音声言語処理技術で書き起こし作業を効率化するWebサービス ToScribe™

ToScribe™ Web Application to Enhance Efficiency of Audio Transcription Work

上野　晃嗣　　芦川　平
■UENO Koji　　■ASHIKAWA Taira

音声をテキスト化したいというニーズに対し，東芝は，音声認識技術を直接用いるのではなく，それをはじめとする様々な音声言語処理技術を組み合わせることで，人手による書き起こし作業を効率化する複数の機能を開発し，インターネット上の無料サービス“音声書き起こしクラウドエディタ ToScribe™（トゥスクライブ）”として一般公開した。

このサービスは，一般ユーザーがWebブラウザ上で利用でき，煩雑で集中力が必要な書き起こし作業において，音声認識の内部結果を用いた音声の自動頭出し，事前の音声特徴量のクラスタリングによる話者推定，及び文章構造解析技術を応用した整文(注1)支援など，高度な音声言語処理技術による支援を得ることができる。

Toshiba has launched ToScribe™, a new, free, cloud-based application that allows users to manually transcribe speeches more efficiently by integrating a number of speech and language processing technologies including automatic speech recognition (ASR) technology. ToScribe™ works with major Web browsers, and offers effective transcription assistance while simplifying troublesome audio player control operations by means of the following high-level speech and language processing technologies: automatic speech position estimation by manipulating the internal results of the ASR, automatic speaker estimation by clustering audio feature values, and proofreading assistance applying our test structure analysis technology.

1　まえがき

スマートフォンをはじめとした録音可能な機器の普及に伴い，音声をテキスト化したいというニーズは大きくなっている。その内容も，議会の議事録や米国における医療記録，コールセンターの会話記録など，証拠性を目的とした従来からあるものだけではなく，一般企業の会議議事録やスマートフォンを利用した音声メモ，及びコンテンツの翻訳目的など，よりカジュアルに，より多様に変化している。

音声をテキスト化するためには，自分で書き起こす以外では，専門家による書き起こしや，クラウドソーシングによる書き起こしなどの手段が取られてきているが，いずれも潜在的なニーズに十分応えられているとは言えない。書き起こし作業は非常に煩雑であり，一般に音声時間の数倍の時間を要する。専門家による書き起こしは良好なテキストが得られる一方，高価であり，書き起こしの依頼から結果を取得するまでに時間がかかる。インターネットを通じて多数のパートタイム作業者に作業を割りふる，クラウドソーシング方式による書き起こしサービスは，安価で比較的高速だが，情報漏えいのリスクが高いため，秘匿性の高いデータには利用できない。

こうした状況を打開するため，音声認識技術に期待が集まっている。音声認識技術は，録音環境や話し方，及び発言内容にある程度の制約をかければ，実用化が可能となる段階まで進歩している。例えば，スマートフォンを使った音声検索は，ひとりの人間がスマートフォンに向かって検索キーワードを話しかける状況に限られるため，条件を自然に絞り込むことができ，十分な認識精度を出せる可能性がある。

しかし，現実の音声データは非常に多様である。会議の録音とインタビューの録音，あるいは動画コンテンツなどでは，録音の質，定常・非定常雑音の種類，話し方，及び話す内容などがまったく異なってしまう。現状の音声認識技術で，このような状況の全てに対応することは難しい。

そこで東芝は，頑健で汎用的な音声テキスト化支援を早期に実現するため，音声認識技術をそのまま用いるのではなく，

図1．ToScribe™における書き起こしのようす — 画面下部のエディタに音声内容を書き起こす。
Scene of usage of audio transcription with ToScribe™

(注1)　書き起こし結果の校正のこと。

音声認識技術をはじめとする様々な音声言語処理技術を組み合わせることで，人手による書き起こし作業を効率化する複数の機能を開発し，インターネット上の無料サービス“音声書き起こしクラウドエディタ ToScribe™（トゥスクライブ）”として一般公開した[1]（**図1**）。

