様式 C-19

科学研究費補助金研究成果報告書

平成２３年　６月２０日現在

機関番号：32675
研究種目：基盤研究（Ｃ）
研究期間：2008～2010
課題番号：20500166
研究課題名（和文）　語彙の確率的構造に基づく符号化による多選択肢タスク用音声認識の高精度化
研究課題名（英文）　Improvement of Very Large Vocabulary Speech Recognition using an encoding based on probabilistic structure of vocabulary
研究代表者
伊藤　克亘　　（ITOU KATUNOBU）
法政大学・情報科学部・教授
研究者番号：30356472

研究成果の概要（和文）：音声認識において、言語モデルの貢献を度外視した場合には、音素単位では認識率が悪くなる可能性は観測されたが、特定の音素系列において統計的に有意なレベルで認識率が悪化することはないことがわかった。一方で、話者認識・話者識別においては、言語モデルを利用しないのが一般的であるため、単語認識と同様の問題が生じることがわかった。さらに、話者識別のほとんどの応用では、学習データを十分に得ることが期待できない。さらに認識対象の音素系列を事前に想定できないことも多く、音声認識より高性能化が難しい問題であることが明らかになった。

研究成果の概要（英文）：For speech recognition, in a large vocabulary task, any phone sequence didn't induce statistically significant deficient performance without contribution of language models. For speaker recognition/verification, on the other hand, it seems to be difficult other than increasing of training data. Most applications of speaker recognition, it cannot be expected sufficient training data. Moreover, it is difficult to assume a target phone sequence in advance. Therefore, a new method is required for speaker recognition, because many previous methods for improving speech recognition cannot be efficient for speaker recognition.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2008 年度	1,600,000	480,000	2,080,000
2009 年度	1,000,000	300,000	1,300,000
2010 年度	900,000	270,000	1,170,000
総　計	3,500,000	1,050,000	4,550,000

研究分野：音声情報処理
科研費の分科・細目：情報学・知覚情報処理・知能ロボティクス・音声情報処理
キーワード：音声認識・話者認識・音響ライフログ

１．研究開始当初の背景

(1)　　コンピュータの高速化と大容量化にともない、多くの学習データを利用することができる確率モデルに基づく音声認識の枠組みが発展してきた。この枠組みで、音響モデルに関しては、学習データを増やすことができる不特定話者のタスクの性能が飛躍的に向上した。また、言語モデルに関しては、言語モデルが有効であるディクテーションなどのタスクにおいて性能向上が見られた。

(2)　　音声認識の有力な応用の一つにインタフェースでの利用があげられる。院ｔフ

ェースの利用としては、いわゆる自然な対話を行うような応用が主として対象とされている。しかし、実用的には、ポータブルプレイヤーのように小型の端末において、沢山の楽曲を選択するシステムやカーナビのようにハンズフリー、アイズフリーの環境で、目的地を選択するような多数の選択肢から選択する、というタスクが重要である。これらは GUI で多用されるボタンやメニューが有効ではないタスクであるため、音声認識が期待されるタスクである。

(3)　これまでの音声認識の枠組みは、音声の伝達を雑音あり通信路をモデルとしてとらえるものだった。認識誤りを通信路の雑音としてとらえ、言語モデルを利用したベイズ推定により雑音による誤りを修正しようとするものであった。

２．研究の目的

(1)　10 万語以上の固有名詞の高精度な音声認識手法を情報源符号化手法に基づき構築する。これまでの音声認識手法は、音声の伝達時に生じる誤りを修正するモデル化であった。本研究では、それとは異なり、情報源のエントロピーを下げることで問題を疑似的に簡単にして性能向上を図る。

(2)　具体的には、語彙が定まったときに生じる利用される音素系列の偏りを確率的構造としてとらえ、その構造を効率的に表現する符号化を検討する。

(3)　音声認識で標準となっている確率モデルを用いた認識は、話者認識・話者識別などにも応用されている。しかし、それらの応用では、音声認識ほどの成功を収めていない。それらに対しても、本研究の問題意識が適用可能であるかどうかを検証する。

３．研究の方法

(1)　タスクの評価のためのコーパスを整備する。具体的には、以下の 4 タスクについて整備する。
一つ目は、多選択肢タスクである。これについては、ポータブル音楽プレイヤーの楽曲選択タスクを想定し、楽曲のタイトルを読み上げた発話を収録したコーパスを構築する。発話環境が認識性能におよぼす影響を考慮するために、収音デバイスや発話環境についてもそれぞれいくつかのバリエーションを試す。さらに有用な応用があれば、それらについても評価できるようにする。
二つ目はセキュリティ用途の話者識別・話者認識である。このタスクについては、評価データと学習データの収録時期の違いが性能に与える影響が大きいことが知られているため、収録時期が異なるデータを収集し、学習・評価に利用できるようにする。
三つ目は、映像コンテンツへの話者ラベリングのための話者識別・話者認識である。このタスクについては、一人当たりの学習データが少ないこと、量に偏りがあることが性能に与える影響が多いと想定される。したがって、音声データ以外に利用できる周辺データである、字幕データや映像データもあわせて整備することで、性能向上を可能にする周辺情報があるかについても検証する。100 種類のコンテンツのデータを整備することを目標とする。
四つ目は、音響ライフログ中の音声区間検出である。このタスクについては、タスク設定や収録方法についても、ほとんど先行研究がなされていないため、まずは、収録デバイスや収録方法を検討するために、合計 70 時間程度のデータを収録する。

