情報処理学会論文誌　Vol. 50　No. 10　2574–2582 (Oct. 2009)

日本語母語話者による英語発話認識のための言語モデル適応化

原　田　貴　史[†1]　　山　本　誠　一[†1]

様々な言語を母語とする人が集まる国際的な会議やビジネスの場などでの発表・交渉・議論は主に英語でなされることから，英語発話ドキュメント処理においては，英語母語話者の発話に加えて第二言語話者による英語発話認識の高性能化も重要である．第二言語話者による発音や表現は母語の干渉を受け，英語母語話者の発音や表現とは異なる点があるため，英語母語話者の音声データを用いて学習を行った英語音声認識装置では，第二言語話者による英語発話に対する音声認識性能は劣化する．本論文は，日本語母語話者による英語発話の認識率を向上させる取り組みとして言語モデル適応手法を提案し，その認識実験結果について記述している．日本語を母語とする英語学習者による英訳文を収集した学習者コーパスを用いて学習を行った言語モデルを，同一ドメインの大規模な英文コーパスで学習した言語モデルに線形補間した言語モデルを開発した．この言語モデルの使用により，日本語母語話者により英語表現された英語発話の認識率が向上することが確認された．また，学習者コーパスの作成方法の違いが認識性能に与える影響についての検討結果も記述している．

Language Model Adaptation of an English Speech Recognizer for Japanese

Takashi Harada[†1] and Seiichi Yamamoto[†1]

English is spoken during presentations and discussions at international conferences and business meetings by many people who have different mother tongues, and there is a need for automatic speech recognition which transcribes English speech by the second language speakers as well as native speakers. English as a second language (ESL) speakers often have a distinct accent, as well as different lexical and syntactic characteristics. As the language model of a speech recognition system is usually trained with transcribed speech data or text data collected in English-native countries, and speech recognition performance is severely affected when the lexical or syntactic characteristics in the training and recognition tasks differ, speech recognition performance is expected to be degraded by mismatch of lexical and syntactic characteristics between native speakers and ESL speakers, as well as the distinction between their accents. This paper proposes a language model adaptation to compensate for mismatch between those characteristics of native speakers and ESL speakers. A language model was created by linearly interpolating a language model trained with a learner corpus and a language model trained with a large native English corpus of the same domain. Some experiments verified that language model adaptation was effective for compensating for the mismatch between the lexical and syntactic characteristics of native speakers and ESL speakers. This paper also describes some relations between the methodology of creating the learner corpus and recognition performance.

1. はじめに

近年，音声認識性能の向上を受け，音声認識技術の新たな活用分野として，音声発話の検索や要約などを行う音声ドキュメント処理技術の研究開発が注目をあびている[1)–3)]．一方，世界経済のグローバル化にともない，英語を第二言語とする人が発表・交渉・議論を英語で行う機会は増大しており，英語音声ドキュメント処理においては英語母語話者の発話に加えて第二言語話者による英語発話認識の高性能化は重要な技術課題である．しかし，第二言語話者による発音や表現は母語の干渉を受け，英語母語話者の発音や表現と異なる点が多々あるため，英語を母語とする話者のデータを用いて学習を行った英語音声認識装置では，第二言語話者による英語に対する音声認識性能は大きく低下する．特に音節の構造や構文構造が大きく異なる日本語を母語とする日本人による英語発話を正しく認識し，音声ドキュメント処理を行うことは困難である．このような視点に立ち，筆者らの研究グループは日本語母語話者の英語発話を認識する技術の研究開発を進めてきた[4),5)]．

音声認識システムでは，音響的な特徴を表現する音響モデルと単語列の出現確率などを表現する言語モデルを準備し，入力音声に対する両モデルの出力確率を最大化する仮説を認識結果として出力する．このため日本語母語話者による英語音声の認識性能の向上を図るには，日本語母語話者の英語発音の特徴を表現した音響モデルの開発に加えて，日本語母語話者による英語表現における言語特性を考慮した言語モデルを構築することが必要である．日本語を母語とする話者の英語音響モデルの開発やその基盤となる英語発話データの収集などは，主に情報処理技術の適用による外国語教育の高度化・効率化などの視点から行われ

†1 同志社大学
Doshisha University

てきており，「外国語学習用読み上げデータベース」などとして開発されている[6),7)]．一方，日本語母語話者による英語表現の言語特性を表現する確率的言語モデルの開発には，日本語母語話者による英語表現を多量に収集したコーパスを構築することが必要となる．第二言語話者による英語表現を収集したコーパスとしては，外国語教育への利用を主目的として，英語学習者の誤用などを実証的に研究するための学習者コーパスが多くの言語の母語話者に対して開発されており[8),9)]，日本語母語話者による学習者コーパスも開発されている[10)]．しかし，確率的な言語モデルの構築を目的として第二言語話者による種々のドメインでの発話を収集したコーパスはほとんど例がなく，また母語話者の発話を収集したコーパスと比較して今後とも大規模なコーパスの開発は困難であると予想される．

日本語母語話者の英語発話の認識性能を向上させる方法として，日本語母語話者による英語発話の誤用を規則化して，その規則により言語モデルの修正を行う方法も提案されているが[11)]，発話中に同じ語を繰り返して使用することや無生物主語を用いる表現が少ないなどの必ずしも誤用とは見なされない発話構成上の傾向などを言語モデルとして組み込むことは困難である．

