論文 / 著書情報
Article / Book Information

論題(和文)	話し言葉音声の認識精度向上のために
Title(English)	
著者(和文)	古井 貞熙, 中村 匡伸, ポール ディクソン, 大西 翼, 岩野 公司
Authors(English)	SADAOKI FURUI, Masanobu Nakamura, Paul Dixon, Oonishi Tasuku, Koji Iwano
出典(和文)	人工知能学会資料 AIチャレンジ研究会, SIG-Challenge-A702-1, Vol. , No. , pp. 1-7
Citation(English)	, Vol. , No. , pp. 1-7
発行日 / Pub. date	2007, 11

話し言葉音声の認識精度向上のために

Toward increasing spontaneous speech recognition accuracy

○古井 貞熙, 中村 匡伸, ポール ディクソン, 大西 翼, 岩野 公司
東京工業大学大学院情報理工学研究科計算工学専攻
* Sadaoki Furui, Masanobu Nakamura, Paul Dixon, Tasuku Oonishi, Koji Iwano
Tokyo Institute of Technology, Department of Computer Science
{furui, masa, dixonp, oonishi, iwano}@furui.cs.titech.ac.jp

Abstract – This paper reports progress on two major fronts in our research on spontaneous speech recognition. First, by comparing various features of spontaneous speech and read speech, we have found that spectral space reduction and linguistic complexity are two major sources of decreases in spontaneous speech recognition accuracy. Second, by introducing on-the-fly composition, disk-based search and acoustic likelihood calculation using GPU hardware, we have constructed a high-flexibility and high-performance WFST-based decoder.

1．はじめに

大語彙の連続音声認識でも、テキストを読み上げた音声であれば、かなり高い精度で音声認識できるようになったが、普通の話し言葉、すなわち考えながら（あるいは考えるよりも前に）話している自発性の高い音声では、認識性能が大幅に下がってしまう。これは、読み上げ音声と話し言葉音声は、音響的にも言語的にも大きく違うことを示している。音声の現象を調べ、音声のモデルを作り、音声認識システムを構築するためには、大規模の音声コーパスを作成することが不可欠である。ところが、大規模の話し言葉音声コーパスを作成するには、実際の話し言葉音声を大量に録音し、人手で書き起こし、さらに形態素解析を行って形態素（単語）に区切ったり、実際の発音を与えたりしなければならないため、多大な人手とコストがかかる。このため、話し言葉音声の研究は、世界的にも、10年くらい前まではほとんど行われず、そのために話し言葉と読み上げ音声の違いも強く意識されず、タスクを限定した対話音声システムを除くと、大語彙連続音声認識では、もっぱら読み上げ音声を用いた研究が行われていた。対話音声システムでも、ユーザは通常、コンピュータの能力を意識した発声をするため、その音声は、ある意味できちんとした読み上げ音声に近く、自然な話し言葉音声とは異なる。

しかし、ニュース音声を用いた研究[1]が進展するに伴って、アナウンサーがテキストを読み上げている音声から、現場からの中継や、スポーツの実況中継などの音声を認識するようになって、自然な話し言葉音声の認識の難しさが強く意識されるようになった。米国では、DARPAによって、Switchboardと呼ばれる、電話での会話音声が大量に録音され、音声コーパスが作成されるようになったが[2]、わが国では、その動きが遅れていた。そこで、国語研、通総研（当時）と東工大で、1998年秋に、科学技術振興調整費の開放的融合研究推進制度によるプロジェクトに、「話し言葉の言語的・パラ言語的構造の解明に基づく『話し言葉工学』の構築」事業を提案し、多数の方々のご支援を得て、1999年から5年間、プロジェクトが推進された[3]。このプロジェクトには、上記3機関の研究者だけでなく、京都大学など種々の大学や研究機関からの研究者が参加し、極めて密度の高い研究開発が行われた。そのプロジェクトでは、「日本語話し言葉コーパス（CSJ: corpus of Spontaneous Japanese）」を構築したが、その中では、共通の話者による話し言葉音声と読み上げ音声の違いが、解析できるようになっている[4]。

