Vol.2011-SLP-85 No.15
2011/2/5

日本人英語発話からの文法誤り検出

安斎拓也[†1]　咸　聖俊[†1]　伊藤彰則[†1]

本研究では，外国語学習者が対話形式で文法や表現の学習をすることができる，音声対話型 CALL システムについての検討を行っている．システムによって文法的な誤りを指摘するためには，学習者の発話に発音の誤りや文法的な誤りが含まれていても，発話された通りに認識する必要がある．そこで我々はまず，様々な学習者の発音レベルを考慮した音響モデルについての検討を行った．次に，日本人学習者の文法的な誤り傾向を反映して生成されたテキストから N-gram 言語モデルを学習した．学習者の発音レベルを考慮し，複数のモデルを用いて音声認識することで，誤りの指摘の精度を最大で 3.7 ポイント改善することができた．

Grammatical Error Detection from English Utterances Spoken by Japanese

Takuya Anzai,[†1] Seongjun Hahm[†1] and Akinori Ito[†1]

We focus on a voice-interactive CALL system that enables a foreign language learner to make grammar or expression practice. For realizing correction of grammatical mistakes by the CALL system, the system is required to recognize the learner's utterances as is, including prounciation errors and grammatical mistakes. So, we first developed an acoustic model considering pronunciation proficiency among several learners. Next, we employed a language model based on an N-gram trained by generated texts that reflect tendency of grammatical mistakes made by Japanese learners. We obtained 3.7 point gain of the error detection accuracy at maximum by considering pronunciation proficiency and recognition using multiple models.

†1 東北大学大学院工学研究科
Graduate School of Engineering, Tohoku University

1. はじめに

近年の国際化や情報処理技術の発達とともに，コンピュータを利用して外国語を学習する CALL（Computer-Assisted Language Learning）システムが注目され，様々な研究が行われている[1),2)]．従来の CALL システムの多くは，読む，書く，聞く能力に焦点を当てたものであり，話す能力を伸ばすための CALL システムはほとんど存在しなかった．最近では話す能力を伸ばすための CALL システムの研究も進んできてはいるが，それらのシステムは発音や韻律を評価するものが多く，また，学習者の発話内容があらかじめ決められた受動的なシステムとなっていることが多い．

しかし，話す能力を向上させるためには，学習者が自分で話す内容を考え，自発的な発話をすることが重要である．また，正しい発音の練習だけではなく，ロールプレイなどによって，正しい文法や表現を実際に使う練習をする必要があると考えられている．そこで近年では，学習者が自発的に発話を考え，コンピュータと簡単な会話をしながら学習を行うことができる，音声対話型の CALL システムの研究が進められてきている[3),4)]．

音声対話型 CALL システムを実現するためには，学習者の発話を認識して円滑に対話を進め，さらに，誤りが含まれていたならその箇所を指摘するなどのフィードバックを行う，という 2 つの機能が必要であると考えられる．しかし，一般的に外国語学習者の発話には発音の誤りや文法的な誤りが含まれる場合が多く，それによって音声認識精度の向上が妨げられている．また，習得しようとしている言語の理解を深め，上達を促す最適なフィードバックの与え方，その役割についてもさまざまな議論が行われてきており，これについても検討が必要である[5)]．

外国語学習者の発話の高精度な音声認識，そしてフィードバックの方法についての 2 つが検討課題であるが，コンピュータとの対話形式で学習することを想定しているので，適切なフィードバックを与えるためには，学習者が発話した内容を，たとえ誤りが含まれていたとしてもその通りに認識することが前提条件となってくる．そこで本稿では，文法や表現の学習ができる音声対話型 CALL システムを実現するために，日本人による英語発話の音声認識の高精度化に関する検討を行った．

2. 音声対話型 CALL システム

本研究で想定している CALL システムでは

(1)　学習者が英会話に使う基本的なフレーズ（文法，表現）を学ぶ

(2)　そのフレーズを用いた会話をコンピュータと行う
(3)　コンピュータが文法的な誤り箇所を指摘する

という流れで英語の学習を進めていくこととする．まず事前学習を行うことで，学習者は対話で必要となるキーワードを知ることができるため，システムへの応答にもそれと同じか近い表現を使うことが予想できる[4]．学習者が発話すべき文法的に正しい文を**正解文**と呼ぶことにする．

