情報処理学会論文誌 Vol.53 No.10 2349–2358 (Oct. 2012)

多言語音声の同時認識システムにおける翻訳モデルとスコア計算の高速化

大村 絵梨[1] 南條 浩輝[2,a)]

受付日 2012年1月26日, 採録日 2012年7月2日

概要：国際会議やニュースでは，複数の言語で同一内容の発話がなされていることが多い．我々は，これまでにこのような多言語音声の音声認識の枠組み，具体的には，ある言語の音声とそれに対応する他言語の音声を，翻訳モデル（TM）を用いてお互いに情報を補い，同時に認識する枠組みを提案している．本論文では，多言語音声の同時認識システムにおける TM モデル化手法および TM スコア計算の高速化について研究を行った．IBM モデル 1，モデル 2 およびモデル 3 のすべてについて，多言語音声の同時認識用の TM として有効であることを明らかにした．ドメインの一致するコーパスを大量に用いて学習することの重要性を確認した．TM スコア計算では，スコア近似手法を提案し，音声認識精度の性能低下を抑えつつ高速化が行えることを示した．

キーワード：音声認識，多言語音声処理，翻訳モデル，高速化

A Study of Translation Models and Score Calculation on Bilingual ASR Framework

Eri Ohmura[1] Hiroaki Nanjo[2,a)]

Received: January 26, 2012, Accepted: July 2, 2012

Abstract: This paper addresses automatic speech recognition (ASR) for multilingual audio contents. Conventionally, ASR has been performed independently, namely, language by language, although multilingual speech, which consists of utterances in several languages representing identical meaning, is available. We previously proposed a bilingual ASR framework based on statistical ASR and machine translation in which bilingual ASR is performed simultaneously and complementarily. In this simultaneous recognition framework, ASR systems use not only acoustic and language model scores but also a translation model (TM) score. In this study, we investigate a suitable TM modeling and an efficient calculation method of TM scores. We compared several TM models, which are trained with matched/unmatched domain corpus, and TM score calculation methods. We confirmed the effectiveness of IBM model-1, model-2 and model-3 based TMs and the significance of TM training with large amount of matched domain corpus. We significantly reduced processing time for TM score calculation without any degradation of ASR accuracy.

Keywords: automatic speech recognition, multilingual audio processing, translation model, fast calculation algorithm

1. はじめに

音声認識技術はニュース字幕や会議録を効率良く作成することを可能とする重要な技術であり，高い認識精度が求められている．我々は，ニュースをはじめとする放送コンテンツや国際会議などにおいて，同一内容の発話が複数の言語でなされている状況（たとえばニュース放送では，主

[1] 龍谷大学大学院理工学研究科
Graduate School of Science and Technology, Ryukoku University, Otsu, Shiga 520–2194, Japan
[2] 龍谷大学理工学部
Faculty of Science and Technology, Ryukoku University, Otsu, Shiga 520–2194, Japan
[a)] nanjo@rins.ryukoku.ac.jp

音声とその通訳の副音声が利用できる）が多く存在することに着目し，音声認識の高精度化を目的として，このような複数の入力音声を統計的機械翻訳のモデルを用いて同時に処理する枠組みを提案している [1]．たとえば，「友達がくるまで待っていた」という発話の「くるまで」は「車で」か「来るまで」かが曖昧であり，提案する音声認識手法は，このような曖昧性を他の言語の音声（たとえば英語）を聞くことで解消を目指すものである．また，一部の語句に対して，雑音や速い発声，小さい声などにより，音声認識が難しい場合でも，他の言語の音声情報からその部分の書き起こし精度の向上を試みるものである．同時通訳が存在するニュースの音声字幕作成支援のための音声認識手法や，国際会議の議事録作成のための音声認識手法として位置づけられる．

文献 [1] において，我々は枠組みが正しく機能することを示した．しかし，用いたモデルおよび同時処理デコーディングアルゴリズムについての検討がなされていないという問題があった．本論文では，これらの問題への対応を行う．具体的には，はじめに複数言語の発話の情報を互いに補うアルゴリズム，すなわちデコーディングアルゴリズムの高速化と認識精度についての調査を行い，適切な方法を明らかにする．このような調査はこれまでなされていない．次に，複数言語の発話の情報を互いに補って同時に音声認識するための統計的機械翻訳モデル（Translation Model; TM）について種々のモデルを比較し，有効なモデル化手法を明らかにする．統計的機械翻訳の研究 [2], [3] や翻訳支援の研究 [4] では種々の TM が調査されているが，多言語音声の同時認識における調査はこれまでに行われておらず，本研究は新しい．その際，TM の学習データについての調査もこれまで行われていないため，本論文では種々の学習データを用いて学習した TM を比較し，学習データに関する新たな知見を得る．本論文では，2 章で多言語音声の同時認識の概要について述べ，3 章で実験環境と実験データについて述べる．4 章では同時音声認識における他言語情報統合アルゴリズムについて述べる．5 章では翻訳モデル化手法について述べる．6 章では翻訳モデルの学習データの影響について述べる．7 章で結論を述べる．

2. 多言語音声の同時認識

本章では，我々が提案している同時音声認識の枠組み [1] について述べる．具体的には，複数言語間でなされている同一内容発話の対応関係が与えられた条件のもとで，TM を用いて 2 言語音声を同時に処理する枠組みについて述べる．概要を**図 1** に示す．この条件のもとでの複数音声の同時認識は，ある言語（たとえば日本語）の音声 X および他の言語（たとえば英語）の対応する音声 Y が与えられたときに，それらを最もよく説明する日本語文字列 $\hat{J}$ および英語文字列 $\hat{E}$ を求める問題として定式化できる．ここでは，

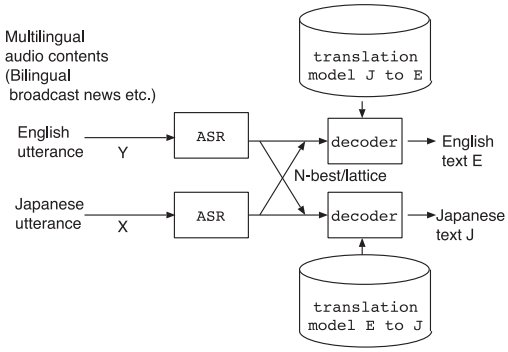

図 1 2 言語音声の同時認識枠組みの概観
Fig. 1 Overview of bilingual speech recognition framework.

