情報処理学会論文誌 Vol. 49 No. 8 2900–2909 (Aug. 2008)

音声とテキストを用いた認識単語辞書の自動構築

倉田岳人[†1] 森信介[†2]
伊東伸泰[†1] 西村雅史[†1]

大語彙連続音声認識（LVCSR）システムを新しい分野に導入する場合，その分野特有の単語を認識単語辞書に追加しなければならないが，計算量や認識単語辞書のメンテナンスを考慮すると，必要な分野特有の単語とその正しい読みのみを選択的に追加することが望ましい．しかし，日本語は，単語間に空白が置かれず，読みにもあいまい性があるため，対象分野のテキストデータのみから分野特有の単語とその読みを正しく自動的に獲得することは困難である．本論文では，対象分野のテキストデータに加えて，音声データも利用することで，対象分野特有の単語とその読みを含む小さいサイズの認識単語辞書を自動構築する方法を提案する．提案手法では，最初にテキストデータから再現率を重視した方法で大きいサイズの認識単語辞書を作成した後，その中から，音声データを利用して必要な単語と読みを選択し，小さいサイズの認識単語辞書を構築する．実験により，音声データを利用することで，最初の認識単語辞書のサイズの 10%以下の小さい認識単語辞書の追加で，対象分野の LVCSR システムを構築することができ，それが従来手法と比較して良い認識精度を示すことを確認した．

Unsupervised Construction of Speech Recognition Lexicon from Speech and Text

Gakuto Kurata,[†1] Shinsuke Mori,[†2] Nobuyasu Itoh[†1] and Masafumi Nishimura[†1]

When introducing a Large Vocabulary Continuous Speech Recognition (LVCSR) system to a specific domain, it is preferable to add the necessary domain-specific words and their correct pronunciations selectively to the lexicon. In this paper, we propose an unsupervised method of building a domain-specific lexicon in the Japanese language, where no spaces exist between words. In our method, by taking advantage of the speech of the target domain, we selected the domain-specific words from among an enormous number of word candidates extracted from the raw corpora. The experiments showed that by exploiting the acquired lexicon, whose size was negligible, an LVCSR system for the target domain was constructed efficiently and its performance was superior to the performance achieved by the conventional method, in which new words were acquired based on the results of automatic word segmentation.

1. はじめに

大語彙連続音声認識（LVCSR）システムは大学講義のアーカイブ[1),2)]や字幕作成[3)]，国会での議事録作成[4)]，コールセンタでの書き起こし[5),6)]のような様々な場面で利用され始めている．後段のアプリケーションで認識結果を直接利用する場合も，人手による修正作業が介在する場合も，音声認識の誤りができるだけ少ないことが期待されている．しかし，大学講義や国会答弁，コールセンタでの会話のような専門的な分野の音声を考えた場合，新聞のような一般分野の単語を網羅した認識単語辞書には含まれない，その分野特有の単語が出現する可能性が大きい．このような音声に対して，高い認識精度を得るためには，分野特有の単語とその読みを認識単語辞書に追加する必要がある．つまり，対象分野ごとに専用の LVCSR システムを構築することが必要となる．様々な分野への導入を考えると，人手を介さずに自動構築する方法が望ましい．

ある分野を対象とする LVCSR システムを構築するために，その分野のテキストデータを利用することは一般的であろう[7),8)]．たとえば，大学講義を考えてみると，その講義の教科書などが利用可能である．これらのテキストデータに対して，形態素解析を行い，その結果を利用して分野特有の単語とその読みを獲得し[9),10)]，それを認識単語辞書に追加する方法がある．しかし，形態素解析器は，その学習データや辞書に含まれない単語を必ずしも正しく解析できるわけではない[11)]．また，読みについても，辞書に含まれない単語に対しては何らかの推定方法で一意に決定されるが，この読み推定の精度は必ずしも十分ではなく[12)]，不適切な読みが付与されることもある．

認識精度を高くするためには，認識単語辞書の対象分野特有の単語に対する高い再現率が必要条件となる．これを考慮して，対象分野のテキストデータに出現する文字列に読みを付与して，大量に認識単語辞書に追加する方法も検討されている[13)]．しかし，大量に追加された文字列とその読みの中で，認識精度向上に寄与するものは一部だと考えられる．計算量

†1 日本アイ・ビー・エム株式会社東京基礎研究所
IBM Research, Tokyo Research Laboratory, IBM Japan Ltd.
†2 京都大学学術情報メディアセンター
Academic Center for Computing and Media Studies, Kyoto University

や認識単語辞書のメンテナンスを考慮すると，必要十分の対象分野特有の単語とその読みを含む小さいサイズの認識単語辞書が望ましい．

本論文では，専門的な分野への LVCSR の導入を想定し，小さいサイズの認識単語辞書の追加で文字誤りを減らすことを目的として，音声データとテキストデータを併用する方法を提案し，それが認識精度の観点から有効であることを示す．提案手法では，最初に対象分野のテキストデータから再現率を重視した大きなサイズの認識単語辞書を作成する．その後，同じ分野の音声データを活用してそのサイズを絞り込むことにより，必要十分なその分野特有の単語と読みを含む小さいサイズの認識単語辞書を作成し，それに基づいた対象分野の LVCSR システムを自動構築する．つまり，提案する手法では，対象分野のものであるということが分かっている音声データとテキストデータを用意すれば，自動的にその分野の単語とその読みを獲得することができ，それを利用して，対象分野の LVCSR システムを構築することができる．ここで，対象分野の音声データは，LVCSR システムで処理しようとしているものであり，利用可能である．また，音声データには対応する書き起こしや何らかのタグ付けが必要ではなく，これを収集することも可能である．利用可能なテキストデータと音声データを用いて，その分野の LVCSR システムを事前に自動構築することができれば，たとえば大学講義などを考えると，字幕作成や講義書き起こしに汎用的に利用できる．ここでの文字誤りの削減は，字幕の可読性の向上や，書き起こしの修正作業の軽減に寄与する．

