User Generated Dialogue Systems: uDialogue

徳田 恵一[1] tokuda@nitech.ac.jp
李 晃伸[1] ri@nitech.ac.jp
南角 吉彦[1] nankaku@nitech.ac.jp
大浦 圭一郎[1] uratec@nitech.ac.jp
橋本 佳[1] hashimoto.kei@nitech.ac.jp
山本 大介[1] yamamoto.daisuke@nitech.ac.jp
内匠 逸[1] takumi@nitech.ac.jp
打矢 隆弘[1] t-uchiya@nitech.ac.jp
堤 修平[1] tsutsumi.shuhei@nitech.ac.jp
Steve Renals[2] s.renals@ed.ac.uk
山岸 順一[3] jyamagis@nii.ac.jp

[1] 名古屋工業大学
[2] エジンバラ大学
[3] 国立情報学研究所

概 要

ユーザによる音声対話コンテンツ生成という新しい概念を導入し，それが実際に機能するための仕組みや条件を実証的に探究する試みについて紹介する．音声インタフェースの主要な「魅力」の一つとして，テキスト処理だけでは実現することのできない，生き生きとしたインタラクティブ感のあるやりとりが挙げられる．本研究では，音声対話システム全体をユーザが制作・改変可能なコンテンツと，それを駆動するシステムとに分離し，1）「魅力的」な音声対話を成立させうるシステムの要件と，2） ユーザによって「魅力的」な音声対話コンテンツが次々と生成されるための条件を解明し，同時にそのための仕組みを確立することを目指す．音声対話デバイスとして公共空間に設置した双方向音声案内デジタルサイネージを想定し，インターネットを介したコンテンツ生成環境を実現した上で，ユーザによる音声対話コンテンツ生成の実証実験を行う．本研究で提案する音声対話コンテンツ生成の枠組みの実現は，音声技術普及のブレークスルーにつながるものと期待される．

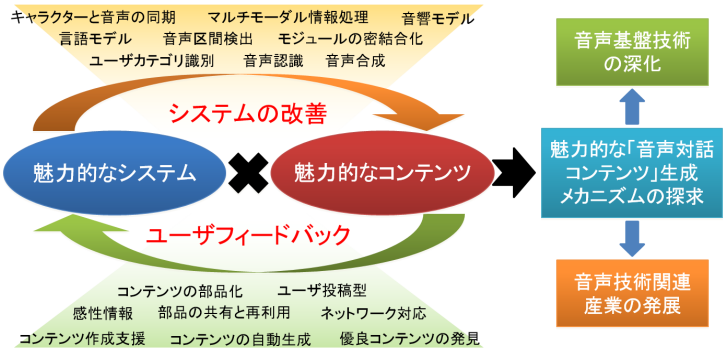

図 1. コンテンツ生成の循環系

1　まえがき

人間中心の情報環境とは，あらゆる人間が情報発信源となり，また自然に情報を享受できる環境である．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ということから，高度なネットワーク情報通信機器があらゆる場所に偏在する中で，人々が音声によって自然かつ自在に情報を発信し，またやりとりする環境が広く社会に実現されることは，ひとつの理想と考えられる．音声認識，音声合成，音声対話の要素技術に関しては，実用化のための要件を満たす水準に達しつつあるが，そのような情報環境が実社会で広く普及しているとは言い難い．個別の要因として，音声認識精度の向上を始めとした既知の課題を列挙することができるが，このような技術的な積み重ねだけでは解決できない要因があると考えるのが一般的であろう．それらの一つは，音声対話システムのユーザにとってのある種の「魅力」に関するものである．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，テキスト的な処理だけでは実現することのできない音声インタフェースの主要な「魅力」のひとつである．それを実現するためには，表情やしぐさ・声質や間合いなど，人間的かつ高度な音声信号処理が必要とされる領域まで踏み込む必要があるが，現実の音声対話システムは，ハードウェア，ソフトウェアを含む様々な制約のために，柔軟性がなく無味乾燥なものになりがちである．

本研究では，音声対話システム全体をユーザが制作・改変可能なコンテンツとそれを駆動するシステムに分離することにより，音声技術を広く人々の間に普及するための「魅力的なコンテンツ」,「魅力的なシステム」がそれぞれどのような要件を具備すべきかを解明することを目的とする（図 1）．しかしながら,「魅力」は機械的に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．従って，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することにより，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見出す必要があると考えられる．このことから，本研究における基本的な戦略は，図 1に示す「コンテンツ生成の循環系」を構成し，そのループゲインを１以上にするための諸要因を実証的に解明することで，そのような状態を容易に実現するための仕組みの構築技術を確立しようとするものである．

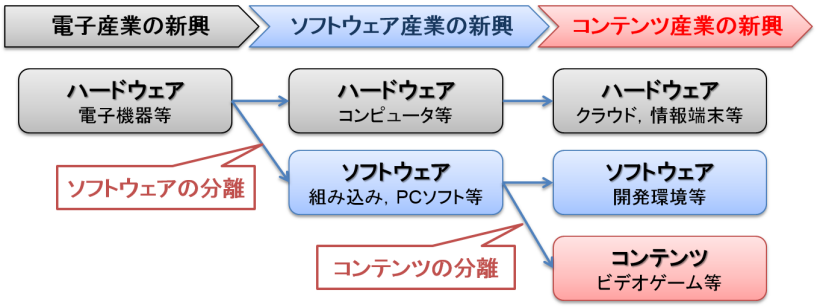

図 2. 情報通信に関連した産業構造の変遷

2 研究の背景と方針

2.1 産業構造との関係

ユーザが音声対話コンテンツを容易に生成・評価する仕組みを確立するためには，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．このことは，図 2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につなげるためには，幅広くより多くのクリエータ・ユーザが「コンテンツ生成」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．

2.2 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注目を集めている．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, Facebook, Twitter, Instagram 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（図 3）．

2.3 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デバイスとして，通常の PC や情報家電，スマートフォン等など様々な形態があり，デバイス特性や使用環境によって，生成される音声対話コンテンツも異なったものになると考えられる．本研究では，実証実験の一形態として公共空間に設置されるデジタルサイネージに着目した．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，近年は，近接センサー，タッチパネル，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いる．これをさらに発展させ，双方向の音声インタラクション機能を付与すること

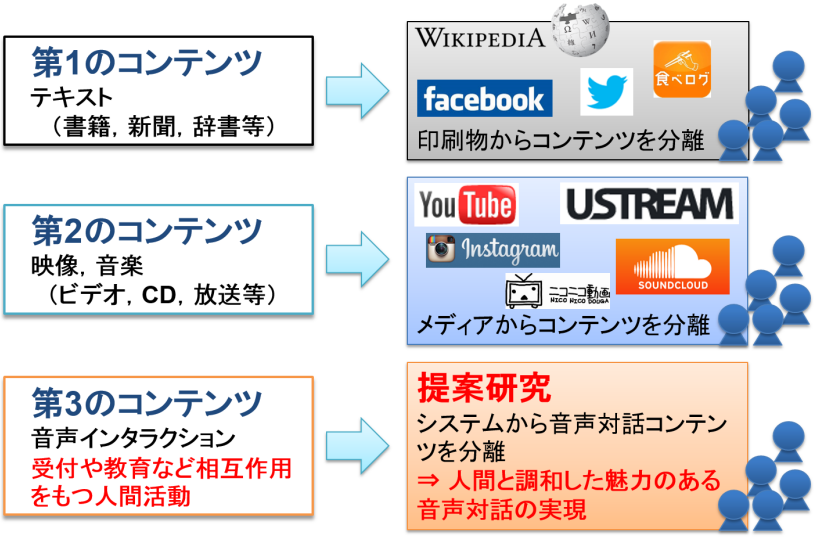

図 3. コンテンツの分離

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を大学キャンパス，観光案内所，市役所等に設置し，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，実証実験の基盤とした．

