〈ソリューション〉

話し言葉認識技術とその応用

Spontaneous Speech Recognition Technology and Its Applications

磯谷亮輔* Ryosuke Isotani
畑崎香一郎* Kaichiro Hatazaki
服部浩明* Hiroaki Hattori
奥村明俊* Akitoshi Okumura
渡辺隆夫* Takao Watanabe

要　旨

ITとネットワークの融合により企業における業務効率化をめざすNECのUNIVERGEソリューションでは，音声とデータがIPで統合したコミュニケーション環境が提供されますが，そのような環境では今後，音声情報を他のデータと連携して有効活用することが必要になってくると考えられます。そのための技術として，NECでは話し言葉の音声をテキスト化する話し言葉認識技術を開発しています。本稿では，話し言葉認識技術と，コンタクトセンタを始めとするいくつかの応用事例について紹介します。

UNIVERGE proposed by NEC aims at promoting business efficiency by “IT/Network Integrated Solutions”. It provides communication environments where voice and data transmissions are integrated. In such environments, it will become important to effectively utilize voice communication data. In order to make voice data easy to handle, we are developing the speech recognition technology which transforms spontaneously spoken utterances into text. This paper describes our speech recognition technology and its applications including contact center solution.

1．まえがき

NECは，ITソリューションとネットワークシステムを融合し，TCO削減，業務効率化などの可能なブロードバンドオフィスを実現するUNIVERGEソリューションの開発を進めています。UNIVERGEでは，音声とその他のデータをIPで統合し，さらに各種ITシステムと連携することにより，最適なコミュニケーション環境，コラボレーション環境の提供をめざしています。このような環境においては，コミュニケーションの手段として用いられる音声データの有効活用が，今後課題となってくると考えられます。

音声データは，そのままでは内容を一覧したり検索したりすることは困難ですが，これを音声認識技術によりテキスト化したり注釈をつけたりすることで，テキスト情報との連携などの様々な応用が可能となります。しかしここで扱う音声は，これまでの音声認識が対象としてきた，人が機械に向かって丁寧に話しかける単語発声ではなく，主に人と人とが会話をしている「話し言葉」であり，その認識には高い技術が必要となります。

NECは，このような話し言葉の音声認識の技術開発を進めています。本稿では，話し言葉認識の実用化に向けて開発中のスケーラブル大語彙連続音声認識技術と，その応用事例について紹介します。

2．スケーラブルな大語彙連続音声認識フレームワーク

話し言葉の認識では，様々な単語や表現を含む文発声を扱う，大語彙連続音声認識が必要となります。大語彙連続音声認識は一般に多くの処理能力やメモリなどのリソースを必要としますが，サーバの豊富なリソースを用いて高精度な認識を行う用途だけでなく，モバイル環境で使用する携帯端末上でローカルに認識する用途や，1台のサーバで多回線を同時に処理する用途もあります。そこでNECは，サーバなどではリソースに応じてより高精度な認識を行うとともに，PDAクラスの端末でも動作する，コンパクトでかつスケーラブルな大語彙連続音声認識フレームワークを開発しました[1)]。

大語彙連続音声認識は，**図1**に示すように，声の見本をモデル化した「音響モデル」を用いて入力音声との「距離計算」を行い，その結果と「単語辞書」，語の並びの規則をモデル化した「言語モデル」を用いて，無数にある単語の組合せから入力された音声に最も近い単語列を効率よく探し出す（「最適単語列探索」），という処理を行います。以下，それぞれの技術について述べます。

＊ メディア情報研究所
Media and Information Research Laboratories

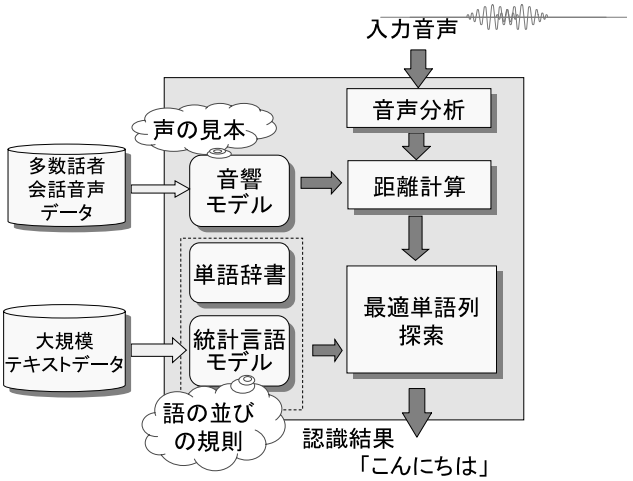

図1　大語彙連続音声認識の構成
Fig.1　Structure of large vocabulary continuous speech recognition system.