以下，2 章では関連する研究について触れる．3 章では提案手法について説明する．4 章では評価実験とその結果を示し，考察を加える．最後に，5 章で本論文をまとめる．

2. 関連する研究

認識単語辞書に含まれない単語への対応方法として，そのような単語をモデル化する方法や，音声を利用して認識単語辞書に追加する方法が提案されてきた．

単語をモデル化する方法としては，音素や音素連接をサブワードとしてモデル化する方法がある[14)–19)]．これらの方法は，たとえば，日本人姓名や会社名などのような，辞書を用意することによる解決が難しかった単語の認識において効果的なことが示されている[14),15)]．

認識単語辞書に単語を追加するために，音声データから読みを生成する方法も検討されてきた[20)–22)]．しかし，これらの方法は，主に人名などの孤立単語発声を対象としており，読みを付与する必要がある単語が分かっている状態で，その単語を発声した音声データが必要となる[⋆1]．しかし，LVCSR システムを新しい分野に導入する場面では，読みが必要となる単語が最初から分かっている状況はあまりなく，また，各々の単語に対して，対応する発声データを用意することも，現実的ではない．

連続発声された入力音声によって認識単語辞書への単語追加を行う方法も検討されてきた．たとえば，IBM ViaVoice[23),24)] では，以下のような手順で単語登録を行うことができる．

i. ユーザが登録したい単語を含む文章を発声する．

ii. IBM ViaVoice がユーザの発声を認識し，認識結果をユーザに提示する．しかし，この段階では認識単語辞書に登録されていない単語は正しく認識されず，提示される認識結果は正しくない．

iii. ユーザが認識誤りを，単語区切りも含めて修正する．

iv. ユーザ発声と修正された認識結果を照合することで，ユーザが意図した単語に対して読みが付与される．

この方法でも，人手による単語の修正入力，および単語分割の修正が必要となる．

3. 提 案 手 法

本章では，適応対象分野の与えられたテキストデータと音声データから，その分野の単語と対応する読みを自動的に獲得し，それらを利用して，小さいサイズの認識単語辞書の追加のみで対象分野の LVCSR システムを構築する方法を説明する．また，後述する実験において比較する既存手法についても説明する．

3.1 提案手法の処理手順

提案手法では，最初に，対象分野の単語分割されていないテキストデータから大量の単語候補を抽出し，それらに表記から考えられる読みを複数割り当て，認識単語辞書を作る．次に，テキストデータを確率的に単語分割し，その結果に基づき対象分野の言語モデル（LM）を推定し[25)]，対象分野用の LVCSR システムを構築する．この LVCSR システムを利用して，対象分野の音声を認識し，その際に認識結果の 1 位候補の中に出現した単語候補と読みを，分野特有の単語として読み付きで獲得する．最後に，獲得した単語とテキストデータから LM を再推定し，対応する読みも利用して認識単語辞書を変更し，LVCSR システムを構築する．**図 1** に提案手法の処理の手順を示した．

⋆1 たとえば，“Stephen” という単語の読みを登録するために，“*Stephen*” と発声したデータを用意して，表記から考えられる読みの中で，音声データと最も近いものを選択する．

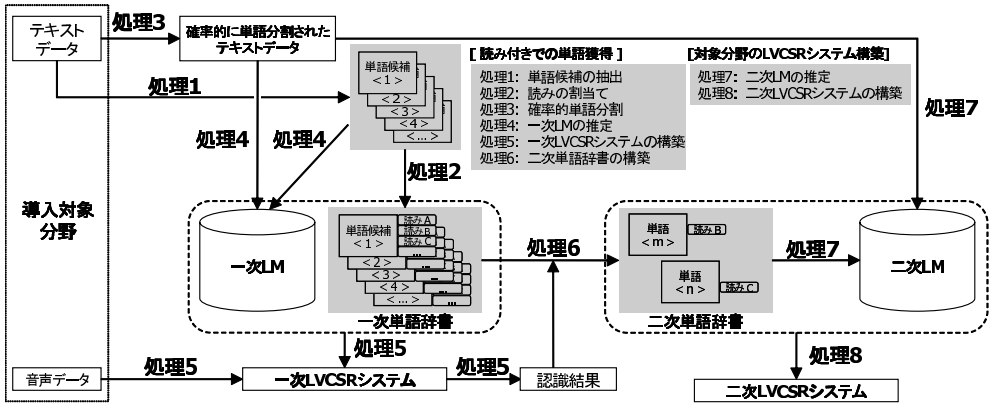

図 1　提案手法の概要
Fig. 1　Overview of proposed method.