3. 日本人英語音響モデル

3.1 先行研究

学習者が誤った発音で英語を発話してしまう問題について，我々は ERJ データベース[6] の日本人による読み上げ英語音声から音響モデルを学習することによって認識精度の向上を試みた[7]．また，子音の直後に母音を挿入する誤りに対して，HMM の状態数を増やすことによって対応し，発音がばらつくことに対して，HMM の出力確率分布の混合数を大幅に増やすことによって対応していた．

しかし，これらの対策だけでは外国語の発音が学習者の習熟度によって大きく異なることに対して不十分であると考えられる．特に，母音の挿入や英語子音の日本語子音発声などの発音誤りは，学習者によって誤り方が異なり，そもそも誤りが出現しない場合もある．このような非母語話者間の発音のばらつきに対しては，HMM の出力確率分布の混合数を増やすだけでは不十分であると考え，各話者の発音レベルが異なることを考慮して音響モデルの構築を行っていく．

3.2 発音レベル

ERJ データベース[6] には，“発声者が意図した音素が適切に生成されているか否か” という発音評定ラベルが含まれている．この評定点に応じて学習データをいくつかに分割し，複数の音響モデルを学習することによって，様々な学習者の発音レベルに対応することを試みる．データの分割には様々な方法が考えられるが，本稿では事前実験の結果より，ラベル付けされた全文発話，単語発話を 3 分割（*Low*: 1.86∼2.78，*Middle*: 2.78∼3.26，*High*: 3.27∼5.00）して 3 つの音響モデルを構築することにする．また，分割された各データ群の発話単語数，つまり学習データ量がほぼ等しくなるようにしている．音響分析条件と音響モデルの学習条件を**表 1**，**表 2** に示す．

3.3 評価実験

構築した音響モデルの性能を評価するために，音素認識実験を行った．

3.3.1 テストデータ

想定しているシステムと同じような条件にするため，実験に用いるテストデータの収集は以下の流れで行った．

(1)　システム側で簡単な基本フレーズを含んだ正解文を準備する．
(2)　学習者が発話練習をして正解文を覚える．制限時間は設けない．
(3)　正解文の日本語訳を見ながら発話，録音するが，その中には一部を入れ換えた，発話時に初めて考えてもらう文も含まれている．また，コンピュータから合成音声による質問文もしくは応答文を発話させ，対話形式にする．

収集したテストデータのうち，正解文が “I’m an office worker. I work at a car company.” であるものを**表 3** に示す．

3.3.2 実験結果

全データから学習した音響モデルを *All*，分割したデータから学習したそれぞれの音響モデルを *Low*，*Middle*，*High* とする．さらに，この三つのモデルを同時に用いて並列に認識を行い，デコーダが出力するスコアが最も高い仮説文を認識結果とする手法を，*Score* と呼ぶことにする．**表 4** のような条件で，HMM の出力確率分布の混合数を変化させながら実験を行った結果を**図 1** に示す．

この結果を見ると *Low* と *Score* がほぼ同じ精度であることが分かるが，これはテスト

表 1　音響分析条件
Table 1　Condition of speech analysis

フレーム間隔	10 ms
サンプリング周波数	16 kHz
分析窓	ハミング窓（25 ms）
特徴量	MFCC，ΔMFCC，ΔΔMFCC，対数パワー，Δ 対数パワー，ΔΔ 対数パワー（計 39 次元）

表 2　音響モデル学習条件
Table 2　Condition of acoustic model training

種類	monophone HMM
状態数	5
遷移パターン	Left-to-Right
全学習データ	ERJ データベース[6] の男女それぞれ 95 人分（約 21 万単語）

表 3 収集した発話文の例
Table 3 Example of the test data

頻度	発話文
6	I'm an office worker. I work at a car company.
2	I'm an office worker. I worked at a car company.
1	I'm an office worker. I work at car company.
1	I'm a worker. I work at a car company.
1	I'm a office worker. I work at a car company.
1	I'm a office worker. I worked at an car company.

