Vol.2015-SLP-109 No.21
2015/12/3

Web ユーザによる音声対話コンテンツ生成環境の構築とそれに基づく実証実験の評価

山本 大介　堤 修平　打矢 隆弘　内匠 逸[†1]

概要：我々は、Web ブラウザ等を用いて多数のユーザらが手軽に参加可能な音声対話コンテンツ生成環境を構築した．本システムでは、複数のユーザが画像の表示や 3D キャラクタのモーション・音声の声色なども設定可能な一問一答形式の対話を投稿可能であり、生成されたコンテンツは、既存の音声インタラクションシステム構築ツールキットである MMDAgent で利用可能である．提案システムを、名古屋工業大学の学生スペースに 1 年以上設置し、その有効性について検証した．また、同様のシステムを、観光協会、市役所や放送局のイベント、学会の大会等での運用も実施したので、それらの有効性についても報告する．

キーワード：音声対話システム, Web サービス，ユーザ生成型コンテンツ

Web-based user generating environment for voice dialogue content and its evaluation experiments

Daisuke Yamamoto　Syuhei Tsutsumi　Takahiro Uchiya　Ichi Takumi[†1]

Abstract: We developed the web-based user generating environment for voice dialogue contents. This system enables users to edit a question and answers-type dialogues, which can be selected 3D model's motion and voice styles. This system can generate a voice dialog content for a toolkit for building voice interaction systems MMDAgent. We opened the proposed system more than 1 year at student room in Nagoya Institute of Technology. Moreover, our systems were installed at a tourism center, city hall, broadcasters, and society events. We report these open examinations.

Keywords: Voice dialogue system, Web service, User generated content

1. はじめに

近年、音声信号処理技術の発展に伴い、Apple の Siri[1]など多くの音声対話システムが実用化されている．しかしながら、現状の音声対話システムの多くは技術者や研究者が音声対話シナリオや関連する素材(本論文では、これらをまとめて音声対話コンテンツと呼ぶ)を作成している場合が多いのが現状である．技術者や研究者が音声対話シナリオを記述することは、その知見を活かして複雑なシナリオを適切に記述できる利点がある一方で、以下に述べる問題点が挙げられる．

1) 必ずしも、利用者視点での音声対話コンテンツを作成できるとは限らない．技術者や研究者の思い込みにより、利用者の要求を正確に反映できないことがある．

2) 時間的・地理的制約により利用者や管理者からの要望があってもすぐに対応できるとは限らない．

3) 音声対話システムが広く普及した場合、技術者や研究者の人的リソースの問題から、音声対話コンテンツの内容の全てを作成していくことは難しい．

そこで、音声対話システムを管理する現場の事務員(管理者)や、音声対話システムを利用する人(利用者)に音声対話シナリオの記述を開放すれば、上記の問題を解消できると考えた．また、音声対話コンテンツを作成する人の層が厚くなるため、より魅力的な音声対話コンテンツが生まれやすくなるだろう．いわゆる、YouTube やブログなどに代表されるユーザ生成型コンテンツの概念を音声対話コンテンツに適用したと言い換えることもできる．

本論文の目的は、ユーザ生成型コンテンツの概念を音声対話コンテンツに適用し、管理者や利用者などの一般の人が手軽に音声対話コンテンツを生成できる環境を構築し、その効果を検証することにある．そのためには、いかに容易に、かつ、ユーザ満足度の高い音声対話コンテンツをつくれる環境を用意できるのかが重要であると考えた．

提案システムを実現する上で直面した課題とその解決法について以下に述べる．

課題 1．不特定多数のユーザが、音声対話コンテンツを編集可能にする必要がある．その際、声色やモーションなどを組み合わせた魅力的な音声対話が実現できると望ましい．

課題 2．一つの音声対話コンテンツを複数のユーザで編集すると、コンテンツの整合性を保つのが難しい．たとえば、二重編集やデッドロックなどの問題が発生する．

†1 名古屋工業大学
Nagoya Institute of Technology

課題 3. ユーザ投稿型の音声対話システムでは、多様で多くの対話が投稿されることが期待されるが、その個々の対話の内容を利用者が認識する方法がない.

