効率的なビジネス活動を支援する会議音声活用システム

Conferencing Audio Utilization System Supporting Efficient Business Operations

高橋　麻理子　　近藤　修明
■TAKAHASHI Mariko　■KONDO Nobuaki

近年，音声や画像を活用したシステムが企業や社会インフラで多く利用されるようになっている。特に，ビジネス活動の現場では映像を活用するビデオ会議やWeb会議の普及が目覚ましく，講演や会議の音声を有効に活用することが求められている。

そこで東芝ソリューション（株）は，東芝で開発された，話しことばに対して高い認識精度を持つ大語彙音声認識技術を利用し，従来は効果的に活用することが難しかった講演や会議の音声を活用できる会議音声活用システムを開発した。会話内容のPC（パソコン）への表示による聞漏らし防止や，重要発話前後の会話再確認，会議の議事録作成支援，会議の記録と振返りなどにより日々のビジネス活動を支援する。

The use of audio and video systems has been expanding in both the corporate and social infrastructure fields in recent years. In particular, the widespread dissemination of video conferencing systems and Web conferencing systems with image display functions has led to a growing need for business sites to make more effective use of the audio data produced by these systems.

To meet such customer requirements, Toshiba Solutions Corporation has developed a conferencing audio utilization system that realizes efficient use of recorded voice data by means of a PC. This system, incorporating a large-vocabulary speech recognition technology with high speech recognition accuracy developed by Toshiba, has the potential for application to various lectures and meetings. It supports daily business operations through a variety of functions, including a function that displays the contents of conversations to prevent missing words, a replay function to confirm conversations before and after important utterances, and functions to support the preparation of minutes of meetings and reviews by participants at past meetings.

1　まえがき

複数の遠隔地を結んで画像及び音声による会議を行うことができるビデオ会議システムやWeb会議システムは，日常の会議や海外拠点との情報交換などで広く使われている[1]。これらのシステムを導入することにより，出張費の削減や，コミュニケーションの活性化，迅速な意思決定と情報共有などが実現されている。しかし，その画像や音声の品質に対して不満を持ち，コミュニケーションが阻害されていると考えるユーザーも多い。

従来，音声認識技術は，主に人と機械との間のインタフェースを実現することを中心に適用されてきたが，近年，人と人との間での音声認識についても適用が進められている。しかし，人間どうしのコミュニケーションにおける音声認識は，機械相手の音声認識とは異なり，考えながら行われる言語的にも音響的にも明瞭とは限らない発話を対象としなければならず，その適用は限定的であると言われている。また，人間どうしの話しことばの音声においても，テレビニュースや議会の会議録などのていねいな発話に対しては音声認識を実用的に行うことができるが，会議などの発話に対しては実用化が困難であると言われている[2]。

このように，人間どうしのコミュニケーションにおける音声認識には課題が多い。そこで東芝ソリューション（株）は，東芝で開発された，話しことばに対して高い認識精度を持つ大語彙音声認識技術（注1）を利用し，会議におけるコミュニケーションを支援する会議音声活用システムを開発した（**図1**）。これ

