3.5.1　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

室長　　河井　恒　ほか 16 名

グローバルコミュニケーション計画の実現に向けた音声認識、音声合成、音声対話の研究

【概　要】

本研究室では、人間にとって自然で簡便な情報伝達手段である音声によるコミュニケーションを用いた音声対話・音声翻訳システムの実現に向けて、音声認識、音声合成、対話処理の研究開発を行っている。平成 26 年度は、音声翻訳に関する国際連携組織 U-STAR（Universal Speech Translation Advanced Research consortium）の枠組みを利用したアジア系言語の音声認識性能改善、災害時等の高雑音環境下でも合成音声を聞き取りやすくする手法の研究及び音声対話エンジンの多言語化を行った。また、平成 26 年 4 月に総務省のグローバルコミュニケーション（GC）計画が開始されたことを受け、大規模な多言語音声コーパスの収集に着手するとともに、日英中韓の基本 4 言語の音声認識性能を強化した。今年度の特筆すべき成果として、長文音声認識の研究に関して評価型国際ワークショップ IWSLT（International Workshop on Spoken Language Translation）で昨年度に引き続き第 1 位となり、三冠を達成した。

【平成 26 年度の成果】

● U-STAR による国際連携とアジア系言語の音声認識・合成

U-STAR は、国際連携による多言語音声翻訳技術の研究の促進を目的とする組織であり、NICT は設立時から事務局として強力にサポートしてきた。U-STAR は、平成 27 年 3 月末現在、25ヵ国、30 機関の大規模な組織に成長し、音声翻訳国際共同実験システム VoiceTra4 U の運用、ワークショップ開催など活発に活動を続けている。

NICT では、連携先研究機関より研究員等を受け入れてロシア語、ミャンマー語、ネパール語の初期的な音声認識システムの試作、ベトナム語音声認識性能の改善を行った。特にベトナム語に関しては、発音書き起こしデータのクリーニング、声調を考慮した特徴量の導入、DNN（Deep Neural Network）に基づく特徴量抽出の導入等により、VoiceTra4U 利用ログ音声に対する単語誤り率が 61 % から 28 % へと大幅に改善された。また、ミャンマー語に関して、初期的な HMM（Hidden Markov Model）型音声合成システムを世界で初めて試作し、音節了解度 85 %、文了解度 70 % を得た。

●騒音下でも聞き取りやすい音声合成の研究

災害時の緊急放送等、騒音が大きく騒然とした環境下で合成音声を再生した場合でも、聞き取りやすく、意図を明確に伝えられるようにするために、合成音声のスペクトルと韻律の両面について研究を行った。

スペクトルに関しては、フォルマント（パワースペクトル包絡上のエネルギー集中部）のピークが雑音レベルに埋もれず、なおかつ補正特性の時間変動が滑らかになるような適応的フィルタを自動設計するアルゴリズムを提案した。合成音の明瞭性を先行研究と比較評価したところ、既存のどの手法よりも優れていることが確認された。

メリハリのよい音声では、韻律の物理的特徴量の 1 つである F0（声帯振動の基本周波数）パターン上で、話題の焦点が置かれた部分の起伏が顕著である。NICT が採用している HMM 音声合成では、大量のメリハリのよい音声を用いて音響モデルを作成すれば、メリハリのよい合成音声を得られる。しかしながら、そのような音声を大量に収録することは困難である。そこで、特に焦点を置かずに発声した音声に対して焦点を設定されたかのように F0 パターンを変形し、それを用いて音響モデルを学習する方法を開発した（図 1）。主観評価実験を行ったところ、自然性に基づく選好スコアが従来法と比べて 15.2 % から 62.5 % へ大幅に改善された。

●音声対話技術の研究

複数の言語が混在する音声対話をクロスリンガル音声対話と呼ぶ。平成 26 年度は、音声対話エンジン WFSTDM 及びシステム構築ツール DMBuilder を日英中 3 言語のクロスリンガル対話に対応させた。これらのソフトウェアは、『MCML 音声インタラクション SDK』として一般公開の予定である。

