7-3　VoiceTra実証実験の概要

7-3 VoiceTra Field Experiments

松田繁樹　安田圭志　河井　恒
MATSUDA Shigeki, YASUDA Keiji, and KAWAI Hisashi

要旨

我々は、自分の発話した音声を外国語の音声に自動翻訳するネットワーク型多言語音声翻訳アプリケーション「VoiceTra」を、スマートフォン用アプリとして開発し、AppStoreやAndroid Marketにおいて無料公開している。本稿では、本アプリケーションで用いられている音声翻訳技術について概説する。また、このアプリケーションを通して収集された音声翻訳システム利用時の実データの分析及びそれらを利用した音声翻訳性能の改善について述べる。

We have developed a network-based speech-to-speech translation system “VoiceTra” for smart-phones that interprets users' speech into speech of foreign languages, and made it available to the public at no charge. This article briefly introduces the technologies of speech-to-speech translation and shows performance improvement obtained by using huge amount of real speech data collected by the “VoiceTra”.

[キーワード]
音声翻訳，音声認識，言語翻訳，スマートフォン
Speech to speech translation, Speech recognition, Language translation, Smart-phone

1　まえがき

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及び多言語翻訳研究室では、多言語の自動音声翻訳技術の研究成果を広く周知し、利用データによる性能改善を行うための実証実験として、アップル社のスマートフォンiPhone向けに、ネットワーク型多言語音声翻訳アプリケーション「VoiceTra」（以下、VoiceTraと略する）を、2010年7月末日より無料公開した。2011年4月からは、Android OSが導入されたスマートフォン向けにも実験を開始した。本システムは、主に旅行で用いられる会話を支援するために用いられる。たとえば、日本に来た外国人とのコミュニケーションや、海外旅行中の現地の人との会話で利用されることを想定している。本稿では、VoiceTraの構成及び、システムで用いられている音声認識、言語翻訳システムについて概説する。

2　多言語音声翻訳アプリケーション「VoiceTra」

VoiceTraは、iPhoneやAndroid OSが導入されたスマートフォン用のネットワーク型多言語音声翻訳アプリケーションである。図1左側にVoiceTraの起動画面、中央にVoiceTraの翻訳時の画面、右側に言語選択画面を示す。画面の例は日本語から英語への翻訳の例である。ユーザの発話した「道に迷いました駅はどこですか」の音声認識結果が上段、下段に英語への翻訳結果“I'm lost. Where is the station?”が表示されているのがわかる。中段の日本語は、英語から日本語への逆翻訳の結果である。翻訳方向の変更は、画面上部に表示されている矢印をタップすることにより行われ、相手の発話した外国語音声を日本語に翻訳する。また、翻訳言語の変更は「日本語」や「英語」と書かれた部分をタップすることにより、図1右側の画面が表示され、希望の言語を簡単に選択することができる。

図 1 VoiceTra の起動画面（左側）、翻訳画面（中央）、言語選択画面（右側）

翻訳可能な言語のリストを表 1 に示す。表に示すように 6 つの言語について、音声認識による入力及び音声合成による出力が可能である。また、これら 6 言語を含む合計 21 言語についてテキスト入力による翻訳が可能である。

図 2 に VoiceTra のシステム構成図を示す。図に示すように、ユーザが発話した音声はインターネットを介して多言語音声翻訳サーバへ送信される。サーバでは、音声認識処理、言語翻訳処理、音声合成処理が行われ、各々の結果がクライアントであるスマートフォンへ送信される。

図 3 に実験を開始した 2010 年 8 月からの累計アクセス数のグラフを示す。図に示すように、アプリ公開時より順調にアクセス数を増しており、2012 年 5 月現在、累計アクセス数 750 万である。アクセス数の内訳は、日本語が 76 %、英語が 19%、中国語が 4%である。現在、収集された音声データに対して、音声を実際に聴取し、男性、女性、ネイティブ、ノンネイティブ等の話者属性や、VoiceTra 利用場面や利用形態、利用場所等の分類作業を行っている。