ToScribe™は，煩雑な頭出し作業を自動化する自動頭出し機能[2]，音声中の話者を推定する話者推定機能，及び書き起こし結果を校正する整文支援機能などを備え，人手による書き起こし作業を効率化する。ユーザーは，ToScribe™によって，秘匿性の高いデータや様々な録音環境の音声についても，より頻繁に書き起こすことが可能になる。

2　音声書き起こしクラウドエディタ ToScribe™

ToScribe™は，音声書き起こしを目的としたWebアプリケーションで，ブラウザ上で動作する。このアプリケーションを用いることで，音声書き起こし作業を従来よりも効率的に行うことができる。以下に，そのシステム構成と動作について述べる。

2.1　システム構成

ToScribe™のシステム構成を**図2**に示す。ユーザーのWebブラウザと，メインシステム内のファイアウォール，Webサーバ，及びデータベースから成るシステムは，通常のWebアプリケーションと同様の仕組みである。このアプリケーションの特徴として，計算負荷の高い音声処理を必要とするため，専用のエンジンサーバを複数用意し，Webサーバは，キューを介して随時エンジンサーバに計算処理を依頼する。メインシステムは固定的な物理サーバクラスタとして構築されているが，ユーザーの一時的増加の際には，VPN（Virtual Private Nerwork）接続したパブリッククラウドにエンジンサーバを更に増やすことが可能になっている。

2.2　ToScribe™の機能と動作

このアプリケーションのユーザーは，Webブラウザで全ての操作を行う。音声データや作業内容は全てサーバ側に自動保存されるため，ユーザーは作業中断後も別のパソコン（PC）から作業を再開できる。

まず，ユーザーは書き起こしたい音声をWAV（Audio Waveform）やMP3（MPEG-1（Moving Picture Experts Group-phase 1）Audio Layer 3）などのファイル形式でアップロードする。システムはエンジンサーバに音声解析要求を送り，エンジンサーバは書き起こし作業に備えて，アップロードされた音声を事前に解析する。解析内容には，後述のとおり音声認識や話者推定が含まれ，全音声の解析には，音声の長さと同程度の時間がかかる。

解析が完了した部分から書き起こしが可能になる。ユーザーは，作業中に随時，音声の再生及び停止とテキストの入力を行うが，ToScribe™は更に，音声の自動頭出し，話者推定，声のピッチを変えない話速変換，及び音声のノイズ除去など高度な機能を提供する。また，これらに必要な計算もエンジンサーバ上で行われる。

書き起こし作業が完了すると，ユーザーは結果をダウンロードする。ダウンロード前にユーザーは語尾の不統一などをチェックする整文支援機能を用いて，文章の校正を行うことができる。

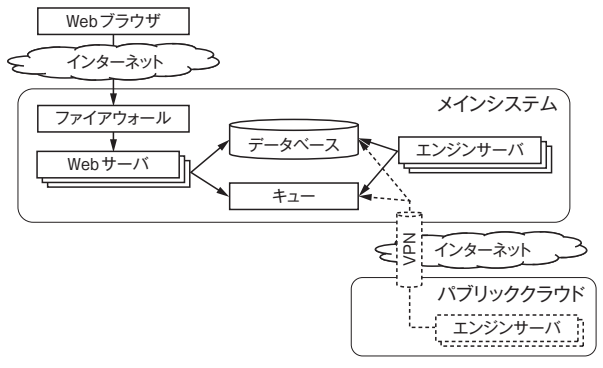


図2．ToScribe™のシステム構成 ― 複数用意されるエンジンサーバが音声言語処理を受け持つ。
System architecture of ToScribe™

3　自動頭出し機能

3.1　粗起こし作業と頭出し位置推定技術

専門家の書き起こし作業は，事前確認や，粗起こし，細部の聞取り，整文などの要素に分かれる。このうちもっとも時間を要するのが，聞いたままをテキストに起こす粗起こし作業である。粗起こし作業では，作業者は次の操作を何度も繰り返す。

