修　士　論　文

題　　目

論文題目

指導教員

松永　昭一 教授

平成29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

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キ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、インデキシングには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のみの音声認識を行うことによって、より高精度なインデキシングが実現可能であると考えた。これまでに先行研究として音響特徴毎に発話者群を二分化していき、段階的に分類していく手法によって話者識別を行う研究が行われた [1]。しかし、この先行研究では話者の発話区間は既知であるとして行われたため人手によって発話区間を切り出す必要があった。そこで、本研究では、話者情報と発話区間が未知の場合でも用いることが可能なアンカーの発話区間抽出手法を提案する。

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

理論的背景

2.1 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。

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

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。

$Y_c(c)$

2.1.1 UBMに対するBaum-Welch統計量

すすす

2.1.2 全因子 w の確率分布とi-vectorの抽出

あ

2.1.3 因子分析モデルパラメータの推定

あ

2.1.4　コサイン類似度

あ てすと

第3章

提案手法

ていあんしゅほう

図 3.1: ぴあの

図 3.2: ぴあの

第4章

評価実験

ひょうかじっけん

4.1 予備実験

4.1.1 実験目的

しきいちきめるよー

4.1.2 実験条件

おおざっぱだよー

4.2 本実験

4.2.1 実験目的

ひょうかするよー

4.2.2 実験条件

よびといっしょだよー

表 4.1: 実験結果

被験者	実験 1	実験 2	実験 3
AAA	3	5	3
BBB	1	2	4
CCC	5	5	6
DDD	5	0	8

4.3　考察

こうさつするよー

第5章

結論

けつろん

第6章

追加

追加

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] Google