修　士　論　文

題　　目

ニュースアンカーの発話間隔と環境を考慮した i-vector を用いた発話検出

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

序章

1.1 はじめに

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、これらの情報全てが必要な情報とは限らず、ほとんどの場合取捨選択をする必要がある。ニュース番組で例えると、ニュース番組はスポーツ、経済、社会、天気予報など様々なジャンルのトピックで構成されていて、視聴者は時間の都合や興味の違いに応じて必要なトピックのみを早送りなどで視聴する。そこで、これらのニュース番組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、所望のトピックのみを視聴することができる。また、テレビ局などの管理する側も、データベースの構築が容易になるというメリットがある。しかし世の中には膨大な量のニュース番組が存在しており、それら全てに人手でインデクスを付与することは事実上不可能であるため、自動的にインデクシングすることが望まれる。

自動でニュース番組のインデクシングを行うためには、ニュース番組内の発話区間、発話者、発話内容の特定が必要であり、これらを推定する技術をダイアライゼーションと呼ぶ。ニュース番組には発話数が多く、ニュース番組の司会およびトピックの切り替えを行う話者としてアナウンサー(ニュースアンカー) が存在する。ニュースアンカーの発話にはトピックのキーワードが多く含まれており、インデクシングに重要な情報を持つ。つまり、アンカーの発話区間を検出し、音声認識を行うことはニュース番組のダイアライゼーションの実現に有効であると考えられる。

ニュースアンカーの発話を音声認識するためには、まずニュースアンカーの発話を検出する必要がある。特定話者の話者識別には話者特徴量 (i-vector) が一般的に用いられている [1][2]。i-vectorとは、ある発話から得られた音響特徴量を因子分解して抽出された話者固有の特徴量である。特徴抽出に因子分解を用いているため、次元を削減して特徴を表現することが可能である。近年の話者認識システムの多くは i-vector に基づいて構築されており、この領域における最高水準となっている。しかし、これまでの i-vector を用いた話者識別に関する研究は、事前に対象のデータに登場する話者の情報を学習し、その学習データを用いて話者を識別する手法をとっているため、発話区間、発話者の情報が未知の場合が多いニュース番組にそれらの手法を用いることは出来ない。そこで先行研究 [3][4] では、ニュース番組音声の発話区間、発話者が未知の場合においても用いることができるニュースアンカーの発話検出手法を提案した。これらの研究でも i-vector を用いており、ニュースアンカーの発話検出を行った結果、ニュースアンカーの発話を約 70%の精度で検出することが出来た。

しかし、短い発話から抽出された i-vector は話者の特徴を十分に表すことが難しいことが知られており [5]、先行研究 [3] でもニュースアンカーの発話検出精度の原因として述べられている。つまり、短い発話から i-vector を抽出した場合、ニュースアンカーの発話検出精度が低下する可能性がある。そこで、話者識別に必要な話者の特徴を短い発話から抽出することが可能になればニュースアンカーの発話区間検出精度の向上に有効であると考えた。

本研究では、発話から抽出される i-vector に加えて、「発話の時間間隔」と「発話環境」を考慮し、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した。これによって、長い発話を擬似的に作成し、短い発話の i-vector の抽出精度向上を目指した。検証の結果、結合した発話区間から抽出した i-vector を用いることでニュースアンカーの発話区間検出精度が約 6%向上した。よって、「発話の時間間隔」と「発話環境」を考慮して発話区間を結合することで短い発話から抽出する i-vector が話者の特徴を抽出出来たことによってニュースアンカーの発話を検出することができることを示した。

1.2 論文構成

次章以降における本論文の構成は、まず 2 章で、インデクシングの流れと i-vector に関係する基本知識の説明を行う。次に 3 章で、本論文で着目したニュース番組における「発話の時間間隔」と「発話環境」の調査を行う。4 章では提案手法によって結合した発話区間から抽出した i-vector を用いてニュースアンカーの発話検出実験を行い、提案手法によるニュースアンカーの発話検出精度への効果を検証する。

第2章

基礎知識

本章では、ニュース番組音声からインデックス作成までの流れと、ニュースアンカーの発話区間検出に必要な知識について説明する。

2.1 ニュース番組のインデクシング [6]

ニュース音声のインデクシングは図 2.1 の手順で行われる。

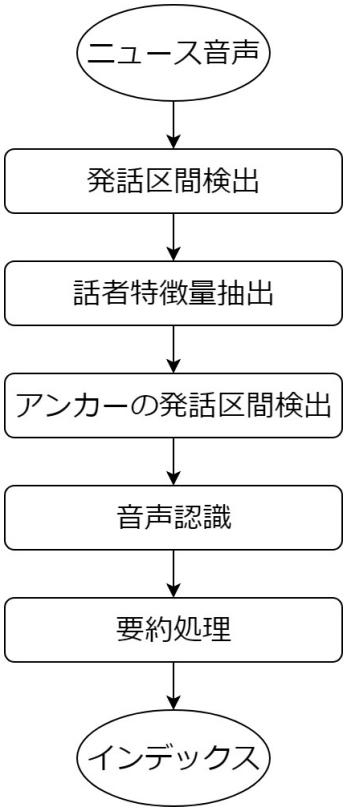

図 2.1: インデクシングの手順

入力であるニュース音声はどの区間で発話されているか未知であるため、まず音源識別によって発話区間検出を行う。次に検出した発話区間の中からアンカーの発話区間を検出するために、発話

から話者特徴量を抽出し、発話をクラスタリングすることでニュースアンカーの発話区間を検出する。検出されたニュースアンカーの発話区間の音声認識を行い、認識結果の要約処理を行うことでインデックスが作成される。ニュースアンカーの発話区間検出に使用する i-vector について、2.2 節で説明を行う。

2.2 i-vector の概要

近年の話者認識システムの多くは i-vector[10] 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。

$$M_u = Tw_u + m \tag{2.1}$$

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、

$$N_{u,c} = \sum_{t=1}^{L} \gamma_t(t) \tag{2.2}$$

$$F_{u,c} = \sum_{t=1}^{L} \gamma_t(c)(o_t - m_c) \tag{2.3}$$

$$S_{u,c} = diag\left[\sum_{t=1}^{L} \gamma_c(c)(o_c - m_c)(o_t - m_c)'\right] \tag{2.4}$$

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

$$\gamma_c(c) = p(c \mid o_t, \Omega) = \frac{\pi_c p(o_t \mid m_c, \sum c)}{\sum_{k=1}^{C} \pi_k p(o_t \mid m_k, \sum k)} \tag{2.5}$$

である。更にこれらを用いて、

$$N_u = \begin{pmatrix} N_{u,1}, I_{D_F} & 0 & \dots & 0 \\ 0 & 0 & \dots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & N_{u,C}, I_{D_F} \end{pmatrix} \in R^{CD_F \times CD_F} \tag{2.6}$$

$$F_u = \begin{pmatrix} F_{u,1} \\ F_{u,2} \\ \vdots \\ F_{u,C} \end{pmatrix} \in R^{CD_F} \tag{2.7}$$

$$S_u = \begin{pmatrix} S_{u,1}, 0 & 0 & \dots & 0 \\ 0 & S_{u,2} & \dots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & S_{u,C} \end{pmatrix} \in R^{CD_F \times CD_F} \tag{2.8}$$

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。

$$p(w) \propto \exp(-\frac{1}{2} w' w) \tag{2.9}$$

- 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

$$P(O_u|w_u) = \exp\left(\sum_{t=1}^{L} \sum_{c=1}^{C} \gamma_t(c) \log(2\pi)^{-\frac{D_F}{2}} |\Sigma_c|^{-\frac{1}{2}} - \frac{1}{2} \sum_{t=1}^{L} \sum_{c=1}^{C} \gamma_t(c) D(o_t; \theta_c) \right) \tag{2.10}$$

  のように書ける。ここで、

$$D(o_t; \theta_t) = (o_t - M_{u,c})' \Sigma_c^{-1} (o_t - M_{u,c}) \tag{2.11}$$

$$M_{u,c} = m_c + T_c w_u \tag{2.12}$$

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 (2.10) の exp の内部を Baum-Welch 統計量を用いて整理すると、

$$\sum_{t=1}^{L}\sum_{c=1}^{C}\gamma_t(c)log(2\pi)^{-\frac{D_F}{2}}\left|\Sigma_c\right|^{-\frac{1}{2}} - \frac{1}{2}\sum_{t=1}^{L}\sum_{c=1}^{C}\gamma_t(c)D(o_t;\theta_c) = G_u^{\Sigma} + H_u^{\Sigma T} + \text{Const.} \tag{2.13}$$

ここで、G_u^{Σ} 及び $H_u^{\Sigma T}$ は、

$$G_u^{\Sigma} = \sum_{c=1}^{C}\left[\frac{1}{2}N_{u,c}\log\left|\Sigma_c^{-1}\right| - \frac{1}{2}tr\left(\Sigma_c^{-1}S_{u,c}\right)\right] \tag{2.14}$$

$$H_u^{\Sigma T} = w_u' T' \Sigma^{-1} F_u - \frac{1}{2} w_u' T' N_u \Sigma^{-1} T w_u \tag{2.15}$$

- 事後分布

  w の事後分布は (2.10)〜(2.15) を用いると、

$$\begin{aligned} p(w_u|O_u) &\propto p(O_u|w_u)p(w_u) \propto \exp(w_u' T' \Sigma' T w_u - \frac{1}{2} w_u' w_u) \\ &\propto exp(w_u' T' \Sigma' F_u - \frac{1}{2} w_u' N_u \Sigma^{-1} T w_u - \frac{1}{2} w_u' w_u) \\ &\propto \exp(-\frac{1}{2}(w_u - G_u T' \Sigma^{-1} F_u)' G_u^{-1} (w_u - G_u T' \Sigma' F_u)) \end{aligned} \tag{2.16}$$

  と書ける。ここで、

$$G_u = (I + T' \Sigma^{-1} N_u T)^{-1} \tag{2.17}$$

  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、

$$E[w_u] = G_u T' \Sigma^{-1} F_u \tag{2.18}$$

$$cov[w_u] = G_u \tag{2.19}$$

  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 (2.18) 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

$$Q = \sum_{u=1}^{U} E[logp(O_w w_u|\theta)] \tag{2.20}$$

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、

$$logp(O_w w_u) = logp(O_u|w_u, \theta) + logp(w_u) \tag{2.21}$$

と書けるので、式 (2.10)〜(2.15) を用いると、式 (2.20) は以下のように整理できる。

$$\begin{aligned} Q = \frac{1}{2}\sum_{u=1}^{U}\sum_{c=1}^{C}(N_{u,c} log\left|\Sigma_c^{-1}\right| - tr(\Sigma_c^{-1} S_{u,c})) \\ + \sum_{u=1}^{U} tr\left(\Sigma^{-1}\left(F_u E[w_u']T' - \frac{1}{2}N_u T E[w_u w_u']T'\right)\right) \\ - \sum_{u=1}^{U}\frac{1}{2}tr(E[w_u W_u']) \end{aligned} \tag{2.22}$$

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。

$$E[w_u] = G_u T' \Sigma^{-1} F_u \tag{2.23}$$

$$E[w_u w_u'] = G_u + E[w_u]E[w_u'] \tag{2.24}$$

M ステップでは、式 (2.22) をパラメータに関して最大化する。まず、(2.22) 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。

$$\sum_{u=1}^{U}\Sigma^{-1} F_u E[w_u'] = \sum_{u=1}^{U}\Sigma^{-1} N_u T E[w_u w_u'] \tag{2.25}$$

これより、T の推定式が、

$$T^i = \left(\sum_{u=1}^{U} F_u^i E[w_u']\right)\left(\sum_{u=1}^{U} N_{u,c} E[w_u w_u']\right)^{-1} \tag{2.26}$$

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、$i = (c-1)\times D_F + f$、$1 \le f \le D_F$ である。また、Σ の推定式は、

$$\Sigma = N^{-1}\left(\sum_{u=1}^{U} S_u - diag\left[\sum_{u=1}^{U} F_u E[w_u']T'\right]\right) \tag{2.27}$$

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。

$$\cos(w_x, w_y) = \frac{w_x \cdot w_y}{\parallel w_x \parallel \parallel w_y \parallel} \tag{2.28}$$

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。

第3章

予備調査

本章では、本研究で用いる i-vector の性質とニュース番組の特徴について調査を行った。

3.1 コサイン類似度を用いた i-vector の性質の調査

本節では、発話の長さによって同一話者の発話間の場合と異なる話者の発話間の 2 つの場合で i-vector のコサイン類似度がとる値についての調査を行う。

3.1.1 使用する音声データ

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 [7] を使用した。読み上げ音声には、男女各 110 人 × 50 発話分が収録されている。

3.1.2 調査方法

各話者の音声データから一発話を取り出し、それ以外の音声データとの i-vector のコサイン類似度を算出する。また、同一話者の発話間の場合と異なる話者の発話間の 2 つの場合で発話の長さごとのコサイン類似度の平均値を調査する。

3.1.3 コサイン類似度の算出条件

i-vector の抽出には、ALIZE と LIR RAL[8] を用いる。読み上げ音声 [7] に収録されている各発話データから i-vector を抽出する。発話データから抽出する音響特徴パラメータを表 3.1 に示す。また混合数は 32 とした。

表 3.1: 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

本稿では、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]

