社団法人 情報処理学会 研究報告
IPSJ SIG Technical Reports

2006－SLP－64（27）
2006／12／21

多言語音声コミュニケーションプラットホームと音声翻訳への応用

清水 徹†‡ 葦苅 豊†‡ 竹澤 寿幸†‡

†独立行政法人 情報通信研究機構 知識創成コミュニケーション研究センター
‡ATR 音声言語コミュニケーション研究所

†‡〒619-0288 京都府「けいはんな学研都市」光台 2-2-2
E-mail: †‡{tohru.shimizu, yutaka.ashikari, toshiyuki.takezawa}@{nict.go.jp, atr.jp}

あらまし 音声言語によるコミュニケーション支援を目的とした実験システムを効率的かつ短期間で構築するための仕組みとして、モジュール間のデータフロー制御情報を書き換えるだけでシステム構成を容易に変更することが可能な、多言語音声コミュニケーションプラットホームを構築した。本プラットホームは、アプリケーション構築の際のカスタマイズにかかる工数削減のみならず、無線LANや携帯電話等のデータ通信網などネットワークインフラの使用、多言語の同時使用、音声や画像データなど大量のバイナリデータのリアルタイム処理など、音声コミュニケーションシステムの開発に必要な機能を実装している。本稿では、効率的なシステム構築のためのプラットホームの概要ならびに同プラットホーム上に構築した音声翻訳システムとその性能評価結果について述べる。

キーワード 音声言語アプリケーション、システム構築、コーパスベース多言語音声翻訳

Multi-lingual speech communication platform and development of multi-lingual speech translation system on the platform

Tohru SHIMIZU†‡ Yutaka ASHIKARI†‡ and Toshiyuki Takezawa†‡

† Knowledge Creating Communication Research Center, National Institute of Information and Communication Technology
‡ ATR Spoken Language Communication Research Labs.

†‡ 2-2-2 Hikaridai, Keihannna Science City, 619-0288 Japan
E-mail: †‡{tohru.shimizu, yutaka.ashikari, toshiyuki.takezawa}@{nict.go.jp, atr.jp}

Abstract This paper describes a multi-lingual spoken language communication platform. This platform enables easy assembly of speech communication experimental system from the corresponding software modules by simply modifying data flow definition (data flow definition between corresponding modules). This platform enables decreasing of system development cost, easy handling wireless communication infrastructure (e.g. wireless LAN, 3G mobile phone), simultaneous multi-language processing, real-time processing of speech / image data, that are necessary for developing speech communication system. We have developed a speech-to-speech translation system on this multi-lingual speech communication platform. Recent evaluation of the overall system is also described in this paper.

Keyword spoken language application, system development, corpus-based multi-lingual speech-to-speech translation system

1. まえがき

筆者らは音声言語によるコミュニケーション支援を目的とした研究開発を行うための実験システムを開発している[1,2]。良く知られているように、音声言語によるコミュニケーション支援アプリケーションとしては、音声から音声への翻訳、音声による情報検索、議事音声の書き起こしなど様々なものがあるが、一般的にこれらのアプリケーションは、音声と言語に関する様々な要素技術から構成されており、これらのシステムをアプリケーション毎に独立して開発・性能評価実験を行うのは、非常に効率が悪い。そこで、要素技術を容易に組み合わせ、カスタマイズを効率的かつ短期

間で実現する仕組みが求められている。

例えば、音声認識、言語翻訳、音声合成の３つのモジュールから構成される既存の音声翻訳システムに、文脈処理や音声検索を新たなモジュールとして追加する場合に、各モジュールの通信制御機能を特定のアプリケーション毎に構築することなく、モジュール間のデータフロー制御情報を書き換えるだけでシステム構成を容易に変更できればシステム構築にかかる工数を大幅に削減できることが期待できる。

また、多言語音声コミュニケーションプラットホームに要求される機能としては、無線LANや携帯電話等のデータ通信網などのネットワークインフラを使用することができる、多言語の同時使用が可能である、音声や画像データなどの大量のバイナリデータをリアルタイムに扱えるなどが挙げられる。

