Vol.2011-SLP-85 No.8
2011/2/4

ライブストリーミングのための協調的音声書き起こしシステム

浮田俊輔[†1]　緒方淳[†2]
後藤真孝[†2]　小林哲則[†1]

本稿では，ustream のようなライブストリーミングの配信動画中の音声を，不特定多数のユーザが協調してリアルタイムに書き起こすことが可能なシステム「Yourscribe」を提案する．従来，人手による書き起こしは労力が大きく，音声認識での書き起こしは精度が不十分であった．また，その精度不足を不特定多数のユーザによる訂正で補う PodCastle は，リアルタイムでの書き起こしには適していなかった．Yourscribe では，各ユーザは動画視聴を中断せずに楽しみながら，気が向いたときに聴き取った断片的なテキストを入力するだけでよい．それらを多数のユーザから随時集めた後に，リアルタイムに音声認識した結果と照らし合わせることで，書き起こしテキストに自動的にまとめ上げることができる．これは「音声認識研究 2.0」の新たな事例に位置づけられる．

A Collaborative Speech Transcription System for Live Streaming

Shunsuke Ukita ,[†1] Jun Ogata ,[†2] Masataka Goto [†2]
and Tetsunori Kobayashi [†1]

In this paper, we propose a real-time transcription system, *"Yourscribe"*, that enables anonymous users to collaboratively transcribe speech in live video streaming like *ustream*. In previous approaches, transcription by human was laborsome and transcription by speech recognition was error-prone. PodCastle was developed to overcome such errors of speech recognition by having anonymous users correct errors, but was not appropriate for real-time transcription. To use Yourscribe, each user can just voluntarily type a short phrase that was heard while enjoying the live video without any interruption. Those phrases from many users were aggregated at all times and matched with results of real-time speech recognition to automatically form the transcription. This is a new instance of our research approach, *Speech Recognition Research 2.0"*.

1. はじめに

動画コンテンツを多人数へリアルタイムに中継・配信できるライブストリーミングが普及したが，その利活用技術はまだ不十分である．ustream[⋆1] やニコニコ生放送[⋆2] 等の Web サービスによって誰もが手軽に番組を中継・配信可能となり，人気のある番組は数百〜数万人に視聴されている．また，視聴しながら，その内容に関連したテキストをタイプする活動も活発である．ustream の場合には，twitter[⋆3] 等のマイクロブログと呼ばれる短いテキスト共有によるコミュニケーション用 Web サービスが併用されることが多く，ニコニコ生放送の場合には，入力したテキストが動画コンテンツの上を重なり合って流れることでコミュニケーション可能な機能が提供されている．しかし，これらの動画コンテンツは見逃すと後からの迅速な内容把握が難しい．そこで，そのコンテンツ中の音声に時刻同期した書き起こしテキストが作成できれば，見逃した人々にとって読むだけで内容把握ができて役に立つ．その上，視聴した人々にとっても構造化や検索が可能になり，内容を振り返り要約しやすくなって，さらなる利活用が促せる．

このように書き起こしテキストは有用なため，従来，関心の高い動画コンテンツの一部は，ボランティアによって後から人手で書き起こされて公開されていたが，多大な労力を要していた．自動的に書き起こしを生成するために音声認識を用いる試みもあったが[1)2)3)4)5)6)]，高い音声認識率を得るには環境を整える必要があり，一般的な動画コンテンツへの適用は難しかった．そうした音声認識において，認識率は今後向上しても 100%にはならない問題への解決策として，不特定多数のユーザに音声認識誤りを訂正してもらう Web サービス PodCastle[7)8)9)] を我々は提案し，2006 年から一般公開している．当初は音声コンテンツのみに対応していたが，2009 年からは動画コンテンツにも対応した[10)]．しかし，過去の録音・録画のみに対応し，ライブストリーミングには対応していなかった．仮に PodCastle を高速化してライブストリーミングに対応させようとしても，ユーザが訂正をしている間にコン

†1 早稲田大学
Waseda University
†2 産業技術総合研究所
National Institute of Advanced Industrial Science and Technology (AIST)
⋆1 http://www.ustream.tv/
⋆2 http://live.nicovideo.jp/
⋆3 http://twitter.com/