3　音声インタラクションシステム構築ツールキット MMDAgent

音声インターフェースにおける魅力等の諸要素を統合的に研究するためには，表情やしぐさ，声質，間合いなどの人間的かつ高度な処理が必要とされる領域までシステムが踏み込む必要がある．そのためには，音声処理系のみならず画像処理やエージェント表現部と密統合したプラットフォームが必要である．また，様々な状況やタスクにおいて実証実験を行い，データを蓄積する必要があるため，ユーザおよびシステム開発者がそれぞれ音声対話コンテンツおよびシステムの各部に自由に手が入れられるような，高度かつ柔軟なシステムでなければならない．

これまでに我々は，HMM 音声合成ツールキット HMM-based Speech Synthesis System (HTS), HTS をベースに構築された日本語テキスト音声合成システム Open JTalk, 音声認識エンジン Julius 等，音声技術に関連したオープンソースの研究基盤ソフトウェアを継続的に開発・公開してきた．これらのソフトウェア群をベースに，物理演算を含む高度な 3DCG によるエージェント表現をリアルタイムに描画可能なグラフィック表示部，および，有限状態トランスデューサ (Finite State Transducer; FST) に基づくインタラクション記述部を新たに実装し，密結合することで，「音声インタラクションシステム構築ツールキット MMDAgent」を構築し，オープンソースのフリーソフトウェアとして公開した（図 4）．音声認識や音声合成に用いる辞書やモデルだけでなく，エージェントの 3D モデルやそれを動かすモーションデータ，FST の定義にもオープンフォーマットを採用しており，既存のツールを用いてあらゆる構成要素をユーザあるいは開発者が自由に作成，

図 4. 音声インタラクションシステム構築ツールキット MMDAgent

編集し，入れ換えることが可能である．

MMDAgent は，音声技術の専門家だけではなく，一般の人々により広く音声技術を利用したものづくりを楽しんでもらうこと，また更には，魅力あふれる音声対話コンテンツを次々と作成してもらうことを強く意識した設計方針が採られており，これが MMDAgent の大きな特徴となっている．MMDAgent は，高速で高精度，かつ CG 表示を伴った表情豊かな音声対話を実現可能であり，本研究の技術基盤として用いることができる．

本研究では，音声インタラクションシステム構築ツールキット MMDAgent を土台に，音声対話システム全体をユーザに提供するコンテンツとそれを駆動するシステムに分離することにより，音声技術を広く人々の間に普及するための「魅力的なコンテンツ」,「魅力的なシステム」がそれぞれどのような要件を具備すべきかを解明することを目的とする．

本目的を達成するための課題は数多く列挙されるが，これらを以下の 3 つの課題にまとめた．

1. 基盤技術と基盤ソフトウェアの高度化
2. コンテンツ生成のための仕組みづくり
3. コンテンツ生成の実証実験

これらの課題に取り組むことで，ユーザが音声対話コンテンツを容易に作成・共有できる仕組みを確立し，音声対話コンテンツのユーザによる作成・共有・評価の中から魅力的な音声対話コンテンツが生成されるメカニズムを解明する．

4 基盤技術の高度化

「魅力的」な音声対話コンテンツを生成するためには，音声対話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．ここでは，音声認識，音声合成等の音声対話システムにおける基盤技術の高度化，音声対話システムを実現するための基盤ソフトウェアの高度化，及びコーパス／エージェントの設計について紹介する．

4.1 基盤技術

音声認識，音声合成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．以下で代表的なものを述べる．

- HMM 音声合成のための特徴抽出とモデリングの統合 [10]
  近年の音声合成では，音声の特徴と言語の特徴の関係性をモデル化する音響モデルに隠れマルコフモデル (hidden Markov model; HMM) という統計モデルが用いられる．本研究では，音声の特徴抽出と HMM の学習を統合することで，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モデル化が行われ，合成音声の品質を大きく改善した．
- 加算構造によるスペクトラムモデルの改良 [11]
  言語の特徴と音声の特徴の関係性に加算的な構造があると仮定し，音声合成のための加算モデルを提案した．また，加算モデルを効率的に学習するための学習アルゴリズムを提案した．加算モデルを用いたスペクトラムのモデル化によって合成音声の品質を大幅に改善した．
- 対話調の音声合成手法 [12]
  人と人との音声対話には，言い淀みやフィラー（会話の隙間を埋める「あー」等の発話）がしばしば存在する．そこで，音声対話における言い淀み等の影響について検証し，合成音声にこれらを自動挿入する手法を検討した．これにより，言い淀み等の挿入箇所を高精度に予測することがかのうとなり，より自然な対話調の音声合成が可能になった．
- リカレントニューラルネットワークに基づく言語モデルを用いた音声認識の高精度化 [13]
  高性能な音声認識のために，言語的な特徴をモデル化する言語モデルに，リカレントニューラルネットワークを用いる研究が進められている．本研究では，音声認識に適したニューラルネットワーク言語モデルの学習方法を提案し，音声認識精度を改善した．さらに，音響的な特徴を付加情報を利用可能にすることによって，より高精度な音声認識が可能となった．
- 音声インタラクションシステム構築ツールキット MMDAgent の対話記述言語の拡張 [20]
  MMDAgent の対話記述には有限状態トランスデューサを採用している．有限状態トランスデューサは状態遷移とシンボルの入出力で表現されるが，正規表現によるマッチングや変数の取扱いが可能となるように仕様を拡張することで，複雑な対話をより単純なスクリプトで記述することができるようになった．
- ディープニューラルネットワークに基づく音声合成の改善 [14, 15, 16]
  近年，音声認識だけでなく，音声合成にもニューラルネットワークを導入することで合成性能の大きい改善が報告されている．本研究では，音声合成の問題により適したニューラルネットワークの学習手法や，複数のニューラルネットワークを統合する手法を提案し，合成音声の品質を改善した．
- 発音情報が未知の言語におけるテキスト音声合成システムの構築法 [17]
  通常の音声合成は，テキストから読みを予測するテキスト解析部と読みから音声波形を生成する波形生成部からなるが，発音情報が未知の言語はテキスト解析部を構築することができない．そこで，異なる言語の音声認識システムを用いることによる，発音情報が未知の言語におけるテキスト音声合成システムの構築法を提案した．この構築法を用いることで様々な言語のテキスト音声合成システムを構築することが可能となった．

- 多様な声質を表現可能な因子分析に基づく音声合成の改善 [18]
  これまで，様々な声質の表現のために，音声合成に因子分析を導入した手法を提案してきた．本研究では，因子分析に基づく音声合成手法におけるモデル学習・モデル構造選択の改善に取り組み，合成音声の品質を改善した．
- トピックモデルを用いた音声対話コンテンツの分析
  音声対話コンテンツの作成／管理のために，トピックの検索や類似コンテンツの検出，人気コンテンツの推薦などができるような枠組みが必要である．音声対話コンテンツは多種多様であり，規則に基づくタグ付け等は困難である．本研究では，言語モデルの一種であるトピックモデルを応用することで，音声対話コンテンツの統計的に自動分類する手法を提案した．

