論文 / 著書情報
Article / Book Information

論題(和文)	大語彙連続音声認識の現状と展望
Title(English)	
著者(和文)	古井貞熙
Authors(English)	SADAOKI FURUI
出典(和文)	日本音響学会 1998年春季講演論文集, Vol. , No. 1-6-10, pp. 19-22
Citation(English)	, Vol. , No. 1-6-10, pp. 19-22
発行日 / Pub. date	1998, 3

スペシャルセッション〔音声認識のための統計的言語モデル〕（招待講演）

1—6—10　大語彙連続音声認識の現状と展望*

○古井　貞熙（東工大）

1．まえがき

音声認識技術の応用場面を考えてみると、数十語程度の少数語彙で、しかも単語入力ですむなら、キーやタッチパネルの方が容易で確実な場合が多い。言い換えると、大語彙で連続音声が認識できる技術が確立して初めて、音声認識技術の用途が大きく広がると予想される。このような期待と、データベースの拡充や計算機の発達に支えられて、近年、大語彙連続音声認識の研究が大きく進展している。その研究の流れは、ディクテーションと対話（会話）音声認識の認識・理解を対象とした研究に分けられる。前者は書き言葉に近い音声を発声し、すべての言葉を文字に変換するものである。認識対象は文法的にはほぼ正確であると想定することができるが、認識しなければならない語彙や文体が極めて多様であるという点に難しさがある。いわゆる音声ワープロがこれに含まれる。後者の対話音声の場合には、対象（タスク）をある分野の情報案内などに限定すれば、ある程度語彙を制限することができるが、話し言葉特有のあいまいな文章や、文法的に誤った文も対象としなければならない難しさがある。「大語彙」の定義は決まっていないが、ここでは語彙数が2,000語以上の研究を取り上げる。その中で、世界最大のプロジェクトは、米国のARPA（Advanced Research Projects Agency）によるものであり、近年の音声認識研究の牽引車となっている。

2．大語彙連続音声認識システムの基本的構成

図1に、典型的な大語彙連続音声認識システムの構成を示す[1]。最近のシステムでは、特徴パラメータとして（メルあるいはLPC）ケプストラム係数とその1次および2次の回帰係数を用い、音素モデルとしては、音素間の調音結合を考慮した混合ガウス分布型トライフォンHMM [2] を用いる。このトライフォンHMMは、大量の音声から学習して作成するが、パラメータの自由度を減らして推定精度を上げるため、異なるモデル間で類似した状態を共有化する。音素モデルの話者適応の方法として、MAP推定[3]やMLLR法[4]を用いた教師なし適応がよく用いられる。

言語処理には、単語のバイグラムやトライグラムのような統計的言語モデルを用いることが多い。言語モデルも大量なテキストを用いて学習するが、テキスト量の不足を補うため、バックオフ平滑化を行う。いきなりトライグラムを用いると計算量が膨大になってしまうため、バイグラムまでを用いて最も可能性の高い多数の文仮説（N-best仮説）あるいは単語ラティスを作成し、これにさらにトライグラムを適用して、仮説のスコアを計算し直すという方法（2-pass decoder）をとる。このスコアの再計算の際に、音素モデルとしてさらに高度なモデル（適応化や単語間の調音結合を考慮）に置き換える方法も用いられている。

認識性能の評価には、（総単語数から置換、脱落、挿入の誤りを引いた数）／総単語数×100%＝単語認識率（正解精度）が用いられる。

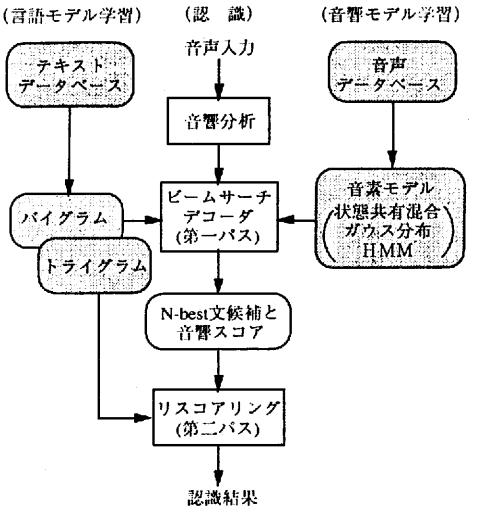

図1．典型的な大語彙連続音声認識システムの構成

3．新聞記事読み上げタスク

ARPAプロジェクトの内、ディクテーションについては、Wall Street Journal新聞を中心として、北米の種々の経済新聞（NAB：North American Business News）の記事を読み上げた文章を認識するプロジェクト（Hub3）が強力に進められた。65,000語の語彙を対象にした場合、93%の単語認識率が得られている [5]。

