Vol.2015-SLP-108 No.7
2015/10/30

音声対話によるテレビコンテンツ検索システム

河村　聡典[†1]

概要： 東芝のレグザシリーズは，「タイムシフトマシン」機能を搭載しており，過去数日の多数チャネルのテレビ番組を録画しておくことにより，「見たい番組を見逃さない」新しいテレビ視聴スタイルを提案している．しかし，限られた時間の中で録画された全ての番組を見ることは困難なため，ユーザが見たい番組へ簡単にたどりつくための機能が必要となる．そこで我々は，音声対話によりテレビコンテンツを簡単に検索できる対話型インタフェースを開発，製品化した．本発表では，2014 年 10 月に発売されたレグザ Z10X シリーズに搭載されている，音声対話によるテレビコンテンツ検索システムについて紹介する．

キーワード：音声対話，音声認識，音声合成，クラウドソーシング，テレビ

Spoken Dialogue System for TV Contents Retrieval

AKINORI KAWAMURA[†1]

Abstract: TOSHIBA TVs, REGZA, have a ‘time-shift machine’ function which record almost all TV programs and enables users to watch their favorite TV programs whenever they like. In order to retrieve users’ favorite TV contents easily, we developed a spoken dialogue system for TV contents retrieval. We introduce the system on REGZA Z10X released in Oct. 2014.

Keywords: spoken dialogue, automatic speech recognition, text-to-speech, crowdsourcing, television

1. はじめに

近年のテレビ放送は，放送チャネル数も増え，24 時間放送のチャネルも多く，膨大な数のコンテンツを視聴することが可能になっている. 東芝のテレビ REGZA シリーズは，過去数日の複数チャネルの番組を自動録画しておき，過去番組として視聴できる「タイムシフトマシン」機能を搭載しており, 放送時間を気にせず, しかも録画予約なしで「見たい番組を見逃さない」視聴スタイルを実現している．その一方で，見たい番組を探す手段としては，リモコンのテンキーから検索キーワードを入力して検索する手段か，過去の電子番組表から所望の番組をカーソルキーで選択する手段しか用意されておらず，見たい番組を簡単に探すことは困難である．

そこで我々は，音声対話により簡単にテレビコンテンツを検索できるシステムを開発し，2014 年 10 月に発売されたテレビ REGZA Z10X に搭載した．ユーザは自由な発声で見たい番組の条件を伝えるだけで，システムがユーザの意図を理解し，適切な検索処理を行い，ユーザの発話意図に沿った番組やシーンを検索し，候補を提示する（図 1）．本稿では，REGZA Z10X に搭載された音声対話によるテレビコンテンツ検索システムについて紹介する．

2. 検索機能概要

本システムでは自由な発話により様々な条件の番組検索

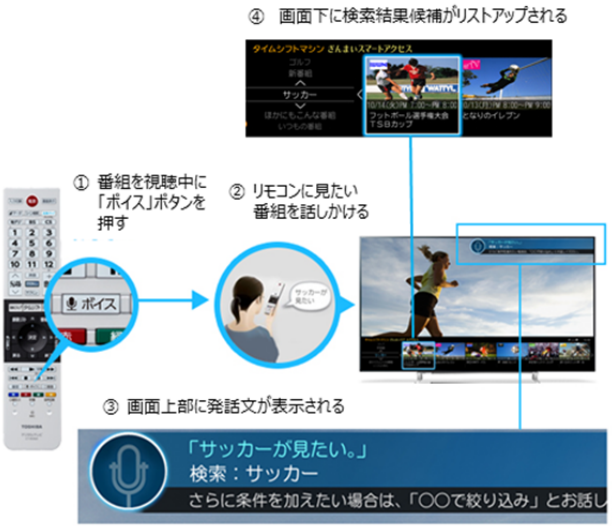

図 1　音声対話による番組検索
Figure 1 Spoken dialog interface for TV contents retrieval

が可能となる（表 1）．番組検索はもちろんのこと，検索候補をさらに絞り込む絞り込み検索や, 番組中のシーン検索, 番組情報の表示も可能である．検索対象となる TV コンテンツは，タイムシフトマシン機能で自動録画された過去番組, ユーザが明示的に録画予約した番組, 電子番組表(EPG)内の番組（これから録画予約したい未来番組），および YouTube[a]である．検索キーワードは，番組名，出演者名，ジャンル名，放送局名などが利用できる．ユーザの自由な発話から，検索対象コンテンツや検索キーワードを特定し