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．

4.2 基盤ソフトウェア

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，以下のオープンソースソフトウェアを公開した[1]．

- 音声インタラクションシステム構築ツールキット MMDAgent
  (http://www.mmdagent.jp/)
  （61,000 ダウンロード）
- 音声認識エンジン Julius
  (http://julius.sourceforge.jp/)
  （216,000 ダウンロード）
- HMM 音声合成ツールキット HTS
  (http://hts.sp.nitech.ac.jp/)
  （371,000 ダウンロード）
- HMM 音声合成エンジン hts_engine API
  (http://hts-engine.sourceforge.net/)
  （41,000 ダウンロード）
- 日本語音声合成システム Open JTalk
  (http://open-jtalk.sourceforge.net/)
  （47,000 ダウンロード）
- 音声信号処理ツールキット SPTK
  (http://sp-tk.sourceforge.net/)
  （42,000 ダウンロード）
- 日本語歌声合成システム Sinsy
  (http://sinsy.sourceforge.net/)
  （1,400 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継続的に行っており，新バージョンを順次公開している．その中でも，音声インタラクションシステム構築ツールキット MMDAgent は，Windows, Mac OS, Linux, Android OS, iOS で動作可能な形で開発した．スマートフォンやタブレット単体で動作可能であり，応答遅延の少ない音声対話システムをスマートフォンやタブレットにおいて利用可能となった（図 5，6）．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は世界初の成果と考えられる．

[1]ダウンロード数は 2011 年 10 月から 2016 年 3 月までの累計である．

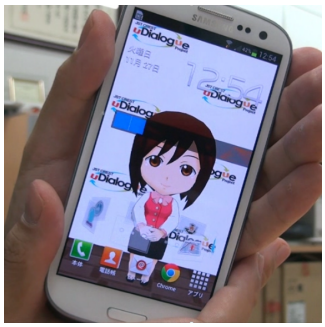

図 5. Android 版 MMDAgent

図 6. iOS 版 MMDAgent

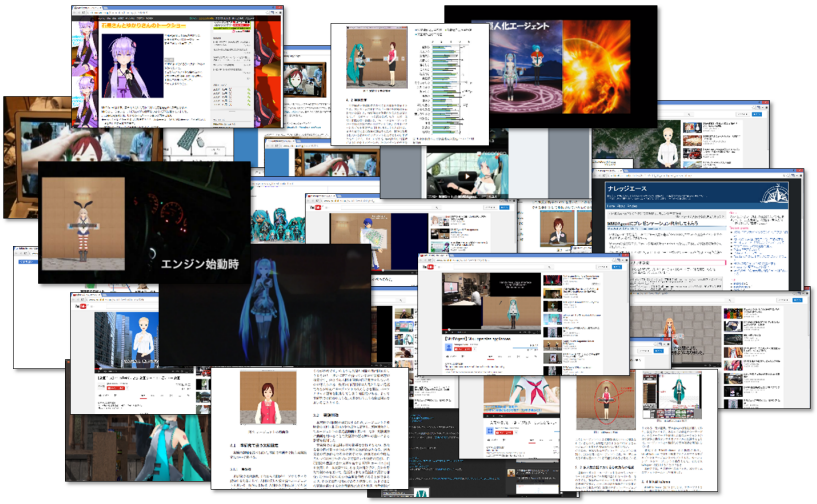

図 7. 公開したオープンソースソフトウェアの利用事例

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，ダウンロード数が示すように，既にデファクトスタンダードの一つとしての地位を確立している．実際に，図 7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．

4.3 コーパス／エージェント

言語に依存しない普遍的な知見を得るためには，様々な言語による実証実験を平行して進める必要がある．この際には，言語的な違いだけでなく，文化的な違いに関しても十分考慮する必要がある．そこで，日本語とイギリス英語を対象として音声対話システムを開発し，日本語と英語の音声対話システムによる実験を行うことにより，言語や文化に依存する点，依存しない点について検証を行った．

まず，音声対話システムの開発に必要な日本語音声コーパスとイギリス英語音声コーパスを構築した．

- 日本語による男性声優コーパス

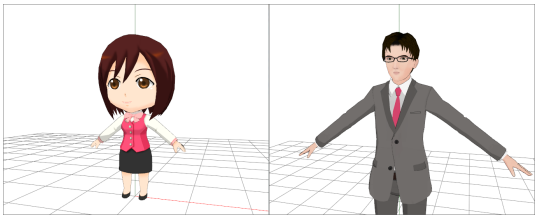

図 8. アニメティックな音声対話エージェント

図 9. リアリスティックな音声対話エージェント

- CSTR VCTK コーパス [19]
  本コーパスは，様々なイギリス英語アクセントをもつ話者 109 名による合計約 60 時間の大規模音声コーパスである．収録内容は各話者 400 文程度であり，新聞記事を含む複数のドメインから選択した．本コーパスは，イギリス英語研究用音声データベースとして現在標準的である WSJCAM0（LDC より有償で配布）の 1.5 倍の規模であり，今後様々な分野において多様な用途に利用されると期待される．
- イギリス・エジンバラ英語による女性声優コーパス
  エジンバラ訛りのイギリス英語を話すプロの女性声優による 4600 文を収録した．2 種類の話速の音声が収録されており，収録音声のうち，800 文は話速の早い音声，800 文は話速の遅い音声である．また，100 分に及ぶ自発的な会話を収録している．さらに，通常発話，早口発話などの 4 つの異なるスタイルの同一文を 400 文収録した．

次に，音声対話に適した対話エージェントを作成した．ユーザに受け入れられる音声対話システムにするためには，文化的な違いを考慮したシステムを作る必要がある．日本ではアニメティックな 3D エージェントに抵抗が少ないと考えられため，2.5 頭身のエージェントや，男性のエージェントを作成した（図 8）．一方，様々な議論の結果から，ヨーロッパ圏においてはリアリスティックな 3D エージェントが，より受け入れられやすいと考え，イギリス英語版の音声対話システム用の対話エージェントを作成した（図 9）．

5　コンテンツ生成および共有のための仕組み作り

本章では，音声対話コンテンツにユーザ生成の概念を取り入れて，音声対話コンテンツを作成および共有するための仕組みについて述べる．

音声対話コンテンツは，図 10に示すように，３つの階層から構成される．マテリアル層では，音声モデル・言語モデルなどといった専門的なモデルデータや，画像・音楽・３Dモデル・モーションなどのバイナリファイルが含まれる．アクション層では，一問一答形式の簡単な挨拶や，天気予報のパネルの表示など，一続きの短い動作が含まれる．シナリオ層では，アクション層の動作を組み合わせることによって，より複雑な音声対話シナリオを実現する．シナリオ層やアクション層は FST 形式で記述され，マテリアル層はそれぞれの素材の独自の形式で格納されている．

また，図 11が示すように，音声対話システムは単体で閉じているのではなく，複数のシステムが互いに連携しあうことも想定している．特に，スマートフォン版の MMDAgent はネットワークやスマートフォンとの連携を容易にする仕組みを実現した [9].

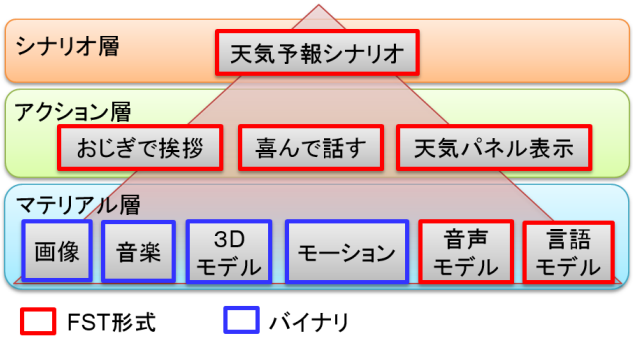

図 10. 音声対話コンテンツの階層

5.1 コンテンツ制作支援ツールの構築

音声インタラクションシステム構築ツールキット MMDAgent では，FST 形式のスクリプトファイルを記述することによって，音声対話シナリオの制御を行っている．しかしながら，FST 形式のスクリプトファイルを一般のユーザが直接記述することは難しい．そこで，音声対話コンテンツを手軽に編集できる仕組みを実現した．