解決法 1. Web サービスの技術を用いて音声対話コンテンツを編集する仕組みを提案. (課題 1 に対応)

解決法 2. 一問一答方式の対話に限定することにより、課題 2 の問題を回避.

解決法 3. 認識可能なキーワードを示唆するバルーンパネルを生成することにより、課題 3 を回避.

音声対話システムの利用環境としては、音声インタラクションシステム構築ツールキット MMDAgent[2]を採用した.

提案システムを用いた実証実験を複数実施することにより、より実践的・実証的な研究を実施した. 良いコンテンツ無しで良い音声対話システムが構築できないという理念のもと、良い音声対話コンテンツをユーザがいかに作成するかに着目した.

2. 音声インタラクションシステム構築ツールキット MMDAgent

音声インタラクションシステム構築ツールキット MMDAgent は、音声対話システムを実現するために必要な機能、たとえば、音声合成、音声認識、3D モデル描画、物理演算に基づく 3D モデル制御、対話制御などを統合したシステムである. Windows や Mac OS、Linux、Android などの PC[3]やスマートフォン[4]での動作が可能である. 音声合成エンジンとして OpenJTalk を、音声認識エンジンとして Julius を、3D モデル形式として MikuMikuDance 形式を、物理演算エンジンとして Bullet Physics を採用している. これらの複数の機能はプラグインとして独立性の高い形式で実装されており、Global Message Queue を介して、互いにメッセージをやり取りすることにより連係動作する.

MMDAgent は、FST(Finite State Transducer)形式の音声対話スクリプトと関連する素材、および、音響モデル・言語モデルに基づいて動作する. これらの音声対話スクリプトと関連する素材・モデルをまとめて音声対話コンテンツと呼んでいる. MMDAgent では、複数の FST スクリプトを並列・独立的に動作させることができる. ただし、複数の FST スクリプトを並列的に動作させる場合には、資源やメッセージの干渉によるデッドロック等が起きる可能性があるので注意深く設計する必要がある.

FST スクリプトによる対話の記述例を図 1 に示す. FST スクリプトでは、各機能部(音声認識部、音声合成部など)から発生するイベントを入力、各機能部への命令コマンドを出力とした、状態遷移機械として記述可能である. 図 1 の例では、音声認識機能部が「こんにちは」と認識すると、状態が 1 から 10 に変化する. 続けて、モーション開始コマンド(MOTION_ADD)、音声合成開始コマンド(SYNTH_START)を出力し、状態 12 まで一気に遷移する. 合成音声の再生が終了するまで状態 12 で待機状態になり、音声合成終了イベント(SYNTH_EVENT_STOP)が発生すると状態 1 に遷移する.

FST スクリプトは、オートマトンと同等であると考えられるので、シーケンシャルな対話制御だけでなく、割り込み(バージインなど)や、文脈に応じた処理など、複雑な制御が可能である. その一方で、人手でこれらの複雑なスクリプトを多量に記載することは困難が予想され、データベースと連携した自動生成や、専用のツールの開発、およびそれらの組み合わせなどの手法を開発する必要がある.

```
 1  10    RECOG_EVENT_STOP|こんにちは <eps>
 1  10    RECOG_EVENT_STOP|おはよう   <eps>
10  11    <eps>                       MOTION_ADD|mei|greet|greet.vmd
11  12    <eps>                       SYNTH_START|mei|normal|こんにちは
12  1    SYNTH_EVENT_STOP|mei        <eps>
```

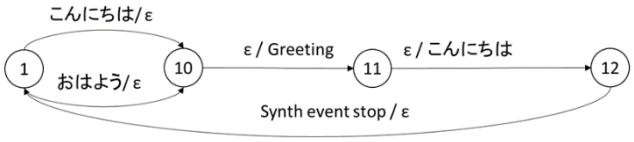

図 1 FST スクリプトの例（上）とその状態遷移図（下）
Figure 1 FST Script (top) and its state transition diagram (bottom)

3. 提案システム

3.1 システム構成

提案システムは、図 2 に示すように、音声対話共有サービス(Web サーバ)と、FST 生成機能からなる.