3.1.4　調査結果

同一話者間の i-vector の特徴

同一話者間の i-vector のコサイン類似度を図 3.1、その標準偏差を図 3.2 に示す。

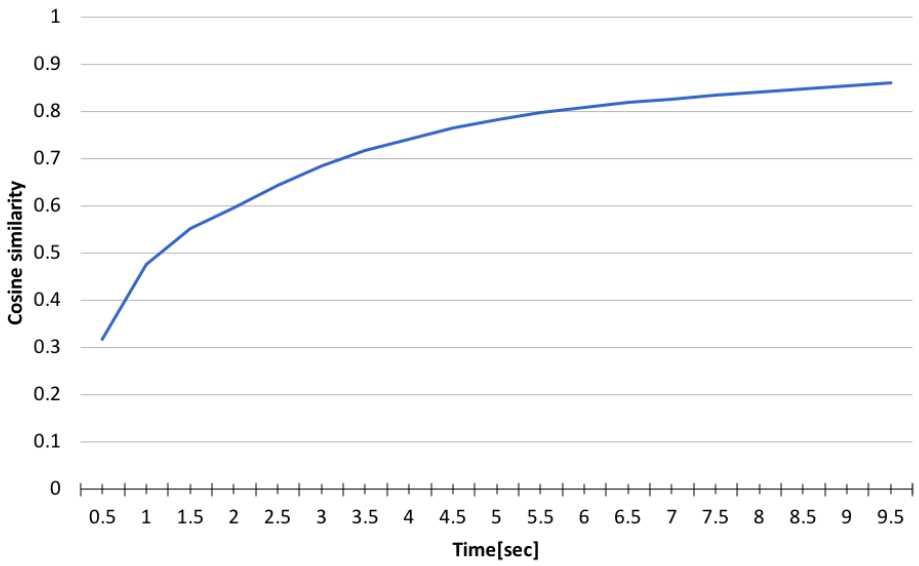

図 3.1: 同一話者間の i-vector のコサイン類似度

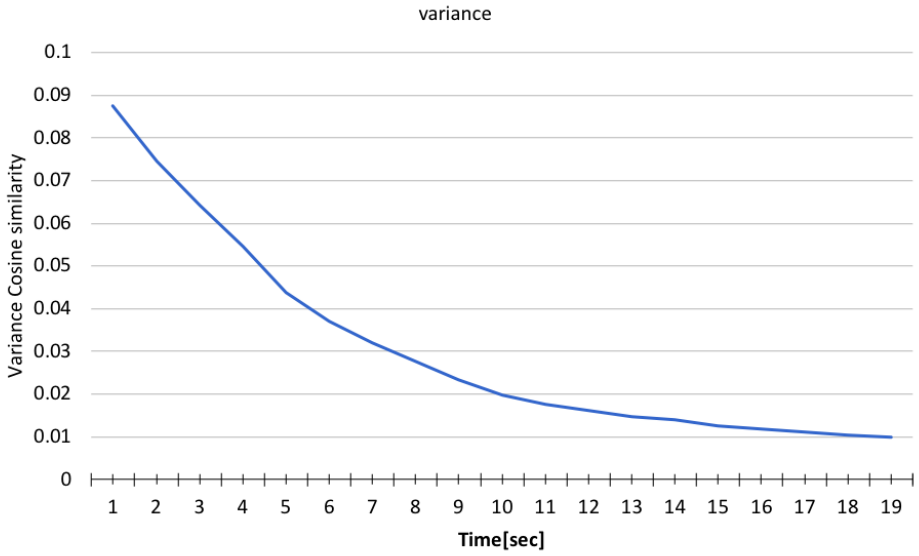

図 3.2: 同一話者間の i-vector のコサイン類似度の標準偏差

図 3.1 より、発話が長いほどコサイン類似度の値が高くなる傾向にある。また、図 3.2 より、発話が短い場合はコサイン類似度の標準偏差が大きく、長くなるにつれて収束する傾向にある。

異なる話者間の i-vector の特徴

異なる話者間の i-vector のコサイン類似度を図 3.3、その標準偏差を図 3.4 に示す。

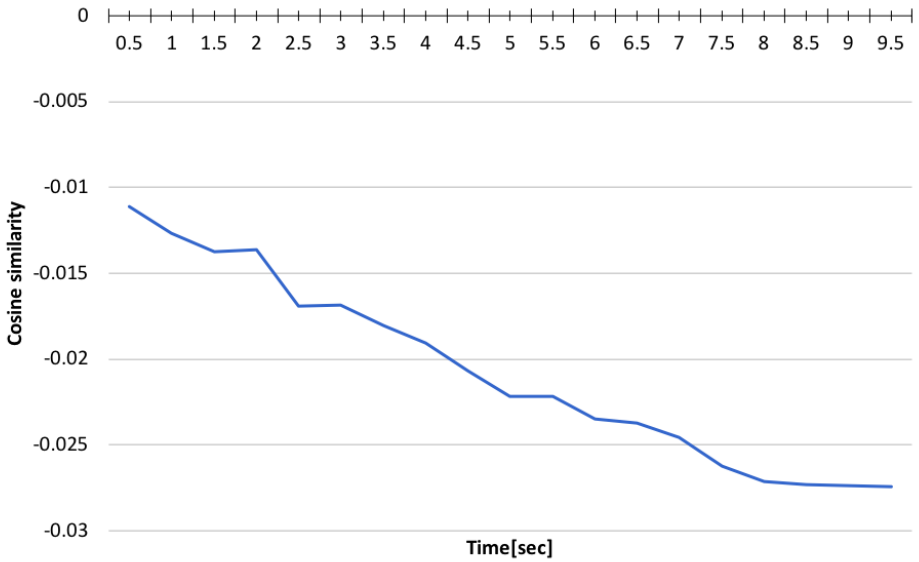

図 3.3: 異なる話者間の i-vector のコサイン類似度

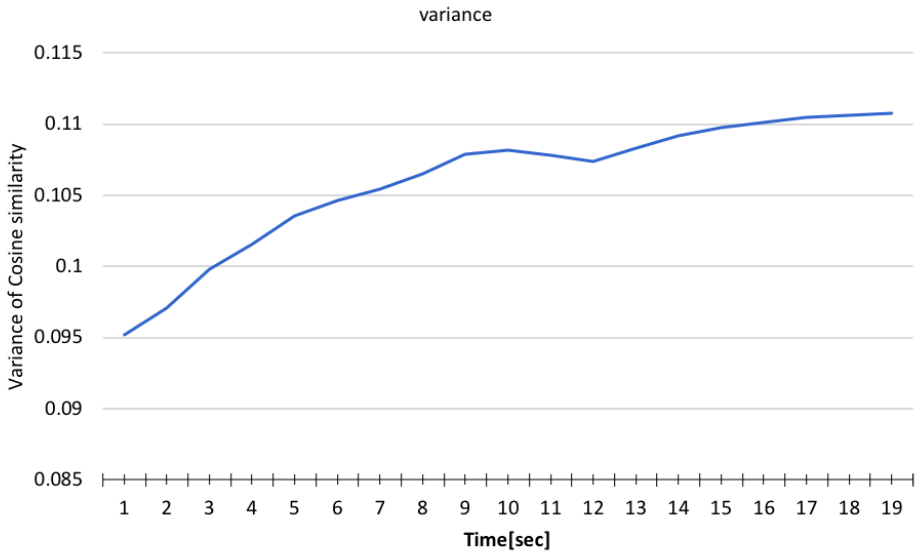

図 3.4: 異なる話者間の i-vector のコサイン類似度の標準偏差

図 3.3 より、発話が長くなるほどコサイン類似度が低くなる傾向にある。また、図 3.4 より、発話が長くなるごとにコサイン類似度の標準偏差が大きくなる傾向にある。

3.1.5　考察

発話が短い時、同一話者間のコサイン類似度の標準偏差が大きい理由として、短い発話からは話者の特徴が十分に抽出できていないためであると考えられる。また発話が短い時、異なる話者間のコサイン類似度の標準偏差が小さいことから、話者の特徴を抽出できていない場合、話者が異なっていても i-vector のコサイン類似度が似た値をとることがわかる。つまり、i-vector を用いた話者識別、話者照合を行う場合、できるだけ長い発話を用いる必要がある。

3.2　ニュース番組音声における発話間隔の調査

本節は、ニュース番組音声における発話間隔の調査を行う。発話と発話の間の区間を非発話区間として、同一話者間の非発話区間の長さの平均と異なる話者間の非発話区間の長さの平均をそれぞれ計測する。

3.2.1　使用する音声データ

本調査では、ニュース番組の音声データ 12 個を用いる。各音声データには、事前に人手で 3 種類 (音楽、音声、雑音) 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。表 3.2 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切り出し、それぞれを一発話とした。

表 3.3 に調査に用いるデータの詳細を示す。

表 3.2: 「音声」の音源ラベルの例

time	speaker
18.526910	-1 male1_INT_S
20.793192	-1 male1_INT_E
21.293665	-1 male1_INT_S
23.116141	-1 male1_INT_E
23.654385	-1 male1_INT_S
26.270058	-1 male1_INT_E
27.799800	-1 male_S
29.811134	-1 male_E
30.368265	-1 male_S
34.277610	-1 male_E

表 3.3: 調査音声データの詳細

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

3.2.2　調査方法

表 3.3 のニュース番組音声と付与された「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、発話と発話の間の非発話区間の長さを計測する。また、発話が重なっている場合は非発話区間を 0 秒とした。