対話スクリプト作成ツール

ユーザ生成型の音声対話コンテンツを実現するためには，ユーザがネットワークを介して，手軽に音声対話コンテンツを作成できることが重要である．そこで，我々は，初心者向けのタッチインタフェースを用いた音声対話コンテンツ作成ツール EFDE（図 12）と，Web ブラウザを用いた上級者向けの音声対話コンテンツ作成ツール MMDAE（図 13）を開発した．

- 音声対話コンテンツ作成ツール EFDE [7]
  初心者を対象とした Android 端末向けのインタフェースであり，状態遷移図で表された FST スクリプトをタッチインタフェースで編集可能である．編集結果をオンデマンドに実行可能にしたことにより，一般ユーザが手軽に音声対話コンテンツの動作を確認しつつ編集が可能である．また，よく使われる FST スクリプトをテンプレート化することにより，少ない状態数で複雑な音声対話を実現可能にした．さらに，音声による対話文やキーワードを入力可能な機能を追加することにより，音声とタッチインタフェースのみで音声対話コンテンツを容易に作成可能な仕組みである．
- 音声対話コンテンツ作成ツール MMDAE [1]
  上級者を対象としたインタフェースであり，Web ブラウザを用いた FST レベルの詳細な編集が可能である．FST スクリプトの入力補助と，FST スクリプトの構造化を実現することにより，容易な編集が可能になった．また，Web ブラウザを基盤として用いることで様々なプラットフォームで動作可能となり，ネットワークを介して，手軽にコンテンツを作成できる環境が整った．

音声インタラクションビルダ

ユーザによる魅力的で優秀なコンテンツの創造を推進するためには，音声対話コンテンツを詳細な部分まで自分の思い通りに作り込めるようなクリエイ

図 11. 音声対話システムとのネットワーク連携

ター向けの環境が必要である．そこで，インタラクティブな音声対話コンテンツを発話のタイミングや詳細な部分まで作りこみ，動作検証を行うことが可能な開発環境として音声インタラクションビルダを試作した（図 14）．この音声インタラクションビルダは，(1) 状態遷移図の 3 次元空間における可視化・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．

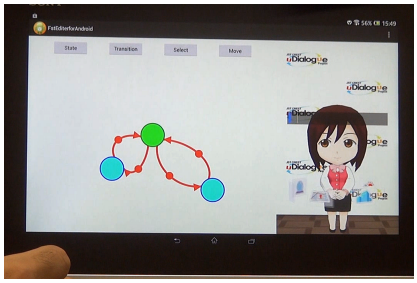

図 12. タブレットを用いた音声対話コンテンツ編集インタフェース

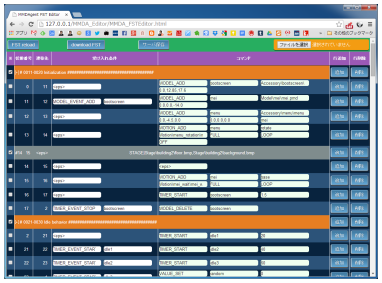

図 13. Web ブラウザを用いた音声対話コンテンツ編集インタフェース

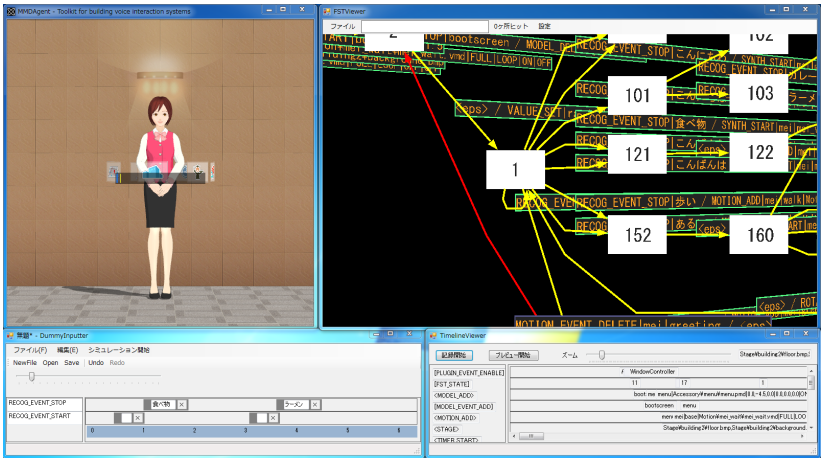

図 14. 音声インタラクションビルダ

5.2 クラウドでのコンテンツ共同編集環境の構築

ユーザ生成型メディアの特徴の一つは，他者とコラボレーションとして作品を構築することが頻繁に行われることである．音声対話コンテンツにおいても，ネットワークを介して音声対話コンテンツを共同作成・編集する環境を構築することで，他者と協力して大きく複雑なコンテンツを作り上げたり，既存のコンテンツを流用・拡張することが容易となる．本節では，クラウド環境を利用した音声対話コンテンツの共同編集環境の構築について述べる．

履歴付き音声対話コンテンツの共同編集システムの構築

音声対話コンテンツを共同構築する環境・システムの構成を考えるとき，最も簡便な形態の一つは，人工無能に代表されるような，想定質問とそれに対するシステムの応答の組をユーザが web 経由で自由に登録していく一問一答タイプのコンテンツ登録・共有である．しかし，この形式は履歴を伴わない１回きりの単純なやりとりしか扱えず，連続した会話やシナリオを持ったやりとりのような，対話本来の楽しさや魅力を提供するシステムとしては明らかに不十分である．一方，あらゆるユーザの応答を想定して長大な対話シナリオを定義するのは難しい．

そこで，履歴を持つ会話を容易に共同構築可能なユーザ生成型音声対話システムを開発した（図 15,16）．対話の登録単位はキーワードと応答の組であるが，親と子を設定できるようにすることで，複数回のやり取りを連続した一問一答としてユーザが協同構築できるようにした．登録された対話コンテンツ全体は，キーワード-応答のペアを単位とする木構造となる．登録画面では，誰でも対話の流れを直感的に把握できて編集できるよう，SNS のチャット風の Web インタフェースを採用するなどの工夫を施した．このシステムは，連続した長い対話を作り込んだり，他者が登録した対話に別の対話分岐や後続の対話を追加することができる．主観評価実験より従来法よりも大幅に高い面白さと期待感をユーザに与えることが示され，対話本来の魅力をもつシステムをユーザ生成的に構築することの可能性が示された．

クラウド方式の音声対話コンテンツの編集環境の構築

一般に MMDAgent における対話のボキャブラリーを増やすためには多数の対話シナリオを記述することが重要であるが，大規模な対話シナリオを一人

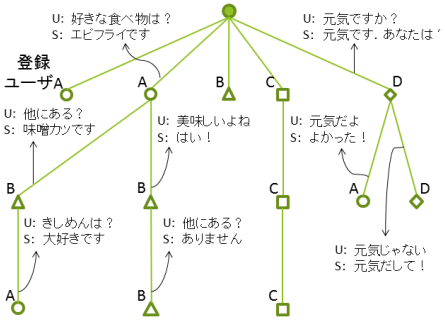

図 15. 履歴付き対話コンテンツの概念図

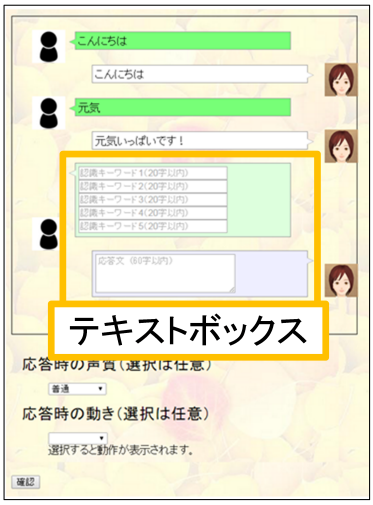

図 16. 履歴付き対話の登録インタフェース

で作成することは困難である．そこで，音声対話シナリオ作成に特化したクラウドソーシングシステムを開発した (図 17)．これはクラウドソーシングの概念に基づき，Web を介して複数人が対話シナリオ作成タスクを受発注するシステムである．これにより複数人の分担作業によるシナリオ作成が可能となった．また，対話シナリオ作成タスクを受注したユーザが手軽にシナリオ作成を行えるよう，前述の MMDAE と Skype 版エージェント動作確認ツールの連携機能を強化した．開発したシステムの評価実験を通して，クラウドソーシングに基づく対話シナリオ作成の有効性を確認した [2, 3].

