Vol.2015-MUS-107 No.6
2015/5/23

表現豊かな音声の認識・合成と Affective Speech-to-Speech Translationへの応用

赤木 正人[1,a)]

概要：Speech to Speech Translation (S2ST) システムは，ある言語の発話された音声を他の言語の音声に変換するための重要な技術である。音声には言語情報のみならずパラ言語・非言語情報が含まれているが，これまでに提案された S2ST システムは主に言語情報を扱ってきており，変換された音声は，感情などのパラ言語・非言語情報が含まれていない合成音声に限定されている。このため S2ST システムで，本来含まれている感情などのパラ言語・非言語情報を扱えるようにするためには，これらを含んだ音声の認識／合成システムが必要となる。本講演では，我々の研究グループで取り組んでいる“affective S2ST”と呼ばれるシステムを紹介する。“Affective S2ST”は，S2ST システムにおいて話者の感情に特に焦点をあて，多言語間でのパラ言語・非言語情報の認識・変換・合成を目指すシステムである。講演では，(1) 音声中の感情をどのように表現し，音声生成・知覚のモデルを構築するか，(2) 多言語間での感情知覚の共通性・相違は何か，そして，(3) これらの議論にもとづいてどのようにシステムを構築するか，について話題を提供する。

Abstract: Speech-to-speech translation (S2ST) is the process by which a spoken utterance in one language is used to produce a spoken output in another language. The conventional approach to S2ST has focused on processing linguistic information only by directly translating the spoken utterance from the source language to the target language without taking into account paralinguistic and non-linguistic information such as the emotional states at play in the source language. This paper introduces activities of JAIST AIS lab1 that explore how to deal with para- and non-linguistic information among multiple languages, with a particular focus on speakers' emotional states, in S2ST applications called “affective S2ST”. In our efforts to construct an effective system, we discuss (1) how to describe emotions in speech and how to model the perception/production of emotions and (2) the commonality and differences among multiple languages in the proposed model. We then use these discussions as context for (3) an examination of our “affective S2ST” system in operation.

1. 講演の内容

1.1 グローバルでユニバーサルな音声コミュニケーション

近年，一層の国際化が進むにあたり，言語・民族・文化を越えた（グローバルな），また，言語・民族・文化のみならず老人，幼児，あるいは障害者との障壁のない（ユニバーサルな）コミュニケーションの重要性が増している。インターネットを含む通信手段の発達は，人と人のコミュニケーションにおいて距離の概念を変えた。たとえば，この発達は，我々に地球の裏側の人たちとも瞬時にコミュニケーションできる環境を与えることができた。しかしながら，直接的で最も有効な手段の一つであるはずの音声によるコミュニケーションでは，依然として，共通の言語を持たない限りお互いのコミュニケーションは可能ではない。これは，グローバルでユニバーサルなコミュニケーションを実現するために解かれなければならない重要な問題である。

1.2 Speech to Speech Translation (S2ST) システム

この問題に対する一つの解決案は，Speech to Speech Translation (S2ST) システムを構築することである。S2ST システムは，ある言語で発話された音声を他の言語の音声に変換して生成する技術であり，図 1 に示すように，三つの主要技術から成る [1][2]。

[1] 北陸先端科学技術大学院大学
1-1 Asahidai, Nomi, Ishikawa 923-1292, Japan
[a)] akagi@jaist.ac.jp

(1) **音声認識 (ASR) システム**により，ある言語で発話された源音声がテキストに変換される。
(2) **機械翻訳 (MT) システム**により，テキストが目的の言語に翻訳される。
(3) **音声合成 (TTS) システム**により，翻訳されたテキストが目的の言語の音声として再合成される。

ところが，音声は表現豊かであり，次のような情報を含む [3]。

- 言語情報：言語によって表記できるあるいは文脈によって一意に推測できる離散的情報
- パラ言語情報：言語情報を変形あるいは補完するために話者によって付加される離散的もしくは連続的情報
- 非言語情報：話者の感情，性別，年齢のような話者によって一般には制御できない情報

従来の S2ST システムは，この中で言語情報のみに焦点をあてており，本来，音声コミュニケーションにおいて重要であるパラ言語情報や非言語情報には，目を向けられていない。たとえば，従来の S2ST システムでは，源音声は“怒った”音声であったとしても，目的言語の音声として平静音声（neutral speech）が合成出力される。自然な音声コミュニケーションにおいて，源音声で表出された表現は，保たれたまま伝えられるべきである [4] が，S2ST システムでは未だ実現されていないのである。