X と Y，J と E のペアはそれぞれ対称であるため，日本語文字列 $\hat{J}$ を求める過程について述べる[*1]．これは式 (1) と表せる．

$$\hat{J} = \underset{J}{\operatorname{argmax}} P(J|X,Y) \tag{1}$$

式変形を行うと，以下の式が得られる [1]．

$$\hat{J} = \underset{J}{\operatorname{argmax}}\Big(\log P(X|J) + \alpha \log P(J) + \beta N_w + \gamma \log P(J|E)\Big) \tag{2}$$

ただし，E は英語音声認識結果の第 1 候補である．

先行研究 [1] では，式 (2) が用いられていた．この $P(J|E)$ は単語列の翻訳スコアであり，単語の翻訳確率スコアの積である．そのため，J の文字列（単語数）の増加にともなって $P(J|E)$ は低下し，長い仮説文候補は不利である．したがって，長さ調節のためのパラメータ δ を導入することを提案し，本研究で用いる．すなわち以下の式 (3) を用いてデコーディングを行う．

$$\hat{J} = \underset{J}{\operatorname{argmax}}\Big(\log P(X|J) + \alpha \log P(J) + \beta N_w + \gamma(\log P(J|E) + \delta N_w)\Big) \tag{3}$$

式 (3) より，多言語の音声認識には，1）翻訳スコアを算出する翻訳モデル（TM），2）音声認識結果とそのスコアを生成する音声認識システム，3）翻訳スコアと音声認識スコアの積（ログスケールでは和）を最大化する単語列 $\hat{J}$ を探索するデコーダ，の 3 つの構成要素が必要であることが分かる．本論文では，1）の TM と 3）のデコーダのアルゴリズムについて述べる．

3. 実験環境とデータ

実験環境と実験に用いたデータについて述べる．本論文では，多言語音声認識の枠組みにおける TM とデコーダの評価を行う．他言語の情報を用いる手法を純粋に比較するために，現在着目している言語（ここでは日本語）とは別

[*1] 英語文字列 $\hat{E}$ を求めたい場合は，X と Y，J と E をそれぞれ入れ替えればよい．

の言語（ここでは英語）の音声認識の認識誤りの影響を除いて [1] 評価を行う．具体的には，日本語の音声認識時に対訳英語テキストと日英機械翻訳モデルを用いて評価を行う．すなわち評価は次の手順で行う．

(1) 単一言語（日本語）の音響モデルと言語モデルのみを用いてゆう度最大化基準で音声認識を行って N-best 仮説（N = 300）を生成する．
(2) 対訳英語テキストと TM を用いて式 (3) に基づいて N-best のリスコアを行う．その過程において，種々の TM およびデコーディングアルゴリズムの比較を行う．

なお，本研究では，式 (3) における重みパラメータとして $\alpha = 6$，$\beta = 1$，$\gamma = 1$，$\delta = 5$ を用いる．この値に設定した経緯は次のとおりである．α は本実験で用いる音声認識エンジン（3.2 節参照）が持つデフォルト値とした．β は $\alpha = 6$ のときに本実験の評価データ（3.1 節参照）に対する音声認識精度が最大となる値とした．γ は特に考慮せず 1 とし，δ は評価データの一部を用いて TM（IBM モデル 3）を利用した音声認識を行い，高い認識精度が得られた値に設定した．

3.1 評価データ

評価データを**表 1** に示す．これは先行研究 [1] と同様の評価データである．日本語テキストとその対訳となる英語テキストは，『The NEWS HOUR リスニング』[5], [6] から選択された双方の単語数が 60 以下のもの 50 ペア（英語ニュース音声の書き起こしとその和訳；政治，経済，社会のドメイン）である [1]．日本語音声は，日本語母語話者 5 名が各 50 文を読み上げたデータ（計 250 発話）である．

3.2 ベースライン音声認識システム

本研究では，認識エンジン Julius rev.3.4.2 [7] と ATR 多数話者音声データベースから学習したモノフォン音響モデル，新聞記事データから学習した言語モデル（逆向き単語 3-gram モデル）を用いて音声認識システムを構成した．

本実験でモノフォンを使用する理由は，1）本実験タスクにおけるトライフォンモデルを使用した音声認識ではゆう度最大の第 1 仮説での認識率が約 95%と高く [1]，TM を用いる効果を十分に調査できないため，2）実際の会議やニュースの話し言葉音声の認識 [8], [9] では今回のタスク（読み上げ音声認識）に比べて一般的に精度が低く，本タスクでモノフォンモデルを用いた音声認識システムが実際の話し言葉の認識のタスクと近いと考えられ，望ましいと考えるためである．

表 1 評価データ
Table 1 Evaluation data.

テキスト	日本語，英語各 50 文
単語数	711（日本語），476（英語）
音声	5 名（日本語母語話者）のテキスト読み上げ（計 250 発話）

3.3 翻訳モデルと学習データ

3.3.1 翻訳モデルの学習データ

TM の学習データには，ATR-SLDB 会話表現および対話データベース [10], [11]（会話）とロイター日英記事対訳コーパス [12]（記事）と日英新聞記事対応付けデータ [12]（新聞）の 3 種類を用いた．一覧を**表 2** に示す．なお，学習データには 1 文 100 単語以下の文のみを採用した[*2]．また，学習データ内の出現頻度 2 以下の単語は未知語単語 UNK として学習を行い，テストデータ中の未知語にはこの UNK の確率を与えた．

3.3.2 翻訳モデル

本研究では，TM として IBM モデル 3 [13] を採用した．本タスクにおいて TM に望まれるのは，対訳の言語に対応する単語がある場合に高い TM スコアを与えることである．たとえば「ゆうじんがくるまでまっている」という日本語発話に対応する英語発話中に “car” が存在するときには仮説「友人が車で待っている」に対して高いスコアを，“come” が存在するときには仮説「友人が来るまで待っている」に対して高いスコアを与えることである．さらに，対訳に対応する単語が存在しない場合，たとえば「友人が」の「が」に対応する単語がない場合，および，対訳中の単語が複数の単語に対応する場合，たとえば，“wait” が「待っ て いる」に対応する場合にも適切に高いスコアを与えることが求められる．語順の入れ替えについては，機械翻訳において種々のモデル化が提案されており，複雑なモデルも存在する．日英のような言語ペアの機械翻訳では特に重要であるものの，多言語音声の同時認識には，語順入れ替えについては簡易なモデルで十分とも考えられる．このような理由から，これらを満たす TM として IBM モデル 3 を採用した．IBM モデル 3 の詳細については次段落で述べる．なお，本論文では，より簡易なモデルでも十分かを検討するために IBM モデル 1 と IBM モデル 2 を調査する．IBM モデル 1，2 および 3 の違いについては 5 章で詳しく述べる．

次に，IBM モデル 3 について述べる．翻訳モデルは，原言

表 2 TM 学習データ
Table 2 Training data for TMs.

