講演音声翻訳のための多言語音声合成技術に関する研究開発（091706004）

Research and Development of Multilingual Speech Synthesis
for Lecture Speech-to-Speech Translation

研究代表者
徳田恵一　名古屋工業大学
Keiichi Tokuda　Nagoya Institute of Technology

研究分担者
李晃伸[†]　南角吉彦[†]　河井恒[††]　倪晋富[††]　志賀芳則[††]　津崎実[†††]
Akinobu Lee[†]　Yoshihiko Nankaku[†]　Hisashi Kawai[††]　Jinfu Ni[††]　Yoshinori Shiga[††]
Minoru Tsuzaki[†††]
[†]名古屋工業大学　[††]情報通信研究機構　[†††]京都市立芸術大学
[†]Nagoya Institute of Technology
[††]National Institute of Information and Communications Technology
[†††]Kyoto City University of Arts

研究期間　平成 21 年度～平成 23 年度

概要
講演、演説等、モノローグの同時翻訳システム実現のために必須となる多言語音声合成技術の研究開発を行った。音声合成方式としては、多様な話者の声を容易に実現可能、多言語化が容易、携帯デバイスでも実現が容易、等の利点があることから、近年、次世代の音声合成方式として注目を集める HMM 音声合成方式を利用した。本研究開発により、新たに「元話者と異なる様々な言語の合成音声を、元話者の声のまま出力する」ための技術基盤が確立された。

Abstract
We have developed multilingual speech synthesis techniques for realizing speech-to-speech translation systems for lectures. We employed as the basis of the development the HMM-based speech synthesis method that recently receives much attention for its advantages over the conventional methods including 1) voice characteristics of synthesized speech can be easily controlled, 2) it can work on a language independent framework, and 3) it can generate smooth and stable speech under a small footprint. The developed techniques allow the translation system to output speech in the target language with the same speech individuality as the speaker of the source language.

１．まえがき
音声翻訳技術は目覚ましい進歩を遂げつつあり、近年は、旅行対話等の限られたタスクドメインに関しては、ほぼ実時間で動作可能な音声翻訳システムが構築されている。しかしながら、翻訳結果として出力される合成音声は、予め定められた話者の声でしか出力することはできないという制限があった。本研究開発課題では、講演や演説等のモノローグを対象とした音声翻訳の実現に向けて、図 1 に示すような「元話者が話さない様々な言語の合成音声を、元話者の声のまま出力する」ための技術の研究開発を行った。本研究開発では、多様な話者の声を容易に実現可能、多言語化が容易、携帯デバイスでも実現が容易、等の利点により、近年、次世代の方式として注目を集めている HMM 音声合成方式を基盤として採用した。

２．研究内容及び成果
２．１．バイリンガル音声データベースの構築
バイリンガル音声データベースを構築した。収録した音声は、話者数 42、合計発話数 27,189、発話時間は合計 42 時間であり、全発話について発声誤りの検査、発話切り出し、ポーズ挿入位置の記録が行われている。発話内容は、音素バランス文、合文法無意味文、米国大統領演説の 3 種類からなる。これらのデータは研究・開発用途に限定して、情報通信研究機構(NICT)が運営する高度言語情報融合フォーラム(ALAGIN)の会員を対象として配布を開始した。

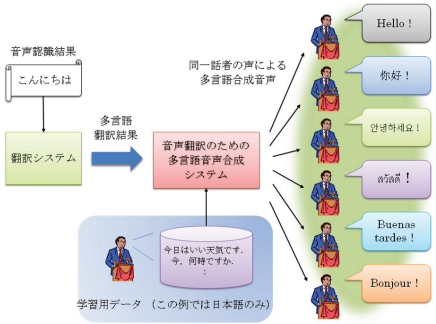

図 1. 音声翻訳のための多言語音声合成システム

２．２．言語間話者適応
入力言語の話者性を出力音声に反映させるために、状態マッピングに基づく言語間話者適応手法を開発した。実用的なシステムへの応用において必要となる教師無し適応の評価を行ったところ、教師有り適応と同等の話者性を再現できることが示された（図 2）。

しかし、状態マッピングに基づく手法では、音響的特徴に含まれる話者性だけでなく、言語性までも変換してしまうという問題がある。そこで、バイリンガル音声データから学習した固有声に基づく言語間話者適応を新たに開発した。評価の結果、固有声に基づく手法(FA, FAS)により状態マッピングに基づく手法(SM)と比較して自然性が大きく改善され（図 3）、話者性についても同程度以上の性