最近、中国語に関しても、多数話者による話し言葉音声と読み上げ音声を収録したコーパスが作られている。世界的には、会議、講演、講義、国会の討論など、多様な話し言葉を対象とした研究が活発に行われつつある[5]。以下に、日本語と中国語の話し言葉コーパスを用いた、話し言葉音声と読み上げ音声の違

いに関する分析結果について述べる。

話し言葉音声の認識のためには、種々の知識を容易に導入でき、スケーラビリティがよく、性能のよいデコーダを開発することが必要である。我々は、このような観点から、WFST（重みつき有限状態トランスデューサ，Weighted Finite State Transducer）に着目し、デコーダを開発してきた[6]。世界的にも、WFSTによるデコーダが主流になりつつある。我々のデコーダの開発現状についても、以下に述べる。

2．日本語話し言葉コーパスを用いた話し言葉の音響的・言語的特徴の分析

2.1 日本語話し言葉コーパス

上記の「日本語話し言葉コーパス（CSJ）」は、タスクを限定しない独話、特に講演（学会講演と模擬講演）を主たる対象として、のべ約 650 時間、単語（形態素）にして約 700 万語規模を有し、世界最大かつ高品質の話し言葉コーパスである[4]。独話、特に講演を対象としたのは、準備された講演音声は、音響的および言語的に、読み上げ音声と対話音声の中間に位置していると考えられ、話し言葉の特性を有しながら、対話音声よりもモデル化しやすいと考えられるためと、講演音声をコンテンツ化したり、自動的に字幕をつけたいというニーズは、極めて大きいためである。全体のコーパスの大きさは、音声認識のための統計的言語モデルを構築するために、最低限必要と思われるデータ量に基づいて決めた。研究の幅を広げるため、コーパスの一部に、インタビューなども含めた。

コーパス全体について、セグメンテーション、書き起こし（正書法による基本形と発音形）、形態素解析などを行ったが、全体の約 8%（50 万語）の「コア」については、人手による形態素解析結果、構文構造、要約の他、韻律情報などパラ言語情報まで含めたコーパスとした。コアの形態素解析結果を用いて、コーパス全体を自動的に形態素解析するためのツール（解析プログラム）の構築も行った。

2.2 音響的特徴の分析

CSJ には、同じ話者が学会講演（AP）、模擬講演（EP）、対話音声（学会講演の内容に関するインタビューと自由対話）（D）、および読み上げ音声（自分の学会講演の書き起こしや、対話形式エッセーを読み上げた音声）（R）がデータベース化されているため、この中の男性・女性話者各 5 名による音声を用いて、話し言葉と読み上げ音声の音響的特徴の違いに関する分析を行った[7]。異なる発声スタイルで、話者が共通しているため、話者による声質の違いの影響が除去できる。

話し言葉音声においては、読み上げ音声に比べて、いわゆる発声のなまけの効果で、スペクトル空間が縮小する傾向がある。これを確認するため、各話者の各音素の 3 状態 1 混合モノフォン HMM を作成し、音素ごとにその中心状態のケプストラムの平均ベクトルの、全音素の分布の中心（平均値）からの距離の縮小率を求めた。発話単位（400ms 以上の無音区間で区切られた約 10 秒長程度の区間）ごとに CMS 処理を行っている。実験の結果、ほとんどすべての音素において、読み上げ音声に比べて、話し言葉音声の平均ベクトルの中心からの距離が減少する傾向が見られ、対話音声において特に顕著になることが確認された。図 1 に、話し言葉音声の場合の、全音素の分布の中心からの各音素までの距離を、読み上げ音声の場合の値で割った縮小率を示す。発話スタイルごとに、母音・子音別に、音素と話者に関して平均して示した。

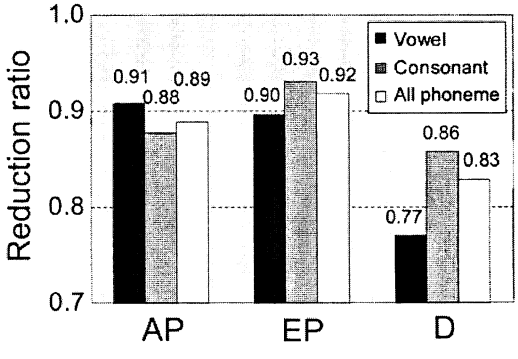

Fig. 1 – Mean reduction ratio of vowels, consonants and all phonemes, respectively, for each speaking style.