3.2.3　調査結果

同一話者間の発話の時間間隔を図 3.5、異なる話者間の発話の時間間隔を図 3.6、図 3.7 に示す。

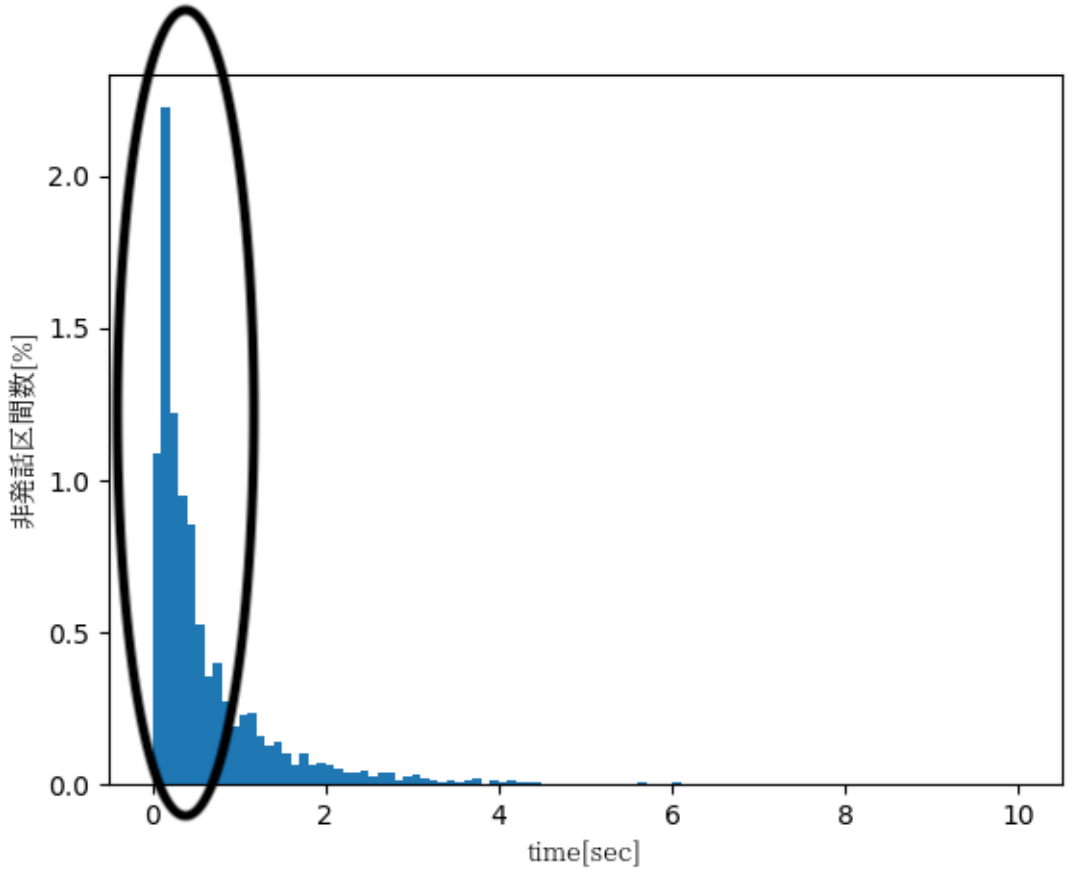

図 3.5: 同一話者間の発話の時間間隔

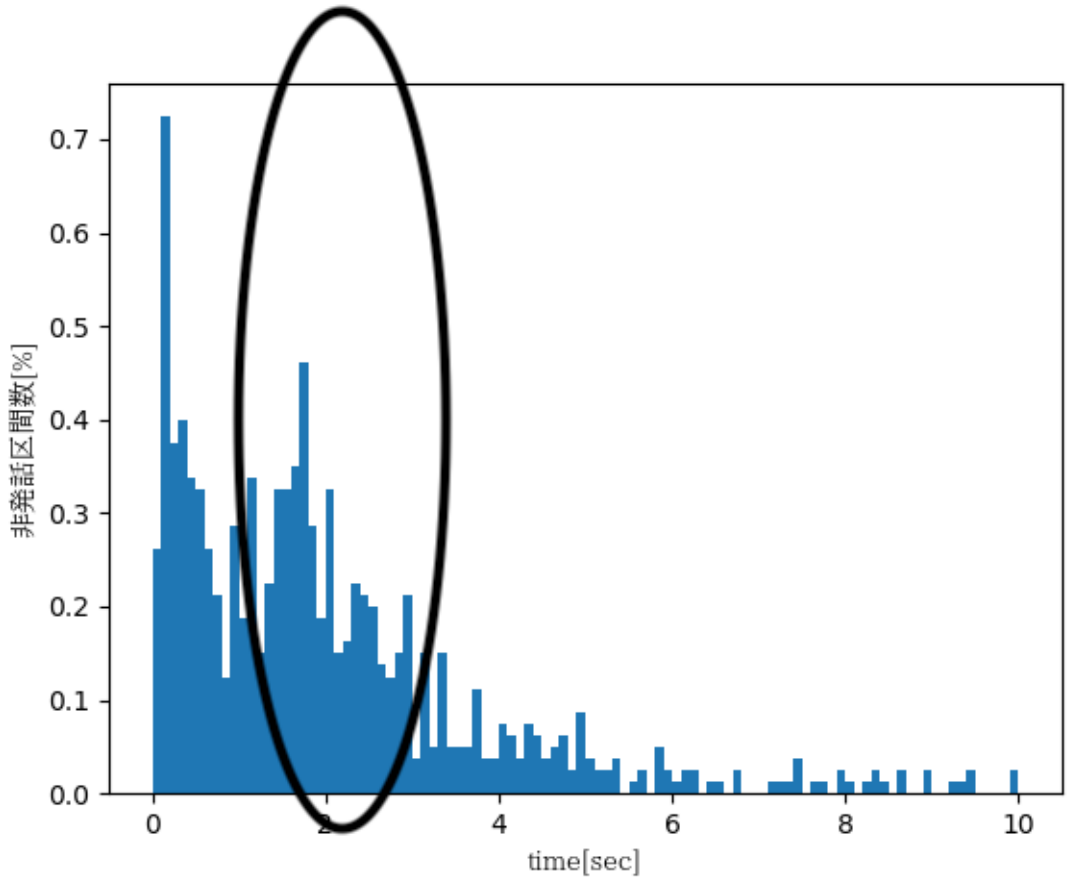

図 3.6: 異なる話者間の発話の時間間隔

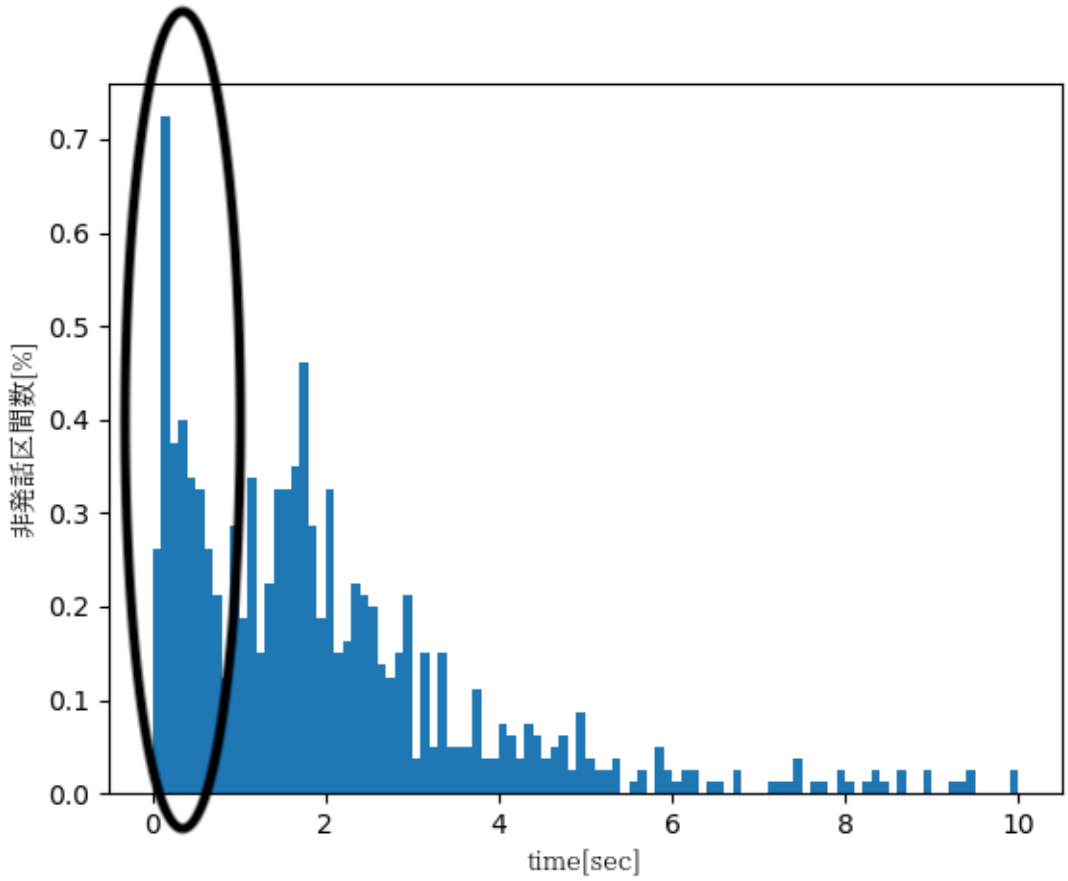

図 3.7: 異なる話者間の発話の時間間隔

図 3.5 より、同一話者の発話の時間間隔は 1 秒以下の割合が多い。また、図 3.6 と図 3.7 より、異なる話者間の発話の時間間隔は 1 秒以下と 2 秒程度の割合が多い。

3.2.4　考察

図 3.5 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い。つまり同じ話者が連続で発話する場合、間髪入れずに発話することがわかる。

また、図 3.6 より、話者が切り替わる場合は非発話区間が比較的長くなる。しかし、図 3.7 で示されているように、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。

第4章

アンカーの発話検出実験

本章では、発話区間結合を行い、結合した発話区間から抽出した i-vector を用いてニュースアンカーの発話検出を行った。

4.1 実験条件

i-vector の抽出には、ALIZE と LIR RAL[8] を用いる。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 [7] を使用する。読み上げ音声に収録されている各発話データから i-vector を抽出する。発話データから抽出する音響特徴パラメータを表 4.1 に示す。また混合数は 32 とした。

表 4.1: 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

評価データ

「音声」の音源ラベルと発話の書き起こしが付与されているニュース番組 5 番組分を用いてニュースアンカーの発話区間検出を行う。ニュース番組音声の詳細を表 4.2 に示す。また、ニュース番組 5 番組分に音源識別を用いて検出した発話区間の詳細を 4.3 に示す。音源識別の詳細は付録 A に記載する。

表 4.2: 評価データの詳細

ニュース ID	ニュースアンカー数	発話区間数	ニュースアンカーの発話区間数
ニュース 1	1	345	165
ニュース 2	2	519	149
ニュース 3	2	608	258
ニュース 4	2	518	219
ニュース 5	2	520	285

表 4.3: 検出した発話区間数とニュースアンカーの発話区間数

ニュース ID	発話区間数	ニュースアンカーの発話区間数
ニュース 1	345	165
ニュース 2	519	149
ニュース 3	608	258
ニュース 4	518	219
ニュース 5	520	285

4.2　i-vector の抽出精度向上のための発話区間結合手法

i-vector は発話ができるだけ長いほうが正確に話者の特徴を抽出することができる。そこで、時系列順に並んでいる発話区間のうち、前後の発話が同一話者である可能性が高い発話区間を結合、擬似的に長い発話を作成する。本稿では発話から抽出できる i-vector に加えて、「発話の時間間隔」「発話環境」を考慮した 2 通りの手法で発話区間を結合した。図 4.1 は本稿の提案手法を組み込んだインデックス付与までの流れである。

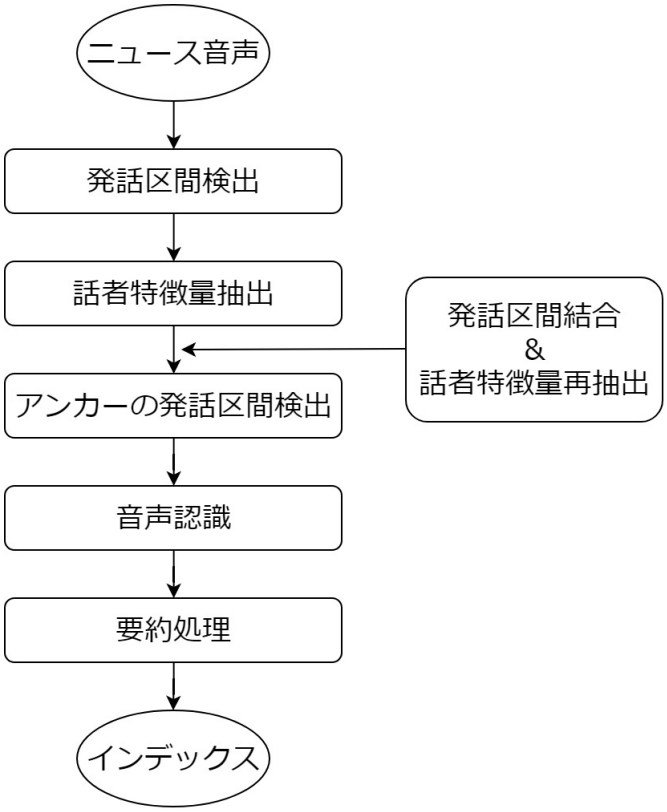

図 4.1: 提案手法を組み込んだインデクシング手法

4.2.1　発話の時間間隔を考慮した発話区間の結合手法

同一話者が連続で発話する場合、間をおかずに次の発話を行うことが多い。そのため、発話区間と発話区間の間 (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ていると考えられる。また、インタビューイや中継アナウンサーなど話者が切り替わった場合、ニュースの画面も切り替わるため非発話区間は長くなる。そこで本手法では、発話から抽出されるコサイン類似度が一定値以上を示し、かつ非発話区間が非常に短いとき、同一話者の発話と判別して発話区間を結合する。

4.2.2　発話環境を考慮した発話区間の結合手法

ニュース番組にはスタジオにいるアンカーのほか、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。本稿で使用する音源識別システムはニュース番組音声を「音声」「背景雑音」「音楽」「無音」のいずれかに分類する。そのため、「音声」以外の区間、つまり非発話区間の音源識別結果である「背景雑音」「音楽」「無音」の検出結果が変化した時、発話環境の変化したと識別することができる。そこで本手法では、発話から抽出されるコサイン類似度が一定値以上を示し、かつ発話環境が変化していないとき、同一話者として発話区間を結合する。

4.3　i-vector を用いたニュースアンカーの発話区間検出手法 [3]

ニュースアンカーの発話区間検出のために、i-vector を用いて話者クラスタを作成し、クラスタに含まれる発話が多いクラスタをニュースアンカーのクラスタとして発話区間を検出した。従来は話者クラスタの作成に k-means が多く用いられていたが、ニュース番組ではニュース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ため、あらかじめクラスタ数を決定する必要がある k-means クラスタリングを用いることは適切ではないと考えた。そのため、ニュースアンカーの発話数は非ニュースアンカーと比較して多いことと、i-vector はベクトル空間上で話者ごとに局所的に分布することに着目した。多くの発話の i-vector が局所的に分布している部分のみをクラスタリングすることで、同一ニュースアンカーの発話区間を検出する。

本手法では、2 つの発話データの i-vector のコサイン類似度が閾値 Th_{cos} 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 Th_{cos} 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 Th_{cos} 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 Th_{cos} 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。以上の手順を繰り返してクラスタリングを行い、クラスタに含まれる発話が一定数以下となった時、クラスタリングを終了する。

4.4　実験方法

同一話者の可能性が高い発話区間を結合し、結合した発話区間から i-vector を再抽出した。次に、再抽出した i-vector を用いてアンカーの発話区間検出を行った。発話区間の結合手法は以下の通りである。

- 手法 1 : 前後の発話の i-vector のコサイン類似度と発話の間隔情報を考慮して発話区間の結合する
- 手法 2 : 前後の発話の i-vector のコサイン類似度と発話環境を考慮して発話区間の結合する
- 手法 3 : 手法 1 + 手法 2

また、Baseline として、音源識別によって得られた各発話区間から抽出した i-vector を用いてニュースアンカーの発話区間検出を行う。

前後の発話区間を結合する際の i-vector のコサイン類似度の閾値を表 4.4 に示す。ここで、T は発話区間の秒数である。これは、図 3.1 で示されるように、同一話者間の発話であっても発話の長さによってコサイン類似度の値が大きく異なり、図 3.2 で示されるように発話が短い時、同一話者間の i-vector のコサイン類似度の標準偏差が非常に大きいためである。そのため、本実験では、3.5 秒と 7 秒で閾値を変更し、発話区間の結合に用いた。

表 4.4: 発話区間の結合の閾値

時間条件	コサイン類似度の閾値
$T < 3.5$	0.2
$3.5 \leqq T < 7$	0.6
$7 \leqq T$	0.75

手法 1、および手法 3 で用いる非発話区間の長さの閾値 Th_{time} は 0.8 秒から 1.5 秒までの範囲を 0.1 秒刻みで行う。これは、図 3.5 で示されているように、同一話者間の非発話区間の長さが約 1 秒以下の割合が大きく、図 3.6 より、異なる話者間の非発話区間の長さが 2 秒前後の割合が高いためである。

i-vector を用いたニュースアンカーの発話区間検出におけるニュースアンカーか否かを判別する Th_{cos} は、先行研究 [3] と同様に 0.5 から 0.8 までの範囲を 0.1 刻みで変更して実験を行う。

4.5　評価方法

評価は、検出されたニュースアンカーの発話区間と正解ラベルを比較して行う。

表 4.5: ニュース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	
		ニュースアンカーの発話区間	ニュースアンカー以外の発話区間
判定結果	正	TP	FP
	誤	FN	TN

表 4.5 に示すニュースアンカーの発話区間の正誤判定を行い、P（適合率（Precision））と R（再現率（Recall））を式 4.1 と式 4.2 のようにそれぞれ計算する。

$$P = \frac{TP}{TP + FP} \tag{4.1}$$

$$R = \frac{TP}{TP + FN} \tag{4.2}$$

ここで P と R はそれぞれ適合率、再現率を表す。適合率が高い値を取るとき、識別結果に含まれる「誤り」の割合が少ないことを示している。また再現率が高いとき、識別結果に「漏れ」が少ないことを示している。一般的に、再現率の高いシステムは適合率が低く、逆に適合率が高いシステムは再現率が低い傾向にある。評価指標が 2 つあるとどちらのシステムが優れているかの判断が難しいため、適合率と再現率の調和平均を取り、ひとつのスカラ値に変換した F 値（F-measure）がある。

$$F = \frac{2 \times P \times R}{P + R} \tag{4.3}$$

また、検出したニュースアンカーの発話区間の時間の割合を式 4.4 を用いて評価する。

$$Acc_{time} = \frac{\text{検出したニュースアンカーの発話の時間数}}{\text{ニュースアンカーの発話の時間数}} \tag{4.4}$$

本実験では、評価指標として適合率、再現率、F 値、Acc_{time} を用いる。

4.6　実験結果

Th_{time} を変更してニュースアンカーの発話検出精度が最も高い F 値を示した条件の結果を図 4.2 ～ 図 4.5 に示す。その他の条件の結果は付録 B で記載する。また、図 4.2 ～ 図 4.5 に示された結果の中で、最も高い F 値をとった Th_{cos} のときの各ニュース番組のニュースアンカーの発話検出精度とニュースアンカーとして検出したクラスタ数を表 4.6 ～ 表 4.9 に示す。

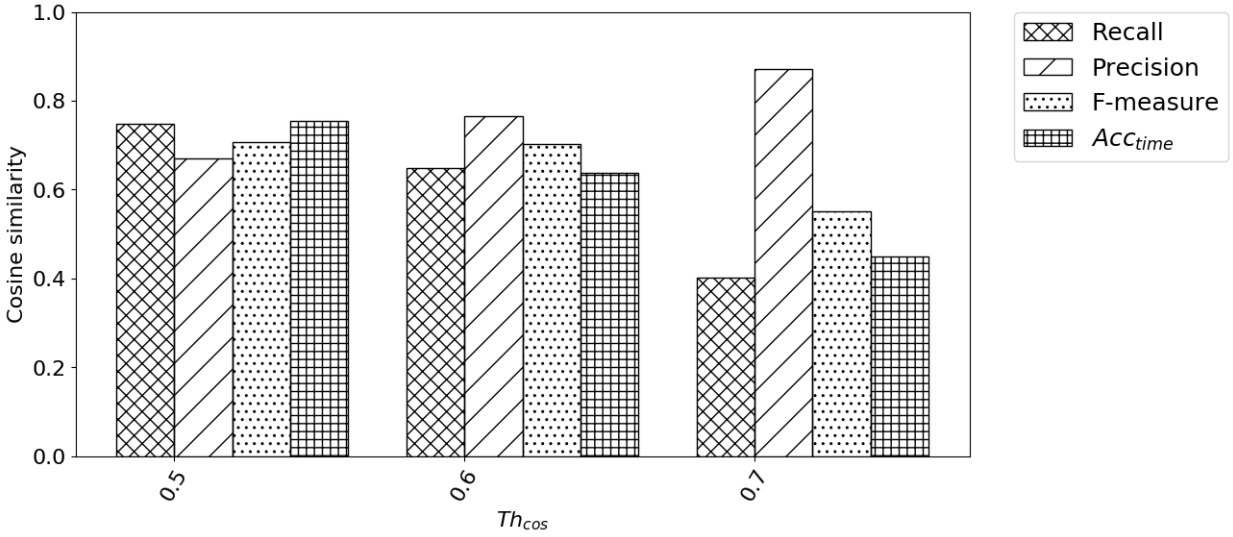

図 4.2: Baseline による発話区間検出精度

表 4.6: Baseline による各ニュース番組音声のニュースアンカーの発話検出精度 ($Th_{cos} = 0.5$)