一方，音声認識の対象となるドメインでの大規模な発話の収集が困難な場合に利用される言語モデル構築手法として，言語モデル適応化技術がある．言語モデル適応化技術は，音声認識対象となるドメインでの少量の発話から学習した言語モデルと，音声認識対象のドメインと関連するドメインでの多量のコーパスから学習した言語モデルとの補間処理などを施す技術であり，音声認識対象のドメインで収集された少量の発話データからでも比較的高性能な言語モデルを構築し認識性能向上を実現している[12)]．

本論文では，日本語母語話者による英語発話の認識性能を向上させる手法として，言語モデル適応化手法，すなわち，日本語母語話者の英語発話認識用の言語モデルとして，音声認識対象のドメインでの日本語母語話者による少量の英語翻訳文から学習した言語モデルと，同一ドメインでの多量の英語テキストコーパスから学習した言語モデルとを線形補間した言語モデルを用いることを提案し，その言語モデルによる認識性能評価について記述している．まず，旅行会話などに関する多言語パラレルテキストコーパスである BTEC（Basic Travel Expression Corpus）[13)] より選択した一部の日本文に対して，日本語母語話者が英訳した文を収集したテキストである学習者コーパス[14)] を用いて学習を行った英語言語モデルと，BTEC での英語テキストを用いて学習した言語モデルとを線形補間することにより，旅行会話などに関する日本語母語話者用の言語モデルを構築した．構築した言語モデルを用いて認識実験を行い，同一ドメインでの日本語母語話者による英語発話に対する認識性能の向上を確認した．さらに，他のドメインでの発話に対しても認識性能向上をもたらすことを確認した．また，学習者コーパスの構築法やコーパスの量による認識性能の違いを検証した．なお，評価実験に使用した被験者の英語発話には語彙選択上の誤りや文法的な誤りが含まれている．正確な意思疎通を重視し，語彙選択上の誤りや文法的な誤りを取り除いた発話のみを認識対象として使用すべきという考え方もあるが，発話に含まれるこの種の誤りがどの程度意思の疎通に支障をきたすかは，文脈情報や韻律情報などの他の要素も複雑にからむため詳細は不明である．このため，本研究ではこの種の誤りが含まれる発話も含めて，日本語母語話者による英語発話を高精度で認識する技術の開発に主眼を置いている．

以下，次章では言語モデル適応化処理の検討に使用した音声認識システムの概要について記述する．3 章では本検討に使用した学習者コーパスとそれを利用して開発した言語モデルについて述べる．4 章では 3 章で述べた言語モデルを用いた各種の認識実験結果について記述する．5 章では認識性能と学習者コーパスの構築法との関連について考察する．6 章は本論文で記述した内容のまとめである．

2. 日本人用英語音声認識システムと音響モデル

言語モデル適応化処理による英語母語話者と日本語母語話者の英語発話の語彙選択上の違いや構文的な違いを補償する効果を検証するため，ATR（Advanced Telecommunications Research）で開発された多言語音声認識システム ATRASR（ATR Automatic Speech Recognizer）[15)] を基本とした試験システムを開発した．以下，ATRASR の概略と日本語母語話者の英語音声を認識するために開発した音響モデルについて記述する．

(1)　音響分析

音声分析は，16 kHz サンプリングで線形 16 bit 量子化された音声データに対し，フレーム長 20 msec，フレームシフト 10 msec の Hamming 窓関数を用いて行われる．音響パラメータはフレームごとに 12 次元 MFCC，12 次元 ΔMFCC，Δ パワーを計算し，計 25 次元の特徴ベクトルとして求められる．音響モデルの構成単位としては，**表 1** に示すように，Wall Street Journal コーパスで使用されている 43 種類の音素モデル[16)] に無音モデルを加えた計 44 種類が使用される．

(2)　音響モデル

音響モデルとして，英語母語話者の発話データから学習した native 音響モデルと日本語母語話者の英語発話データから学習した non-native 音響モデルの 2 種類の音響モデルを準備した．native 音響モデル，non-native 音響モデルともに HMM（Hidden Markov Model）

表 1　音素リスト
Table 1　List of phonemes for an English ASR (Automatic Speech Recognizer).

Stops	B, D, G, P, T, K, DX
Affricates	JH, CH
Fricatives	S, SH, Z, ZH, F, TH, V, DH
Nasals	M, N, NG
Semivowels	L, R, W, Y, HH
Vowels	IY, IH, EH, EY, AE, AA, AW, AY, AH
	AO, OY, OW, UH, UW, ER, AX, IX, AXR