本稿では、効率的なシステム構築のためのプラットホームの概要ならびに同プラットホーム上に構築した日英中多言語音声翻訳システムとその性能評価結果について述べる。

2. 多言語音声コミュニケーションプラットホームの構成

プラットホームは図１に示す４つの構成要素からなる。

モジュール(Module)：

音声認識、言語翻訳、音声合成、ユーザインタフェースなどのプログラムである。各モジュールは、モジュールマネージャとのみ接続している。

フロー定義(Data flow definition)：

ユーザがプラットホーム上のデータフローを柔軟に定義することを目的とするものである。モジュール間のデータの流れ等を定義したXML形式のファイルで実行時に読み込まれる。入力モジュール群(from)、出力モジュール群(to)、メッセージ種別(event)、配信種別(dispatch)、の集合からなる。配信種別としては、空いているモジュールへの配信(例：サーバリクエスト)、同一のモジュールへの配信(例:文脈処理の一貫性確保)、全てのモジュールへの配信(例：処理の取消)などの定義が可能である。

フロー定義の記述例

```
<arc>
        <from node_name="PDAja2en"/>
        <to node_name="SRja" />
        <event type="SR_IN" />
        <dispatch type="ready" />
</arc>
..
..
```

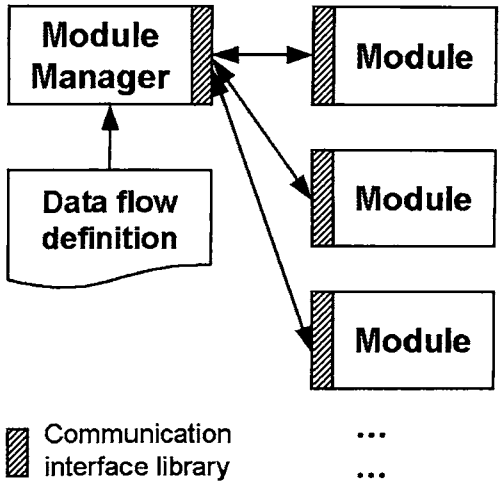


図1. 多言語音声コミュニケーションプラットホームの構成

モジュールマネージャ(Module Manager)：

各モジュールから出力されるメッセージをフロー定義ファイルに従ってモジュールに配信する。配信先は送受信データの種別と送信元の情報を基に決定される。

通信ライブラリ(Communication interface library)：

通信ライブラリは、モジュール間のメッセージ通信に関わる部分を特定のアプリケーションに依存しない汎用的な仕組みを提供するもので、モジュールマネージャと全モジュール間で共通の通信制御を行う。通信ライブラリには、メッセージパケットの生成と解析、メッセージの送受信、送受信エラー回復処理、バイナリデータ等大量データの分割送受信、各種ログファイル出力などのサービスが含まれている。TCP/IPで接続された計算機であれば、各モジュールを分散して配置することが可能であり、アプリケーションを構成する各モジュールは、通信環境(無線LAN、携帯電話のデータ通信等)を意識する必要がない。多言語を同時に扱うため文字コードはUTF-8に統一されている。

これらの機能の実装にあたっては、多言語かつ複数話者(マルチユーザ)環境において、

- モジュールの処理負荷によらず、言語・話者・セッションの一貫性が保たれ、時系列順に正しく処理が行われること
- 耐雑音対策に用いるマイクロフォンアレイの多チャンネル音声情報の送受信が可能であること

・ 同時入力に対するスループット向上
・ ユーザ意思による既発話の処理の取消し
・ モジュールがビジーの場合のデータのバッファリング

が配慮されている。

3. 音声翻訳システムへの応用

3.1. システム構成

本プラットホームを用いて、

・ 全機能を1台の携帯型ＰＣに収容した一体型
・ サーバと端末に処理を分散したサーバクライアント型

の２つの携帯型多言語音声翻訳(STST)システムを試作した。図２にクライアントサーバ型の構成例を示す。

システム全体は、音声翻訳全体を制御するモジュールマネージャを中心に、日英中の言語間の双方向音声翻訳に必要な音声認識(ASR)、翻訳(MT)、音声合成(SS)の各モジュール群、ユーザインタフェース(UI)モジュールを結んだ構成である。メッセージフロー定義は、ユーザインタフェース(UI)から入力された音声信号を、音声認識(ASR)が単語列に変換、UI を経由して言語翻訳(MT)がターゲット言語の単語列に変換、UI を経由して音声合成(SS)が合成音に変換するように記述されている。

