーパーベクトルを用いる。発話 u から作成された GMM スーパーベクトル $M_u \in R^{CD_F}$ は以下で定義される。

$$M_u = Tw_u + m \tag{2.9}$$

ここで M_u は大量の不特定話者の発話データから作成される UBM（Universal Background Model）を事前情報として事後確率最大化（MAP）法により推定された GMM を用いる。また m は UBM から得られる話者及びチャネル非依存の GMM スーパーベクトルである。C は GMM (UBM) の混合数，D_F は音響パラメータの次元数、$T \in R^{CD_F * D_r}$ は低ランクの矩形行列 $D_r \ll CD_F$ で、全変動空間を張る基底ベクトルで構成される固有声行列である。$W_u \in R^{D_r}$ は発話ごとに与えられる潜在変数であり、平均ベクトルが $0 \in R^{D_T}$ で共分散行列行列が単位行列 $I \in R^{D_T * D_T}$ のガウス分布 $N(w; 0, I)$ に従う。この w は total factor(全因子) と呼ばれ、各発話に対する i-vector である。つまり、i-vector は GMM スーパーベクトル空間における平均的な話者 (UBM の平均) から「差 (を次元圧縮したもの)」として各話者を表現したものと言える。

2.2.1 UBM に対する Baum-Welch 統計量

準備として、UBM に対する Baum-Welch 統計量を計算することから始める。$O_u = o_1, o_2, o_3, \cdots, o_L$、$o_t \in R^{D_F}$、を発話 u から得られる L フレームの音響パラメータ系列 $c = 1, 2, 3 \cdots, C$、を UBM (GMM) の混合要素を表す添え字、$\Omega = \{\pi_c, m_c, \sum_c\}_{c=1}^{C}$ を UBM のパラメータ (混合重み、平均ベクトル、対角共分散行列) とする。このとき、発話 u に対する 0 次、1 次、2 次の Baum-Welch 統計量は、

$$N_{u,c} = \sum_{t=1}^{L} \gamma_t(t) \tag{2.10}$$

$$F_{u,c} = \sum_{t=1}^{L} \gamma_t(c)(o_t - m_c) \tag{2.11}$$

$$S_{u,c} = diag\left[\sum_{t=1}^{L} \gamma_c(c)(o_c - m_c)(o_t - m_c)'\right] \tag{2.12}$$

と書ける。ここで、$y_c(c)$ は、o_t が UBM の c 番目の要素分布から生成される事後確率

$$\gamma_c(c) = p(c \mid o_t, \Omega) = \frac{\pi_c p(o_t \mid m_c, \sum c)}{\sum_{k=1}^{C} \pi_k p(o_t \mid m_k, \sum k)} \tag{2.13}$$

である。更にこれらを用いて、

$$N_u = \begin{pmatrix} N_{u,1}, I_{D_F} & 0 & \cdots & 0 \\ 0 & 0 & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & N_{u,C}, I_{D_F} \end{pmatrix} \in R^{CD_F \times CD_F} \tag{2.14}$$

$$N_u = \begin{pmatrix} F_{u,1} \\ F_{u,2} \\ \vdots \\ F_{u,C} \end{pmatrix} \in R^{CD_F} \tag{2.15}$$

$$N_u = \begin{pmatrix} S_{u,1},0 & 0 & \dots & 0 \\ 0 & S_{u,2} & \dots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & S_{u,C} \end{pmatrix} \in R^{CD_F \times CD_F} \tag{2.16}$$

ここで、$I_{D_F} \in R^{D_F \times D_F}$ である。

2.2.2 全因子 w の確率分布と i-vector の抽出

本節では、w に関する種々の確率分布を導出する。このとき、w の事後分布の導出過程において i-vector の具体的な計算方法を示す。

- 事前分布

  w の事前分布は $p(w)$ 平均 0、共分散行列を持つガウス分布であり、以下のように書ける。

$$p(w) \propto exp(-\frac{1}{2}w'w) \tag{2.17}$$

- 条件付き分布

  $M_{u,c}$ を混合要素に対する M_u の部分ベクトルとする。直感的には、$M_{u,c}$ は発話 O_u で学習した GMM の混合要素 c に割り当てられた O_c の各フレームは、平均 $M_{u,c}$、共分散行列 $\sum_c$(UBM のまま) に従うと仮定する。すなわち、w の値で条件付けられた観測データ O の条件付き分布は

$$P(O_u|w_u) = etp\left(\sum_{t=1}^{L}\sum_{c=1}^{C}\gamma_t(c)log(2\pi)^{-\frac{D_F}{2}}|\Sigma_c|^{-\frac{1}{2}} - \frac{1}{2}\sum_{t=1}^{L}\sum_{c=1}^{C}\gamma_t(c)D(o_t;\theta_c)\right) \tag{2.18}$$

  のように書ける。ここで、

$$D(o_t;\theta_t) = (o_t - M_{u,c})'\Sigma_c^{-1}(o_t - M_{u,c}) \tag{2.19}$$

$$M_{u,c} = m_c + T_c w_u \tag{2.20}$$

  である。$T_c \in R^{D_F \times D_T}$ は、混合要素 c に対する T の部分行列である。式 (2.18) の exp の内部を Baum-Welch 統計量を用いて整理すると、

$$\sum_{t=1}^{L}\sum_{c=1}^{C}\gamma_t(c)log(2\pi)^{-\frac{D_F}{2}}|\Sigma_c|^{-\frac{1}{2}} - \frac{1}{2}\sum_{t=1}^{L}\sum_{c=1}^{C}\gamma_t(c)D(o_t;\theta_c) = G_u^{\Sigma} + H_u^{\Sigma T} + Const \tag{2.21}$$

ここで、G_u^{Σ} 及び $H_u^{\Sigma T}$ は、

$$G_u^{\Sigma} = \sum_{c=1}^{c} \left[\frac{1}{2} N_{u,c} log \left| \Sigma_c^{-1} \right| - \frac{1}{2} tr \left(\Sigma_c^{-1} S_{u,c} \right) \right] \tag{2.22}$$

$$H_u^{\Sigma T} = w_u' T' \Sigma^{-1} F_u - \frac{1}{2} w_u' T' N_u \Sigma^{-1} T w_u \tag{2.23}$$

- 事後分布

  w の事後分布は (2.18)〜(2.23) を用いると、

$$\begin{aligned} p(w_u|O_u) &\propto p(O_u|w_u)p(w_u) \propto exp(w_u' T' \Sigma' T w_u - \frac{1}{2} w_u' w_u) \\ &\propto exp(w_u' T' \Sigma' F_u - \frac{1}{2} w_u' N_u \Sigma^{-1} T w_u - \frac{1}{2} w_u' w_u) \\ &\propto exp(-\frac{1}{2}(w_u - G_u T' \Sigma^{-1} F_u)' G_u^{-1} (w_u - G_u T' \Sigma' F_u)) \end{aligned} \tag{2.24}$$

  と書ける。ここで、

$$G_u = (I + T' \Sigma^{-1} N_u T)^{-1} \tag{2.25}$$

  である。w の事後分布もガウス分布であることに注意すると、平均及び分散は、

$$E[w_u] = G_u T' \Sigma^{-1} F_u \tag{2.26}$$

$$cov[w_u] = G_u \tag{2.27}$$

  となる。前述のとおり、確率的潜在変数モデルのもと、i-vector は w の事後分布の平均として得られる。つまり、発話 u の i-vector は、Baum-Welch 統計量 N_u、F_u 及び推定済みのパラメータ T,Σ を用いて、式 (2.26) により計算することができる。

2.2.3 因子分析モデルパラメータの推定

因子分析モデルのパラメータ T 及び Σ は、EM アルゴリズムにより求められる。すなわち、完全データ $(O_u, w_u)_{u=1}^{U}$ に対する対数尤度の期待値

$$Q = \sum_{u=1}^{U} E[logp(O_w w_u|\theta)] \tag{2.28}$$

の最大化問題を解くことで求める。ここで、θ はパラメータ T、Σ を表す。完全データの対数尤度は、

$$logp(O_w w_u) = logp(O_u|w_u, \theta) + logp(w_u) \tag{2.29}$$

と書けるので、式 (2.18)〜(2.23) を用いると、式 (2.28) は以下のように整理できる。

$$
\begin{aligned}
Q = \frac{1}{2}\sum_{u=1}^{U}\sum_{c=1}^{C}(N_{u,c}log\left|\Sigma_c^{-1}\right| - tr(\Sigma_c^{-1}S_{u,c})) \\
+\sum_{u=1}^{U} tr\left(\Sigma^{-1}\left(F_u E[w'_u]T' - \frac{1}{2}N_u T E[w_u w'_u]T'\right)\right) \\
-\sum_{u=1}^{U}\frac{1}{2}tr(E[w_u W'_u])
\end{aligned}
\tag{2.30}
$$

以上より、E ステップにおいては古いパラメータを使って、w 空間の事後分布の統計量を以下のように計算する。

$$E[w_u] = G_u T'\Sigma^{-1}F_u \tag{2.31}$$

$$E[w_u w'_u] = G_u + E[w_u]E[w'_u] \tag{2.32}$$

M ステップでは、式 (2.30) をパラメータに関して最大化する。まず、(2.30) を T に関して微分して 0 と置くことで、以下の関係式を得る。

$$\sum_{u=1}^{U}\Sigma^{-1}F_u E[w'_u] = \sum_{u=1}^{U}\Sigma^{-1}N_u T E[w_u w'_u] \tag{2.33}$$