1.3 Affective S2ST システム

本講演では，特に音声に含まれる感情に焦点をあて，多言語間でパラ言語情報・非言語情報をどのように処理するのかについて，我々の研究室で取り組んでいる“affective S2ST”と呼ぶ研究を紹介する。

「多言語の環境において，源音声に含まれる感情で色づけされた音声をどのように目的言語の音声として合成するのか？」

Affective S2ST システムでは，図 2 に示すように，(1) 源音声に含まれる感情情報を抽出し，この情報に従って (2) TTS システムで合成された目的言語の平静音声が感情を含むように変換する。講演では，Affective S2ST システムの構築のために，我々が議論した内容；

(1) 音声中の感情をどのように記述するか，そして，その記述にもとづいて感情の知覚と生成をどのようにモデル化するか [5][6][7][8]
(2) 感情の知覚と生成におけるモデル上での多言語間での共通性・相違性とは何か [9][10]，そして，
(3) 感情の知覚と生成におけるモデルをどのように工学的に実現するか

についても説明する。

謝辞 本研究の一部は，科研費（基盤 A，25240026）および A3 Foresight Program (JSPS) の援助を受けて行われている。

参考文献

[1] Nakamura, S., “Overcoming the language barrier with speech translation technology,” NISTEP Quarterly Review, 31, 35–48, 2009.
[2] Shimizu, T., Ashikari, Y., Sumita, E., Zhang, J.S., and Nakamura, S., “NICT/ATR Chinese-Japanese-English Speech-to-Speech Translation System,” Tsinghua Science and Technology, 13, 4, 540–544, 2008.
[3] Fujisaki, H., “Information, Prosody, and Modeling – with Emphasis on Tonal Features of Speech –,” Speech Prosody 2004, 23–26, 2004.
[4] Szekely, E., Steiner, I., Ahmed, Z., and Carson-Berndsen, J., “Facial Expression-based Affective Speech Translation,” Journal on Multimodal User Interfaces, DOI: 10. 1007/s12193-013-0128-x, 2013.
[5] 赤木正人, “音声に含まれる感情情報の認識 –感情空間をどのように表現するか–”, 日本音響学会誌, 66, 8, 393-398, 2010.
[6] Elbarougy, R. and Akagi, M., “Speech Emotion Recognition System Based on a Dimensional Approach Using a Three-Layered Model,” Proc. APSIPA2012, Hollywood, USA, 2012.
[7] Elbarougy, R. and Akagi, M., “Cross-lingual speech emotion recognition system based on a three-layer model for human perception,” Proc. APSIPA2013, Kaohsiung, Taiwan, 2013.
[8] Hamada, Y., Elbarougy, R. and Akagi, M., “A Method for Emotional Speech Synthesis Based on the Position of Emotional State in Valence–Activation Space,” Proc. APSIPA2014, Siem Reap, Cambodia, 2014.
[9] Elbarougy. R., Han.X., Akagi, M., and Li, J., “Toward relaying an affective speech-to-speech translator: Cross-language perception of emotional state represented by emotion dimensions,” Proc. O-COCOSDA2014, Phuket, Thailand, 48-53, 2014.
[10] Han, X., Elbarougy, R., Akagi. M. Li, J., Ngo, T. D., and Bui, T. D., “A study on perception of emotional states in multiple languages on Valence-Activation approach,” Proc NCSP2015, Kuala Lumpur, Malaysia, 2015.