なお，一般分野の単語分割済みコーパスと読み付きの認識単語辞書（一般分野単語辞書）はすでに用意されているものとし，これらから一般分野 LM の推定を行う．音声認識実行時には，一般分野 LM と対象分野のテキストデータから推定する LM を補間して利用する．また，対象とする音声と同じ環境の音声データから学習された音響モデル（AM）を音声認識実行時に利用する．一般分野単語辞書，一般分野 LM，および AM については，図が煩雑になることを避けるため，図 1 には示していない．

以下では，図 1 に従って，提案手法の各処理について説明を行う．

処理 1：単語候補の抽出

対象分野のテキストデータから，対象分野特有の単語に対して再現率の高い方法を利用して，単語候補を抽出する．ここでは，文字列の頻度に基づく方法を採用する[13),26)]．以下に処理の流れを示した．

i. 句読点，鍵括弧，記号に加え，「が」「を」などの助詞，「けれども」のような頻出接続詞，合計 78 個を，必ず当該位置で分割すべきストップワードと定義し，それに基づき対象分野のテキストデータを分割する．

ii. すべての文字列の頻度を計数する．

iii. ある文字列の頻度よりも，左右どちらかに 1 文字を追加した文字列の頻度が小さくなった場合，追加前の文字列と追加した 1 文字の文字境界を単語境界候補とする．

iv. 両端が単語境界候補となっている文字列を，単語候補として抽出する．

図 2 に例を示した．なお，この例では，「ア」の左側が単語境界候補となっている状態で，右側に 1 文字ずつ追加していき，単語境界候補を探している．

この方法は，高い再現率が期待できるが，多くの無意味な文字列も単語候補として抽出してしまうため，この段階では単語候補の数は非常に多くなる．ここで抽出された単語候補の中から，この後の処理で，音声データを利用して適切なものを獲得するため，この段階では再現率を重視した方法を採用する．

処理 2：読みの割当て

単語候補に対して音声合成システムで利用される未知語読みモデル[12)]を利用して読みを付与する．この未知語読みモデルは，文字と読みの組合せを単位とした n-gram で未知語をモデル化し，確率付きで読みの候補を出力する．文献 12) によると，未知語読みモデルは 100%の精度で読みを推定できるわけではない．予備実験として，一般分野単語辞書の単語の中で，一般分野の単語分割済みコーパス内での頻度下位 1,000 単語を対象として，この未知語読みモデルで読みの推定を行った．**図 3** に，横軸に順位を，縦軸にその順位までに正解が含まれる割合（累計精度）を示した．図 3 から，1 位のみを利用した場合の精度は 0.79 程度であるが，複数の読みを利用することで累計精度が上昇し，上位 10 位程度で累計精度が飽和していることが分かる．よって，ここでは各々の単語候補に対して，文献 13) のように

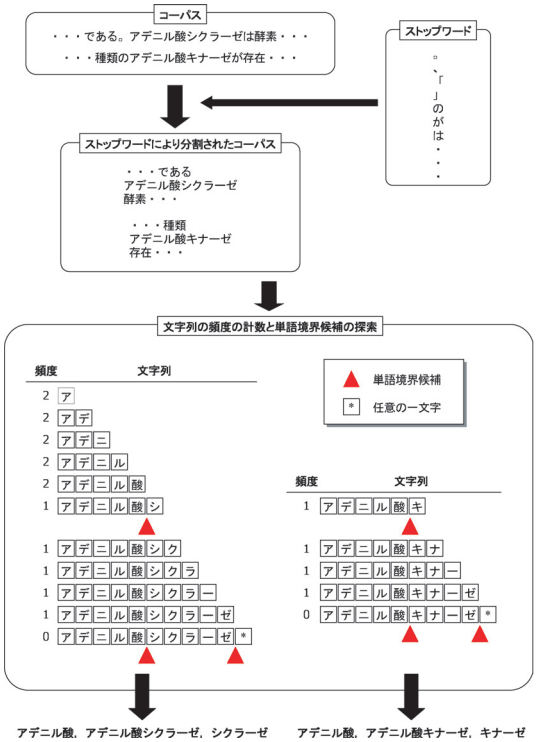

図 2　単語候補抽出の例
Fig. 2　Extraction of word candidates.

未知語読みモデルが出力する最尤の読みだけを付与するのではなく，上位 10 個の読みを付与する．**処理 1** で得た単語候補と各々に付与した複数の読みの集合を一次単語辞書と呼ぶ．

処理 3：確率的単語分割

対象分野の LM を推定するために，対象分野のテキストデータを確率的に単語分割する[25]．確率的単語分割では，単語境界位置を一意に決定せず，すべての文字間に，ある確率で単語境界が存在すると見なす．つまり，n_r 文字からなるテキストデータを文字列 $\boldsymbol{x} = x_1 x_2 \cdots x_{n_r}$ と見なし，i 番目の文字 x_i のあとに単語境界が存在する確率 p_i（「単語境界確率」と称する）がすべての $i \in \{1, 2, \cdots, n_r - 1\}$ について計算される．ここでは，まず自動単語分割器の単語境界推定精度[⋆1] α を算出し，その自動単語分割器でテキストデータを単語分割する．そして，単語境界と判定された部分の単語境界確率を α に，判定されなかった部分の単語境界確率を $1 - \alpha$ とすることにより，確率的単語分割を行う．

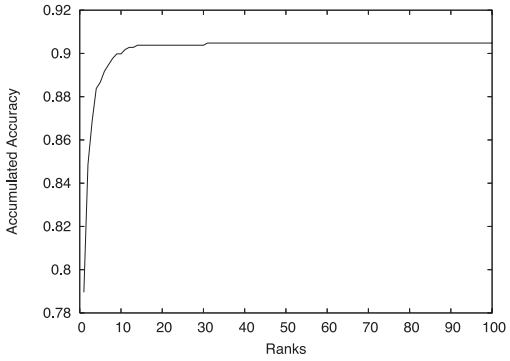

図 3　未知語読みモデルの性能
Fig. 3　Performance of unknown word pronunciation model.

