3　音声コミュニケーション技術

3　Spoken Language Communication Technology

3-1　音声コミュニケーション技術の概要

3-1　Overview of Spoken Language Communication Technologies

柏岡秀紀
KASHIOKA Hideki

要旨

NICT 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では、真に人との親和性が高いコミュニケーション技術の創造を目指し、誰が、いつ、どこで、どのような表現で、何語で話そうとも、息の合ったコミュニケーションを実現する多言語コミュニケーションの研究開発を推進している。本稿では、その構成技術として、音声に関わる音声認識技術、音声合成技術、および対話処理技術についてその概要を示す。

The goal of Spoken Language Communication Laboratory, Universal Communication Research Institute, NICT, is to realize multi language communication technologies with spoken language regardless of who, where, when, how and in which language users speak. Toward this goal, we will intensively develop ICT for a human-machine interface, such as multilingual speech recognition, multilingual speech synthesis, and spoken dialogue technology. In this paper, we indicate these technologies overview.

[キーワード]
音声認識，音声合成，対話処理
Speech recognition, Speech synthesis, Dialogue processing

1　まえがき

情報通信技術の進歩により、遠く離れた場所にいる人や異なる言語を使う人の間など、様々な環境・状況において様々な人の間でコミュニケーションの実現が望まれている。これらのコミュニケーションを実現するためには、誰が、いつ、どこで、どのような表現で、何語で話そうとも、息のあったコミュニケーションを実現する多言語音声コミュニケーションの研究開発が必要不可欠であり、人が最も自然に行うコミュニケーション手段の1つである音声によるコミュニケーションにかかわる音声コミュニケーション技術の研究開発は重要な課題の1つである。日常生活においても、スマートフォンの急速な普及により音声による多様な情報へのアクセスを実装したサービスが行われ、多くの人々に使われ始めている。

本稿では、音声コミュニケーション技術を構成する主要な技術である音声認識技術、音声合成技術、対話処理技術についてその概要を示す。

2　音声コミュニケーション技術

音声コミュニケーション技術は、我々の身の回りで日常行われている音声を介した様々なコミュニケーションの情報を記録、活用するとともに、

コミュニケーションの障害を乗り越えコミュニケーションを実現するための技術である。主な技術として、話された音声発話をテキストに変換する音声認識技術、テキストの情報を音声として出力する音声合成技術、音声によるインタラクションを支える対話処理技術がある。

2.1 音声認識技術

我々の日常生活において、人と会話して得られる情報のみならず、様々なアナウンス、また、テレビ、ラジオから流れる音声情報、さらにネットワーク上にある動画に付随する音声など、音声により入ってくる情報は非常に多い。これらの音声をテキストに変換する技術が音声認識技術である。

音声をテキストに変換するために、音としての特徴を捉え文字化するためのモデルと文字化された文字列を単語やフレーズ、文として言語化するためのモデルを大量のコーパスから学習し、入力された音声をそのモデルと照合することで実現している。音としての特徴を捉えて文字化するためのモデルが音響モデル、文字化された文字列を単語やフレーズ、文として言語化するためのモデルが言語モデルである。現在、音声コミュニケーション研究室では、音声とモデルの照合結果を探索するために、音響モデル、言語モデルを同様の探索モデルとして、重み付き有限状態トランスデューサ（Weighted Finite State Transducer: WFST）を用い表現し、最適化を行うことにより、高速で高精度な音声認識システムを構築し[1]、音声対話システムや音声翻訳システムにおいて利用している。現在、日本語については、65万語相当の辞書を持ち、6単語程度の短い発話であれば、実時間計数（Real Time Factor, RTF）1以内で処理できる。

様々な環境に含まれる音声には、音声以外の音が入力に含まれている。そのため、音声認識を行うためには、上述した音声からテキストへ変換する処理技術だけでなく、音声に含まれる音声以外の音を雑音として処理する技術、対象とすべき音声が含まれている音声区間を切り出すための発話区間検出技術も重要な技術となる。また、入力音声が必ずしも音声認識の対象として理想的なマイクを通じて入力されているわけではない。身近な例では、スマートフォン等の携帯端末を利用した様々なアプリケーションが利用される様になってきている。様々な雑音が音声と同時に入力されており、音声認識の前処理として、雑音を抑圧する処理や、音声認識のモデル（特に音響モデル）を、雑音を含む音声データから構築し、耐雑音性を高めた認識モデルの構築により対処されている。

具体的な音声認識が利用される応用事例としては、現在、スマートフォンなどで普及しつつある音声翻訳、対話システムが最も身近な応用アプリケーションである。また、コールセンター用のシステムへの期待も高い。さらに、ニュースなどテレビ番組やネットワーク上の動画等の字幕付与は、障害者への対応や記録など様々な理由から望まれている。これら応用事例において音声認識を実用的に利用できるようにするためには、雑音処理、長文への対処、精度向上といった課題に取り組む必用がある。また、多言語への対応も、重要な課題である。

2.2 音声合成技術

様々な状況において音声で情報を発信すべきことは多い。特に公共交通機関や防災のアナウンスに音声合成が実際に利用されている。音声で伝達することが期待される情報がテキストであるときに、テキストの情報を音声として出力する技術が音声合成技術である。