TM 学習データ	文数	単語数		語彙サイズ	
		日	英	日	英
(1) 会話	63k	567k	484k	37k	34k
(2) 記事	56k	1.6M	1.3M		
(3) 新聞	179k	5.3M	4.9M		

[*2] 本実験で用いた学習ツールはデフォルトでは 101 単語以上の文を扱えないため．

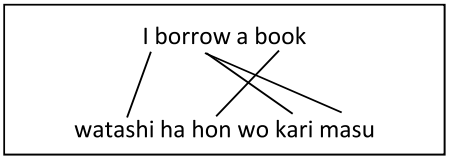

図 2 アライメント $A = (1\ 0\ 4\ 0\ 2\ 2)$ の例
Fig. 2 Example of alignment: $A = (1\ 0\ 4\ 0\ 2\ 2)$.

語の単語列 E と目的言語の単語列 J の対応スコア $P(J|E)$ を与えるモデルである．IBM モデルでは，E と J のアライメント（単語対応）A を考え，$P(J|E)$ をすべてのアライメントに関する条件付き確率 $P(J, A|E)$ の和として式 (4) で表す．

$$P(J|E) = \sum_{A} P(J, A|E) \tag{4}$$

原言語単語と目的語単語の 1 対 1 または 1 対多の対応しか考えないこととすると，アライメントは目的言語の単語数 v を要素数とするベクトル A として表すことができる．A の i 番目の要素 A_i の値は，目的言語の i 番目の単語が対応している原言語の単語列中の単語番号である．すなわち目的言語の i 番目の単語が原言語の j 番目に対応していたとき，$A_i = j$ となる．なお，原言語に対応する単語がない場合はダミー単語 NULL に対応しているとし，$A_i = 0$ とする．アライメント $A = (1\ 0\ 4\ 0\ 2\ 2)$ の例を**図 2** に示す．

IBM モデル 3 では，式 (4) を計算するために以下のモデルを用いる．

Fertility model

Fertility model は，原言語の単語 E_i が ϕ_i 個の目的言語の単語に対応する繁殖率 $n(\phi_i|E_i)$ を与えるモデルである．Fertility model スコア F_s は式 (5) で表される．

$$F_s = \prod_{i} n(\phi_i|E_i) \tag{5}$$

NULL generation model

目的言語に対応づけるべき原言語の単語が存在しない場合，原言語に NULL を挿入し，NULL と目的言語の単語が対応づけられていると考える．NULL Generation model は，ある単語の後に NULL が挿入される確率 P_{NULL} を与えるモデルである．NULL Generation model スコア N_s は式 (6) で表される．m は原言語の単語数である．なお，本研究では $P_{NULL} = 0.02$ とした．

$$N_s = {P_{NULL}}^{\phi_0} \cdot (1 - P_{NULL})^{m-\phi_0} \tag{6}$$

Lexicon model

Lexicon model は，原言語の単語 E_{A_i} が目的言語の単語 J_i に翻訳される翻訳確率 $t(J_i|E_{A_i})$ を与えるモデルである．Lexicon model スコア L_s は式 (7) で表される．

$$L_s = \prod_{i} t(J_i|E_{A_i}) \tag{7}$$

Distortion model

Distortion model は，単語数 u 個の原言語の j 番目の単語が，単語数 v 個の目的言語の i 番目に移動する確率 $d\left(\frac{i}{j, u, v}\right)$ を与えるモデルである．ただし，「NULL」は原言語で 0 番目にあるとする．あるアライメント A が与えられたときの Distortion model スコア D_s は，定義より $j = A_i$ であるので，式 (8) で表される．

$$D_s = \prod_{i} d\left(\frac{i}{A_i, u, v}\right) \tag{8}$$

IBM モデル 3 では，アライメント A が与えられたときの対応度 $P(J, A|E)$ は次のように計算される．

$$P(J, A|E) = F_s \cdot N_s \cdot L_s \cdot D_s \tag{9}$$

本研究では，TM の学習に GIZA++ [3] を用いる．

4. 同時音声認識における他言語情報統合アルゴリズム

複数言語の同時音声認識においてある言語の音声認識時に他言語の情報を統合するアルゴリズム，すなわちデコーディングアルゴリズムについて調査を行う．このようなアルゴリズムについて比較・調査された例はみられず新規性を有する．具体的には，対象とする言語の認識候補（テキスト）と他言語の認識候補（テキスト）間の TM スコアの算出時の近似計算とその効果について述べる．

4.1 翻訳モデルスコアの近似計算手法

原言語の単語列 E，目的単語の単語列を J としたとき，IBM モデルに基づく翻訳スコア計算では，式 (4) に示すとおり，すべての可能なアライメント A に対して $P(J, A|E)$ を求めてその和をとる必要がある．原言語単語と目的語単語が 1 対 1 または 1 対多の対応しか考えない場合は，原言語が u 単語，目的言語が v 単語からなる場合には $(u+1)^v$ 通りのアライメントが存在する．このことは単語列の長さによって爆発的に計算量が増加することを示している．本研究では，TM スコアを効率的に計算する方法を提案する．