音声認識性能に直接関係するのは、異なる音素間の距離と考えられるので、次に、各音素モデル間のマハラノビス距離の分布を調べ

た。AP、EP、Dに加えて、新聞読み上げ音声（R）について分析した。図2に、各音素間のマハラノビス距離の相対累積度数を、発話スタイルごとに示す。朗読（R）、自然独話（AP, EP）、対話（D）の順に、つまり自発性が高くなるに従い、音素間のマハラノビス距離が小さくなる傾向がある。

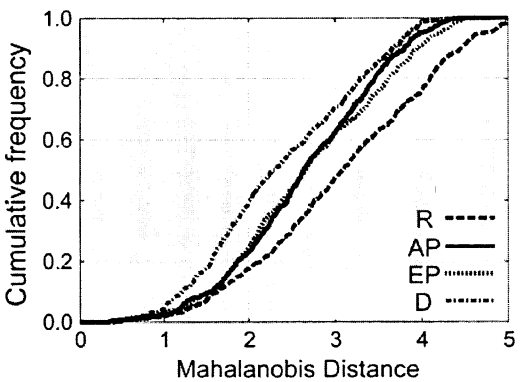

Fig. 2 – Distribution of Mahalanobis distances between phonemes for each speaking style.

次に、音素間のマハラノビス距離の平均値と音素認識正解精度の関係を調べた。音素モデルは、CSJ の学会講演と模擬講演の男女計100 名の音声データを用いて作成し、各発話スタイルの音声を認識した。結果は図3に示す通りで、音素間のマハラノビス距離の平均値と音素正解精度の間には、高い相関があることがわかった。このことは、音素間のマハラノビス距離を調べれば、音素正解精度が予測できることを示している。

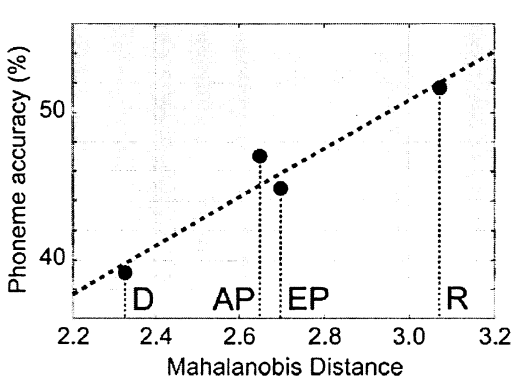

Fig. 3 – Relationship between Mahalanobis phoneme distance and phoneme recognition accuracy.

2.3 言語的特徴の分析

CSJ 中の種々の発話スタイルのコーパスに加えて、書き言葉およびそれに近いタスクとして、毎日新聞記事（NP）および放送のニュース解説（NC）のコーパスを用いて、それらの言語モデルの比較を行った[8]。各コーパスについて、トライグラムを作成し、コーパス相互間のテストセットパープレキシティと未知語率を調べた。コーパスによって語彙数や未知語率が異なるため、厳密な比較はできないが、テストセットと同じコーパスから言語モデルを作成した場合でも、書き言葉である新聞記事に比べて、講演、対話などの話し言葉では、パープレキシティが約5倍に大きくなることがわかった。各コーパス間のパープレキシティ行列を、距離行列化し、多次元尺度構成法により、言語モデル間の距離の可視化を行った結果を、図4に示す。第1軸（横軸）が、ほぼ自発性の度合いに対応している。

音素モデルを共通にし、これらの言語モデルを用いて各タスクの音声認識実験を行った結果、パープレキシティと単語正解精度の間に、高い相関があることが確認されている。

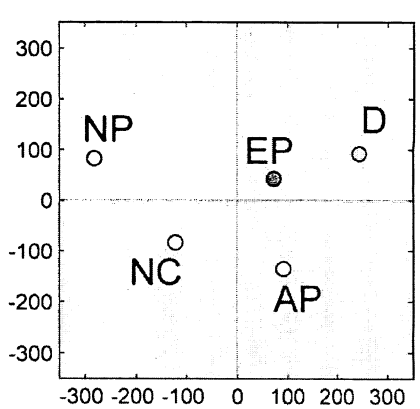

Fig. 4 – Relationship between the language models derived from the perplexity matrix.

