社団法人　情報処理学会　研究報告
IPSJ SIG Technical Report

2005－ FI － 80（16）
2005－ NL － 169（16）
2005／9／30

話し言葉認識に向けた基本技術と応用

磯谷 亮輔　畑崎 香一郎　服部 浩明　奥村 明俊　渡辺 隆夫

NEC メディア情報研究所

あらまし　音声認識の対象は，コマンドや読み上げ文から，自然な会話調の文発声や人が人に向かって話しかけている音声など，いわゆる「話し言葉」へと広がってきている．話し言葉認識の基盤となるのは大語彙連続音声認識技術である．大語彙連続音声認識は一般に多くの計算量とメモリを必要とするが，携帯端末で処理を行いたい場合や，１台のサーバで多回線の処理を行いたい場合などには，リソースの削減が必要となる．そこで我々は，サーバから PDA クラスの端末までリソースに応じて動作するスケーラブルな大語彙連続音声認識のフレームワークを開発した．本稿では，その基本技術と，さまざまな応用事例を紹介する．

Basic Technologies for Spontaneous Speech Recognition and Its Applications

Ryosuke Isotani, Kaichiro Hatazaki, Hiroaki Hattori,
Akitoshi Okumura and Takao Watanabe

Media and Information Research Laboratories, NEC Corporation

Abstract　The targets of automatic speech recognition are now expanding from voice commands and read speech to “spontaneous speech” such as naturally spoken utterances and human-to-human communications. The basis of spontaneous speech recognition is the large vocabulary continuous speech recognition (LVCSR) technology, which generally requires much computational capacity and memory. We have developed a scalable LVCSR framework which works on PDA-class terminals as well as sever computers depending on their available resources. It enables LVCSR on mobile terminals and multi-channel processing on a PC server. This paper presents the basic technologies of our LVCSR and its applications.

1.　はじめに

音声認識の対象は，離散単語や定型文によるコマンド入力から読み上げによるディクテーション，さらに近年は，機械との自然言語文による対話や，講演など人が人に向かって話しかけている音声など，いわゆる「話し言葉」の認識へと広がりつつある．話し言葉認識の基盤となるのは大語彙連続音声認識技術である．大語彙連続音声認識は，混合連続分布 HMM によるトライフォン音響モデル，木構造単語辞書，N-gram 統計言語モデル，フレーム同期ビームサーチを含む複数パスサーチ，などの技術を組み合わせて実現されている[1]．しかし，これらの方式は処理に高速 CPU と大容量メモリを

必要とする. 実用化において携帯端末や多回線同時処理を行うサーバなどに搭載するためには大幅なコンパクト化が必要である．そこで我々は組み込み向けからサーバ用途まで, 対象システムの要件にあわせてスケーラブルにコンパクト化が可能な大語彙連続音声認識フレームワークを開発した[2]．本稿ではそのフレームワークと, いくつかの応用事例について説明する.

2. スケーラブルな大語彙連続音声認識フレームワーク

2.1. 音響モデル

特徴ベクトルとしてメルケプストラムとエネルギーおよびそれらの時間変化量を LDA 変換により次元圧縮して使用している. 定常な加算性および乗算性の歪を抑圧するためにスペクトルサブトラクションとケプストラム平均正規化を行っている. 音響モデルは混合ガウス分布を出力確率とするトライフォン HMM で, 音素環境の決定木による状態クラスタリング法を用いて状態を共有化している. 音響モデルの使用メモリ量と計算量を削減するために, 以下の 3 手法を用いている.

2.1.1. MDL 規準に基づく混合ガウス分布数削減 [3]

認識率の劣化を抑えつつ音響モデルのサイズを縮小するために，記述長最小（MDL）基準を用いて出力確率の混合ガウス分布から冗長なガウス分布を削減した. その手順は以下のとおりである.

(1) 学習データを用いて混合ガウス分布 HMM を学習し，各ガウス分布の十分統計量（平均ベクトル, 共分散行列, 事後確率）を算出する.
(2) HMM の状態ごとに要素ガウス分布を木構造にクラスタリングする．ここではガウス分布間の（対称化）KL 距離尺度を用いてトップダウンに 2 分木を作成した．リーフノードが元のガウス分布に対応する.
(3) 各 2 分木の中間ノードに対して, そのノードに属するリーフのガウス分布の十分統計量を合算して, 中間ノードの十分統計量とする.
(4) ルートノードからはじめて, 記述長の減少が止まるか, リーフノードに到達するまで順番に下位のノードへ分割を続ける.
(5) 上記で得られた分割後のノード集合のガウス分布を各状態の混合ガウス分布として，学習データを用いて再推定する.