データ ID	Recall	Precision	F-meature	作成したクラスタ数
ニュース 1	0.970	0.623	0.758	1
ニュース 2	0.709	0.437	0.540	2
ニュース 3	0.736	0.719	0.727	2
ニュース 4	0.728	0.661	0.693	2
ニュース 5	0.683	0.947	0.793	2

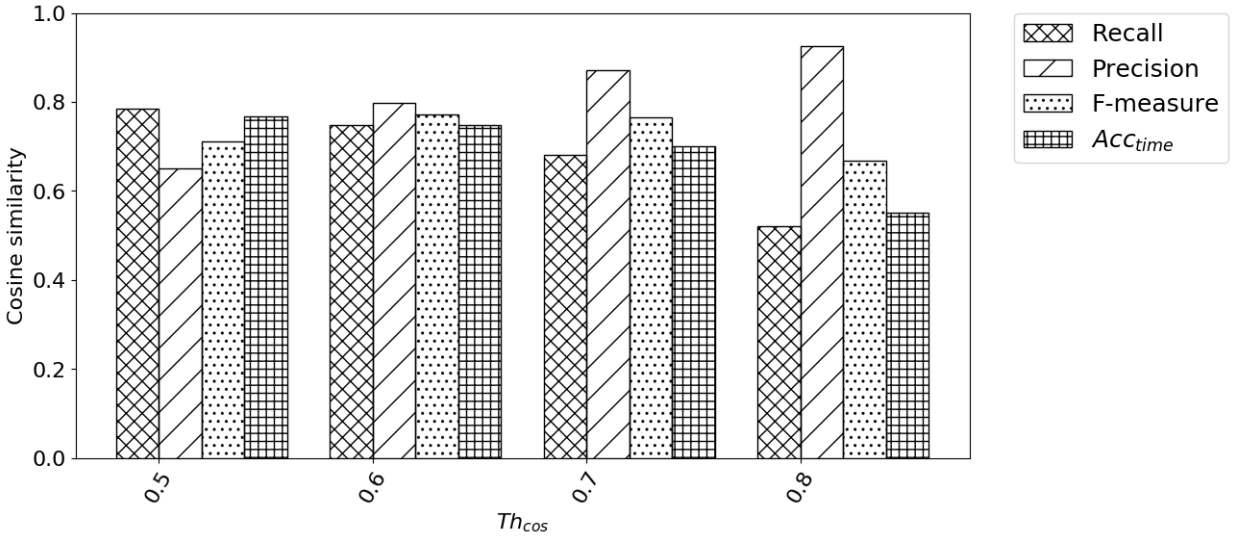

図 4.3: 手法 1 によるアンカーの発話区間検出精度 ($Th_{time} = 1.2$)

表 4.7: 手法 1 による各ニュース番組音声のニュースアンカーの発話検出精度 ($Th_{cos} = 0.6, Th_{time} = 1.2$)

データ ID	Recall	Precision	F-meature	作成したクラスタ数
ニュース 1	0.964	0.707	0.815	1
ニュース 2	0.764	0.685	0.722	2
ニュース 3	0.729	0.860	0.789	2
ニュース 4	0.683	0.741	0.711	2
ニュース 5	0.695	0.978	0.813	2

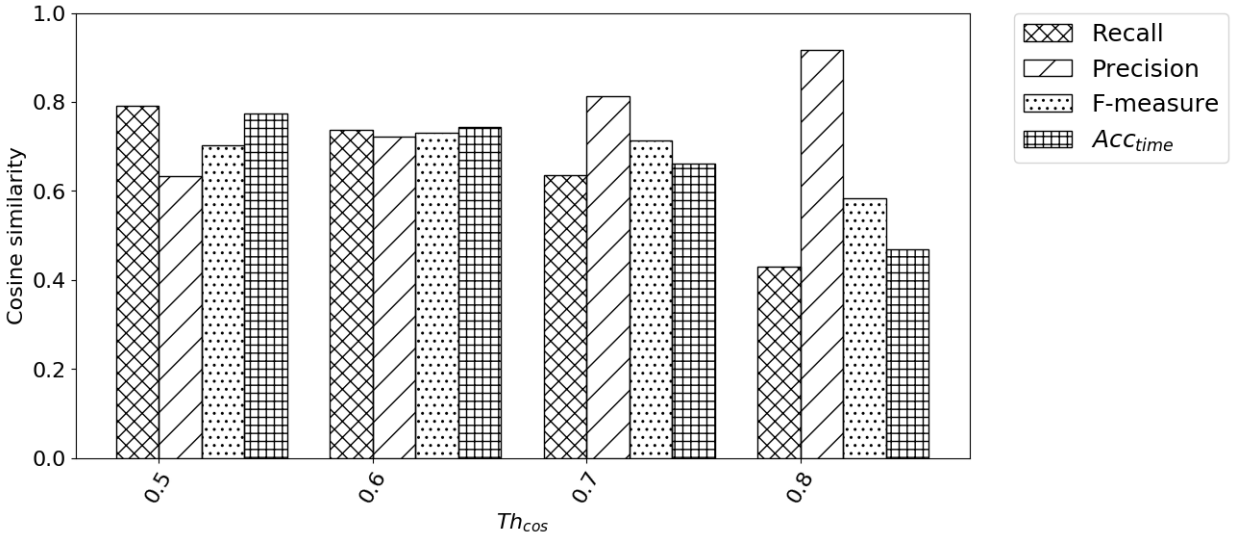

図 4.4: 手法 2 によるニュースアンカーの発話検出精度

表 4.8: 手法 2 による各ニュース番組音声のニュースアンカーの発話検出精度 ($Th_{cos} = 0.6$)

データ ID	Recall	Precision	F-meature	作成したクラスタ数
ニュース 1	0.958	0.617	0.751	1
ニュース 2	0.788	0.512	0.621	2
ニュース 3	0.695	0.823	0.754	2
ニュース 4	0.683	0.732	0.707	2
ニュース 5	0.679	0.960	0.796	2

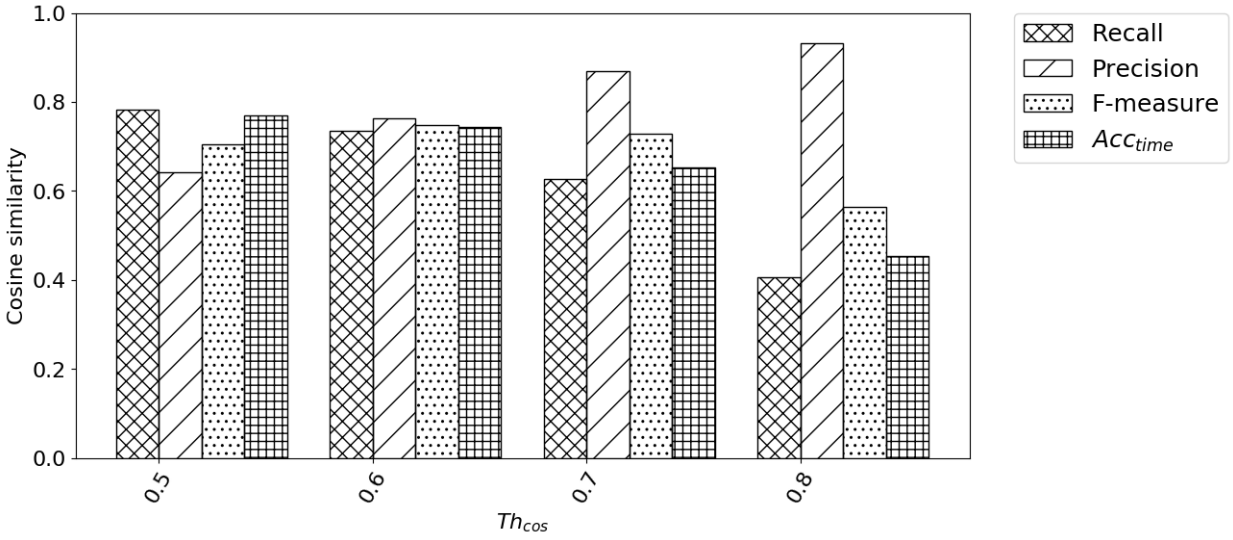

図 4.5: 手法 3 によるニュースアンカーの発話検出精度　($Th_{time} = 1.3$)

表 4.9: 手法 3 による各ニュース番組音声のニュースアンカーの発話検出精度 ($Th_{cos} = 0.6, Th_{time} = 1.3$)

データ ID	Recall	Precision	F-meature	作成したクラスタ数
ニュース 1	0.958	0.702	0.810	1
ニュース 2	0.758	0.590	0.663	2
ニュース 3	0.725	0.846	0.781	2
ニュース 4	0.646	0.755	0.696	2
ニュース 5	0.683	0.973	0.802	2

実験の結果、発話区間を結合して再抽出した i-vector を用いた手法が全体的に高い精度を示した。Baseline は Th_{cos} が 0.8 のときニュースアンカーの発話区間を検出できなかった。これは、同一話者間の発話においてコサイン類似度が 0.8 を上回る i-vector が少なく、ニュースアンカーと考えられるクラスタを作成できなかったためである。また、Baseline は Th_{cos} が 0.5 のときが F-measure が最も高い値をとるのに対して、手法 1 ～ 手法 3 では Th_{cos} が 0.6 のとき、F-measure が最も高い値をとった。

本実験の提案手法では、手法 1 が最もニュースアンカーの発話検出精度が高く、F 値が 0.772 であった。また、いずれの手法においても Th_{cos} が小さい時には Recall が高く、大きい時には Precision が高くなる傾向が確認された。

手法 1 を除く全ての手法においてニュース 2 のニュースアンカーの発話検出精度が最も低い。

4.7　考察

Baseline と提案手法によって結合した発話区間から抽出した i-vector を用いた手法を比較したとき提案手法のほうがニュースアンカーの発話検出精度が最大で約 6%向上したことから、従来と比較して i-vector が話者の特徴をより正確に抽出できたと考えられる。また、いずれの手法でもニュース 2 はニュースアンカーの発話検出の精度が他のニュースと比較して低い。これは、表 4.2 や表 4.3 で示されているように、アンカーの発話の割合が少ないためである可能性が高い。これによって、インタビューイや天気アナウンサーなどのニュースアンカー以外の発話の割合が高くなり、ニュースアンカーの発話を検出することが困難であったと考えられる。

提案手法の中では手法 1 が最も高い F 値を示した。これは手法 2、手法 3 において、ニュースアンカーが発話中に参考映像などの背景雑音、音楽が鳴った場合、発話環境の変化と誤認識してしまい、話者の切り替わりと判別して発話区間の結合ができなかったためであると考えられる。

また、結合した発話区間から抽出された i-vector を用いて、ニュースアンカーの発話区間が未知の場合と既知の場合でニュースアンカーの発話の音声認識を行った。音声認識手法として i-vector と木構造話者クラスタを用いた音声認識手法 [15] を用いた。音声認識システムの概要、実験の詳細は付録 C に記載する。実験の結果、アンカーの発話区間が未知であった場合、Baseline と比較してニュースアンカーの発話の音声認識精度が向上した。これは、ニュースアンカー以外の発話を音声認識した場合、全て誤りと計算したためである。ニュースアンカーの発話区間が既知であった場合、提案手法による音声認識精度への効果が確認されなかった。これは、BGM や背景雑音が発話に含まれていた場合、いずれの提案手法を用いた場合でも認識できなかったためである。しかしいずれの手法にしてもニュース番組音声の音声認識精度が非常に低いため、音声認識精度向上のために、雑音除去や雑音に頑健な音声認識システムが必要である。

第5章

結論

本稿では、ニュース番組音声のインデクス自動付与に向けたダイアライゼーション実現のために、発話間隔と発話環境を考慮した i-vector を用いたニュースアンカーの発話検出精度向上を目指した。

特定話者の識別には i-vector が一般的に用いられるが、短い発話からは話者の識別に必要な十分な話者の特徴を抽出できない。そのため、本稿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し、長い発話を擬似的に作成した。次に、結合した発話区間から i-vector を抽出することで短い発話から得られる i-vector の抽出精度向上を目指した。発話区間の結合には、同一話者が連続で発話する場合間をおかずに発話すること、話者が切り替わった時に発話環境が変化することに着目し、発話と発話の時間間隔を考慮する手法と、発話者の発話環境を考慮する手法を用いた。以上の手法を用いて結合した発話区間から抽出した i-vector を用いてニュースアンカーの発話検出を行った結果、発話検出精度が約 6%向上し、ニュースアンカーの発話検出への有意性を示した。しかし、ニュースアンカーの発話数が少ないニュース番組においてはアンカーの検出精度が著しく低下した。このため、ニュースアンカーの発話が少ない場合においても発話を検出する手法を提案する必要がある。

また、本研究で抽出した i-vector を用いてニュースアンカーの音声認識を行なった。音声認識はニュースアンカーの発話区間が既知の場合と未知の場合で行い、発話区間が既知のときは音声認識精度の向上が確認できなかった。ニュースアンカーの発話区間が未知の場合、従来と比較してニュースアンカーの発話区間検出精度が向上したことが音声認識精度の向上に繋がった。

今後の課題として、ニュースアンカーの発話が少ない番組におけるニュースアンカーの発話検出精度の向上が必要である。これは、発話内容など i-vector 以外のニュースアンカーの特徴を考慮する必要があると考えられる。また、音声認識においては、ニュース音声の背景雑音の除去、雑音に頑健な音響モデルの作成が挙げられる。ニュースアンカーはスタジオで発話しているため基本的には音声以外は入らないが、参考映像などの音が発話中に流れることがある。そのため、雑音、音楽に対して音源分離による雑音除去や、雑音や音楽が含まれた学習データを用いて音響モデルを学習することで、音声認識精度が向上すると考えられる。

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，高田 寛之助教, 山下 優博士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] 小川哲司, 塩田さやか,”i-vector を用いた話者認識”, 日本音響学会誌 70(6), 332-339(2014)

[2] 西史人, 井上中順, 篠田浩一,”マルチモーダル i-vector を用いた話者ダイアライゼーション”，情報処理（2015）

[3] 野崎大智,”ニュース音声における i-vector を用いた同一アンカーの発話群の検出”, 電気情報通信学会九州支部学生会講演会 (2018)

[4] 安達大輔,”ニュース番組における発話者群の段階的分類の検討”, 電気情報通信学会九州支部学生会講演会 (2012)

[5] 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 (2015)

[6] 奥，佐藤，小林，本間，今井,“ マルチ音素クラスのベイズ情報量基準に基づくオンライン話者ダイアライゼーション ”, 信学論，Vol.J95-D，No.9，pp.1749-1758（2012）

[7] 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”

[8] ”ALIZE”,http:/alize.univ-avignon.fr

[9] 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)