3 多言語音声認識システム

高精度かつ頑健な音声認識を実現するには、話者の違いや発話スタイルの変動、背景雑音などによる歪み、クリッピング等、様々な歪みに対して適切にモデル化することが重要である。1980 年代より、このような変動や歪みに対して確率モデルを適用することで音声認識を行う統計的音声認識手法の研究が盛んに行われてきた。VoiceTra も同様に、統計的音声認識を基礎とした手法により音声認識を行っている。音声の時間的な変化がモデル化された「音響モデル」として隠れマルコフモデル [1]、単語の並び等の言語情報がモデル化された「言語モデル」として N-gram モデル [2] を用い、入力された特徴ベクトル時系列 O に対して最も高い条件付き確率 $P(W \mid O)$ が得られる単語列 W^*が探索される。この処理を数式で表すと次のようになる。

$$W^* = \arg\max_w P(W \mid O)$$
$$= \arg\max_w \frac{P(O \mid W)P(W)}{P(O)}$$

式中の $P(O \mid W)$ は音響モデルを表し、単語列 W に対する音響特徴ベクトル時系列 O の音響尤度が計算される。また、$P(W)$ は言語モデルを表し、単語列 W に対する言語確率が計算される。arg max は、$P(O \mid W)P(W)$ で計算される確率値が最大となる単語列 W^*の探索を表し、音声認識ソフトウェアがこの処理を行う。分母の

表 1　翻訳可能な言語

音声入力、音声出力が可能な言語	テキストによる翻訳が可能な言語
日本語、英語、中国語、インドネシア語、ベトナム語、韓国語	日本語、英語、中国語、台湾華語、韓国語、フランス語、ドイツ語、ヒンディ語、インドネシア語、イタリア語、マレー語、ポルトガル語、ポルトガル語（ブラジル）、ロシア語、スペイン語、タガログ語、タイ語、ベトナム語、アラビア語、オランダ語、デンマーク語

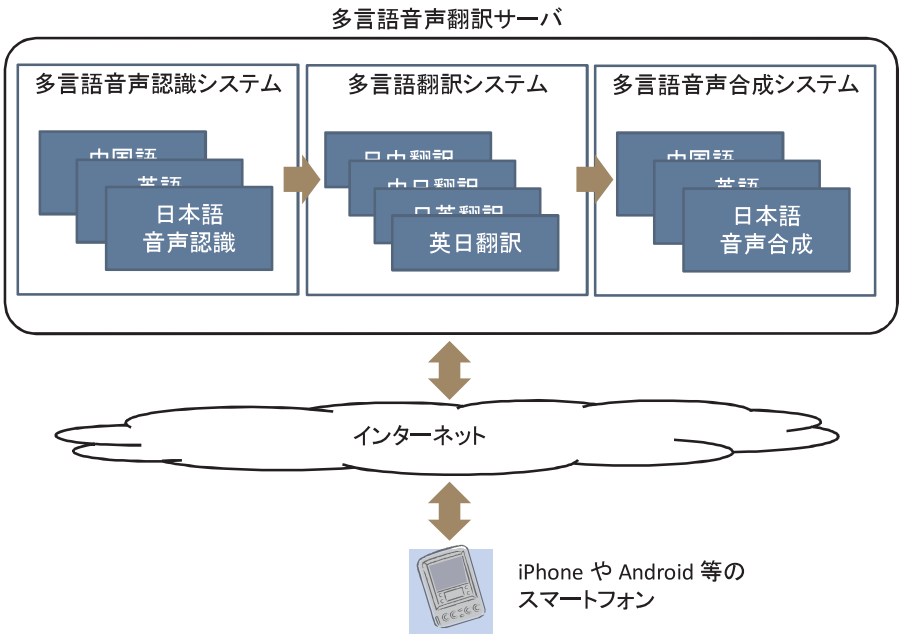

図 2　VoiceTra のシステム構成

$P(O)$ は定数であり、arg max の計算では考慮する必要はない。統計的音声認識で用いられるモデルは、大量の音声や、大量のテキストコーパスから推定される。

VoiceTra サービス開始時における音声認識システムでは、日本語の場合、成人 4,200 名、高齢者 300 名による旅行会話文の読み上げ音声約 400 時間、及び、音声翻訳の日本全国 5 地域での実証実験で収集された音声のうち、人手で書き起こした約 6 万文を用いて音響モデルの推定を行った。この全国 5 地域での実証実験では、旅館やホテル、イベント会場において、旅行者に音声翻訳システムを貸し出し、実際にシステムを利用した時の音声を収集しており、読み上げ音声だけでは観測されない多様な発話スタイルを含んでいる。