†1 (株)東芝　研究開発センター　知識メディアラボラトリー
Toshiba Corporation, Corporate Research & Development Center, Knowledge Media Laboratory

a) YouTube は Google Inc.の商標

検索する（表 2）．以下，受理可能な発話例と検索条件の例を挙げる．

- 「マッサンが見たい」⇒過去番組からマッサンというタイトルの番組を検索
- 「火曜日の WBS を見せて」⇒過去番組から，この前の火曜日の WBS（正式名称：ワールドビジネスサテライト）を検索
- 「新番組を予約したいのだけど」⇒ERG から新番組を検索
- 「ドラマに絞り込んで」⇒その中から更にドラマの新番組に絞りこむ
- 「有村架純の CM が見たい」⇒番組中のシーンとして有村架純の出ている CM を検索

表 1 提供機能
Table 1 List of functions of spoken dialogue system

機能	発話例
番組検索	「恋仲が見たい」
絞り込み検索	「ドラマに絞り込んで」
シーン検索	「有村架純の CM が見たい」
番組情報表示	「これのタイトルは」

表 2 検索条件
Table 2 Search criteria

条件	内容
検索対象	自動録画番組，手動録画番組，番組表，シーン，YouTube
検索キーワード	番組名，出演者名，コーナ名，ジャンル名，放送局名など

3. 音声対話システムの構成

3.1 概要

図 2 に音声対話によるテレビコンテンツ検索システムの構成を示す．本システムは，クラウド上に配置された音声認識エンジン，対話エンジン，音声合成エンジン（発音生成エンジン）と，テレビ（REGZA Z10X），マイク付リモコン，テレビに内蔵された音声合成エンジン（音声生成エンジン）から構成されている．また，随時更新されるテレビコンテンツの検索ができるように，音声認識，対話，音声合成の各エンジン用の辞書を更新するための語彙獲得エンジ，語彙獲得を支えるクラウドソーシングシステムも動作している．

ユーザがリモコンの発話ボタンを押してリモコンのマイクに対して発話すると，音声は Blutooth により TV に送信される．発話音声は TV を介して更にクラウド音声認識エンジンに送信され認識結果としての発話文が返送される．次に，発話文がクラウド対話エンジンに送信される．対話エンジンは発話文・テレビの画面状態・設定情報を受け取り，ユーザ意図を推定してテレビが行うべき動作を決定し，動作に応じたテレビ制御コマンド，応答文，及び応答文の発音情報を生成しテレビに返送する．テレビは対話エンジンから受信したコマンドを実行すると共に，発音情報から TV 上の音声合成器で生成された応答音声を再生する．語彙獲得エンジンは，テレビコンテンツ検索に必要な語彙（番組名，出演者名，トレンドワードなど）を収集する．収集した語彙は，音声認識，対話，音声合成用の語彙辞書の作成に使用される．

本システムにおいて実用的な機能・性能を実現するために，システム構成要素である音声認識，対話，音声合成，語彙獲得の全てのエンジンを自社開発している．以下，音声認識エンジン，対話エンジン，音声合成エンジン，語彙獲得エンジン，クラウドソーシングシステムについて詳細に説明する．

