音声ユビキタス情報環境実現のための次世代音声技術基盤の検討

Next-generation voice technology foundation for voice ubiquitous information environment realization

徳田 恵一 (Keiichi TOKUDA, Dr. Eng.)
名古屋工業大学大学院 教授
(Pprofessor, Nagoya Institute of Technology)

電子情報通信学会　日本音響学会　人工知能学会　情報処理学会　IEEE SCA

受賞：IEEE Fellow (2014)　ISCA Fellow (2013)　(社)情報処理学会「喜安記念業績賞」(2013)　The 2013 EURASIP-ISCA Best Paper Award (2013)　科学技術分野の文部科学大臣表彰「科学技術賞（研究部門）」(2012)　(社)電子情報通信学会「平成 20 年度情報・システムソサイエティ活動功労賞」(2008)　(社)電子情報通信学会「平成 19 年度電子情報通信学会情報・システムソサイエティ論文賞（連作論文）」(2007)　(財)電気通信普及財団「第 23 回電気通信普及財団賞（テレコムシステム技術賞）」(2007)　(社)電子情報通信学会「第 57 回論文賞」(2000)　(社)電子情報通信学会「第 7 回猪瀬賞」(2000)　(財)電気通信普及財団「第 16 回電気通信普及財団賞（テレコムシステム技術賞）」(2000)

著書：Heiga Zen, Keiichi Tokuda, "7.3 The HMM-based speech synthesis system (HTS)," in Computer processing of Asian spoken languages, Editors: Shuichi Itahashi, Chiu-yu Tseng, Consideration Books, Los Angeles, March 2010. (ISBN 978-0-935047-72-1)

統計的パターン認識, 分担, 数理工学社（予定 執筆依頼受領）

徳田恵一, "4.3 HMM 音声合成における統一的な韻律の制御," 韻律と音声言語情報処理—アクセント・イントネーション・リズムの科学—, 広瀬啓吉編著, 丸善, pp.118–127, Jan. 2006. (ISBN 978-4-621-07674-3)

Keiichi Tokuda, Heiga Zen, Alan W. Black, "An HMM-Based Approach to Multilingual Speech Synthesis (Chapter 7)," Text-to-Speech Synthesis: New Paradigms and Advances, Shrikanth Narayanan, Abeer Alwan (Eds.), Prentice Hall, pp.135–153, Aug. 2004. (ISBN 978-0131456617)

Shin-ichi Kawamoto, Hiroshi Shimodaira, Tsuneo Nitta, Takuya Nishimoto, Satoshi Nakamura, Katsunobu Itou, Shigeo Morishima, Tatsuo Yotsukura, Atsuhiko Kai, Akinobu Lee, Yoichi Yamashita, Takao Kobayashi, Keiichi Tokuda, Keikichi Hirose, Nobuaki Minematsu, Atsushi Yamada, Yasuharu Den, Takehito Utsuro, Shigeki Sagayama, "Galatea: Open-source Software for Developing Anthropomorphic Spoken Dialog Agents," Life-Like Characters: Tools, Affective Functions, and Applications, Series: Cognitive Technologies, Helmut Prendinger, Mitsuru Ishizuka (Eds.), Springer-Verlag, pp.187–211, 2004. (ISBN 978-3540008675)

小林隆夫, 徳田恵一, "14.3 音声のスペクトル分析," スペクトル解析ハンドブック, 日野幹男総編集, 朝倉書店, pp.481–492, Feb. 2004. (ISBN 978-4-254-20108-6)

研究専門分野：音声言語情報処理　マルチモーダル情報処理　統計的学習理論

あらまし　音声は、人間にとって最も基本的なコミュニケーション手段であるということから、音声対話システムを含む音声インターフェースの実用化を目指し、これまでに長い間、研究開発が続けられてきたが、実社会における音声インターフェースの広い普及には未だ至っていない。音声特有の生き生きとしたインタラクティブ感のあるやりとりは、手による操作を主とする従来型のインターフェースでは実現することができない音声インターフェースの固有の「魅力」の一つであるが、未だ広く普及していない原因の一つに、音声ならではの「魅力」を最大限に活かした音声対話コンテンツを十分に提供することができていない点が考えられる。そのため、音声インターフェースの「魅力」の解明と、それを実現するシステムがどのような要件を具備すべきかを解明することが必要になるが、「魅力」というものは容易に評価できるものではなく、人間が持つ感性や知見の積み重ねによって創られていくものである。本稿では、ユーザが音声対話コンテンツを容易に作成する仕組みを確立し、ユーザが大量の音声対話コンテンツを生成・評価する中から、帰納的にその本質を探究する試みについて紹介する。