これより、T の推定式が、

$$T^i = \left(\sum_{u=1}^{U}F_u^i E[w'_u]\right)\left(\sum_{u=1}^{U}N_{u,c}E[w_u w'_u]\right)^{-1} \tag{2.34}$$

のように得られる。ここで、T^i、F_u^i は、おのおの T、F_u の i 行目を表し、$i = (c-1)\times D_F + f$、$1 \leq f \leq D_F$ である。また、Σ の推定式は、

$$\Sigma = N^{-1}\left(\sum_{u=1}^{U}S_u - diag\left[\sum_{u=1}^{U}F_u E[w'_u]T'\right]\right) \tag{2.35}$$

となる。ここで、$N = \Sigma_{u=1}^{U}N_u$ である。

2.2.4 コサイン類似度

発話 x から抽出した i-vectorw_x と発話 y から抽出した i-vectorw_y の比較を行うための方法としてコサイン類似度を用いる。

$$\cos(w_x, w_y) = \frac{w_x \cdot w_y}{\parallel w_x \parallel \parallel w_y \parallel} \tag{2.36}$$

類似度の値の範囲は、$-1 \leq \cos(w_x, w_y) \leq 1$ であり、類似度が最も高い値は 1 である。

2.3　音声認識

2.3.1　音声認識システムの流れ

音声認識の流れを図 2.1 に示す。まず、入力された音声データから前処理として雑音区間と無音区間を除去し、発話区間を検出する。次に検出した発話区間の音響的特徴量を抽出し、デコーダへと渡す。デコーダではこの音響的特徴量をもとに、音響モデルと言語モデル、単語辞書を参照しながら単語列の尤度を算出し、最も尤度の高いものを認識結果として出力する。言語モデルと単語辞書については 2.3.2 節、音響モデルについては 2.3.3 節で説明する。

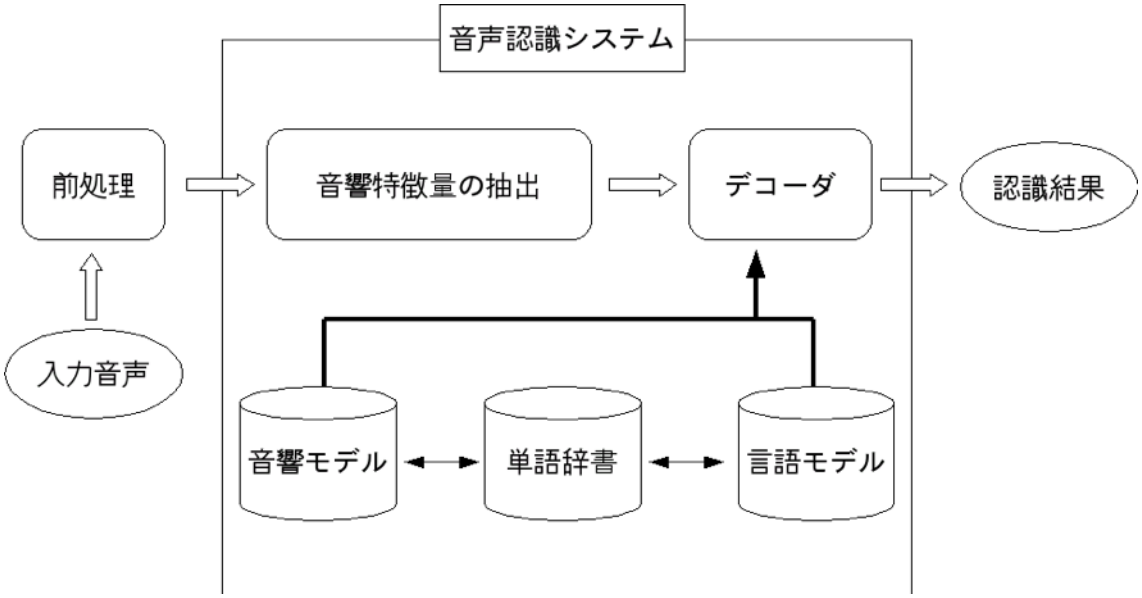

図 2.1: 音声認識の流れ

2.3.2　単語辞書と言語モデル

単語辞書

単語辞書には、一般的に学習データに出現する単語のなかで出現頻度の高い単語を登録する [7]。言語モデルもその単語辞書に登録された単語を用いて構築する。単語辞書の例を表 2.3 に示す。単語辞書には表記、発音形、原型、品詞番号、出現表記、音素表記などを登録する。

表 2.3: 単語辞書の例

表記+発音形+原形+品詞番号	出力表記	音素表記
あか+アカ＋アカ＋ 14	あか	a k a
技術＋ギジュツ＋ギジュツ＋ 1	技術	g i zh j u ts u
新聞＋シンブン＋シンブン＋ 1	新聞	sh i ng b u ng

音声認識では、言語モデルは、「表記+発音形+原型+品詞番号」を、音響モデルは「音素表記」の部分を用いて最尤の単語を算出する。辞書に登録している単語が少ない場合、入力された単語が辞書に登録されていないことが多くなり、他の誤った単語を出力し認識率が低下してしまう。一方、辞書に登録している単語が多すぎる場合、認識処理に時間がかかるだけでなく、認識候補が増えるため認識率が低下してしまう。よって適切な単語の登録数を検討する必要がある。

言語モデル

音声認識における言語モデルとは、文の品詞や単語と単語の関係性、音素の並びの制約などを定式化したもののことである。言語モデルの主流はサンプルデータから統計的な手法によって確率推定を行なう統計的言語モデルである。その中でも最も広く使われているのが N グラムモデルである。

N グラムモデル

N グラムモデルとは、与えられた単語列 $\omega_1, \omega_2, \cdots, \omega_n$ に対して、その出現確率 $p(\omega_1, \omega_2, \cdots, \omega_n)$ を推定する場合に、

$$P(\omega_1, \omega_2, \cdots, \omega_n) = \prod_{i=1}^{n} p(\omega_i \mid \omega_{i-N+1} \cdots \omega_{i-1}) \tag{2.37}$$

のような近似を行なうモデルである。N グラムモデルでは、i 番目の単語 ω_i の生成確率が、直前の $N-1$ 単語 $\omega_{i-N+1} \cdots \omega_{i-2}\, \omega_{i-1}$ だけに依存すると考える。特に $N=1$ のときユニグラム（unigram）、$N=2$ のときバイグラム（bigram）、$N=3$ のときトライグラム（trigram）という。

文や発話中の単語の生成確率は文脈に依存することから、N グラムモデルの推定確率は、N が大きいほど高くなる。しかし、N グラムモデルは語彙の N 乗のコストがかかることから、N を大きくするためには、膨大な量のテキストを用意しなければならない。しかし、自由発話を記述したテキストは極めて少ない。本研究では、$N=3$ の trigram を用いる。

2.3.3 音響モデル

音響モデルとは、音声の最小単位である音素または、単語や音節の音響特徴パラメータの時系列をモデル化したものである。この音素の特徴は、発話者や発話内容などによって変化するが、発話者ごと、または発話タスクごとにモデル化することは、膨大なコストがかかり汎用性がないため好ましくない。そのため、音響モデルの構築方法としては、音素ごとに様々な学習音声で学習を繰り返し、最尤の音響モデルを作ることが一般的である。本研究では、隠れマルコフモデル（Hidden Markov Model）を用いて最尤の音響モデルを構築する。

以下に音響モデルを構築する際に、必要となる知識について述べる。

MFCC

メル周波数ケプストラム係数（Mel - Frequency Cepstrum Coefficient : MFCC）とは、メル周波数という人間の音の高低に対する感覚尺度で音声スペクトルから係数スペクトルを抽出したものである [7]。これは一般的に、音声の特徴を抽出するパラメータとして用いられる。 MFCC の計算では、スペクトル分析は周波数軸上に L 個の三角窓を配置し、フィルタバンク分析により行なう。すなわち、窓の幅に対応する周波数帯域の信号のパワーを、単一スペクトルチャネルの振幅スペクトル $|S'(k)|$ の重み付けの和 $m(l)$ で求める。

$$m(l) = \sum_{k=lo}^{hi} W(k;l) \mid S'(k) \mid \quad (l = 1, \cdots, L) \tag{2.38}$$

$$W(k;l) = \begin{cases} \frac{k - k_{lo}(l)}{k_c(l) - k_{lo}(l)} & \{k_{lo} \leq k \leq k_c(l)\} \\ \frac{k_h i(l) - k}{k_{hi}(l) - k_l(l)} & \{k_c \leq k \leq k_{hi}(l)\} \end{cases} \tag{2.39}$$

ただし、$W(k;l)$ は重み、$k_{lo}(l)$、$k_c(l)$、$k_{hi}(l)$ はそれぞれ l 番目のフィルタの下限、中心、上限のスペクトルチャネル番号であり、隣り合うフィルタ間で

$$k_c = k_{hi}(l-1) = k_{lo}(l+1) \tag{2.40}$$

なる関係がある。さらに、$k_c(l)$ はメル周波数軸上で等間隔に配置される。メル周波数は

$$Mel(f) = 2595 \log_{10}(1 + \frac{f}{700}) \tag{2.41}$$

により計算される。ただし、f の単位は [Hz] にとる。

最終的にフィルタバンク分析により得られた L 個の帯域におけるパワースペクトルを離散コサイン変換することで、式 (2.42) のように MFCC が得られる。

$$c_{mfcc}(i) = \sqrt{\frac{2}{N}} \sum_{i=1}^{L} \log ml \cos\left\{\left(l - \frac{1}{2}\frac{i\pi}{L}\right)\right\} \tag{2.42}$$

隠れマルコフモデル (HMM)