5.3 モジュール化およびエージェント間連携基盤の構築

音声対話コンテンツのパッケージ化

一般に音声対話コンテンツはユーザや技術者が個々に作成しているのみで，それらを効果的に流通させるための手段はあまり考慮されていない．そこで，ユーザ生成の概念に基づき，ユーザが作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について提案した．まず，FST スクリプトを拡張し，複数並列化に基づく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 18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．

また，対話スクリプトだけでなく単語辞書やボイス，3D モデルやモーションといった音声対話コンテンツ全体の各構成要素をユーザが自由に組み替えて選択・構築できる環境の実現を目指し，ユーザ生成型音声対話システムのアーキテクチャ提案および構築を行った（図 19, 20）．音声認識や合成・対話管理・エージェントモデル等の全てのコンテンツを汎用モジュールとして統一的に扱い，パッケージとして動作の整合性やモジュール間の依存関係を扱えるようにする．実際に MMDAgent に対してモジュールをパッケージで管理する仕組みを提案し，対話スクリプト内のモデル名指定のようなモジュール名指定に対するスコープの設定や，パッケージ間の依存性や競合をヘッダ記述により自動的に検出する仕組みを実装した．

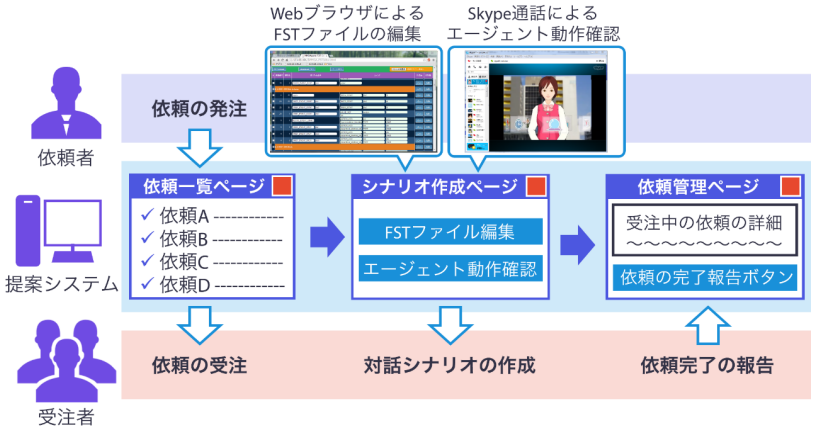

図 17. 対話シナリオ作成に特化したクラウドソーシングシステム

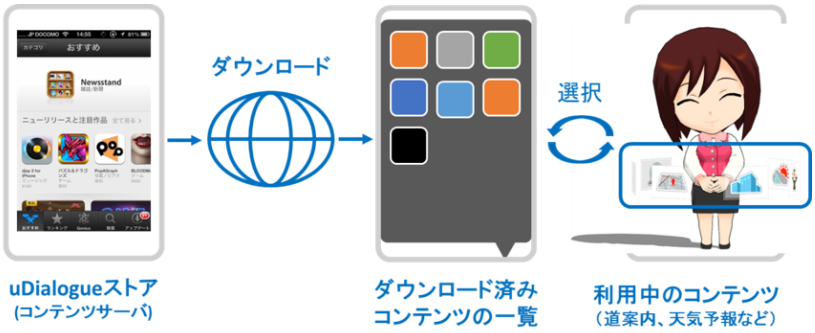

図 18. 音声対話コンテンツのパッケージ化

ネットワークエージェント技術を用いた音声対話システム間の連携基盤

既存のスマートフォン向け音声対話システムは，端末単体で動作するか，サーバとの通信しかできず，効率的に複数台の端末が連携した複雑な対話シナリオを構築することができなかった．そこで，下記のシステム連携基盤を導入した．

- 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を構築した．具体的には，エージェント/NFC/Bluetooth 技術に基づく対話システムのネットワーク接続機構を開発した (図 21) [4]．スケジュール調整システムを試作し，複数対話システム連携型サービスの有効性を確認した．
- 複数のサイネージ上の MMDAgent が協調動作するための環境を構築した．この環境は，多数のサイネージの接続を考慮し規模拡張性にも優れた設計となっている．対話回数共有システムを試作し，サイネージ間システム連携の有効性を確認した．

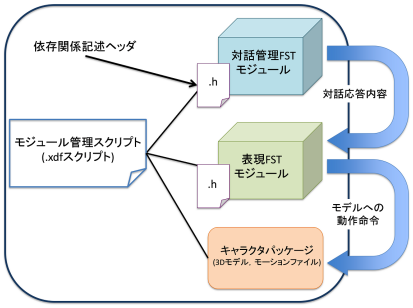

図 19. 音声対話モジュールのパッケージ化

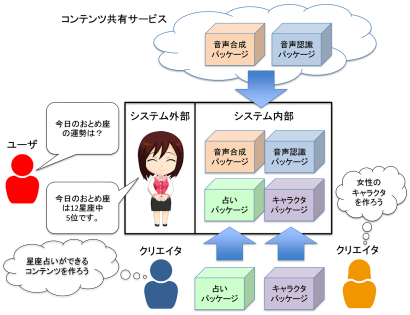

図 20. モジュールベースの音声対話コンテンツ共有

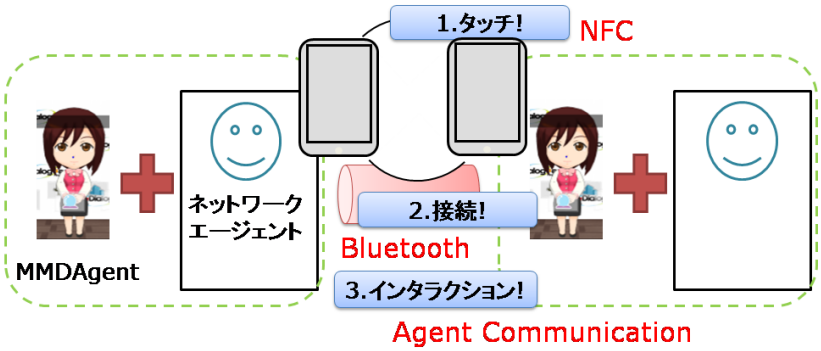

図 21. P2P 型ネットワーク接続機構

6　コンテンツ生成の実証実験

本章では，提案システムを用いた実証実験について述べる．

6.1　コンテンツ利用を促進する仕組みづくり

ユーザ生成型の音声対話コンテンツでは，ユーザによる多種多様で多量の音声対話コンテンツが作成されることが期待できる一方で，それらのコンテンツ群を適切に分類・関連付けすることができなければ，それらの利用価値を最大限に高めることは困難である．ここでは，コンテンツの推薦や検索等の技術を実現し，音声対話システムの利用支援や作成支援等の実現するための音声対話コンテンツ群の分析・分類・関連付け手法の研究成果について述べる．

音声対話システムの利用履歴に基づく関連ワード推薦

音声対話システムの利用に不慣れな初心者に対する対話システムの利用促進を目指して，ユーザに次に話すべきキーワードの候補群を提示する仕組みを検討した．そして，サーバに蓄積されたユーザの対話履歴から，情報推薦技術等を用いて分析し，関連ワードを獲得する手法を実現した（図 22）．被験者実験より，関連ワードを提示することにより，ユーザの音声対話システムの円滑な利用を支援することを示唆する結果を得た．また，関連ワードの推