(1)　音声の再生を開始する。
(2)　聞き取れた部分を書き起こす。
(3)　音声が先に進んでしまうため，書き起こせた部分の終端位置まで音声を巻き戻す。

従来，(1)と(3)における音声の再生と巻戻し操作が作業者にとって煩雑であり，負担になっていた。また，音声の中で書き起こせた部分の終端位置が自明ではないため，正確な巻戻しは難しく，(3)においても巻戻し操作は複数回必要であった。専門家は，フットペダルを用いることで肉体的負荷を軽減しているが，作業時間自体の短縮にはつながっていない。

この問題に対し，ToScribe™は自動頭出し機能を提供する。キーによって頭出しを指示すると，自動的に書き取った部分の終端位置に音声を巻き戻し，再生を再開する。頭出し指示は初期状態でEnterキーに割り当てられているため，作業者は書き起こし中に余分にEnterを押すだけで，音声位置を常に適正に保つことができる。また，書き起こし終わったテキスト中の任意の位置にカーソルを置き，Enterキーを押すことでカーソル位置から音声を再生することもできる。

自動頭出し機能を実現するため，既存の音声言語処理技術を応用した頭出し位置推定技術を新たに開発した。

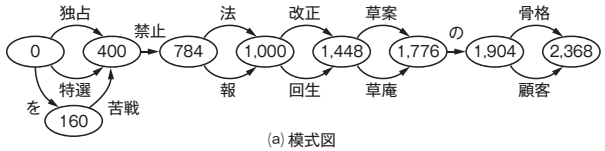
頭出し位置推定技術は，書き起こし済みのテキストと音声の組から，テキスト中の任意の位置（実際はカーソル位置）と対応する音声位置を推定する技術である。ここで，書き起こしテキストは，必ずしも音声内容を正確に表していない。現実の書き起こしでは，文章の倒置や削除が行われるためであり，例えば，「晴れていますね，えー，きょうは」という発言を「きょうは，晴れていますね」と書き起こす。聞き取れなかった箇所を●などの記号で代替することもある。

この技術の開発にあたっては，こうした不正確なテキストと多様な条件の音声との対応を推定するため，ラティス方式と強制アライメント方式という二つの異なる方式を併用することで，頑健性を上げる方針とした。

3.2　ラティス方式

ラティス方式は，事前の音声認識によってラティスを生成し，このラティスの中から書き起こしテキストを単語単位で検索する方式である。ラティスとは音声認識の出力形式であり，候補単語を連結したグラフ構造のデータを指す。候補単語は音声中の出現位置を含むので，目的のテキストが見つかれば，自動的にその音声位置が得られる。ラティスの例を**図3**に示す。

ラティスは，音声認識によって推測された音声言語構造の情報を最大限に含む，一種のマスタデータとなる構造体である。単語単位や文章単位の信頼度上位n件など，一般的な音声認識結果の形態は，前段で内部的に生成されたラティス構造から変換，抽出されて作られている。ToScribe$_{TM}$では，更に音声認識エンジンのパラメータを変更し，通常よりも大きいラティスを生成することで，信頼度の低い候補単語や語尾の揺らぎなどのバリエーションを増やし，多様な音声入力に対する頑健性を上げている。



(a) 横式図

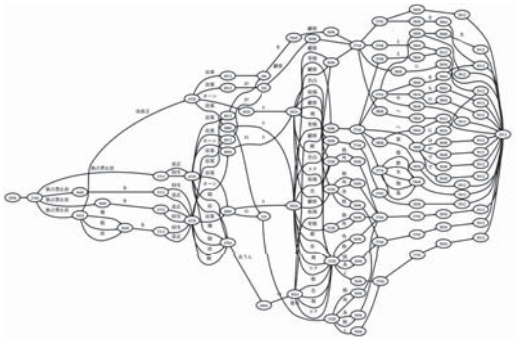

(b) 実際に生成されるラティスの例

図3．音声認識結果ラティス — ラティスは，音声から推測される単語例の候補と出現位置情報から成る。
Example of lattice estimated by ASR results