HMM は時系列信号の確率モデルであり、複数の定常信号源の間を遷移することで、時系列に適応させ、音響モデルを構築する [7]。図に HMM の例を示す。a、b、c は状態、矢印は状態遷移を示す。HMM は次の状態への遷移と現在の状態の遷移を行なう。このように現在の状態への遷移があるため、様々な長さの時系列信号に対応できる。

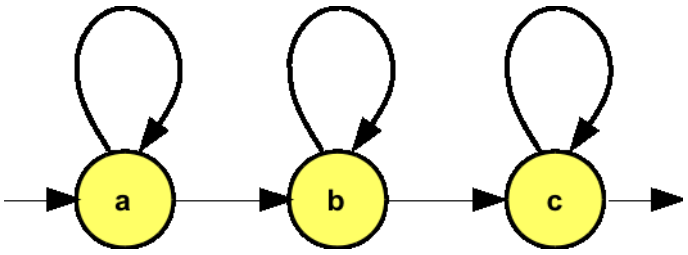

図 2.2: HMM の例

調音結合

音素の音響的な特徴は、周辺前後の音素の影響を受けて同じ音素でも様々に変化することが知られている [7]。この現象を調音結合という。特に、音素から音素への渡りの部分ではスペクトル特徴が時間とともに連続して大きく変化するため、音声を扱う分野ではこの調音結合への対応が重要である。この調音結合に対する最も直接的な対応策として、前後の音素を考慮した 3 つ組音素（トライフォン）を認識の処理単位として用いるものがある。

MFCC は、フレームと呼ばれる数十 ms 程度の音声区間を定常とみなした上で得られる静的な特徴量である。しかし調音結合があるため、フレーム分析により得られた静的な特徴に加え、時間とともに変化する動的な特徴を特徴量に加えて音声認識を行なうことで、認識の精度が大きく向上することが知られている。動的な特徴には式 2.43 や式 2.44 で示される一次差分か二次差分を利用することが多い。ここで、K は回帰係数を計算する範囲であり、一般的に 20~40ms である。

$$\Delta c(n;l) = \frac{\Sigma_{K=-K}^{K} k_c(n;l+k)}{\Sigma_{K}^{K=-K} K^2} \tag{2.43}$$

$$\Delta\Delta c(n;l) = \frac{\Sigma_{K=-K}^{K} k_c(n;l+k)}{\Sigma_{K}^{K=-K} K^2} \tag{2.44}$$

2.3.4　DNN の概要

本研究では DNN をベースとした会議音声認識を行なう。DNN とは、多層ニューラルネットワークを使った機械学習のことである。DNN は図 2.3 のように、auto-encoder または Restricted Boltzmann Machines (RBM) などを積み重ねた深い構造をもつ。入力に近い層では、単純に特徴抽出しかできないが、それらの重み付け和をとると表現能力が上がり、それをさらに上位の層の入力にしていくことで、モデルの表現力がさらに上がるとされている。

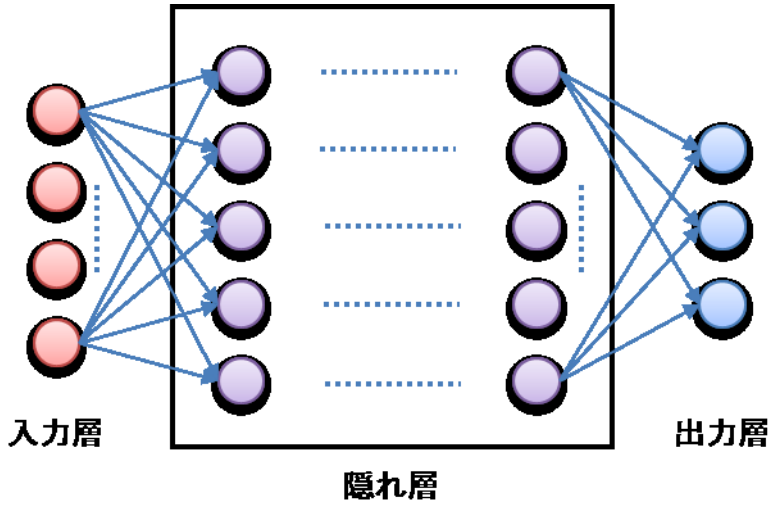

図 2.3: DNN の構造図

音声認識においては、入力層の入力は MFCC などの音響特徴量となり、出力層は HMM の各状態となる。

第3章

予備調査

3.1　ニュース番組音声における非発話区間の調査

本章では、ニュース番組音声の発話間隔の調査を行う。

3.1.1　使用する音声データ

パラメータの学習用にニュース番組の音声データ 13 個を用いる。各音声データには、事前に人手で 4 種類（音楽、音声、雑音、無音) の音源ラベルが付与されている。「音声」の音源ラベルが付与された区間においては、更に発話者の情報が付与されている。図 3.1 は音声の音源ラベルの一例である。また「音声」の音源ラベルをもとに対象の音声データから発話区間を抽出し、それを一発話とした。

表 3.1.1 に調査に用いるデータの詳細を示す。

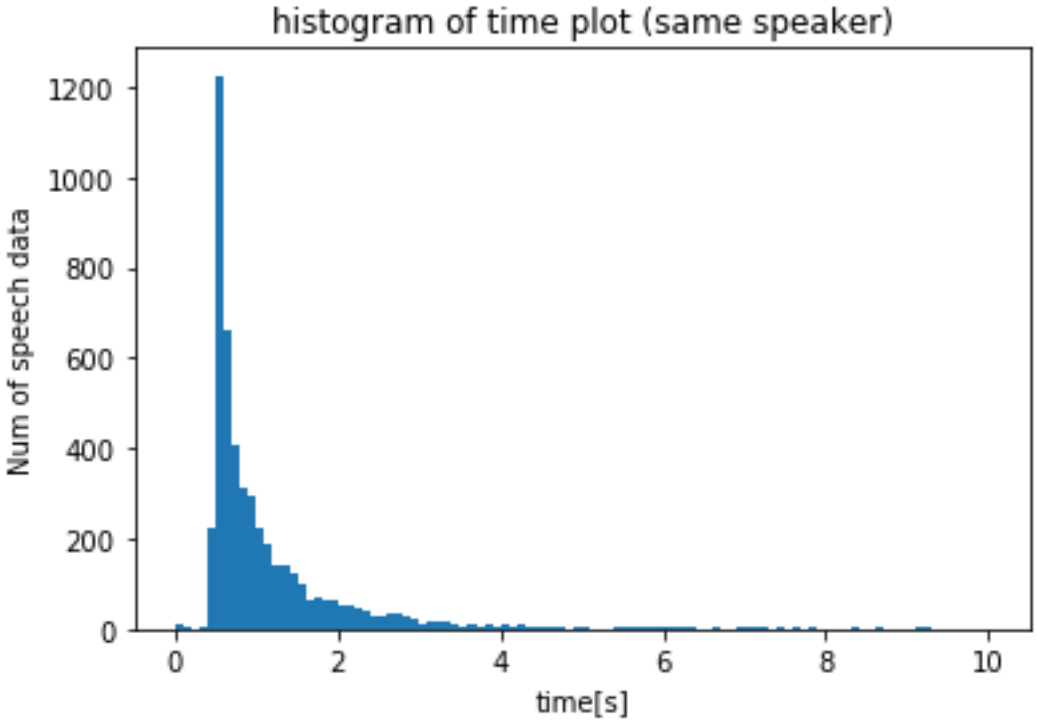

図 3.1: 「音声」の音源ラベルの例

表 3.1: パラメータの学習用音声データの詳細

データ ID	収録時間	話者数	全発話数
ニュース A	30 分 3 秒	20	337
ニュース B	30 分 3 秒	31	312
ニュース C	30 分 3 秒	21	324
ニュース D	30 分 4 秒	20	324
ニュース E	20 分 3 秒	13	159
ニュース F	30 分 3 秒	22	343
ニュース G	30 分 4 秒	22	313
ニュース H	30 分 4 秒	20	315
ニュース I	30 分 4 秒	17	321
ニュース J	30 分 4 秒	16	337
ニュース K	30 分 4 秒	20	363
ニュース L	30 分 4 秒	26	345
ニュース M	30 分 4 秒	26	314

3.1.2　調査方法

本調査は、図 3.1 のような「音声」の音源ラベルを用いて行う。ラベル付けがされていない区間を非発話区間として、発話が終了後、次の発話までの時間を計測する。また、発話が重なっている場合は発話が終了していなくても次の発話が開始されるため、非発話区間を 0 秒とした。