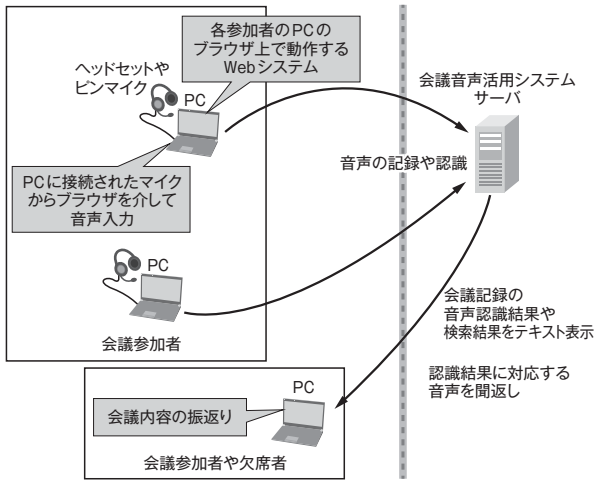

図1．会議音声活用システムの概要 ― 各参加者のPCに接続されたマイクからブラウザを介して音声を入力し，音声を記録，認識するWebシステムである。

Overview of conferencing audio utilization system

（注1）　語彙に関する制約をできるだけ取り除き，多くの語彙を扱うことができる音声認識技術。

により，音声だけの会議における聞漏らしを防止することや，発話内容を確実に相手に伝えること，今までの議論の思い返しを支援することなどが可能となり，会議の音声を価値ある情報として活用するシステムを実現した。

2　東芝の大語彙音声認識技術

会議音声活用システムに利用した東芝の大語彙音声認識技術について述べる。

2.1　音声認識技術の発展

従来，音声認識技術には限られたグラマ（文法）に沿った音声を認識するグラマ型音声認識が用いられていた。グラマ型音声認識は，限定的な表現を認識するのには適しているが，認識可能な語句は数語から数千語規模に限られていた。しかし近年，ハードウェアの高性能化及び音声認識エンジンの性能改善に伴い，数十万語以上の大語彙連続音声認識が可能になってきている。

また従来は，機器に対する命令音声や原稿の読上げ音声など，はっきりとした発声の読上げ調音声を対象とした音声認識技術が主流であったが，近年は，音声認識エンジンの大語彙化及び音韻識別性能の高精度化に伴い，自由な話しことばに対する音声認識も可能になってきた（**図2**）。

2.2　大語彙音声認識技術

音声認識エンジンの大語彙化のためには，新語や，造語，口語表現なども含んだ大量の語彙を収集し，音声認識エンジンが用いる単語辞書及び単語モデルに追加する必要がある。

東芝は，不特定多数の人に業務を発注するクラウドソーシングを用いた語彙収集方法を確立し，大量の語彙収集を可能にした。また，大量の語彙には多くの類似単語が含まれ，それらを判別するための音韻識別性能の向上が必要となる。これに対しては，新たな音響特徴量であるSATC（Sub-Band Average Time Cepstrum：帯域別平均時間ケプストラム）を開発し，音韻識別性能の向上を図っている[3]。

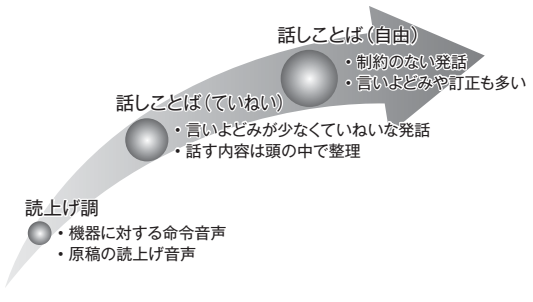

図2．読上げ調音声認識から話しことば認識への性能改善 ― 従来の読上げ調音声に加え，自由な話しことばに対する音声認識が可能になってきた。
Improvement in performance of speech recognition from controlled speech to free speech

2.1.1　大規模な語彙収集方法の確立

テキストコーパス(注2)からの未知語自動獲得技術は従来からあるが，獲得された語彙が必ずしも適切でない場合があり，人手によるチェックが必要である。多くの語彙に対するチェックを行うためにクラウドソーシングを活用している。しかし，クラウドソーシングには，作業結果の精度が低いという課題がある。これを解決するため，作業者の行動履歴から作業者の正解率と経験値及びスキルを算出することで，ある作業を適切な作業者に割り当てることができるPCSS（Private Crowdsourcing System）を開発し，語彙収集に利用している[4]。PCSSにおける正解率は“正解数／作業数”で算出され，一定値以下の作業者には作業を割り当てない。経験値は“正解数－不正解数”で算出され，一定の経験値を持つユーザーに対しては高難度の作業を割り当てる。スキルは作業者の得意とする作業の種別を表し，作業の種別ごとに適切なスキルを持つ作業者に割り当てることで，精度を向上させる。

