7-2　全国音声翻訳実証実験の概要

7-2 Speech-to-Speech Translation System Field Experiments in All Over Japan

安田圭志　松田繁樹
YASUDA Keiji and MATSUDA Shigeki

要旨

平成 21 年度に全国 5 地方で実施された「地域の観光振興に貢献する自動音声翻訳技術の実証実験」について概説する。次に、ここで得られたデータを用いて音声翻訳システム性能を改善する実験について述べる。実験の結果から、あらかじめ開発セット等を用意しておき、データをフィルタリングする際の閾値や、アダプテーションを適用するモデルを適宜決めて行く必要があるものの、人手による書き起こしや、対訳作成無しにシステム性能の改善が得られることが示された。

We explain field experiments conducted during the 2009 fiscal year in five areas of Japan. We also show the experiments of evaluation and data selection method from speech translation field data. The data selection method selects useful data from filed data by using a development data set. According to the experimental results, the proposed data selection method gives the improvement of the speech-to-speech translation systems.

[キーワード]
音声翻訳実証実験，統計翻訳，音声翻訳システム，音声翻訳システム実利用データ
Speech translation field experiment, Statistical machine translation, Speech translation system, Speech translation field data

1　音声翻訳実証実験

本稿では、平成 21 年度に実施された音声翻訳実証実験 [1] について述べる。次に、本実証実験により収集されたデータセットを用いた音声翻訳システム改善手法について述べる。本実証実験は、自動音声翻訳技術の翻訳精度の飛躍的向上及び訪日観光分野における同技術を活用したサービスの早期実用化を図ることを目的としており、総務省が「地域の観光振興に貢献する自動音声翻訳技術の実証実験」（総事業予算額 9.85 億円）を民間法人等に委託して実施した。

実証実験は、日英中韓の 4 ヶ国語を対象とし、図 1 に示す通り、全国 5 地方の観光施設等約 370 箇所に約 1,700 台の端末を設置して行われた。実験期間中には、約 20 万件のアクセスが記録された。このように大規模で、実利用に近い条件下での実証実験は、世界的にも類を見ない。独立行政法人情報通信研究機構（National Institute of Information and Communications Technology）は、実証実験を受託したすべての事業者に対して音声翻訳技術を提供するとともに、実験システム構築、運用、データ分析等の面で全面的にサポートした。

2　システム構成

各地方プロジェクトが構築した実証実験システムの簡略化された構成図を図 2 に示す。音声翻訳端末は、スマートフォン、ノート PC などからなり、台数は 300 ～ 500 である。端末で入力された音声は、16 kHz サンプリングの ADPCM 形式で音声翻訳サーバーに送られる。音声翻訳サーバーは、実際には言語ごとに用意された音声認識、機械翻訳、音声合成用のサーバー群から構成される。翻訳結果は、テキストおよび合成音声の

形で端末に送信される。また、入力音声、音声認識結果、翻訳結果は、日時、端末ID、言語指定等の情報とともに利用ログとしてシステム内に蓄積される。

2.1 音声認識システムの概要

本プロジェクトで使用する音声認識システムは、フロントエンド部とデコーダ部から構成されている。フロントエンド部では、パーティクルフィルタにより時間と共に変化する雑音パワーを逐次推定し、非定常な雑音の抑圧[2]が行われる。デコーダ部では、音響モデルとして隠れマルコフモデル（HMM: Hidden Markov Model）を、言語モデルとして単語クラスN-gramの拡張である多重クラス複合N-gram[3]を用い、2パスで認識を行う。第1パスでは音響モデルと2-gram言語モデルを用いて単語グラフを生成し、第2パスでtrigram言語モデルを用いて単語ラティスのリスコアリングを行い認識結果を探索する。

音響モデルは、高齢者を含む青年男女約4,500名の発話した約400時間の音声コーパスを用いて音響モデルを推定した。本実証実験で使用される音声認識システムは、屋外など騒音環境下で使用されることを想定している。そのため、雑音に対して頑健な音声認識を実現するため、車の走行音や街路、駅コンコースなど様々な場所で収録した雑音を10～30 dBのSN比でランダムに重畳した学習データを用いて音響モデルの推定を行った。

次に、言語モデルは、旅行会話を中心として収集されたテキストコーパス約74万文を用いて多重クラス複合2-gram多重クラス3-gram言語モデルを推定した。語彙サイズは、約5万語である。

2.1.1 各地方向け言語モデルカスタマイズ

各地方の固有名詞や固有表現（方言）を用いて、地方ごとに言語モデルのカスタマイズを行った。各地方の固有名詞（約5,000単語）は、地名や施設名等のカテゴリ毎に、基本辞書中の代表単語の言語確率を付与することにより追加した。また、固有表現（約3,000文）を用いて単語N-gramを学習し、基本モデルと線形結合を行うことにより、言語モデルの適応を行った。