(4)の分割例を図 1 に示す．図 1 ではルートノード 0 が子ノード 1 と 2 に分割され，さらにノード 2 が 5 と 6 に分割されている．ノード 1 はこれ以上記述長の減少がないため分割されていない．最終的にはノード集合 {1,5,6} が得られ，最初の混合数 4 個が 3 個に削減されたことになる. なお上記アルゴリズムにおけるノード分割前後の記述長減少量は，ノード i が, ノード j と k に 2 分割された場合に次式であらわされる.

$$\Delta_{i\to j+k} = \frac{1}{2}\left\{\Gamma_j \log|\Sigma_j| + \Gamma_k \log|\Sigma_k| - \Gamma_i \log|\Sigma_i|\right\} + \alpha K \log N$$

ここで $\Gamma_i, \Gamma_j, \Gamma_k$ が各ノードの事後確率,

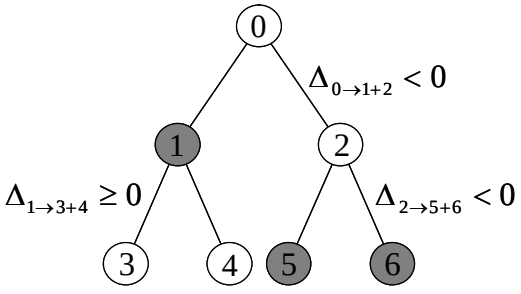

図 1 MDL 基準による混合ガウス分布数の削減

$\Sigma_i, \Sigma_j, \Sigma_k$ が共分散行列，K は特徴ベクトルの次元数，N はデータ数を表す．上式第 2 項の定数 α は，ノード分割を制御するために経験的に導入したパラメータである（α を小さくすると分割が進み，大きくすると抑制される）．以下においてはいくつかの α を試みて，所望の混合数に削減された結果を使用している．

2.1.2. ガウス分布の対角共分散行列の共有化

特徴ベクトルの次元圧縮に用いた LDA 変換は，クラス内分散行列とクラス間分散行列を同時対角化する（クラス内分散行列は白色化，すなわち恒等行列化される）．得られたクラス間分散行列の固有値が小さい成分を除去することにより，クラス分離度の低下を抑えて次元圧縮を行うことができる．ここで HMM の各ガウス分布をクラスとして選んだ場合，クラス内分散行列は各ガウス分布の共分散行列平均値にほぼ一致する．そのため LDA 後に各ガウス分布の共分散行列を恒等行列に共有化しても顕著な認識率低下が生じないことが期待される．

これにより，音響モデルのパラメータ数を約 1/2 に削減することができる（平均ベクトルと対角共分散行列を両方保持する場合に比べて，後者が不要になるため）．また，ガウス分布確率値計算を，入力特徴ベクトルとガウス分布平均ベクトルのあいだのユークリッド距離計算へ還元できるため，計算量も削減できる．

2.1.3. ガウス分布木構造化に基づく高速計算 [4]

フレーム同期ビームサーチにおけるガウス分布確率値計算を高速化するために，音響モデルに含まれる全ガウス分布を木構造状にクラスタリングする．リーフノードは HMM の各状態のガウス分布に対応し，中間ノードは下位のノードを被覆するガウス分布となっている．入力特徴ベクトルに対して，ルートノードからリーフノードに向かってガウス分布確率値計算を行う中で，確率値の大きいノードから一定個数（上位 n 個）をリーフノードへ向かって計算し，それ以外のノードの確率値は上位ノードの値で代用することにより計算量を削減している．

たとえばノード分岐数（子ノードの数）が N 個の 3 階層木構造とすると，ガウス分布確率値の計算回数は $N + nN + nnN = N\left(1 + n + n^2\right)$ 回となり，リーフの全ガウス分布確率値を計算する回数 N^3 回に比べて大幅削減が可能となる．さらにリーフノードでは確率値が計算されなかったガウス分布に対しても，親ノードの確率値を近似値として代用できるため，定数値などを用いる場合に比べて，近似精度の向上が期待できる．

ガウス分布確率値計算の高速化法としては，本手法以外にも Gaussian Selection 法[5]などが知られている．これらに比べて本手法は，高速計算のために作成したガウス分布の木構造を話者適応化[6]にも使用することができるという点で，限られたリソースにシステムを収めたい場合に有効である．

2.2. 言語モデル

言語モデルは利用可能なメモリ量に応じて単語 bigram，クラス bigram，単語 trigram を組み合わせている．クラスは品詞をベースに，タスク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る．またサイズ削減のために，N-gram 確率の対数値を 256 段階（1byte）に量子化して格納している．

