声できくと、声で答える。観光案内アプリAssisTra

ー自然な音声で簡単に情報を取得できるシステムの実現を目指してー

「音声コミュニケーション研究室の技術を用いたスマートフォン用の音声対話システム AssisTra を、Siri しゃべってコンシェルに先駆けて、2011 年 6 月に公開しました。」

翠 輝久（みす てるひさ）
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 研究員

人と人、人とコンピュータのコミュニケーションに関心を持ち、音声対話の研究をしています。

水上 悦雄（みずかみ えつお）
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 主任研究員

コミュニケーションにおける相互調整の性質に関心があり、対話の評価研究をしています。趣味は映画（特に SF）観賞、公園探検と称する娘たちとの散歩、昆虫探索、車でのぶらり遠出旅行。

はじめに

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、わずらわしい操作を覚えなくとも、その人にとって自然なコミュニケーションの手法で、容易に情報システムを利用できる社会の実現を目指して、研究を進めています。中でも私たちは、人に話しかけるような、自然な音声による要求を受け付け、その意図を理解・推測することによって、適切な情報を提示する、高精度対話処理技術を研究しています。これまでの研究成果の実証実験および実データ収集を目的として、観光案内 iPhone 用アプリ「AssisTra」を 2011 年 6 月にリリースしました。本稿では、AssisTra の中心機能である、『はんなのガイド 京都編』で利用されている音声対話処理技術について解説します。

『はんなのガイド 京都編』とは？

ユーザの自然な音声による要求を受け付け、音声と画面で、その要求に答える「音声対話システム」です。図1の例のような音声対話をすることができ、ユーザは京都の観光スポットやレストランなど観光に役立つ様々な情報を調べることができます。

音声対話処理技術

一般に音声対話システムは、図 2 のような構成をしており、大きく分けて、音声認識、音声言語理解、対話制御、応答文生成、音声合成の 5 つのモジュール(要素技術)で構成されます。『はんなのガイド 京都編』に用いられているこれらのモジュールは、すべて当研究室で開発したものです。以下では、これらのモジュールについて概説します。

図1　『はんなのガイド 京都編』対話例

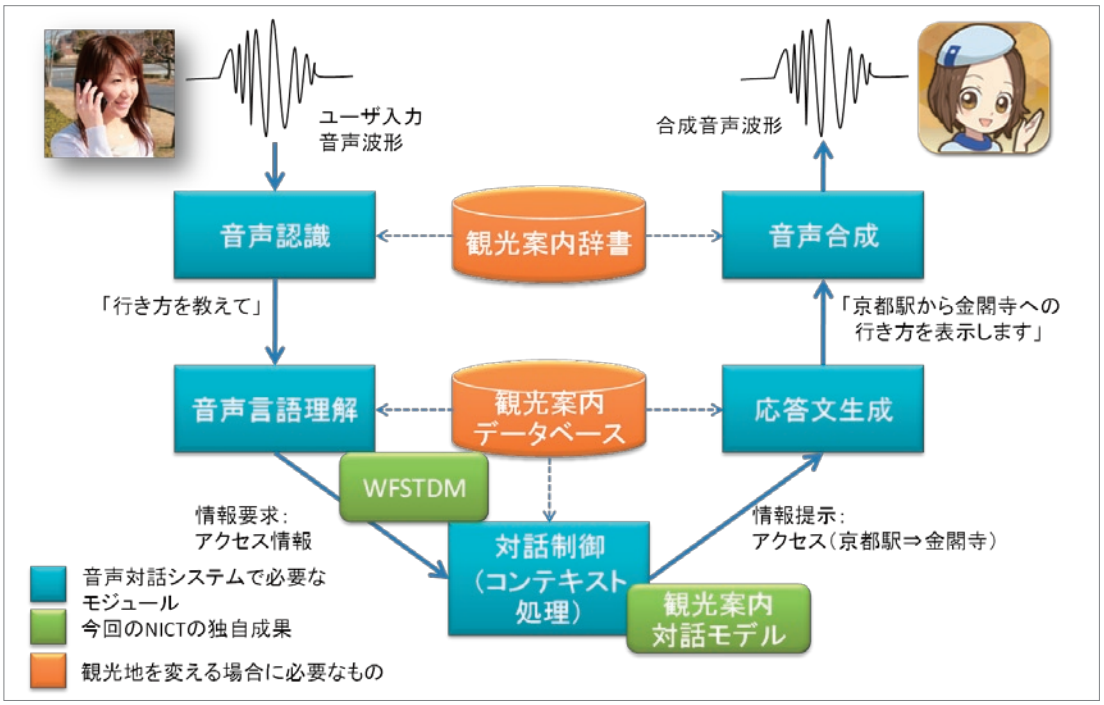

図2　音声対話システム構成図

・音声認識・音声合成・応答文生成

音声認識・音声合成は、隠れマルコフモデル*に基づく統計的手法を利用しており、大量の音声データを学習に用いることで、自然で連続的な音声発話文を認識し、人の発話音声に近い合成音を作成することができます。後述の大量の観光案内対話データを利用して、観光案内用に特化したモデルを作成することにより、高い音声認識率と、ユーザに話しかけるような自然な合成音声を実現しています。さらに、応答文生成で利用するテキストとして、プロのガイドの発話内容をもとに、桜、紅葉など様々な観点からの観光スポットの説明文を整備しました。