なお、携帯型の音声翻訳システムは屋外での利用が前提であることから、音声認識モジュール(ASR)には、雑音環境下におけるユーザ音声の高品質収音ならびにパーティクルフィルタと GMM による雑音抑圧前処理法[3]が実装されている。

3.2. サーバクライアント型における分散処理

サーバクライアント型の音声翻訳システムでは、端末の処理能力と無線ネットワークの帯域・遅延などを考慮した柔軟な構成としている。サーバと端末は既存の無線インフラを利用して接続されており、利用者は無線のサービスエリア内であれば自由に移動して音声翻訳機能を利用することができる。

広帯域のネットワークが利用可能な場合：

サンプリング周波数 16kHz、量子化ビット数 16 ビットの PCM(Pulse Code Modulation)でデジタル化された音声をそのまま高速な無線インフラ（現在は IEEE802.11b）を利用する。

種々の環境雑音の中からユーザの音声を選択的に集音、追従するために入力デバイスとしてマイクロフォンアレイを用いる場合には、マルチチャネル音声信号を PCM 形式のままサーバに転送し、適応信号処理に基づく指向性制御と雑音抑圧処理[4]の後、音声認識、言語翻訳、音声合成はサーバで処理される。

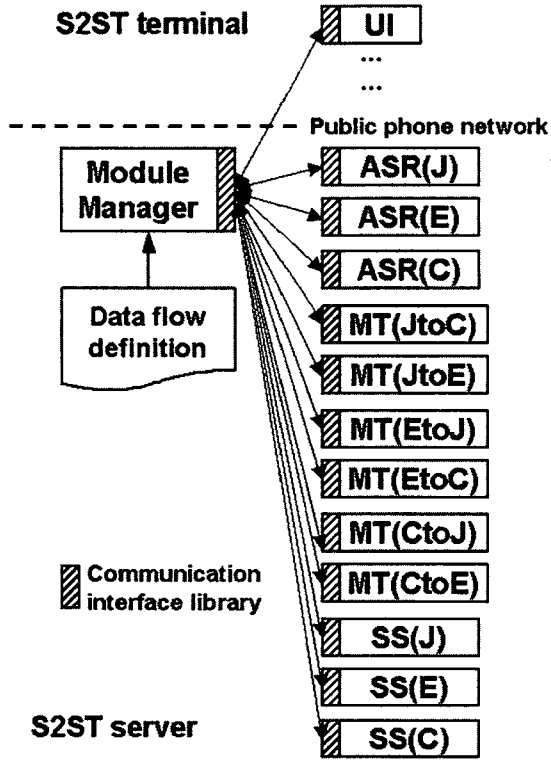


図 2. 多言語音声コミュニケーションプラットホームの適用例
(日英中音声翻訳、クライアントサーバ型)

狭帯域のネットワークを利用する場合：

ADPCM(Adaptive Differential PCM)(64kbps)で圧縮、あるいは分散型音声認識のフロントエンド処理によって圧縮・雑音除去を行ない[5,6]、PHS あるいは第３世代携帯の回線交換方式（帯域保証）またはパケット交換方式を用いて端末とサーバを接続する[1]。

種々の環境雑音の中からユーザの音声を選択的に集音、追従するために入力デバイスとしてマイクロフォンアレイを用いる場合には、端末で簡単な指向性制御（Fixed beam former）を行ってマルチチャネル音声信号からユーザの声を選択的に集音した１チャンネルの音声信号に変換した後、上述の圧縮処理を行ない、サーバでは、雑音抑圧、音声認識、言語翻訳、音声合成を行う。なお、音声合成の出力音声の通信についても、音声認識への入力と同様に ADPCM 形式に対応している[1]。

3.3. 各モジュールの特徴

音声認識、言語翻訳、音声合成の3つのモジュールは、すべてコーパスベースの統計的な手法により構築されている[7]。コーパスベースのアプローチを採用することのメリットは、新しい別の言語に拡張したり、応用分野を変更したりすることが、コーパスの追加と適応技術により、容易に可能な点にある。

3.3.1. 音声認識(ASR)