図 22. 関連ワード提示手法

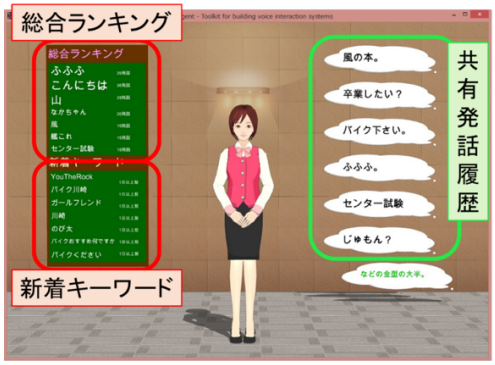

図 23. 発話履歴の共有と音声対話コンテンツのランキング提示

薦手法として，履歴ベース推薦，内容ベース推薦，ランキング推薦，ランダム推薦の 4 つの方式，および，それらを統合した複合推薦の方式を比較検討した．被験者実験により，複合推薦方式が内容ベース推薦やランダム推薦単体よりもよい総合結果が得られることを確認した．

利用履歴のユーザ間共有による相互インセンティブの強化

ユーザ生成型メディアにおいては，ユーザ間で活動履歴がシェアされることでユーザが互いに刺激を受け，システムの利用やコンテンツの作成がユーザ間で誘引される現象が見られる．音声対話コンテンツにおいても，対話システムとのやりとりやコンテンツの利用履歴をユーザ間で共有することで，ユーザ同士で相互に利用意欲を高めることが期待される．このため，一問一答型のユーザ参加型音声対話システムにおいて，(１）ユーザが何を発話しているかのオンライン逐次共有，(２）対話の再生回数やキーワード発話回数を記録しランキング表示，によって相互にインセンティブを強化する仕組みを構築した．これらを図 23に示すように音声対話コンテンツの利用者と作成者に提示することにより，利用者が必要としているキーワードを作成者が知ることが可能になった．また，コンテンツ作成者が，ランキング結果に基づきユーザが必要としている対話を手軽に登録する仕組みも実現した．被験者実験により，これらの仕組みが利用者とコンテンツ作成者の両方のモチベーション（インセンティブ）の向上に繋がる可能性を確認した．

6.2　公共空間での実証実験

提案システムを名古屋工業大学正門前や半田市観光協会など，様々な場所で設置・運用することにより，提案システムの社会実装の可能性を検証した．また，これらの実証実験で収集された音声対話コンテンツは，「魅力的な音声対話コンテンツ」の実現法を解明するためのデータセットとして利用可能である．以下で代表的なものを述べる．

学内での実証実験

我々が開発した音声インタラクションシステム構築ツールキット MMDAgent を用いて構築した「全天候型双方向音声案内デジタルサイネージ（メイちゃん）」を，大学正門入ってすぐのオープンスペースに設置した（図 24, 25）．複数のカメラを用いた顔画像認識技術によるキャラクターの視線制御や，焦

図 24. 双方向音声案内デジタルサイネージ「メイちゃん」

図 25. 名古屋工業大学正門

図 26. 情報基盤システムとの連携

電センサーを用いた人体検知によるキャラクター側からの能動的な呼びかけ等の機能が実現されている．表示テキストや画像，発話テキストに加えて，案内時のキャラクターのモーションや発話スタイルまでも統合したコンテンツを，サーバから動的に更新することが可能であり，情報基盤センターのデータベースに登録された学内イベント情報の中から，季節や時間，内容に応じて適切に取捨選択された情報がタイムリーに案内される（図 26）．想定されるユーザは，来学者，学生・教職員である．

2011 年 4 月 6 日にシステムの稼働を開始し，2011 年 6 月 15 日にコンテンツ登録システムを学内に公開した．2011 年 11 月 15 日までに 100 件以上のコンテンツが登録されており，1 日当たりのユーザー発話も，休日を含めた平均も約 350 発話を超えている．学生や教職員から 243 件（内学生 43 件）の音声対話コンテンツの投稿があった．図 27に投稿されたパネル画像の例を示す．

また，2014 年 9 月からは，新たに名古屋工業大学のオープンスペース（夢ルーム）にも設置した（図 28, 29）．学生スペースに設置した提案シス

図 27. パネル画像

テムは，PC だけでなく NFC を用いてスマートフォンからも容易に音声対話コンテンツを投稿可能である．設置から半年の間に学生などから計 437 件の音声対話コンテンツの投稿があった．

また,「メイちゃん」について，学内での知名度や使用頻度などについて，Moodle のアンケートシステムを用いて調査を行った．この結果，262 件の有効回答があり，メイちゃんの知名度は 99%とほぼ全員が認知していた．また，対話経験も 77%と高く，MMDAgent を利用した音声対話システムの実用性の高さを示している．自由記述アンケートへの回答率は 33%であり，三分の一のユーザが積極性を持って，音声対話システムの改良点などについての意見を寄せた．

学外での実証実験

2013 年度は，新たに，室内型双方向音声案内デジタルサイネージを試作し，半田市観光協会 蔵のまち案内所（国登録有形文化財小栗家住宅内）に設置した（図 30, 31）．小さいキャラクターを画面下部に置き，画面上部にパネル情報等を提示することによって，狭い室内でも快適に利用可能になるように工夫した．利用者は，観光客を中心とした一般の来訪者である．

また，スマートフォンや Web ブラウザを用いてユーザ（ここでは，観光案内所の職員）が簡単に音声対話コンテンツを追加・更新できる仕組みを実現することで，ユーザの創意工夫で，現場に即した，より面白い音声観光案内が実現可能になった．実際に，観光協会の職員によって，多数の観光案内に関するコンテンツが登録されている．なお，本実証実験システムは，テレビや新聞などで紹介される，半田市の公式広報誌の表紙に掲載されるなど，観光 PR に繋がるものとして地域からも注目を集めた．

図 28. 名古屋工業大学オープンスペース（夢ルーム）

図 29. 名古屋工業大学オープンスペース（夢ルーム）

図 30. 半田市観光協会蔵のまち案内所

図 31. 半田市観光協会蔵のまち案内所（外観）

さらに，同様のシステムを，半田市市役所新庁舎や NHK 名古屋放送局，国立情報学研究所などに音声対話システムを設置した（図 32, 33, 34）．NHK 名古屋放送局では TV 番組やイベント等において実証実験を実施した．その際，複数のキャラクターを連携させて動作できるようにするなどの拡張を行った．国立情報学研究所では，国立情報学研究所のキャラクター「ビット」の 3D モデルを利用した音声対話システムを構築した（図 34）．このキャラクターに合わせた音声データベースも構築し，キャラクター声による音声合成システムを構築した．

その他の実証実験

モバイル環境における音声対話システムとユーザ生成型音声対話コンテンツ作成環境に関する研究を実施した．

- iPhone などのスマートフォンを用いて CG キャラクターと音声による対話を可能にするシステムを開発した（図 35）[5]．本システムは，Skype のビデオ電話機能と MMDAgent を連携させることによって実現している．さらに，第 74 回情報処理学会全国大会において公開実験とアンケート調査を実施し，120 名分のアンケート結果から，モバイル向け対話システムはデジタルサイネージ版より対話の応答速度は劣るものの，親しみやすさが向上していることが判明した．
- 中学生を対象として，音声対話に基づくモバイル学内見学支援システ

図 32. 半田市市役所

図 33. NHK 名古屋

図 34. 国立情報学研究所

図 35. Skype

図 36. モバイル学内見学支援システム

ムの実証実験を 2014 年 1 月に行った (図 36) [6]．スマートフォン上の音声対話エージェントが，中学生に対して名古屋工業大学のキャンパスや設備に関して音声案内を行うシステムである．なお，実験で用いた音声対話コンテンツは，名古屋工業大学の案内担当の職員 2 名によって作成された．主観評価実験の結果，提案システムに基づく音声案内の有効性が示された．