2.2 機械翻訳システムの概要

図3は、音声認識、機械翻訳、音声合成からなる音声翻訳システムの内、機械翻訳部の処理の詳細である。機械翻訳部は、主に統計的機械翻訳と2つの翻訳メモリから構成されている。統計翻訳システムは、フレーズベース型統計翻訳[4]の枠組みを利用した。本手法は、翻訳対象の原言語の単語列（f）に対する目的言語の単語列（e）の確率を次式により求める。

$$p(e|f) = \frac{\exp\left(\sum_{i=1}^{M} \lambda_i h_i(e,f)\right)}{\sum_{e'} \exp\left(\sum_{i=1}^{M} \lambda_i h_i(e',f)\right)} \qquad (1)$$

ここで、e'は、fに対する翻訳候補文を表す。$h_i(e, f)$は、学習コーパスから得られる素性関数で、目的言語から原言語、原言語から目的言語の単語やフレーズ単位の翻訳確率（翻訳モデル）や、目的言語の言語モデル等からなる8つの素性関数[5]である。また、λ_iとMは、それぞれ、各素性関数に対する重みと素性関数の数（8）を表す。

式（1）の分母は一定とし、式（2）により翻訳結果$\hat{e}$を求める。

$$\hat{e}(f, \lambda_1^M) = \mathrm{argmax}_e \sum_{i=1}^{M} \lambda_i h_i(e,f) \qquad (2)$$

各モデルの学習には、MOSESツールキット[5]とSRILMツールキット[6]とを用いて、翻訳モデルと言語モデルの学習を行っている。

実証実験において、実験実施地方毎に、以下の2種類のデータを事前に収集した。

- 固有名詞：数千語からなる地域特有の名詞。訳語（英中韓）やその語が属するカテゴリの情報も合わせて整備されている。
- 固有表現：設置場所への事前のヒアリング等により収集した数千文からなる地域固有の表現（テキスト）とその英中韓訳。ここには、設置店舗などにおいて、業務上不可欠な表現なども含まれる。

固有名詞の利用法としては、[7]で提案された手法を用い前処理で固有名詞からカテゴリのトークンに置き換え翻訳を行っている。また、実際の各モデルの学習には、BTECコーパス[8]に加え、固有表現を用いている。データの使い方は、まず、前述のツールキットを用いて、データ毎に