ラティス方式のアルゴリズムは次のとおりである。ここでは，カーソル位置直前の書き起こし文をSEN，音声の再生開始位置をSS，及び現在の音声再生位置をSCとする。また，ラティス中の候補単語情報はアーク（図3における矢印）に含まれるものとする。

(1)　SSからSCを推定区間Sとする。
(2)　S内に存在するラティスのアーク群A'を得る。
(3)　A'中の2文字以上の単語のアーク群Aを得る。
(4)　AからSEN中に単語が含まれるアーク群Pを得る。
(5)　Pから終端時間位置が最大であるアークP_{max}を得る。
(6)　A からP_{max}と同じ候補単語を検索する。
(7)　(6)が存在しなければ，P_{max}をP_{out}とする。
(8)　(6)が存在すれば，その終端時間位置が最小であるアークをP_{out}とする。
(9)　P_{out}の終端時間位置を推定頭出し位置とする。

ラティス方式は事前に音声認識を実行できるため，実行時の処理時間が短く，テキストの倒置や削除に強い反面，ラティスに書き起こしテキストの単語が一つも含まれない場合は，頭出し位置を推定できないという問題がある。

3.3　強制アライメント方式

ラティス方式の問題点を解消するため，頭出し位置推定技術では，ラティス方式による推定に失敗した場合には強制アライメント方式を用いる。強制アライメント（Forced Alignment）とは，音素表記列と音声の組から，各音素の音声中の出現位置を推定する技術である。例えば，音声認識や音声合成などで使われる音響モデル学習用データを自動的に音素単位に整備するために使われる。ここでは，ユーザーによる書き起こし文と音声データの間の強制アライメントを可能にするため，音声合成技術で用いる，漢字かな交り文からかな文字列への変換器を強制アライメントへの入力に組み合わせた。

強制アライメント方式のアルゴリズムは次のとおりである。変数名は前述のラティス方式と同様とする。

(1)　SSからSCを推定区間Sとする。
(2)　SEN中の単語群Wを得る。
(3)　Wのそれぞれについて推定区間Sに対する強制アライメントを行い，出現位置群Tを得る。
(4)　Tのうち最大のものを推定頭出し位置とする。

強制アライメント方式は，音声認識で認識できず，ラティス方式では頭出し位置が推定できなかった音声であっても，書き起こしテキストが実際に発音されたことばであれば音声位置を推定できる。ただし，実行時に音声からの音響特徴量算出と音素位置推定を行うため，処理時間が長くなる。

3.4　頭出し位置推定技術の評価

頭出し位置推定技術の評価を行った。10分前後の音声

表1．頭出し位置推定技術の評価結果
Results of evaluation of automatic speech position estimation

ID	頭出し誤差（ms）	従来手法からの改善（ms）
1	−2,623.9	2,225.8
2	−3,323.2	2,112.4
3	−2,303.2	2,606.5
4	−3,598.2	1,986.2
5	−1,108.0	3,726.3
6	−1,247.8	3,358.0

ID : Identification

データと正解データ（書き起こし及び単語出現位置）のペア6件をテストセットとした。音声内容は会議室での会議，展示会会場でのプレゼン会話，及びラジオで，話者数は2～10人である。出現単語数は701～1,257で，総計5,971である。実験では，全ての単語に対して頭出し位置推定を試行した。音声再生開始位置は語の出現位置から5.0秒前，現在の音声再生位置は語の出現位置から0～20.0秒後のランダムな値とした。また，従来手法として，現在の音声再生位置から固定で5.0秒前を頭出し位置とする場合との比較を行った。ただし，開発した手法でも，ラティス方式と強制アライメント方式の両方が失敗した場合には，固定長の巻戻しを行う。