[10] N．Dehak，P．Kenny，R．Dehak，P．Dumouchel and P．Ouellet，”Front-end factoranalysis forspeaker verifiCatiOn” IEEE Trans. Audio Speech Lang. Process, 19, 788-798(2011)

[11] 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)

[12] :鹿野清宏, 伊藤克亘, 河原達也, 武田一哉, 山本幹雄,”音声認識システム”, 情報処理学会, オーム社 (2003)

[13] 新美康永,“ 音声認識 ”, 共立出版株式会社 (1979)

[14] 河原達也,”音声認識システム”, 情報処理学会 (2016)

[15] 吉村竜哉,”話者クラスタ音響モデルを用いた会議音声認識のための話者適応”, 電気情報通信学会九州支部学生会講演会 (2014)

[16] 小島和也,”会議音声認識のための DNN を用いた高精度な音響モデルの構築法の検討”, 長崎大学工学部情報システム工学科 平成 25 年度修士論文 (2013)

[17] ”KALDI”,http://kaldi.sourceforge.net/

付録 A

音源分離実験

A.1　音源分離システムの概要

音響データ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。

本システムでは、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 [9] 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、喧騒

(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 8 混合のガウス分布を用いている。本研究の音響特徴パラメータを表 A.1 に示す。

表 A.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	

以降、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。

A.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 の窓関数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。

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

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 $s_\omega(n)$　$(n = 0, 1, \cdots, N-1)$ 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。また、窓関数とは、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。

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

$$\text{ハミング窓} : \omega(n) = 0.54 - 0.54\cos(\frac{2\pi n}{N-1}) \qquad (n = 0, 1, \cdots, N-1) \tag{A.3}$$

(2) 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。

$$S(n) = \sum s_\omega(n)e^{-j2\pi\frac{nk}{N}} \qquad (k = 0, 1, \cdots, N-1) \tag{A.4}$$

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。

$$S(n) = \frac{1}{N}\sum S(k)e^{j2\pi\frac{nk}{N}} \qquad (k = 0, 1, \cdots, N-1) \tag{A.5}$$

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 (A.2),(A.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、

$$S\prime(n) = S(e^{j\frac{2\pi}{n}k}) = \sum s_\omega(n)e^{-j2\pi\frac{2\pi}{N}kn} \qquad (k = 0, 1, \cdots, N-1) \tag{A.6}$$

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。

(3) パワースペクトルの算出: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 (A.7) で表される。

$$|S'(k)|^2 = \frac{1}{N}[\mathrm{Re}\{S'(k)\}^2 + \mathrm{Im}\{S'(k)\}^2] \tag{A.7}$$

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ログラムという。スペクトロ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。

A.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 [11]

(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ため、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 [13]。

$$E(l) = \sum_{n=0}^{N-1} \{x_l(n)\}^2 \tag{A.8}$$

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。

A.2　調査方法

本調査では、3.2 節の調査で用いたニュース番組の音声データ 12 個を用いて音源分離技術による音源分離を行う。検出する区間は「音声」「背景雑音」「音楽」「無音」の 4 つである。表 A.2 は音源識別の調査条件である。

表 A.2: 音源識別実験の実験条件

FFT の窓幅 (フレーム長)	2048point(約 0.046[sec])
FFT のシフト幅 (フレーム間隔)	1024point(約 0.023[sec])
窓関数	ハミング窓

A.2.1　評価方法

評価は、検出された各区間と正解ラベルを比較して行う。

表 A.3: 検出した区間の正誤判定

		正解ラベル	
		ラベルが付与された区間	ラベルが付与されていない区間
判定結果	正	TP	FP
	誤	FN	TN

表 A.3 が得られると P（適合率（Precision））と R（再現率（Recall））、F 値 (適合率と再現率の調和平均) は式 A.9 と式 A.10、式 A.11 のようにそれぞれ計算する。

$$P = \frac{TP}{TP + FP} \tag{A.9}$$

$$R = \frac{TP}{TP + FN} \tag{A.10}$$

$$F = \frac{2 \times P \times R}{P + R} \tag{A.11}$$

本実験では、評価指標として適合率、再現率、F 値を用いる。

A.2.2　調査結果

表 A.4 ～ 表 A.7 に音源識別による識別精度を示す。

表 A.4: 音声区間検出精度

データ ID	Recall	Precision	F-meature
ニュース A	0.892	0.966	0.928
ニュース B	0.888	0.963	0.924
ニュース C	0.883	0.963	0.921
ニュース D	0.902	0.952	0.927
ニュース E	0.884	0.970	0.925
ニュース F	0.907	0.974	0.939
ニュース G	0.907	0.961	0.933
ニュース H	0.843	0.966	0.900
ニュース I	0.886	0.982	0.932
ニュース J	0.902	0.980	0.939
ニュース K	0.875	0.963	0.917
ニュース L	0.886	0.963	0.923
平均	0.888	0.967	0.926

表 A.5: 音楽区間検出精度

データ ID	Recall	Precision	F-meature
ニュース A	0.467	0.565	0.511
ニュース B	0.508	0.640	0.566
ニュース C	0.507	0.687	0.583
ニュース D	0.429	0.661	0.520
ニュース E	0.481	0.633	0.547
ニュース F	0.627	0.699	0.661
ニュース G	0.611	0.936	0.740
ニュース H	0.570	0.406	0.474
ニュース I	0.481	0.648	0.552
ニュース J	0.531	0.776	0.631
ニュース K	0.718	0.381	0.498
ニュース L	0.672	0.471	0.554
平均	0.537	0.622	0.576

表 A.6: 背景雑音区間検出精度

データ ID	Recall	Precision	F-meature
ニュース A	0.259	0.835	0.395
ニュース B	0.406	0.681	0.509
ニュース C	0.199	0.857	0.323
ニュース D	0.225	0.678	0.338
ニュース E	0.282	0.783	0.414
ニュース F	0.145	0.587	0.233
ニュース G	0.192	0.855	0.313
ニュース H	0.235	0.803	0.364
ニュース I	0.338	0.817	0.478
ニュース J	0.268	0.746	0.395
ニュース K	0.268	0.906	0.413
ニュース L	0.349	0.511	0.415
平均	0.263	0.756	0.390

表 A.7: 無音区間検出精度

データ ID	Recall	Precision	F-meature
ニュース A	0.883	0.659	0.755
ニュース B	0.334	0.685	0.449
ニュース C	0.923	0.669	0.776
ニュース D	0.581	0.587	0.584
ニュース E	0.807	0.693	0.745
ニュース F	0.859	0.564	0.681
ニュース G	0.934	0.659	0.773
ニュース H	0.788	0.626	0.698
ニュース I	0.907	0.708	0.795
ニュース J	0.763	0.645	0.699
ニュース K	0.887	0.615	0.726
ニュース L	0.602	0.702	0.648
平均	0.787	0.649	0.712

表 A.4 より、音声区間の検出精度の F 値の平均が 0.926 で、他の区間と比較して最も精度よく検出できた。対して、表 A.6 より、背景雑音区間の検出精度は F 値の平均が 0.390 で、次に検出精度が低い音楽区間と比較しても F 値の平均に 0.186 の差が生じた。また、背景雑音区間の検出において、Recall は 0.263 で全ての音源識別精度で最も精度が低かったが、Precision に関しては 0.756 であり、これは音声区間に次いで 2 番目に精度が高かった。

A.3　考察

背景雑音区間の検出精度が音声区間の区間検出精度と比較して大きく下がっている理由として、背景雑音が音声区間と同時に存在していることが多いためである。背景雑音は街頭インタビュー中やアンカーの発話中に流れる映像に多く含まれる。このように、音声と背景雑音が同時に収録されていた場合、基本的に音声は背景雑音や音楽と比較して大きく編集されていることが多い。つまり、音声と背景雑音が同時に収録されている場合、音声区間として検出される可能性が高いため、背景雑音区間の検出精度が低下したと考えられる。つまり、背景雑音区間の検出精度において P 値が非常に高いことから、背景雑音のみが収録されていた場合は高い精度で検出できていることがわかる。しかし、音源が複数同時に収録されている場合は検出が出来ないため、検出精度の向上のためには、複数音源の検出を同時に行うシステムを構築するか、音源分離を行う必要がある。

付録B

アンカーの発話検出精度

本文で記載した条件以外のアンカーの発話検出精度の詳細を以降に記載する。

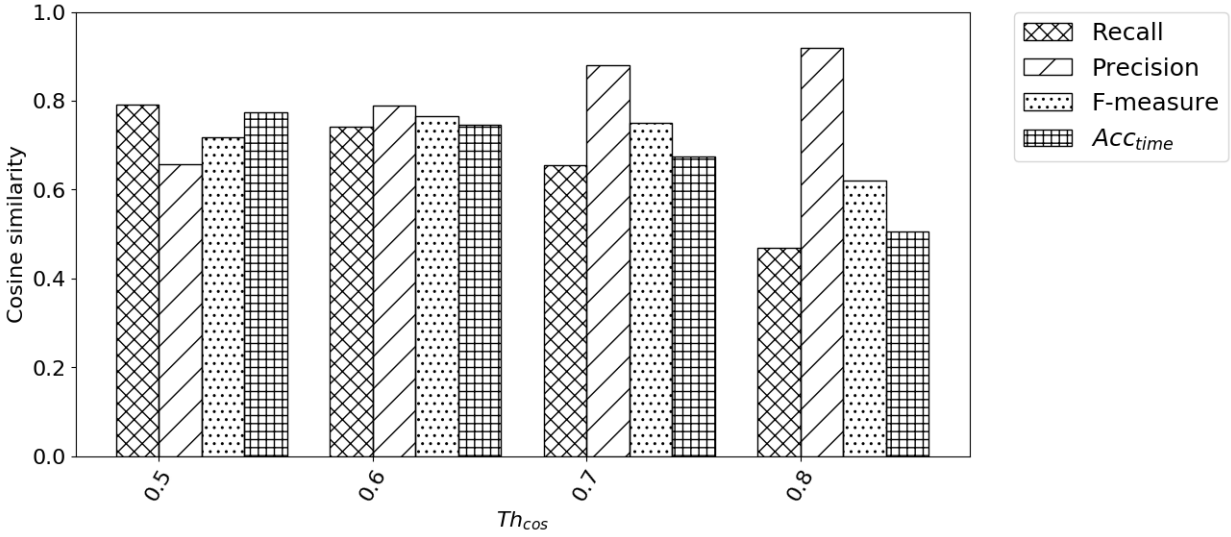

図 B.1: 手法 1 によるアンカーの発話区間検出精度 ($Th_{time} = 0.8$)

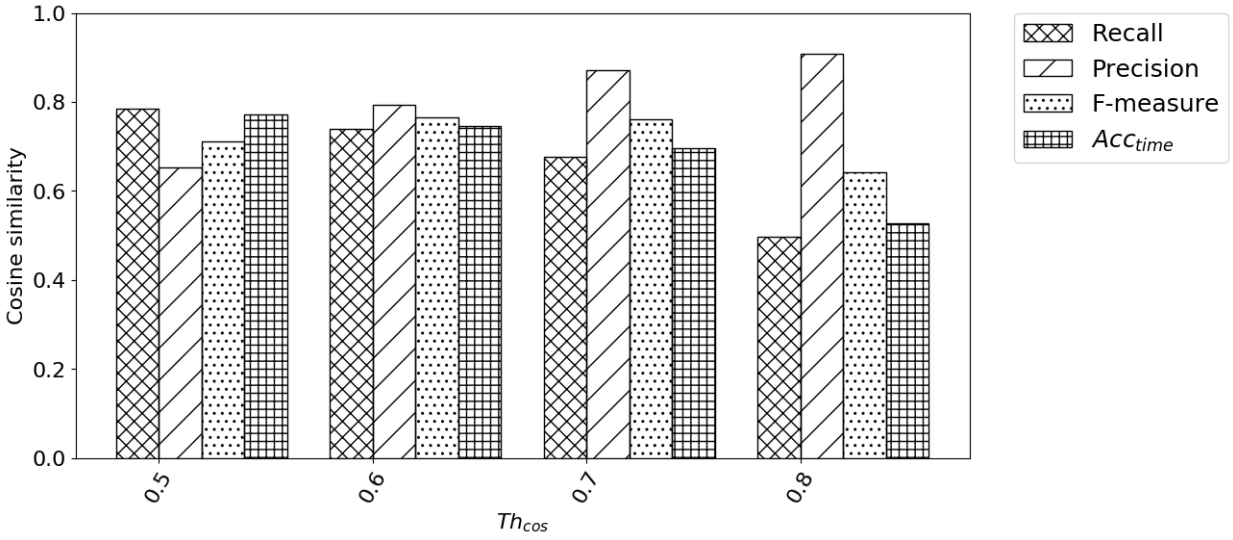

図 B.2: 手法 1 によるアンカーの発話区間検出精度 ($Th_{time} = 0.9$)

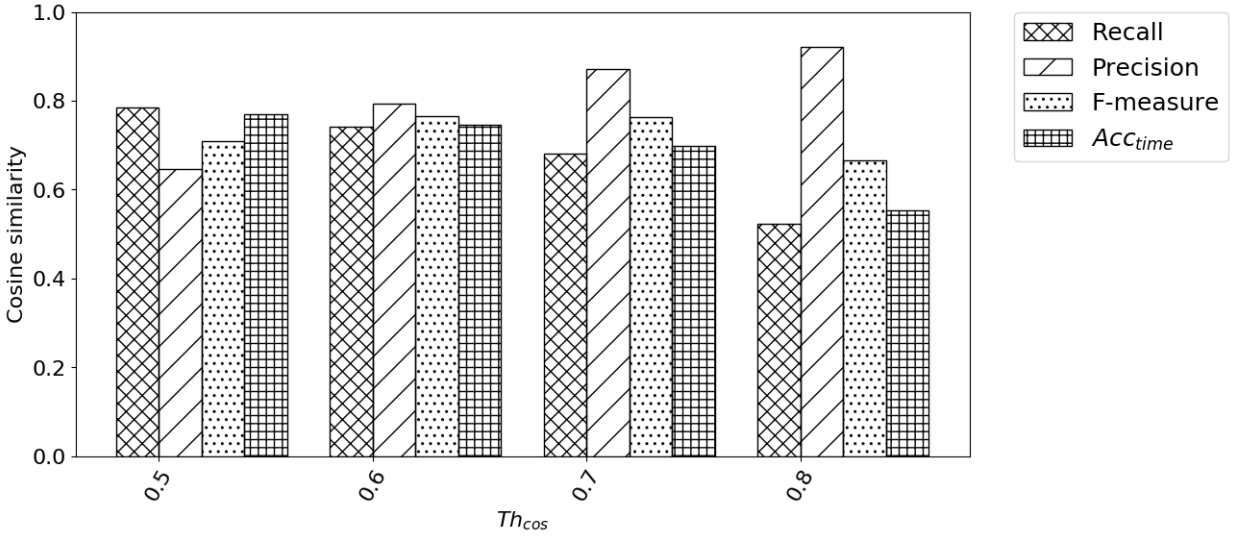

図 B.3: 手法 1 によるアンカーの発話区間検出精度 ($Th_{time} = 1.0$)