音声対話共有サービスでは、スマートフォンや PC に搭載されている一般的な Web ブラウザを用いて、不特定多数のユーザが音声対話コンテンツを編集・投稿するための機能である. 提案システムでは、編集できる音声対話コンテンツは、定型的な一問一答形式に限定している. 一問一答方式ならば、複数のユーザが「一問一答」単位で独立・並列に編集・投稿しても、全体としてシステムの健全性を保ちやすい利点があると考えた. 編集・投稿された情報はデータベース上に格納される.

FST 生成機能では、音声対話共有サービスにより収集されたデータベース上に格納された情報と後述する FST テンプレートから MMDAgent で利用可能な音声対話コンテンツを自動生成する機能である.

また、FST 生成機能以外に、静的に作成された FST スクリプトを動かすこともできる. 静的 FST スクリプトは、主に、技術者によって手作業で記述することを想定しており、挨拶や天気予報・自己紹介などを記述することを想定している. 前述の通り、自動生成機能では、一問一答方式の対話しか生成できないが、静的 FST スクリプトでは、任意の対話を記載できるので、複雑な対話はこの部分に記載することが可能である.

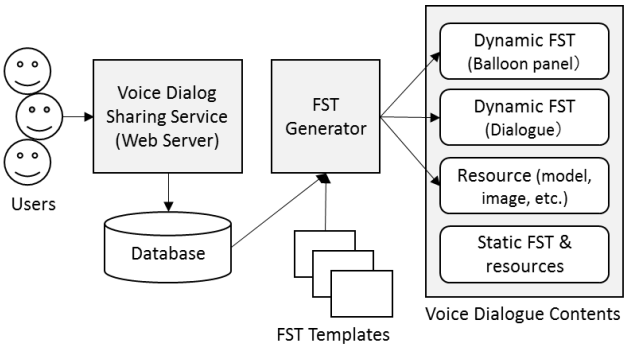

図 2 システム構成図

Figure 2 System configuration.

3.2 音声対話のモデル

提案システムで想定する音声対話モデルは、一問一答形式の対話の集合とする．一問一答形式の対話とは、一つないし複数からなる認識用キーワードを含む音声を認識すると、それらのキーワードに対応した音声合成を含む一連の動作が実行される形式の対話である．この一連の動作には、音声合成の声色の調整、3D モデルの動作などが含まれる．

想定される課題としては、1)音声だけでは詳細な情報が伝わらない、2)ユーザが何を話したらよいか分からない、という問題点が考えられる．

前者の問題に関しては、音声説明時にポスター画像(A1 縦サイズのポスターを想定)を画面上に提示して、ユーザの理解の手助けをする仕組みが考えられる．後者の問題について、次節で述べるバルーンパネル方式を採用した．

3.3 音声対話への導入(バルーンパネル)

前述した通り、一般に、音声対話システムの課題として、ユーザが何を話しかけてよいのかわからないという問題がある．この解決法として、認識可能なキーワードを示唆するイメージ画像を画面内に表示させる方法が考えられる．イメージ画像としては、図 3,4,5 に示すように、アイコンのように抽象度が高いもの(アイコン形式、)、ポスター画像のように様々な情報を掲載したもの(ポスター形式)、認識キーワードを直接文字として記載したもの(キーワード形式)が考えられる．アイコン形式とポスター形式は正門メイちゃん[5]で採用しているが、それらの情報から認識キーワードを類推する必要があり、それが一般ユーザには困難であった．キーワード形式は、パネルに表示される文字を読み上げれば良く、話しかけやすいという利点がある．また、キーワードとともに画像を表示することで、より対話内容をイメージしやすくする工夫もしている．また、複数のキーワードを想定した場合、全ての情報を画面に掲載することは難しいので、図に示すように、バルーンパネルとして、順次情報を提示する方式を採用した．

バルーンパネルの副次的な効果として、タッチパネルを用いることにより、直接バルーンをタッチすれば音声が認識しにくい人たち(子供、お年寄り、ハンディキャップのある人)でも使いやすくできる．

図 3 アイコン形式による対話内容の提示

Figure 3 A presentation of interactive content by icon format.