表 4 実験条件
Table 4 Condition of experiment

デコーダ	Julius-4.1.5
テストデータ	男性 14 人，女性 1 人， 42 種類，計 441 発話 平均単語正解精度：88.3 %

データの多くが *Low* のモデルで認識したときのスコアが一番高くなり，発話者別で見ても *Low* で認識したときに結果が最も良い話者が大部分を占めるからである．テストデータの発話者のうち，128 混合の各モデルで認識したときの 7 人分の結果を**表 5** に示す．それぞれのモデルで最も認識精度が高い発話者がおり，発音レベルに応じて学習データを分割したことによる有効性を確認することができた．次に，*Score* と *All* を比較してみると，128 混合までは全データで学習した場合よりも 2.1~3.9 ポイント改善しているが，256 混合ではあまり効果が出ていない．これはデータを分割したことによって，各音響モデルの学習データ量が減ってしまったことが原因だと考えられる．この問題に対応するために，次節では全データで学習したモデルを各レベルに適応させる方法について検討していく．

3.4 MLLR 法を用いたレベル別学習

全データで学習したモデルを各レベルに適応させることによって学習データ不足の問題を解決し，精度の向上を試みる．ここでは適応手法として MLLR 法を用いた．学習の手順を以下に示す．

(1) 全データから一つの音響モデル (*All*) を学習する．
(2) 学習データを *Low*，*Middle*，*High* の三つに分けてそれらを適応データとし，1. で作成したモデルを各レベルに適応させて，三つの音響モデルを構築する．
(3) 2. で作成した各モデルに対して，それぞれの適応データを用いて学習を繰り返す．

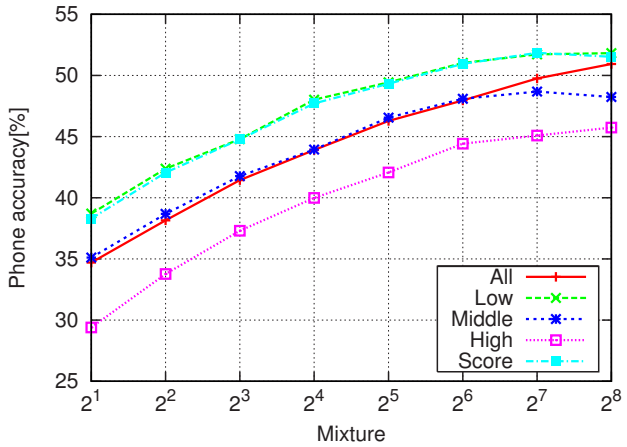

図 1 各モデルの音素認識精度
Fig. 1 Comparison of each models

表 5 発話者別に見た音素認識精度
Table 5 Phone accuracy with respect to the learners

発話者	*Low*	*Middle*	*High*
A	**59.4%**	54.6%	53.0%
B	**60.9%**	54.2%	50.5%
C	**57.7%**	49.5%	48.3%
D	49.8%	**51.5%**	40.6%
E	**51.5%**	49.6%	46.5%
F	**57.9%**	54.6%	48.0%
G	43.3%	43.5%	**51.9%**

このような手順でモデル *MLLR-Low*，*MLLR-Middle*，*MLLR-High* を作成した．また，各モデルを同時に用いて並列に認識を行い，スコアが最も良い仮説文を認識結果とする方法を *MLLR-Score* とする．

256 混合のとき，学習回数による音素認識精度の変化を**図 2** に示す．*MLLR-Middle* と *MLLR-High* は学習を繰り返すと認識精度が下がっていくが，これは 3.3 節でも述べた通り，テストデータ中の発話者の多くが *Low* のモデルで最も良い結果となるため，むしろ精度良くモデル化できているということになる．そのため，入力された発話文ごとにスコアが