4.1.1 統計的機械翻訳における翻訳モデルスコア計算

統計的機械翻訳は，単語列 J が与えられたときに，事後確率 $P(E|J)$ を最大とする単語列 $\hat{E}$ を見つける問題として，式 (10) で定式化される．

$$\hat{E} = \underset{E}{\operatorname{argmax}}\, P(E|J) \tag{10}$$

$P(E|J)$ はベイズ則を用いて $P(J|E)P(E)/P(J)$ と書ける．分母 $P(J)$ は式 (10) の右辺の最大化に影響を与えないため省略することができ，式 (10) は式 (11) に変形できる．

$$\hat{E} = \underset{E}{\operatorname{argmax}}\, P(J|E)P(E) \tag{11}$$

IBM モデルを用いた場合は，式 (4) を代入して，以下の

	(NULL)	I	borrow	a	book
私	0	$7.25 * 10^{-3}$	0	0	$1.25 * 10^{-4}$
は	$2.47 * 10^{-1}$	$1.24 * 10^{-2}$	0	0	$1.07 * 10^{-3}$
本	0	0	$1.3 * 10^{-5}$	0	$1.19 * 10^{-2}$
を	$3.07 * 10^{-6}$	0	0	$2.02 * 10^{-4}$	0
借り	0	0	$1.40 * 10^{-3}$	0	0
ます	0	0	$1.07 * 10^{-3}$	0	$3.14 * 10^{-2}$

図 3 行列 M の例

Fig. 3 Example of matrix M.

注：行列中の要素の値は実際の値ではなく，説明が分かりやすくなるようなものとしている．

式を解くことを目指す．

$$\hat{E} = \operatorname*{argmax}_E P(E) \sum_A P(J, A|E) \tag{12}$$

実際には，さらに和を最大値で置き換えた式 (13) を解く問題として，実現される [14], [15], [16], [17].

$$\hat{E} = \operatorname*{argmax}_E P(E) \cdot \max_A P(J, A|E) \tag{13}$$

右辺第 2 項の $P(J, A|E)$ を求めることに関しては，本研究と同じである．IBM モデル 1 と 2 以外では $\max_A P(J, A|E)$ を 1 回の手続きで求められないため，機械翻訳では初期アライメントを作り，そこから一部の単語対応を入れ替えながら最良の A と E の組合せの探索をすすめる方法 [15] や，初期アライメントから開始し，最適な E と A を交互に繰り返し探索をすすめる方法 [16], [17] などが採用されている．初期アライメントは，たとえば，E の各単語はその単語の翻訳確率が最大の J 中の単語に対応していると見なして作成する．なお，機械翻訳では $P(J, A|E)$ の最大値を求めることが主な目的であり，$\sum_A P(J, A|E)$ を求めることやその近似を目的としていない．

4.1.2 提案近似手法

次に，本研究で提案する IBM モデル 3 のスコア $\sum_A P(J, A|E)$ の近似計算アルゴリズムについて述べる．

(1) 目的言語文 E，原言語文 J の単語数をそれぞれ行と列の数とし，$v \times (u + 1)$ 行列 M を作る．ただし，一番左の列は 0 列目とする．

(2) 行列 M の各要素 M_{ij} の値に Lexicon model スコアと Distortion model スコアの積，$t(J_i|E_j) \times d\left(\frac{i}{j, u, v}\right)$ を代入する．

(3) 各行につきゼロでない要素を 1 つ選択する．定義より，これはスコアが 0 にならないアライメント A を選択することに相当する．

(4) (3) で得られた要素の積を求め，A に対応する Fertility model スコアおよび NULL generation model スコアとの積をとり，$P(J, A|E)$ とする．

(5) (3) に戻り，別のゼロでない要素の組（アライメント）を選択する．候補がなければ，次へ進む．

(6) (4) で求まったすべての値 $P(J, A|E)$ の和をとり，$P(J|E)$ とする

原言語文 “I borrow a book” と目的言語文「私は本を借ります」を例にとり説明する．これらから行列 M を生成すると**図 3** のようになる．図 3 のように行列 M の要素の多くは値が 0 となるため，可能なアライメント $(u + 1)^v$ すべてを計算する必要はなく，計算量を大幅に削減できる．

この行列だけに関して考えればよいのであれば，すべての可能な組合せの確率の和は，各行について和をとってそれらの積を求めればよい．しかし，ステップ (4) で示すとおり，Fertility model スコアと NULL Generation model スコアは A を決めたのちに求まるため，これらのスコアを事前に行列 M に挿入できず，この計算アルゴリズムは適用できない．

このような背景に基づき，本研究では行列 M の計算時に相対閾値 $THRES_{rel}$ を導入して近似計算を行うことで，計算時間の短縮を図る．具体的には，先ほどのアルゴリズムの (2) の直後に，以下の手順を追加する．

(2.1) 行 i での最大値 $\max_j M_{ij}$ を求める．

(2.2) 行 i において，各列要素 M_{ij}（$j = 0, \cdots$）に対して以下を実行する．

$$M_{ij} = 0 \quad \text{if} \quad \log M_{ij} - \log \max_j M_{ij} < THRES_{rel} \tag{14}$$

たとえば，$THRES_{rel} = -2$ と設定すると，図 3 の例では 2 行目については最大値は $M_{20} = 2.47 * 10^{-1}$ であるため，$M_{24} = 0$ と見なされて計算から除かれる．同様に 3 行目についても $M_{32} = 0$ と見なされる．このように，計算量の削減が得られる．

さらに，最も単純な $\sum P(J, A|E)$ の近似は総和を最大スコアに置き換えることである．実際には，F_s と N_s の計算も必要であるが，本研究では，L_s と D_s からなる行列 M の各行の最大値をそれぞれ選択することで最適なアライメント A を選択したと見なし，導出する．なお，これは上記手順 (2.2) において，$THRES_{rel} = 0$ とすることで実現できる．

本手法は，初期アライメントを作成し，そこから一部の単語対応を入れ替えながら新たなアライメントを評価して

いる点では，統計的機械翻訳でのスコア計算と同じである．ただし，統計的機械翻訳では，最適なアライメントとそのときのスコアを求めるのが目的であるため，スコアが下がる方向への単語の入れ替え（アライメントの作成）とそのようなアライメントでのスコア計算を避けようとする．本研究では，$\sum P(J, A|E)$ の近似，すなわち，初期アライメントから確率計算を開始し，累積スコアが $\sum P(J, A|E)$ の近似となることを目的としている．そのため，提案手法は，現在計算されている中で最適なアライメントのときのスコアよりも小さいスコアであっても，累積スコアへの寄与が高そうなものであれば計算を行うという手法となっている．