１．研究の目的と背景

１．１　産業構造との関係

音声が人間にとって最も基本的なコミュニケーションメディアであることから、音声インタラクションによるユビキタスな情報環境の構築は、「人間と情報環境の調和」という意味で到達すべきひとつの理想形態であるが、このような社会が未だ到来していない原因のひとつは、ユーザにとって魅力的なコンテンツを提供することができていないという点であろう。また、ユーザによる評価や改善へのアイディア・工夫があったとしても、それらが速やかにフィードバックされる仕組みのない閉じた系となっていることが多いと考えられる。システムの設計に関しても、インタラクティブ感の実現を主題において構築されたものは多くはないと思われる。このような問題を解決するためには、音声コンテンツ制作においても、コンテンツ制作部分をソフトウェア制作部分から分離し、クリエータおよび

一般ユーザに広く解放する必要があると考えられる。このことは、図 1 に示す産業構造の変遷と符合させることができる。情報通信産業は、電子機器によって成り立っていたが、まず、ソフトウェア制作が分離され、更に、コンテンツ制作が分離され、それらが大きな産業分野を形成するに至っている。典型例として、携帯ゲーム等が挙げられ、これらの分野では、コンテンツ制作の分離が進んでいる。音声技術に関しても、このような変革を起こすためには、幅広くより多くのクリエータ・ユーザが「コンテンツ生成の循環系」に関わることにより、魅力的なコンテンツの生成が加速される必要があると考えられる。

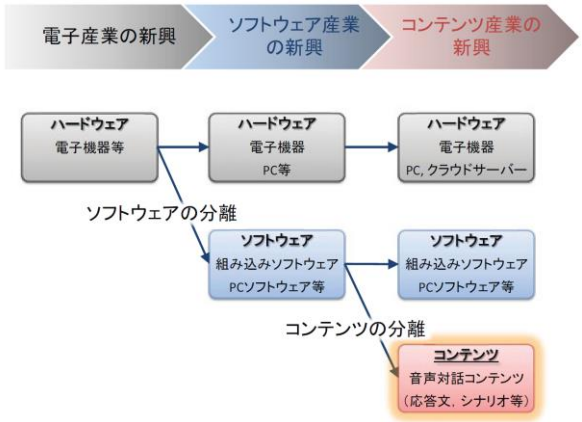

図１　情報通信に関連した産業構造の変遷

１．２ UGC アプローチとの関係

現在、CGM（Consumer Generated Media）、 UGC（User Generated Content） 等の用語で参照されるユーザが作成したコンテンツが活用されている。これは、ユーザが主体となってコンテンツを作成するシステムであり、Wikipedia、YouTube、ニコニコ動画、食べログ、クックパッドなどがよく知られている。これらのシステムでは、ユーザの手で次々と新しいコンテンツが作られることや、ユーザの評価や要望がそのままコンテンツに反映されることが大きな特徴である。本研究のアプローチは、これらの音声対話コンテンツ版と捉えることもでき、ユーザが情報発信する情報環境を実現しようとするものとなっている。

１．３ 音声ユビキタス情報環境実現のためのデバイス

音声ユビキタス情報環境を実現するための装置として、通常の PC から情報家電、スマートフォン等の携帯デバイス等、様々な形態を考えることができるが、本研究では、デジタルサイネージに着目する。デジタルサイネージとは、大型液晶ディスプレイ等の表示技術とデジタル通信技術により実現される情報提供媒体・広告媒体のことであり、表示内容をデジタル通信によって随時変更できる、動画等を表示できる等の利点があり、その将来性と可能性が大きな注目を集めている。近年は、近接センサー、顔画像認識等を駆使し、インタラクティブに表示を制御することが試みられており、更に、双方向の音声インタラクション機能を付与することにより、自然で印象深いインタラクティブ性を演出することが可能と期待される。本研究では、このような音声機能を備えたデジタルサイネージ（図 2）をキーデバイスと位置付け、キャンパス、あるいは観光施設等の公共空間に複数設置し、更に、インターネットを介したネットワーク連携の仕組みを考慮しながら、実証実験の基盤とする。

図２　音声対話デジタルサイネージ

以下、2 章で基盤技術に関連した成果について述べ、3 章でユーザによるコンテンツ生成環境の構築について紹介する。4 章で実証実験における成果について紹介し、5 章で将来展望を述べる。