図 4 ポスター形式のバルーンパネルの表示

Figure 4 A presentation of balloon panel by poster format.

図 5 キーワード形式のバルーンパネルの例

Figure 5 A presentation of balloon panel by keyword format.

4. 提案システムの実装

提案システムは Java を用いて開発した．データベースは PostgreSQL を採用し、Web フレームワークとして、Apache Click を採用した．

4.1 Web 編集機能

ユーザはログインする．Google アカウントなど外部のアカウントでログインすることも可能である．

Web インタフェースの対話投稿画面を図 6 に示す．主要な項目として、キーワード、よみがな、対話文、声、表情、モーション、バルーンパネル、ポスター画像からなる．キ

ーワードは一問一答形式のキーワードであり、固有名詞などに対応できるように､「よみがな」をひらがなで記載する．対話文には、音声合成によって再生されるテキストを記載する．声は音声合成する際の声色を選択可能であり、表情とモーションは 3D キャラクタの動きを選択可能である．バルーン画像には、前述のバルーンパネルに表示される画像を表示し、ポスター画像には、音声案内とともに表示される画像を入力可能にしている．対話文の入力項目は二か所あり、二か所目の対話文に対話文を入力すると、その対話文の合成音声を再生しているときにのみ、感情音声合成やモーションなどが適用され、1 か所目の対話文を再生中は標準の声色・モーションのままである．これにより、対話中に声色や動きに変化を与えることができ、音声対話における魅力の向上につながると考えた．

投稿されると対話一覧の画面に掲載される．それぞれの対話は、投稿したユーザと管理者のみ修正や削除などが可能である．

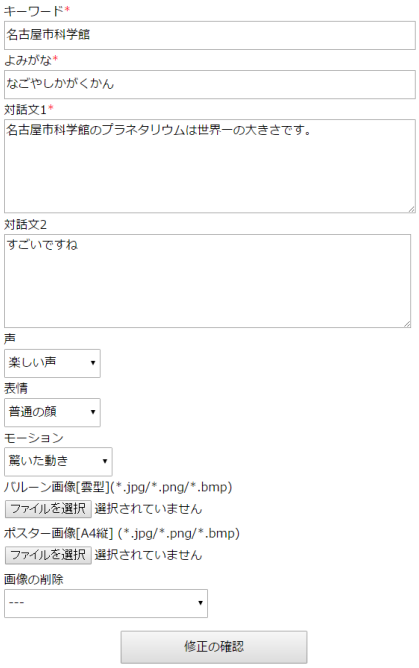

図 6　Ｗｅｂ編集画面の例

Figure 6 A capture of Web editing screen.

4.2　自動生成機能

前節でデータベースに格納された音声対話コンテンツの情報に基づき、FST を自動生成する．自動生成方式としては、FST テンプレート方式[6]を採用した．FST テンプレート方式は、対話の種類に応じて FST テンプレートを使い分け、テンプレートの変数にデータベースの値を当てはめる方式である．ただし、FST の場合は状態番号の一貫性を保つ必要があるため、その部分は自動生成機能部で管理している．バルーンパネル用の FST スクリプトと、音声認識用の FST スクリプトの二つを自動生成する仕組みを備えた．また、FST テンプレートの書式を工夫することにより、バージインにも対応した．

5.　実証実験と考察

ユーザ生成型の実証実験のシステムをいくつか開発・運用したので、それに関するシステムについて紹介する．既存研究である正門メイちゃんも含めた比較を表にまとめる．

5.1　半田市への設置

2014 年 2 月より、図 7 に示す、半田市観光案内所へ設置した．主に、半田市の観光案内に使うことを想定し、観光案内所の職員が対話内容の登録を行った．

図 7　半田市観光協会設置の音声対話システム

Figure 7 Proposed system at Tourist center in Handa city.