2.1.2　音韻識別性能の向上

音韻識別性能の向上のために開発した，新たな特徴量であるSATCは，従来の特徴量と比べ長時間の音声情報を表現できる。従来の短時間の音声特徴量であるMFCC（Mel-Frequency Cepstrum Coefficient：メル周波数ケプストラム係数）などと併用することで音韻識別性能の向上が期待できる。実際，SATCを併用した特徴量を用いて大語彙連続音声認識を行うことで，コンタクトセンターの対話における精度向上[3]や雑音環境下での精度向上が確認されている[5]。

3　会議音声活用システムの特長と構成

会議音声活用システムは，前述の大語彙音声認識技術を会議音声に適用することで，今まで効果的に活用することが難しかった講演や会議の音声を活用できる。

3.1　会議音声活用システムの特長

会議音声活用システムは，会議音声活用システムサーバにブラウザを介してアクセスする，Webシステムとして構築されている。

会議参加者がそれぞれPCを用いて会議音声活用システムにアクセスするとともに，PCに接続したヘッドセットに備わるマイクやピンマイクを用いて音声を入力する。入力された音声は会議音声活用システムサーバに送信され，音声認識エンジンで音声認識が行われる。認識結果は会議参加者全てのPCに送信され，ブラウザ上にテキストで表示される。また，事前に設定した重要キーワードを含む会議中の発話は重要発話として判断され，自動でリストアップされる。そして，その重要発話の前後の発話を聞き返すこともできる。

会議終了後には，会議参加者，若しくは会議欠席者がPCから会議音声活用システムサーバにアクセスし，記録された発話

（注2）　自然言語の文章を大規模に収集し，言語的な情報を付与したもの。

の参照や検索，聞返しなどを行いながら議事録を作成できる。

3.2　会議音声活用システムの構成

会議音声活用システムは複数のコンポーネントから構成される（**図3**）。

クライアント（PC）は，認識文字表示・会議操作・議事録操作・発話再生部及び音声入力部から構成される。

(1)　認識文字表示・会議操作・議事録操作・発話再生部　HTML5（Hypertext Markup Language 5）によるUI（ユーザーインタフェース）であり，会議の設定や発話再生，発話の認識結果（テキスト）の表示などを行う。

(2)　音声入力部　マイクから入力される音声やUIからの音声認識開始・終了操作に応じたトリガを送信する。また，サーバから受信した発話の認識結果（テキスト）を認識文字表示部に送信する。

サーバは，発話区間検出部，発話文字化部，発話・認識文字・会議管理部，及び議事録管理部から構成される。それぞれが持つ機能を以下に示す。

(1)　発話区間検出部　クライアント（PC）からのトリガを受信して発話開始時間と発話終了時間を取得し，音声データを切り出す。切り出した音声データと発話開始・終了時間を発話文字化部に送信する。

(2)　発話文字化部　受信した音声データと発話開始・終了トリガを音声認識エンジンに送信する。また，音声認識エンジンから受信した認識結果（テキスト）を発話・認識文字・会議管理部に送信する。

(3)　発話・認識文字・会議管理部　受信した音声データはWAV（Audio Waveform）形式のファイルとしてサーバ上に保存し，認識結果（テキスト）は認識文字データベース（DB）に保存する。WebSocket(注3)を使用して認識結果（テキスト）を認識文字表示部に送信する。

(4)　議事録管理部　議事録操作部から受信した議事録の情報を議事録DBに保存する。また，議事録操作部から受信した検索条件に対応した認識結果（テキスト）を認識文字DBから取得し，認識文字表示部に送信する。