端末に8つのマイクロフォン(横方向2cm間隔に5つ、縦方向4cm間隔に4つ、そのうち1つは両方向で共用)を装着したマイクロフォンアレイユニット(W93mm x D33mm x H132mm)の外観写真を図3に示す。

音声認識エンジン(ATRASR)は、学習データ量に応じて最適な状態数を割り当てる最小記述長による隠れ状態網音響モデル[8]と、前後の文脈を分離し前方文脈と後方文脈とを別々に考慮する複合N-gram、言語モデルの単位を可変長にする可変長N-gramを用いている[9]。上記音響モデルのパラメータは、現在、日本語話者4000人、英語、中国語各約500人の音声を地方のアクセントを考慮した形で収集したデータにより推定している。また、言語モデルは、旅行表現集や実旅行対話の書き起こしテキストなど(日本語約85万発話、英語約71万発話、中国語約50万発話)[10]を用いて学習を行っている。

また、音声認識結果の信頼性尺度として一般化単語事後確率(GWPP)に基づく信頼性尺度を発話単位に拡張したもの[11]をユーザへのフィードバックに用いているほか、認識結果に類似した文をコーパスから抽出し、ユーザに次候補を提示する機能を実装している。

3.3.2. 言語翻訳(MT)

言語翻訳モジュールの特徴は、アプローチの面では、大規模旅行会話コーパスから翻訳エンジンを自動構築した点にあり、フレーズベースの統計翻訳手法に基づいている。この手法は同じく統計をベースとする音声認識モジュールとの親和性が高いものとなっている。学習に用いた大規模旅行対話コーパスは、日英言語ペア約71万発話、日中言語ペアで約50万発話であり、音声認識モジュールの言語モデル学習コーパスと共有されている。

音声翻訳の核となる翻訳モデルは、log-linearモデルに基づいており、翻訳元言語から翻訳先言語への単語翻訳確率、翻訳先言語における単語入れ替え確率等を構成要素としている。また、翻訳モデルとともに用いられる言語モデルでは、バイリンガル文クラスタリングに基づく逐次タスク適応を行っている[12]。

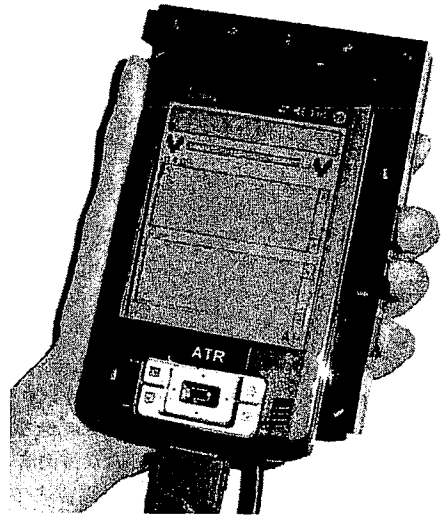


図3. マイクロフォンアレイを用いた端末

3.3.3. 音声合成(SS)

音声合成モジュール XIMERA[13]は、大規模コーパス(日本語男性 110 時間、日本語女性 60 時間、中国語女性 20 時間)から構築されている。コーパスを大規模化することで、自然性が保持されやすく、音声の個人性などが維持される。また、音声認識でもよく用いられる HMM (Hidden Markov Model) を用いた韻律 (イントネーションなど) パラメータのモデル化および生成[14]、知覚実験に基づく素片選択コスト関数の最適化[15]によって、自然性の高いパターン生成、人間が自然と感じる音声により近い合成音を得ている。

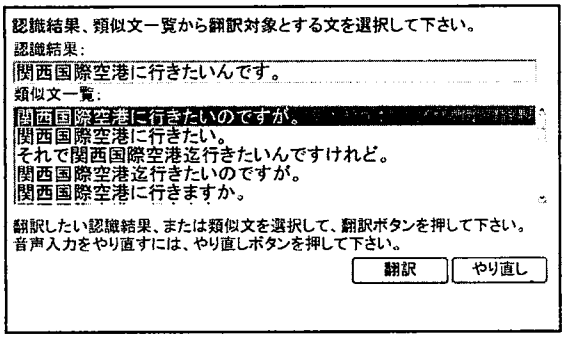
3.4. ユーザインタフェース(UI)