最初の登録内容は、観光案内所の職員から観光資源の写真と説明文をもらい、22 項目の対話を筆者らが登録した．その後、6 項目の対話が削除され、15 項目の対話内容が更新され、6 項目の対話が新たに追加された．対話内容の更新に関して、当初、対話文の文字数が長めで標準偏差が大きかった(平均 91.8 文字、標準偏差 29.3)が、変更後は、短い長さで標準偏差が小さく(平均 73.1 文字、標準偏差 16.3)変更されていた．また、単語の間に「､」を追加することで、合成音声の聞き取りやすさを向上させる、文章を話し言葉に変更する、モーションや声を変更させるなどの調整も自発的に行っていた．これらの結果から、管理者(この場合は観光案内所の職員)を対象にしたユーザ生成の仕組みを導入すると、現場のカイゼン活動により、より良い音声対話システムが実現できることが分かった．

また、利用者(観光案内所にくるお客さん)に対して、アンケートを実施した．39 名から有効回答を得た．アンケート結果を図 8 に示す．おおむね、良い結果であり、特に「楽しさ」の項目が高かった．自由記載の意見として、返答のバリエーションが多く良い、楽しみながら使えるのが良いなどのポジティブな意見が多かった．声色やモーションが変更できる機能が効果的に働いたと考えらえる．一方で、音声のみでは分かり難い、説明時に地図や写真などの説明があったほうが良い、タッチパネルで反応すると良いとの改善意見もあった．バルーンパネルに写真は掲載されてい

るものの、それだけでは不十分な場合があることが分かった．

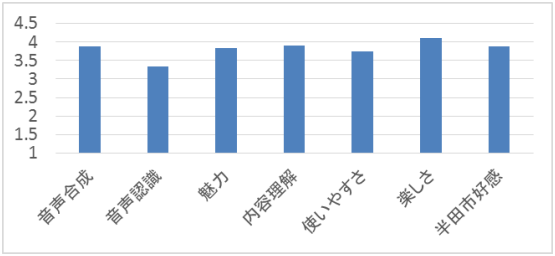

図 8　半田市利用者へのアンケート結果．5 段階評価．
Figure 8 Questionnaire results for tourists at Handa city.

5.2　名古屋工業大学夢ルームへの設置

名古屋工業大学の夢ルームとは、学生の談話スペースであり、そこに設置した（図 9）．特徴として、全学生が名古屋工業大学の情報基盤センターが提供するアカウントでログインして、全学生に対話を投稿可能にした点にある．

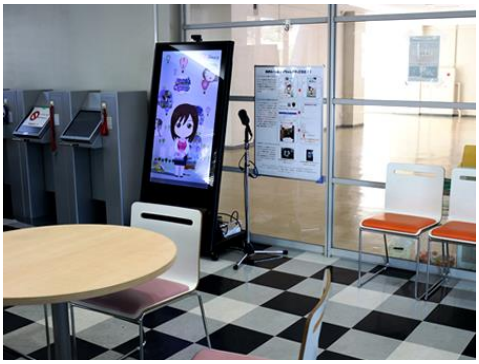

図 9　名古屋工業大学夢ルーム設置の音声対話システム
Figure 9 Proposed system at Nagoya Institute of Technology.

2014 年 10 月 1 日から正式運用を開始して、2015 年 10 月 22 日までに 502 件の投稿があった．月別の投稿数を図 10 に示す．ただし、初期運用時に運営側で 28 件のサンプルを投稿してあるので、全体としては、530 件ある．なお、ユニークユーザ数は 47 人であるが、表 1 に示すように、一人で 363 回投稿しているユーザもいた一方で、半数の 24 人は 1 回のみの投稿であった．

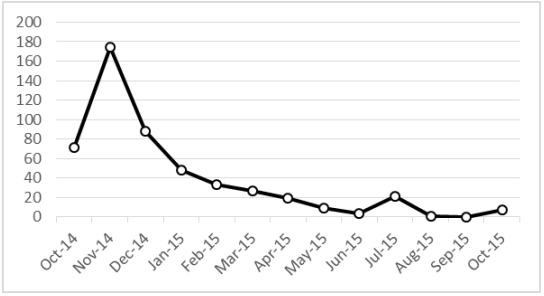

図 10　月別の投稿数．縦軸の単位は人
Figure 10 Number of submitting per month.