Vol.2011-SLP-85 No.8
2011/2/4

テンツの内容が先に進み，コンテンツ自体をリアルタイムに楽しめなくなる問題が生じる．

そこで本研究では，ライブストリーミングに特化して，動画コンテンツ中の音声を多数のユーザが協調してリアルタイムに書き起こすことが可能なソーシャルアノテーションシステム「Yourscribe」を提案する．我々は，ユーザ自身がコンテンツ視聴を楽しみながら貢献できることを重視する．そのため，無理のない範囲で気が向いたときに，聞き取った断片のテキストを書き起こしとして入力できるクライアントインタフェースを提供する．上述のように既に普及している視聴中のテキスト入力同様，ユーザは数十秒から数分ごとに好きなタイミングで好きな長さだけ，気軽に入力すればよい．Yourscribe のサーバでは，それらを入力時刻情報を伴う形で数百〜数万人分集積することで，一つの書き起こし文章に自動的にまとめ上げる．しかし，各断片テキストは文の一部で不完全で，入力の時間遅れも不均一なため，文章にまとめるのは容易でない．そこで音声認識をサーバ側でリアルタイムに実行して，その認識結果と断片テキスト群を統合していくことで，書き起こしとしてまとめ上げることを可能にする．

以下，2 章において Yourscribe が実現する機能の特長を議論し，3 章で具体的な実現方法を述べる．4 章で小規模な予備実験の結果について述べ，5 章で関連研究を紹介して今後の課題を議論する．最後に 6 章でまとめを述べる．

2. Yourscribe の概要と特長

本研究で提案する「Yourscribe」は，動画コンテンツのライブストリーミング中の音声を集団で書き起こすためのソーシャルアノテーションシステムである．ライブストリーミングを提供する Web サービスに関しては，Yourscribe のサーバ側でストリーミングされた音声を音声認識用にリアルタイムに取得ができ，ユーザのクライアント上で再生ができれば，任意の Web サービスを用いてよい（動画に限らず音声のみの配信でもよい）．本研究ではその第一段階として，個人が自由に動画（映像と音声）を多人数とリアルタイムに共有できる Web サービス ustream を対象に研究を進めた．ustream では各動画の配信者の意向によって，ライブストリーミングを後から閲覧することが可能な場合と不可能な場合があるが，可能な場合を対象とした．ただし後から閲覧できるとは言っても，リアルタイムに中継・配信されているときに，より多くのユーザが高い関心を持って視聴すると考えられる．そこで我々は，そうしたユーザが多く集まる中継時に焦点を絞り，ユーザが視聴しながら，同時に twitter 等でテキストをタイプしてコミュニケーションしていることに注目した．

参加の敷居を低くするために，各ユーザは Yourscribe のクライアント上で，ustream の

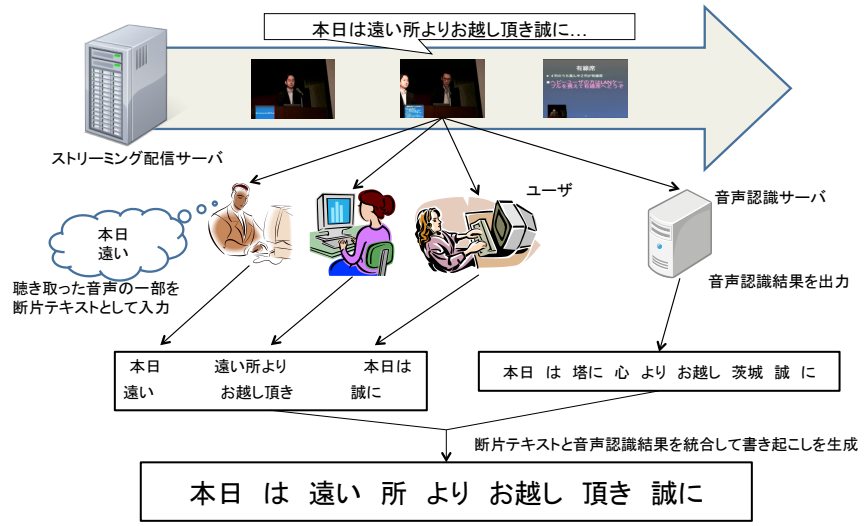

図 1 Yourscribe のコンセプト：不特定多数のユーザの各々が書き起こしの断片となるテキストを入力すると，それらが集められて音声認識結果と統合され，最終的な書き起こしテキストが生成される．