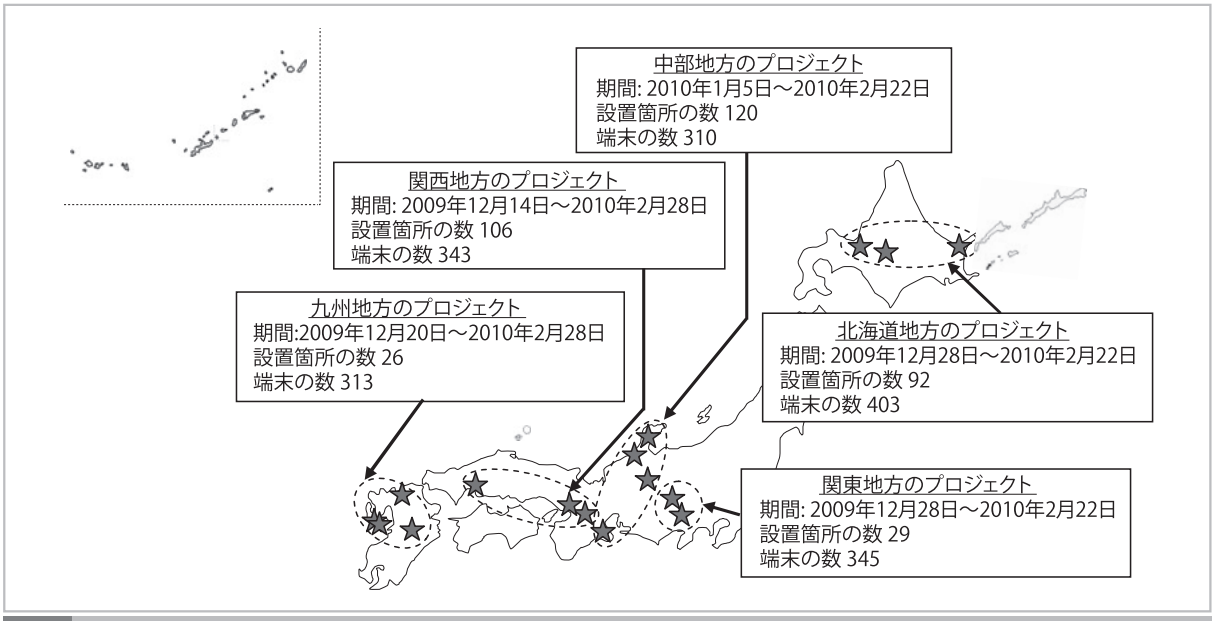

図 1　5 地方における実証実験の概要

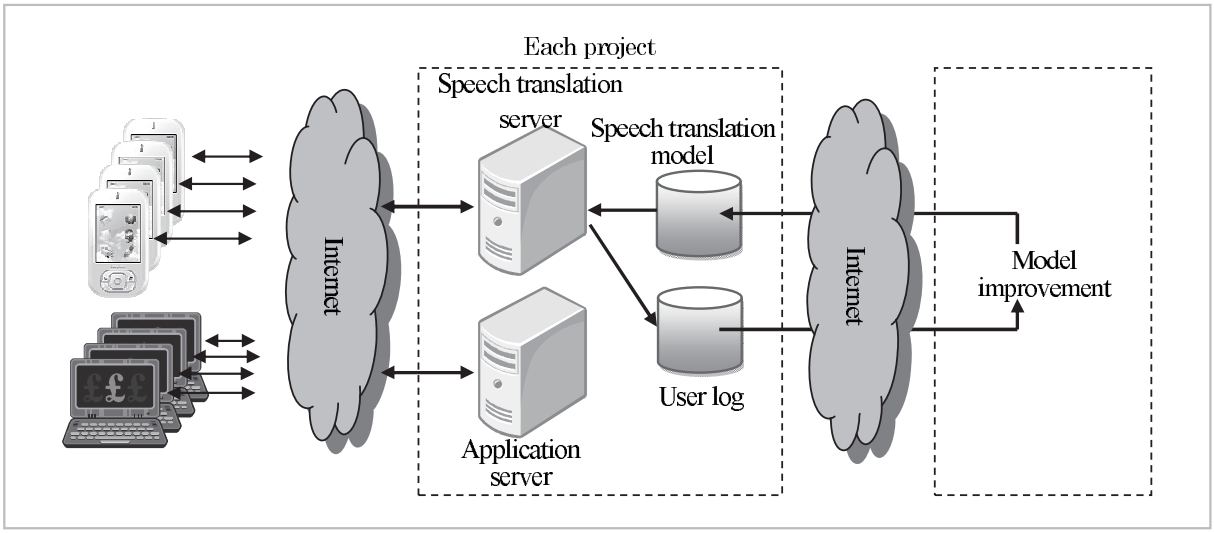

図 2　音声翻訳実証実験におけるシステム構成図

個別に素性関数を学習し、次に、両素性関数を次式により線形結合して利用する。

$$\begin{aligned}\mathbf{h}_{baseline}(e,f) &= \mu\mathbf{h}_{btec}(e,f) \\ &\quad +(1-\mu)\mathbf{h}_{regional}(e,f)\end{aligned} \tag{3}$$

ここで $\mathbf{h}_{btec}$ は、BTEC コーパスを用いて学習した式（1）の 8 つの素性関数を、$\mathbf{h}_{regional}$ は、固有表現を用いて学習した 8 つの素性関数をそれぞれ表す。$\mathbf{h}_{baseline}$ は、これらを線形結合することにより得られた 8 つの素性関数を表す。また、重み μ の値については、**4.1** で述べる。

固有表現の利用法として、コーパスレベルでデータを 1 つにまとめ、1 つのモデルを学習する方法も考えられる。このような方法では、固有表現の追加等が生じた場合に、再度、全データを用いたモデル学習を行う必要がある。一方、式（3）による手法では、固有表現のみの小規模データでの再学習のみが必要となり、メンテナンスが容易であるというメリットがある。

また、固有表現は翻訳メモリにおいても用いられている。図 3 の翻訳メモリ 1 では、前述の

BTEC コーパスを、翻訳メモリ 2 では、各地方ごとに収集した固有表現をそれぞれ用いている。

3 実証実験収集データ利用法

提案手法では、機械翻訳結果を原言語に逆翻訳した結果を用い、モデルのアダプテーションに用いるデータの取捨選択を行う。**3.1** では、提案手法と 2 つの従来法による、教師無しフィルタリング手法について述べ、**3.2** では、従来法による教師有りフィルタリングについて述べる。最後に、**3.3** では取捨選択されたデータをモデルアダプテーションに用いる方法について説明する。