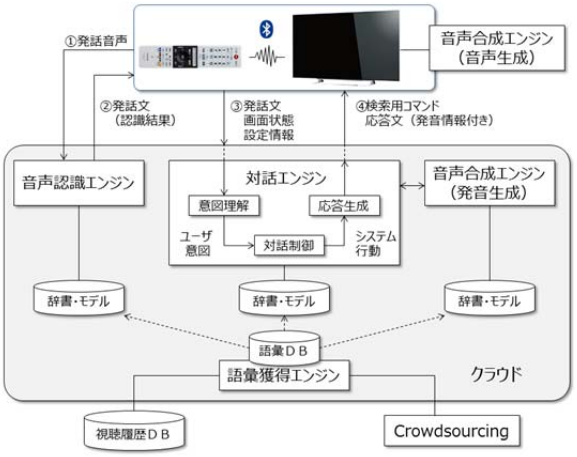

図 2 音声対話システムの構成
Figure 2 Spoken dialogue system

3.2 音声認識エンジン

自由な発話によるテレビ音声の検索を実現するために，あらかじめ定められた形式の発話文しか認識できないグラマ型音声認識ではなく，任意の表現を認識可能な大語彙連続音声認識を用いた．しかし，テレビコンテンツ検索では，番組名や出演者名やトレンドワードなどの認識すべき語彙の数が多く，発音のよく似た語彙が増加する．また，自由発話のため不明瞭な発音も多くなるため，音声認識精度の低下が懸念される．

そこで，音声特徴として SATC（Sub-Band Average Time Cepstrum）[1]を従来の振幅スペクトルと併用した DNN（Deep Neural Network）による音響モデル学習・照合技術を適用することにより認識精度の高精度化を図っている．SATC は従来よりも長い時間の分析窓で切り出した信号における周波数帯域毎の時間軸上での重心位置情報に相当する特徴量であり，従来の短時間分析窓による振幅スペクト

ル情報に基づく特徴量と相補的な性質を持つように開発された特徴量である（図 3）．従来の特徴量と SATC を併用することで，発音が不明瞭になりやすい場合の認識性能が向上することは，従来の GMM を用いる音声認識で確認されていたが, DNN への入力特徴として用いた場合にも認識精度の向上を達成している（図 4）．

また，レグザクラウドサービス「TimeOn」利用者の視聴履歴情報を活用することで，語彙獲得エンジンで獲得した語彙の言語モデルへの単語登録時に，単語出現事前確率の重みづけを施すことにより，高精度な音声認識を実現している．

3.3 対話エンジン

対話エンジンは，これまでに我々が開発したエンジン[2]を用いており，図に示すように意図理解部，対話制御部，応答生成部から構成されている．テレビから音声認識結果の発話文・テレビの画面状態（映像表示中，番組表表示中など）・設定情報（視聴可能チャンネル，HDD 接続など）を受け取り，まず意図理解部でユーザの発話意図を解析する．次に対話制御部でユーザの発話意図からテレビが行うべき動作を決定する．最後に，応答生成部で，システム動作に応じたテレビ制御コマンド（検索，推薦など）及び応答文を生成しテレビに返送する．

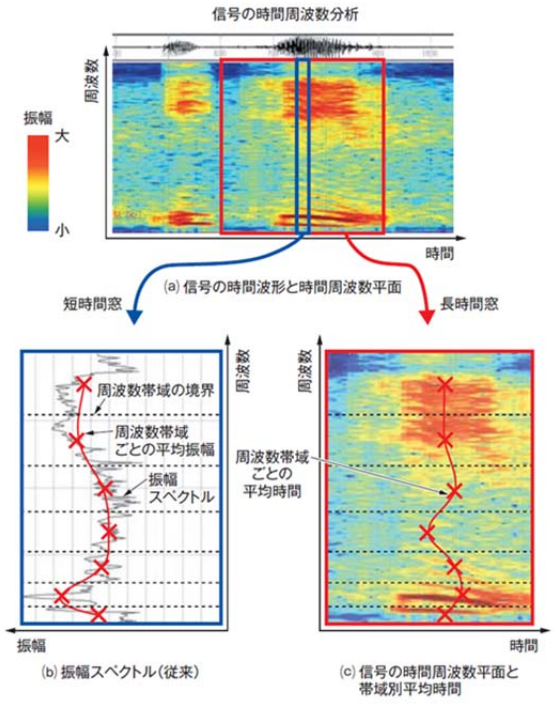

図 3 振幅スペクトルと帯域別平均時間
Figure 3 Amplitude spectrum envelope and sub-band average time