としては HMNet[17] を用いている．native 音響モデルは，Wall Street Journal コーパス中の不特定話者モデル学習用として推奨されている男性 143 名および女性 141 名による合計約 37,000 発話を利用して学習されている[18]．non-native 音響モデルは，「日本語母語話者による英語読み上げ音声 DB」[7]（男性 100 名，女性 101 名）の合計約 24,000 発話を用いて学習されている．後述するように，TOEIC スコアのきわめて高い一部の被験者にも対応するように，native 音響モデルと non-native 音響モデルを同時に音響モデルとして使用し，尤度の高い仮説を認識結果として選択する構成を採用している．この際，使用する音声認識エンジンの構成上の制限により，音響モデルのトポロジは同一である必要があるため，native 音響モデル用に決定されたトポロジ構造をそのまま non-native 音響モデルにも採用しており，non-native 音響モデルでは出力分布確率と遷移確率のみをパラメータとして学習している．なお，non-native 音響モデルの学習に使用した「日本語母語話者による英語読み上げ音声 DB」では表 1 に示す音素表記のうち，弾音 DX と母音 IX は使用されていないため，これらの音響モデルについては native 音響モデルの初期モデルがそのまま使用されている．日本語母語話者用の英語音響モデルの開発に使用した「日本語母語話者による英語読み上げ音声 DB」でトポロジから学習した場合と，英語母語話者用の音響モデルのトポロジをそのまま用いてパラメータ学習のみを行った場合の音響モデルを比較した場合，読み上げ発話セット RSS に関しては同等の性能を示すことを確認済みである．

(3)　**単語発音辞書**

単語発音辞書には旅行会話に関連する約 35,000 語が登録されている．単語発音辞書の発音記号は，表 1 に記載された音素表記と無音を表す表記により表現されている．なお，複数の発音記号列が付与されている単語が単語発音辞書中には存在するが，これら複数の発音記号列の生起確率は考慮されておらず，同一の生起確率と見なして音響尤度が計算される．

(4)　**言語モデル**

言語モデルは 2 種類準備した．一方の言語モデルは旅行会話などに関する大規模対訳コーパス BTEC の約 500,000 文（語彙数：約 3 万語）から学習した単語 bi-gram および tri-gram である．以下，この言語モデルを native 言語モデルと略称する．他方の言語モデルは学習者コーパスから学習された言語モデルと native 言語モデルとを線形補間した non-native 言語モデルであり，詳細は次章で記述する．

(5)　**認識エンジン**

認識エンジンには ATR で開発された ATRASR を用いた．本認識エンジンでは 2 パス探索が採用されており，第 1 パスでの言語モデルに単語 bi-gram を使用し，認識結果はワードグラフとして出力される．第 2 パスでは単語 tri-gram を使用し，ワードグラフで表現された仮説の尤度の再計算を行い，認識結果として出力する．

3. 英語学習者コーパスと言語モデル

3.1　学習者コーパス

日本語母語話者の英語表現と英語母語話者による英語表現の違いを補償して音声認識率を向上させる手段としての言語モデル適応化処理の効果を検証するために，500 名の日本語母語話者の被験者により翻訳された英訳文約 150,000 文を収集した学習者コーパス[14] を利用した．この学習者コーパスは，各被験者が辞書を利用しないで英訳した文と，日英バイリンガルによって与えられた各英訳文に対する主観的な訳質評価値などの付属情報から構成されている．各被験者の英語能力は同時期に受験した TOEIC スコアで評定されており，被験者の TOEIC スコアは 300 点から 990 点にわたっている．

日本語母語話者による英語学習者コーパスの構築に際して，提示された英訳課題文は，**表 2** に示すように，BTEC からランダムに選択された旅行会話文 296 文，その他の会話文 568 文，そして中学・高校で使用されている英語教科書の英訳問題から抜粋した 636 文の計 1,500 文である．これらの英訳課題文 1,500 文を 5 セットに分割（300 文 × 5 セット）し，各被験者は各々300 文を翻訳している．

なお，その他の会話にはビジネス会話用フレーズブック，留学用フレーズブックなどの会話から選択された課題文が含まれている．旅行会話，その他の会話，教科書の語彙数は表 2 に示すように，各々約 5 K，11 K，8 K であるが，その他の会話および教科書は各々BTEC 中の語彙の約 54%，約 48%をその語彙として含んでいる．

学習者コーパスの総単語数は約 120 万語で，語彙数は約 17,000 語である．学習者コーパ

表 2　学習者コーパスのタスクと文数および語彙数
Table 2　Domains of the learner corpus and some features.

タスク	課題文数	有効英訳文数	総単語数	語彙数
旅行会話（BTEC）	296	28 K	149 K	5 K
その他の会話	568	53 K	530 K	11 K
教科書	636	65 K	527 K	8 K

ス中の語彙はすべて発音単語辞書中に含まれており，未登録語は存在していない．なお，翻訳作業終了後に翻訳文は検査され，スペルミスのみが修正されている．

3.2　日本語母語話者用英語言語モデル

日本語母語話者用の言語モデルとして，旅行会話文を課題文とした有効英訳文約 28,000 文について形態素解析を行い，解析結果に対し back-off スムージングにより単語 bi-gram および tri-gram モデルの開発を行った．言語モデルの開発に際しては，英語母語話者用言語モデルと同一の単語辞書（約 35,000 語）を使用した．さらに，得られた言語モデル（単語 bi-gram と単語 tri-gram）$P_j(w_i|h_i)$ と BTEC から構築された英語母語話者用言語モデル $P_e(w_i|h_i)$ とを，式 (1) に示すように確率値の線形補間を行い，non-native 言語モデル $P_n(w_i|h_i)$ を開発した．ここで，$P_j(w_i|h_i)$ は時刻 i での単語 w_i 以前の単語履歴 h_i による単語 w_i の条件付き確率を示しており，単語履歴 h_i は単語 bi-gram では $h_i = w_{i-1}$，単語 tri-gram では $h_i = w_{i-2}w_{i-1}$ で表現される．

$$P_n(w_i|h_i) = \lambda P_j(w_i|h_i) + (1-\lambda)P_e(w_i|h_i) \quad (1)$$

なお，補間処理の際に言語モデルにかかる重み係数 λ は，クロスバリデーション法[19] でセットを変えて算出したが，得られた重み係数 λ の値は各セットによって 0.29 から 0.52 の範囲にわたっており，平均値である $\lambda = 0.46$ を使用している．