4.2 翻訳モデルスコアの近似計算の効果

TM スコアの近似計算の効果を調査した．具体的には，対訳英語情報を利用した日本語音声認識において，種々の $THRES_{rel}$ を用いて，計算時間と日本語音声認識の認識率の関係を調べた．ここでは，データ量が少なくドメインの一致するデータ（表 2 の (2) 記事）を用いて学習した TM（語彙サイズ：日本語 13k，英語 15k）と，大きいサイズのデータ（表 2 の (1)，(2)，(3) すべて）を用いて学習した TM（語彙サイズ：日本語 37k，英語 34k）の 2 種類を用いた．

結果を**図 4** と**図 5** に示す．図 4 の横軸は $THRES_{rel}$，左の縦軸は 1 対訳あたりの TM スコア計算時間（秒），右の縦軸は目的言語 1 単語あたりの対数 TM スコアを表している．学習データおよび語彙サイズが大きい TM のほうが TM スコアが高いこと，および同じ $THRES_{rel}$ であってもより計算時間を必要とすることが分かる．また，TM のサイズにかかわらず，近似計算を行うことでスコア計算時間を短縮できるものの，TM スコアが低下することが分かる．特に $THRES_{rel}$ が -0.5 よりも大きいところでは，計算時間はほとんど変わらないものの，TM スコアの低下度合いが大きいことが分かる．

図 5 の横軸は $THRES_{rel}$，左の縦軸は 1 対訳あたりの TM スコア計算時間（秒），右の縦軸は音声認識結果の単語誤り率（WER; word error rate）を表している．英語情報を用いない場合の日本語音声認識（ベースライン）の結果（WER）は 12.58%であった．TM スコア自体はスコア計算近似パラメータ $THRES_{rel}$ に影響を受けるものの，英語情報（TM）を用いた音声認識ではその影響は大きくなく，WER はそれぞれ，およそ 11.8%から 12.0%の間，およそ 10.4%から 10.8%の間に収まっており，TM を用いる効果がみられる．特に，大きな $THRES_{rel}$ を用いることで認識精度を保ったまま計算時間の大幅な短縮が可能であることが分かる．提案した TM スコアの近似計算方法により，認識精度を大きく低下させることなく計算コストを削減できることが分かる．$THRES_{rel}$ を -0.5 以上に設定した場合には，TM の語彙サイズによる計算時間の増大の影響も小さく抑えられることが分かる．なお，TM スコア計算時に近似を行わずに厳密に求めようとする場合（$THRES_{rel}$ を限りなく小さく設定する場合）は，原言語を u 単語，目的言語を v 単語として $O((u+1)^v)$ の計算コストを必要とし，文が長い場合に現実的な時間で計算できない．実際に本タスクでは数週間かけても計算が終わらなかった．

次に，文の長さごとの TM スコア計算時間と WER の影響を調べた．具体的には，目的言語の 1 文あたりに含まれる単語数によって 4 つのグループに分類した（**表 3**）．ここでは，表 2 の (1)，(2)，(3) すべてを用いて学習した TM（語彙サイズ：日本語 37k，英語 34k）を用いた．結果を**図 6** と**図 7** に示す．図 6 では，横軸が $THRES_{rel}$，縦軸が TM スコア計算時間（秒）である．図 7 では，横軸が

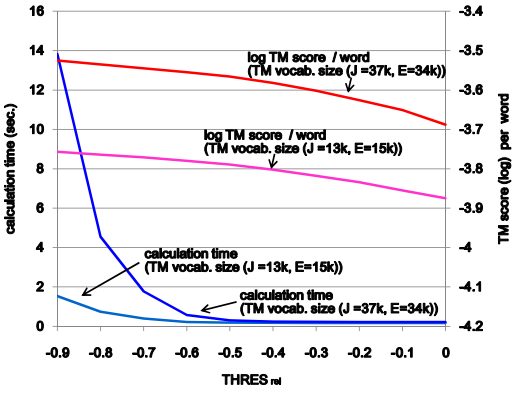

図 4 TM スコア計算時間と 1 単語あたりの対数 TM スコア
Fig. 4 Times for TM score calculation and averaged TM scores.

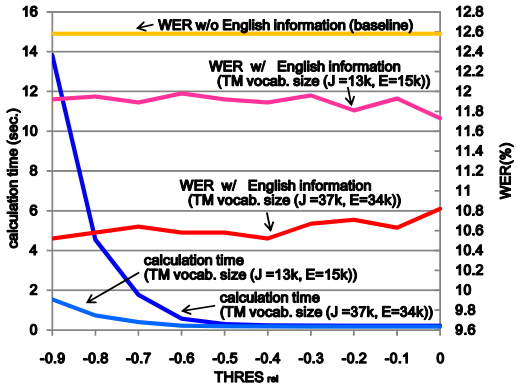

図 5 TM スコア計算時間と音声認識精度
Fig. 5 Times for TM score calculation and speech recognition performances.

表 3 評価データの分布（目的言語文の単語数ごと）
Table 3 Evaluation data distribution (Number of words in target language sentence).

1 文あたりの単語数	文数
1–10	75
11–15	85
16–20	55
21–30	35

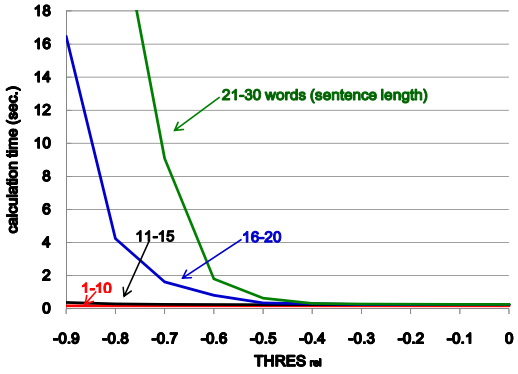

図 6 文の長さごとの TM スコア計算近似と計算時間
Fig. 6 Times for TM score calculation for each length group.