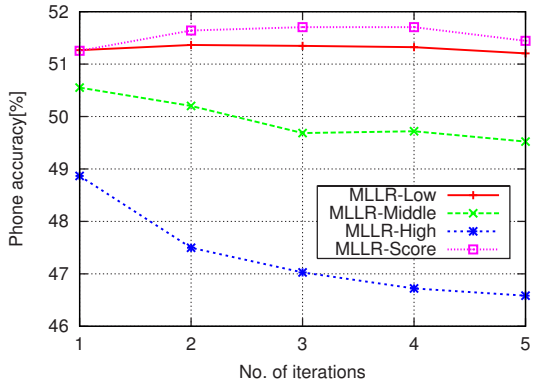

図 2 学習回数と音素認識精度
Fig. 2 Phone accuracy with respect to no. of iteration

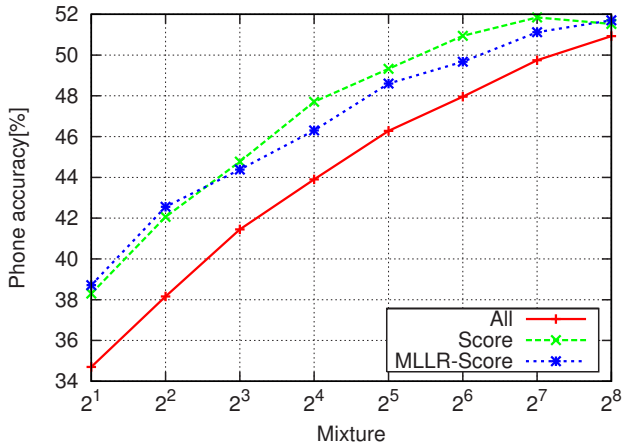

図 3 各モデルの音素認識精度
Fig. 3 Comparison of each models

高い仮説文を認識結果とする *MLLR-Score* では，ある程度学習を繰り返す方が結果が良くなっている．

次に，*All*，*Score*，*MLLR-Score* を比較した結果を**図 3** に示す．*MLLR-Score* では，それぞれの混合数ごとに適切な学習回数となるようにしている．この結果より，混合数が大きいときならば，全データで作成したモデルを MLLR 法でレベル別に学習させることが有効であることが分かる．128 混合のとき *Score* で 51.84 ポイント，256 混合のとき *MLLR-Score* で 51.71 ポイントであるが，混合数をさらに増やすと *Score* よりも良くなることが期待できる．

4. 生成したテキストからの N-gram の学習

4.1 誤りルール

我々は，正解文に誤りルールを適用することによって，日本人が犯しそうな誤りを含む文を大量に生成し，それを用いて N-gram を学習する方法を提案した[8]．この誤りルールには 3 種類あり，以下でこれらについて説明していく．

(1) コーパスによる誤りルール

コーパスによる誤りルールは，日本人による英語発話データ[9] から，どの語句をどのように誤るかという情報を取り出したものである．コーパス中の誤り箇所には，正しくはこう言うべきであった，というタグが人手で付与されているため，これによって日本人英語発話の誤り傾向をルール化することができる．

(2) 一般的な誤りルール

コーパスに登場する誤りのみでは不十分であり，全ての誤りに対応できるとは言えない．なぜなら，コーパスに出現していない，あるいは非常に頻度が少ない単語の活用変化，単数・複数の誤り等には対応できないからである．そのため，正解文に品詞のタグ付けをし，品詞ごとにさまざまな誤りルールを適用して生成する誤り文のバリエーションを増やすことを考えた．このようなルールを一般的な誤りルールを呼ぶことにする．なお，単語の品詞タグ付けには Brill's Tagger[10] を用いている．

(3) シソーラスによる誤りルール

学習者が正解単語とは異なる単語を発話し，その単語がコーパスに登場していなかった場合，認識することは不可能になってしまう．このような未知語問題を解決するため，WordNet[⋆1] を用いて正解単語の類似語と上位語を取得し，認識不可能な単語を減らすことを試みている．このようなルールをシソーラスによる誤りルールと呼ぶ．