- より一般の人に音声対話システムを利用してもらうためには，公衆電話網等の連携が有用である．そこで，MMDAgent と VoIP クライアントを連携させる仕組みを開発した [8]．学内に整備・展開された Skype for business（Lync の新バージョン）に接続するためのミドルウェアを作成し接続した．同時に，名古屋工業大学ユニファイドコミュニケーションシステムの一部である Voice over IP による新しい内線電話システムとも接続するために，Software Phone 連携用ミドルウェアを開発し，相互接続試験を実施した．これらの仕組みを発展・評価すると同時に，学内電話網との連携の仕組みを検討した．

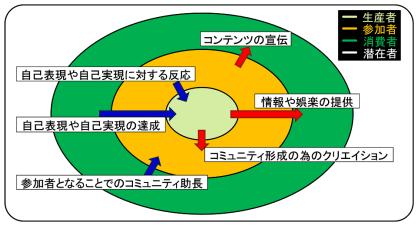

図 37. ユーザ層の分類とインセンティブ

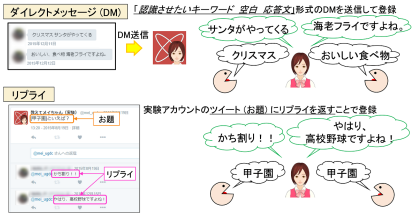

図 38. お題提供型対話コンテンツ生成

6.3 ネットワーク環境での実証実験

より幅広い社会環境におけるユーザ生成型音声対話コンテンツの実現のための手がかりを得るために，一般のネットワーク環境へ広げて実証実験を行った．また，その前段階として，音声対話コンテンツが実際に社会でユーザ生成型メディアとして形成され成長していくために必要な要件の一つである，ユーザの動機および誘引要素（インセンティブ）に関して分析を行った．

まずユーザの利用動機および誘引要素の分析については，ユーザをコンテンツへの関わり度合いから以下の 4 種類の層に分類し，それぞれの相互関係と移動の関係を分析した（図 37）．

- 潜在者 — システムを全く利用したことがない人．
- 消費者 — システムを利用する・したことがある人．受け身なユーザ．
- 参加者 — システムを積極的に利用しコメントやお気に入り，評価を下す人．アクティブなユーザ．
- 生産者 — コンテンツを作成・投稿する人．

そして，ネットワーク環境において音声対話コンテンツがユーザ生成型メディアとして成立するために音声対話システムが満たすべき要件を以下のように定めた．

- 潜在者から消費者へ：スマートデバイス等マルチデバイス対応による音声対話コンテンツ再生環境の拡張
- 消費者から参加者へ：ユーザ動線設計や SNS での定期的な情報提供によるシステム全体への興味の喚起
- 参加者から生産者へ：Twitter のリプライ機能を用いた「お題」提供型対話コンテンツ生成（図 38），

上記の改善を施したマルチプラットフォームの一問一答型対話コンテンツ配信・登録システムを構築し，社会実験を実行した（図 39）．音声対話コンテンツを保持するコンテンツサーバを学外に向けて公開し，マルチプラットフォーム（Windows/Linux/Android）でのシステムを提供した．実際に一般ユーザを誘引するための動線設計やサポート体制を整え，1ヶ月間の公開実験を行った．Twitter のリツイートによって約 30,000 人への周知を行い，6,400 回の発話と 232 個の音声対話コンテンツ新規登録を得た．利用者の 55% は Android からの参加者であり，15% が Twitter のリプライ機能を用いた「お題」提供に参加しライトなコンテンツ登録を行うなど，提案法の有効性が確認でき，まだユーザ動向に関する貴重な資料を収集することができた．

6.4 音声対話コンテンツ共有サービス

音声対話コンテンツの製作者と利用者の間に音声対話コンテンツを共有する枠組みが用意されておらず，作成された音声対話コンテンツは様々な場所に分散した状態になっているのが現状であった．そこで，音声対話コンテンツ

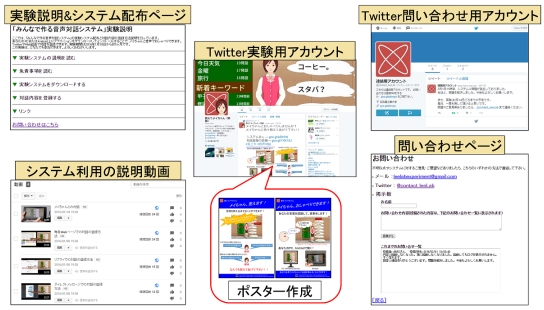

図 39. 社会実験における一般ユーザ向けの説明ページおよび参加動線の設計

を発表・共有する場を提供することで，より多くの音声対話コンテンツを生成することに繋がると考えた．具体的には，音声対話コンテンツをユーザ間で共有可能にするための仕組みとして，音声対話コンテンツ共有サービスを立ち上げた（図 40, 41）．一般に，音声対話コンテンツを容易に共有できるようにするためには様々な課題が考えられるが，提案システムは以下に挙げる方法で解決した．

1. 音声対話コンテンツは多数のファイル群から構成されるなど扱いが煩雑であったため，独自の MMDA 形式を定義し，1 ファイルで利用可能にしている．
2. MMDA 形式のファイルをサーバ側で容易に生成できる機能を作成している．
3. Windows 用の MMDAgent インストーラを新たに開発し，MMDA 形式のコンテンツを 1 クリックで実行可能にしている．
4. サービスの利用を容易にするために Google などのオープン ID との連携を可能にしている．
5. コンテンツの投稿時に自動的にスクリプトのエラーやワーニングを検出する仕組みを開発している．
6. 本サービスの利用規約やプライバシーポリシーを慎重に作成している．
7. コンテンツに含まれるすべての付属文章（README などを）を自動的に本サービスから確認できるようする仕組みを開発するなど，著作権に最大限配慮している．

また，音声対話コンテンツを手軽に Web で作成するためのサービスも試作した．これは，ユーザが一般的な Web ブラウザを用いて，画像やキーワード，文章などを入力していくことで，手軽に音声対話コンテンツを作成するための仕組みである（図 42）．

7　エンサイクロペディア

これまでに開発してきた音声インタラクションシステム構築ツールキット MMDAgent を中心とした基盤ソフトウェア群や音声対話コンテンツ制作支援ツール群だけではなく，キャンパス内外における複数の実証実験，解説論文執筆，MMDAgent 講習会，インターネット上のブログ形式による TIPS の発信等を通して，MMDAgent および音声対話コンテンツ制作に関する指導書・手引書，スライド，リファレンス・マニュアル，サンプルスクリプト等が充実してきた．これらの成果物を統合し，さらに拡張・整備することにより，ソフトウェア群，マニュアル，基本音声対話ライブラリ，設計指針等を

図 40. 音声対話コンテンツ共有サービス

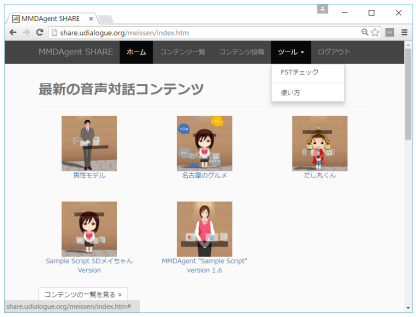

図 41. 共有された音声対話コンテンツ

図 42. 音声対話コンテンツ作成サービス

総合的・統合的にまとめたオールインワン・パッケージ「エンサイクロペディア MMDAgent」を構成した．パッケージは以下の項目を含むものとなる．

1. 音声インタラクションシステム構築ツールキット MMDAgent（マルチプラットフォーム対応）
2. MMDAgent 概説（日英）
3. MMDAgent クリエータ（コンテンツ制作者）向けリファレンス・マニュアル（日英）
4. MMDAgent ディベロッパ（開発者）向けリファレンス・マニュアル（日英）
5. MMDAgent 講習用スライド（日英）
6. MMDAgent 講習用動画（MOOC, OCW）
7. 音声対話コンテンツ制作支援ツール群（Web ブラウザ上で利用可能な音声対話コンテンツ編集ツール，タブレット上で利用可能な音声対話コンテンツ編集ツール，音声インタラクションビルダ，他）
8. 音声対話コンテンツライブラリ（基本対話ライブラリ，サンプル用 3D