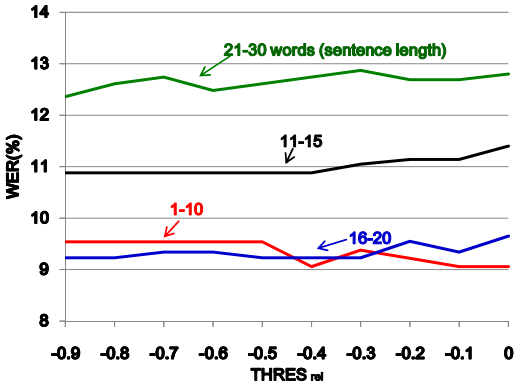

図 7 文の長さごとの TM スコア近似と音声認識結果
Fig. 7 Approximations of TM score and speech recognition performances for each length group.

$THRES_{rel}$，縦軸が音声認識精度（WER）である．グラフ上の数字は，目的言語の 1 文に含まれる単語の数を表している．図 6 からは，$THRES_{rel}$ を -0.5 以上に設定した場合には，TM スコア計算時間について，文の長さ（目的言語の単語数）による影響がほとんどないことが分かる．

$THRES_{rel}$ を -0.5 未満に設定した場合には，小さい $THRES_{rel}$ を用いるのにともなって，長い文においてより多くの計算時間が必要であることが分かる．一方，図 7 からは，大きい $THRES_{rel}$ を用いて強い近似計算を行ったとしても WER の大きな増大はみられないことが分かる．これらのことは，提案する TM スコアの近似計算方法は文の長さ（単語数）に影響を受けず，特に，長い文に対して有効であることを示している．$THRES_{rel}$ を -0.5 から 0 の間に設定して近似計算を行うことで，音声認識の精度を低下させることなく，TM スコア計算時間の高速化が可能であることを示した．

すべての学習データで学習した TM を用いたときに，用いないときに比べて音声認識の精度が向上した例を**図 8** に示す．“How”，“change” によって「こう書いている」を「どう変えていく」に修正できていることや，“room” によって同音異義語「予知」と「余地」から適切な「余地」を選択できていることが分かる．**図 9** には精度が低下した

発声：
規制措置はそのような状況をどう変えていくのでしょうか
（対訳 “How would this change that”）

英語情報なし：
規制措置はそのような状況こう書いているのでしょうか
英語情報あり：
規制措置はそのような状況**どう変えていく**のでしょうか

発声：
それではさらに連邦議会が介入しもっと強い措置を作る余地はあるのでしょうか
（対訳 “And is there room for Congress to still step in and do more itself”）

英語情報なし：
それではさらに連邦議会が介入しもっと強い措置を作る予知は可能でしょうか
英語情報あり：
それではさらに連邦議会が介入しもっと強い措置を作る**余地はあったの**でしょうか

図 8 英語情報を用いた日本語音声認識による精度向上例
Fig. 8 Successful examples of Japanese speech recognition with English information.

発声：
しかし今年はすでに農作物に使われてしまいました
（対訳 “But it’s already been used on food this year”）

英語情報なし：
しかし今年はすでに農作物に使われてしまいました
英語情報あり：
しかし今年はすでに農作物に使われてしまい**ますが**

発声：
そののちに私たちは救済措置に達することでしょう
（対訳 “After that we will get to remedy”）

英語情報なし：
そののちに私たちは救済措置に達することでしょう
英語情報あり：
そののちに私たちは救済措置に**とする**ことでしょう

図 9 英語情報を用いた日本語音声認識による精度低下例
Fig. 9 Unsuccessful examples of Japanese speech recognition with English information.

例を示す．上の例では，“But” と「ますが」という逆説表現の共起が学習されたことによる悪影響と考えられる．下の例では，多義的な語 “get” が英語表現にあり，「～へたどりつく」という意味の「達する」よりも「～する」が選択されたと考えられる．このような多義語については，学習データ量が不足している可能性も考えられる．

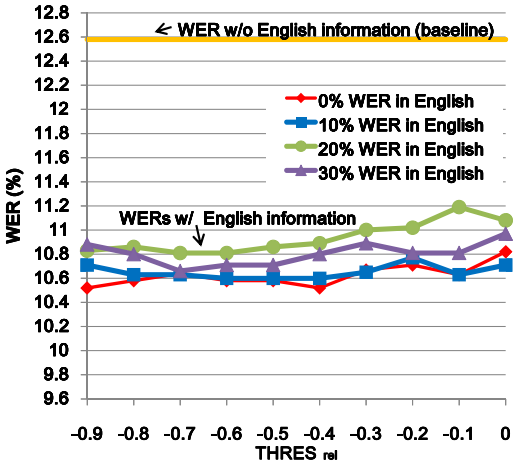

図 10 対訳英語情報に誤りを含むときの TM スコア近似と認識結果
Fig. 10 Approximations of TM score calculation and speech recognition with imperfect information in other language (English).

4.3 翻訳モデルスコアの近似計算における対訳情報の誤りの影響

次に，対訳英語情報を用いた日本語音声認識において，対訳情報に誤りが含まれる場合の，TM スコアの近似計算の有効性を調査した．具体的には，対訳の英語テキスト中の一部の単語を削除または別単語に置換し，本来存在しない単語を挿入することで，誤りを含む対訳英語テキスト（擬似的な音声認識結果に相当する英語テキスト）を生成し，それを用いて日本語音声認識を行った．対訳英語の誤りの割合（WER）はおよそ 10%，20%，30%となるよう設定した．

結果を**図 10** に示す．ここでは，すべての学習データから学習した TM を用いている．誤りを含む対訳英語情報を用いても，対訳英語情報を用いない場合（ベースライン）に比べて，音声認識精度を改善できることが分かる．対訳英語の誤り（割合：0%，10%，20%，30%）は，日本語音声認識時に対訳情報として用いる限りにおいて，影響が大きくないことが分かる．さらに，対訳情報に誤りが含まれている場合のスコア近似計算による日本語音声認識への影響は大きくなく，対訳情報に誤りが含まれていない場合の影響と同程度であることが分かった．これらのことは，提案手法の有効性を示している．

5. 翻訳モデル化手法

次に，TM のモデル化手法の比較を行った．多言語音声の同時認識における TM として，我々はこれまでに IBM モデル 3 を用いてきた [1]．これは 3.3.2 項でも述べたとおり，多言語音声の同時認識において TM にも望まれるのは，語順をあまり気にせずに対応する単語がある場合に高い TM スコアを与えること，および 1 対多の単語対応と対応単語がない場合にも適切にスコアを与えることであるためである．ただし，IBM モデル 3 はある程度複雑なモデルであり，多言語音声の同時認識にとっての TM はより簡易なモデルでも十分な可能性がある．このようなモデル化手法の検討はこれまで十分に行われていない．したがって本章では，IBM モデル 1，モデル 2，モデル 3 の比較を行う．