4.　音声認識実験

4.1　評価用音声データ

TOEIC スコアが既知の日本語母語話者の男性 28 名，女性 27 名の計 55 名（18 歳から 25 歳の大学生もしくは大学院生）から収録した英語発話を，評価用音声データとして使用した．録音条件は，使用したマイクロホン特性は除いて，音響モデルを開発するのに用いた音声データを収録したのと同じ条件である．このため，以下の音声認識実験ではマイクロホン特性の違いを補償するため，Cepstrum Mean Subtraction 法を用いている．

発話内容は BTEC から 14 文および中学・高校の英語教科書から 36 文をランダムに選択

表 3　発話課題文と被験者による英訳例
Table 3　Examples of the task sentences and their translations by the subjects.

そのお祭りはいつですか	What is the day of the festival?
	When is the festival?
	Where is the festival?
家まで送りましょうか	Will you send you to your home?
	Why don't you to take you to your home?
	Shall I take you your home?
授業前にあなたは何をしてましたか	What did you do before your lessons?
	What did you do before the class?
	What were you doing before class?

した日本文を，被験者がその場で英訳した発話である．すなわち，**表 3** に例を示すような日本語の課題文 50 文を順次被験者に提示し，課題文をその場で英語で表現・発声してもらい，録音した英語音声を英語自由発話セットとした．なお，BTEC から選択された課題文を英訳した 770 発話を自由発話セット B（SSS-B; Spotaneous Speech Set と略称），教科書から選択された課題文を英訳した 1,980 発話を自由発話セット T（SSS-T と略称）とした．英語自由発話セットの各発話は録音終了後，各被験者が録音された発話に基づきテキスト化し，その後テキスト表記の検査によりスペルミスのみが修正されている．テキスト化された英語自由発話セットの全話者の発話内容に，単語発音辞書に含まれていない未登録語は存在しなかった．なお，同一被験者 55 名が BTEC からランダムに選択した英文 180 文を読み上げた音声を録音し，読み上げ発話セット（RSS; Read Speech Set と略称）とした．

英語音声読み上げ発話セット RSS を入力し，音響モデルとして native 音響モデルのみを使用した場合，non-native 音響モデルのみを使用した場合，native 音響モデルと non-native 音響モデルの両音響モデルを使用し両者の認識結果から尤度の高い結果を採用した場合の認識結果の比較を**表 4** に示す．なお，この場合は読み上げ発話セット RSS の認識実験のため，言語モデルとしては native 言語モデルを使用している．参考データとして 2 名の英語母語話者による同一英文の読み上げに対する認識実験結果を付記している．表 4 から明らかなように，日本語母語話者の調音特性は英語母語話者とは大きく異なり，non-native 音響モデルの使用は native 音響モデルの使用に比較して音声認識性能を大幅に向上させている．また，性能向上効果の大半は non-native 音響モデルの使用によるものであるが，TOEIC スコアのきわめて高い被験者の一部には non-native 音響モデルより native 音響モデルを使用する方が WER（Word Error Rate）の低い被験者が存在する．このため，以下に述べる言語モデルの適応化処理に関する検討では両音響モデルを使用する．

表 4　音響モデルの違いによる読み上げ発話セット RSS での WER の比較
Table 4　Comparison of WER for the RSS with the native acoustic model, the non-native acoustic model, and both sets of the acoustic models.

	native AM	non-native AM	both sets of AM
男声	70.2%	26.4%	26.4%
女声	63.3%	27.5%	26.4%
英語母語話者	17.7%	-	-

表 5　native 言語モデル，non-native 言語モデルと cross-domain 言語モデルによる性能比較
Table 5　Comparison of performances with the native language model, the non-native language model, and cross-domain language model.

	SSS-B			SSS-T		
	bi-gram	tri-gram	WER	bi-gram	tri-gram	WER
native LM	49.5	33.1	30.6%	89.0	73.2	39.9%
non-native LM	33.6	21.0	26.2%	79.4	55.2	36.7%
cross-domain LM	38.1	21.6	26.6%	32.8	13.8	20.4%

4.2　言語モデル適応化処理の効果

日本語母語話者の英語表現と英語母語話者の英語表現との違いを補償して認識率を向上させる言語モデル適応化処理の効果を検証するために，発話セット SSS-B および SSS-T を入力として，native 言語モデルと non-native 言語モデルを用いた場合のテストセットパープレキシティおよび WER の比較を行った．なお，SSS-B の発話は native および non-native 両言語モデルの学習に用いた BTEC コーパスと同一ドメインの発話であり，SSS-T の発話は異なるドメインでの発話と考えられる．あわせて，他のドメインでの英語発話に対する効果を検証するために，新たに学習者コーパス中の旅行会話以外の会話文約 53,000 文および教科書文約 65,000 文の英訳文を加えて言語モデルを作成し，これを native 言語モデルと線形補間した言語モデルを作成した．作成した言語モデルは日本語母語話者の教科書の英訳文を使用しており，自由発話セット SSS-T に対してはドメイン適応がなされていると考えられるため，この言語モデルを cross-domain 言語モデルと略称する．これら 3 種類の言語モデルによるテストセットパープレキシティと認識率の比較を**表 5** に示す．