3．中国語音声を用いた実験

3.1 中国語音声データベースと実験方法

話し言葉音声としては、電話での対話を想定して、通常のマイクロホンで収録した対話音声を用い、読み上げ音声としては、同一話者が発声したニュース記事読み上げ音声、および本や新聞を朗読した音声を用いた[9]。分

析対象の音響単位としては、トーン（四声および軽声）を区別した全 184 種類の中国語 Initial-Final モデル（以下、簡単のため、音素と呼ぶ）を用いた。

上記の日本語の音素の分析の場合は、簡単のために各音素を、単一ガウス分布を用いた 3 状態のモノフォン HMM でモデル化し、その中心の状態を用いたが、中国語音素の分析では、さらに精密化を図るため、話し言葉と読み上げ音声それぞれの各音素を、混合ガウス分布を用いたモノフォン HMM でモデル化した。日本語の場合と同様に、発話単位（無音区間で区切られた約 10 秒長程度の区間）ごとに CMS 処理を行っている。HMM の学習には、460 名（男女各 230 名）の話者の音声を用い、認識実験の評価話者には、30 名（男性 16 名, 女性 14 名）の話者を用いた。

単一ガウス分布の場合は、マハラノビス距離を用いることによって、分布間の距離を測ることができるが、混合ガウス分布の場合はそれができないので、Kulback-Leibler 擬距離（KLD）を用いた。KLD を定義どおりに計算することは困難なので、ここでは unscented transform に基づく、次のような近似式を用いた[10]。

$$D(s\|\tilde{s}) \approx \frac{1}{2N}\sum_{m=1}^{M}\omega_m \sum_{k=1}^{2N}\log\frac{p(o_{m,k}|s)}{p(o_{m,k}|\tilde{s})}$$

ただし、N は音響特徴量（ケプストラム、対数パワー、およびそれらの動的特徴量）の次元数（N=39）、M は混合数、ω_m は GMM における m 番目のガウス分布の混合重み、$o_{m,k}$（$1 \leqq k \leqq 2N$）は、m 番目のガウス分布の k 番目の sigma point（標準偏差の位置）である。

3.2 全音素間の KLD の分布

話し言葉音声と読み上げ音声において、音素モデルの混合数を、1, 2, 4, 8, 16, 32, 64 の 7 段階に変化させた場合の、全音素相互間の KLD の変化を比較した。各音素に対して、10 個の近傍音素を選び、その KLD を求めた。音素間の KLD の値の相対累積度数を、図 5 に示す。音素モデルの混合数が増えるにしたがって、モデルがより正確になってくるので、全音素間の KLD が大きくなるのは当然であるが、読み上げ音声に比べて、話し言葉音声の場合に、混合数の増加による KLD の増加率が小さいことがわかる。

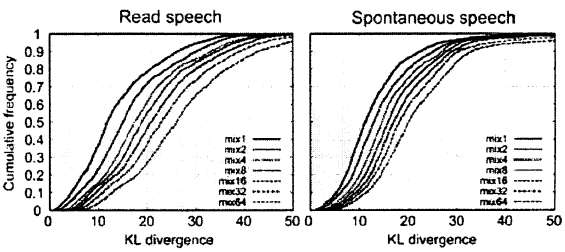

Fig. 5 – Comparison of distributions of KLD distances between phonemes for read and spontaneous speech as a function of the number of mixtures.

3.3 全音素間の KLD の中位値と、音素認識精度の関係

HMM の各状態の混合数を 1 から 64 まで増やしていったときの、音素（Initial-Final モデル）ベースのネットワークを用いた音素認識実験を行った。挿入ペナルティは、発話スタイルごとに最適化した。混合数が 1 の場合に対する音素認識誤りの削減率と、音素間の KLD の中位値との関係を、図 6 に示す。この結果から、KLD と音素認識誤りの削減率の間には、強い相関があることがわかる。ただし、同じ KLD の値でも、読み上げ音声と話し言葉音声では、音素正解精度の絶対値には顕著な違いがある。この原因を明らかにすることは、話し言葉音声の認識精度を向上させるための示唆を与える可能性がある。