音声入出力およびグラフィカルユーザインタフェース(GUI)プログラムは端末に実装されている(図4)。全機能を1台の携帯型PCに収容した一体型もサーバと端末に処理を分散したサーバクライアント型も同一のユーザインタフェースを備えている。

音声認識結果・言語翻訳結果を表示する画面は大きく2つに分かれる。上が音声認識結果およびシステムメッセージ、下が言語翻訳結果である。また、この画面とは別に3.3.1節で述べた次候補提示画面を有している。

(a) 音声認識結果、言語翻訳結果表示画面

(b)



(b) 次候補提示画面

図 4. 音声翻訳の画面例

3. 5. 音声翻訳の性能評価

3.5.1. 評価データ

評価用データは表 1 に示す 3 種類のものからなる。BTEC (Basic Travel Expression Corpus)は旅行会話基本表現コーパスの読み上げ音声である。MAD (Machine Aided Dialogues)は音声自動翻訳システムを介して日本語話者と英語話者が実施した課題遂行対話データであり，音声認識システムの代わりにタイピストが発話を書き起こし，機械翻訳システムに入力する形態で集めたものである。FED (Field Experiment Data)は関西国際空港で実施したモニタ実験から選んだデータである[16-19]。平均発話長は BTEC と FED が同程度で MAD が長い。パープレキシティは BTEC, MAD, FED の順に大きくなっている。

3.5.2. 音声認識性能

処理時間をリアルタイムファクタ(RTF)で表現したときに，RTF=5 となる条件での音声認識性能を表 2 に示す。対話(MAD,FED)の音声認識率は、読み上げの音声認識率(BTEC)に比較して低いが、文献[11]に示す手法により信頼性の低い発話を除いた場合の発話認識率はいずれの評価データでも 83%を越えた。

3.5.3. 言語翻訳性能

日本語の堪能な英語あるいは中国語ネイティブによる翻訳品質の主観評価を実施した。翻訳結果は以下の A から D の 4 つのランクに分類される。

A) 訳文だけでまったく問題なし。
B) 訳文は少し情報が欠けている。
C) 訳文はかなり情報が欠けている。
D) 訳文からは情報が想像もできない。

A, B, C ランクの割合を合計した翻訳率を表 3 に示す。

表 1. 評価に用いたデータ

	BTEC	MAD	FED
特徴	読み上げ	対話 オフィス	対話 空港
話者数	20	12	6
発話数	510	502	155
単語数	4,035	5,682	1,108
平均発話長	7.91	11.32	7.15
パープレキシティ	18.9	23.2	36.2

表 2. 音声認識性能

	BTEC	MAD	FED
%word accuracy			
日本語	94.9	92.9	91.0
英語	92.3	90.5	81.0
中国語	90.7	78.3	76.5
% utterance correct (日本語)			
全発話	82.4	62.2	69.0
リジェクト後	87.1	83.9	91.4

表 3. 言語翻訳性能

	BTEC
% correctly translated	
日本語から英語	92.5
日本語から中国語	88.4
英語から日本語	92.5
中国語から日本語	92.5

4. むすび

本稿では、音声言語によるコミュニケーション支援を目的とした実験システムを効率的かつ短期間で構築するための仕組みとして、モジュール間のデータフロー制御情報を書き換えるだけでシステム構成を容易に変更することができる、多言語音声コミュニケーションプラットホームについて述べた。また、同プラットホーム上に音声翻訳システムを試作し、音声翻訳システムの構成、各モジュールの特徴、性能について述べた。

本プラットホームは、システムのカスタマイズが効率的に行えるだけでなく、音声コミュニケーションに必要とされる機能として、無線LANや携帯電話等のデータ通信網などのネットワークインフラを使用することができる、多言語の同時使用が可能である、音声や画像データなどの大量のバイナリデータをリアルタイムに扱うことができるなどの特徴を有する。

今後は、本プラットホームを、音声翻訳だけでなく、音声対話システム、マルチモーダルインタラクションシステム、音声検索システム等の試作に活用する予定である。

文　献

[1] Tohru Shimizu, Yutaka Ashikari, Toshiyuki Takezawa, Masahide Mizushima, Genichiro Kikui, Yutaka Sasaki, and Satoshi Nakamura, "Developing client-server speech translation platform," In Proc. of TAMC(Workshop on tools and applications on mobile contents), 2006.