意図理解部では，発話文を解釈してユーザ意図を識別する．例えば表 1 提供機能に挙げた４つの機能「番組検索」「絞り検索」「シーン検索」「番組表示」のうち，どの機能が実行されることをユーザは期待しているのか，検索対象コンテンツは過去番組なのか未来番組なのか YouTube なのか，を発話文から識別する．しかし，同じ意図であっても様々な発話表現が存在するため，あらかじめ定めたルールにより発話文から正しく意図を識別することは困難である．そこで，統計的機械学習により意図理解のための識別モデ

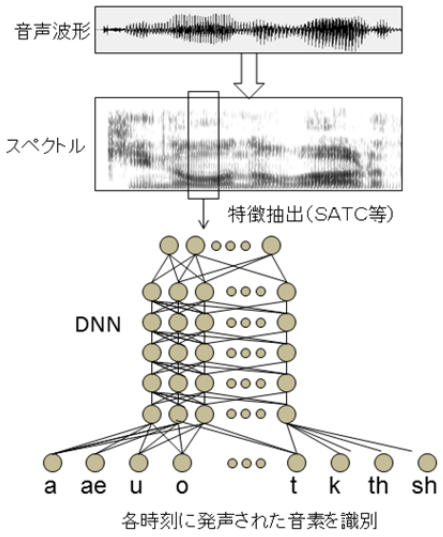

図 4 DNN による音声認識
Figure 4 Speech recognition using DNN

ルを学習している．いろいろな発話表現を集めたコーパスを用いた機械学習により，発話表現の多様性に対してロバストな意図理解を実現している．意図理解精度向上のためには，多様な発話表現を収集することが肝要であるが，当社で独自に開発したプライベートクラウドソーシングシステムを利用することで，多様な発話表現を低コストで収集することが可能となり, 高精度な意図理解を実現している．例えば，シーン検索において発話文に「シーン」という単語を明示的に含まなくても「有村架純が出てるところを出して」という発話で，シーン検索というユーザ意図として解釈できるようになる．

対話制御部では，コンテキスト情報を考慮した対話処理を行っている．①過去の対話履歴を考慮した対話処理と，②画面状態などユーザの置かれている現在の状態を考慮した対話処理である. ①に関しては, 通常であれば,「ドラマ」と発話すれば「ドラマを検索します」と応答が返り，録画番組からドラマ一覧が検索候補として表示されるが,「ＮＨＫの番組が見たい」と発話した後で「ドラマ」と言った場合には，「ＮＨＫのドラマを検索します」となる（表 3）．このように，同じ「ドラマ」という発話であっても，発話履歴を考慮することにより，ユーザの期待に添ったシステム動作を決定する対話処理となっている．②に関しては，対話やリモコン操作によって画面状態が変わった後にユー

ザが発話する時には，画面から分かる情報や操作履歴などの情報は暗黙的なコンテキストとして，発話から省略される事が多いことに対応するための対話処理である．ここでは，表 4 に示した対話例とともにその処理について簡単に説明する. まず, ユーザ 1 でユーザがシーン検索をすると，＜シーン検索画面＞が表示される．この画面を見ながらさらにシーン検索をしたいと思った場合，すでに＜シーン検索画面＞が表示されているため，ユーザ 2 のようにあえてシーン検索を意識した表現をしなくなることがある．もしユーザ 2 の発話だけで対話処理をすると，録画番組を見たいのか，シーンを見たいのかなど特定できない．そこで本システムでは，＜シーン検索画面＞状態で発話を行った場合にはその状態をコンテキスト情報として考慮した対話処理を行い，システム 2 のようにシーン検索として扱えるようにしている．なお本処理は，テレビから受け取った画面状態のほか設定情報も用いることで視聴環境もコンテキストとして扱っている．

