様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成２４年４月１日現在

機関番号：１２６０８
研究種目：基盤研究（B）
研究期間：2009～2011
課題番号：２１３０００６２
研究課題名（和文）　ＷＦＳＴによる音声認識の高度化

研究課題名（英文）　Advancement of speech recognition technology using WFST

研究代表者
古井 貞熙（FURUI SADAOKI）
東京工業大学・名誉教授
研究者番号：90293076

研究成果の概要（和文）：重みつき有限状態トランスデューサ（WFST）による音声認識の高度化と、WFST デコーダの新たな応用展開を目指して研究を行い、以下の種々の成果を上げることができた。WFST デコーダの on-the-fly 合成アルゴリズムの改良を行い、世界最高性能の音声認識デコーダ（T³デコーダ）を開発した。これにさらに音声・非音声情報を組み込み、雑音下での認識性能を向上させた。開発したデコーダを、大規模コーパスを持たない音声や、複数言語が混在して用いられる音声の認識、transliteration などに適用し、効果的なアルゴリズムを提案した。さらに、デコーダ技術の新たな展開となるアイディアを創出した。開発した T³デコーダを、国内外に公開した。

研究成果の概要（英文）：With the aim of improving the performance of automatic speech recognition using the Weighted Finite State Transducer (WFST)-based decoder and developing new applications of the decoder, a wide range of research has been conducted and various achievements have been obtained. The world highest performance speech recognition decoder, "T³ decoder", has been developed by improving the on-the-fly algorithm for the WFST decoder. Recognition performance under noisy environment has been improved by incorporating speech/non-speech information to the decoder. Various new techniques have been developed to apply the decoder to the recognition of resource-deficient languages and code-switching speech, and to transliteration. Innovative ideas have been proposed toward new directions of the decoder technology. T³ decoder has been released to domestic as well as overseas research laboratories.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
２００９年度	5,800,000	1,740,000	7,540,000
２０１０年度	5,200,000	1,560,000	6,760,000
２０１１年度	2,900,000	870,000	3,770,000
総　計	13,900,000	4,170,000	18,070,000

研究分野：総合領域
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
キーワード：音声情報処理、音声認識、ＷＦＳＴ、デコーダ

１．研究開始当初の背景

近年、音声認識技術が進展し、電話サービス、放送ニュースへの字幕の自動作成、議会などでの議事録作成の自動化、音声ドキュメントの検索、カーナビなど、様々なところで使われるようになってきた。しかし、自由な話し言葉音声の認識や、雑音中の音声認識では、依然として低い認識性能しか得ることができない。現在の音声認識では、基本的にHMM（隠れマルコフモデル）で表現された音素情報と、バイグラム・トライグラムなどの統計的言語モデルによる単語連鎖情報の組合せによる認識処理が行われている。一方、人がどうやって音声認識をしているかを考えてみると、相手の声や話し方、周囲の雑音などの影響にすばやく対応し、話題、対話の流れ（コンテキスト）などの多様な情報を巧みに組み合わせて、相手の声を認識している。従って、コンピュータによる音声認識でも、認識性能の向上のためには、多様な情報を有効に組み合わせる必要がある。

近年、新しい音声認識の枠組みとして、重みつき有限状態トランスデューサ（WFST: Weighted Finite State Transducer）を用いた音声認識手法が提案された。この枠組みでは、上記のHMM、統計的言語モデルなどの情報をすべてWFSTの形式で表現し、それらを合成演算により一つに合成することで、音声認識結果としての仮説を探索するネットワークを構築する。合成演算には最適化の機能があるため、従来の2パスからなる音声認識手法に比べて、効率的で精度の高い認識を行うことができる。

しかしWFSTによる音声認識では、すべてのモデルを組み合わせて一つのネットワークに合成するため、そのネットワークが肥大化し、認識時に大きなメモリーが必要となる欠点があった。また、話題の変化などに対応して言語モデルなどを変更する際には、探索ネットワークを再構築する必要があるため、大きなオーバーヘッドを発生する問題があった。これらの欠点を克服する方法として、あらかじめ複数のネットワークに分けておいて、認識の過程で合成するon-the-fly合成法が提案されているが、これによって合成処理のオーバーヘッドが発声するだけでなく、ネットワーク全体としての最適化ができなくなるため、探索効率と認識精度が低下する問題があった。