バルーン用の画像の投稿は任意であるが、内、65%(327 件)にバルーン用の画像が含まれていた．一方で、表 2 に示すように、声色の設定は 78%(391 件)が、モーションの設定は 80%(402 件)が初期設定である「ふつう」のままであった．バルーン画像の登録は当初ユーザに対する負担となるかと考えたが、多くのユーザにとって自然な行為であることが分かった．

表 1 投稿数別投稿人数
Table 1 Number of person by number of submitting

Number of Submitting	Number of person
1	24
2	9
3	2
4	4
From 5 to 10	5
From 10 to 20	2
363	1

表 2　声色およびモーションの種類別設定数
Table 2 Number of setting of voice style and motions.

Voice Style Type	
Normal	391
Anger	20
Embarrassment	16
Happiness	63
Sadness	12

Motion Type	
Embarrassment	18
Guts	27
Bye	3
Greeting	3
Happiness	31
Normal	402
Sadness	12
Surprise	6

音声合成用の対話文の文字数は平均 71.2 文字、標準偏差 53.8 であった．半田市観光協会のカイゼン後の平均文字数(73.2 文字)とほぼ同じであるが、標準偏差が極めて大きかった．これは、半田市は投稿内容が観光案内に限定されているのに対して、夢ルームは自由度が高いのが原因であると考えらえる．中には 300 文字を超える長さのものもあった．音声合成が長いと待ち時間が多く大変なので、バージイン機能が有効であると考えらえる．また、長い投稿は短くする、話速を調整して早口言葉で話をする、複数の対話に分割することが有効であると考えられるが、今後の課題である．

投稿されたコンテンツの種類を筆者が分類したところ、アニメ・ゲーム・漫画関係が多く(151 件)、一般知識(83 件)、雑談(81 件)、学校(53 件)、人物(38 件)、音楽・娯楽(40 件)、店舗・施設(23 件)、科学(8 件)であった．その他不明(25 件)もあった．

次に、対話の利用状況について調査した．図 11 に、1 投稿あたりの実際の対話回数について、対話回数の多い順に並べたものを示す．いわゆる、ユーザ生成型コンテンツの特徴の一つである、ロングテール型の傾向を示している．ちなみに、人気のある順に、榛名(162 回)、留年(122 回)、ぐぐる(119 回)、ルイズ(113 回)、台風(108 回)、押すな(74 回)、熱風(68 回)、ガブリアス(66 回)、天鳳(66 回)、うんざうんざ(54 回)となった．ユーザ生成型ではない、一般的なキーワードである、メイちゃん(78 回)、こんにちは(31 回)などよりも多くなっていた．バルーンパネルの効果だと考えられる．530 件中、472 件が期間中に実際に対話として利用された．1 投稿あたり平均 11.4 回利用された．

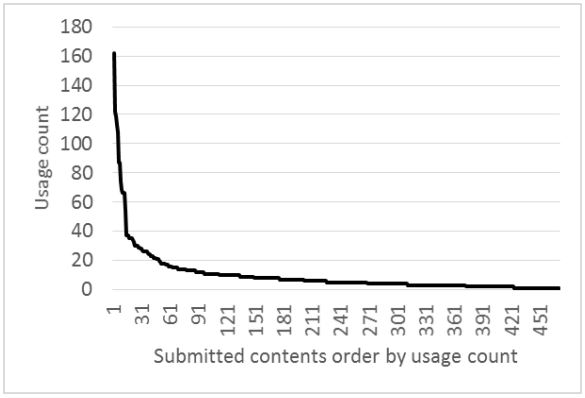

図 11　音声対話コンテンツの対話別利用状況
Figure 11 Usage of voice dialogue content per dialogue

5.3 その他の実証実験例

その他、NHK 名古屋局の公開イベントや学会案内等にも利用したので簡単に報告する．NHK 名古屋局においては、コンテンツの登録と運営を NHK の職員が行った．また、学会(平成 27 年度電気・電子・情報関係学会東海支部連合大会)においては、学会でのセッション案内に利用した．バルーンパネルに、対話キーワードだけでなく、セッション番号や日時などの情報も表示し、さらに、音声合成による説明に加えて、会場の部屋までの屋内地図を表示した．また、より大型のディスプレイを用いた音声対話システムを半田市役所にも設置した．