動画コンテンツのリアルタイム配信を視聴しながら，自分が聞き取った音声の断片をテキスト入力するだけで貢献できるようにする点が，従来の書き起こし活動にはない特長である．ユーザの一人一人がコンテンツ自体を楽しみながら書き込める量には限界があるが，**図 1** のように，そうしたユーザからの小さな貢献を，多くのユーザから集めてまとめ上げることで，書き起こしの質を上げることができる．そのためには，貢献するユーザ数が多いことが重要になる．コンテンツに対する社会的関心が高いほど視聴者は多くなり，Yourscribe 上で協力してくれるユーザ数も増えることが期待できる．そうした人気の高いコンテンツは書き起こしの潜在需要も高いが，そうなるほど的確な書き起こしが作成できる可能性が高くなるのは，Yourscribe のアプローチの重要な特長と言える．一方，視聴者がほとんどいないコンテンツでは，Yourscribe では的確な書き起こしは作成できない．なお，ユーザによる断片的なテキストの入力は，常に本来の音声の時刻よりも遅れるが，音声認識結果と統合することで，各単語が音声の時刻と同期した状態の書き起こしテキストを生成する点も特長である．

近年，ライブストリーミングで提供される動画コンテンツが増えてその価値と人気は高

まっており，音声の書き起こしを作成する重要性も増している．見逃した人々あるいは途中から視聴し始めた人々にとって，書き起こしは読むだけで内容把握ができて役に立つ．それだけでなく，リアルタイムに視聴した人々にとっても，音声と同期した書き起こしのテキストは有用であり，コンテンツ内の構造化やブラウジング，キーワード等による検索に利用できる．後から内容を振り返り，言及したり要約したりする際にも便利である．他にも，書き起こしテキストに対するマイニングやマッチング広告等も可能になる．もし書き起こしを自動的に字幕化して，元の動画と共に再度提供できれば，聴覚障害者等がコンテンツを楽しむ上でも活用できる可能性がある．このように多様なメリットがあり重要であるため，気軽な協力であれば，視聴者の一部は Yourscribe に書き起こしの断片を提供してくれると我々は仮定した．

3. 実 現 方 法

Yourscribe のプロトタイプ段階での簡易的な実装状況を説明する．Yourscribe の最終的な実装形態は Web サービスを想定しており，エンドユーザが Web ブラウザでアクセスすると，利用できることが望ましい．しかし現段階では，スタンドアロンのクライアントプログラムとサーバプログラムで構成され，多数のユーザが利用できる状態には至っていない．そこで，Web サービスとしての実装の詳細は稿を改めて紹介することとし，本稿では，後述する 4 章の予備実験で使用した，既に実装済の中核となる処理を紹介する．

3.1 断片テキストの入力

ユーザが，Yourscribe のクライアント上で聴き取った断片テキストをタイプ入力すると，それが入力された際に記録するコンテンツ上の時刻を伴って，サーバ側にネットワークを通じて集められる．視聴を楽しみながら気軽に入力するテキストであるため，入力しやすい名詞，特に固有名詞の 1 単語か，数単語からなる複合名詞が主な入力対象となり，それらを含むフレーズが中心になると考えられる．コンテンツの書き起こしに高い意欲を持つ一部のユーザは，より多様で長い語句を入力する可能性もある．しかし，入力された断片テキスト群は，コンテンツ中の全ての音声区間をカバーするものだとは期待できない．