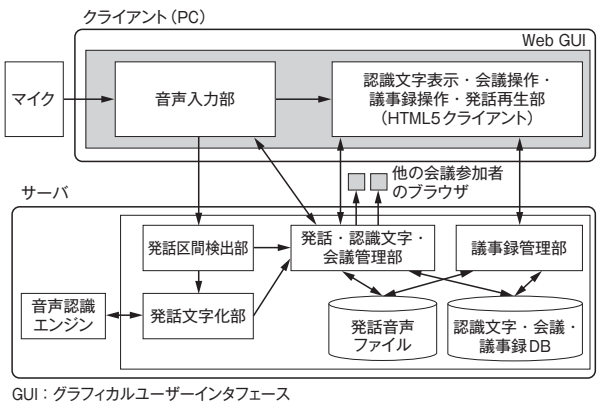

図3．会議音声活用システムの構成 ― クライアント（PC）上及びサーバ上の複数のコンポーネントから構成される。

Configuration of conferencing audio utilization system

4　ビデオ会議システムとの連携

会議音声活用システムは，ビデオ会議システムと連携して動作することもできる。その一例としてテレプレゼンスシステムと連携動作する場合を**図4**に示す。会議音声活用システムサーバ上に，テレプレゼンス連携アプリケーションを設置し，サーバ上で動作させる。テレプレゼンス連携アプリケーションは，次に示す会議連携部と字幕送信部から構成される。

(1)　会議連携部　会議音声活用システム内の会議管理情報とテレプレゼンスシステム内の会議管理情報との対

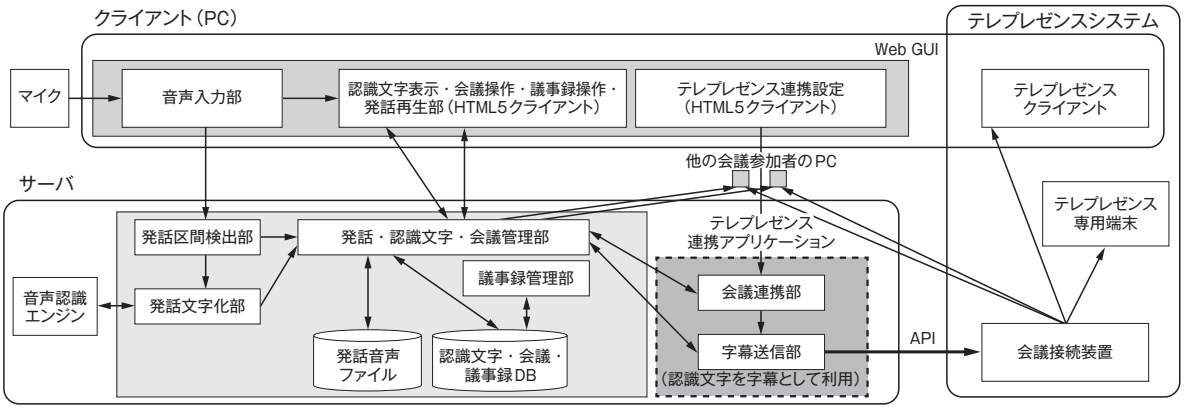

図4．ビデオ会議システム連携時の構成 ― ビデオ会議システムであるテレプレゼンスシステムと連携して動作する。

Configuration of conferencing audio utilization system used in conjunction with video conferencing system

(注3)　サーバがレスポンスを返した後も，コネクションが切断されない双方向通信技術。

応関係を設定し，管理する。

⑵　字幕送信部　　発話・認識文字・会議管理部からWebSocketを使用して送信された会議の発話の認識結果（テキスト）のうち，会議連携部で対応関係が設定された会議の発話の認識結果（テキスト）だけを受信し，MCU（Multipoint Control Unit：多地点接続装置）のAPI（Application Programming Interface）を用いて字幕として表示する。字幕はMCUに接続しているPC上のテレプレゼンスクライアントやテレプレゼンス専用端末上で閲覧できる。

5　効果

会議音声活用システムを導入することによって得られる効果について述べる。

まず，会議中の効果は以下の2点である。

⑴　遠隔会議など音声だけでのコミュニケーションによる情報伝達に失敗する可能性がある場面において，認識した音声をテキストとして表示することで，発話内容を確実に相手に伝えられる。