表 5 に示すように，同一ドメインの発話である SSS-B が入力された場合，non-native 言語モデルによるテストセットパープレキシティは native 言語モデルを用いた場合と比較して減少しており，WER は 14%以上削減されている．non-native 言語モデルは日本語母語話者の英語表現と英語母語話者の英語表現との違いを補償して認識性能を向上させており，言語モデル適応化処理は日本語母語話者による同一ドメインでの英語音声の認識性能を向上させるのに有効であることが確認された．なお，native 言語モデルの学習と比べた場合，non-native 言語モデルの学習には学習者コーパスも利用されているため，学習に使用したデータ量が増加している．一方，non-native 言語モデルを用いた際の読み上げ発話セット RSS のパープレキシティは，bi-gram の場合で 43.9 から 44.9 にわずかながら増加している．このことから，non-native 言語モデルによるパープレキシティの改善はデータ量の増加ではなく，学習者コーパスの利用により日本語母語話者の英語表現を適切にモデル化していることによると考えられる．

異なるドメインの発話である SSS-T が入力された場合も，non-native 言語モデルは認識性能を向上させている．日本語母語話者が英語発話を行う際の母語の干渉には様々な現象があるが，特に数多い現象としては冠詞の欠落などが指摘されており[10]，これらの現象は多くの被験者に共通に見られると想定できる．このため，これらの現象は種々のドメインにわたって共通に現れ，この現象をモデル化している non-native 言語モデルは他のドメインでの日本語母語話者の英語発話についても認識性能を向上させることができると想定できる．

本提案手法は，日本語母語話者による他のドメインでの英語音声についても認識性能を向上させるが，ドメイン適応がなされている cross-domain 言語モデルによる性能と比較すると認識率は低い．言語モデルのドメイン依存性は音響モデルより大きいことが指摘されており[20]，日本語母語話者の英語認識用に適応化された言語モデルについても同様の結果が得られていると考えられる．一方，SSS-B が入力された際の non-native 言語モデルと cross-domain 言語モデルでの性能比較結果から見られるように，ドメイン外の発話を加えた場合には性能が劣化している．このため様々なドメインについての発話に関する認識性能の向上を図るために，個々のドメインごとに英語母語話者の発話を収集したコーパスに対応した学習者コーパスの構築が必要となると考えられる．

4.3　話者の英語能力と認識性能

被験者 55 名の TOEIC スコアが 325 点から 930 点に広く分布することが示すように，各被験者の英語能力は大きく異なる．被験者の英語能力が低いほど一般にその英語表現は英語母語話者による表現から乖離する度合いは大きいと推測される．このため non-native 言語モデルによる認識性能の向上度合いは被験者の英語能力により異なると予想される．TOEIC で示される被験者の英語能力と，non-native 言語モデルによる native 言語モデルと比較した場合の性能の向上度合いとの関係を，**図 1** では tri-gram の場合の両言語モデルでのパープレキシティの差で，**図 2** では WER の差で示す．なお，図中の直線は回帰直線を示して

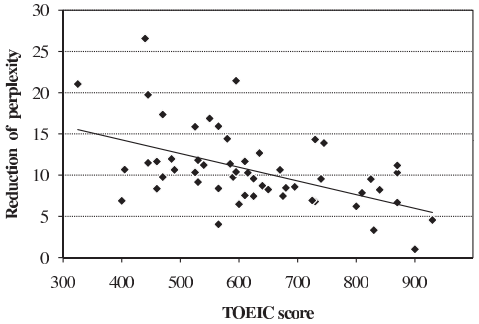

図 1　non-native 言語モデル（tri-gram）の使用によるパープレキシティの減少と被験者の TOEIC スコアとの関係
Fig. 1　Relation between TOEIC score of each subject and reduction of perplexity with the tri-gram of the non-native language model.

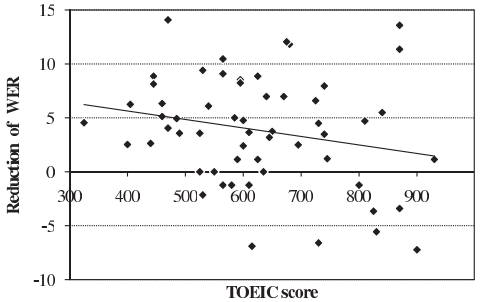

図 2　non-native 言語モデルの使用による WER の削減と被験者の TOEIC スコアとの関係
Fig. 2　Relation between TOEIC score of each subject and reduction of WER with the non-native language model.