2.1　音響モデル・距離計算

入力音声信号は，音声分析により1/100秒程度の一定時間ごとに計算される，特徴ベクトルの時系列に変換されます。音響モデルは声の見本を多数の確率分布の集合により表現しており，距離計算では特徴ベクトルの出力確率を各分布に対し計算します。音響モデルの使用メモリ量と距離計算の計算量を削減するために，①記述長最小基準に基づく効率的な分布数削減[2]，②対角共分散行列の共有化による分布の簡易化，③分布の木構造化に基づく出力確率の高速計算[3]の3つの手法を開発・導入しています。

2.2　単語辞書・言語モデル

対象分野の大規模なテキストデータから，単語辞書と統計言語モデルを構築します。言語モデルとしては，単語のn個組の連鎖確率であるn-gramモデルを用いています。利用可能なメモリ量や処理能力に応じて単語2-gram，クラス2-gram，単語3-gramを組み合わせています。クラスは品詞をベースに，対象分野に応じて意味的なクラスや自動クラスタリングにより細分化して用いています。

2.3　最適単語列探索

最適単語列探索は距離計算結果と言語モデルを用いて，入力音声に同期して可能性の高い候補のみに絞り込みながら，辞書中の単語との照合を行います。辞書は先頭の共通部分を束ねることで木構造化して圧縮した表現で保持し，動的に展開して用います。一定間隔ごとのワークメモリのガベージコレクションや，言語モデルの計算結果の再利用などにより，メモリ量，計算量を削減しています。

3．応用事例

3.1　コンタクトセンタ向け音声認識

コンタクトセンタにおける，オペレータ通話音声を認識するシステムを開発しました。通話音声をテキスト化することによって，ナレッジ検索キーワード入力，応対記録作成などのオペレータ業務効率化や，特定単語検出によるリアルタイム状況検知，モニタリング業務での通話内容確認などのスーパーバイザ業務支援が可能になります（**図2**）。

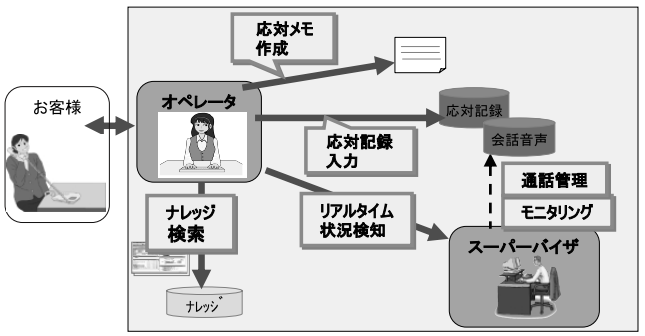

図2　コンタクトセンタへの応用
Fig.2　Contact center application.

通話音声は人どうしの自然な話し言葉ですので，発音があいまいになったり，発声速度が大きく変化するなど，様々な音響的な変動が生じます。本システムでは多人数の長時間の通話音声を用いて，話し言葉の発音変動や多様性を高精度に表現する音響モデルを開発し，用いています。また，個々のオペレータの声の特徴を学習し，認識性能をさらに向上することも可能です。

さらに，各コンタクトセンタでの言い回しや用語に合わせた言語モデルや辞書を作成することにより，認識率を向上させます。ここでは，テキストマイニングや機械翻訳で培ったテキスト解析技術を拡張することにより，「えー」「あのー」などの付加語や，「～していただいてもよろしかったでしょうか」のような話し言葉特有の表現に対応することが可能になりました。

3.2　電話大語彙姓名認識システム[4]

電話音声自動応答によるチケット予約システムなどを実現するために，日本人の姓名を認識する電話大語彙音声認識システムを開発しました。約10,000種類（読みの異なり数）の姓と約5,000種類の名を認識辞書に持ち，人口の約98%をカバーしています。出現頻度による重み付けをした照合を行うことにより，処理を効率化しています。音節単位に区切った発声や，母音を伸ばした強調発声も受理できるように，辞書エントリに読みを加えています。

3.3　携帯電話マニュアルの音声検索システム[5]

最近の携帯電話端末はメール，ブラウザに加えてカメラなど豊富な機能を有しています。そこで外出先でもその操作マニュアルを簡単に検索・参照できるようにするシステムを開発しました（**図3**）。

ユーザは電話音声認識サーバに電話をかけて，音声で携帯電話端末の操作に関する質問を行います。サーバは音声認識結果を用いて電子化された端末操作マニュアルを検索し，得られた結果候補をWebページとして電話端末上に表示します。ユーザは端末画面上で複数の検索結果候補から，目的の情報を選んで参照することができます。

図3　音声検索システム
Fig.3　Speech-activated text retrieval system.