(2)　整備したコーパスを用いて、典型的な応用の性能評価を行う。その際には、音響モデルの構築単位や、モデルパラメータの数、モデルの形状などを評価できるような実験をおこなう。

４．研究成果

(1)　多選択肢タスクの具体的な例として、携帯音楽プレーヤ用の音声インタフェースを取り上げ、一人の話者が同一単語を多数発話したコーパスのプロトタイプを構築した。具体的には、25 語を 10 発話ずつ発話したものを、のべ 10 名分収録した。このコーパスを評価した結果、次のような知見が得られた。この程度のコーパスの規模の場合に、話者に関わらず特定の語の認識性能が悪い、つまり認識が困難な語がある、という統計的に有意な結果は得られなかったが、一部の子音は認識率の悪化に関与していることがわかった。また、その要因としては、音響モデル構築時にあることがわかった。
また、携帯音楽プレーヤの曲目選択というタスクでは、5000 語程度で、現状のニーズには十分であることがわかり、実際の使用状況にあわせて 10 万語を超える評価を行うのは現実的ではないことがわかった。

(2)　多選択肢タスクの具体的な例として携帯音楽プレーヤ用の音声インタフェース

を評価するために、収録環境による認識精度への影響を考慮するために複数のマイク、複数の発話環境での収録も行った。具体的には、USB スピーカーホン、bluetooth ヘッドセットで収録した。発話環境としては、高騒音下の環境として、交通量の激しい路上、自動車内、テレビ視聴時の居室などで収録した。さらに、日常生活における音声インタフェースの利用可能性を検証するため、実際の日常生活下でのデータ収録も行った。
その結果、音響モデルの性能が悪化するような環境では、タスクの規模が小さくても、本研究で想定したような状況が生じることがわかった。

(3)　音声認識以外にも、確率的な音響モデルを利用する応用はいくつかある。その例として、ボトムアップな音響信号の分類問題や、話者識別・話者認識があげられる。それらの問題にも、どのような単位でモデル化を行うか、どのようなモデル形状をとるかということについては、本研究の枠組で対応できることを見出した。
特に、テキスト独立型話者識別を対象として検証に取組んだ。テキスト独立型話者識別では、学習用の音声データが、音素の種類・出現回数ともに非常に偏りがある。したがって、識別性能を高精度化するためには、音素ごとに、モデル形状やパラメータ数を適切に選択することが望ましい。従来は、それらの点についての検証が余り行われていなかった。本研究では、まず、モデル形状や、モデル化の単位を検討した。インタフェースでのセキュリティ応用という、収録音の特性はある程度制御でき、初期学習データは、数分程度という環境下では、話者ごとに音素などを区別せず単一のモデルを用意して、モデル形状も最も単純な一状態が最もよい性能であることが明らかになった。
また、その条件下では、15%程度の誤り率を達成できた。しかし、この性能では、単独デバイスとしては、セキュリティへの応用には不十分な性能である。
そこで、音響ライフログと映像コンテンツの話者ラベルアノテーションを新たな応用として検証することにした。

(4)　映像コンテンツの発話データを用いた話者ラベリングについては、まず、タスクの検証をおこなった。従来研究では、ニュース番組へのラベリングが主であった。このようなタスクでは、主たる話者の発話量が非常に多く、また、コンテンツ内に出現して識別しなければならない話者の数が数名と非常に少ない。これらの条件下で従来研究においては、70%程度の話者ラベル付与率を得られている。
映像コンテンツで主流たるテレビ番組においては、ニュース以外のバラエティ、ドラマ、アニメーションなどのコンテンツの方が数は多い。したがって、本研究では、それらのコンテンツでも話者ラベリングが可能かどうかを検討した。上記コンテンツの多くでは、10名以上と、発話者が多いため、従来手法では、識別率が 20%程度と極めて低かった。
ディジタルテレビでは、字幕情報が半数以上のコンテンツで整備されている。字幕情報には、ドラマなどでは、主登場人物(1 名)については、ほとんどの発話に対して話者ラベルが付与されているものの、その他の人物については、最初の発話で話者ラベルが付いている他は、話者ラベルが付与されていない。その結果、全体の発話の 50%は話者ラベルが付与されている。したがって、話者ラベルが付与されている発話で話者モデルを構築し、残りの発話に話者ラベルを付与するという問題設定が可能となる。
話者ラベルが付与されている発話を学習データとし、字幕情報を元に音素単位の話者モデルを構築した。このモデルでは、識別率 40%がえられた。
音素単位のモデル化であるため、音素モデルごとの信頼性に偏りがある。したがって、本応用では、モデル選択が効果的な情報源の符号化として位置づけられるかもしれない。また、特徴量選択も情報源の符号化として位置づけられる可能性があろう。