正解文に誤りルールを適用することによって生成したテキストで，N-gram を学習する過

⋆1 WordNet, http://wordnet.princeton.edu/

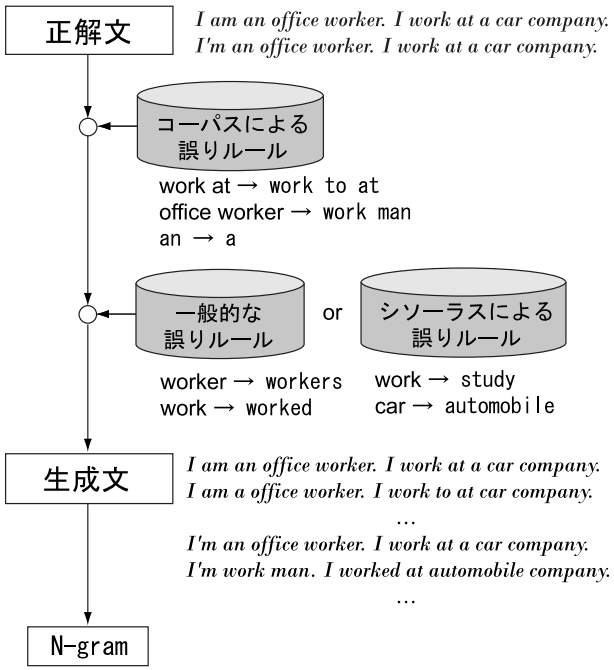

図 4 学習テキスト生成の手順
Fig. 4 Procedure of text generation

程を**図 4** に示す．もし正解文に短縮形，または短縮しうる表現が含まれている場合，短縮形を含む文，含まない文の両方を正解文としている．また，誤りルールを適用するかどうかは誤り単語出力確率によって決定しており，適用する際には，最初にコーパスによる誤りルールを適用し，それでも正解単語である場合は一般的な誤りルールもしくはシソーラスによる誤りルールを適用する．この誤り単語出力確率の値を変えることによって，学習テキストの文法的な誤り具合を調整することができる．

4.2 単語認識

第 3 節で構築した各音響モデルと，4.1 節で説明した手法によって学習された N-gram 言語モデルを用いて，表 4，**表 6** のような条件で単語認識実験を行った．ここで，生成したテキストにおける短縮形を含む文，含まない文の割合は正解文ごとに適切な値に設定し，一般的な誤りルールを適用する確率は，1 − シソーラス適用確率 である．

表 6 実験条件
Table 6 Condition of experiment

学習テキスト生成数	100,000/正解文
シソーラス適用確率	0.4

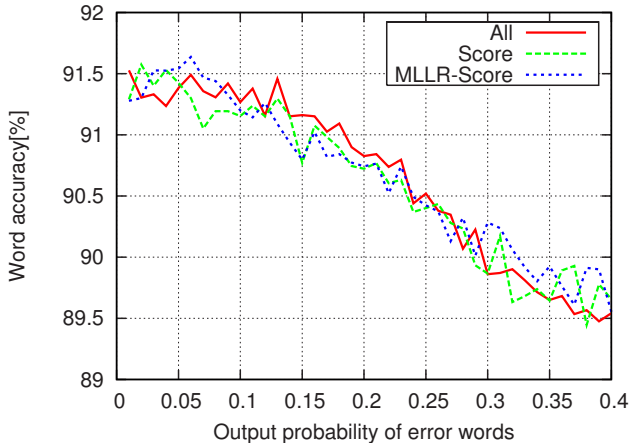

図 5 各モデルの単語認識精度
Fig. 5 Comparison of each models

誤り単語出力確率を変化させながら単語認識した結果を**図 5** に示す．*Score*，*MLLR-Score* ともに発音レベルが *Low*，*Middle*，*High* の 3 種類の音響モデルを用いているが，事前実験の結果より，それぞれのモデルは同じ混合数にし，各手法で事後的に最適な混合数を設定している（*All*：256 混合，*Score*：128 混合，*MLLR-Score*：64 混合）．また，テキスト生成時における誤り単語出力確率も，3 つのモデルで同じ（0.01〜0.4）としている．図 5 より，確率が 0.06 のときに *MLLR-Score* が最も良い値になっているものの，*All* と比べてほとんど同じ結果になってしまった．

実際に学習システムとして使用する際には，学習者の発話中に含まれる誤りをどの程度適切に指摘できるかが重要となる．そこで，単語認識精度ではなく，誤りの指摘の精度で改めて評価を行う．再現率（recall）と適合率（precision）を