なお，ここで利用した自動単語分割器は，新聞などの一般分野の単語分割済みコーパス（約 5 万文）を利用して構築されており，一般分野のテキストに対する単語境界推定精度は 0.952 であった．つまり，$\alpha = 0.952$ とする．

処理 4：一次 LM の推定

確率的単語分割済みのテキストデータ，**処理 1** で得られた単語候補，および一般分野単語辞書から対象分野の LM を推定する[⋆2]．確率的単語分割を利用することで，コーパス中の任意の文字列に対して，単語 n-gram 確率を推定することができるため，**処理 1** で抽出したすべての単語候補に対して，文脈を考慮した適切な単語 n-gram 確率が与えられることになる．この LM を一次 LM と呼ぶ．

処理 5：一次 LVCSR システムの構築

一次単語辞書，一般分野単語辞書，一次 LM と一般分野 LM の補間 LM を利用して一次 LVCSR システムを構築し，対象分野の音声データに対して音声認識を行う．

⋆1 すべての文字間に対して，自動単語分割器が「分割する」，「分割しない」を正しく判断した割合
⋆2 確率的に単語分割されたテキストデータからの単語 n-gram 確率推定方法は，文献 25) に示されている．

処理 6：二次単語辞書の構築

一次単語辞書の中で，認識結果の 1 位候補の中に出現した単語候補とその読みの組合せを獲得する．つまり，一次単語辞書に含まれる大量の単語候補の中から出現した単語候補のみが，さらにその単語候補に複数割り当てられている読みの中から出現した読みのみが獲得される．これらを二次単語辞書と呼ぶ．

処理 7：二次 LM の推定

処理 3 で確率的に単語分割したテキストデータと二次単語辞書，および一般分野単語辞書から LM を再推定する．ここで得られる LM を二次 LM と呼ぶ．

処理 8：二次 LVCSR システムの構築

二次単語辞書，一般分野単語辞書，二次 LM と一般分野 LM の補間 LM を利用して二次 LVCSR システムを構築する．二次 LVCSR システムは，一次 LVCSR システムによる認識結果から獲得された単語候補と読みのみ，つまり小さいサイズの認識単語辞書のみを追加して構築された対象分野の LVCSR システムということができる．

3.2　比較する既存手法

対象分野のテキストデータを利用して，その分野の LVCSR システムを構築する方法としては，自動単語分割器を利用する方法が一般的である．この方法では，3.1 節で示した提案手法に対して，以下の点を変更する．

単語候補の抽出　対象分野のテキストデータを自動単語分割器で単語分割する．この結果として得られる単語分割済みテキストデータに出現するすべての単語の中で，一般分野単語辞書に含まれないものを，その分野の単語候補とする．

LM の推定方法　一次 LM，二次 LM ともに，自動単語分割器で単語分割された対象分野のテキストデータに基づいて，推定する．

他の処理，つまり，単語候補に未知語読みモデルを利用して上位 10 個の読みを与えて一次単語辞書を構築する**処理 2** や，一次 LVCSR システムによる認識結果に基づいて二次 LVCSR システムを構築する**処理 5**，**6**，**8** は，提案手法と同様である．本論文では，この方法を，従来手法と位置づける．

ここで述べた従来手法と，3.1 節で述べた提案手法を定性的に比較する．提案手法では，再現率の高い方法でテキストデータから抽出される多くの単語候補の中から，音声データによって分野特有の単語とその読みが獲得される．それに対して，従来手法では，自動単語分割の結果得られた単語候補の中から獲得されることになり，自動単語分割の誤りに対して頑健ではない．

4. 実　　験

1 章で示したように，字幕作成などへの利用も考慮し，テキストデータと音声データを利用して，必要十分の分野特有の単語とその読みを含む小さいサイズの認識単語辞書を追加し，事前に LVCSR システムを構築した．そして，3.2 節で示した自動単語分割器を利用する従来手法と比較して，提案手法に基づく LVCSR システムが，対象分野の未知の音声データに対して，どの程度の認識精度を得ることができるのかを定量的に検証した．本章では，実験についてまとめ，その結果を示し，考察を加える．

4.1　実験の対象分野

専門的分野への LVCSR システムの導入という観点から，我々は放送大学の講義音声を対象として，実験を行った．放送大学はテレビやラジオを通じて講義を配信している．各々の講義の内容は専門的であり，新聞記事などの一般分野のテキストには出現しない専門的な単語が頻出し，これらを認識単語辞書に追加しなければ，高い認識精度を得ることは難しい．今回の実験では，生物（講義 $\mathcal{B}$）と地球科学（講義 $\mathcal{G}$）の講義を対象とした．

3.1 節で述べたように，提案手法では対象分野のテキストデータと音声データを利用する．分野特有の単語を獲得するためのテキストデータとして，放送大学が発行している各々の講義の教科書などを用意した．利用したテキストデータのサイズは**表 1** に示した．また，音声データとして，放送大学の講義音声の中の講師の発話部分を利用した．放送大学の 1 回の講義時間は 45 分間であり，講義 $\mathcal{B}$ については 1 回分，講義 $\mathcal{G}$ については 2 回分の講義音声を利用した．音声データのサイズについても表 1 に示した．

4.2　実験の手順

各々の講義について，対応するテキストデータから一次単語辞書，一次 LM を作成し，一次 LVCSR システムを構築した．各々の講義音声を 10 個に分割し，このうちの 9 個を一次 LVCSR システムで認識し，認識結果の 1 位候補の中に出現した単語候補とその読みから，講義ごとの二次単語辞書を作成した．最後に，講義ごとの二次 LVCSR システムを構築し，

表 1　講義音声とテキストデータのサイズ
Table 1　Statistics of the lectures.