3.1 教師無しフィルタリング

3.1.1 正規化翻訳スコアを用いたフィルタリング

正規化翻訳スコアを用いた従来法 [9] では、次式を用いて正規化翻訳スコア（S_{trans}）を計算し、この値が閾値以上の場合は、アダプテーション用データとして利用する。

$$S_{trans} = p(e|f)^{\frac{1}{n_e}} \tag{4}$$

ここで n_e は、翻訳結果における単語数を、$p(e|f)$ は式（1）により計算される翻訳確率をそれぞれ表す。

3.1.2 原言語パープレキシティを用いたフィルタリング

原言語パープレキシティを用いた従来法 [10] では、次式を用いて原言語パープレキシティ（$S_{pp}(f)$）を計算し、この値が閾値以下の場合は、アダプテーション用データとして利用する。

$$S_{pp} = p(f)^{-\frac{1}{n_f}} \tag{5}$$

ここで n_f は、入力文における単語数を、$p(f)$ は原言語側の言語モデルにより与えられる入力文のパープレキシティをそれぞれ表す。本論文では、**2** で述べた BTEC コーパスの原言語側を利用して学習した言語モデルを用いる。

3.1.3 提案手法

提案手法では、まず、順方向の機械翻訳結果を、再度原言語に機械翻訳する。次に、順方向の機械翻訳への入力である音声認識結果（f）を参照訳とみなし、逆翻訳の結果（f'）の翻訳自動値を計算する。本実験では、翻訳自動評価値として、次式で計算される PER（Position independent word Error Rate）を用いる。

$$S_{src\_per} = PER(f, f') \tag{6}$$

ここで PER は、音声認識の評価等で計算される単語誤り率を、語順を無視して計算した値である。機械翻訳の評価においては、自動評価値として BLEU [11] がしばしば用いられるが、発話単位の BLEU を本タスクにおいて適用すると、多くのスコアが 0 となってしまい、優劣がつけられないという問題が生じる。これは、評価対象の翻訳の質と評価手法の分解能のミスマッチの問題であるが、本研究では比較的整合性の高い PER を用いている。

3.2 教師有りフィルタリング

教師有りフィルタリングは、人手により作成した参照訳を用い、機械翻訳結果に対して文単位の翻訳自動評価を行うことにより、訳質の高い翻訳を選択する方法である。本論文で取り扱う教師無しアダプテーションの目的は、人手による参照訳を作成しないでアダプテーションを行うことであるため、フィルタリング時に参照訳を利用するというのは非現実的である。しかしながら、本論文では、従来研究 [12] の追実験や、教師の有無でのフィルタリング性能の比較といった目的から、教師有りフィルタリングの実験も行っている。

従来研究の研究 [12] では、翻訳自動評価として BLEU [11] が用いられていたが、本研究では、次式で計算される $PER(S_{tgt\_per})$ を用いる。

$$S_{tgt\_per} = PER(e, e') \tag{7}$$

3.3 アダプテーション手法

選択された実利用データは、**2** で述べた固有表現データとともに、次に述べる方法で用いる。

Step 1 得られた実利用データと、**2** で述べた固有表現とを結合し、アダプテーション用コーパスとする。

Step 2 Step 1 で得られたコーパスを用いて、式（1）の 8 つ素性関数（$\mathbf{h}_{field}(e, f)$）の学習を行う。

Step 3 BTEC コーパスを用いて学習したモデル（$\mathbf{h}_{btec}(e, f)$）と $\mathbf{h}_{field}(e, f)$ を、式（8）によ

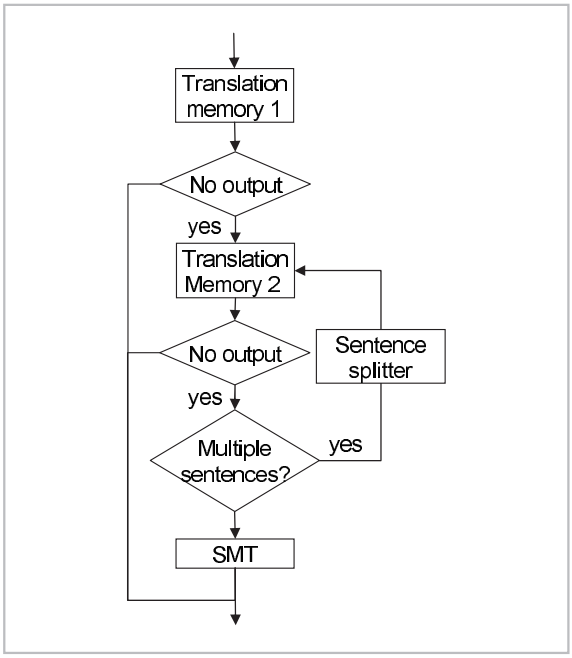

図３　機械翻訳部の処理の流れ

り線形結合し、アダプテーションモデル（$\mathbf{h}_{adapted}(e,f)$）とする。

$$\mathbf{h}_{adapted}(e,f) = \mu\mathbf{h}_{btec}(e,f) + (1-\mu)\mathbf{h}_{field}(e,f) \quad (8)$$