日本における大語彙連続音声認識の研究としては、Wall Street Journalに似ていると考えられる日経新聞の記事を読み上げた文章を認識する研究が行なわれている [6][7][8]。42種類の音素に対して構築されたトライフォンの総状態数は2106、1状態あたりの混合数は4である。これらの構築には、男性53名による13,270文の音声を用いた。特徴パラメータベクトルはLPCケプストラムと（正規化）対数パワーおよびそれらの1次微係数の計34次元である。約5年分の新聞記事（のべ約750万文、1億8000万語）を用いて、形態素解析プログラムによりテキストを形態素単位に区切り、形態素を単位とする統計的言語モデルを作成した。頻度の高い7,000語（形態素）を対象としたトライグラムに基づく言語モデルと、トライフォンのHMMを用いた認識実験で、90%の

* Recent advances and perspectives of large-vocabulary continuous-speech recognition
By Sadaoki Furui (Tokyo Institute of Technology)

単語（形態素）認識率が得られている。図2に、種々の条件における単語認識率を示す。統計的言語モデルが、日本語に対しても極めて有用であることが分かる。

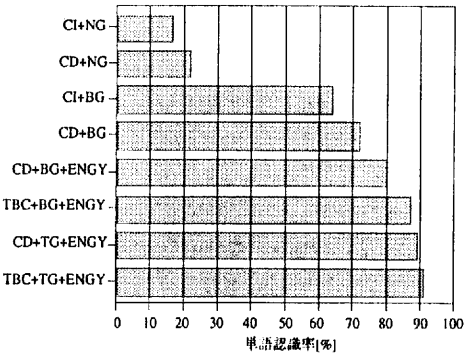

図2: 種々の条件における新聞記事読み上げタスクの単語認識率
CI: 音素文脈独立音響モデル
CD: 音素文脈依存音響モデル
NG: 言語モデルなし
BG: bigram言語モデル　　TG: trigram言語モデル
ENGY: 音響特徴量にパワー情報を追加
TBC: tree-based clusteringによる音響モデル

その後日本語に関しても、「日本人が潜在的に意識している」単語を単位とする統計的言語モデルを用いた研究が行なわれ[9]、この方が単位当たりの音素数が多いので音響処理的に有利であり、言語処理能力も変わらないと報告されている。また最近、毎日新聞を用いた日本音響学会新聞記事読み上げ音声コーパスが構築され[10]、今後日本における関連研究の共通データベースとして使われると期待されている。

4．ニュース音声認識タスク

4.1 ARPA の Hub4 タスク

ARPA プロジェクトの現在の中心は、Hub4 と呼ばれる放送ニュースのディクテーションである[11][12][13]。上記の新聞読み上げタスクとの基本的違いは、対象とする音声が 'real-world speech data' であることである。この技術が確立すれば、実際の放送に自動的に字幕を付けることができると期待されている。表1に示す7種類の実験条件（*focus conditions*）に分けられており、native のアナウンサーが原稿を元に静かな環境で発声した音声を認識するF0が最も基本的な条件である。各条件に分けられている音声を対象とするPE(*Partitioned Evaluation*)と、分けられていない UE(*Unpartitioned Evaluation*)の評価があり、前者は必須、後者はオプションである。

PE に参加しているのは、BBN、Cambridge 大の Connectionist と HTK グループ、CMU、IBM、LIMSI、Rutgers 大、および SRI である。NYU は SRI の音声認識結果のリスコアリングという形で参加している。UE に参加しているのは、BBN、CMU、および IBM である。後者については、音楽・雑音・音声が重畳したり連続している音を、tied mixture HMM を用いて、きれいな音声、音楽重畳音声、電話系音声、