講義	講義内容	テキストデータの文字数	講義音声の書き起こしの文字数
$\mathcal{B}$	生物	73,437	12,600
$\mathcal{G}$	地球科学	88,003	19,974

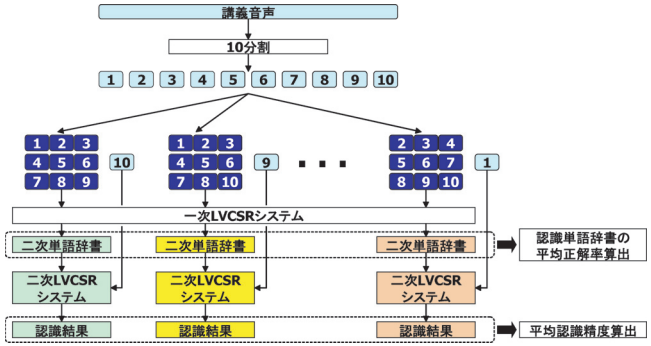

図 4　10 分割交差検定の概要
Fig. 4　Overview of 10-fold cross validation.

表 2　自然発話音声コーパスの概要
Table 2　Overview of the training corpus for AM.

話者数	97 人
発話時間	83 時間
文数	27,135 文
単語数	1,098,888 単語
異なり単語数	23,929 単語

一次 LVCSR システムでは認識対象としなかった残りの 1 個に対する認識実験を行った．正確な評価を行うため，一次 LVCSR システムで認識する 9 個と，評価で利用する 1 個を順次入れ替えて実験を行い，10 回の平均を算出する交差検定を行った．**図 4** に，その概要を示した．提案手法と，自動単語分割器を利用する従来手法で二次 LVCSR システムを各々構築し，その認識性能を比較した．

参考のため，提案手法と従来手法でそれぞれ作成した各々の講義の一次 LVCSR システムを利用して同じ音声データを認識する実験，および，一般分野 LM，一般分野単語辞書から構築した一般 LVCSR システムを利用して同じ音声データを認識する実験も行った．

なお，すべての実験条件で共通して利用した AM，一般分野 LM，および一般分野単語辞書については，以下のとおりである．

AM[27]　**表 2** に，今回の実験で利用した AM を構築するために利用した，自然発話音声コーパスのサイズを示した．各音素は環境依存で 3 状態の left-to-right 型の HMM として表現されている．HMM の各状態は音素環境決定木によりクラスタリングされており，決定木のリーフ数は 2,728 である．また，HMM の各状態は，11 混合の混合正規分布でモデル化されている．

表 3　一般分野 LM を構築するためのコーパスの概要
Table 3　Overview of the training corpus for general LM.

コーパスの総文字数	46,022,380 文字
コーパスの総単語数	24,442,503 単語
一般分野単語辞書のサイズ	45,402 単語

表 4　従来手法（A）と提案手法（P）を利用した場合の一次，二次 LVCSR システムの比較
Table 4　Comparison of initial and purified LVCSR systems using proposed and conventional method.

講義	LVCSR System	一般分野単語辞書		追加した単語辞書		未知語率	CER
		単語数	読みの総数	単語数	読みの総数	[%]	[%]
$\mathcal{B}$	一般	45,402	53,225	—	—	13.43	26.7
	一次 (A)			1,365	4,439	3.74	14.1
	二次 (A)			86.8	87.7	4.48	14.5
	一次 (P)			3,980	25,994	3.18	11.5
	二次 (P)			218.4	221.1	3.99	11.5
$\mathcal{G}$	一般	45,402	53,225	—	—	5.43	30.2
	一次 (A)			1,446	7,440	2.52	29.2
	二次 (A)			59.7	60.5	2.80	29.4
	一次 (P)			5,008	37,533	1.87	28.6
	二次 (P)			194.0	195.8	2.40	28.4

一般分野 LM と一般分野単語辞書　一般分野のコーパスとして，主に新聞記事から構成されているコーパスを利用した．**表 3** に，このコーパスの概要を示した．そして，このコーパスに基づき一般分野 LM を構築した．なお，一般分野のコーパスの単語分割に関しては，一部分は専門家により正確に分割されており，残りの部分に関しては，自動単語分割器により分割された後，専門家により大まかに点検されている．次にこのコーパスに出現する単語の 95%を網羅するように，出現頻度降順に単語を選択した．さらにこれらの単語に適切な読みを人手で付与し，一般分野単語辞書とした．

4.3　実 験 結 果

表 4 に行った実験の結果を示した．左から 2 列目は利用した LVCSR システムを示しており，“(A)” は自動単語分割器の結果に基づき単語を獲得して二次 LVCSR システムを構築する従来手法を利用した場合，“(P)” は提案手法を利用した場合を表している．左から 3,

4 列目には一般分野単語辞書のサイズを，5，6 列目には各々の LVCSR システムで，一般分野単語辞書とは別に追加した認識単語辞書のサイズを示している．なお，一般 LVCSR システムでは，一般分野単語辞書のみを利用するため，該当する上から 3，8 行目の，左から 5，6 列目を「—」としている．また，二次 LVCSR システムに追加した認識単語辞書のサイズは，交差検定の平均のため，整数ではなくなっている．そして，7 列目には各々の LVCSR システムの認識単語辞書に対する，評価用の講義音声の未知語率を文字数換算で示している．最後に，8 列目には評価用の講義音声に対する認識実験の結果得られた文字誤り率（CER）を示している．