２．基盤技術と関連ソフトウェアの高度化

「魅力的」なコンテンツを生成するためには、システムがユーザにとって「魅力的」であり得る技術基盤をもたなければならない。このため、克服しなければならない個別の課題が数多くある。本章では、課題克服のための基盤技術の高度化、また、その技術を導入したソフトウェアの高度化について紹介する。

２．１ 基盤技術

音声合成、音声認識等の音声情報処理技術に関する多くの基礎研究を実施し、基盤技術の高度化に取り組んだ。ここで得られた研究成果については、学会等で積極的に発表した。以下に代表的なものを述べる。

- HMM 音声合成*1 のための特徴抽出とモデリングの統合手法
  音声の特徴量抽出と HMM の学習を統合した新たな音声のモデル化手法を提案した。提案法により、統一された基準で音声波形を直接モデル化することが可能となり、システム全体としてより適切な音声のモデル化がされることから、合成音声の品質を大きく改善することができた[1]。
- 加算構造によるスペクトラムモデルの改良
  コンテキストによる音響特徴量の加算構造を仮定し、音声合成のための加算モデルを提案した。提案法は、合成音声の品質を改善すると同時に多様な声質を生成することが可能であり、音声対話システムにおける多様な声質の再現に利用可能である[2]。
- 音声対話システム構築ツールキット MMDAgent の FST スクリプト*2 の拡張
  対話記述方式を拡張し、新たに変数を取扱い可能な FST スクリプトを定義した。これにより、正規表現によるマッチングや変数などを統一的に扱うことが可能になり、複雑な対話をより短い FST スクリプトで記述可能になった。

この他にも、多様な感情を表す音声合成や音響モデルの適応手法等、音声対話システムに必要な様々な基盤技術の改善に取り組んだ。

２．２ 関連ソフトウェア

前述の基盤技術の高度化において得られた成果を研究基盤ソフトウェアとしてまとめ、オープンソースソフトウェアとして公開した。以下に、筆者らが開発・公開したオープンソースソフトウェアを示す。

- 音声対話システム構築ツールキット MMDAgent（38,805 ダウンロード）
- 音声認識エンジン Julius （146,841 ダウンロード）
- HMM 音声合成ツールキット HTS（252,516 ダウンロード）
- HMM 音声合成エンジン hts_engine API（28,604 ダウンロード）
- 日本語音声合成システム Open JTalk（33,252 ダウンロード）
- 音声信号処理ツールキット SPTK（27,873 ダウンロード）

これらのオープンソースソフトウェアの開発は、初年度から継続的に行っており、本研究課題期間中にも、新バージョンを順次公開してきた。特に、音声対話システム構築ツールキット MMDAgent については、Android OS 上での動作を可能とした形で公開した（図 3）。

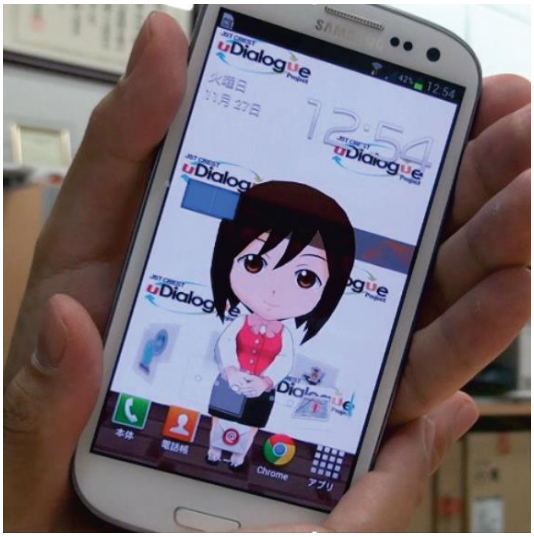

図３　Android 上で動作する MMDAgent

MMDAgent を Android OS に移植することにより、スマートフォンやタブレット単体で動作可能な、応答遅延の少ないモバイル音声対話システムが実現可能となり、音声対話システムを様々な環境で利用することが可能となった。なお、スマートフォンやタブレット単体で動作可能な、3D エージェント付き音声対話システムを実現するオープンソースソフトウェアの実現は、世界初の成果と考えられる。これらのソフトウェアは最先端の技術を含む研究基盤ソフトウェアであり、ダウンロード数からわかる通り、既にデファクトスタンダードのひとつとしての地位を確立している。実際に、図 4 に例示するように、学術論文やソフトウェア開発、イベント等、様々な場面で広く利用されている。