例えば，コンテンツ中に「小笠原っていうのは東京都でありながらですね沖縄以上に南国みたいなところあるわけで」のような発声があった場合，それを不特定多数のユーザが「小笠原」,「東京都でありながら」,「沖縄以上に」,「南国みたいなところ」等のような断片テキストとして入力すると考える．ユーザ間で入力内容に重複も起き，例えば上記の例で,「東京都でありながら」と入力するユーザが数十人いる一方で,「東京都」と入力するユーザも数十人いる可能性がある．

ただし，各ユーザは聴き取った後の好きなタイミングで断片テキストを入力するので，本来の発声のコンテンツ上の時刻（該当する音声区間）には時間的に対応付いていない．つまり，断片テキスト入力の際に記録される時刻は，該当する音声区間の時刻から常に遅延 (タイムラグ) が生じた状態となる．このような「不完全な」テキストを利用して，最終的にコンテンツ中の全音声区間に対する書き起こしを生成することを試みる．

3.2 断片テキストを利用した書き起こし生成手法

ライブストリーミングされたコンテンツをユーザが視聴するの同時に，サーバはそのコンテンツ中の発声を音声認識し，その音声認識結果のテキストを断片テキストと統合することで，極力すべての発声区間の書き起こしテキストを生成する．上述のように各断片テキストの時刻は不正確なため，最終的な書き起こしを生成するための基本的な方針としては，音声認識結果のテキストに対して，ユーザから入力された個々の断片テキストを，本来発声された区間に適切に対応付け（アラインメント）することを考える．ここで，断片テキストが入力されたときに記録される，コンテンツ（音声データ）上での時刻を t_e とする．その元となる発声の先頭から，断片テキストの入力が完了する時刻 t_e までの遅延の上限を T_{delay} とすると，断片テキストは，t_e から一定の時間幅 T_{delay} だけ遡った音声データ中に，含まれるはずである．そこで，その区間内の音声データ中から，断片テキストに該当する区間を探索して見つけ，その区間の書き起こしとして時間的に対応づけていく（割り当てていく）．

このような時間的な対応付けの課題に取り組む上で，簡便な方法の一つは，対象となる探索区間の音声データに対応する音声認識結果 (単語列) と各断片テキストの単語列との間で，DP マッチングによるテキストアラインメントを行うことである．ただし，基本的に音声認識結果は認識誤りを含んでいるため，単語表記の類似度を考慮する DP マッチングではアラインメントのエラーが避けられない．特に本研究の対象であるライブストリーミングのコンテンツは，実環境の多様な音声データであり，現状の大語彙連続音声認識システムではコンテンツによっては誤認識の数が著しく増えてしまう．そこで本手法では，音声認識結果のテキストに対してアラインメントするのではなく，元の音声の音響信号に対して直接アラインメントをする．以下のように音響モデル (HMM: 隠れマルコフモデル) を利用して，断片テキストが音声データ中のどこに含まれるかを対象とする音声データ中から見つけだし，断片テキストを適切な区間へ割り当てていく．

3.2.1 断片テキストの単語分割と読みの獲得

音声の音響信号に対する HMM を用いたアラインメントの前処理として，まず入力され

た断片テキストを形態素解析により単語分割する．ここでは単語分割の際に，不特定多数のユーザによって日々整備され，更新されている「Web キーワード辞書」を活用した形態素解析を行う[11]．これにより，新出語の分割誤りを低減することができ，さらに今回のアラインメントで特に重要な，読み (発音) の情報も獲得することができる．

3.2.2 HMM によるアラインメント

断片テキスト入力時刻 t_e から一定の時間幅 T_{delay} だけ遡った音声データ中に，断片テキストが出現する区間を求める．具体的には，断片テキストという認識対象が一つに絞られている状況で，音響モデル (HMM) を用いたアラインメントをする．これは，音声データ中からある特定のキーワードが発話された区間を特定するキーワードスポッティング法に近い枠組みで実現できる．具体的には，入力された断片テキストの音素列に沿って音素 HMM を連結することで，断片テキストに相当するキーワード HMM を作成する．キーワード HMM の前後には，キーワード以外の音声区間を割り当てるためのガベージモデルを付与することで，上記の音声区間に対する認識ネットワークを構成する．ここでガベージモデルとしては，任意の音素のループを用いる．これを利用して Viterbi デコーディングを行うことで，コンテンツ中の断片テキスト（キーワード）の存在区間（開始時刻と終了時刻) を推定する．