3.1.3　調査結果

調査結果を図 3.2、図 3.3 に示す。

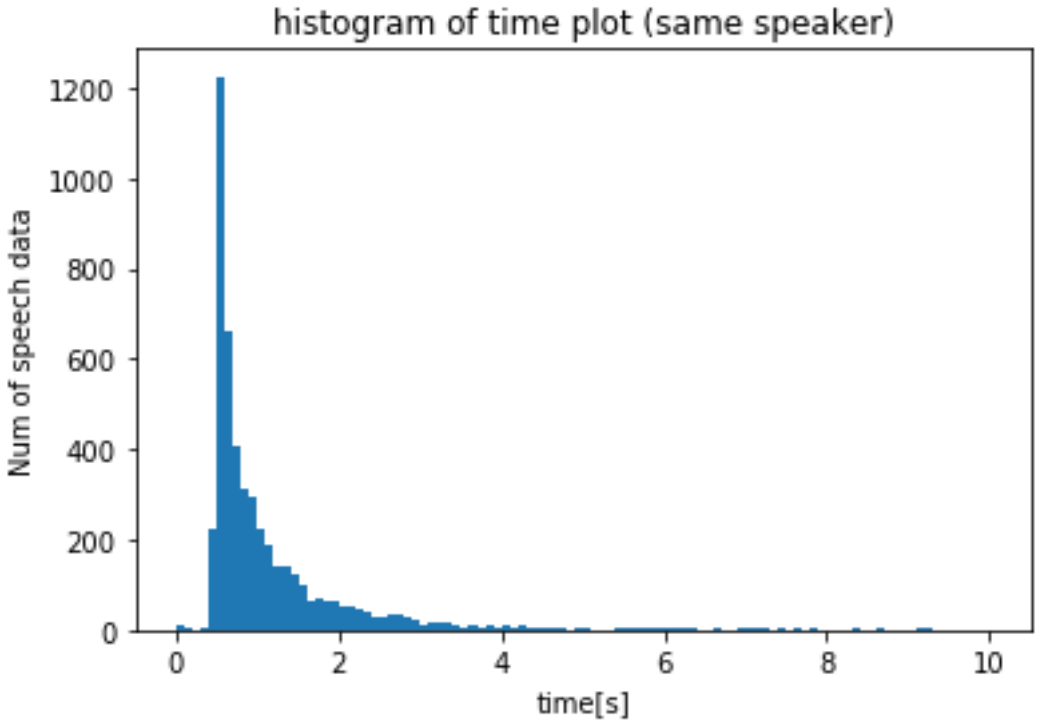

図 3.2: 同一話者間の非発話区間の時間情報

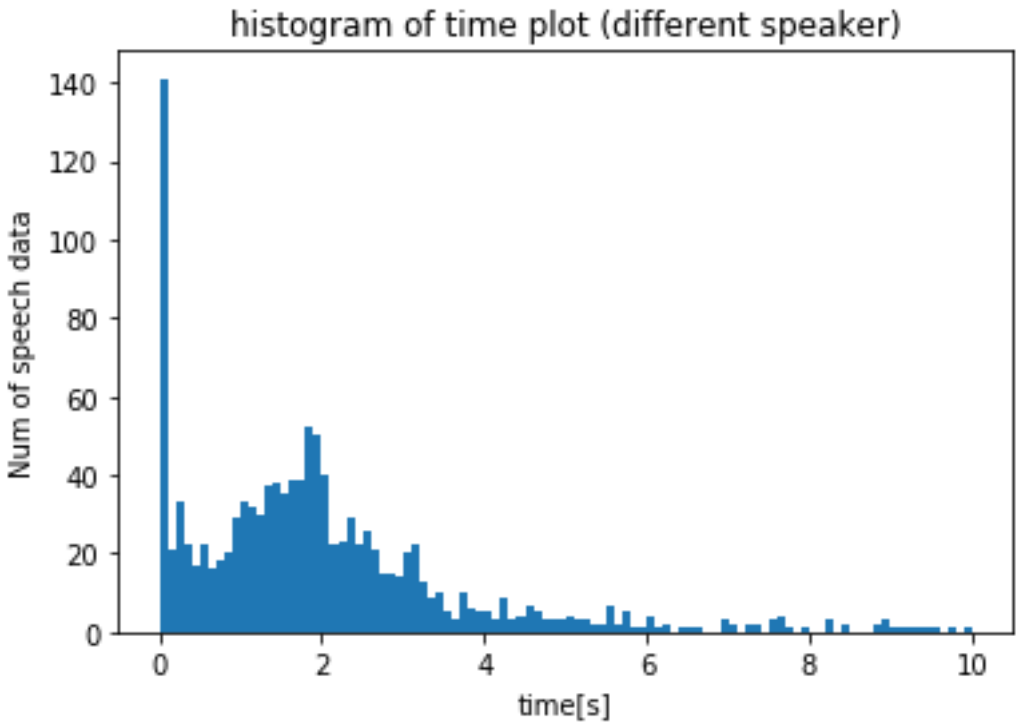

図 3.3: 異なる話者間の非発話区間の時間情報

以上の結果より、同一話者の発話は連続して行われるため、非発話区間は非常に短く、話者が切り替わる場合は非発話区間が比較的長くなることがわかる。しかし、話者が切り替わる場合でも非常に非発話区間が短くなる場合がある。これは、

- 対話者による発話中の相槌
- 対話中の素早い応答
- インタビューイの切り替わり

があるためである。

3.2　コサイン類似度を用いた i-vector の性質の調査

発話データの長さを変更して i-vector を抽出、コサイン類似度を算出してヒストグラムにすることで、発話の長さによる i-vector の性質を調査する。

3.2.1　使用する音声データ

本研究では、UBM モデルの学習データおよびコサイン類似度を用いた i-vector の性質の調査に読み上げ音声 [6] を使用した。

3.2.2　コサイン類似度の算出条件

対象の音声データからある発話を取り出し、それ以外の発話との i-vector のコサイン類似度を算出する。それを同一話者の発話間の場合と異なる話者の発話間の 2 つの場合に分けてヒストグラムに表し、コサイン類似度の性質の調査を行った。

i-vector の抽出に使用する UBM モデルの学習には読み上げ音声 [6] を使用し、発話から抽出する音響特徴パラメータを表 3.2 に示す。また混合数は 32 とした。

表 3.2: 使用する音響特徴パラメータ

特徴量	次元数
MFCC	19
POW	1
ΔMFCC	19
ΔPOW	1
ΔΔMFCC	19
ΔΔPOW	1
計	60

本研究では、音響特徴量のひとつとしてメル周波数ケプストラム係数 (MFCC) を用いる。メル周波数ケプストラム係数（Mel - Frequency Cepstrum Coefficient : MFCC）とは、メル周波数という人間の音の高低に対する感覚尺度を考慮した特徴量であり、音声スペクトルから係数スペクトルを抽出したものである。これは一般的に、音声の特徴を抽出するパラメータとして用いられる。[5]

3.2.3　長い音声データから抽出した i-vector の性質

本節では、比較的長い音声データとして、20 秒間の音声データから i-vector を抽出した。その結果、図 3.4 より、同一話者の発話間の場合はコサイン類似度が高い値に多く分布し、図 3.5 より、異なる話者の発話間の場合はコサイン類似度が全体的に分布していることが分かる。