重みμの値については、**4.1** で述べる。

4　実験

次に、実証実験で収集された実利用データを用いた、音声翻訳システム性能の改善方法について検討する。通常の音声翻訳システムの研究では、得られたデータを人手によって整備し（書き起し、対訳付与等）、整備されたデータを用いてシステムを再学習する。この方法は、非常に効果的であるが、データ整備に時間やコストがかかるという問題がある。この問題を解決するため、ここでは、人手による整備を必要としないデータ利用方法について検討する。

4.1　実験条件

実験では、５つの実証実験実施地方の内、予備実験における訳質の主観評価において、訳質が最も低かった北海道と、最も高かった九州のデータを用いた。翻訳方向は、全地域共通してニーズが高く、最も潤沢なデータが収集された日英方向とした。

モデル学習に用いた日英 BTEC コーパスは、691,829 文からなり、北海道と九州地区の固有表現はそれぞれ、3,000 文と 5,095 文からなる。

機械翻訳の評価では、それぞれの地方のデータに対してランダムに抽出した 100 文を評価セットとして用いた。北海道と九州のテストセットパープレキシティは、日本語 BTEC コーパス単体で学習した言語モデルでの、北海道と九州のテストセットパープレキシティは、それぞれ、40.98 と 19.36、また、式（3）により得られるベースライン言語モデルでは、それぞれ、42.64 と 17.17 であった。これらのことからも、九州のテストセットと比較し、北海道のテストセットの難易度が高いことが分かる。

予備実験ならびに、本実験の訳質の評価としては、５段階 S（Perfect）、A（Correct）、B（Fair）、C（Acceptable）、D（Nonsense）の主観評価を実施した。また、**4.2** で示す全ての評価において、テスト文を機械翻訳する際の入力は、音声認識を含まないテキスト入力としている。

式（2）における、各素性関数に対する８つの重みλ_iについては、ベースラインモデルと（$\mathbf{h}_{baseline}$）と、500 文からなる BTEC 開発セットを用いた MERT（Minimum Error Rate Training）[13] により値を決定し、全ての実験条件において同一の値を用いている。また、式（3）と式（8）におけるμについても同様、BTEC 開発セットを用いた予備実験により、0.9 としている。これらの設定は、実証実験実施時と同様の設定である。実証実験実施時にこのような設定にした理由を以下に示す。

- 実証実験実施前には、実利用データの十分な量の開発セットが得られない。
- 雪祭り等のイベントなどにも利用されることがあり、利用時期によって、音声翻訳システムが利用される場面が大きく異なることがある。そのため、ある一定の時期にサンプリングした小規模な開発セットでのパラメータチューニングでは、過適応になる恐れがある。
- BTEC 開発セットを用いたパラメータチューニングにより、最低でも BTEC のド

メインの発話についての性能は維持することができる。

より最適なパラメータ設定にするためには、事後的に、実証実験実施全期間の実利用データから、ランダムにサンプリングした開発セットを用いて、個別にλ_iとμの値をチューニングすべきであるが、本論文では、提案手法が利用されるであろう状況に合わせた実験を行うため、実証実験実施時と同様の設定にしている。

4.2 実験結果

表1は、実利用データの取捨選択を行わなかった場合の結果である。各地方の結果において、1行目はベースライン、2行目は利用可能な実利用データ全てを用いて教師無しアダプテーションを行った結果、3行目は入力音声の書き起こしは人手で行い、アダプテーションに用いる目的言語側の情報として機械翻訳の出力を用いた場合の結果（一部教師有りアダプテーション）を表す。4行目は書き起こしも対訳作成も全て人手で行った結果（教師有りアダプテーション）で、アダプテーションによる性能改善の上限を表す。

同地域においても、条件により実利用データのサイズが異なるのは、音声認識や機械翻訳の過程で出力が得られなかったデータはアダプテーションに利用していないためである。

先に述べたように、九州のデータにおいては、北海道よりも固有表現の文数が多くなっている。固有表現の文数の影響をみるため、表1の最終行（下線部）では、ランダムサンプリングにより、九州の固有表現の数を北海道と同じ3,000文にした場合の結果を示している。ここでは3度のランダムサンプリングを行い、主観評価した結果の平均を示している。