表1．1996 Hub4における実験条件

F0：ベースライン放送音声（スタジオで録音した原稿のある音声）

F1：対話放送音声（スタジオで録音した対話音声）

F2：電話系音声（帯域制限された音声）

F3：背景音楽中の音声（SNRがA特性で10～20dBの原稿あり／なし音声）

F4：劣化音響条件下の音声（SNRがA特性で10～20dBの、加算性雑音、環境雑音、あるいは非線形歪みのある原稿あり／なし音声）

F5：non-native話者の音声（米語のnon-native話者—英国のnative話者の英語を含む—のスタジオ品質の音声）

FX：その他（上に属さないものや、上の2つ以上の条件を満たすもの—例えば、背景音楽中のnon-native話者の音声—）

雑音、音楽に分け、異なる音響モデルを用いる研究が行なわれている。LIMSIでは、セグメントベースのケプストラムの平均値と分散の正規化、filler words と breath noise のための音素モデルの追加などを行っている。Cambridge 大の HTK グループでは、特徴パラメータとして MF-PLP（メル周波数フィルタバンクを用いた PLP）を用いて、ロバスト性を高めている。

次の3種類の音響データベース（上の F0 から F5 をカバー）が NIST の協力を得て LDC から配付されている。なお、学習には1996年6月末以前のあらゆる公開データベースを用いてよいことになっている。

・学習用：11のニュース番組（ABC, CNN, CSPAN, NPR, PRI）約50時間の音声（130時間の音声から選んで書き下し、annotation 付き）
・開発用：1996年7月に録音された3時間のラジオとテレビの音声
・評価用：1996年9月に録音された2.5時間の放送ニュース音声（学習および開発用と番組の一部重なりあり。スポーツと CM は除外。）

テキストデータベースは、110の放送（ABC, CNN, NPR, PBS）から抽出した1億2200万語（学習用）および1900万語（開発用）のセットである。標準の言語モデルはないが、CMU で開発されたツールが広く使われている。トライグラムモデルの test-set perplexity は表1の各条件によって異なるが、平均150（Cambridge 大）～180（LIMSI）である。Cambridge 大では4グラムまで用いており、このときの test-set perplexity は平均140である。

認識性能の評価は、NIST の評価プログラム SCLITE[14]で行なわれている。語彙数は65,000語が標準的（未知語率0.66～0.74%）で、PE の場合、全評価セットに対して46～73%、F0 に対しては57～81%の単語認識率が得られている。同じ研究機関の PE と UE の結果の差は極めて小さく、UE でも PE と同程度の結果が得られることを示している。

4.2 日本における研究状況

日本でも最近、ニュース音声のディクテーション

の研究が行なわれている [15] [1]。評価用音声の条件としては、ほぼF0に相当するが、学習用と評価用音声の収録系が異なるので、それに対処するため、各文ごとにCMN法によりケプストラムの正規化を行っている。言語モデルとしては、約5年間のニュース原稿（のべ約2400万語）を用いて、頻度の高い20,000語（約98%をカバー）を対象としたトライグラムを作成した。表2に、ニュース原稿、日経新聞、およびNABの各学習テキストの単語数とカバー率の比較を示す。20,000語でも、ニュース原稿と日経新聞では、約70%（14,000語）しか単語が重複していない。表3に、ニュース原稿と日経新聞の学習テキストに含まれるn-gramの種類数と平均出現回数を示す。トライグラムになると出現回数が極めて少なくなり、ほとんどが1回以下しか出現しない。このためバックオフ平滑化が不可欠である。

実験の結果、80%程度の単語（形態素）認識率が得られている。ニュースの一文当りの平均単語数は約50語で、日経新聞の一文の2倍程度に長い。ニュース音声は原稿を元に発声されているが、文頭、文中における「えー」や、言い誤り、言い直しなどの自然発話現象を多く含んでおり、これらが認識の難しさの原因となっている。日経新聞とニュース原稿から作成した言語モデルを比較したところ、元のテキストの量がはるかに少なくても、ニュース原稿のみから作成したモデルの方が有効であることが分かった。表4にtest-set perplexityの比較を示す。

表2: 各学習テキストの単語数およびカバー率の比較

	ニュース	日経新聞	NAB
学習テキスト(単語数)	24M	180M	237M
異なり単語数	114k	623k	476k
5kカバー率	91.5%	88.0%	90.6%
7kカバー率	93.7%	90.3%	-
20kカバー率	98.0%	96.2%	97.5%
30kカバー率	99.1%	97.5%	-
65kカバー率	99.7%	99.0%	99.6%