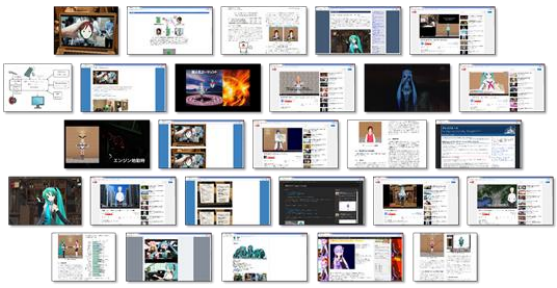

図４　MMDAgent 等が利用された事例

３．ユーザによるコンテンツ生成環境の構築

音声対話コンテンツは、挨拶のように短いものから、数分程度のストーリー性をもったものまで、様々な粒度のものが考えられる。また、できるだけ簡便に登録・共有できることが求められる。ここでは、音声対話コンテンツを図 5 に示すような階層にわけて考え、ユーザによるコンテンツ生成環境の構築に取り組んだ事例を紹介する。

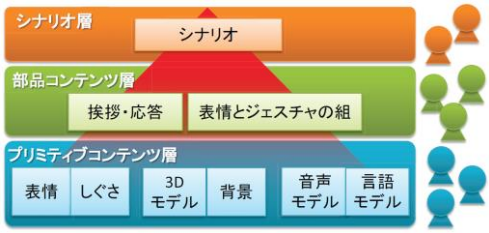

図５　音声対話コンテンツの階層

３．１　音声インタラクションビルダ

まず、ユーザがもっともコンテンツ生成に参加しやすいと思われる「部品コンテンツ」の登録・共有法を確立し、その後、ある程度専門知識をもったユーザのための「プリミティブコンテンツ」、続いて、大規模なコンテンツを自作するユーザのための「シナリオ」について、順次検討を行う必要がある。クライアント環境で、より詳細に音声対話コンテンツを作りこめる環境として、音声インタラクションビルダの試作を行った。このビルダは、図 6 に示すように、

(1) 状態遷移図の三次元空間における可視化とブラウジングによる FST スクリプトの構造把握機能
(2) イベント入力シミュレーションによる動作確認機能
(3) 入出力イベント系列の保全と再現による時系列インタラクションの検証機能

の 3 つのコンポーネントから構成されている。主観実験の結果、ユーザからは既存の環境に比べてコンテンツが作成しやすいとのフィードバックが得られた。

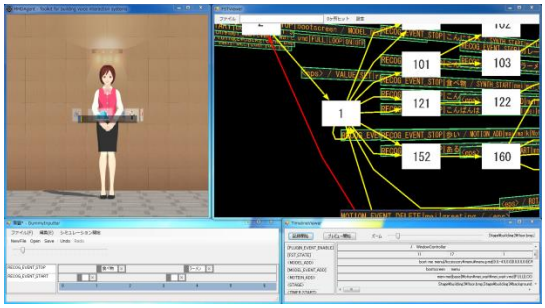

図６　音声インタラクションビルダ

３．２　コンテンツの循環環境

Web インタフェース等を用いて、ユーザが容易に音声対話コンテンツを生成可能なシステムの開発を行う必要がある。音声対話コンテンツには、シナリオや話し言葉だけでなく、エージェントの動きやパネル画像・動画や地図などの様々なマルチメディアコンテンツが含まれる。一般に、これらのマルチメディアコンテンツを Web ブラウザを用いて編集することは困難

である。そこで、コンテンツの実データとメタデータを分離し、メタデータに基づく編集を行う仕組みを検討した。コンテンツの循環を促進するために、提案システムや音声対話コンテンツをインターネットから利用可能な音声対話クライアントを提供することによって、幅広いユーザが手軽に参加可能な仕組みを検討し、サーバ・クライアント型の連携と、エージェント技術を用いた P2P 型の連携の二つの方式から取り組んだ。前者は、ユーザが生成した音声対話コンテンツを、サーバを介してネットワーク上で配信・共有するための仕組みである。まず、FST スクリプトを拡張し、パッケージ化手法について検討した。これにより、FST スクリプトの保守性が向上すると同時に、従来困難であった FST スクリプトの部分更新や機能単位の配信が可能になった。さらに、図 7 に示す音声対話コンテンツのパッケージ化に基づく配信の仕組みを試作することによって、パッケージ単位でのコンテンツ循環の仕組みについて検討した。後者は、図 8 に示すように、複数のスマートフォン端末上の MMDAgent が協調動作するための環境の構築である。具体的には、Android 版 MMDAgent と連動して動作するネットワークエージェントを開発し、このエージェントを用いて P2P 通信を行った。P2P 通信機構において、通信対象となる端末の認識には NFC を用いており、端末を接触させて画面をタップするだけで通信を開始することができる。また、NFC を用いて通信を開始した後は、Bluetooth を用いて通信を行うことで、端末を離しても通信が可能である。試作システムでは、2 台の端末間でのネットワーク連携に成功した。NFC を用いた相手端末の認識や、Bluetooth を用いた通信路の確立も、実用的な時間に収まっており、良好な結果が得られた。