これより、比較的長い音声データでは、コサイン類似度の値が高いほど、i-vector を照合した発話の話者が同一の話者である確率が高いということが分かる。

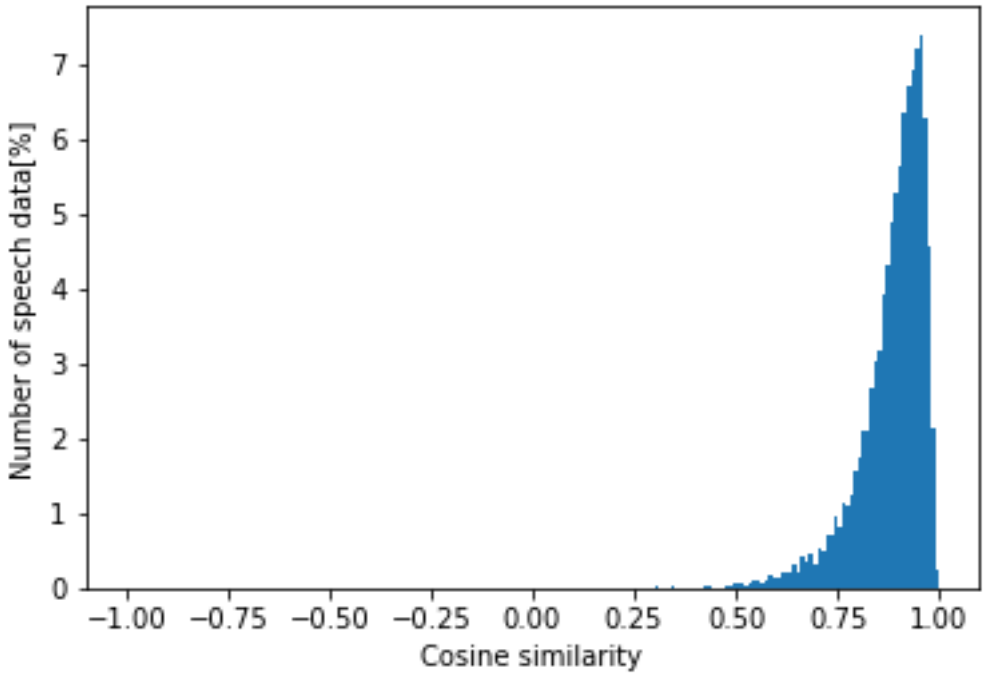

図 3.4: 長い音声データから抽出した同一話者間の i-vector のコサイン類似度

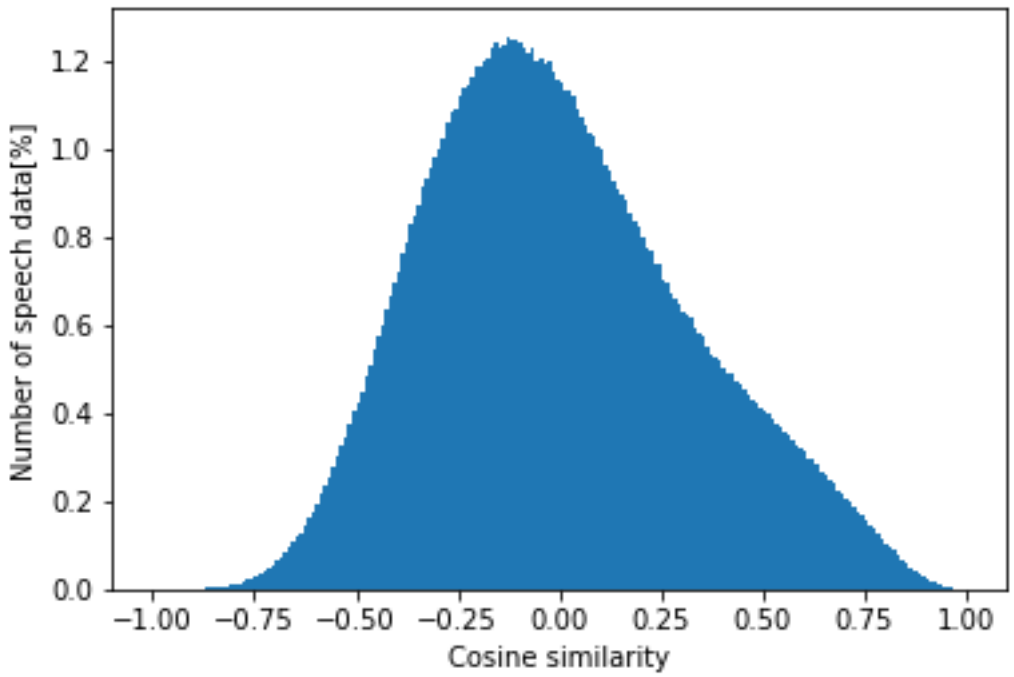

図 3.5: 長い音声データから抽出した異なる話者間の i-vector のコサイン類似度

3.2.4　非常に短い音声データから抽出した i-vector の性質

本節では、非常に音声データとして、0.3 秒間の音声データから i-vector を抽出した。その結果を図 3.6 と図 3.7 に示す。

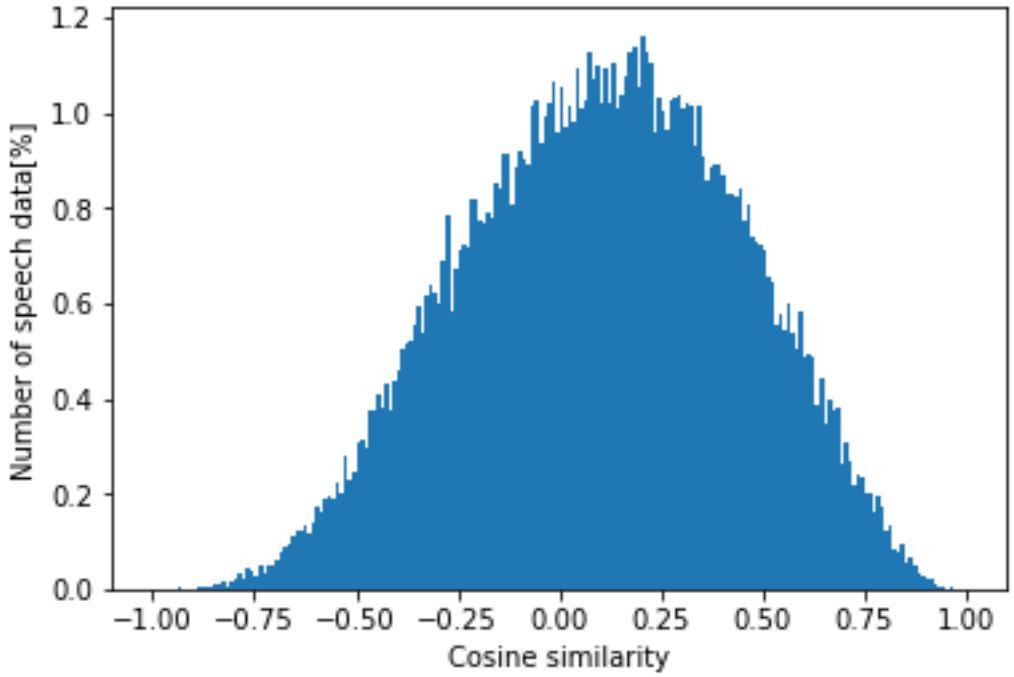

図 3.6: 非常に短い音声データから抽出した同一話者間の i-vector のコサイン類似度

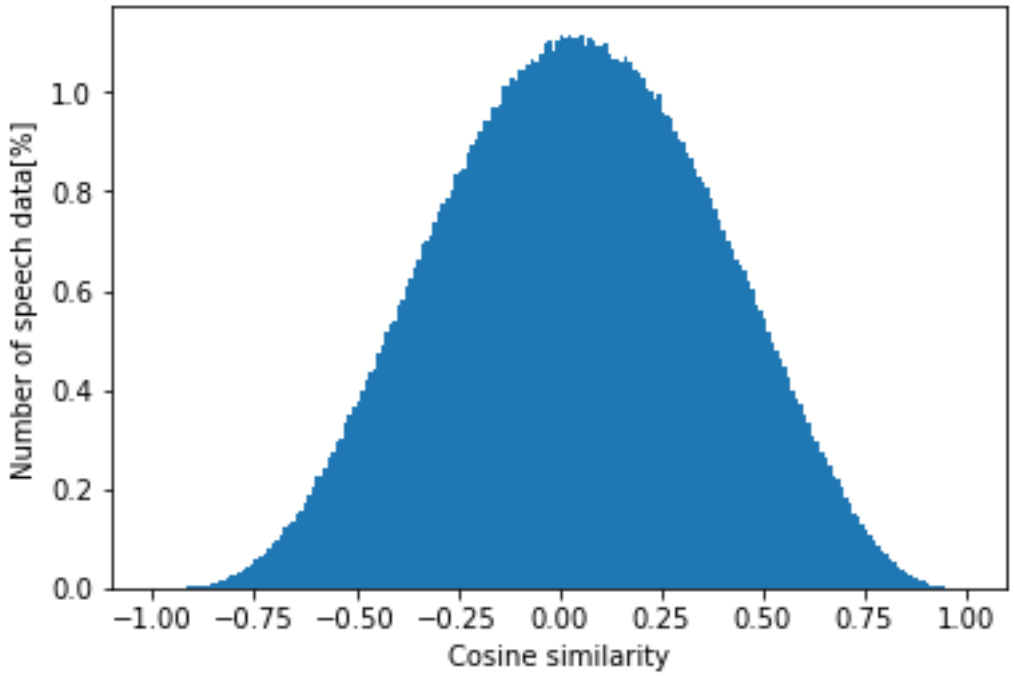

図 3.7: 非常に短い音声データから抽出した異なる話者間の i-vector のコサイン類似度

非常に短い音声データから抽出した i-vector は同一話者間ではコサイン類似度が 0.2 付近、異なる話者間では 0 付近に多くのデータが集まっている。しかし、同一話者間のコサイン類似度と異なる話者間のコサイン類似度のヒストグラムは非常に似た形をしており、本検証で抽出された i-vector では話者の識別をすることは非常に難しい。これより、i-vector を用いた話者照合、話者識別を行う場合はできるだけ長い発話データを用意することが必要である。

第4章

提案手法

本研究では 2 通りの方法で前後の同一話者と考えられる発話区間を結合し、擬似的に長い発話データを作成する。1 つ目は発話間の時間情報を考慮した発話区間の結合手法である。2 つめは、話者の発話環境を考慮した発話区間の結合手法である。

4.1 発話間の時間情報を考慮した発話区間の結合手法

本手法では、発話区間と発話区間の間 (非発話区間) が短い場合、発話区間を結合する。これは図 3.2 で示されるように、同一話者が連続で発話する場合は間をおかずに次の発話を行うことが非常に多く、非発話区間が非常に短いためである。つまり、非発話区間が非常に短いとき、高い確率で同一話者の発話が行われると考えられる。しかし、話者が切り替わった場合でも対話中やインタビューイの存在により非発話区間が非常に短い場合がある。この場合、同一話者が連続で発話しているか話者が切り替わったかの判別は非発話区間の時間情報のみでは不可能である。そこで、図 3.6 と図 3.7 より、発話が非常に短い場合でも若干のコサイン類似度の差があることに着目した。これらより、非発話区間が非常に短く、発話から抽出された i-vector のコサイン類似度が一定値以上のとき、同一話者の発話と判別し発話区間を結合する。

4.2 発話環境を考慮した発話区間の結合手法

本手法では、発話環境の変化を音源識別によって検出し、同一話者の可能性が高い前後の発話区間を結合する。ニュース番組にはスタジオにいるアンカーや天気アナウンサーのほか、台風の状況を中継する中継アナウンサー、騒音の中でインタビューを受けるインタビューイなどが存在する。そこで、アンカーから中継アナウンサー、インタビューイからアンカーなど話者が切り替わった場合発話環境が変化することに着目した。しかし、音源識別では、発話の背景雑音がどの種類であるかを特定することが非常に難しい。そこで、非発話区間の音源識別結果を用いることで、発話環境の変化を検出する。発話環境の変化の検出方法を以下の図に示す。

例の図 4.2 では、最初の話者の非発話区間が「無音区間」であることから、雑音が少ない環境で発話していることがわかる。ここで、次の話者が発話を始めたとき、発話環境が「雑音区間」に変化している。つまり、雑音の多い屋外などで発話をしているインタビューイや中継アナウンサーに話者が切り替わったと考えられる。ここで、前の非発話区間と現在の非発話区間の音源識別結果を参照し、同一の結果であった場合発話区間を結合、異なった場合、話者の切り替わりと識別する。