表3: n-gramの種類数と平均出現回数

学習テキスト	n-gram	種類数	平均出現回数
ニュース原稿	ユニグラム	20k	1160
	バイグラム	0.9M	24
	トライグラム	4.4M	5
日経新聞	ユニグラム	20k	8747
	バイグラム	3.6M	44
	トライグラム	24M	6

表4．日経新聞とニュース原稿から作成した言語モデルのニュース音声に対するtest-set perplexityの比較

学習テキスト	言語モデル	
	バイグラム	トライグラム
ニュース原稿	105	48
日経新聞	190	63

4.3 話題抽出

CMUの'Informedia Digital Video Library Project'では、自然言語理解、画像処理、音声認識、および情報検索の技術を組み合わせて、膨大な映像と音声データベースを検索するシステムを構築している[16]。音声認識技術は、音声データベースのディクテーションと、音声による検索の両方に用いられる。このシステムでは、ディクテーションの結果をそのまま検索に用いている。

ディクテーションの結果から、さらにその内容を表す語（話題、トピックス）を推定しようとする研究が、日米の両方のタスクで試みられている。具体的方法としては、HMM[17]あるいはχ^2-法[15]が用いられている。ニュースをデータベース化し検索を容易にする上で、このような技術が役に立つと期待される。

5．ATISタスク

ARPAの対話音声の認識・理解プロジェクトとしては、ATISシステム（Air Travel Information System）が強力に進められた。これは、米国内の航空路線を用いて旅行をすることを想定して、コンピュータによる案内システムに、音声で問い合わせや予約をするシステム（語彙数：1,500～3,000語）である[18]。自由に発声した話し言葉を対象とするため、言い間違い、繰り返しなど難しい問題が多数含まれている。どこまで難しい音声を対象とするかによって異なるが、90～95%程度の単語認識率が得られている。このようなタスクでは、すべての単語が正しく認識される必要性は必ずしもなく、意味が正しく抽出され、ユーザが要求した動作が駆動されることが大切である。音声認識結果の自然言語を、システムを駆動する意味言語に変換するための言語モデルを、学習テキストから自動的に獲得する研究も行なわれている[19]。

ARPAの対話音声に関するプロジェクトは、最近では'Switchboard'や'Call Home'タスク（Hub5）と呼ばれる日常的な電話対話の認識へと展開している。ただし今のところ、これまでのディクテーションを目的とした方法をそのまま踏襲しているため、音響モデルの適応化や発音辞書の改良などによって認識率は徐々に向上しているが、対話音声特有の問題への打開策が見つからない状況にあると言える。

6．日本における大語彙対話音声認識

電話番号案内を対象とした研究が行なわれている[20]。電話番号の問い合わせを目的として、極めて多数の人の名前や住所を、かなり自由な文体で発声した音声の認識ができる必要がある。このため約80,000語を語彙とし、文脈自由文法を用いた認識システムの実験が進められている。このようなタスクの場合も、実際の問い合わせ音声を大量に録音して、それを文字に書き起し、それを用いて統計的言語モデルを構築することは原理的に不可能ではないが、膨大な労力を必要とする。このため、実際の対話例を参考にしながら人が手作業で作成した文脈自由文法を用い、言語処理には一般化予測LRパーザ

を用いている。ここで用いられるLRテーブルは、文脈自由文法の書き換え規則から、自動的に生成することができる。認識対象語彙数が大きくなると、LRテーブルの状態数が増え、テーブルの作成や修正が困難になるため、2段LRパーザを用いている。実験の結果、89%のキーワード認識率が得られている。

ATRにおける音声翻訳システムでは、旅行案内をタスクとして（語彙数：5,000～6,000）、可変次数単語n-gramとタスクに依存した言語制約を組み合わせた2-pass decoderを用いている[21]。

7．今後の展望

大語彙連続音声認識の研究は、近年大きく進歩しているが、任意の話題についての、あらゆる人による、あらゆる環境での音声が理解できるという目標からは、まだ程遠い状況にある。このためには、まず実環境での認識性能、すなわち耐性の向上、その具体的方法としての自動適応機能の研究が重要である。さらに、実際のシステムを実現するだけでなく、研究の効率化のためにも、認識処理速度の向上、いいかえると効率的な認識処理（探索）方法の実現が必須である。また、認識対象（タスク、言語など）が変わったときに始めからシステムを作り直すのでは、無駄が多すぎるので、いわゆるportabilityを向上させる研究も重要である。