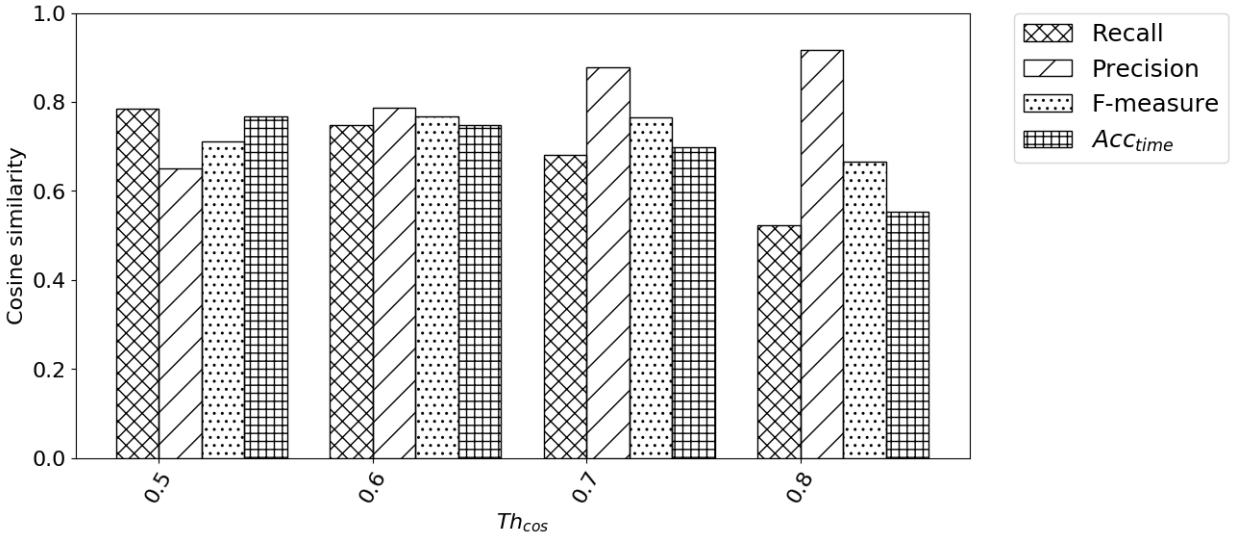

図 B.4: 手法 1 によるアンカーの発話区間検出精度 ($Th_{time} = 1.1$)

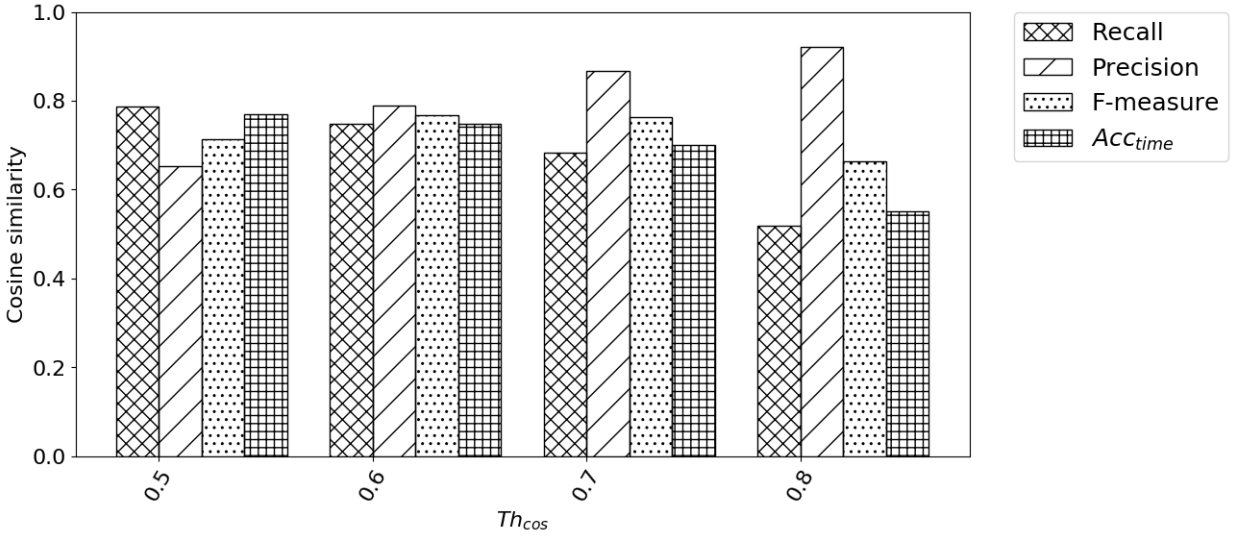

図 B.5: 手法 1 によるアンカーの発話区間検出精度 ($Th_{time} = 1.3$)

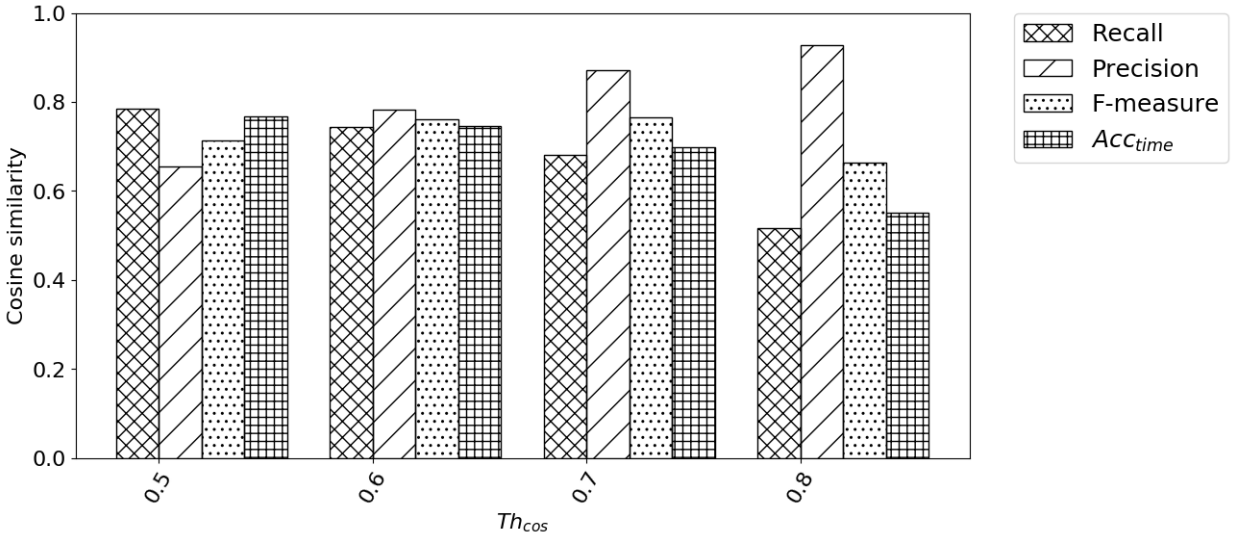

図 B.6: 手法 1 によるアンカーの発話区間検出精度 ($Th_{time} = 1.4$)

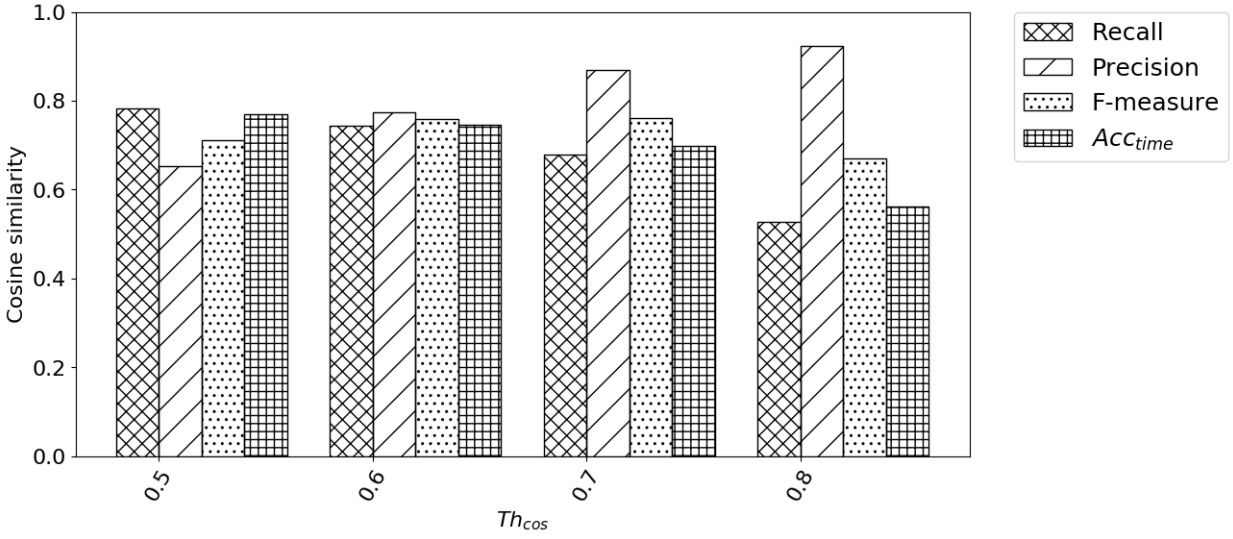

図 B.7: 手法 1 によるアンカーの発話区間検出精度 ($Th_{time} = 1.5$)

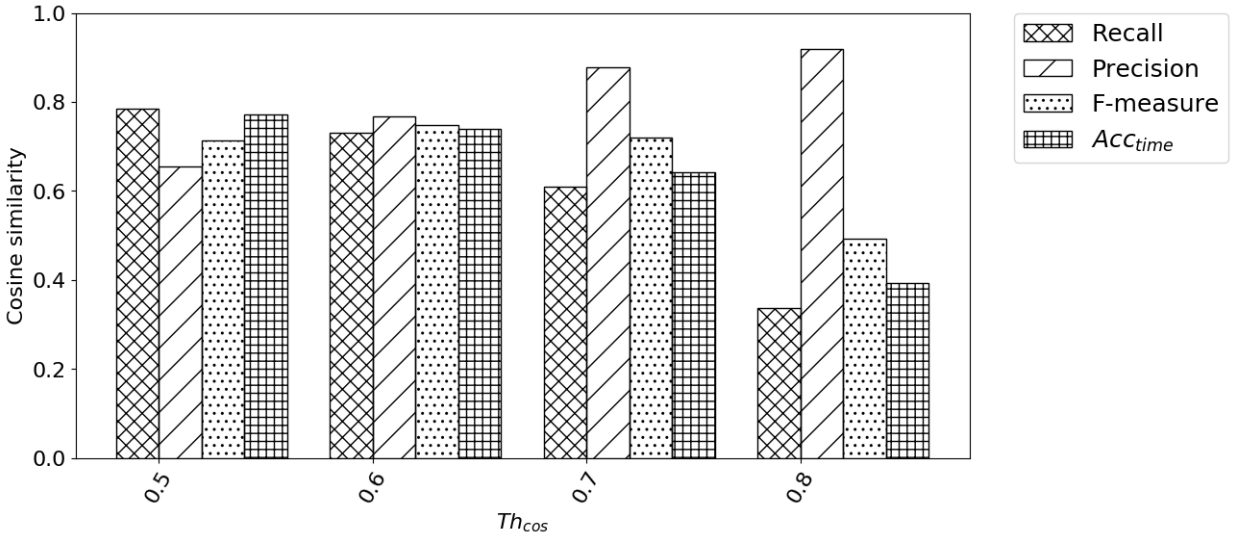

図 B.8: 手法 3 によるアンカーの発話区間検出精度 ($Th_{time} = 0.8$)

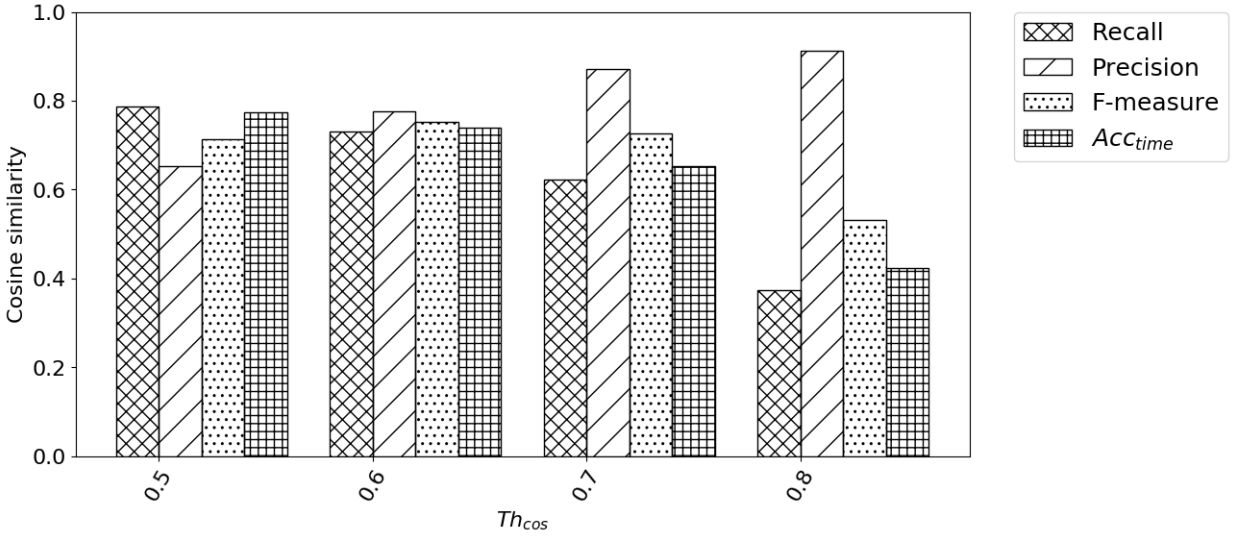

図 B.9: 手法 3 によるアンカーの発話区間検出精度 ($Th_{time} = 0.9$)