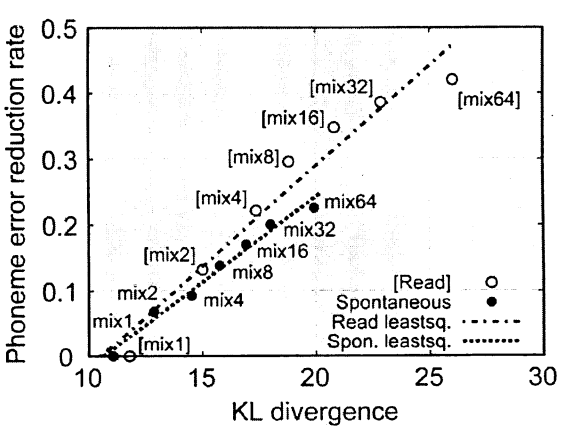

Fig. 6 – Relationship between median of KLD phoneme distances and phoneme recognition error reduction rate.

4．話し言葉音声の分析に関する考察

話し言葉音声の認識はどうして難しいのか？　その疑問への回答と、認識精度の向上を目指して、日本語と中国語の音声を用いて分析を行ってきた。これまでに、言語に依存せずスペクトル空間が縮小することなど、いくつかの定量的な事実が明らかになってきたが、まだ不明なことが多い。これまでの分析結果は、残念ながらまだ認識性能の向上には結びついていない。話し言葉音声の認識は、音声認識応用の展開において、最も重要なテーマの一つであるので、データベースの整備を含め、言語横断的な協力によって、分析を強力に進めていく必要がある。

5．話し言葉音声認識のためのデコーダ

5.1 WFST デコーダ

我々は、経済産業省「情報家電センサー・ヒューマンインターフェイスデバイス活用技術開発・音声認識基盤技術」プロジェクトの一環として、WFST (Weighted Finite State Transducer)を利用したデコーダを構築している。WFST に基づくデコーダは、高速で高性能、かつフレキシビリティの高い方法として注目されており、これまでに種々のデコーダが実際に構築されて、有効性が確認されている[11][12][13]。

WFST に基づく音声認識では、探索に先だち、音響モデルや言語モデル、単語発音辞書などの構成要素を合成してまとめあげ、一つの巨大な WFST 形式のネットワークを構築する。認識はこのネットワークを探索することで進められる。従来の認識手法に比べて、探索ネットワークを保持するためのメモリ量を多く必要とする反面、モデルの融合が事前に行われることから、探索時に動的なモデル融合を行う必要がなく、高速なデコーディングが可能となる。また、様々な形式のモデルや辞書を扱う必要が生じても、最終的に WFST の形式でネットワークに変換できれば、モデルに応じてデコーダ自体を変更するといった必要がないため、柔軟なデコーダが実現できる。我々は、スケーラビリティの向上を狙い、省メモリ化や高速化などのための、様々な機能の実装と検討を行っている。

5.2 WFST デコーダの構成

WFST は、与えられた入力記号列に対して状態遷移を繰り返し、それに対応した出力記号列と重みを出力する有限状態オートマトンの一種である。WFST を利用した音声認識では、まず音素モデルや言語モデル、単語発音辞書などをそれぞれ個別に WFST の形式に変換する。次に、基本演算の一つである合成(composition)演算を施して WFST 同士をまとめ、複数のモデルを組み込んだ一つの WFST を生成する。合成に際して、最小化(minimization)や決定化(determinization)などの演算を施すことにより、すべてのモデルを考慮したネットワーク全体に対して最適化が行われ、効率的な探索ネットワークを生成することができる。これにより、高速で高精度な音声認識を実現することができる。

入力音声は、フロントエンドを通して特徴ベクトルに変換され、デコーディングに利用される。本デコーダでは、Sphinx[14]で利用されている、多段フィルタによるフロントエンド設計を採用している。例えば、入力音声は、「窓掛け」や「FFT」などの個別の処理フィルタに順次通されることで、MFCC などの特徴ベクトルへ変換される。ユーザは、利用目的に応じて設計した処理フィルタを、容易に取り入れることができる。例えば、動画から画像特徴量への変換フィルタを作成することで、デコーダを動画認識やマルチモーダル音声認識に利用することが容易にできる。