また、WFSTによるデコーダには、その優れた原理から、音声認識に用いるだけでなく、他の種々の分野への適用可能性があるが、その観点からの検討は十分に行われていなかった。

２．研究の目的

上記の課題に対応するため、本研究では、on-the-fly 合成法の高度化と、雑音耐性の向上を中心とする、WFST 音声認識デコーダの機能の総合的な高度化を目指した。さらに、音声以外の新たな応用分野の開拓を目指し、多様な用途に適用可能なフレキシブルなデコーダを実現するとともに、その可能性を実証することを目的とした。

３．研究の方法

上記の目的を達成するため、WFST デコーダの on-the-fly 合成アルゴリズムの改良、WFST による音声認識デコーダの環境適応機能の高度化、その評価、WFST による音声認識デコーダを用いた新たなアプリケーションの開発、次世代の音声認識デコーダのアイディアの創出、音声認識以外の分野への適用について研究を行った。さらに、開発した音声認識デコーダを、我が国のみならず、世界中の研究者、技術者に利用・活用・評価してもらう仕組みを構築することを目指した。

４．研究成果

(1) WFST デコーダの On-the-fly 合成アルゴリズムの改良

音声認識で利用するモデルの大規模化を実現するため、認識時に探索ネットワークを動的に合成する方法（on-the-fly 合成）の高速化を実現した。過去に提案した最適化付き on-the-fly 合成手法に高速化のための技術を追加した。具体的には、WFST のトポロジーの最適化、合成演算で利用する半環演算の最適化、二つのラベル集合の高速な積集合計算法を実装した。その結果、ネットワークの拡大を抑えて探索効率の低下を減らすとともに、合成処理によるオーバーヘッドを大幅に減らすことに成功した。

評価実験の結果、日本語話し言葉コーパス(CSJ)を用いた大語彙音声認識タスクにおいて、大幅な認識速度の改善が得られることが確認できた。また、これにより、数十万語の超大語彙タスクにおいて、実時間の音声認識を実現することに成功した。

(2) 音声・非音声情報のデコーダへの組み込み

高雑音環境下において頑健な音声認識を実現するため、音声・非音声性スコアを組み込んだデコーダを実現した。音声・非音声特徴を表現する二つの混合ガウスモデル（GMM）により、音声・非音声の信頼度をフレーム毎に算出し、その値を用いて、音声・非音声に対応する認識仮説の音響尤度を調整する。この手法は、従来のフロントエンドで非音声のフレームを棄却する手法に比べて、音声フレームを誤って棄却するエラーを削減することができ、このため、高雑音環境下など、音声・非音声の判定が難しい環境下において、認識精度を改善することができる。運転中カーナビ音声コーパス(DJSC)を用いた音声認識タスクにおいて、本手法により、従来の一般的なフロントエンド型VAD(音声区間検出)手法（零交差とパワーの閾値による手法、および音声・非音声GMMの尤度比を利用する手法）と比べて、大幅な認識率の改善が確認され、本手法の有効性が確かめられた。

また、上記のGMMを、雑音環境や話者の音声の変化に自動的に適応させることにより、雑音環境での音声認識性能が大幅に向上することを確認した。

さらに本手法を、国際的な標準データベースである Aurora コーパスに適用し、種々の雑音環境において、従来法を上回る音声認識性能を有することを確認した。

(3) T^3デコーダの性能評価

開発した WFST に基づく T^3 デコーダ（Tokyo-Tech Transducer-based decoder）の性能を、国際的に定評のある３つのデコーダ（Juicer、HDecode、Sphinx3）の性能と比較し、実時間比（認識時間）に対する認識精度において、T^3デコーダが最も優れていること、さらに音響尤度計算にGPUを用いることによって、その特徴がさらに顕著になることを確認した。

(4) アイスランド語の音声認識

英語、日本語、中国語、フランス語、ドイツ語などの主要言語を除く、世界中のほとんどの言語において、統計的言語モデルを作成するのに必要な、十分な音声データベース（コーパス）が存在しないという問題がある。この問題に対処するため、文法的に英語に比較的近いが、大規模コーパスが存在しないアイスランド語を取り上げ、大規模コーパスが存在する英語で作成した言語モデル、当該言語（アイスランド語）に対して収集した小規模のコーパスから作成した言語モデル、および両言語間の翻訳モデルを、WFSTの枠組みで組合せ、音声認識を行う方法を実現した。実際の音声を用いた認識実験の結果、提案法の有効性が確認された。