・音声言語理解

人間の自然発話には、ユーザや状況によって様々な言い回しが存在します。たとえば、「観光スポットへのバスを利用したアクセス方法」が知りたいと考えている場合を考えますと、図3の例をはじめとして、多種多様な言い回しが存在します。このような発話の意図を解釈することは人にとっては難しいことではありませんが、コンピュータがこれらの発話を理解するためには、これらの表現を同一のシンボル(コンピュータが処理可能な言葉)に変換する必要があります。これが音声言語理解の役割です。

この機能を実現するためには、ユーザが実際に使用する表現を収集するとともに、高精度な音声言語理解アルゴリズムを研究・開発することが重要になります。会話の中で実際に利用される言い回しを収集するために、私たちはプロの観光ガイドと旅行者の模擬会話を150時間300対話収録しました。これは、現在収集されている単一状況での音声対話データとしては世界的にも大規模なものです。さらに、プロトタイプ音声対話システムを構築して、被験者実験を行い、実際のシステム利用を想定した状況での発話表現を収集しました。これらのデータをもとに、私たちの研究室で独自に開発した音声言語理解・対話制御フレームワークである『重み付き有限状

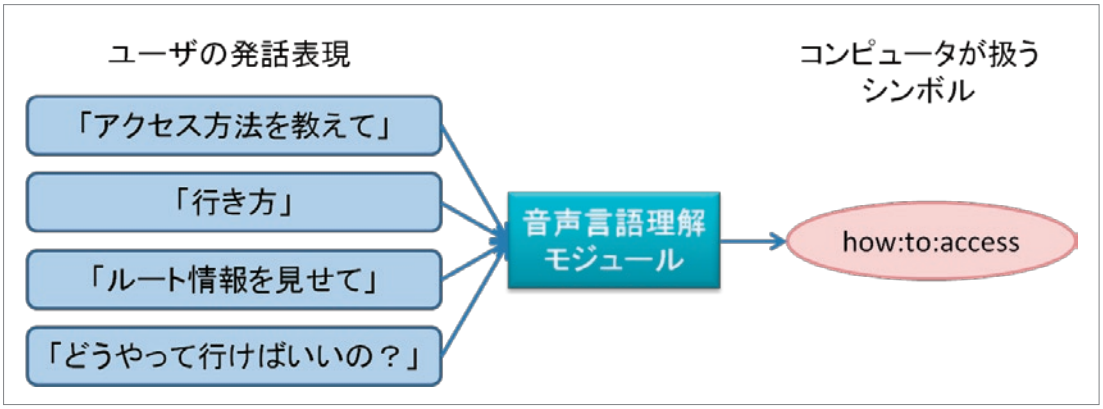

図3　言語音声理解の例

態トランスデューサ対話制御機構(WFSTDM: Weighted Finite-State Transducer-based Dialog Manager)』を用いて、WFST 表現による音声言語理解モデルを作成することで、高速かつ高精度な音声言語理解を実現しています。

・対話制御

まったく同じ発話が入力された場合でも、状況や発話履歴に応じて発話に含まれるユーザの意図が異なる場合があります。たとえば、「アクセス方法を教えて」という入力があった場合には、直前の対話内容に基づいて「どこから、どこまで、どのような交通手段で」などの情報を補完する必要があります。これらの発話に隠れた意図を適切に補って応答内容を決めることが対話処理の役割です。

このような対話履歴処理は、対話システムが利用される状況や、ユーザがシステムを使う目的に対する依存性が高いものです。そこで、ユーザの実際の利用状況に近い、前述の大規模対話データをもとに観光対話用の履歴処理モデルを作成し、対話履歴を適切に処理しています。

おわりに

今回、アプリを公開し、収集されたログデータを分析していますが、システムの応答の精度はまだ十分ではありません。人間の発話や意図の種類・言い回しのバリエーションが 150 時間程度の学習データではカバーしきれないほど多様で複雑なものであり、コンピュータが人の意図を正確に理解するためには、より大きな対話データを収集するとともに、音声言語理解や対話履歴処理の精度の改善が必要であることが分かりました。今後はシステム運用により収集した発話データを追加して各モジュールのモデルを再構築するとともに、より柔軟に発話を理解し対話を制御するアルゴリズムの研究を進めていきます。また、システムの利用の拡大を目指して、訪日観光支援に利用できるように『はんなのガイド 京都編』の英語版を 2012年 3月に無料公開しました。さらに、チケット予約や、コールセンター業務など、実世界で必要とされている様々なタスクを扱う音声対話システムを構築し、対話処理技術の実用性を証明していきたいと考えています。

用語解説

＊　隠れマルコフモデル

観測される記号列(音声認識の場合、音声の特徴量)が、直前の m 個の記号から決定されるマルコフ過程であると仮定し、それを出力するような状態遷移系列が非決定的である(隠れている)とする確率モデルです。音声認識の場合、状態遷移確率などのパラメータが、大量のデータから学習され、最も高い確率で記号列を出力するような単語列や音素列が認識結果となります。