表中の白いセルは、ベースラインの性能を上回った場合、ライトグレーのセルはベースラインと同じ性能の場合、ダークグレーのセルは、ベースラインの性能を下回った場合をそれぞれ表す*。表1を見ると、教師無しアダプテーションでは全く改善が得られていない。一部教師有りアダプテーションにおいては、一部の条件で性能の改善が得られているものの、性能が劣化することもある。一方、教師有りアダプテーションでは、全ての場合において、性能の改善が得られている。これらの結果から、特に教師無しアダプテーション時においては、実利用データをそのまま用いると、翻訳性能の劣化を招く可能性があることが分かる。実利用データのフィルタリングは、このような劣化を防ぐことを目的としており、表2では提案手法の効果を検証している。

＊　表2～3においても同様の配色を用いる。

表1 教師無しアダプテーションと教師無しアダプテーション時における評価結果（データフィルタリング無し）

Project Area	System Type	Additional Field Data			Ratio（%）			
		Transcription	Translation	Size（# of sentences）	S	S, A	S, A, B	S, A, B, C
Hokkaido	Baseline	N/A	N/A	0	29	38	55	62
	Baseline + unannotated data	ASR	MT	9602	29	38	53	61
	Baseline + unannotated data 1	Manual	MT	10009	31	39	51	62
	Baseline + unannotated data 2 (Upper bound)	Manual	Manual	10335	34	44	61	68
Kyushu	Baseline	N/A	N/A	0	50	62	72	76
	Baseline + unannotated data	ASR	MT	9722	50	60	71	76
	Baseline + unannotated data 1	Manual	MT	10337	49	62	72	77
	BL + unannotated data 2 (Upper bound)	Manual	Manual	14138	55	64	74	79
	Baseline with 3000 regional expressions	N/A	N/A	0	47.7	60.0	69.0	73.3

最後に、各地域のベースライン間の比較を行うと、九州のデータでは、固有表現のサイズを3,000文に縮小することにより、性能の劣化がみられている。しかしながら、同じく固有表現のサイズが3,000文の北海道よりも大幅に訳質が高いことが分かる。これらのことを考慮すると、九州のテストセットが北海道のテストセットよりも翻訳が容易なセットになっていると言える。この違いは、九州の実証実験では、空港や観光地において説明員が配置され、比較的制御された状態で利用されていたことに起因すると考えられる。

表2は、音声認識誤りと機械翻訳誤りが含まれる、教師無しアダプテーションの条件で、従来手法と提案手法により、アダプテーションに用いる音声翻訳実利用データのフィルタリングを行った結果を表している。ここでの音声認識システムの性能は、北海道と九州のテストセットで、それぞれ単語誤り率29.9%と20.3%である。

提案手法では、フィルタリングの閾値を、それぞれ0.1、0.2、0.4の場合での結果を示している。公正な比較を行うため、従来法においては、提案手法における閾値が0.1、0.2、0.4のそれぞれの場合と同じ文数のデータが得られる様、閾値を調整している。

まず、提案手法における閾値を0.1とした場合に得られるデータ量を地域ごとに比較すると、北海道では、1,224文（北海道実利用データ全体の12.7%）であるのに対し、九州では4,560文（九

表2　教師無しアダプテーション時における評価結果（データフィルタリング有り）

Project Area	Additional field data		Ratio（%）			
	Filtering function	Size（# of sentences）	S	S, A	S, A, B	S, A, B, C
Hokkaido	N/A（Baseline）	0	29	38	55	62
	S_{trans}（eq. 4）	1244	32	41	54	64
		1861	30	40	52	61
		3565	32	42	53	63
	S_{pp}（eq. 5）	1244	30	41	53	65
		1861	30	41	53	65
		3565	30	40	53	62
	S_{src_per} <= 0.1（eq. 6）	1244	31	40	55	66
	S_{src_per} <= 0.2（eq. 6）	1861	32	41	56	69
	S_{src_per} <= 0.4（eq. 6）	3565	32	41	56	66
	S_{tgt_per}（eq. 7）	1244	32	40	54	64
		1861	30	40	53	63
		3565	31	41	53	63
Kyushu	N/A（Baseline）	0	50	62	72	76
	S_{trans}（eq. 4）	4560	49	62	72	75
		5274	49	62	72	75
		6699	49	61	71	77
	S_{pp}（eq. 5）	4560	48	60	70	74
		5274	48	59	69	74
		6699	48	59	71	74
	S_{src_per} <= 0.1（eq. 6）	4560	49	60	70	74
	S_{src_per} <= 0.2（eq. 6）	5274	49	61	71	75
	S_{src_per} <= 0.4（eq. 6）	6699	51	61	71	74
	S_{tgt_per}（eq. 7）	4560	50	62	72	76
		5274	50	61	71	75
		6699	49	60	71	75