たとえば，最下行は提案手法を利用した場合の講義 $\mathcal{G}$ に関する二次 LVCSR システムについてのデータであり，単語数が 45,402，読みの総数が 53,225 の，各条件で共通の一般分野単語辞書に対して，追加した二次単語辞書の単語数が平均 194.0 個，読みの総数が平均 195.8 個であった．また，未知語率は 2.40%，CER は 28.4%であった．

本論文では，認識精度の評価尺度として，英語などで一般的に用いられる単語誤り率（WER）ではなく，CER を採用した[28]．この理由は，以下のとおりである．日本語には単語分割にあいまい性が存在し，このあいまい性は分野特有の単語についてはさらに大きくなる．また，提案手法においては，文字単位での処理で単語候補を選択し，それらの中から対象分野特有の単語を選択しており，正解の書き起こしデータにおける単語分割との不一致が生じるため，WER では正確な認識精度の比較が期待できない．

4.4　実験結果に対する考察

4.4.1　認識精度と認識単語辞書のサイズ

提案手法を利用した場合，いずれの講義においても，二次単語辞書は一次単語辞書[⋆1]と比較して，単語の数が 10%以下，読みの数が 1%以下に削減されたが，一次 LVCSR システムと二次 LVCSR システムでは同じ程度の認識精度が得られた．この結果は，一次 LVCSR システムによる認識結果の 1 位候補に現れた単語のみを選択的に利用することで，対象分野の音声を認識するために必要十分な単語とその読みのみを含む小さいサイズの二次単語辞書を構築することができたことを意味する．

また，従来手法を用いた場合の二次 LVCSR システムと，提案手法を用いた場合の二次 LVCSR システムを比較すると，いずれの講義においても，提案手法を利用した場合の方が二次単語辞書のサイズが大きくなったが，認識精度は向上した．たとえば講義 $\mathcal{B}$ では，CER が 14.5%から 11.5%に減少し，改善がみられている．それぞれの講義における改善は，有意水準 1%で統計的に有意であった．なお，提案手法を利用した場合の，従来手法に対する二次単語辞書のサイズの増加分は，一般分野単語辞書のサイズの 0.3%以下であり，得られた CER から判断すると悪影響はなかった．

これらの結果から，提案手法を利用することにより，小さいサイズの認識単語辞書の追加で，従来手法よりも高い認識精度を得ることができる LVCSR システムを構築することができた．

4.4.2　未 知 語 率

提案手法と従来手法を比較すると，二次 LVCSR システムの未知語率は，提案手法の方が低く抑えることができた．これは，提案手法を利用することにより，従来手法よりも多くの単語を獲得できたことが要因である．

評価用の音声データに対する未知語率は，提案手法を利用した場合も，従来手法を利用した場合も，一次 LVCSR システムと比較して，二次 LVCSR システムでは高くなった．提案手法でも従来手法でも，単語とその読みを獲得するためには，その単語がテキストデータと音声データの両方に現れる必要がある．しかし，テキストデータに現れ，交差検定を行う際の評価用の音声データにも現れたが，一次 LVCSR システムで認識対象とした音声データには現れなかった単語も存在し，そのような単語は二次単語辞書には追加されなかったため，二次 LVCSR システムの未知語率は一次 LVCSR システムよりも高くなった．

従来からテキストデータの音声データに対する被覆率は，LVCSR において高い認識精度を得るために重要であったが，提案手法を有効に活用するためには，テキストデータの音声データに対する被覆率と音声データのテキストデータに対する被覆率の両方が高いことが重要となる．

4.4.3　二次単語辞書として獲得された単語とその読み

字幕作成だけではなく，インデクシングなどの他のアプリケーションでの応用を検討した場合，認識結果が直感的に単語単位と見なせる単位から構成されていることが望ましい．これを考慮して，二次 LVCSR システムに追加された二次単語辞書を調べ，それらが対象分野特有の単語と読みとしてふさわしいかどうかを検証した．追加された単語が，分野特有の単語としてふさわしく，その読みが正確であった場合に，その組合せを正解と判断した．複合語が追加される場合もあるが，それらについては，複合語内の係り受け関係に基づいて正誤を判断した[29]．**表 5** に，正解・不正解と判断した単語と読みの組合せの例，およびその理由を示した．全体の平均の正解率に加えて，一次 LVCSR の結果の 1 位候補に複数回現れた

⋆1 単語長により読みの候補が 10 個以下の場合もあるため，読みの総数は単語候補の数の 10 倍以下となっている．

表 5　正解・不正解と判断した単語と読みの組合せの例
Table 5　Example of proper and improper acquired lexicon.

講義	単語	読み	判断	理由
$\mathcal{B}$	受容体	ju yo o ta i	○	
	リン酸化	ri n sa n ka	○	
	サブユニット	sa bu yu ni tto	○	
	百日咳	hya ku ni chi ze ki	○	
	残基	za n ki	○	
	ナーゼ	na a ze	×	「キナーゼ」の一部．単語として意味がない．
	模式化した	mo shi ki ka shi ta	×	単語としてふさわしくない．
$\mathcal{G}$	古生代	ko se e da i	○	
	三葉虫	sa n yo o chu u	○	
	フズリナ	fu zu ri na	○	
	顕生累代	ke n se e ru i da i	○	
	筆石	fu de i shi	○	
	ある意味	a ru i mi	×	単語としてふさわしくない．
	多細胞	ta sa i **ho** o	×	読みが誤っている．

表 6　二次単語辞書の正解率（正解の数 / 獲得された単語と読みの組合せの総数）
Table 6　Accuracy of acquired lexicon.