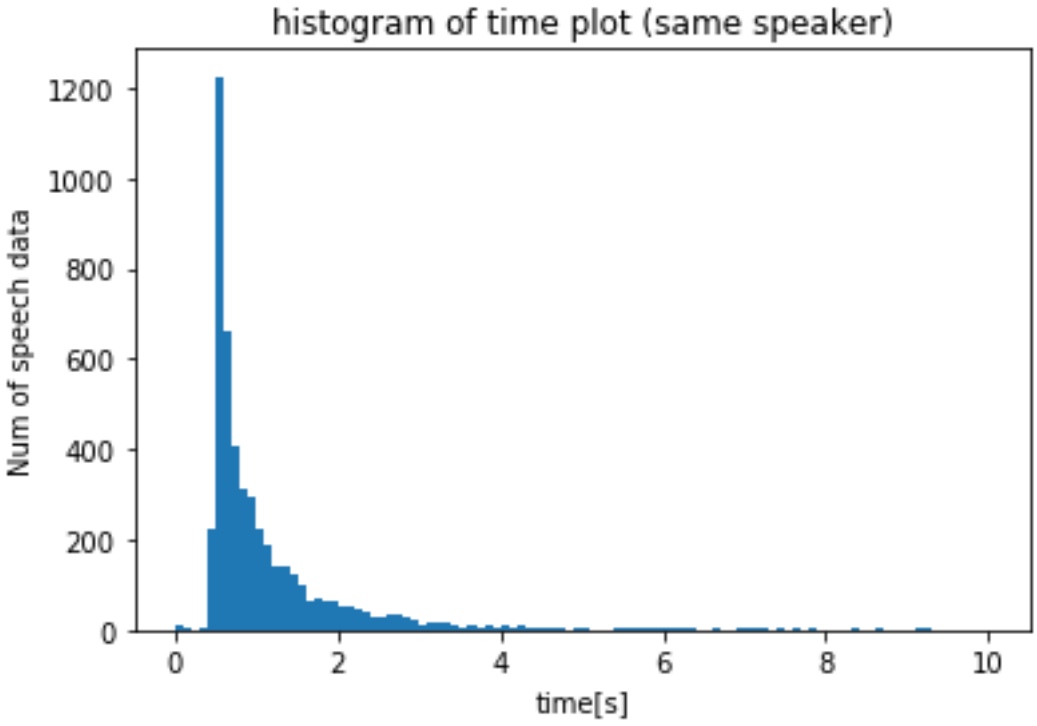

図 4.1: 同一話者間の非発話区間の時間情報

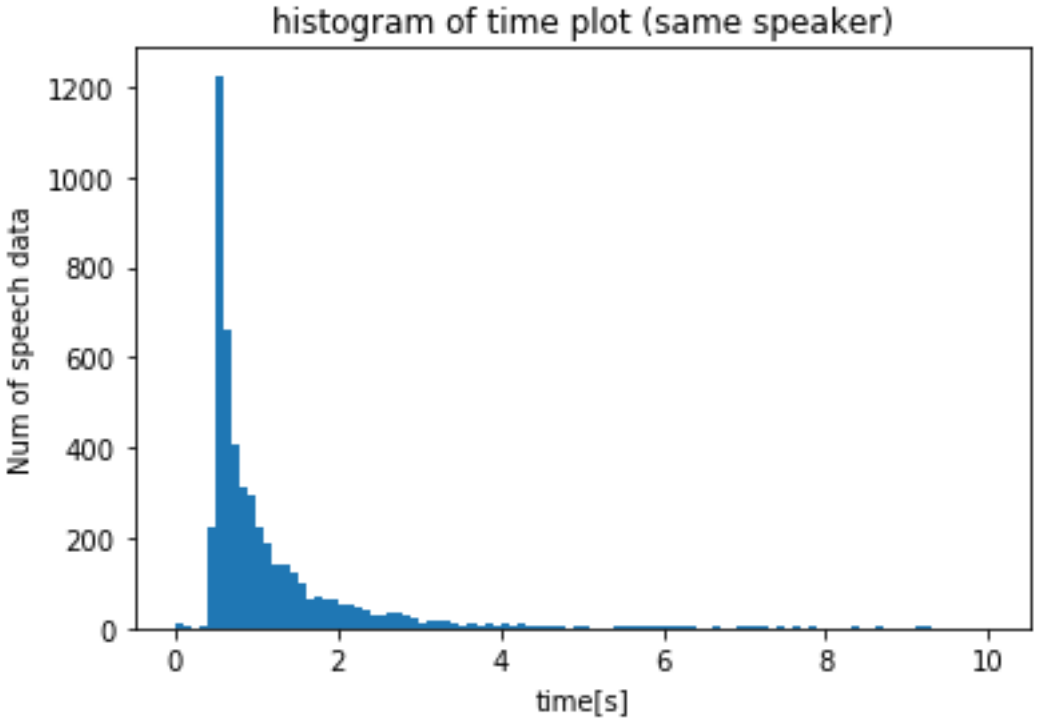

図 4.2: 同一話者間の非発話区間の時間情報

第5章

実験

5.1 使用する音声データ

評価用にニュース番組の音声データ 5 個を用いる。本来のニュース番組の音声には「音声」の音源ラベルが付与されていない。そこで、音源識別を用いて発話区間を検出、切り出しを行なった。また、切り出された発話区間それぞれを一発話し、各発話区間に人手で発話者の情報が付与した。

表 5.1 に検証に用いるデータの詳細、表 5.2 に音源識別による発話区間検出精度を示す。

表 5.1: 評価用音声データの詳細

データ ID	収録時間	話者数	全発話数
ニュース 1	30 分 3 秒	20	337
ニュース 2	30 分 3 秒	31	312
ニュース 3	30 分 3 秒	21	324
ニュース 4	30 分 4 秒	20	324
ニュース 5	20 分 3 秒	13	159

表 5.2: 評価用音声データの発話区間検出精度 [%]

データ ID	Recall	Precision	F-meature
ニュース 1	89.49	91.60	90.53
ニュース 2	84.09	95.54	89.45
ニュース 3	88.30	85.99	87.13
ニュース 4	90.06	83.33	86.56
ニュース 5	90.95	90.30	90.63

5.2 発話区間の結合実験

本節では、ニュース番組音声の発話区間を対象として前後の発話区間が同一話者である可能性が高いとき発話区間を結合し、i-vector 抽出精度の向上を目指す。

5.2.1 実験方法

4 章で述べた手法を用いて結合実験を行う。手法は以下の通りである。

- 手法 1 : 発話間の時間情報を考慮した発話区間の結合手法
- 手法 2 : 発話環境を考慮した発話区間の結合手法
- 手法 3 : 手法 1 + 手法 2

また、手法 1 では非発話区間の長さの閾値 T によって結合するか否かを決定するため、閾値 T によって発話区間の結合精度が変化する。本実験では図 3.2 より、0.8 秒から 1.5 秒までを 0.1 秒刻みで閾値 T を変更して行う。また、非発話区間が閾値 T の時間より短いかつ、i-vector のコサイン類似度が 0.2 以上の時、発話区間を結合するとした。

5.2.2 評価方法

表 5.3 を用いて、同一話者の発話区間の結合精度 Acc_{same}(式 5.1) と話者の切り替わりの検出精度 Acc_{diff}(式 5.2) を定義する。

表 5.3: 発話区間の結合の正誤判定

		「発話者」のラベルが付与された発話区間	
		前の発話が同一話者	前の発話が異なる話者
判定結果	正	①	②
	誤	③	④

$$Acc_{same} = \frac{①}{① + ③} \tag{5.1}$$

$$Acc_{diff} = \frac{②}{② + ④} \tag{5.2}$$

また、同一話者の発話区間の結合精度 Acc_{same} に対して、結合した発話区間の時間の合計を Acc_{time} を式 5.3 として定義する。

$$Acc_{time} = \frac{\text{結合した発話区間の時間の合計}}{\text{発話区間の時間の合計}} \tag{5.3}$$

5.2.3 実験結果

発話区間の結合精度の結果を表 5.4、表 5.5、表 5.6 に示す。

表 5.4: 手法 1 による発話区間の結合結果

T	Acc_{same}	Acc_{time}	Acc_{diff}
0.8	0.634	0.679	0.877
0.9	0.660	0.706	0.869
1.0	0.680	0.724	0.853
1.1	0.699	0.745	0.845
1.2	0.710	0.756	0.834
1.3	0.717	0.766	0.826
1.4	0.727	0.722	0.815
1.5	0.736	0.783	0.796

表 5.5: 手法 2 による発話区間の結合結果

Acc_{same}	Acc_{time}	Acc_{diff}
0.827	0.831	0.259

表 5.6: 手法 3 による発話区間の結合結果

T	Acc_{same}	Acc_{time}	Acc_{diff}
0.8	0.566	0.601	0.886
0.9	0.584	0.619	0.883
1.0	0.597	0.631	0.877
1.1	0.611	0.648	0.879
1.2	0.622	0.658	0.869
1.3	0.626	0.663	0.861
1.4	0.631	0.668	0.856
1.5	0.636	0.674	0.847

手法 1、手法 3 はともに、閾値 T を上げると $Acc_{same}(Acc_{time})$ が向上するが、Acc_{diff} は低下する傾向にある。また、Acc_{diff} が同じ値をとる時、$Acc_{same}(Acc_{time})$ は手法 1 の方が高い精度で発話区間を結合できている。

手法 2 は手法 1 や手法 3 と比較して同一話者の発話区間の結合精度が高いが、話者の切り替わりの検出精度が大きく低下している。

5.2.4 考察

手法 2 は他の手法と比較して大きく Acc_{diff} が低下したが、これは話者の切り替わりが起こる時、必ずしも発話環境が変化するわけではないためである。例として、以下の場合がある。