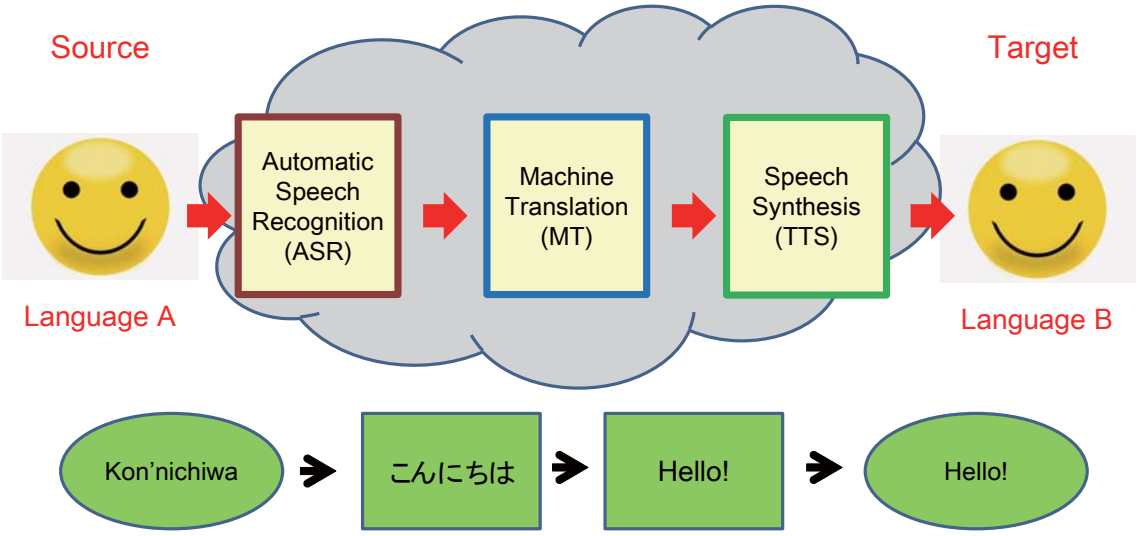

図 1 Traditional Speech-to-Speech Translator

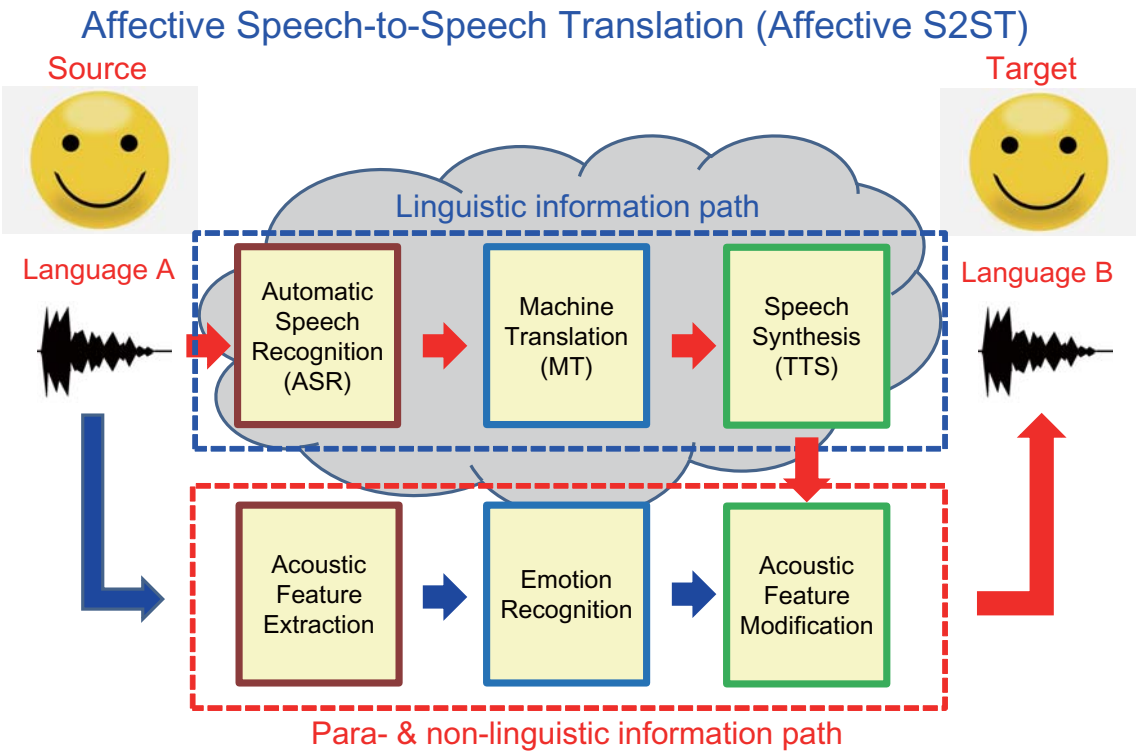

図 2 Affective Speech-to-Speech Translator