講義	手法	1 位候補に現れた回数ごとの正解率		平均
		複数回	1 回のみ	
$\mathcal{B}$	従来手法	88.5%　(416 /　470)	74.2%　(302 /　407)	81.9%　(　718 /　877)
	提案手法	96.0%　(972 / 1,012)	74.4%　(892 / 1,199)	84.3%　(1,864 / 2,211)
$\mathcal{G}$	従来手法	76.5%　(151 / 198)	55.3%　(225 /　407)	62.1%　(　376 /　605)
	提案手法	82.6%　(404 / 489)	57.4%　(843 / 1,469)	63.7%　(1,247 / 1,958)

場合と，1 回しか現れなかった場合とを分類して正解率を算出し，結果を**表 6** に示した．なお，この結果は交差検定の各検定時に獲得された単語と読みの組合せの総計を表している．

いずれの講義においても，提案手法を利用した場合の方が正解率が高く，また多くの単語と読みの組合せを獲得することができた．従来手法では，自動単語分割器が分野特有の単語を正確に単語分割できなかった場合，その単語は一次単語辞書に登録されず，獲得することはできなかった．それに対して，提案手法では，再現率の高い方法で単語候補を選択し，さらに確率的単語分割に基づいて，それらに適切な LM 確率を付与しているため，従来手法よりも多くの単語を獲得できた．

また，複数回 1 位候補に現れた単語の正解率は，1 回しか現れなかった単語の正解率よりも高かった．これは，提案手法だけでなく，従来手法でも同様であった．認識結果の 1 位候補に複数回現れたということは，学習用音声中に当該単語の当該の読みが，LM 確率が高い文脈で複数回現れた可能性が高いということであり，これらの単語の正解率が高いことは妥当な結果である．逆に，1 回しか現れなかった単語に関しては，単なる挿入・置換誤りの可能性もあり，これらの正解率が下がっていることも妥当である．

すべて自動的に獲得されたことを考慮すると，提案手法を利用して得られた二次単語辞書の正解率は高く，他のアプリケーションでの応用も期待できる．

5. おわりに

LVCSR システムを様々な分野に導入する場合，導入対象分野の必要な単語とその正しい読みのみを自動的に獲得し，認識単語辞書に追加することが望まれている．本論文では，対象分野の音声データとテキストデータを併用して，その分野特有の単語とその読みを含む小さいサイズの認識単語辞書を自動構築する方法を提案した．

実験により，提案手法を利用することで，小さいサイズの認識単語辞書の追加で対象分野の LVCSR システムを構築することができ，また，この LVCSR システムは，自動単語分割に基づく従来手法を利用して構築した LVCSR システムよりも，対象分野の未知の音声データに対して良い認識精度を示すことを確認した．つまり，提案手法は，専門的分野の字幕作成や書き起こしに利用する LVCSR システムの自動構築への貢献が期待できる．また，構築した認識単語辞書に含まれる単語とその読みは，高い確率で分野特有の単語とその読みとしてふさわしいことも確認した．これらは，インデクシングなどに今後利用できる可能性がある．

提案手法を利用するためには，音声データのテキストデータに対する被覆率が重要となる．つまり，提案手法で獲得できる単語と読みは，音声データとテキストデータの両方に現れる単語に限られる．高度なテキスト処理技術[11),29)]と融合させて，たとえばテキストデータ中での頻度が非常に高い文字列については，音声データに現れなくても無条件に分野特有の単語として獲得するということも考えられる[⋆1]．これについては今後検討していきたいと考えている．