おり，相関係数はパープレキシティの場合で −0.51，WER の場合で −0.31 となっている．TOEIC スコアと両言語モデルによるパープレキシティの差について相関がないとの仮説は，t 検定により有意水準 1%で棄却される．WER は言語モデルの改善効果以外に音響モデルと話者の発声との整合性など，他の多くの要因が関連する．このため，TOEIC スコアと両言語モデルによる WER の差についての相関は小さくなり，TOEIC スコアと両言語モデルによる WER の差に関して相関がないとの仮説は，有意水準 5%では棄却されるが，パープレキシティの場合と同じ有意水準 1%では棄却されない．これらのことから，(1) TOEIC スコアの低い被験者ほど線形補間された言語モデルによる改善効果は若干大きい，(2) 各被験者によって改善効果はばらついており，特に WER の削減値のばらつきは大きく，被験者によっては逆に WER が増加する場合もある，という傾向が見られる．

表 6　言語モデルによる認識性能の向上効果の被験者の英語能力による差異
Table 6　Comparison between reductions of WER for all subjects and subjects of high proficiency.

被験者	native LM	non-native LM
全被験者	30.6%	26.2%
TOEIC スコア 730 点以上	30.2%	26.8%

英語能力が高く英語でのコミュニケーションを行う機会が多いと考えられる日本語母語話者の英語発話に対する non-native 言語モデルの有効性を検証するため，「どんな状況でも適切なコミュニケーションができる素地を備えている」と評価[21] されている TOEIC スコア 730 点以上の被験者による英語発話の音声認識性能評価を行った．TOEIC スコア 730 点以上の被験者による英語発話に対する non-native 言語モデルと native 言語モデルの認識性能の比較を**表 6** に示す．

表 6 に示されるように，non-native 言語モデルを用いた場合の WER は，TOEIC スコア 730 点以上の被験者の発話に限った場合，全被験者の発話に対する場合と比較して若干低下するものの native 言語モデルと比較すると 11%程度削減されており，英語表現能力が高い話者に対しても効果があることが示されている．しかし，各被験者の改善効果はかなりばらついており，その原因の追及とともに話者適応などの導入の検討も必要と考えられる．

5. コーパス作成方法

以上述べたように，日本語母語話者による学習者コーパスから求めた言語モデルを同一ドメインの大規模英文テキストから学習した native 言語モデルに線形補間処理した non-native 言語モデルは，日本語母語話者による同一ドメインでの英語発話の認識性能を向上させる効果を有する．学習者コーパスの開発・整備は人手と費用のかかる作業であるため，開発に際しては被験者および課題文の選択方法やその規模など，音声認識性能に影響を与えると考えられる様々な要因をあらかじめ検討しておく必要がある．ここでは，言語モデルの性能に影響を与える複数の要因について検討を行う．

5.1 学習者コーパス中の重複文の影響

本実験に使用した学習者コーパスの開発に際しては，1 課題文あたり 100 名の被験者が翻訳作業を行っているため，学習者コーパス中には同一の課題文についてまったく同じ英語表

表 7　重複文を除いて学習した言語モデルの特性比較
Table 7　Comparisons among the native LM, the non-native LM, and a non-native LM which was trained excluding the same translation by other subjects.

	SSS-B			SSS-T		
	bi-gram	tri-gram	WER	bi-gram	tri-gram	WER
native LM	49.5	33.1	30.6%	89.0	73.2	39.9%
non-native LM	33.6	21.0	26.2%	79.4	55.2	36.7%
non-native LM (no_dup)	35.8	22.2	26.5%	79.2	54.8	36.6%

記である重複文が数多く存在する．対象とする話題に関する確率的な言語モデルの開発に際して用いられる対象分野でのテキストや発話を収集したコーパス中には同一表記で表される文や発話も出現する．コーパス中のこれらの同一表現の出現率は，その表現の言語運用上の出現頻度に依存して決まると考えられる．一方，学習者コーパス中の重複文の出現率は，必ずしも表現そのものの言語運用上の出現頻度を表しておらず，課題文の英訳の容易性など他の要因も関係している可能性がある．これは対象分野のコーパス中の一部の表現を取り出し，それを用いて学習者コーパスを開発するという手法自身に内在する課題であり，この課題の解決には対象コーパス中から英訳課題文を選択する基準，英訳文を作成する被験者の選択基準，被験者数など多くの検討課題がある．ここでは，言語運用上の出現頻度の影響も考慮されないことになるが，被験者を 1 名と見なして各課題文に対応する英語表現の重複を認めない場合の効果を検証するために，重複文を取り除いて言語モデルを学習した場合の認識実験を行った．重複文を取り除いた場合，旅行会話文の総英訳文数は約 28,000 文から約 19,000 文となり，データ量は約 32%減少する．

non-native 言語モデルを構築する際に学習者コーパスから重複文を取り除いて学習したモデル non-native LM（no_duplication）を用いた場合のパープレキシティおよび WER を**表 7** に示す．同一ドメインの発話 SSS-B と異なるドメインでの発話 SSS-T によって若干傾向は異なるが，同一ドメインの発話 SSS-B の場合は，重複文を除かない場合と比較してパープレキシティ，WER は増加するが，native 言語モデルを用いる場合に比べて特性は改善されている．すなわち，異なる被験者による重複した英訳文を取り除いて言語モデルを学習した場合，同一ドメインの発話に関して若干性能は劣化するが，その劣化割合は小さく，本実験で使用した学習者コーパスのように被験者数が 100 名との多人数の被験者は不要であり，1 名の被験者による英訳文を使用しても効果があるように見える．しかし，重複文以外の文については複数の被験者の英訳文が使用されているため，各課題文につき何名程度の被験者が必要であるかを改めて検討する必要がある．