探索は、フレーム同期型の１パス探索であり、第１位仮説からの尤度差と保持仮説数の上限値を用いた枝刈りを行っている。認識結果は 1-best や単語ラティス形式で出力する。認識結果の出力方式には、探索の終了時に一度に出力を行う「バッチ型」と、探索途中で確定した単語列を順次出力する「逐次型」を選択することができる。

音声への字幕付与などのアプリケーションでは、発話から単語列確定までの遅れ時間の短縮が非常に重要になる。逐次デコーディングにより、全ての発話が終了した後に最尤となる単語列を確定するのではなく、発話途中で早期に単語列を確定することができる。具体的手法として、保持している複数の仮説の

単語履歴に対し、履歴中の先頭からの部分単語列が共通になった段階で、その単語列を出力する手法[15]、過去の仮説単語履歴と比較する手法[16]、推定された無音区間毎に単語列を出力する手法[17]などがある。本デコーダでは[15]の手法を用いている。

本デコーダは音声検索、リスコアリング処理を利用したアプリケーションなどとの親和性を高めるため、ラティス形式での出力を行うことを可能にしている。ラティスは WFST 形式であり、仮説展開の際に同時に生成される。ラティスを構成する要素単位は、事前に合成されたネットワークに依存しており、HMM の状態を構成要素とした単語ラティスや、(文脈依存) 音素を構成要素とした単語ラティスが得られる。

5.3 省メモリ化

WFST による音声認識では、肥大化した探索ネットワークの読み込みに伴う、メモリ使用量の増大がしばしば問題となる。その対策として、1) 事前のネットワーク構築の段階ですべての WFST を合成せず、一部の WFST については、探索中に動的に合成するようにして、読み込む探索ネットワークの肥大化を防ぐ手法 (on-the-fly 合成[18][19][20])、2) 認識時に探索ネットワーク全体をメモリ上に読み込むのではなく、ディスク上に展開しておき、必要分だけを随時メモリ領域に読み込んで利用する方法 (disk-based search[21]) の２つについて検討を行った。On-the-fly 合成に関しては、事前の探索ネットワークの合成に利用する WFST を換えて、様々な方式について検討した。

5.4 高速化

混合ガウス分布を音響モデルとして利用する音声認識では、ガウス分布の混合数の増加に伴い、音響尤度計算に多くの時間を要する。このため音響尤度計算を効率的に行うことは、高速な音声認識において非常に重要である。

近年、グラフィックスカードに搭載された GPU(Graphics Processing Unit)の浮動小数点速度が、 CPU のそれと比較して飛躍的に向上しており、将来的には、汎用的な計算プロセッサとして GPU が広く利用されることが予想される。我々は高速に音響尤度を計算する一つのアプローチとして、GPU を利用した音響尤度計算手法を提案し、その実装を行った。このアプローチでは、高速な演算ユニットを利用して正確に音響尤度計算を行うため、近似的な計算により計算量を削減するアプローチと違い、認識率の劣化なしに高速に音響尤度を計算することができる。

5.5 評価結果

On-the-fly 合成において、オーバヘッドの増大を防ぐため、HMM としてスキップなしの left-to-right 型を扱うこととした。その結果、最大 60%以上のメモリ消費量の削減が実現できた。また disk-based search を行うことで、最大で 60%以上のメモリ消費量の削減を確認することができた。さらに、それらを組み合わせることで、全ての WFST を事前に合成した場合と比べて、80%以上のメモリ消費量の削減を確認することができた。

また GPU の利用により、最大で 25%程度の認識時間が削減できることが確認できた。

6. むすび

話し言葉音声の認識性能の向上を目指して進めている研究の中から、二つの最近の研究内容を紹介した。一つ目は、話し言葉音声と読み上げ音声の違いに関する分析で、話し言葉音声では、読み上げ音声に比べて、顕著にスペクトル（ケプストラム）空間が縮小するとともに、言語モデルの複雑さが増して、それらが音声認識誤りの増大の原因になっていることを明らかにした。