各 IBM モデル [13] の特徴を以下に示す．

- IBM モデル 1：$P(J, A|E)$ 計算時に，Lexicon model（単語の翻訳確率モデル）と，均一な Distortion model（すべての単語の語順入れ替えの確率（Distortion model スコア）を等確率とするモデル）を用いるモデル
- IBM モデル 2：$P(J, A|E)$ 計算時に，Lexicon model と，単語の語順入れ替え確率を考慮する Distortion model[*3] を用いるモデル．モデル 1 と Distortion model のみが異なる．
- IBM モデル 3：$P(J, A|E)$ 計算時に，Lexicon model，Distortion model，NULL Generation model，Fertility model を用いるモデル．モデル 1，モデル 2 と比べると，NULL Generation model（対訳中に対応する単語が存在しない場合）と Fertility model（1 対多の単語対応）を考慮している点が異なる．

なお，IBM モデル 1 とモデル 2 では，原言語文の長さと目的言語の文の長さで求まる値 ϵ を考慮する必要がある．具体的には $P(J, A|E)$ のスコア計算時に ϵ を乗じる必要があるが，本実験では用いない（1 と見なす）．これは，ϵ は J と E が固定されれば定数となり，本実験では式 (3) の δ で反映できると考えるためである．

5.1 TM モデル化手法の比較

TM の種類の違いによる影響について調査した．正確な比較のために TM の語彙はすべてのデータで設定したものとした（日本語 37k，英語 34k）．ここでも各 TM を用いた実験において，式 (3) の各重みとして 3 章で述べたデフォルト値（$\alpha = 6$，$\beta = 1$，$\gamma = 1$，$\delta = 5$）を用いた．また，式 (14) の閾値を $THRES_{rel} = 0$ とした．すなわち各 TM で $P(J|E) = \max_A P(J, A|E)$ と近似を行い，アライメント計算を 1 回としたとき（最高速設定）の結果を比較した．

表 4 に各 TM を用いた音声認識の結果をまとめる．ベースラインシステム，すなわち他の言語の情報を用いない従来の音声認識システムによる WER は 12.58%である．これに対し，対訳テキストと TM を用いた音声認識（提案手法）では，どの TM を用いても WER の改善が得られることが分かる．対訳情報に誤りを含まない場合（英語の単語誤り率 0%）では，最も単純な TM である IBM モデル 1 を用いた場合でも WER が 6.1%（12.58%→11.81%）改善された．各 TM を用いた多言語音声認識を比較したところ，IBM モデル 3，モデル 2，モデル 1 の順に低い WER が得

[*3] モデル 3 の Distortion model と基本的に同じであるが，言語ペアの向きが逆方向のモデル．Alignment model と称される．

表 4 多言語音声認識における TM のモデル化手法の比較（日本語 WER（%））
Table 4 Comparison of TM modelings in multilingual speech recognition (WER (%) in Japanese speech recognition).

TM	対訳英語の単語誤り率			
	0%	10%	20%	30%
IBM モデル 1	11.81%	11.87%	11.81%	11.93%
IBM モデル 2	11.47%	11.73%	11.82%	11.99%
IBM モデル 3	10.82%	10.71%	11.08%	10.97%
なし（音声認識のみ）	12.58%			

TM 学習データ：会話 + 記事 + 新聞

表 5 話者ごとの音声認識精度（WER）
Table 5 Speech recognition performance for each speaker (WER).

話者	翻訳モデルなし	翻訳モデルあり（IBM モデル 3）
A	15.69%	13.87%
B	14.25%	11.82%
C	11.22%	7.97%
D	5.68%	4.40%
E	16.03%	16.01%
Ave.	12.58%	10.82%

※：対訳英語情報の誤りなし

られた．IBM モデル 3 を TM として用いることで，WER の 14.0%（12.58%→10.82%）の改善が得られた．対訳英語に誤りを含む場合（英語の単語誤り率，10%，20%，30%）でもすべての TM で WER の改善が得られ，IBM モデル 3 を用いたときに最も低い WER が得られた．

表 5 に，IBM モデル 3 を使用した場合の各話者（5 名）に対する結果（対訳英語の誤りなし）をまとめる．すべての話者に対して提案手法により音声認識性能が向上している．

IBM モデル 1，モデル 2，モデル 3 とも TM として有効であることが分かった．本実験では，IBM モデル 3 による精度改善が最も大きかった．IBM モデル 1 とモデル 2 は，語順の入れ替えモデルの複雑さについてのみ異なるが，両者を用いた多言語音声認識では大きな差はみられなかった．このことは，IBM モデル 3 から語順入れ替えのみをさらに複雑にしたモデルである IBM モデル 4 およびモデル 5 [13] を用いた多言語同時音声認識では，さらなる精度向上を得られにくい可能性を示唆している．IBM モデル 2 とモデル 3 の比較からは，両者の大きな相違点が対応する単語がない場合と 1 対多の単語対応を考慮している点であることから，これらのモデル化が多言語同時音声認識にとって重要であることが示唆された．多対多の単語対応を考慮できるモデル（たとえばフレーズ翻訳モデル）と IBM モデル 4，モデル 5 を含めた各種 IBM モデルの比較を行い，適したモデルを明らかにしていく必要性を確認した．

表 6 TM の学習データと音声認識結果
Table 6 Comparison of TMs trained by different training data.