評価結果を**表1**に示す。頭出し位置は平均で正解から約1.1～3.3秒巻き戻った場所となった。これを従来手法と比較すると約2.0～3.7秒の改善となり，いずれのデータセットでも有意水準5％で有意差が認められ，従来手法に対して有効であることが示された。

4　仕上げ支援機能

議事録やインタビューの書き起こしでは，粗起こし結果はそのままアウトプットとして使えず，必ず話者の記載や文章の校正が必要であり，専門家にとってもこれらは神経を使う作業となっている。ToScribe™は，これらの仕上げ作業の支援機能として，話者推定機能と整文支援機能を提供する。

4.1　話者推定機能

話者推定機能は，音声中の各発言の話者を推定する機能である。ユーザーがこれから書き起こそうとする範囲の音声の話者を推定し，その結果が，既に入力された発言の話者と同一である場合，話者名が自動的に入力される。

ToScribe™は，音声特徴量のクラスタリングに基づく話者推定技術によって話者推定機能を実現している[3]。事前の教師データをいっさい与えなくても推測可能であり，部分的な教師データを与えることで結果を改善できるという特長を持つ。

この技術は，音声データを特徴量の時間系列に変換するステップと，特徴量群をクラスタリングによって話者別に弁別するステップで構成される。特徴量の時間系列への変換は比較的計算量を要するが，音声データごとに1回だけ計算すればよいため，ToScribe™では，アップロード後の音声解析段階で事前に計算しておき，その結果を書き起こし作業中にも再利用するようにした。

4.2　整文支援機能

前述のとおり，整文とは書き起こし結果の校正のことである。通常の校正は原稿と出力結果の比較により行われるが，書き起こしでは，音声の話しことばと書き起こし結果テキストとの比較になるため，実際には，結果テキスト自体が読みやすく，文章として問題ないことが基準となる。

ToScribe™では当社の機械翻訳技術を応用し，話者ごとの文末表現の不一致，かな表記の揺れ，誤字脱字，及び同音語の誤り，の4点を自動的に判断し指摘する整文支援機能を提供する。

5　あとがき

当社は，煩雑な頭出し作業を自動化する自動頭出し機能をはじめ，人手による音声書き起こし作業の効率化を実現する複数の機能を開発し，インターネット上の無料サービス“音声書き起こしクラウドエディタ ToScribe™”として一般公開した。

このサービスは，専門家やクラウドソーシングによる書き起こしと将来の汎用音声認識の間をつなぎ，現状技術の組合せによって，増大する音声テキスト化ニーズに応えるものである。この目的のために，音声認識技術のほか，音声合成や機械翻訳に用いられる技術をも応用している。

今後は，運用中のサービスの更なるユーザビリティ向上に取り組むほか，当社の音声言語処理研究の最新成果を応用して支援機能を向上させるなど，ユーザーの音声テキスト化をより効率化するため研究開発を進める。

文　献

(1) 東芝．“ToScribe−音声書き起こしクラウドエディタ−”．<http://www.toscribe.toshiba.co.jp/>，（参照2013-08-09）．
(2) 芦川　平 他．音声書き起こし支援システムに向けた自動頭出し機能の開発と評価．研究報告 音声言語情報処理（SLP）．2012-SLP-90，23，2012，p.1−6．
(3) 広畑　誠 他．“話者交代検出のためのモデル学習区間推定法”．日本音響学会 2007年春季研究発表会．東京，2007-03，日本音響学会．3-10-3．

上野　晃嗣　UENO Koji
研究開発センター 知識メディアラボラトリー研究主務。
音声認識サービスの設計・開発に従事。情報処理学会会員。
Knowledge Media Lab.

芦川　　平　ASHIKAWA Taira
研究開発センター 知識メディアラボラトリー研究主務。
音声認識サービスの設計・開発に従事。
Knowledge Media Lab.