6. 関連研究

より、高度に音声対話コンテンツを編集する方式として、我々はいくつかの方式を提案してきたので、はじめにそれを紹介する．

EFDE[6]は、タブレット端末を用いて音声対話コンテンツの状態遷移図をタッチパネルと音声を用いて直感的に編集することが可能なインタフェースである．一般に、状態遷移図を直接編集しようとすると状態数が増加して操作性が悪化する問題があるが、定型的なテンプレートを用意し、その一つのテンプレートを一つの状態として扱うことが可能である．提案システムより複雑な状態を編集可能である利点がある一方で、複数人で同時編集する仕組みを適用しにくいという問題がある．MMDAE[7]は、Web 上で FST を直接編集できるようにした Web サービス型インタフェースである．FST を直接編集できることで自由度の高い FST の編集・共有可能になる利点がある一方で、FST に関する高度な知識がなければ扱えないという問題がある．正門メイちゃん[5]は、提案システムと同様に、Web ブラウザをもちいて、イベント情報の登録が可能なシステムである．提案システムは、正門メイちゃんのシステムを発展的に改善したものである．

7. おわりに

本研究では、ユーザ生成型の音声対話コンテンツ編集システムを提案した．Web サービスの技術を採用することにより、誰でも簡単に音声対話の編集が可能になった．また、音声対話の形式を一問一答形式に限定することにより、二重編集やデッドロック等の問題を回避した．バルーンパネルを導入することにより、音声対話への導入を図った．さらに、半田市観光協会、名古屋工業大学学生スペースでの実証実験の結果を報告し、また、NHK 名古屋局、学会等での取り組みを紹介した．特に、名古屋工業大学の学生スペースでは、500 件を超える自由な対話の投稿があった．また、そのコンテンツの利用状況を調査すると、ロングテールの傾向があることが分かった．

今後の課題としては、音声案内時に音声だけでは分かり難い、音声認識がしにくいという意見が多かった．正門メイちゃんで採用している音声合成時にポスター画像を提示と組み合わせる方式や、タッチパネルとの併用手法についても検討していく．

謝辞 本研究は国立研究開発法人科学技術振興機構 CREST の支援を受けた．

参考文献

1) Siri, http://www.apple.com/jp/ios/siri/
2) MMDAgent, http://mmdagent.jp/
3) Lee, A., Oura, K., and Tokuda, K.: MMDAgent - A fully open-source toolkit for voice interaction systems. In Proc. ICASSP 2013, pp.8382-8385 (2013).
4) Yamamoto, D., Oura, K., Nishimura, R., Uchiya T., Lee, A., Takumi, I., Tokuda, K.: Voice interaction system with 3D-CG virtual agent for stand-alone smartphones, In Proc. HAI 2014, ACM, pp 323-330 (2014).
5) 大浦圭一郎，山本大介，内匠逸，李章伸，徳田恵一: キャンパスの公共空間におけるユーザ参加型双方向音声案内デジタルサイネージシステムの構築，人工知能学会誌, Vol. 28, No. 1, pp.60-67, (2013).
6) Wakabayashi, K., Yamamoto, D., Takahashi, N.: A Voice Dialog Editor Based on Finite State Transducer Using Composite State for Tablet Devices, Proc. of IEEE/ACS ICIS 2015, pp.125-139 (2015)
7) Nishimura, R., Yamamoto, D., Uchiya, T., Takumi, I.: Development of a dialogue scenario editor on a web browser for a spoken dialogue system, In Proc. HAI 2014, ACM, pp.129-132 (2014)