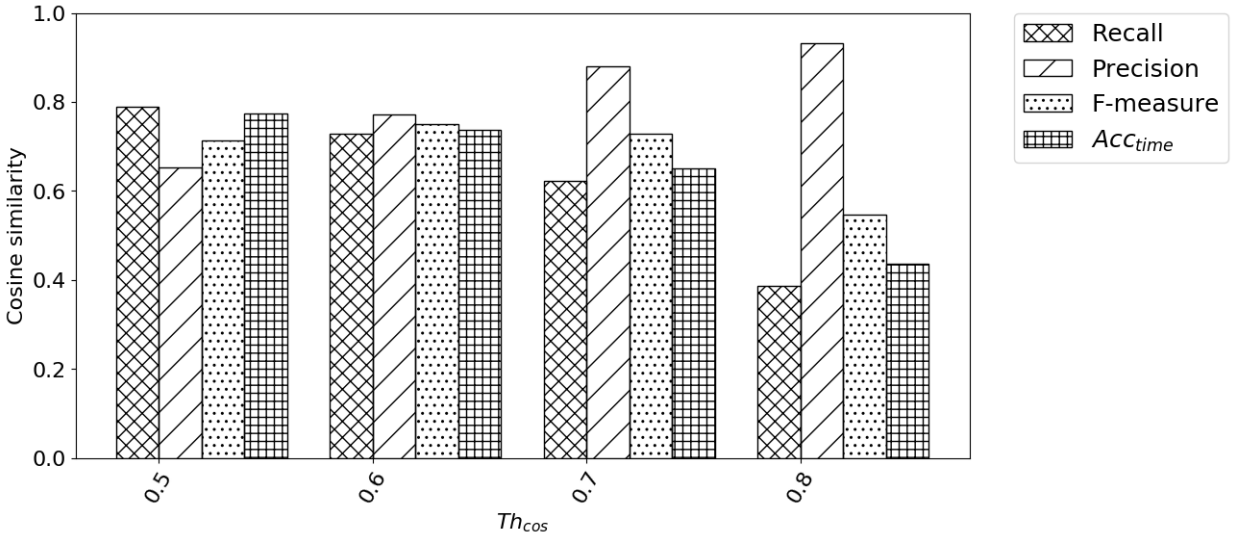

図 B.10: 手法 3 によるアンカーの発話区間検出精度 ($Th_{time} = 1.0$)

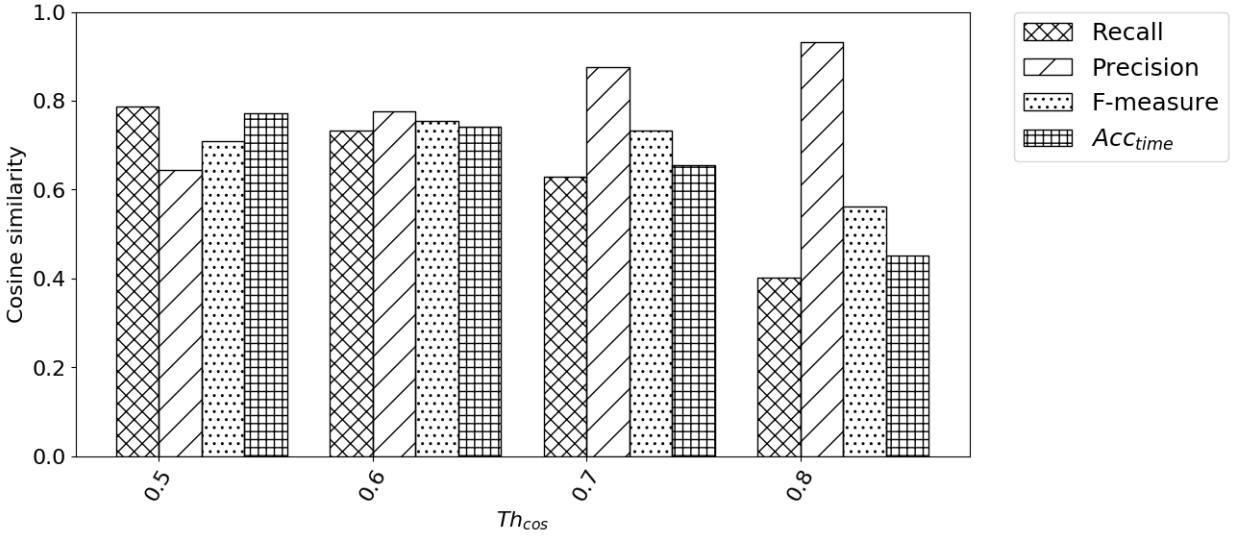

図 B.11: 手法 3 によるアンカーの発話区間検出精度 ($Th_{time} = 1.1$)

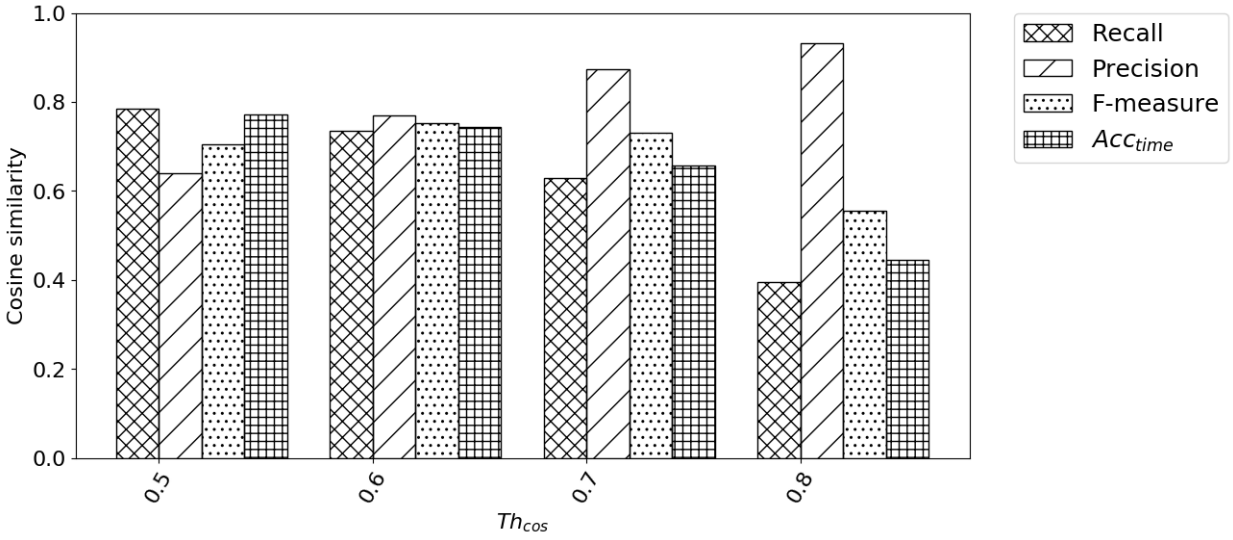

図 B.12: 手法 3 によるアンカーの発話区間検出精度 ($Th_{time} = 1.3$)

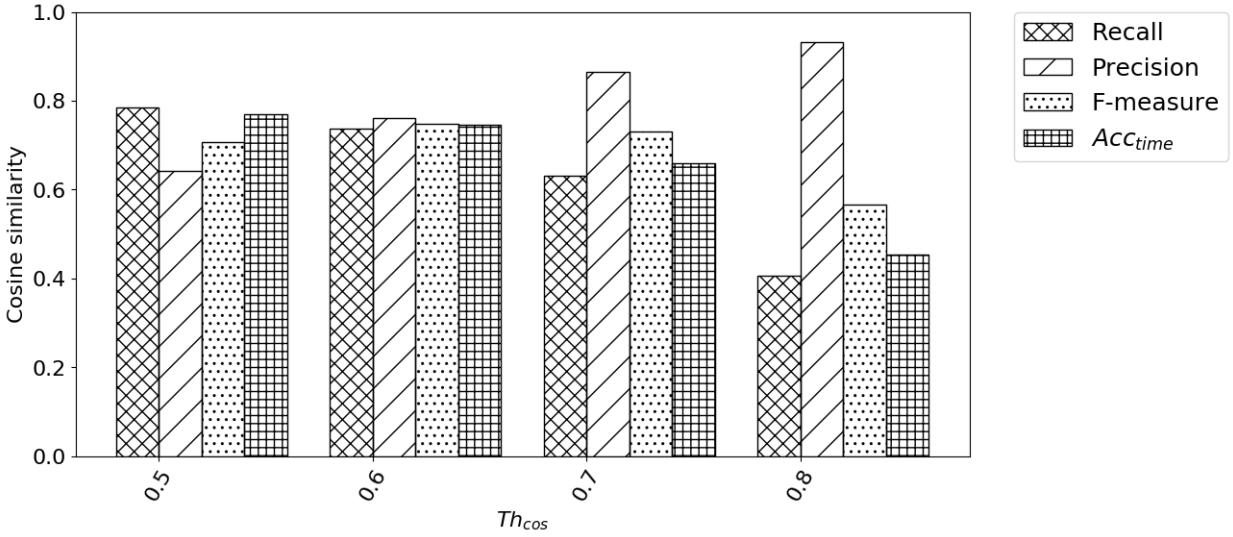

図 B.13: 手法 3 によるアンカーの発話区間検出精度 ($Th_{time} = 1.4$)

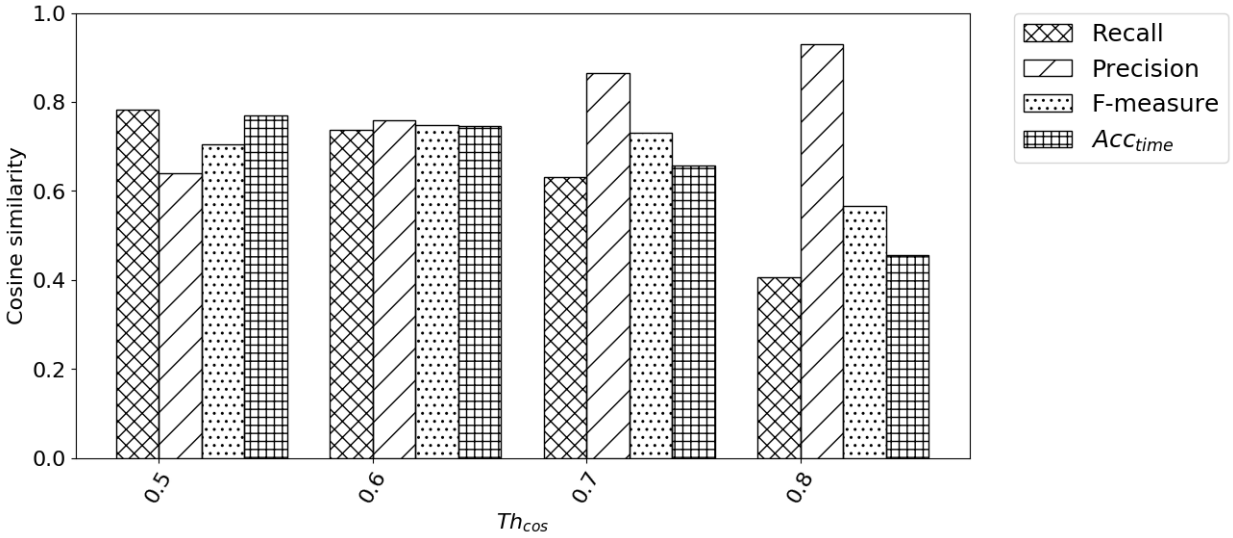

図 B.14: 手法 3 によるアンカーの発話区間検出精度 ($Th_{time} = 1.5$)

付録C

ニュースアンカーの発話の音声認識実験

C.1　音声認識システムの概要

C.1.1　音声認識の流れ

音声認識の流れを図 C.1 に示す。まず、入力された音声データから前処理として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 C.1.2 節、音響モデルについては C.1.3 節で説明する。

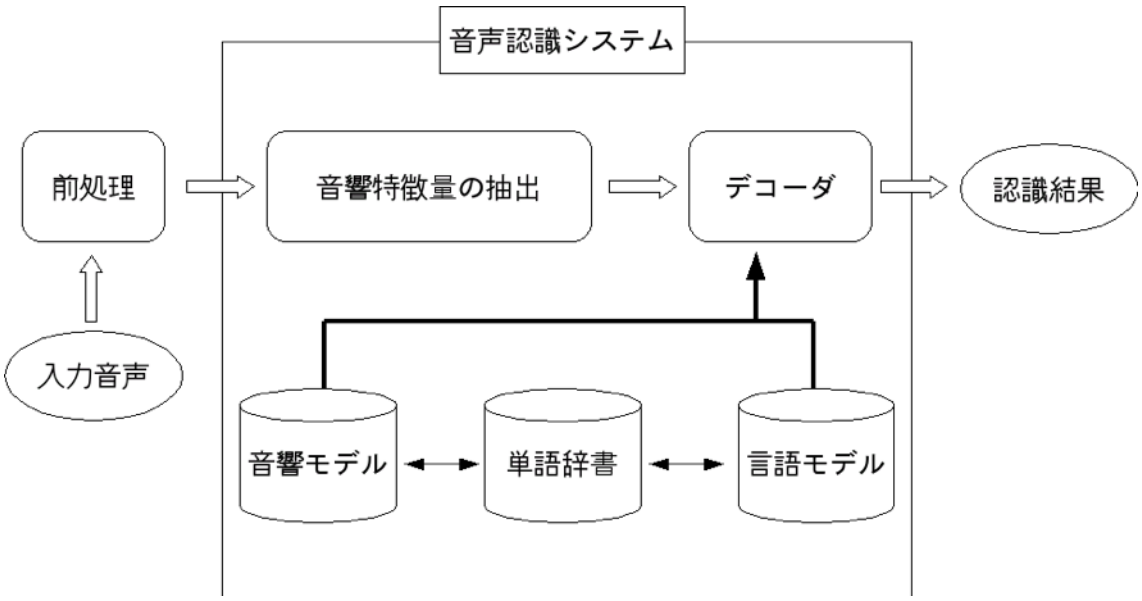

図 C.1: 音声認識の流れ

C.1.2　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 [12]。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 C.1 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 C.1: 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、

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

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。

C.1.3　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 [12]。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。

$$m(l) = \sum_{k=lo}^{hi} W(k;l) \mid S'(k) \mid \quad (l = 1, \cdots, L) \tag{C.2}$$

$$W(k;l) = \begin{cases} \frac{k-k_{lo}(l)}{k_c(l)-k_{lo}(l)} & \{k_{lo} \leq k \leq k_c(l)\} \\ \frac{k_hi(l)-k}{k_{hi}(l)-k_l(l)} & \{k_c \leq k \leq k_{hi}(l)\} \end{cases} \tag{C.3}$$

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

$$k_c = k_{hi}(l-1) = k_{lo}(l+1) \tag{C.4}$$

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

$$Mel(f) = 2595 \log_{10}(1 + \frac{f}{700}) \tag{C.5}$$

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 (C.6) のように MFCC が得られる。

$$c_{mfcc}(i) = \sqrt{\frac{2}{N}} \sum_{i=1}^{L} \log ml \cos\left\{\left(l - \frac{1}{2}\frac{i\pi}{L}\right)\right\} \tag{C.6}$$

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 [12]。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。

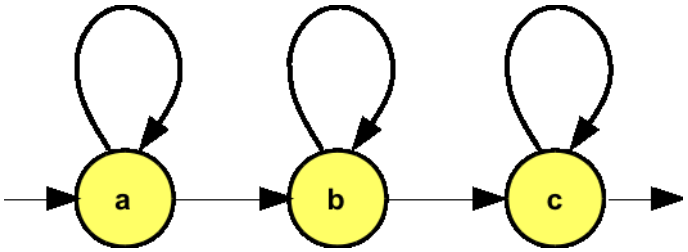

図 C.2: HMM の例

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 [12]。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 C.7 や式 C.8 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。

$$\Delta c(n;l) = \frac{\Sigma_{K=-K}^{K} k_c(n;l+k)}{\Sigma_{K}^{K=-K} K^2} \tag{C.7}$$

$$\Delta\Delta c(n;l) = \frac{\Sigma_{K=-K}^{K} k_c(n;l+k)}{\Sigma_{K}^{K=-K} K^2} \tag{C.8}$$