(1) インタビューイの切り替わり

(2) アンカーから天気アナウンサーへの切り替わり

(3) アンカーから屋内にいる中継アナウンサーへの切り替わり

(1) の場合、街中など人が多い場所、かつ同じ場所で発話していることが多い。また、(2) と (3) の場合、同じスタジオもしくは雑音が殆どない環境で発話している。以上のことから、音源識別による発話環境の変化のみで話者の切り替わりの検出は難しいと考えられる。

また、手法 3 は手法 1 と比較して Acc_{diff} が向上しているが、$Acc_{same}(Acc_{time})$ が大きく低下している。これは、アンカーの発話中に映し出される映像の音が重なることによって発話環境が変化したと判別したためであると考えられる。また、同じ Acc_{diff} の値に対して $Acc_{same}(Acc_{time})$ は手法 1 が高い数値を示している。このことから、発話区間の結合、および話者の切り替わりの検出は手法 1 が最も有効であると考えられる。

Acc_{diff} と $Acc_{same}(Acc_{time})$ は反比例の関係になっていることから、発話区間を多く結合したい場合は T を下げ、話者の切り替わりを多く検出したい場合は T を上げるなど、使い分けができると考えられる。

5.3 アンカーの発話群検出実験

本節では、i-vector を用いてアンカーの発話区間検出を行う。

5.3.1 実験方法

本節では、5.2 節の手法 1 で得られた i-vector を用いてアンカーの発話区間検出を行う。i-vector を用いたアンカーの発話区間抽出方法を 5.3.2 節に示す。

5.3.2 i-vector を用いたアンカーの発話区間抽出方法

ニュース番組では、アンカー以外にインタビューイ (インタビューの受け手) や中継の有無によって話者数が大きく異なる。そのためクラスタ数を決定した場合、クラスタ数と話者数に不一致が起こり同一アンカーの発話群検出精度が低下する場合がある。そこで，同一話者の発話データの i-vector はベクトル空間上で局所的に分布することに着目した。アンカーの発話数は非アンカーと比較して多いことから多くのアンカーの発話が局所的に集まると考えたため、同一アンカーの発話データをより精度よく検出できると考えた。

そこで、2 つの発話データの i-vector のコサイン類似度が閾値以上の場合、その 2 つの発話データの話者は同一話者であると仮定した。まず、全ての発話データ間の i-vector のコサイン類似度を求める。次に、このコサイン類似度が閾値以上となる発話データ数が最も多い発話データを同一アンカーの発話データ群 O のセントロイドとし、閾値以上 (話者性が類似している) の全データをそのデータ群 O の初期要素とする。

一方、i-vector を抽出する発話データの発声の抑揚が大きい場合、同一話者の発話間の i-vector であってもコサイン類似度が閾値以下になる場合がある。そこで、発話データ $u_i(\in O)$ と発話データ群 O の距離が一定距離以内であるとき、発話データ u_i は発話データ群 O の要素として追加する。

5.3.3 評価方法

評価は、正解ラベルを用いて検出された話者の発話区間と比較して行う。

表 5.7: アンカーの発話区間の正誤判定

		「発話者」のラベルが付与された発話区間	
		アンカーの発話区間	アンカー以外の発話区間
判定結果	正	TP	FP
	誤	FN	TN

表 5.7 が得られると P（適合率（Precision））と R（再現率（Recall））は式 5.4 と式 5.5 のようにそれぞれ定義できる。

$$P = \frac{TP}{TP + FP} \tag{5.4}$$

$$R = \frac{TP}{TP + FN} \tag{5.5}$$

すなわち、適合率とは識別結果にどれだけ「ゴミ」がないかを表している。一方、再現率は識別にどれだけ「漏れ」がないかを表している。一方、したがって、適合率と再現率は大きい値ほど性能がよいことになる。ここで、2 つのシステムを比較する場合は 1 次元のスカラ値によって、2 値的な判断ができたほうが便利である。適合率と再現率をひとつのスカラ値に変換する手法として F 値（$F - measure$）がある。

$$F = \frac{1}{\frac{1}{P} + \frac{1}{R}} \tag{5.6}$$

ここで P と R はそれぞれ適合率、再現率を表す。本研究では、評価方法として適合率、再現率、F 値を用いる。

5.3.4 実験結果

5.3.5 考察

5.4 アンカーの発話区間の音声認識実験

5.4.1 実験方法

5.4.2 音響モデルの仕様

本実験で用いた DNN-HMM 音響モデルの仕様を表 5.8 に示す。この仕様に関しては小島らの研究 [10] で使用されたもので、状態数は 3000、音響特徴の次元数は 39 次元 (表 5.9)、隠れ層の数は 6 層、各層における繰り返し学習数は 5 回、隠れ層のノード数は 1024 とした。以下に、DNN を用いた際の学習の手順を示す。

表 5.8: 音響モデルの仕様

状態数	使用した音素	混合数
3,000	27	16

表 5.9: 使用する音響特徴パラメータ

特徴量	次元数
MFCC	12
POW	1
ΔMFCC	12
ΔPOW	1
ΔΔMFCC	12
ΔΔPOW	1
計	39

構築手順

DNN を用いた音響モデルの構築や、この音響モデルを用いた音声認識に必要な学習テキストや言語モデルを作成する為に Kaldi ツールキットを用いた [11]。このツールキットの大きな流れを図 5.1 に示す。まず学習や評価に必要なデータを用意し、言語モデルと単語辞書の Weighted Finite State Transducer （WFST）を作成する。WFST とは重み付き有限トランスデューサといい、状態遷移機械モデル有限オートマトンの一種である。次に音声データから特徴量を抽出したデータを準備し、このデータと書き起こしを用いて GMM-HMM による音響モデルの WFST を作成する。これらの WFST を、合成等を行ない 1 つの WFST とする。この WFST を用いて音声認識を行ない、学習データのアライメント（フレームごとの音素情報）をとる。このアライメントを用いて DNN を用いた音響モデルの学習（プレトレーニングと微調整）を行ない、最終的な音声認識を行なう。

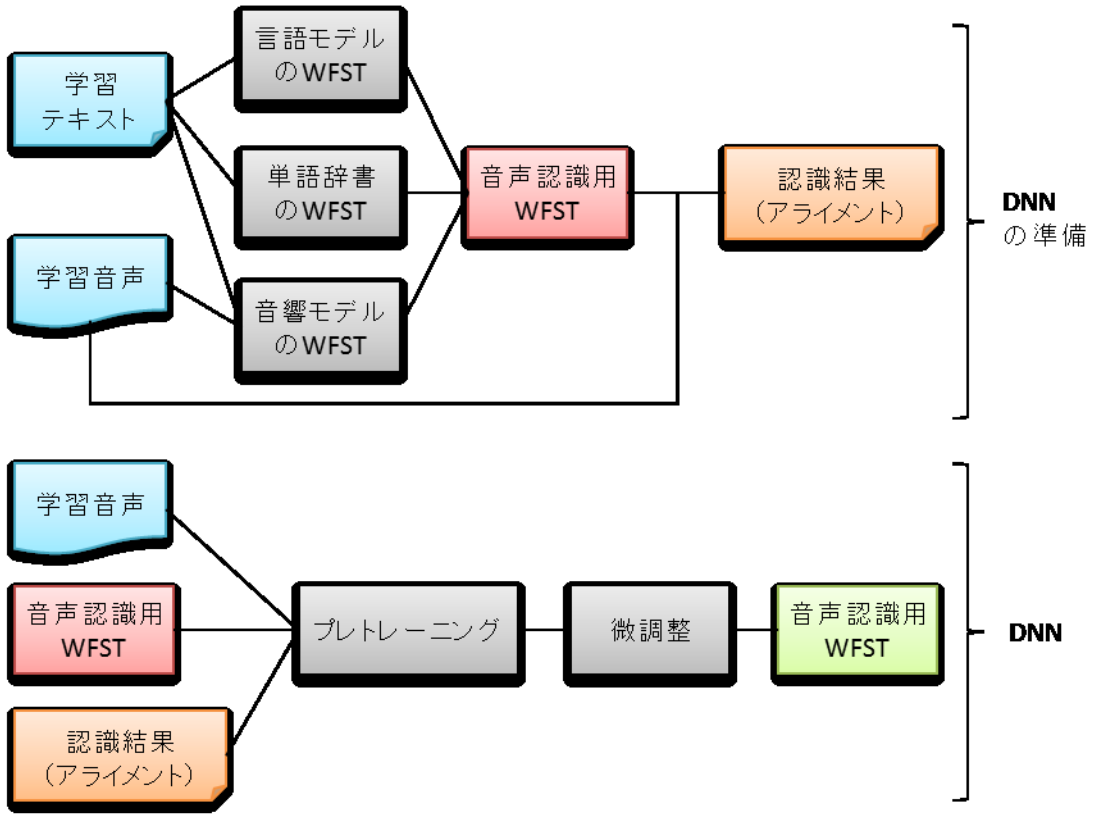

図 5.1: DNN を用いる際の学習の流れ

使用コーパス

音声認識は統計的モデルを用いるため、大量の音声・言語素材が必要である。本研究では 2004 年、国立国語研究所・情報通信研究機構・東京工業大学が共同開発した「日本語話し言葉コーパス」(Corpus of Spontaneous Japanese：CSJ) を使用する。この CSJ は日本語の自発音声を大量に集めて多くの研究用情報を付加した話し言葉研究用データベースである。コーパスとは様々な研究機関において共通に利用可能な大量のデータのことである。全体で約 660 時間の自発音声 (語数にして約 700 万個) が格納されている。