⑵　会議中の発言を探して聞き返すことや重要発話を参照することで，聞漏らしの確認や会議内容の把握ができる。

次に，会議後の効果は以下の2点である。

⑴　記録された発話の検索や聞直しを行うことで，時間経過後の振返りを容易に行えるとともに，議事録の元になる発話内容を簡単に整理でき，議事録の作成にも生かせる。

⑵　会議中の全ての発話を聞き返すことはユーザーの負担になるが，重要発話を参照し，その発話の前後の音声を再生することで，音声聞返しの負担を低減できる。更に，次回会議での前回会議の振返りや会議欠席者への情報伝達などにも生かせる。

会議後における会議の振返りを支援する例を**図5**に示す。

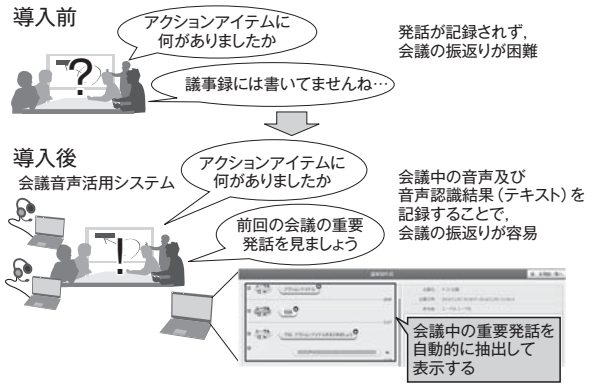

図5．会議の振返りを支援する例 — 前回の会議での重要発話を参照し，その前後の音声を聞き返すことで確実な振返りができる。

Example of use of function supporting reviews of past meetings

この例では，前回の会議での重要発話を参照し，その前後の音声を聞き返すことで確実な振返りを行うことができている。

6　あとがき

話しことばに対して高い認識精度を持つ大語彙音声認識技術を利用し，従来効果的に活用することが難しかった講演や会議の音声を活用できる会議音声活用システムを開発した。

今後は，更なる認識精度の向上や，様々なビデオ会議システムとの連携，他のアプリケーションパッケージやソリューションとの連携などを図っていく。

また，東芝ソリューション（株）では，東芝で開発された同時通訳技術を利用して翻訳連携を行う機能も開発中である。海外拠点とのビデオ会議などの際に，音声認識により可視化した音声をリアルタイムに翻訳，字幕化することで，多言語会議でノンネイティブな言語に対して理解力が不足している参加者を補助することが可能になる。

会議の音声を活用したシステムにおいて，“翻訳連携”と“ビデオ会議連携”を備えたシステムは，グローバル化が進むビジネス活動の現場での活用シーンが広がると考えている。

文　献

⑴　リクルートマーケティングパートナーズ．“Web会議システムの導入状況”．キーマンズネット．<http://www.keyman.or.jp/at/30003806/>，（参照2014-12-08）．

⑵　河原達也．話し言葉の音声認識の進展－議会の会議録作成から講演・講義の字幕付与へ－．メディア教育研究．**9**，1，2012，p.S1－S8．

⑶　益子貴史 他．同時通訳や音声対話の実用化に向けた大語彙音声認識技術．東芝レビュー．**68**，9，2013，p.6－9．

⑷　中田康太 他．“PrivateCrowdSourcing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～”．2013年度人工知能学会全国大会（第27回）論文集．富山，2013-06，人工知能学会．2013，3M3-OS-07d-2．（CD-ROM）．

⑸　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年春季研究発表会講演論文集．横浜，2012-03，日本音響学会．2012，p.135－136．（CD-ROM）．

高橋　麻理子　TAKAHASHI Mariko

東芝ソリューション（株）プラットフォームセンター ソフトウェア開発部主任。メディアインテリジェンスソフトウェアの開発に従事。

Toshiba Solutions Corp.

近藤　修明　KONDO Nobuaki

東芝ソリューション（株）プラットフォームセンター ソフトウェア開発部グループ長。メディアインテリジェンスソフトウェアの開発に従事。

Toshiba Solutions Corp.