州実利用データ全体の46.9％）と非常に多くのデータが得られている。これには以下の2点が起因すると考えられる。

- 北海道と比較し、九州のデータでは単語誤り率が低いため、訳質への音声認識誤りの悪影響が少なく、その結果として、逆翻訳結果と入力文の一致度が高くなる。
- 北海道と比較し、九州のデータでは翻訳の難易度が低いデータが多く含まれており、順方向逆方向とも比較的正しく翻訳され、その結果として逆翻訳結果と入力文の一致度が高くなる。

次に、北海道のデータセットについて注目すると、提案手法では、閾値を0.1とした場合の一部で性能が劣化しているものの、ほぼ全ての場合において、システム性能が向上している。正規化翻訳スコア（S_{trans}）と、原言語パープレキシティ（S_{pp}）においても、改善が得られているものの、提案手法による改善の方が大きい場合が多い。また、教師有りフィルタリングの結果（S_{tgt_pp}）と提案手法（S_{src_per}）とを比較すると、提案手法は教師無しにも関わらず、教師有りフィルタリング利用時と同等以上の性能改善が得られていることが分かる。また、$S_{src_per} <= 0.2$の条件においては、S，A，B，Cの割合が、表1で示したUpper boundを超えている。翻訳誤りや音声認識誤りが含まれない場合においては、ドメイン外の学習文を除去することにより、訳質性能が向上しうることが示されている[14]。$S_{src_per} <= 0.2$においては、偶発的にこのような条件と重なった可能性があるが、表1のUpper boundにおいても[14]のような手法によりデータの取捨選択を行うことにより、Upper boundを超える性能が得られると考えられる。

4.2.1 九州データセットについての詳細分析

表3は、提案手法において、言語モデルのみ、

表3 モデルごとのアダプテーション時における評価結果（データフィルタリング有り）

Project Area	Filtering function	Additional field data		Ratio（%）			
		Used for LM training	Used for TM training	S	S, A	S, A, B	S, A, B, C
Hokkaido	N/A（Baseline）	No	No	29	38	55	62
	$S_{src\_per} <= 0.1$	Yes	Yes	31	40	55	66
	$S_{src\_per} <= 0.2$	Yes	Yes	32	41	56	69
	$S_{src\_per} <= 0.4$	Yes	Yes	32	41	56	66
	$S_{src\_per} <= 0.1$	Yes	No	32	40	54	63
	$S_{src\_per} <= 0.2$	Yes	No	32	41	56	64
	$S_{src\_per} <= 0.4$	Yes	No	31	41	55	64
	$S_{src\_per} <= 0.1$	No	Yes	31	40	54	64
	$S_{src\_per} <= 0.2$	No	Yes	32	40	55	64
	$S_{src\_per} <= 0.4$	No	Yes	32	40	54	63
Kyushu	N/A（Baseline）	No	No	50	62	72	76
	$S_{src\_per} <= 0.1$	Yes	Yes	49	60	70	74
	$S_{src\_per} <= 0.2$	Yes	Yes	49	61	71	75
	$S_{src\_per} <= 0.4$	Yes	Yes	51	61	71	74
	$S_{src\_per} <= 0.1$	Yes	No	47	57	68	73
	$S_{src\_per} <= 0.2$	Yes	No	47	57	68	73
	$S_{src\_per} <= 0.4$	Yes	No	47	57	68	73
	$S_{src\_per} <= 0.1$	No	Yes	51	62	73	76
	$S_{src\_per} <= 0.2$	No	Yes	50	61	73	76
	$S_{src\_per} <= 0.4$	No	Yes	52	63	73	76

または、翻訳モデルのみのアダプテーションに実利用データを用いた場合の結果を示している。表3を見ると、北海道のデータセットでは、言語、翻訳両モデルのアダプテーションを行った場合に最も性能の改善が大きくなっており、言語モデル単体または翻訳モデル単体へのアダプテーションでは、改善が小さくなっている。

九州のデータセットにおいては、言語モデル単体へのアダプテーション時に、より大きな劣化が生じている。その反面、翻訳モデル単体へのアダプテーション時には改善が得られている。これらのことから、実運用時において、あらかじめ開発セット等を用意しておき、アダプテーションを適用するモデルを決定する方法などをとることにより、性能の劣化を防ぎ、改善が得られることが示された。

5　まとめ

音声認識結果と機械翻訳結果の対からなる音声翻訳の実利用データを用いた機械翻訳アダプテーション手法を提案した。提案手法では、機械翻訳結果を原言語側へ逆翻訳して音声認識結果と比較し、入力文（音声認識結果）と逆翻訳結果が近いデータのみをアダプテーションに用いる。