2.3. サーチ

サーチは 2 パスで構成されている．1 パス目は木構造単語辞書を用いたフレーム同期ビームサーチを行っている．発声の始端に対しては，直前の発声の認識結果を言語的コンテキスト

として与えて仮説を絞り込んでいる. コンパクト化のために木構造単語辞書は音素をノードとし, 各ノードを 1byte で表現している. 各ノードはサーチ時に動的にトライフォンに展開される. 単語間に対してもトライフォンを使用している. 各フレームで単語終端に到達した仮説を単語終端テーブルに書き出し, 発声終端においてワードグラフとして結果を出力する.

単語終端テーブルを一定フレーム間隔ごとにガベージコレクションすることにより, 発声長に依存したワークメモリの増加を抑えている. また処理量削減のため, 単語終端における言語スコア計算結果の再利用を行っている. すなわち連続するフレームで同じ言語スコア計算が繰り返し行われることに着目し, 一度計算した言語スコアを保持して再利用している. とくに, 一つの仮説が複数の言語コンテキストに対応する場合, 単語終端ですべてのコンテキストに対し, 言語スコア計算, 累積スコア計算を行い, その最良値を選択する必要があるが, フレームが違っても異なるのは現単語の音響スコアのみであるため, それを除いた値を保持して再利用することで, 全コンテキストに対する処理は一度で済む. これにより, たとえば先行単語が無音である仮説を木構造辞書のルートノードで一つにマージしてサーチし, 単語終端で無音の前の単語コンテキストを考慮した展開を行うような場合の処理が効率化される.

2 パス目では, 発話ごとに得られたワードグラフをつなげてリスコアし, 最尤パスを認識結果として出力する.

2.4. コンパクト化評価実験

語彙 5000 語の文入力タスクでコンパクト化の評価実験を行った. 新聞記事 2 年分(約 90M 単語)を言語モデルの学習コーパスとして用い, 認識辞書は同コーパスの頻度上位 5000 単語で構成した. 評価データはこの 5000 単語でカバーされる文を, 男性話者 6 名が文節程度に区切って発声した読み上げ発声である.

2.4.1. 音響モデル

上記評価条件で性別依存音響モデルをコンパクト化して評価した. 言語モデルは後述の b300 に固定し, 話者適応は行っていない. 結果を表 1 に示す.

ベースライン（A0）の音響モデルは状態数 1000 状態, 混合ガウス分布数 8000 個の状態共有トライフォン HMM で, 男性 1000 名の約 15 万文から学習した（特徴ベクトルは 23 次元）. MDL 基準を用いた混合ガウス分布数削減(A1)ではガウス分布数を 8000 個から 6000 個に削減している.（A3）でそれを 3 階層（各ノードの最大分岐数 20）の木構造にクラスタリングした. ガウス分布確率値計算では第 1 層のノードのうち確率値が大きい順に上位 4 個のノードの子ノード（第 2 層）の確率値を計算し, さらにその上位 6 個のノードの子ノード（第 3 層）の確率値だけを計算した. これによりガウス分布の確率値計算回数の最大値は 580 回となる. 結局（A0）に比べてわずかに認識率の劣化はあるものの, 保持すべきパラメータ数を約 1/3 に, 確率値計算回数を約 1/10 以下に削減することができた.

表 1 不特定話者音響モデルのコンパクト化

		単語正解精度 (%)	パラメータ数（個）	確率値計算回数
A0	ベースライン	96.4	368000	8000
A1	A0+MDL基準による混合ガウス分布数削減	96.2	276000	6000
A2	A1+ガウス分布の共分散行列共通化	95.5	138000	6000 (*)
A3	A2+ガウス分布の木構造化	95.0	115000	580 (*)

(*)1 回の計算が分散を考慮しないユークリッド距離計算に軽減.