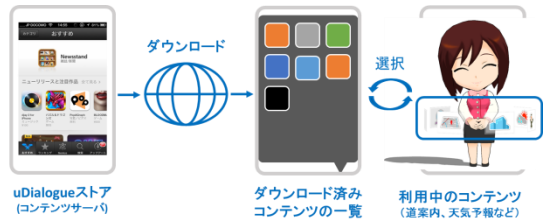

図７　コンテンツのパッケージ化の概念

図８　ネットワークエージェントによる P2P 通信

４．実証実験

デジタルサイネージをキャンパス内外に複数台設置することによって、さまざまな利用環境やターゲットを想定した実験を行った（図 2、図 9）。

図９　様々な実証実験

ここでは、名古屋工業大学と半田市観光協会に設置・運用してきた音声対話デジタルサイネージのアンケート結果を紹介する。これらのシステムは、実際にユーザが対話を利用すると同時に、Web ブラウザ等を用いて音声対話コンテンツを投稿する仕組みも含まれる。アンケート結果を図 10 に示す。全ての項目において、名古屋工業大学のアンケート結果よりも半田市の実証実験の結果の方が、より良い結果となった。当初は、主に大学生を対象とした名古屋工業大学の方が良い結果となると予想していたが、一般の人の利用者

が多い半田市の方が良い結果となった。これらの結果は、提案システムは一般の人にも受け入れられることを示唆する。

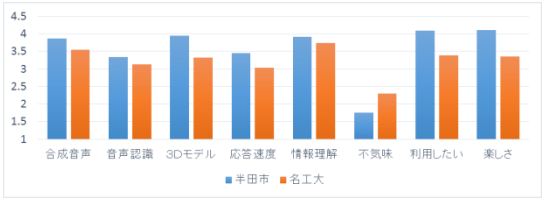

図１０　アンケート結果

５．将来展望

本研究は、ユーザによるコンテンツ生成環境の構築という新しい切り口から音声技術を考えるものであり、今後の音声インタフェース構築のために有用な知見が得られるだけでなく、音声インタフェース普及のブレークスルーに繋がることが期待される。また、公共空間におけるデジタルサイネージの形での実装・実験は新しい形のユビキタス情報環境の具現化となっており、近い将来における実用化等、波及効果が期待できる。さらに、大量の音声対話コンテンツ例や実際の音声対話例を大量に収集することが可能となるため、これらの大量のデータに基づいた音声対話の統計的モデル化手法へと発展させていくことが可能と期待される。なお、研究成果はオープンソースの研究基盤ソフトウェアとして、これまで同様に、公開していく予定である。

用語解説

*1 HMM 音声合成
音響モデルとして、隠れマルコフモデル（Hidden Markov Model; HMM）を用いた音声合成手法。高い多様性の表現能力や低い言語依存性が注目されている。

*2 FST スクリプト
有限状態トランスデューサ（Finate State Transducer; FST）を用いた対話記述言語。状態遷移と入出力シンボルから成り、柔軟な音声対話を実現できる。

参考文献

[1] Kazuhiro Nakamura, Kei Hashimoto, Yoshihiko Nankaku, and Keiiti Tokuda, “Integration of spectral feature extraction and modeling for HMM-based speech synthesis,” IEICE Transactions on Information and Systems, vol. E97-D, no. 6, June 2014.

[2] Shinji Takaki, Yoshihiko Nankaku and Keiichi Tokuda, “Spectral modeling with contextual additive structure for HMM-based speech synthesis,” IEEE Journal of Selected Topics in Signal Processing, vol. 8, issue. 2, pp. 229-238, April 2014.

この研究は、平成２３年度ＳＣＡＴ研究助成の対象として採用され、平成２４～２６年度に実施されたものです。