テキストを音声として出力するために、テキストがどのような構成であるかを解析するテキスト解析部と、テキストを構成する各語やフレーズをどのようなイントネーション、リズムで音として生成するか等の処理を行う合成エンジンを構築し実現している。テキスト解析部では、単語の情報やフレーズの情報を取り出し、合成エンジンでは、音声合成用音響モデルを用いて合成する。現在、音声コミュニケーション研究室では、HMM音声合成による方式を採用し、日本語だけでなく、英語、中国語、韓国語、インドネシア語、ベトナム語、マレー語に対応した音声合成を実装[2]している。また、音声合成用音響モデルによって、言語の種類、声質、発話スタイルが異なるため、モデルを切り替えることで発話スタイルや声質を変えることができることに着目し、音声

翻訳などにおいて原発話者の音声特徴に類似したモデルを選択することで、翻訳結果の音声を原発話者に類似した音声で出力するボイスセレクターを開発した。さらに、モデル構築時のフィルターを改善し合成音声の自然性を改善している。

音声翻訳や音声対話システムを構成するためには、音声合成システムは必要不可欠であり、人間との会話で利用するため、自然性の豊かな合成音声が望まれている。音声合成用音響モデルも同一人物の音声コーパスから構築されるが、テキストを読み上げた音声を収録したコーパスから構築するより、会話している音声を収録したコーパスから構築した方が、自然性が上がると言う研究結果も報告されている。また、音声合成用音響モデルの構築はコストがかかるため、その自動化も重要な課題である。

2.3 対話処理技術

音声によって対話を継続して進めていくためには、発話の状況、環境を把握し発話を理解しなければ適切な応答、質問を行うことができない。音声によるコミュニケーションでは、発話内容だけでなく、音声が持つ情報がこれらの状況、環境に付随する情報を伝えることがある。また、連続する発話を考慮することにより得られる情報もある。様々な状況において対話を総合的に管理し、発話を理解し、次発話の予測、生成を行う技術が対話処理技術である。

対話処理技術は、発話を理解するための発話理解技術、発話理解により得られた発話意図により周りの情報サービス等との関連を考慮し応答内容を生成するコンテキスト処理技術、および応答内容から応答文を生成する発話生成処理技術に分けることができる。発話を理解するためには、発話内容と発話意図を理解する必用がある。発話内容の理解は、固有名等の概念の把握、多様な表現の把握、同音異義語や同一対象の異なる表現の把握に基づいて論理的な命題として発話内容を記述することで実現される。発話意図の理解は、発話表現やイントネーションなどの音声の持つ情報により、依頼や質問、情報提供などの発話意図であることを認識する。音声コミュニケーション研究室では、観光案内の音声対話コーパスを収集し、WFST を用いた対話制御機構 [3] を開発し、高速かつ高精度な音声言語理解を実現している。

対話処理技術は、直接的に音声対話システムの構築に利用され、一問一答の質問応答システムとは異なり対話を継続することによって、適切な情報を得ることが可能となる。また、対話システムに限らず文脈を理解し予測する機構に応用することができる。これは、様々な音声コミュニケーション技術において文脈を考慮することで適切な処理を実現していくための重要な技術と考えられる。

3 むすび

音声コミュニケーション技術を構成する主要な技術である音声認識技術、音声合成技術、対話処理技術についてその概要についてまとめた。NICT 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では、これら要素技術を単独で研究開発するだけではなく、多研究室の技術と組み合わせることにより、実社会で活用できる統合システムの研究開発を行い、実際にシステムを利用することで、要素技術の抱えている課題、進むべき方向を見極め、研究開発を推進している。具体的には、音声認識技術、音声合成技術に、多言語翻訳技術を統合した音声翻訳システム VoiceTra を、また、音声認識技術、音声合成技術、対話処理技術を統合した音声対話システム AssisTra を開発し、実証実験としてスマートフォン上で利用可能なアプリケーションとして公開している。今後、音声アーカイブを構築する技術の研究開発を進め、ネットワーク上の情報源として、音声によるデータや映像情報に含まれる音声情報を、テキスト情報と同等に活用可能とすることで、コミュニケーションを阻害する壁を乗り越えた息の合ったコミュニケーションを実現する多言語コミュニケーションの研究開発を推進していきたい。

参考文献

1 Dixon Paul Richard, Chiori Hori, and Hideki Kashioka, “A COMPARISON OF DYNAMIC WFST DECODING APPROACHES,” In Proc. ICASSP, 2012.

2 Yoshinori Shiga, “EFFECT OF ANTI-ALIASING FILTERING ON THE QUALITY OF SPEECH FROM AN HMM-BASED SYNTHESIZER,” In Proc. ICASSP, 2012.

3 C. Hori, K. Ohtake, T. Misu, H. Kashioka, and S. Nakamura, “Statistical Dialog Management Applied to WFST-based Dialog Systems,” In Proc. ICASSP, pp. 4793–4796, 2009.

（平成 24 年 6 月 14 日 採録）

柏岡秀紀（かしおかひでき）
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室長
博士（工学）
音声言語処理、音声翻訳、音声対話
hideki.kashioka@nict.go.jp