ディクテーション（音声ワープロ）の他、電話などによる情報案内（電話番号案内、株価案内）の自動化、種々の電話サービスの自動化、各種予約サービスなど、コンピュータシステムとの対話という、より広くかつより有用な音声認識応用を実現するためには、言い直し、余剰語、未知語（OOV）などを沢山含む対話音声の言語モデルを構築し、自然な音声から、その意味内容を抽出し理解するための研究が極めて重要である。その際、すべての言葉を認識（ディクテーション）しようとするのではなく、意味を理解（意味検出、意味抽出）しようとするアプローチへの移行が必要である。

ARPAの1998年のワークショップ（2月8日～11日、Lansdowne、Virginia）では、Hub4に関して米語に中国語とスペイン語が加わる他、昨年以上にニュース音声の理解、トピック抽出、インデクシング、検索、さらに対話音声認識（Hub5）に重点が置かれる模様である。日本語への関心も高まっている。

参考文献

[1] 小林、他："ニュース音声認識システムの検討"、秋季音学講論、3-1-9 (1997)

[2] S. Young, et al.: "Tree-based state tying for high accuracy acoustic modeling", Proc. ARPA Human Language Technology Workshop, pp. 307-312 (1994)

[3] J.-L. Gauvain, et al.: "Maximum *a posteriori* estimation for multivariate Gaussian mixture observations of Markov chains", IEEE Trans. Speech and Audio Processing, SA-2, 2, pp. 291-298 (1994)

[4] C. J. Leggetter, et al.: "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models", Computer Speech and Language, 9, pp. 171-185 (1995)

[5] Proc. ARPA Speech Recognition Workshop, Morgan Kaufmann Publishers, San Francisco (1996)

[6] 松岡、他："新聞記事データベースを用いた大語い連続音声認識"、信学論、J79-D-II、12、pp. 2125-2131 (1996)

[7] 吉田、他：" 単語trigramを用いた大語彙連続音声認識"、音学音声研資、SP96-82 (1996)

[8] 大附、他："高次n-gramを用いた大語彙連続音声認識の検討"、音学講論、2-6-2 (1997)

[9] 西村、他："単語を認識単位とした日本語大語彙連続音声認識"、秋季音学講論、3-1-5 (1997)

[10] 板橋、他："日本音響学会新聞記事読み上げ音声コーパスの構築"、秋季音学講論、2-Q-36 (1997)

[11] R. Stern: "Specification of the 1996 HUB4 broadcast news evaluation", Proc. ARPA Speech Recognition Workshop, Morgan Kaufmann Publishers, San Francisco, pp. 7-10 (1997)

[12] J. L. Gauvain, et al.: "Transcribing broadcast news shows", Proc. ICASSP97, pp. 715-718 (1997)

[13] P. C. Woodland, et al.: "Broadcast news transcription using HTK", Proc. ICASSP97, pp. 719-722 (1997)

[14] J. S. Garofolo, et al.: "Design and preparation of the 1996 HUB-4 broadcast news benchmark test corpora", Proc. ARPA Speech Recognition Workshop, Morgan Kaufmann Publishers, San Francisco, pp. 15-21 (1997)

[15] 大附、他："ニュース音声を対象とした大語彙連続音声認識と話題抽出"、音学音声研資SP97-27 (1997)

[16] M. Witbrock, et al.: "Speech recognition and information retrieval: Experiments in retrieving spoken documents", Proc. ARPA Speech Recognition Workshop, Morgan Kaufmann Publishers, San Francisco, pp. 160-164 (1997)

[17] T. Imai, et al.: "Improved topic discrimination of broadcast news using a model of multiple simultaneous topics", Proc. ICASSP97, pp. 727-730 (1997)

[18] Proc. ARPA Spoken Language Systems Technology Workshop, Morgan Kaufmann Publishers, San Francisco (1995)

[19] 松岡、他：" テキストコーパスを用いた音声理解のための言語モデル自動獲得"、信学論、J79-D-II、12、pp.2070-2077 (1996)

[20] 南、他：" 番号案内を対象とした大語彙連続音声認識アルゴリズム"、信学論、J77-A、2、pp.190-197 (1994)

[21] T. Shimizu, et al: "Fast word-graph generation for spontaneous conversational speech translation", Proc. ICASSP97, pp. 95-98 (1997)