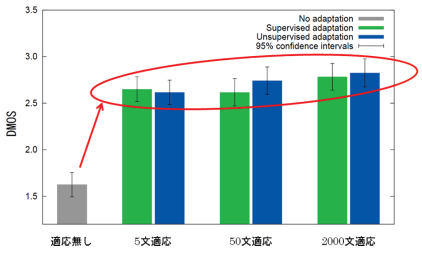

図 2. 状態マッピングに基づく言語間話者適応の評価結果

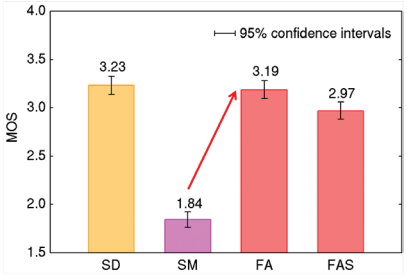

図 3. 固有声に基づく言語間話者適応の評価結果

能が示された。これより、言語性と話者性を分離したことによって、韻律的特徴などの言語性を損なわずに合成音声を生成することが可能となった。

２．３．音声翻訳システムの試作

2.2 節で示した固有声に基づく言語間話者適応手法の近似として音声合成用モデル(voicefont)を自動選択するネットワーク型音声翻訳システムを試作した。サーバーでは入力音声と同等の話者性を有すると推定される翻訳先言語の voicefont を選択し、これを用いて音声を合成し、Android 2.2 上で動作するクライアントアプリケーション（図 4）に送信する。

２．４．主観評価法の確立と評価

システムが合成する元話者の話者性を反映した音声は元話者が話さない異言語となるため、聴取者は通常の再認方略が使えない。従って、元話者の声に聞こえるかどうかという安直な印象を答えることは困難であった。そこで、2 話者の自然音声を対提示し、その一方を模擬した合成音声を聞かせていずれを模擬したと思うかを選択させる実験方法を開発し、聴取実験を行った。その際に 3 種類の話者モデルの選択法について比較した結果、いずれの手法もチャンス・レベル以上の正答率となり、提案する合成手法により話者性を伝達可能であることが示された。

２．５．研究基盤ソフトウェアの整備

以下の研究基盤ソフトウェアの新機能の開発を行い、一般に公開した。

- 大語彙連続音声認識エンジン Julius
- 音声信号処理ツールキット SPTK
- HMM 音声合成ツールキット HTS
- ランタイム用音声合成エンジン hts_engine API
- 英語テキスト音声合成システム flite+hts_engine
- 日本語テキスト音声合成システム Open JTalk

これらのソフトウェアは国内外の多くの研究機関や企業において利用されており、当該研究分野に特筆すべき大きな貢献があったといえる。

３．むすび

本研究開発課題では、講演や演説等のモノローグを対象とした音声翻訳の実現に向けて、「元話者が話さない様々な言語の合成音声を、元話者の声のまま出力する」ための技術を開発した。同時に、バイリンガル音声データベースの構築、元話者の声をよりよく再現するための合成音声の自然性の向上、主観評価法の確立と評価を行った。

図 4. 試作した音声翻訳クライアントアプリケーション

本成果は、国内外でその成果が認められ、59 件の査読付き論文、72 件の口頭発表、7 件の受賞を得た。また、研究基盤ソフトウェアをオープンソースとして多数公開する、6 件の報道発表を行う、関連ワークショップを 5 件主催する等、研究成果普及に関しても十分な活動を行った。特にオープンソースソフトウェアは NTT ドコモの携帯電話 20 機種以上に採用される等、情報通信産業界へ特筆すべき大きな貢献があった。人材育成の面に関しても、学部生 25 名、修士課程学生 30 名、博士課程学生 7 名が修了及び学位取得に至り、大きな貢献があったといえる。

【誌上発表リスト】

[1] K. Hashimoto, J. Yamagishi, W. Byrne, S. King, K. Tokuda, “Impacts of machine translation and speech synthesis on speech-to-speech translation,” Speech Communication, vol.54, Issue 7, pp.857-866 (2012 年 9 月)

[2] K. Oura, J. Yamagishi, M. Wester, S. King, K. Tokuda, “Analysis of unsupervised cross-lingual speaker adaptation for HMM-based speech synthesis using KLD-based transform mapping,” Speech Communication, vol.54, Issue 6, pp.703-714 （2012 年 7 月）

[3] M. Tsuzaki, K. Tokuda, H. Kawai, J. Ni, “Estimation of perceptual spaces for speaker identities based on the cross-lingual discrimination task,” Interspeech 2011, pp.157-160 （2011 年 8 月）

【受賞リスト】

[1] 大浦圭一郎、情報処理学会 2011 年度山下記念研究賞 “Sinsy:「あの人に歌ってほしい」をかなえる HMM 歌声合成システム”、2012 年 3 月 7 日

[2] 彭湘琳、日本音響学会東海支部優秀発表賞、“バイリンガルデータを用いた固有声手法に基づくクロスリンガル話者適応”、2011 年 12 月 21 日

[3] 鹿住恭介、日本音響学会第 3 回学生優秀発表賞、“多様な声質を表現するための因子分析に基づく HMM 音声合成”、2011 年 9 月 21 日

【報道発表リスト】

[1] “Singing Synthesizers: The Technology Behind a Digital Popstar”、NHK WORLD TV、 Science View、2012 年 3 月 11 日

[2] “思いが伝わる声をつくれ”、NHK 総合、 クローズアップ現代、2012 年 2 月 28 日

[3] “最新研究！しゃべる・歌うコンピューター”、NHK ラジオ第一放送、渋マガ Z、2010 年 2 月 12 日

【本研究開発課題を掲載したホームページ】

http://www.sp.nitech.ac.jp/project/scope/