(5) 複数言語混在音声の認識への適用

インドネシア語の音声認識において、英語とインドネシア語が文間、あるいは文内で入れ替わる状況（code-switching）に対応するため、code-switching言語モデルと、単独の言語の言語モデルを組み合わせる２種類の方法を検討し、それぞれ認識タスクの特徴に対応して特長があることが確認された。

(6) 音声認識誤り訂正の容易なインタフェースの検討

音声認識を用いた入力インタフェースにおいて、ユーザが認識結果候補を参照しながら逐次的に誤りを訂正する過程で、更新された言語モデルを用いて、候補単語リスト中での正しい単語のランクを自動的に上げることにより、誤り訂正を容易にする方法を提案し、その有効性を実験的に確認した。

(7) デコーダ技術の新たな展開

T^3デコーダをSilverlight plugin中で動作させることにより、Web ブラウザで音声認識が実現できることを示した。また、純粋関数型言語で WFST デコーダをプログラミングすることによって、デコーダのプログラムが桁違いにコンパクトになり、デコーダを含む音声認識システムの拡張などを容易に行う環境が構築できることを示した。

(8) TransliterationへのWFSTデコーダの適用

Joint source channel model (JSCM)を用いた transliteration（固有名詞を別の言語の文字に置き換えること、例えば、英語の固有名詞のカタカナ読み）に、WFSTデコーダを用いることにより、処理の高速化を実現した。

(9) 眼電位入力インタフェースへの適用

筋委縮性側索硬化症（ALS）において、眼球運動だけが最後まで障害されないことに基づき、眼電位を用いて眼球動作を認識する方法について検討した。複数電極からの電位入力に対して音声認識デコーダを用いた認識実験を行い、眼電位を用いたコミュニケーションの可能性を確認した。

(10) デコーダの公開

T^3デコーダを NICT（独立行政法人 情報通信研究機構）に譲渡し、NICTから国内・国外の音声認識研究者を対象に公開を開始した。今後のメンテナンスをNICTに委託した。

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）

〔雑誌論文〕（計８件）

① 吉川正祥、篠崎隆宏、岩野公司、古井貞熙、軽量な画像特徴量を用いたマルチモーダル音声認識、電子情報通信学会論文誌、査読有、Vol. J95-D、2012、pp. 618-627
② Yuzo Hamanaka, Koichi Shinoda, Takuya Tsutaoka, Sadaoki Furui, Tadashi Emori, Takafumi Koshinaka, Committee-based active learning for speech recognition, 電子情報通信学会英文論文誌、査読有、Vol. E94-D、2011、pp. 2015-2023
③ D. Yang, P. Dixon, S. Furui, A new hybrid method for machine transliteration, 電子情報通信学会論文誌, 査読有, Vol. E93-D, 2010, pp. 3377-3383
④ 大西翼、ディクソン・ポール、古井貞熙、WFST に基づく T^3 デコーダ、情報処理、査読無、Vol. 51、2010、pp. 1440-1448
⑤ 古井貞熙、音声認識技術の実用化への取組み、情報処理、査読無、Vol. 51、2010、pp. 1387-1393
⑥ 古井貞熙、小林哲則、矢頭隆、大淵康成、河村聡典、三木清一、庄境誠、音声認識実用化技術の展開、電子情報通信学会誌、査読無、Vol. 93、2010、pp. 725-740
⑦ P. Dixon, T. Oonishi, S. Furui, Harnessing graphics processors for the fast computation of acoustic likelihoods in speech recognition, Computer Speech and Language, 査読有, Vol. 1, 2009, pp. 510-526
⑧ 大西翼、ディクソン・ポール、岩野公司、古井貞熙、WFST 音声認識デコーダにおける on-the-fly 合成の最適化処理、電子情報通信学会論文誌、査読有、Vol. J92-D、2009、pp. 1026-1035