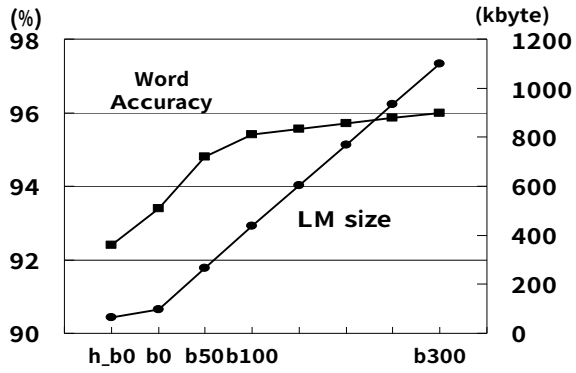

図 2 言語モデルのサイズと単語正解精度

2.4.2. 言語モデル

ここではクラス bigram に比較的少数の高頻度単語 bigram を追加して用いている．品詞をクラスとしているが，一部の品詞については LBG アルゴリズムを用いた自動クラスタリングによってクラスを細分化している．自動クラスタリングの効果，および単語 bigram の個数と単語正解精度の関係について評価した．図 2 において，h_b0 は品詞に基づくクラスを用いたクラス bigram，b0 は自動クラスタリングを用いて品詞クラスの「名詞」「動詞」をそれぞれ 32 クラスずつに分割したクラス bigram であり，いずれも単語 bigram は持たない．b50，b100，b300 は，b0 に頻度上位の単語 bigram をそれぞれ 5 万個，10 万個，30 万個追加したものである．

評価データ（男性 6 名，632 文）の単語 perplexity を表 2 に示した．音響モデルは A3 に 10 分程度の発声による話者適応を行ったものを用いた（話者適応により表 1 の A3 95.0%が図 2 の b300 96.0%に改善した）．

表 2 評価データの単語 perplexity

言語モデル	単語perplexity
h_b0	199.2
b0	161.5
b50	91.5
b100	80.9
b300	74.8

2.4.3. サーチ

サーチの 1 パス目では文節程度に区切って発声された各発話を別々に処理し，各発声始端には直前発声の認識結果を言語的コンテキストとして与えている．得られた各発話のワードグラフを 1 文単位に連結して，2 パス目では同じ言語モデルを用いてリスコアリングを行っている．文の区切りは人手で与えて実験した．

サーチにおけるガベージコレクションの効果を評価したところ，ワークメモリ使用量が約 4 割削減された（0.8Mbyte から 0.5Mbyte に低減）．また単語終端での言語スコア計算結果の再利用により言語スコア計算回数は 1/35 に低減された．

2.4.4. 統合結果

以上を統合して評価を行った．言語モデルは b300，音響モデルは A3 に 10 分程度の発声による話者適応を行ったものを用いた．結果のグラフを図 3 に示す．これは単語正解精度と実行時間の発声時間に対する比を，サーチパラメータを変えてプロットしたもので，実行時間は CPU StrongARM 206MHz の PDA 上で計測したものである．

結果として，PDA 上の実時間処理で単語正解精度 91.6%であった．実時間の数割増で単語正解精度 96%程度まで改善できることから，今後機器の処理能力が向上した場合に認識率

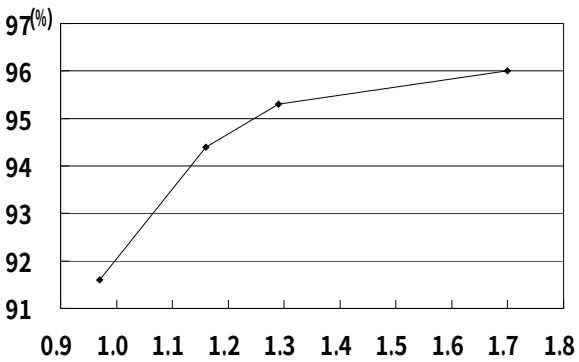

図 3 実時間比と単語正解精度

向上を期待できる. いずれのサーチパラメータでも, サーチに使用する動的なワークメモリは 0.5Mbyte 弱であり，静的なメモリ使用量は言語モデル 1.1Mbyte，認識辞書 0.2Mbyte，音響モデル 0.7Mbyte，ほか実行バイナリなど，すべてあわせて 3.5Mbyte である. 後者は読み出し専用メモリに保持することが可能なため，比較的小さな読み書き可能メモリしか持たない機器への搭載も可能と考えられる.

3. 応用事例

PDA 単体で動作するものからサーバで処理するものまで種々の応用システムを開発した.

3.1. 旅行会話向け自動通訳 PDA [7]

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構築した（図 4）. 音声認識用言語モデルは旅行会話用に収集したテキストコーパス（約 10 万文）を用いて作成した. クラス bigram，単語 bigram に加えて単語 trigram も使用している. 認識辞書は, 言語モデル学習コーパスに出現する単語と一般に利用頻度の高い単語から構成した. 語彙サイズは日本語 5 万語，英語 2 万語である. サーチの 1 パス目はクラス bigram と単語 bigram を使用し，2 パス目に単語 trigram を用いてリスコアリングを行った. 入力音声が文単位の発話（区切りなし）のため，リスコアリングは発話単位で行っている. 音声認識モジュールの使用メモリは，起動時で約 7Mbyte，また処理中のワークメモリは約 1Mbyte に抑えられた（1 言語あたり）. 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 64Mbyte，メモリカード 128Mbyte）上で動作している.