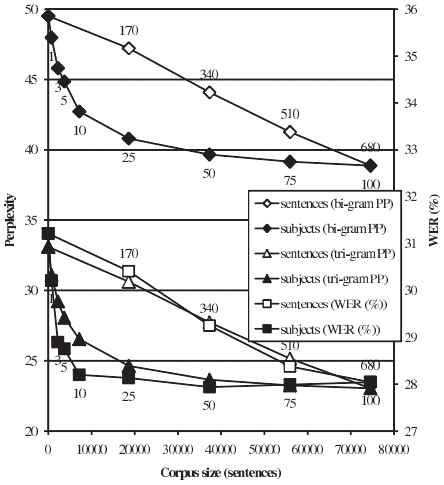

図 3　コーパスサイズとパープレキシティおよび WER の関係
Fig. 3　Relation of WER and perplexity to subjects and target sentences.

5.2　被験者数と課題文数の影響

言語モデルの高性能化の直接的な手段として，コーパスの規模の拡大があるが，学習者コーパスの場合には英訳課題文数の増加と被験者数の増加の 2 通りが考えられる．5.1 節に示したように，被験者数を増加させる場合，各英訳課題文に対し異なる表現の英訳文が得られるが，認識性能改善に効果が薄いと考えられる重複文も増加する．そのため，ここでは被験者数の増加による改善効果を，課題文数の増加による改善効果との比較により，必要と想定される被験者数の検証を行った．検証方法としては，本実験に使用した学習者コーパスの規模を，(1) 被験者数を各セット 100 名に固定し英訳課題文数を変化させた場合と，(2) 各セットでの英訳課題文数を 170 文に固定し被験者数を変化させた場合での改善効果の比較を行った．(1) および (2) の条件の下で総英訳文数を変化させた場合のパープレキシティおよび WER の変化を**図 3** に示す．native 言語モデルとの線形補間処理では，先に述べたように学習者コーパスは 5 セットから構成されているため，言語モデルの学習には 5 セットのうち 4 セットを用い，その中で英訳課題文数および被験者数を変化させ，重み係数 λ は言語モデル学習時に使用していない残り 1 セットを用いて決定した．

図 3 の横軸は言語モデル学習時に使用された総英訳文数を示し，縦軸はパープレキシティ

および WER を表している．なお，図中，白いプロットは条件 (1) すなわち英訳課題文数を変化させた場合を示し，図中の数値は英訳課題文数を示している．黒いプロットは条件 (2) すなわち被験者数を変化させた場合を示し，図中の数字は被験者数を示している．図から明らかなように，総英訳文数を増加させた場合，条件 (1)，(2) ともに言語モデルの特性は改善されている．しかし，条件 (2) の場合すなわち，英訳課題文数を固定し被験者数を変化させた場合の方が，より早く言語モデルの特性改善がなされている．

このことから，日本語母語話者による英語発話認識のための言語モデルの開発を目的として学習者コーパスを開発する場合には，同数の英訳文数において，被験者数を多くするより課題文数を多く設定する方が言語モデルの特性改善に役立つと考えられる．なお，被験者数は約 10 名程度の被験者数で被験者数増加による特性改善はほぼ飽和すると考えられる．

6. 結　　び

英語発話のドキュメント処理技術の高性能化は今後よりいっそう重要となると予想される．しかし，第二言語話者による発音や表現は母語の干渉を受け，英語母語話者の発音や表現と異なる点が多々あるため，英語母語話者の音声データを用いて学習を行った英語音声認識装置では，第二言語話者による英語に対する音声認識性能は大きく低下する．このため日本語母語話者による英語音声の認識率の向上を図るには，日本語母語話者の英語発音の特徴を表現した音響モデルの開発に加えて，日本語母語話者による英語表現の言語特性を考慮した言語モデルを構築することが必要である．この課題の解決のため，英語母語話者のテキストコーパスから学習した言語モデルと，同一ドメインの課題文を日本語母語話者が英訳した英訳文を収集した学習者コーパスから構築した言語モデルとを線形補間処理した言語モデルを用いる手法を提案し，日本語母語話者による英語音声の認識性能を向上させることを確認した．提案した言語モデルは TOEIC の高得点者に対しても認識性能の向上を達成している．このことから，日本語母語話者の英語音声を認識し，ドキュメント処理を行うには，少数の学習者コーパスから学習を行った言語モデルと英語母語話者のコーパスから学習を行った言語モデルとで線形補間処理を行った言語モデルを使用することの有効性が確認された．また，日本語母語話者による学習者コーパスを用いて作成された言語モデルは，他のドメインでの発話に対しても一定の認識性能の向上をもたらすことが確認された．

本研究では，学習者コーパスは日本語母語話者による英訳文から構成されており，評価に用いた英語発話もその場で英語表現がなされた発話であるが，日本語による課題文を提示し，それを翻訳する形式で発話されたものである．このため，実際の英語による会話の状況で自由発話として発話されたものではなく，filled pause は含まれていないなど自由発話による英語表現ではない．今後，実際の状況での英語発話表現を収集し，提案手法の有効性を確認する必要がある．