〔学会発表〕（計 42 件）

① Hiroko Murakami, Koichi Shinoda, Sadaoki Furui, Designing text corpus using phone-error distribution for acoustic modeling, Proc. IEEE ASRU, 2011.12.11, Hawaii（米国）
② Takahiro Shinozaki, Masakazu Sekijima, Shigeki Hagihara, Sadaoki Furui, Compact speech decoder based on pure functional programming, Proc. APSIPA-ASC, 2011.10.18, Xi'an（中国）
③ Takahiro Shinozaki, Sadaoki Furui, Strategies for model training and adaptation based on data dependency control, Proc. APSIPA-ASC, 招待論文, 2011.10.18, Xi'an（中国）
④ Christoph Draxler, Thomas Altosaar, Sadaoki Furui, Mark Liberman, Peter Wittenburg, Speech processing tools - An introduction to interoperability, Proc. INTERSPEECH, 招待論文, 2011.8.28, Florence（イタリア）
⑤ 古井貞熙、コンピュータによる音声認識のこれまでと今後の展望、日本音響学会春季研究発表会論文集、招待講演、2011.3.9、東京（東京都）
⑥ S. Furui, Selected topics from ASR research for Asian languages at Tokyo Tech, Proc. APSIPA-ASC, 招待講演, 2010.12.17, Singapore（シンガポール）
⑦ S. Furui, Automatic speech recognition - Where we are, and where we should go -, Proc. ICALIP, 招待講演, 2010.11.23, Shanghai（中国）
⑧ T. Oonishi, K. Iwano, S. Furui, VAD-measure-embedded decoder with online model adaptation, Proc. INTERSPEECH, 2010.9.30, 幕張（千葉県）
⑨ J. R. Novak, P. Dixon, S. Furui, An empirical comparison of the T^3, Juicer, HDecode and Sphinx3 decoders, Proc. INTERSPEECH, 2010.9.29, 幕張（千葉県）
⑩ P. Dixon, S. Furui, Exploring web-browser based runtime engines for creating ubiquitous speech interfaces, Proc. INTERSPEECH, 2010.9.27, 幕張（千葉県）
⑪ D. Yang, P. Dixon, S. Furui, Jointly optimizing a two-step conditional random field model for machine transliteration and its fast decoding algorithm, Proc. ACL, 2010.7.11, Uppsala（スウェーデン）
⑫ J. Novak. E. Whittaker, S. Furui, Evaluation of a WFST-based ASR system for train timetable information, Proc. APSIPA-ASC, 2009.10.6, 札幌（北海道）
⑬ A. Jensson, T. Oonishi, K. Iwano, S. Furui, Development of a WFST-based speech recognition system for a resource deficient language using machine translation, Proc. APSIPA-ASC, 2009.10.5, 札幌（北海道）
⑭ P. Dixon, T. Oonishi, K. Iwano, S. Furui, Recent development of WFST-based speech recognition decoder, Proc. APSIPA-ASC, 2009.10.5, 札幌（北海道）
⑮ T. Oonishi, P. Dixon, K. Iwano, S. Furui, Robust speech recognition using VAD-measure-embedded decoder, Proc. INTERSPEECH, 2009.9.9, Brighton（英国）
⑯ T. Oonishi, P. Dixon, K. Iwano, S. Furui, Generalization of specialized on-the-fly composition, Proc. ICASSP, 2009.4.22, Taipei（台湾）

⑰ T. Oonishi, P. Dixon, S. Furui, Fast acoustic computations using graphics processors, Proc. ICASSP, 2009.4.22, Taipei (台湾)

〔図書〕(計１件)

① Agnieszka Betkowska Cavalcante, Koichi Shinoda, Sadaoki Furui, Robust speech recognition in the car environment, in LTC 2009, LNAI 6562, Springer, 2011, 11p

〔その他〕
ホームページ等
http://www.furui.cs.titech.ac.jp/top_e.html

６．研究組織
(1)研究代表者
古井 貞熙 (FURUI SADAOKI)
東京工業大学・名誉教授
研究者番号：90293076

(2)研究分担者
篠田 浩一 (SHINODA KOICHI)
東京工業大学・大学院情報理工学研究科・准教授
研究者番号：10343097
篠崎 隆宏 (SHINOZAKI TAKAHIRO)
千葉大学・大学院融合科学研究科・助教
研究者番号：80447903