謝辞　放送大学の番組制作に携わっておられる方々に深謝します．

⋆1 ただし，この場合には読みが決定されない．

参 考 文 献

1) Glass, J., Hazen, T.J., Cyphers, S., Malioutov, I., Huynh, D. and Barzilay, R.: Recent Progress in the MIT Spoken Lecture Processing Project, *Proc. INTERSPEECH*, pp.2553–2556 (2007).
2) Kawahara, T.: Spoken language processing for audio archives of lectures and panel discussions, *Proc. ICKS*, pp.23–30 (2004).
3) Miyamoto, K.: Effective Master-Client Closed Caption Editing System for Wide Range Workforces, *Proc. UAHCI*, Vol.7 (2005).
4) 秋田祐哉，河原達也：統計的機械翻訳の枠組みに基づく言語モデルの話し言葉スタイルへの変換，情報処理学会研究報告，2005-SLP-59-19, pp.109–114 (2005).
5) Chen, S.F., Kingsbury, B., Mangu, L., Povey, D., Saon, G., Soltau, H. and Zweig, G.: Advances in Speech Transcription at IBM under the DARPA EARS Program, *IEEE Trans. Audio, Speech and Language Processing*, Vol.14, No.5, pp.1596–1608 (2006).
6) Soltau, H., Kingsbury, B., Mangu, L., Povey, D., Saon, G. and Zweig, G.: The IBM 2004 Conversational Telephony System for Rich Transcription, *Proc. ICASSP*, Vol.1, pp.205–208 (2005).
7) Bellegarda, J.R.: Statistical Language Model Adaptation: Review and Perspectives, *Speech Communication*, Vol.42, pp.93–108 (2004).
8) Janiszek, D., Mori, R.D. and Bechet, F.: Data Augmentation and Language Model Adaptation, *Proc. ICASSP*, Vol.1, pp.549–552 (2001).
9) Nagata, M.: A Stochastic Japanese Morphological Analyzer Using a Forward-DP Backward-A* N-Best Search Algorithm, *Proc. COLING*, pp.201–207 (1994).
10) 松本裕治：形態素解析システム「茶筌」，情報処理，Vol.41, No.11, pp.1208–1214 (2000).
11) 東　藍，浅原正幸，松本裕治：条件付確率場による日本語未知語処理，情報処理学会研究報告，2006-NL-173, pp.67–74 (2006).
12) 長野　徹，森　信介，西村雅史：*N*-gram モデルを用いた音声合成のための読みおよびアクセントの同時推定，情報処理学会論文誌，Vol.47, No.6, pp.1793–1801 (2006).
13) 倉田岳人，森　信介，西村雅史：講義関連コーパスを利用した音声認識システムの自動適応，電子情報通信学会論文誌，Vol.J90-D, No.9, pp.2530–2540 (2007).
14) 谷垣宏一，山本博史，匂坂芳典：クラスに依存した語彙の確率的記述に基づく階層型言語モデル，電子情報通信学会論文誌，Vol.J84-D-II, No.11, pp.2371–2378 (2001).
15) 山本博史，小窪浩明，菊井玄一郎，小川良彦，匂坂芳典：複数のマルコフモデルを用いた階層化言語モデルによる未登録語認識，電子情報通信学会論文誌，Vol.J87-D-II, No.12, pp.2104–2111 (2004).
16) Bazzi, I. and Glass, J.R.: Modeling Out-of-Vocabulary Words for Robust Speech Recognition, *Proc. ICSLP*, pp.401–404 (2000).
17) Bazzi, I. and Glass, J.R.: Learning Units for Domain-Independent Out-of-Vocabulary Word Modelling, *Proc. EUROSPEECH*, pp.61–64 (2001).
18) Bazzi, I. and Glass, J.R.: A Multi-Class Approach for Modelling Out-of-Vocabulary Words, *Proc. ICSLP*, pp.1613–1616 (2002).
19) Galescu, L.: Sub-Lexical Language Models for Unlimited Vocabulary Speech Recognition, 電子情報通信学会技術研究報告，SP2002-30, Vol.102, No.108, pp.37–42 (2002).
20) Deligne, S., Maison, B. and Gopinath, R.: Automatic Generation and Selection of Multiple Pronunciations for Dynamic Vocabularies, *Proc. ICASSP*, Vol.1, pp.565–568 (2001).
21) Maison, B., Chen, S.F. and Cohen, P.S.: Pronunciation Modeling for Names of Foreign Origin, *Proc. ASRU*, pp.429–434 (2003).
22) Bodenstab, N. and Fanty, M.: Multi-Pass Pronunciation Adaptation, *Proc. ICASSP*, Vol.4, pp.865–868 (2007).
23) Nuance: IBM ViaVoice（オンライン）．http://japan.nuance.com/viavoice/（参照 2008-05-11）．
24) 西村雅史，伊東伸泰：単語を認識単位とした日本語ディクテーションシステム，電子情報通信学会論文誌，Vol.J81-DII, No.1, pp.10–17 (1998).
25) 森　信介，宅間大介，倉田岳人：確率的単語分割コーパスからの単語 *N*-gram 確率の計算，情報処理学会論文誌，Vol.48, No.2, pp.892–899 (2007).
26) Feng, H., Chen, K., Deng, X. and Zheng, W.: Accessor Variety Criteria for Chinese Word Extraction, *Computational Linguistics*, Vol.30, No.1, pp.75–93 (2004).
27) 西村雅史，伊東伸泰：講義コーパスを用いた自由発話の大語彙連続音声認識，電子情報通信学会論文誌，Vol.J83-DII, No.11, pp.2473–2480 (2000).
28) 西村雅史，伊東伸泰，山崎一孝：単語を認識単位とした日本語の大語彙連続音声認識，情報処理学会論文誌，Vol.40, No.4, pp.1395–1403 (1999).
29) Asahara, M. and Matsumoto, Y.: Japanese Unknown Word Identification by Character-based Chunking, *Proc. COLING*, pp.459–465 (2004).

(平成 20 年 1 月 10 日受付)
(平成 20 年 5 月 8 日採録)

倉田　岳人（正会員）

2002 年東京大学工学部電子工学科卒業．2004 年同大学大学院情報理工学系研究科電子情報学専攻修士課程修了．同年日本アイ・ビー・エム（株）入社．以来，同社東京基礎研究所にて，音声認識等の音声言語情報処理の研究に従事．日本音響学会会員．

森　　信介（正会員）

1998 年京都大学大学院博士後期課程修了．同年日本アイ・ビー・エム（株）入社．2007 年 5 月より京都大学学術情報メディアセンター准教授．工学博士．1997 年本学会山下記念研究賞受賞．言語処理学会会員．

伊東　伸泰（正会員）

1982 年大阪大学基礎工学部生物工学科卒業．1984 年同大学大学院博士前期課程修了．同年日本アイ・ビー・エム（株）入社．以来，同社東京基礎研究所にて，文字認識，音声認識の研究に従事．

西村　雅史（正会員）

1981 年大阪大学基礎工学部生物工学科卒業．1983 年同大学大学院物理系博士前期課程修了．同年日本アイ・ビー・エム（株）入社．以来，同社東京基礎研究所にて，音声認識等の音声言語情報処理の研究に従事．同社主席研究員．工学博士．1998 年本学会山下記念研究賞，1999 年日本音響学会技術開発賞各受賞．電子情報通信学会，日本音響学会各会員．