図 4 旅行会話向け自動通訳 PDA

3.2. 携帯電話マニュアルの音声検索 [8]

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る. 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試作した（図 5）.

ユーザー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う. 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

図 5 音声検索システム

結果候補を WEB ページとして電話端末上に表示する. ユーザー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る.

この応用例では, すべての対話を音声だけで行うのではなく, 検索結果候補は閲覧性の高い画面表示で提示することによって, 音声認識や自然言語検索の誤りの影響を抑えて, 快適なインターフェースの実現を試みている.

3.3. AV コンテンツの検索

蓄積された AV コンテンツのアーカイブを音声認識してアノテーション情報を付与し, 検索するシステムを試作した. 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, 不特定話者音声認識を行う. 認識結果テキストと音声区間の話者クラスタリング結果を組み合わせて教師なし話者適応化を行い, 適応後の音響モデルを用いて再度認識を行う. アノテーション情報は, 認識結果テキストと対応する時間情報からなる. キーワードを入力すると認識結果テキストに対し検索を行い, 検索された箇所に対応する時間情報をもとにビデオを再生する.

3.4. ビデオとテキスト情報の同期閲覧 [9]

AV コンテンツの認識の別の応用例として, ビデオ音声の認識結果テキストと, そのビデオに関連するテキスト情報をアライメントして, ビデオとテキスト情報を同期表示する閲覧システムを試作した. コンテンツとしては講演音声と講演に使用したスライドを用いた. そこでは認識結果テキスト中の連続する複数の文と, 1 枚のスライド中の単語を語順自由で対応付けしている. また認識結果中の「さて」,「次に」などのようにスライドの切り替わりに特有な表現（談話指標）がスライド境界候補になるようにマークしてアライメント精度を高めている.

3.5. 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を可能としている（図 6）.

話し言葉の発音変動など人間同士の通話音声に現れるさまざまな音響的な変動に対処するため, 本システムでは多人数の長時間の通話音声を用いて音響モデルを学習している. また, 個々のオペレータについて話者適応を行うことも可能である.

さらに, 各コンタクトセンタでの言い回しや用語に対応するため, 通話音声を書き起こしたテキストを用いて言語モデルおよび辞書を作成している. 書き起こしは,「えー」「あのー」等の付加語や「～して頂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を含む. このようなテキストから言語モデルを学習するため, 話し言葉対応を強化した形態素解析エンジンを開発し, 使用している.

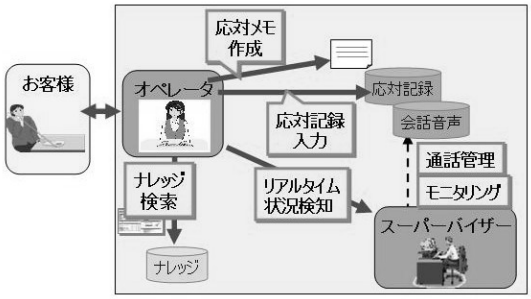

図 6 コンタクトセンタへの応用

4. まとめ

話し言葉認識の基盤として, 携帯端末などの組み込み向けからサーバ用途まで幅広く適用可能な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た. さらに, 本フレームワークを用いた種々の応用システムを開発することで, その実用性を検証した.

今後さらに, 各種タスクへの展開を容易とする環境を構築, 整備していくことにより, より多くの実用場面に適用していきたい.

参考文献

[1] 河原達也 他, "連続音声認識コンソーシアムの活動報告及び最終版ソフトウェアの概要," 情処研報, 2003-SLP-049-057 (2003.12).

[2] 石川晋也 他, "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).

[3] 篠田浩一 他, "Efficient reduction of Gaussian components using MDL criterion for speech recognition," 信学技報, SP2001-83 (2001.10).

[4] T. Watanabe et al., "High speed speech recognition using tree-structured probability density function," ICASSP-95, pp.556-559 (1995).

[5] E. Bocchieri, "Vector quantization for the efficient computation of the continuous density likelihoods," ICASSP-93, pp.692-695 (1993).

[6] 篠田浩一 他, "音声認識における自律的なモデル複雑度制御を用いた話者適応化," 信学論誌, J79-D-II, No.12 (1996.12).

[7] 山端潔 他, "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, 2002-NL-150 (2002.7).

[8] 安達史博 他, "携帯電話向け音声/Web 連動型検索システム," 音学講論, 1-8-20 (2004.3).

[9] 中澤聡 他, "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).