VoiceTra は屋内だけでなく屋外の騒音環境での利用を想定している。雑音に対する頑健性改善のため、フロントエンド処理として観測された音声からウィーナーフィルタを用いた雑音抑圧手法 [3] の適用及び、バックエンド処理として車の走行音や街路、駅コンコースなど様々な場所で収録した雑音を、学習データに重畳して音響モデルの推定を行った。

言語モデルは、旅行会話文章 6.1 M 単語及び、音響モデルと同様に、全国 5 地域での実証実験で得られた書き起こしテキストを用いて推定した。

サービス開始後は、図 3 に示すように実データが日々増加したため、これら大量の音声データを用いた教師無し適応を行うことにより、音響モデ

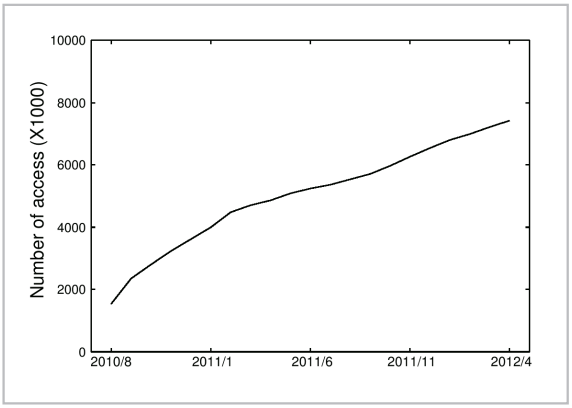

図3 VoiceTra の累計アクセス数

ル、言語モデル両方の性能改善を試みた。教師無し適応とは、通常音響モデルや言語モデルで必要な書き起こしテキストを準備することなくモデル適応を行う手法である。個々の発話の認識結果の信頼度を計算し、信頼度の高い文章や単語を用いてモデル適応が行われる。

4 言語翻訳

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。

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

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、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。

$$\hat{e}(e, \lambda_1^M) = \arg\max_e \sum_{i=1}^{M} \lambda_i h_i(e, f) \quad (2)$$

表2 音声翻訳システムの評価結果

システム	評価結果			
	S	S, A	S, A, B	S, A, B, C
サービス開始時	24%	32%	39%	45%
システムアップデート後	33%	44%	52%	56%

学習データとしては、主に基本旅行会話表現コーパス（BTEC）を用いた。また、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。

5 評価実験

表2に音声翻訳システムの評価結果を示す。評価においては、VoiceTra 実利用データの中から676文をランダムサンプリングし、これをテストセットとした。評価方法は、バイリンガルの評価者による5段階主観評価（S（Perfect）、A（Correct）、B（Fair）、C（Acceptable）、D（Nonsense））である。

表2では、VoiceTra サービス開始時の性能と、システムアップデート後の性能を示している。システムアップデートでは、VoiceTra 実利用データを用いて音声認識システムと機械翻訳システムの再学習を行っている。表2に示す通り VoiceTra の実利用データを用いることにより、テストセットの10%以上に対して、音声翻訳システムの性能が改善されていることが分かる。

6 むすび

2010年8月より公開しているスマートフォン向けネットワーク型多言語音声翻訳アプリケーション VoiceTra の概要について述べた。システム構成や、音声翻訳システムを構成する要素技術（音声認識システム、機械翻訳システム）について説明した。今後は、旅行会話だけでなくビジネス会話への適用や、過去の履歴を用いた音声翻訳、さらに同時通訳への応用について研究開発して行く予定である。

参考文献

1 L. R. Rabiner et al.,“An Introduction to Hidden Markov Models,”IEEE Transactions on Acoustic Speech, Signal Processing, Vol. 3, No. 1, pp. 4–16, 1986.
2 L. R. Bahl et al.,“A maximum likelihood approach to continuous speech recognition,”IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 179–190, 1983.
3 M. Fujimoto et al.,“A Non-stationary Noise Suppression Method Based on Particle Filtering and Polyak Averaging,”IEICE Transactions on Information and Systems, Vol. E89-D, No. 11, pp. 2783–2793, 2006.
4 P. Koehn, F. J. Och, and D. Marcu,“Statistical Phrase-Based Translation,”Proc. of HumanLanguage 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.

（平成 24 年 6 月 14 日 採録）

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

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

河井　恒（かわいひさし）
株式会社 KDDI 研究所主幹研究員／
元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
上席研究員
工学博士
音声情報処理、音声翻訳
hi-kawai@kddilabs.jp