(5)　音響ライフログとは利用者が体験全体を収録して活用するライフログの中でも特に音響データによるログをさす。音響データでもっとも役に立つ側面の一つが発話データである。何を話しているか、誰と話しているかなどが、重要な情報である。そこで、音響ライフログの情報から、人間が発話している部分を検出する問題を設定した。従来から、発話区間検出は様々なアプローチで取り組まれている問題であるが、ライフログの場合、非常に雑多な音響環境下でログが収録されるため、環境全体をモデル化することは困難であり、音声を統計モデルでモデル化することで検出を試みた。いわば、話者認識手法を用いた音声区間検出である。
その結果、ログをとっているユーザ自身が数名の相手と会話するような環境では、70%以上の発話が正しく検出できた。一方で、ユーザ自身以外の発話では、30%程度しか検出できなかった。
検出結果を検証したところ、SN 比が検出性能に大きく影響することがわかった。

(6)　　背景雑音や収録デバイスがばらつく実環境の音声認識においては、頑健な認識を行うために、雑音抑制や音声強調を前処理として用いることが多い。
そこで、それらがモデル化に対してどのような影響を与えるかを検証した。雑音抑圧・音声強調手法としては、スペクトル減算法および変調スペクトルフィルタリング手法を検証した。
音響ライフログにおいては、スペクトル減算法が有効であった。
一方、放送コンテンツでは、変調スペクトルフィルタリング法が有効であった。
この違いは、放送コンテンツにおいては、音声を伝達することが前提となっているため、仮に BGM や効果音が重畳している場合であっても、ある程度の SN 比が担保されていること、また、重畳される音は、音声と区別しやすいものに統制されていることが理由であろう。
一方で、音響ライフログにおいては、必ずしも音声が主たる音量で収録されるとは限らない。また、音量のダイナミックレンジも大きく変化する。変調スペクトルフィルタリングは、フィルタ係数のしきい値の調整で性能が左右されるため、音響ライフログでは、放送コンテンツほどの性能が得られなかった。

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔学会発表〕（計 14 件）

① Kiichiro Yamano and Katunobu ITOU, Detecting Scenes in Lifelog Videos based on Probabilistic Models of Audio data. Acoustics 08, 概要査読有, 2008 年 7 月 3 日
② 山野貴一郎, 伊藤克亘, バイノーラルマイクを用いたライフログ映像のショット識別, 信号処理シンポジウム, 査読無, 2008 年 11 月 13 日
③ 山野貴一郎, 伊藤克亘,　音響情報を用いたライフログデータのインデキシング, 情報処理学会全国大会, 査読無, 2009 年 3 月 11 日
④ Kiichiro Yamano and Katunobu ITOU, Browsing Audio Life-log Data Using Acoustic and Location Information, UBICOMM 2009, 査読有, 2009 年 11 月 15 日
⑤ 山野貴一郎, 伊藤克亘, 音響ライフログへのアノテーションのための話者と場所の自動分類, 情報処理学会全国大会, 査読無, 2010 年 3 月 11 日
⑥ 山室慶太, 伊藤克亘, 携帯端末への話者照合を用いたセキュリティロック, 情報処理学会全国大会, 査読無, 2010 年 3 月 11 日
⑦ 中村一文, 伊藤克亘, コンテンツ制作における収録音編集のための音声強調, 情報処理学会全国大会, 査読無, 2010 年 3 月 10 日
⑧ 田母神恒, 伊藤克亘, 高齢者の加齢による聴力低下に対応する音声強調, 情報処理学会全国大会, 査読無, 2010 年 3 月 11 日
⑨ 松浦健太, 伊藤克亘, Flash コンテンツ操作のための音声認識インタフェース, 情報処理学会全国大会, 査読無, 2010 年 3 月 11 日
⑩ Keita Yamamuro and Katunobu ITOU, Speaker model updating by the conversational sounds in speaker verification, IIWAS2010, 査読有, 2010 年 11 月 4 日
⑪ Kazufumi Nakamura and Katunobu ITOU, Speaker model updating by the conversational sounds in speaker verification, internoise 2010, 概要査読有, 2010 年 6 月 15 日
⑫ 平野邦彦, 伊藤克亘, 話者照合と音声認識を併用したスマートフォン向け認証システムの作成, 情報処理学会全国大会, 査読無, 2011 年 3 月 4 日
⑬ 山室慶太, 伊藤克亘, デジタル放送の字幕情報を用いた発話者のアノテーション, 情報処理学会全国大会, 査読無, 2011 年 3 月 4 日
⑭ 住澤卓也, 伊藤克亘, 音声を用いた農作業日誌システムの構築, 情報処理学会全国大会, 査読無, 2011 年 3 月 4 日

〔その他〕
ホームページ等
http://cis.k.hosei.ac.jp/info/faculty/digital/itou.html

６．研究組織
(1)研究代表者
伊藤　克亘（ITOU KATUNOBU）
法政大学・情報科学部・教授
研究者番号：30356472