3.2.3 最終的な書き起こしの単語列の生成

上記の HMM に基づくアラインメントによって，各断片テキストがコンテンツ中のどの区間に存在しているかが求まる．それらすべての断片テキストを用いて，**図 2** に示すようにコンテンツ全体に対する音声認識結果（単語列）との間で時間的照合を行い，両者が反映された最終的な単語列を求める．時間的照合の際，音声認識結果と断片テキストのアラインメント結果で単語境界が異なる場合には，両区間のオーバーラップ率で閾値処理を行うことで判断した．

4. 予 備 実 験

3 章で述べた断片テキストを利用した書き起こし生成手法の基本的な性能及び効果を確認するための予備実験を行った．実験に使用する動画コンテンツとしては，本システムで想定している ustream のライブストリーミングとは異なるが，Web 上でのポピュラーな音声コンテンツであるポッドキャストを利用した．実験で用いた音声データは，Web 上で公開されている 3 つのポッドキャスト A，B，C の各 1 エピソードである（計 49 分 30 秒）．これら 3 つは，音声認識率が大きく異なる音声データとして選んだ．現段階ではまだ Web サービスとして運用できないため，実験用の断片テキストを不特定多数のユーザから集めるのは

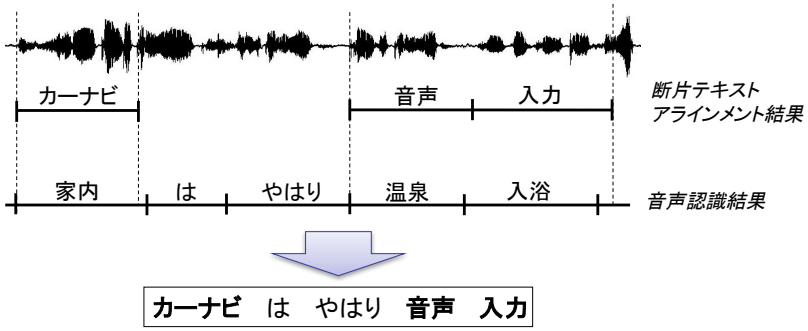

図 2 断片テキストのアラインメント結果と音声認識結果の統合処理

困難である．そこで本予備実験では，著者のうちの 1 人が，コンテンツを視聴しながら多人数によって無理なく入力可能と想定されるような単語やフレーズを，書き起こしの断片テキストとして用意した．用いた断片テキストは，各エピソードともに全体の音声区間の 75%をカバーするデータである（仮名漢字交じりの書き起こしテキストの正解文を人手で用意し，その文字数上での 75%がカバーされる分量に調整した）．これらの断片テキストは，3 つのポッドキャスト全体で計 1886 件を用意し，その仮名漢字交じりの書き起こし文字数は，平均 7.41 文字，標準偏差 3.26 文字であった．

コンテンツ全体に対する音声認識結果のテキスト（単語列）を生成するための音声認識器には，PodCastle 音声認識システム[12][13] を用いた．音響モデルは，CSJ の約 600 時間の講演音声データから学習された，状態数 3000，1 状態あたり混合ガウス分布数 16 の tied-state cross-word triphone モデルである．特徴量には 39 次元 PLP(12 次元 PLP ケプストラム係数と正規化パワー，それらの Δ，ΔΔ)，そして話者，環境の変動に対処するために CMLLR ベースの適応化学習[14] を行っている．言語モデルは，Web キーワードベースの N-gram[11] であり，Web ニューステキスト，CSJ の講演書き起こしを用いて学習したものである．一方，書き起こし生成手法における HMM アラインメント処理には，上記の CSJ から学習した混合数 32 の monophone モデルを用いた．ここで triphone ではなく，monophone を利用することで，アラインメント処理における計算コストを大きく削減することができる．