表 3　対話例（1）
Table 3　Example of spoken dialogue (case 1)

話者	対話内容
ユーザ 1	「NHK の番組が見たいんだけど」
システム 1	「キーワードやジャンルをお話下さい」
ユーザ 2	「ドラマ」
システム 2	「NHK のドラマを検索します」

表 4　対話例（2）
Table 4 Example of spoken dialogue (case 2)

話者	対話内容
システム 0	＜映像表示＞
ユーザ 1	「有村架純が出てるところを出して」
システム 1	「有村架純のシーンを検索します」 ＜シーン検索画面＞
ユーザ 2	「福士蒼汰を見せて」
システム 2	「福士蒼汰のシーンを検索します」 ＜シーン検索画面＞

3.4 音声合成エンジン

テレビから出力される音声として十分堪えうる高い自然性を実現しつつ，音声対話の応答として遅延なく瞬時に音声を合成できるよう，最新のHMM（隠れマルコフモデル）方式による音声合成エンジンを用いた．

HMM方式の音声合成では，声質の特徴を表わすスペクトルや声の高さを表す基本周波数などの音響・韻律パラメータのパターンを，人の音声から抽出したパラメータ系列から統計的にモデル化することで，統計的に妥当でかつ滑らかなパラメータ遷移をする自然な音声を合成することができる．しかし，HMM方式では一般的に，パラメータ系列の生成や生成されたパラメータ系列からの音声波形への変換において複雑な処理を行うため，必要となる計算量は比較的高いが，テレビ内で行われるさまざまな処理に悪影響を及ぼすことなく，音声対話の音声応答として遅延なく音声を合成するためには，この計算量を低く抑えることが課題となる．

当社のHMM方式では，合成時に必要となる周波数帯域別のパルス信号やノイズ信号を，合成時に生成するのではなく，あらかじめ必要な分だけ作成したものを音声辞書に格納しておき，合成時にはこれらを組み合わせて音源信号を作るなどの様々な高速化の工夫によって，自然な音質を保ちつつ低計算量を実現した．この音声合成エンジンを用いることで，音声合成に割り当てられるハードウェアリソースが限られている中で，遅延なく自然な音声応答を実現している[3].

3.5 語彙獲得エンジン

テレビコンテンツ検索向け音声対話システムでは，番組タイトル，出演者名など，多数かつ更新頻度の高い固有表現を取り扱う必要がある．語彙獲得エンジンは，定期的にこれらの情報をインターネット等から取得する．

取得した語彙を音声対話システムで扱えるようにするためには，語彙の読み情報の推定が必要となるが，自動読み推定の技術はあるものの，推定精度には限界がある．また，番組名や出演者名などは，必ずしも正式名称を利用した検索がなされるとは限らない. たとえば,「ワールドビジネスサテライト」というニュース番組を「だぶりゅびーえす」という発声で検索したり,「やはり俺の青春ラブコメはまちがっている。」というアニメ番組を「おれがいる」という略称の発声で検索することは十分に想定される. また「木村拓哉」を「きむたく」,「サザンオールスターズ」を「さざん」で検索というように，出演者の愛称を発声することで検索することは自然である．これらの番組名の略称や人名の愛称を自動推定する技術の研究報告[4][5]はあるものの，やはり推定精度には限界があり，カバー率を上げるためには推定候補数を増やす必要があるため，音声認識にとっては無用な認識語彙数の増加につながり認識精度の低下を招きかねない.