C.1.4　DNN の概要 [14]

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 C.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。

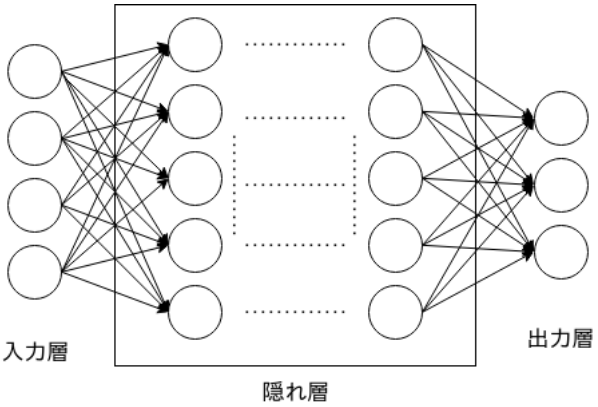

図 C.3: DNN の構造図

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。

C.1.5　モデルの構築手順

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 [17] を用いた。このツールキットの大きな流れを図 C.4 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。

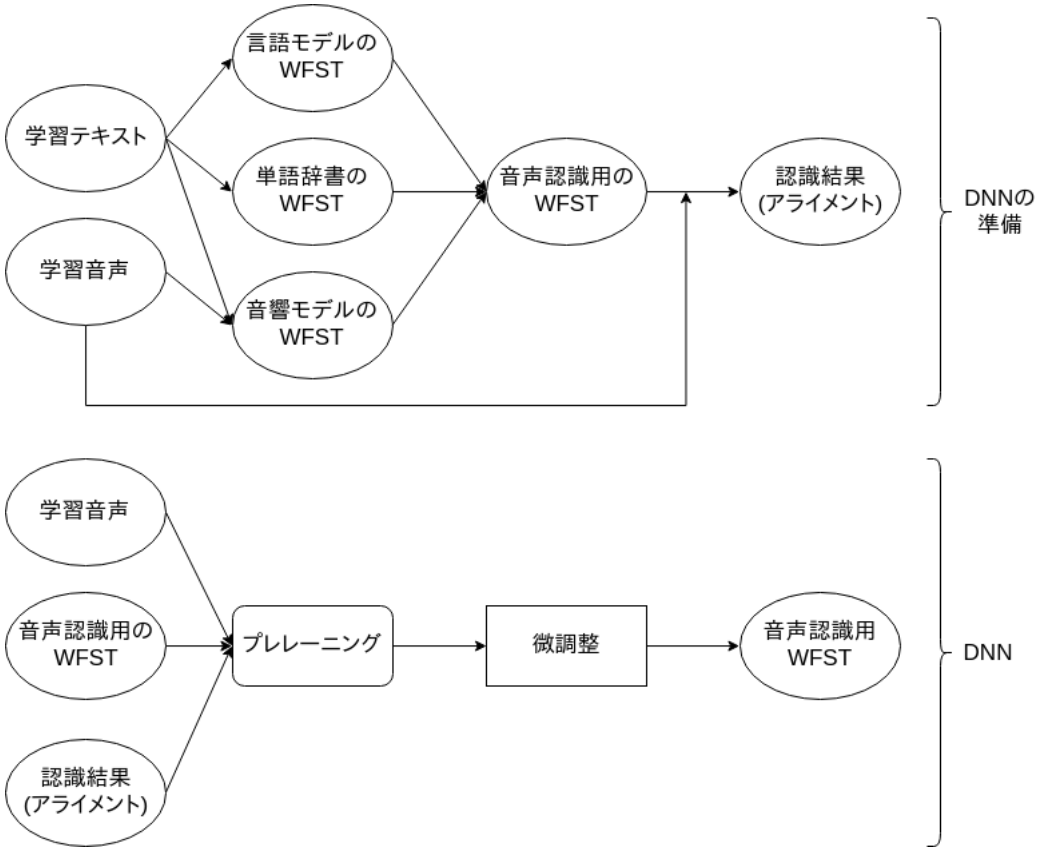

図 C.4: DNN を用いる際の学習の流れ

C.2　i-vector を用いた音声認識手法 [15]

学習データに含まれる話者の音響特徴を考慮して木構造話者クラスタを作成し、各話者クラスタに含まれる学習データを用いて音響モデルを学習した。この木構造話者クラスタは、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。

認識の際は、学習データに用いた話者の i-vector と評価データの i-vector のコサイン類似度を求める。求めたコサイン類似度の高い上位 n 人の学習データを全て含んでいるクラスタを選択し、選択したクラスタに含まれる学習データで学習した音響モデルを用いて音声認識を行なった。

C.3　実験方法

本実験では、C.2 節で述べる木構造話者クラスタを作成し、話者クラスタに含まれる学習データを用いて音響モデルを学習して音声認識実験を行う。各話者クラスタに含まれる男女の発話データ数を図 C.5 に示す。

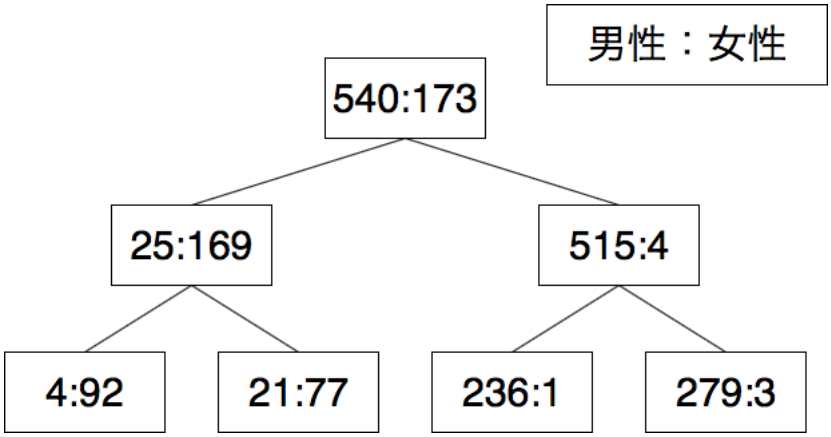

図 C.5: 各話者クラスタに含まれる発話データ数

学習データに用いた話者の i-vector と認識するアンカーの発話の i-vector のコサイン類似度を求め、求めたコサイン類似度の高い上位 5 人の学習データを全て含んでいるクラスタを選択し、選択したクラスタに含まれる学習音声で学習された音響モデルを用いて音声認識を行う。音響モデルの選択に用いる i-vector と認識するアンカーの発話区間は 4 節で検出したアンカーの発話区間のうち、各手法でもっとも F 値の高い条件のものを用いる。

本実験で使用したコーパスについては C.4 節、音響モデルの仕様は C.5 節、言語モデルと単語辞書の仕様は C.6 節で述べる。

C.4　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 C.2 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。

表 C.2: 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

C.5　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 C.3 に示す。この仕様に関しては小島らの研究 [16] 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 C.4)、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 C.3: 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 C.4: 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

使用した音素

本研究で使用した音素 27 個を表 C.5 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 C.6 に示す

表 C.5: 使用した音素

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

表 C.6: 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	#

C.6　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 C.6 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 C.6: 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。

C.7　評価方法

本研究では評価尺度としては式 C.9 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。

$$Acc = \frac{(W - S - D - I)}{W} \tag{C.9}$$

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、本研究ではニュースアンカーの発話区間が既知の場合と未知の場合で音声認識精度の評価を行う。ニュースアンカーの発話区間が未知の時、ニュースアンカーの発話検出において、ニュースアンカー以外の発話区間で検出された単語は全て挿入誤り、ニュースアンカーの発話が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。

C.8　実験結果

各手法で抽出された i-vector を元に、各手法における発話データの音響モデルの選択結果を図 C.7 ～ 図 C.10 に示す。

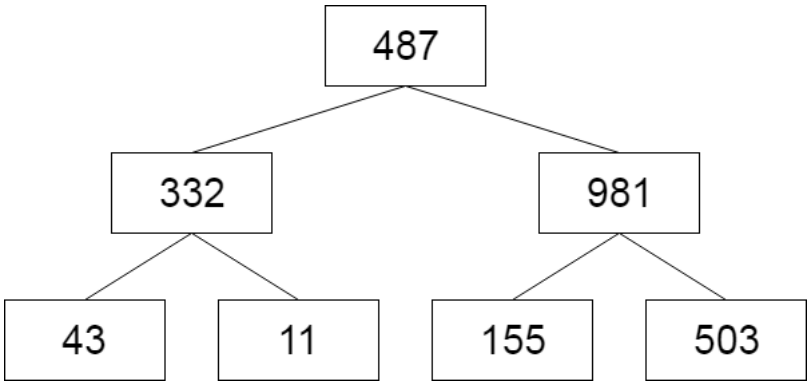

図 C.7: Baseline による音響モデルの選択結果

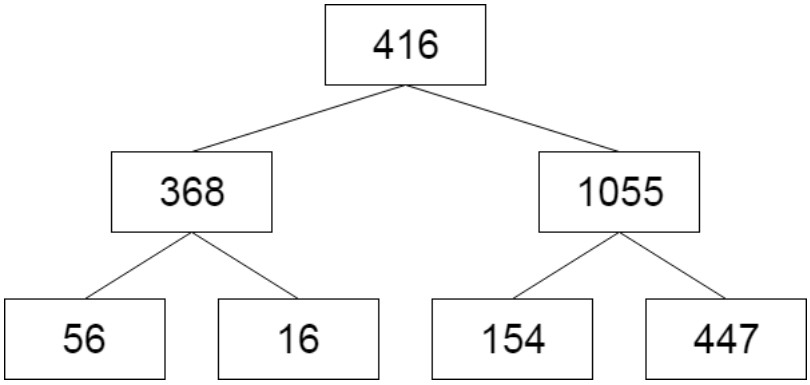

図 C.8: 手法 1 による音響モデルの選択結果

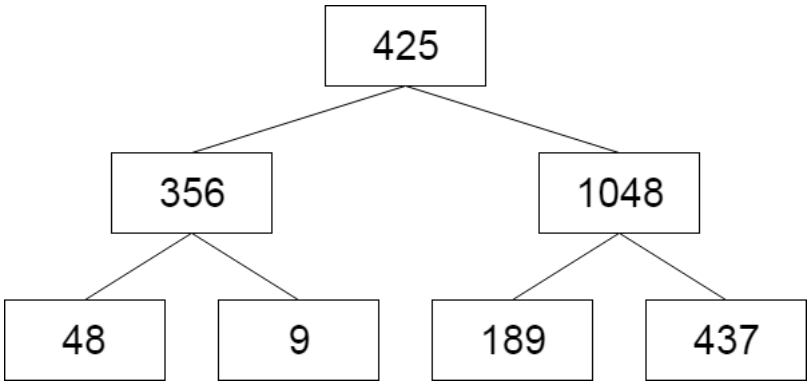

図 C.9: 手法 2 による音響モデルの選択結果

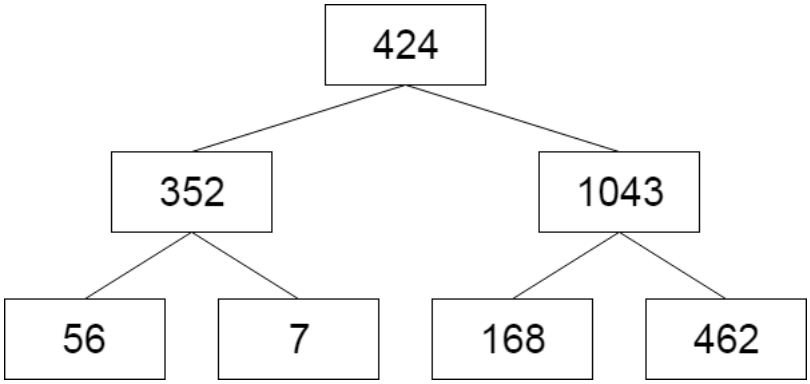

図 C.10: 手法 3 による音響モデルの選択結果

Baseline と比較して、いずれの手法も最上位の音響モデルを選択する発話データ数が減少し、下位クラスタを選択する発話データが増えている。

ニュースアンカーの発話区間が既知の場合

ニュースアンカーの発話区間が既知の場合の音声認識結果を表 C.7 に示す。

表 C.7: ニュースアンカーの発話区間が既知の場合の音声認識結果

手法	*Acc*	Substitution	Deletion	Insertions
Baseline	61.6	463	307	1834
手法 1	61.6	477	318	1813
手法 2	61.7	460	305	1836
手法 3	61.6	453	304	1827

Baseline と比較して、いずれの手法も認識精度の向上は確認できなかった。

ニュースアンカーの発話区間が未知の場合

アンカーの発話区間が未知の場合の音声認識結果を表 C.8 に示す。

表 C.8: アンカーの発話区間が未知の場合の音声認識結果

手法	*Acc*	Substitution	Deletion	Insertions
Baseline	26.7	957	2334	1684
手法 1	35.4	1014	1711	1657
手法 2	29.9	994	2080	1681
手法 3	35.4	930	1997	1682

アンカーの発話区間が未知の場合はいずれも発話区間が既知の場合と比較して大きく音声認識精度が低下した。また、いずれの手法も Baseline と比較して音声認識精度は向上している。

C.9　考察

本研究で提案した手法はいずれも Baseline と比較して下位のクラスタを選択する発話データが増加した。これは、発話区間を結合したことで、i-vector が性別の違いを判別できる程度の特徴を抽出できたためであると考えられる。

アンカーの発話区間が既知の場合に音声認識精度がいずれも変化がなかった理由として、背景雑音、音楽の存在が考えられる。音響モデルの学習に用いた CSJ は基本的に雑音が入らない環境で収録されている。このため、本実験で作成した木構造話者クラスの音響モデルのいずれも認識できない発話が多く存在してしまい、認識精度の違いがなかったと考えられる。音声認識精度の向上のために、雑音除去、もしくは雑音、音楽に頑健な音響モデルの作成が必要であると考えらえる。

ニュースアンカーの発話区間が未知の場合はいずれも発話区間が既知の場合と比較して大きく音声認識精度が低下した理由として、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、ニュース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したためである。また、いずれの手法も Baseline と比較してニュースアンカーの発話区間検出精度が向上していたため、削除誤りと挿入誤りが少なくなり、結果として音声認識精度が向上した。しかし、最も音声認識精度が高い場合でも 35.4%であり、ニュースアンカーの発話区間が既知の場合と

比較して 26.2%低下したため、音声認識精度を向上させるためには、ニュースアンカーの発話区間検出精度を向上させる必要がある。