3.4　講演・放送音声認識システム

蓄積されたビデオアーカイブを音声認識して，注釈を付与するシステムを開発しました。システムはビデオの音声トラックから音声区間を抽出し，音声認識を行います。認識結果テキストと音声区間の話者判別結果を組み合わせて音響モデルの適応化を行い，適応後の音響モデルを用いて再度認識を行うことで，認識精度の向上を図っています。

得られた認識結果テキストは，ビデオに対する注釈として検索や要約[6]に利用することができます。応用例として，講演ビデオの音声の認識結果テキストと，講演に使用したスライドのテキスト情報の対応付けを取ることで，ビデオとスライドを同期表示する閲覧システムを開発しました[7]。

3.5　旅行会話向け自動通訳PDA[8]

コンパクト化によりリソースの少ない環境でも動作する特徴を生かし，PDA向けの日英双方向の旅行会話自動通訳システムを開発しました。旅行会話文約10万文を収集して辞書と言語モデルを構築しました。認識辞書のサイズは日本語5万語，英語2万語です。1言語当たりの音声認識モジュールの使用メモリは，起動時で約7Mバイト，処理中のワークメモリは約1Mバイトに抑えられています。システムは日・英音声認識，日英・英日翻訳，日・英音声合成を統合してPDA（CPU StrongARM 206MHz，メモリ64Mバイト，メモリカード128Mバイト）上で動作しています。

4．むすび

本稿では，話し言葉認識に向けた大語彙連続音声認識技術と，コンタクトセンタなどいくつかの応用事例について紹介しました。今後はブロードバンドオフィスにおける電話や会議の音声の活用のための注釈付与やテキスト化，海外とのTV会議の通訳によるコミュニケーション支援の実現などをめざし，さらに技術開発を進めていく予定です。

＊StrongARMは，ARM Ltd.の登録商標です。

参考文献

1）　石川ほか；「コンパクトなディクテーションの開発」，音学講論，3-5-12，2002-3.
2）　篠田ほか；「音声認識のためのMDL基準を用いた効果的なガウス数削減」，信学技報，SP2001-83，2001-10.
3）　Watanabe, et al.；“High Speed Speech Recognition Using Tree-Structured Probability Density Function”，ICASSP-95，pp. 556-559，1995.
4）　三木ほか；「大語彙姓離散発声電話音声認識の検討」，音学講論，3-Q-11，2003-3.
5）　安達ほか；「携帯電話向け音声/Web連動型検索システム」，音学講論，2004-3.
6）　中澤ほか；「ビデオ音声認識テキストからの文認定」，言語処理学会年次大会，2002-3.
7）　中澤ほか；「談話指標とテキスト長を利用した講演音声-プレゼン資料アライメント」，FIT2003，2003-9.
8）　山端ほか；「PDAで動作する旅行会話向け日英双方向音声翻訳システム」，情処研報，NL-150，2002-7.

筆者紹介

Ryosuke Isotani
磯谷　亮輔（いそたに　りょうすけ）　1987年，NEC入社。現在，メディア情報研究所主任研究員。電子情報通信学会，日本音響学会各会員。

Kaichiro Hatazaki
畑崎香一郎（はたざき　かいちろう）　1982年，NEC入社。現在，メディア情報研究所主任研究員。情報処理学会，電子情報通信学会，日本音響学会各会員。

Hiroaki Hattori
服部　浩明（はっとり　ひろあき）　1985年，NEC入社。現在，メディア情報研究所研究部長。工学博士。電子情報通信学会，日本音響学会各会員。

Akitoshi Okumura
奥村　明俊（おくむら　あきとし）　1986年，NEC入社。現在，メディア情報研究所研究部長。工学博士。人工知能学会理事，情報処理学会，言語処理学会各会員。

Takao Watanabe
渡辺　隆夫（わたなべ　たかお）　1974年，NEC入社。現在，研究企画部/メディア情報研究所エグゼクティブエキスパート。工学博士。電子情報通信学会，日本音響学会，IEEE各会員。