TM 学習データの種類	TM	WER
なし（ASR のみ）	なし	12.58%
会話	IBM-model3	13.19%
記事		11.65%
新聞		10.62%
会話 + 記事 + 新聞		10.82%

※：対訳英語情報の誤りなし

6. 翻訳モデルの学習データの影響

最後に，同時音声認識のための TM の学習データについての調査を行った．具体的には，学習データと評価データとのドメインの一致度の影響，および学習データの量の影響を調査した．このような調査は，同時音声認識の枠組みにおいては行われていない．

ここでは，表 2 に示す 3 つの学習データを 1 つまたは複数個用いて TM の学習を行い，比較を行った．これらのうち会話データは，評価データと大きく異なるドメインのデータであり，記事データと新聞データは評価データと同じ新聞記事のドメインであるがテキストデータサイズが大きく異なるものである．なお，TM には前章で最も高い精度が得られた IBM モデル 3 を用いた．語彙はすべての TM で統一した（日本語 37k 単語，英語 34k 単語）．

これらの TM を用いて音声認識を行った結果を**表 6** に示す．ここでも，式 (14) の閾値は $THRES_{rel} = 0$ である．異なるドメイン（会話）のデータで TM を学習した場合は，TM を用いることによる精度改善が得られないことが分かった．ドメインの一致するデータ（記事，新聞）で TM を学習した場合は，TM を用いることによって精度改善が得られた．記事データよりも新聞データのほうがデータサイズが大きく，新聞データから学習した TM を用いることがより有効であることが分かる．すべてのデータを用いて TM を学習したときの音声認識結果も表 6 の最下段に示されている．ドメインが一致するデータに，ドメインが異なるデータ（会話）を加えても効果がないことが分かる．これらのことは，多言語音声の同時認識においてもドメインが一致する大量のデータで TM を学習することが重要であることを示している．

7. おわりに

多言語音声の同時認識における，TM スコア計算の高速化，TM のモデル化，および TM の学習データについて研究した．TM スコア計算の高速化の提案を行い，音声認識の性能を保ちつつ計算コストを削減できることを示した．TM としては，IBM モデル 1 からモデル 3 まですべて有効であることが分かり，本実験では，IBM モデル 3 により最も高い精度改善が得られた．さらに，TM の学習データと

評価データのドメインが一致した大量のコーパスでTMを学習することの重要性を確認した．

本研究では，あらかじめ対応のとれた対訳発話を用いて実験を行った．また，英語発話には音声認識誤りのないデータと擬似的に誤りを生成したデータを用いた．今後は，実際の話し言葉音声（同時通訳音声）を用いた同時音声認識システムの評価，および複数言語の同一内容発話の自動対応付けとそれに基づく音声認識の評価を行っていく予定である．

謝辞 本研究は科研費（21700210）の助成を受けて行われたものである．

参考文献

[1] 南條浩輝：多言語音声の同時認識枠組みの提案，情報処理学会論文誌，Vol.49, No.12, pp.4044–4048 (2008).

[2] 永田昌明：統計的機械翻訳，オペレーションズ・リサーチ，pp.700–705 (2007).

[3] Och, F.J. and Ney, H.: A Systematic Comparison of Various Statistical Alignment Models, *Computational Linguistics*, Vol.29, No.1, pp.19–51 (2003).

[4] Paulik, M., Fugen, C., Stuker, S., Schultz, T., Schaaf, T. and Waibel, A.: Document Driven Machine Translation Enhanced ASR, *Proc. INTERSPEECH*, pp.2261–2264 (2005).

[5] 宮野智靖：The NEWS HOUR リスニング，語研 (1999).

[6] 宮野智靖：The NEWS HOUR リスニング (2)，語研 (2000).

[7] 河原達也，李 晃伸：連続音声認識ソフトウエア Julius，人工知能学会誌，Vol.20, pp.41–49 (2005).

[8] 秋田祐哉，河原達也：話し言葉音声認識のための汎用的な統計的発音変動モデル，電子情報通信学会論文誌，Vol.J88-DII, No.9, pp.1780–1789 (2005).

[9] 尾上和穂，佐藤庄衛，小林彰夫，本間真一，今井 亨：帯域フィルタ出力の時間変化特徴量を利用したニュース音声認識，情報処理学会研究報告，2005-SLP-59-35 (2005).

[10] Morimoto, T., Uratani, N., Takezawa, T., Furuse, O., Sobashima, Y., Iida, H., Nakamura, A., Sagisaka, Y., Higuchi, N. and Yamazaki, Y.: A speech and language database for speech translation research, *Proc. ICSLP94*, pp.1791–1794 (1994).

[11] 江原暉将，小倉健太郎，篠崎直子，森元 逞，樽松 明：電話またはキーボードを介した対話に基づく対話データベース ADD の構築，情報処理学会論文誌，Vol.33, No.4, pp.448–456 (1992).

[12] Utiyama, M. and Isahara, H.: Reliable Measures for Aligning Japanese-English News Articles and Sentences, *ACL-2003*, pp.72–79 (2003).

[13] Brown, P.F., Pietra, V.J.D., Pietra, S.A.D. and Mercer, R.L.: The mathematics of statistical machine translation: Parameter estimation, *Computational Linguistics – Special issue on using large corpora: II*, Vol.19, No.2, pp.263–311 (1993).

[14] Och, F.J., Ueffing, N. and Ney, H.: An Efficient A* Search Algorithm for Statistical Machine Translation, *Data-Driven Machine Translation Workshop*, pp.55–62 (2001).

[15] Germann, U., Jahr, M., Knight, K., Marcu, D. and Yamada, K.: Fast decoding and optimal decoding for machine translation, *Proc. 39th Annual Meeting on Association for Computational Linguistics*, *ACL 2001*, Stroudsburg, PA, USA, Association for Computational Linguistics, pp.228–235 (2001).

[16] Faruquie, T.A., Maji, H.K. and Udupa, R.: A New Decoding Algorithm for Statistical Machine Translation: Design and Implementation, *ALENEX/ANALCO*, pp.180–194 (2005).

[17] Udupa, R., Faruquie, T.A. and Maji, H.K.: An algorithmic framework for the decoding problem in statistical machine translation, *Proc. 20th International Conference on Computational Linguistics*, *COLING 2004*, Stroudsburg, PA, USA, Association for Computational Linguistics (2004).

大村 絵梨

2010 年龍谷大学理工学部情報メディア学科卒業．2012 年同大学院修士課程修了．在学中，音声認識に関する研究に取り組む．

南條 浩輝 （正会員）

1999 年京都大学工学部情報学科卒業．2001 年同大学院情報学研究科修士課程修了．2004 年同大学院情報学研究科博士後期課程修了．同年龍谷大学理工学部助手．2007 年同助教．音声言語処理，特に音声認識・理解に関する研究に従事．日本音響学会，電子情報通信学会，IEEE，ISCA 各会員．2008 年度日本音響学会粟屋潔学術奨励賞受賞．