$$recall = \frac{\text{システムが正しく指摘した誤り数}}{\text{発話文中の誤り数}} \quad (1)$$

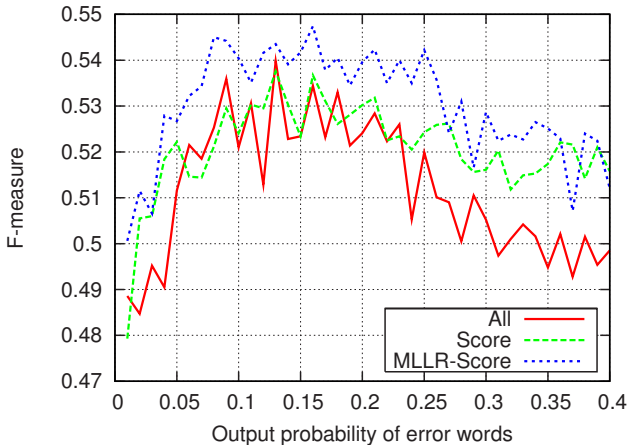

図 6 各モデルの F 値
Fig. 6 Comparison of each models

$$precision = \frac{\text{システムが正しく指摘した誤り数}}{\text{システムが指摘した誤り数}} \quad (2)$$

により求め，これらの調和平均である F 値を以下の式で求める．

$$F\text{-}measure = \frac{2 \cdot recall \cdot precision}{recall + precision} \quad (3)$$

F 値で各手法を評価した結果を，**図 6** に示す．*All* と比較すると，確率 0.04 のときに最大で 3.7 ポイント改善することができた．単語認識精度ではなく誤りの指摘の精度の場合，*MLLR-Score* が最も良いということが確認でき，発音レベルに応じて複数の音響モデルを学習したことによる有効性を示すことができた．文法的な誤りについては，事前実験で各発音レベルに対して様々な文法的な誤り方の言語モデルとの組み合わせを検討してみたが，効果はほとんど見られなかった．これはテストデータの発話者の中で，発音の誤り方には差があるが，文法的な誤り方はそれほど違いがなかったからだと思われる．

5. ま と め

本稿では，日本人英語発話の音声認識の高精度化のために，音響モデルの学習データを発音レベルに応じて分割し，複数のモデルを構築することを試みた．音素認識実験を行った結果，全データで学習したモデルよりも最大で 3.9 ポイント改善することができた．次に，複数の音響モデルと，生成したテキストにより学習した N-gram を組み合わせて単語認識実験を行ったが，効果はあまり見られなかった．しかし，F 値による誤りの指摘の精度で評価した場合，最大で 3.7 ポイントの改善が確認できた．これらの結果より，本手法の有効性を示すことができた．今後は文法的な誤り方において，発話文や正解文ごとに適切な誤り単語出力確率を決定する方法について検討していく．

参 考 文 献

1) 中川聖一，牧野正三，壇辻正剛：音声言語処理技術を用いた語学学習システム，日本音響学会誌，Vol.59, No.6, pp.337-344, (2003).
2) Doremalen, J.V, Cucchiarini, C. and Strik, H.: Optimizing Automatic Speech Recognition for Low-Proficient Non-Native Speakers, Eurasip Journal on Audio, Speech, and Music Processing, pp.1-13, (2009).
3) 阿部一彦，田中和世，河原達也，清水政明，壇辻正剛：対話型英語学習システムにおける日本人英語音声認識精度の検討，音響講論，2-5-20, pp.113-114, (2002).
4) Kweon, O.P., Ito, A., Suzuki, M. and Makino, S.: A grammatical error detection method for dialog-based CALL system, Journal of Natural Language Processing, Vol.12, No.4, pp.137-156, (2005).
5) Vries, B.P.D., Cucchiarini, C., Strik H. and Hout, R.V.: The Role of Corrective Feedback in Second Language Learning: New Research Possibilities by Combining CALL and Speech Technology, Proc. L2WS, O4-05, (2010).
6) 峯松信明，富山義弘，吉本啓，清水克正，中川聖一，壇辻正剛，牧野正三：英語 CALL 構築を目的とした日本人及び米国人による読み上げ英語音声データベースの構築，日本教育工学会論文誌，Vol.27, No.3，pp.259-272，(2004).
7) Ito, A., Tsutsui, R., Makino, S. and Suzuki, M.: Recognition of English Utterances with Grammatical and Lexical Mistakes for Dialogue-based CALL System, Proc. Interspeech, pp.2819-2822, (2008).
8) Anzai, T., Seongjun, H. and Ito, A.: Grammatical Error Detection from English Utterances Spoken by Japanese, Proc. 2nd APSIPA Annual Summit and Conference, pp.482-485, (2010).
9) 和泉絵美，内元清貴，井佐原均：日本人 1200 人の英語スピーキングコーパス，アルク，(2004).
10) Brill, E.: A simple rule-based part of speech tagger, Proc. ANLP-92, 3rd Conf. on Applied Natural Language Processing, pp.152-155, (1992).