実験では、平成21年度に全国で実施された音声翻訳実証実験のデータを用いた。実験の結果、ベースラインの性能が低い北海道地区のデータを用いた場合、提案手法による翻訳性能の改善が得られた。一方、ベースラインの性能が高い九州地区のデータでは、性能の劣化が見られた。しかしながら、九州のデータセットにおいては、言語モデルのアダプテーションを行わず、翻訳モデルのみのアダプテーションを行うことにより、ある程度の性能改善が得られることが示された。

実運用時においては、あらかじめ開発セット等を用意しておき、データをフィルタリングする際の閾値や、アダプテーションを適用するモデルを、適宜決めていく必要があるものの、人手による書き起しや対訳作成無しに、システム性能の改善が得られることが示された。

実証実験におけるアンケート結果を見ると、現状の音声翻訳システムの性能は、誰もが満足する十分な性能であるとは言えない。本提案手法は、実利用データを用いて、各モデルの確率値を調整する機能を持っているものの、実利用データに含まれる未知語に対応できるようになるわけでは無い。今後の更なるシステム改善には、WEB等から自動的に固有名詞を獲得する枠組みを音声翻訳システム内に組み込み、実利用データの収集を続けて行く必要がある。

参考文献

1　河井　恒，磯谷亮輔，安田圭志，隅田英一郎，内山将夫，松田繁樹，葦苅　豊，中村　哲，“H21年度全国音声翻訳実証実験の概要，”日本音響学会2010年秋季研究発表会，pp.99–102，2010.

2　M. Fujimoto and S. Nakamura,“A non-stationary noise suppression method based on particle filtering and polyak averaging,”The IEICE Transactions on Information and Systems, Vol. E89-D, No. 3, pp. 922–930, 2006.

3　山本博史，匂坂芳典，“接続の方向性を考慮した多重クラス複合n-gram言語モデル，”信学論，Vol. J83-D-II, No. 11, pp. 2146–2151, 2000.

4　P. Koehn, F. J. Och, and D. Marcu,“Statistical Phrase-Based Translation,”Proc. of Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL), pp. 127–133, 2003.

5　P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst,“Moses: Open source toolkit for statistical machine translation,”Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions, pp. 177–180, Association for Computational Linguistics, June 2007.

6 A. Stolcke, “SRILM-an extensible language modeling toolkit,” Proceedings of the International Conference on Spoken Language Processing, pp. 901–904, 2002.

7 H. Okuma, H. Yamamoto, and E. Sumita, “Introducint a translation dictionary into phrasebased smt,” The IEICE Transactions on Information and Systems, vol. 91-D, no. 7, pp. 2051–2057, 2008.

8 G. Kikui, S. Yamamoto, T. Takezawa, and E. Sumita, “Comparative study on corpora for speech translation,” IEEE Transactions on Audio, Speech and Language Processing, vol. 14(5), pp. 1674–1682, 2006.

9 N. Ueffing, G. Haffari, and A. Sarkar, “Semi-supervised model adaptation for statistical machine translation,” Machine Translation, vol. 21, no. 2, pp. 77–94, 2007.

10 K. Yasuda, R. Zhang, H. Yamamoto, and E. Sumita, “Method of selecting training data to build a compact and efficient translation model,” Proceedings of the Third International Joint Conference on Natural Language Processing, pp. 655–660, 2008.

11 K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, “Bleu: a method for automatic evaluation of machine translation,” Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 311–318, 2002.

12 N. Bach, R. Hsiao, M. Eck, P. Charoenpornsawat, S. Vogel, T. Schultz, I. Lane, A. Waibel, and A. W. Black, “Incremental adaptation of speech-to-speech translation,” Proceedings of NAACL HLT 2009, pp. 149–152, 2009.

13 F. Och, “Minimum error rate training in statistical machine translation,” Proc. of 41st Annual Meeting of the Association for Computational Linguistics (ACL), pp. 160–167, 2003.

14 K. Yasuda, H. Yamamoto, and E. Sumita, “Training set selection for building compact and efficient language models,” The IEICE Transactions on Information and Systems, vol. 92-D, no. 3, pp. 506–511, 2009.

（平成 24 年 6 月 14 日 採録）

安田圭志（やすだけいじ）
ユニバーサルコミュニケーション研究所
多言語翻訳研究室主任研究員
博士（工学）
機械翻訳、自然言語処理
keiji.yasuda@nict.go.jp

松田繁樹（まつだしげき）
ユニバーサルコミュニケーション研究所
音声コミュニケーション研究室
主任研究員
博士（情報科学）
信号処理、音声認識
shigeki.matsuda@nict.go.jp