[2] 葦苅豊, 木村法幸, 清水徹, "携帯型多言語音声コミュニケーションプラットホーム," 日本音響学会秋季研究発表会講演論文集, 1-2-22, pp.43-44,2006.

[3] Masakiyo Fujimoto and Satoshi Nakamura, "A Non-stationary Noise Suppression Method Based on Particle Filtering and Polyak Averaging," IEICE Transactions on Information and Systems, Vol.J89-ED, No.3, pp.922-930, March 2006.

[4] W. Herbordt, T. Horiuchi, M. Fujimoto, T. Jitsuhiro, and S. Nakamura, "Hands-free speech recognition and communication on PDA using microphone array technology," In Proc. of ASRU, pp. 302–307, 2005.

[5] ETSI standard document. Speech processing, transmission and quality aspects (STQ); distributed speech recognition; front-end feature extraction algorithm; compression algorithm. ETSI ES 201 108, V1.1.2, 2000.

[6] ETSI standard document. Speech processing, transmission and quality aspects(STQ); distributed speech recognition; advanced front-end feature extraction algorithm; compression algorithm. ETSI ES 202 050, V1.1.1, 2002.

[7] S. Nakamura, K. Markov, H. Nakaiwa, G. Kikui, H. Kawai, T. Jitsuhiro, J. Zhang, H. Yamamoto, E. Sumita, and S. Yamamoto, "The ATR multilingual speech-to-speech translation system," IEEE Trans. on Audio, Speech, and Language Processing, 14, No.2:365–376, 2006.

[8] T. Jitsuhiro, T. Matsui, and S. Nakamura. Automatic generation of non-uniform context-dependent HMM topologies based on the MDL criterion. In Proc. of Eurospeech, pp. 2721–2724, 2003.

[9] H. Yamamoto, S. Isogai, and Y. Sagisaka. Multi-class composite N-gram language model. Speech Communication, 41:369–379, 2003.

[10] G. Kikui, E. Sumita, T. Takezawa, and S. Yamamoto. Creating corpora for speech-to-speech translation. In Proc. Of Eurospeech, pp. 381–384, 2003.

[11] Toshiyuki Takezawa, Tohru Shimizu, "Performance Improvement of Dialog Speech Translation by Rejecting Unreliable Utterances," In Proc. of ICSLP 2006, pp.1169-1172, 2006.

[12] Ruiqiang Zhang, Hirofumi Yamamoto, Michael Paul, Hideo Okuma, Keiji Yasuda, Yves Lepage, Etienne Denoual, Daichi Mochihashi, Andrew Finch,Eiichiro Sumita, "The NiCT-ATR Statistical Machine Translation System for the IWSLT 2006 Evaluation," submitted to IWSLT, 2006.

[13] H. Kawai, T. Toda, J. Ni, and M. Tsuzaki. XIMERA: A new TTS from ATR based on corpus-based technologies. In Proc. of 5th ISCA Speech Synthesis Workshop, 2004.

[14] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura. Speech parameter generation algorithms for HMM-based speech synthesis. In Proc. of ICASSP, pp. 1215–1218, 2000.

[15] T. Toda, H. Kawai, and M. Tsuzaki. Optimizing sub-cost functions for segment selection based on perceptual evaluation in concatenative speech synthesis. In Proc. of ICASSP, pp. 657–660, 2004.

[16] T. Takezawa and G. Kikui. Collecting machine –translation-aided bilingual dialogs for corpus-based speech translation. In Proc. of Eurospeech, pp. 2757–2760, 2003.

[17] G. Kikui, E. Sumita, T. Takezawa, and S. Yamamoto. Creating corpora for speech-to-speech translation. In Proc. Of Eurospeech, pp. 381–384, 2003.

[18] T. Takezawa and G. Kikui. A comparative study on human communication behaviors and linguistic characteristics for speech-to-speech translation. In Proc. of LREC, pp. 1589–1592, 2004.

[19] 菊井玄一郎, 竹澤寿幸, 水島昌英, 山本誠一, 佐々木裕, 河井恒, 中村哲, "音声対話翻訳システムの実環境におけるモニタ実験," 日本音響学会秋季研究発表会、1-7-10, 2005.