今回のドメインでの発話については，学習者コーパスの構築の際の被験者数は 10 名程度で十分であり，それ以上の増加は性能向上に寄与しないとの結果が得られたが，被験者の選択方法については今後の研究課題である．また，言語モデル適応による改善効果は，同等の TOEIC スコアを有する被験者についても個々の被験者によってかなりばらつきが存在する．話者適応も含めて個々の話者の違いにどのように対処するかも今後の研究課題である．

謝辞　本研究に際して有意義なコメントをいただいた同志社大学工学研究科柳田益造教授，ATR 安田圭史研究員に感謝します．本研究は科学研究費補助金（基盤研究 B）（課題番号 16300048）による助成研究の一部である．

参　考　文　献

1) 中川聖一：音声ディクテーションから音声ドキュメント処理へ，日本音響学会秋季研究発表会 1-3-1，pp.1–4 (2007).
2) Garofolo, S., Auzanne, C. and Voorhees, E.: The TREC Spoken Document Retrieval Track: A Success Story, *Proc. 8th Text Retrieval Conference*, pp.107–129 (2000).
3) Chelba, C., Hazen, T. and Saraclar, M.: Retrieval and Browsing of Sopoken Content, *IEEE SIGNAL PREOCESSING MAGAZINE*, Vol.25, No.3, pp.39–49 (2008).
4) 山崎博紀，喜多村圭祐，山本誠一：日本語母語話者のための英語音声認識システム用英語言語モデルの検討，信学技報 TL2007-72，SP2007-167, WIT2007-72, pp.1–6 (2008-1).
5) Yamazaki, H., Kitamura, K., Harada, T. and Yamamoto, S.: Creation of Learner Corpus and its Application to Speech Recognition, *Proc. LREC2008* (2008).
6) 中川聖一：科学研究費特定研究（A）「メディア教育利用」，日本音響学会誌，Vol.56, No.11, pp.767–770 (2000).
7) 峯松信明，仁科喜久子，中川聖一：外国語学習用読み上げ音声データベース，日本音響学会誌，Vol.59, No.6, pp.345–350 (2003).
8) http://casls.uoregon.edu/sla.php
9) http://catalog.elra.info/product_info.php?products_id=568
10) 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).
11) 筒井良平，鈴木基之，伊藤彰則，牧野正三：誤り訂正を用いた日本人英語音声認識，日本音響学会春季研究発表会，2–10–22, pp.119–120 (2008).

12) Bellegarda, J.: Statistical language model adaptation: review and perspectives, *Speech Communication*, Vol.42, No.1, pp.93–108 (2004).
13) Takezawa, T., Sumita, E., Sugaya, F., Yamamoto, H. and Yamamoto, S.: Toward a broad-coverage bilingual corpus for speech translation of travel conversations in the real world, *Proc. LREC*, pp.147–152 (2002).
14) 喜多村圭佑，安田圭志，山本誠一，柳田益造：英語学習者コーパスの開発と英語表現能力評価尺度の検討，信学技報 ET2007-87，pp.19–24, 電子情報通信学会 (2008).
15) Nakamura, S., Markov, K., Nakaiwa, H., Kikui, G., Kawai, H., Jitsuhiro, T., Zhang, J., Yamamoto, H., Sumita, E. and Yamamoto, S.: The ATR Multilingual Speech-to-Speech Translation System, *IEEE Trans. ASLP*, Vol.14, No.2, pp.365–376 (2006).
16) http://www.ldc.upenn.edu/Catalog/docs/LDC93S1/PHONCODE.TXT/
17) Takami, J. and Sagayama, S.: A successive state splitting algorithm for efficient allophone modeling, *Proc. ICASSP*, Vol.1, pp.573–576 (1992).
18) Paul, D. and Baker, J.: The design for the wall street journal-based CSR corpus, *Proc. DARPA Speech and Natural Language Workshop*, pp.357–362 (1993).
19) 北　研二：確率的言語モデル，東京大学出版会 (1999).
20) Lefevre, F., Gauvain, J. and Lamet, L.: Improving genericity for task-independent speech recognition, *Proc. Eurospeech2001*, pp.1241–1244 (2001).
21) http://www.ets.org/toeic/

(平成 20 年 12 月 19 日受付)
(平成 21 年 7 月 2 日採録)

原田　貴史

昭和 60 年生．平成 20 年 3 月同志社大学工学部卒業．同年同志社大学大学院工学研究科入学．音声認識の研究に従事．電子情報通信学会学生員．日本音響学会学生員．

山本　誠一（正会員）

昭和 25 年生．昭和 47 年大阪大学工学部卒業．昭和 49 年大阪大学大学院基礎工学研究科修士課程修了．同年国際電信電話株式会社入社．ATR 音声言語コミュニケーション研究所所長を経て現在同志社大学理工学部教授．この間，適応信号処理，音声合成，音声認識，音声翻訳等の研究に従事．工学博士．日本音響学会第 3 回技術開発賞，第 5 回技術開発賞，電子情報通信学会情報システムソサイエティ論文賞，電気通信普及財団テレコムシステム技術賞等を受賞．日本音響学会，言語処理学会，人工知能学会各会員．IEEE Fellow，電子情報通信学会 Fellow.