二つ目は、話し言葉音声認識のための、フレキシビリティの高い、高性能のデコーダの開発で、WFST をベースとして、種々の機能を持たせるとともに、省メモリ化、高速化を実現している。

人が話し言葉音声を理解する際には、文脈などを含む、極めて多様な知識を組み合わせて用いている。フレキシビリティの高いデコーダをベースに、如何にこれらの知識を組み込んだ音声認識の枠組みを構築していくかが、今後の進展の鍵になるであろう。

謝辞

本研究は、21 世紀 COE プログラム「大規模知識資源の体系化と活用基盤構築」および、経済産業省「情報家電センサー・ヒューマンインターフェイスデバイス活用技術開発・音声認識基盤技術」プロジェクトの支援を得て行われている。

参考文献

[1] J.-L. Gauvain et al., "Structuring broadcast audio for information access," EURASIP Journ. on Applied Signal Process.," vol.2003, no.2, pp. 140-150, 2003.

[2] J. J. Godfrey et al., "Switchboard: Telephone speech corpus for research and development," Proc. ICASSP, pp. I-517-520, 1992.

[3] S. Furui, "Recent progress in corpus-based spontaneous speech recognition," IEICE Trans. Inf. & Syst., vol.E88-D, no.1, pp. 1-11, 2005.

[4] 前川, "『日本語話し言葉コーパス』公開版の仕様," 第３回話し言葉の科学と工学ワークショップ講演予稿集, pp. 7-14, 2004.

[5] 古井, "音声認識の動向[1]-話し言葉音声認識," 電子情報通信学会誌, vol.89, no.8, pp. 746-751, 2006.

[6] 大西他, "WFST 音声認識デコーダの開発とその性能評価," 情報処理学会研究報告, vol.2007, no.68, 2007.

[7] M. Nakamura et al., "Analysis of spectral space reduction in spontaneous speech and its effects on speech recognition performances," Proc. Interspeech, pp. 3381-3384, 2005.

[8] 中村他, "読み上げ音声に対する話し言葉音声の言語的特徴の分析," 日本音響学会秋季研究発表会講演論文集, 3-6-10, 2005.

[9] 中村他, "KLD を用いた中国語における読み上げ音声と話し言葉音声の違いの分析," 日本音響学会秋季研究発表会講演論文集, 3-6-10, 2007.

[10] J. Du et al., "Minimum divergence based discriminative training," Proc. Interspeech, pp. 2410-2413, 2006.

[11] M. Mohri et al., "Weighted finite-state transducers in speech recognition, " Computer Speech and Language, vol.16, no.1, pp.69-88, 2002.

[12] D. Moore et al., "Juicer: A weighted finite-state transducer speech decoder," Proc. MLMI, 2006.

[13] T. Hori, "NTT Speech recognizer with OutLook On the Next generation; SOLON," Proc. Communication Scene Analysis, 2004.

[14] P. Lamere et al., "Design of the CMU Sphinx-4 decoder," Proc. ICSLP, pp.1181-1184, 2003.

[15] P. F. Brown et al., "Partial traceback and dynamic programming," Proc. ICASSP, pp.1629-1632, 1982.

[16] 今井他, "最ゆう単語列逐次比較による音声認識結果の早期確定," 電子情報通信学会論文誌 D-II, vol.J84-D-II, no.9, pp.1942-1949, 2001.

[17] 河原他, "話し言葉音声認識のための言語モデルとデコーダの改善," 情報処理学会研究報告, vol.2001, no.55, pp.15-22, 2001.

[18] H. J. G. A. Dolfing et al., "Incremental language models for speech recognition using finite-state transducers," Proc. ASRU, 2001.

[19] T. Hori et al., "Generalized fast on-the-fly composition algorithm for WFST-based speech recognition," Proc. Interspeech, pp. 847-850, 2005.

[20] D. A. Caseiro et al., "A specialized on-the-fly algorithm for lexicon and language model composition," IEEE Transactions on Audio, Speech, and Language Processing, vol.14, no.4, pp. 1281-1291, 2006.

[21] D. Willett et al., "Time and memory efficient Viterbi decoding for LVCSR using a precompiled search network," Proc. Eurospeech, pp.847-850, 2001.