モデル，音声合成用モデル等を含む）
9. 音声対話コンテンツ設計指針

本パッケージは，マルチプラットフォームに対応したソフトウェア，ユーザによるコンテンツ制作を支援するツール群，長期運用に基づいた実証実験に裏付けされた音声対話コンテンツの設計指針に，コンテンツ制作者・開発者向けリファレンス・マニュアルや講習用スライド，moocs・OCW 用チュートリアル動画，音声対話コンテンツライブラリを加え，有機的に統合したものであり，本パッケージによりユーザコンテンツ生成環境を容易に構築可能とするものである．

8 むすび

本研究では，音声対話システム全体をユーザに提供するコンテンツとそれを駆動するシステムに分離することにより，ユーザ，あるいはユーザに近いクリエータにより，魅力的な音声対話コンテンツが次々と生成されることを期待し，基盤技術の高度化，コンテンツ生成のための仕組みづくり，コンテンツ生成の実証実験に取り組んだ．公開したソフトウェアツール群のダウンロード数，インターネット上での多様な利用事例から，多くのコンテンツが生成される様子が観測された．得られた成果を統合的にまとめた「エンサイクロペディア MMDAgent」を作成し，公開した．更に，これらのコンテンツを分析するためにコンテンツ共有サーバーを構築した．今後，コンテンツ共有サーバーの運用により，魅力的なコンテンツが数多く生成されるとともに，更なる知見が得られるものと期待される．

本研究は，ユーザによる音声対話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ただけでなく，今後，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．実際に最近の音声対話システムでは，ウィットの効いたユーザとのやり取りによりその魅力をアピールするものが増えており，このようなコンテンツはシナリオライターと呼ぶべき人たちが制作しており，このような状況は本研究が想定するところと符合している．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．将来的に，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れば，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も可能と期待される．

References

[1] R. Nishimura, D. Yamamoto, T. Uchiya, and I. Takumi, “Development of a Dialogue Scenario Editor on a Web Browser for a Spoken Dialogue System,” Proceedings of the Second International Conference on Human-agent Interaction, pp. 129–132, 2014.

[2] Y. Matsushita, T. Uchiya, R. Nishimura, D. Yamamoto, and I. Takumi, “Crowdsourcing Environment to Create Voice Interaction Scenario of Spoken Dialogue System,” Proc. of the 18-th International Conference on Network-Based Information Systems (NBiS-2015), pp.500–504, 2015.

[3] Y. Matsushita, T. Uchiya, R. Nishimura, D. Yamamoto, and I. Takumi, “Experiment and Evaluation of Crowd Sourcing Model for

Creation of Voice Interaction Scenario," Proc. of the IEEE GCCE 2015, pp.321–322, 2015.

[4] T. Uchiya, R. Nakano, D. Yamamoto, R. Nishimura, and I. Takumi, "Extension with Intelligent Agents for the Spoken Dialogue System for Smartphones," Proc. of the IEEE GCCE 2015, pp.298–299, 2015.

[5] T. Uchiya, D. Yamamoto, M. Shibakawa, M. Yoshida, R. Nishimura, I. Takumi, "Development of Spoken Dialogue Service based on Video Call named "Mobile Meichan"," Proceedings of JAWS2012, 2012. (in Japanese)

[6] T. Uchiya, M. Yoshida, D. Yamamoto, R. Nishimura, and I. Takumi, "Design and Implementation of Open-Campus Event System with Voice Interaction Agent," International Journal of Mobile Multimedia, Vol.11, No.3,4, pp.237–250, 2015.

[7] K. Wakabayashi, D. Yamamoto, and N. Takahashi, "A Voice Dialog Editor Based on Finite State Transducer Using Composite State for Tablet Devices," Computer and Information Science 2015, Studies in Computational Intelligence, Vol.614, pp.125–139, 2016.

[8] R. Nishimura, K. Sugioka, D. Yamamoto, T. Uchiya, and I. Takumi, "A VoIP-based Voice Interaction System for a Virtual Telephone Operator Using Video Calls," Procceedings of the IEEE GCCE 2014, pp.529–532, 2014.

[9] D. Yamamoto, K. Oura, R. Nishimura, T. Uchiya, A. Lee, I. Takumi and Keiichi Tokuda, "Voice Interaction System with 3D-CG Human Agent for Stand-alone Smartphones," Proceedings of the 2nd International Conference on Human Agent Interaction, ACM digital library, pp.320–330, 2014.

[10] K. Nakamura, K. Hashimoto, Y. Nankaku, and K. Tokuda, "Integration of spectral feature extraction and modeling for HMM-based speech synthesis," IEICE Transactions on Information and Systems, vol.E97-D, no.6, pp.1438–1448, 2014.

[11] S. Takaki, Y. Nankaku, and K. Tokuda, "Contextual partial additive structure for HMM-based speech synthesis," 2013 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp.7878–7882, Vancouver, Canada, 2013.

[12] R. Dall, M. Tomalin, M. Wester, W. Byrne, and S. King. "Investigating automatic & human filled pause insertion for speech synthesis." Proceedings of Interspeech, 2014.

[13] S. R. Gangireddy, S. Renals, Y. Nankaku, and A. Lee, "Prosodically-enhanced Recurrent Neural Netwrok Language Models." Proceedings of Interspeech 2015, Dresden, Sep. 2015.

[14] K. Hashimoto, K. Oura, Y. Nankaku, and K. Tokuda, "The effect of neural networks in statistical parametric speech synthesis," Proceedings of 2015 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2015), pp.4455–4459, Brisbane, Australia, April 19–24, 2015.

[15] S. Takaki, S. Kim, J. Yamagishi, and J.J. Kim, "Multiple feed-forward deep neural networks for statistical parametric speech synthesis," Proceedings of Interspeech 2015, pp.2242–2246, 2015.

[16] K. Hashimoto, K. Oura, Y. Nankaku, and K. Tokuda, "Trajectory training considering global variance for speech synthesis based on neural networks," Proceedings of 2016 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2016), pp.5600–5604, Shanghai, China, March 20–25, 2016.

[17] K. Sawada, K. Hashimoto, K. Oura, Y. Nankaku, and K. Tokuda, "Evaluation of text-to-speech system construction for unknown-pronunciation languages," Technical Report of IEICE, vol.115, no.346, SP2015-80, pp.93–98, December 2–3, 2015.

[18] T. Yoshimura, K. Hashimoto, K. Oura, Y. Nankaku, and K. Tokuda, "Simultaneous optimization of multiple tree structures for factor analyzed HMM-based speech synthesis," Proceedings of Interspeech 2015, pp.1196–1200, Dresden, Germany, September 6–10, 2015.

[19] CSTR VCTK Corpus
`http://www.udialogue.org/ja/download-ja.html`

[20] T. Funayachi, K. Oura, Y. Nankaku, A. Lee, and K. Tokuda, "A simple dialogue description based on finite state transducers for user-generated spoken dialog content," Proceedings of ASJ 2013 autum meeting, 2-P-28, pp.223–224, September 25–27, 2013. (in Japanese)