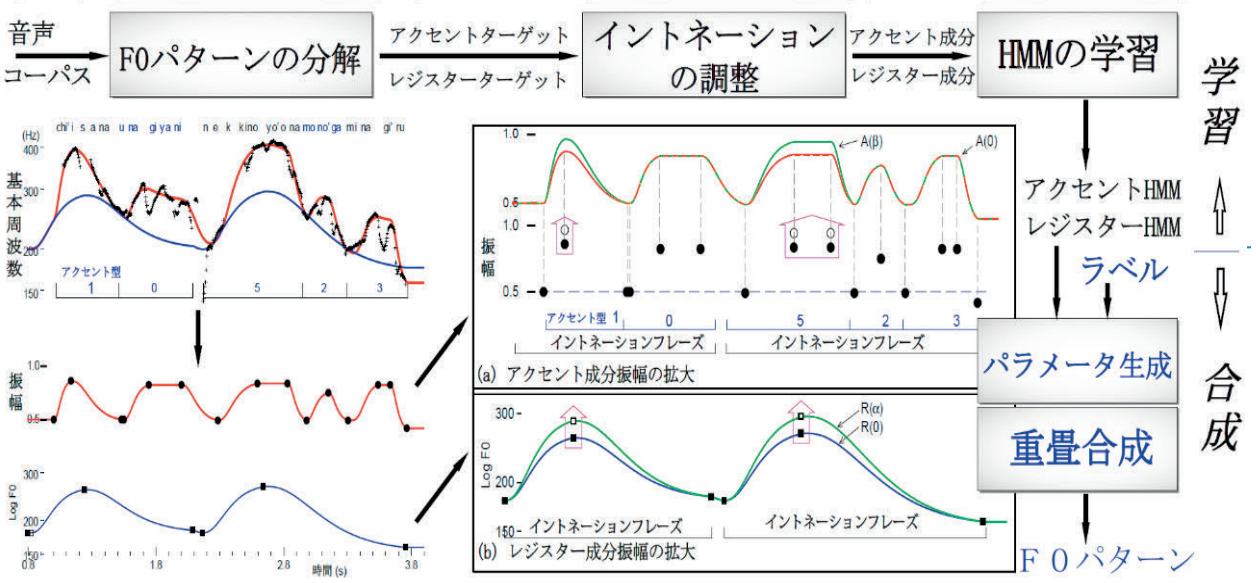
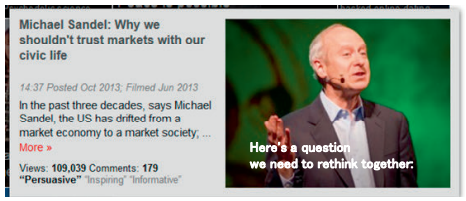


図１　基本周波数 (F0) パターンを操作して焦点を付与した音声データを用いてメリハリのある音声を合成可能な音響モデルを学習する手法

●評価型国際ワークショップ IWSLT で 3 年連続首位獲得

長文の音声認識性能の改善に関して、複数種類の DNN 音響モデルによるアンサンブル型認識技術と複数回話者適応技術を組み合わせる手法を考案した。評価型国際ワークショップ IWSLT において、英語講演 TED の音声認識タスクに対してこの手法を適用したところ、単語誤り率 8.4 % という他を圧倒する性能を達成し、3 年連続の 1 位を獲得した (図 2)。

	2014	2013	2012
NICT	8.4	13.5	12.1
EU-BRIDGE	9.8	-	-
MITLL-AFRL	9.9	15.9	-
KIT	11.4	14.4	12.7
FBK	11.4	23.2	16.8
LIUM	12.3	-	-
UEDIN	12.7	22.1	14.4
RWTH	-	16.0	13.6
NAIST	-	16.2	-
KIT-NAIST	-	-	12.4
MITLL	-	-	13.3



http://www.ted.com

AFRL　：空軍研究所（米）
KIT　：カールスルーエ工科大学（独）
FBK　：ブルーノ・ケスラー財団 研究所（伊）
LIUM　：ル・マン大学（仏）
UEDIN　：エディンバラ大学（英）
RWTH　：アーヘン工科大学（独）
NAIST　：奈良先端科学技術大学院大学（日）
MITLL　：マサチューセッツ工科大学リンカーン研究所(米)
EU-BRIDGE：RWTH, UEDIN, KIT, FBKの連合チーム

図２　評価型国際ワークショップ IWSLT の英語講演音声認識タスクで 3 年連続 1 位を達成

●グローバルコミュニケーション計画に向けた音声認識基盤技術の強化

平成 26 年 4 月に総務省が発表したグローバルコミュニケーション (GC) 計画に従い、多言語音声認識システムの性能向上を図るための基盤として、英、中、韓、タイ、インドネシア、ベトナム、スペイン、フランスの各言語について模擬会話音声の収集を開始し、平成 26 年度末の段階で合計 640 時間分を得た。平成 27 年度末には、合計 6,500 時間となる見込みである。

日英中韓の各言語の音響モデルを従来型の GMM (Gaussian Mixture Model) から DNN に置き換えたことにより、単語誤り率が日本語で 18.9 % から 13.4 % になる等、大幅に改善した。