したがって辞書として利用可能な精度の読み情報をもった語彙を獲得するためには，人手によるチェックは不可欠であるが，大量の語彙について，少数の開発者だけで短期間で読み付け，略称生成，愛称生成を行うことは非現実的である．これを解決する手段として，プライベートクラウドソーシングを活用し，短時間で低コスト，高精度な語彙獲得を実現している．

3.6 クラウドソーシング

対話エンジン，語彙獲得エンジンの節で述べたように，多様な発話表現の収集，番組略称，人名愛称，発音情報の

付与作業を短期間で安価に行う手段として，クラウドソーシングを活用している．図 5 に出題の画面例を示す．

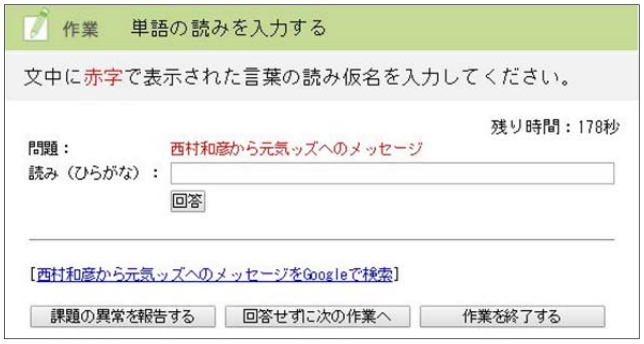

図 5 クラウドソーシングの出題例
Figure 5 Example of crowdsourcing task

クラウドソーシングとは，単純であるにも関わらず機械により自動化することが困難な作業（タスク）を，不特定多数の一般の作業者に委託し実施してもらう手法である．これにより，少数の開発者だけでは膨大な実施時間を要する大量のタスクを短期間・低コストで行うことが可能になる．しかし，既存のクラウドソーシングサービスには，作業結果の精度が低いという問題点がある．クラウドソーシングには，専門家ではない一般作業者にタスクを依頼するため，専門家に比べて精度が低いという問題が存在する．また，実際には，タスクに真面目に取り組まない不誠実な作業者も少なからず存在するため，全体的な精度低下の大きな要因となる．そのため，いかに個々の作業者の作業結果の精度を高めるかが重要な課題となる．

この問題を解決するために, 当社で自ら開発した Private CrowdSourcing System (PCSS) [6][7] を利用している．PCSS では，作業者を募集する際のユーザ属性情報に基づくフィルタリングや，作業者の PCSS での作業履歴をベースとした作業者の正解率と経験値，及びスキルの管理などにより作業精度の向上を図っている．

PCSS は WebAPI により, 作業依頼と作業結果の取得ができるようになっており，語彙獲得エンジンは PCSS の WebAPI を叩くことで人手による作業を含めた語彙獲得を自動的に行えるようになっている．

4. おわりに

本稿では, 2014 年 10 月に発売されたテレビ REGZA Z10X に搭載された音声対話によるテレビコンテンツ検索システムについてその機能と構成について紹介した．本音声対話システムによりユーザは自由な発話で番組検索の要望を伝えるだけで，所望の番組を検索することができ，膨大な録画番組を効率よく検索することができる．

参考文献

1) 中村匡伸，他: 群遅延に基づく音声特徴量の雑音環境下での評価，日本音響学会 2012 年春期研究発表会講演論文集, pp.13-15 (2012).
2) 岩田憲治，他: 課題解決知識を用いた音声アシスタント，人工知能学会言語・音声理解と対話処理研究会資料, Vol. 67, pp. 13-14, (2013).
3) 田村正統，他: HMM 音声合成による英語音声合成システムの開発，日本音響学会 2011 年春季研究発表会講演論文集，pp.313-314 (2011).
4) 若木裕美，他: Web 情報を用いた人物の愛称抽出，日本データベース学会論文誌, Vol.7, Num.1, pp.169-174 (2008).
5) Hiromi Wakaki, et al.: Abbreviation Generation for Japanese Multi-Word Expressions, Identification, Interpretation, Disambiguation and Applications pp.63-70 (2009).
6) 芦川将之，他: Private Crowdsourcing を用いた言語，音声資源の収集～システムの構築と言語収集～，人工知能学会全国大会（第 27 回）,3M3-OS-07d2 (2013).
7) 芦川将之，他: Crowdsourcing を用いた単語への読み付け，アクセント付け手法の提案，電子情報通信学会技術研究報告. AI，人工知能と知識処理. 111, 447, pp.11-16 (2012).