CSJ に収録されている音声の種類と分量を表 5.10 に示す。学会講演は、国内の様々な学会でライブ録音された研究発表音声である。収録された学会は、工学ないし自然科学系が 3 学会、621 ファイル、人文科学系が 4 学会、187 ファイル、社会科学系が 2 学会、169 ファイルであり、理工学系の学会での話者は男性の大学院生であることが多いため、学会講演の話者は年齢と性別に偏りがある。講演時間は、大部分が 12 分から 25 分程度の長さであるが、なかには 1 時間を超える招待講演の類も含まれている。模擬講演は、人材派遣会社によって選定された一般話者による日常話題についての「スピーチ」である。模擬講演の話者は、性別と年齢がほぼ均等に分布されている。話者は三つの大まかなテーマを与えられ、それぞれについて平均 12 分程度のスピーチを行なった。

使用した音素

本研究で使用した音素 27 個を表 5.11 に示す。また、その音素をもとに記したカナ音素対応表を

表 5.10: CSJ の音声の種類と分量

音声の種類	話者数	ファイル数	独話・対話	時間数
学会講演	838	1007	独話	299.5
模擬講演	580	1699	独話	324.1
朗読音声	244	491	独話	14.1
インタビュー話者による模擬講演	16	16	独話	3.4
学会講演インタビュー	10	10	対話	2.1
模擬講演インタビュー	16	16	対話	3.4
課題志向対話	16	16	対話	3.1
自由対話	16	16	対話	3.6
再朗読	16	16	独話	5.5

表 5.12 に示す

表 5.11: 使用した音素

母音	子音	濁音	半濁音	撥音	促音	無音
a i u e o	ch f h j k m n r s sh t ts w	b d g z zh	p	ng	q	#

表 5.12: カナ音素対応表

ア	a	イ	i	ウ	u	エ	e	オ	o
カ	k a	キ	k i	ク	k u	ケ	k e	コ	k o
サ	s a	シ	sh i	ス	s u	セ	s e	ソ	s o
タ	t a	チ	ch i	ツ	ts u	テ	t e	ト	t o
ナ	n a	ニ	n i	ヌ	n u	ネ	n e	ノ	n o
ハ	h a	ヒ	h i	フ	f u	ヘ	h e	ホ	h o
マ	m a	ミ	m i	ム	m u	メ	m e	モ	m o
ラ	r a	リ	r i	ル	r u	レ	r e	ロ	r o
ワ	w a								
ガ	g a	ギ	g i	グ	g u	ゲ	g e	ゴ	g o
ザ	z a	ジ	zh i	ズ	z u	ゼ	z e	ゾ	z o
ダ	d a	ヂ	d i	ヅ	d u	デ	d e	ド	d o
バ	b a	ビ	b i	ブ	b u	ベ	b e	ボ	b o
パ	p a	ピ	p i	プ	p u	ペ	p e	ポ	p o
ヤ	j a	ユ	j u	ヨ	j o				
キャ	k j a	キュ	k j u	キョ	k j o				
ギャ	g j a	ギュ	g j u	ギョ	g j o				
シャ	sh j a	シュ	sh j u	ショ	sh j o				
ジャ	zh j a	ジュ	zh j u	ジョ	zh j o				
チャ	ch j a	チュ	ch j u	チョ	ch j o				
ニャ	n j a	ニュ	n j u	ニョ	n j o				
ヒャ	h j a	ヒュ	h j u	ヒョ	h j o				
ビャ	b j a	ビュ	b j u	ビョ	b j o				
ピャ	p j a	ピュ	p j u	ピョ	p j o				
ミャ	m j a	ミュ	m j u	ミョ	m j o				
リャ	r j a	リュ	r j u	リョ	r j o				
イェ	i e	シェ	sh e	ジェ	zh e	ティ	t i	トゥ	t u
チェ	ch e	ツァ	ts a	ツィ	ts i	ツェ	ts e	ツォ	ts o
ディ	d i	ドゥ	d u	デュ	d u	ニェ	n i e	ヒェ	h e
ファ	f a	フィ	f i	フェ	f e	フォ	f o	フュ	f u
ブィ	b i	ミェ	m e	ウィ	w i	ウェ	w e	ウォ	w o
クヮ	k a	グヮ	g a	スィ	s i	ズィ	z i	テュ	t e j u
ヴァ	b a	ヴィ	b i	ヴ	b u	ヴェ	b e	ヴォ	b o
ン	n g	ッ	q					無音	#

木構造話者クラスタ

先行研究 [2][3] では、話者の音響特徴、話者特徴ごとに作成した木構造話者クラスタ、音響モデルを作成することで音声認識精度の向上を確認したため、本研究でも使用する。このクラスタは、母音の定常状態である HMM の中央の状態の平均と分散を用いた Bhattacharyya 距離による k-means 法によって作成した。クラスタの個数は、最上位のクラスタを 2 分割し、作成された 2 つのクラスタをさらに 2 分割した計 7 つのクラスタを使用する。

5.4.3　言語モデル・単語辞書の仕様

言語モデルはトライグラムモデルを構築した。以下、使用した学習テキストを説明する。

CSJ

CSJ には書き起こしテキストも提供されており、その一部の例を図 5.2 に示す。書き起こしテキストは主に情報部と発話部に区別される。情報部では発話 ID や時間情報等を、発話部では発話内容を「&」の左側に基本形、右側に発音形という形式で記している。発話形はカタカナを用いて実際に発音された音声を忠実に表記したものである。発音の怠けや言い間違い等を書き取れる範囲で忠実に記録している。本研究では、音響モデル構築の際には主に発話部の発音形を用い、このカタカナ表記を音素列に変換し、ラベルファイルとして定義する。

```
0089 00233.188-00234.021 L:
こんな                                  & コンナ
こと                                    & コト
言ってると                              & ユッテルト
0090 00234.587-00235.552 L:
いう                                    & ユー
風な                                    & (W フ;フー)ナ
感じです                                & カンジデス
0091 00236.322-00237.419 L:
ただ                                    & タダ
これだと                                & コレダト
ちょっと                                & チョット
0092 00237.895-00240.618 L:
差分の                                  & サブンノ
データーとして                          & データートシテ
精度が                                  & セードガ
悪いので                                & ワルイノデ
```

図 5.2: 書き起こしテキストの例

本研究ではこの CSJ をベースに学習テキストを構成する。使用するデータは 977 講演分のテキストで、約 14MB である。

拡張したコーパスによる学習テキスト

この学習テキストは江頭らによる、学術講演の書き起こしと新聞記事に拡張されるテキストとして参加者名の入ったテキスト、Web から収集してきたテキスト、そして対話コーパスから作成される対話テキストを追加した未知語の減少に着目した学習テキストである。この学習テキストは会議中に参加者の名前を呼ぶことが多い、会議は対話形式であるなどの会議の特徴を考慮した学習テキストである。テキストサイズは約 100MB である。以降本論文では、このテキストを拡張したコーパスによる学習テキストと呼ぶ。

拡張したコーパスによる学習テキスト

この学習テキストは荒井らによる、会議における発話行為に着目して作成された学習テキストである。学術講演の書き起こしと新聞記事に対話表現に近い特徴を持っていると考えられる Q & A サイトから収集したテキストと対話コーパスを追加した学習テキストである。テキストサイズは約 44MB である。以降本論文ではこのテキストを対話特化テキストと呼ぶ。

5.4.4　評価方法

本研究では評価尺度としては式 5.7 で与えられる単語正解精度 Acc(Word Accuracy) を用いる。ここで W は単語数、S(Substitution) は置換誤り、D(Deletion) は脱落誤り、I(Insertions) は挿入誤りの単語数を表わす。置換誤りとは、正解の単語が別の単語に誤認識された場合の誤りである。脱落誤りとは、単語があるべき部分に認識結果が何も出力されなかった場合の誤りである。挿入誤りは、本来単語がない部分に誤認識結果として単語が出力された場合の誤りである。

$$Acc = \frac{(W - S - D - I)}{W} \tag{5.7}$$

評価は、正解ファイルと認識結果のファイルを DP マッチングを行なうことにより算出する。この正解ファイルは形態素解析した結果の形態素列によって作成したものである。また、本研究ではアンカーの発話区間を対象とした音声認識を行うため、5.3 章で検出した発話区間より、アンカー以外の発話区間で認識された単語は全て挿入誤り、アンカーの発話として検出出来なかった発話区間の単語は全て削除誤りとして計算する。

5.4.5　実験結果

5.4.6　考察

第6章

結論

謝辞

最後に，本研究および本修士論文作成にあたり暖かい御指導および適切な御助言をして頂いた松永　昭一教授，また，関係者各位に心より感謝いたします.

また，同研究室博士前期 (修士) 課程 2 年の博士前期 (修士) 課程 1 年の学士課程 4 年のその他関係各位に心から感謝いたします.

参考文献

[1] 辻川美沙貴, 西川剛樹, 松井知子,”i-vector による短い発話の話者識別の検討”, 電子情報通信学会,18,June,2015

[2] 俵直弘, 小川哲司, 小林哲則,”i-vector を用いたスペクトラルクラスタリングによる雑音環境下話者クラスタリング”, 情報処理学会,28,February,2015

[3] 冨久祐介,“ 音源識別のための音クラスタリングとガウス分布混合数の有効性の検討 ”, 長崎大学工学部情報システム工学科平成,19 年度卒業論文,2008

[4] 水野理, 大附克年, 松永昭一, 林良彦: “ ニュースコンテンツにおける音響信号自動判別の検討 ”, 電気情報通信学会総合大会,2003

[5] 新美康永,“ 音声認識 ”, 共立出版株式会社,1979

[6] 国立情報学研究所データセット集合利用研究開発センター”ATR バランス文”