評価尺度としては，最終的に断片テキストを統合して生成した書き起こしテキストが，実際の正解文の書き起こしテキストをどの程度正しく再現できたか，すなわち音声認識性能

表 1 断片テキストを利用した書き起こし生成手法による性能向上の評価 (単語正解精度)

	通常の音声認識結果（ベースライン）	断片テキスト+音声認識結果（提案手法）
A	77.88%	**85.66**%
B	47.42%	**73.00**%
C	31.90%	**60.25**%

表 2 断片テキストのデータ量 (正解の書き起こしテキストに対するカバー率) の違いによる比較 (単語正解精度)

カバー率	25%	50%	75%
A	80.47%	82.65%	**85.66**%
B	58.57%	65.43%	**73.00**%
C	42.18%	52.53%	**60.25**%

と同様に単語正解精度を用いた．また，アラインメント処理を行う際の時間幅 T_{delay} は 10 秒とした．

表 1 に，ベースラインとなる音声認識器による書き起こし生成結果（通常の音声認識結果）と，提案手法の断片テキストを利用した書き起こし生成結果の，それぞれの単語正解精度を示す．通常の音声認識結果より，ポッドキャストによって認識性能に大きな違いがあることがわかる．特に，C は芸能人による雑談音声であり，発話速度も比較的速く誤認識が著しい．提案する断片テキストを利用した書き起こし生成手法により，3 つ全てのポッドキャストにおいて大きく単語正解精度を改善できていた．認識率 50%以下のような，音声認識が非常に困難なデータに対しての改善率が特に高い結果となった．これにより monophone のような簡易な音響モデルであっても，HMM によるアラインメント処理は有効であり，断片テキストの発声区間をある程度正しく検出できていたことがわかった．

次に，入力される断片テキストのデータ量の違いによる，最終的な書き起こし性能への影響を調べた．ここでは，全体の 75%をカバーする断片テキスト群からランダムに削減することにより，全体の 50%，25%それぞれをカバーする断片テキスト群を作成した．75%をカバーするためには，非常に多くのユーザが同時に視聴しながら，ときどき書き起こしている状況が想定されるが，50%，25%は，そのユーザ数が減った状況を想定した実験条件である．**表 2** にそれぞれのデータ量の断片テキストを利用したときの単語正解精度を示す．結果より，比較的少量の断片テキストが入力された場合でも，本手法により，それらを最終的な書き起こしとして有効利用できる可能性があることがわかった．

5. 議　　論

音声の自動書き起こしについては，現在も様々なタスク・ドメインについて研究が精力的に行われている．文献 15)16)1)2) では放送ニュースを対象とした音声認識に関する研究がなされ，そういった放送コンテンツに対してリアルタイムに字幕を作成するシステムも提案されている[15)16)]．文献 17) では，国会審議音声を対象とした音声認識システムが検討され，話者や話題の変化に追従するための半自動システム更新等[18)] により高精度な認識を実現している．また，話し言葉音声研究のプラットフォームとして，学会講演[19)]，大学講義[3)20)]，会議音声[4)5)6)] といったタスクについても盛んに研究されている．大語彙連続音声認識技術のアプリケーションとして，特に字幕付与や議事録等を考えた場合，音声認識結果の修正や整形といった人手による後処理[21)] は必要不可欠である．大語彙連続音声認識技術が格段に向上した現在でも，このような人手での後処理はコストが大きく，従来のシステムでは，例えば放送ニュースの字幕生成における認識誤り修正[16)] のように少数の人員による専門的な作業 (一人が誤認識箇所を発見，別の一人が正解をタイピング等) を前提としていた．

それに対し，本研究では，少数の人から多大な貢献を期待するのではなく，多数の人から少しずつの貢献を期待する立場を取る．これは，文献 7) で PodCastle が取っていた立場と同じであり，いかに多数の人に参加してもらえるかが鍵となる．PodCastle では，既存のコンテンツを非リアルタイムに視聴する状況を前提に，ユーザの参加の形態が「訂正」であったが，訂正するためには一旦音声認識結果をユーザが確認し，その上で，再生を一時停止して訂正する必要があった．それに対して Yourscribe では，ライブストリーミングをリアルタイムに視聴する状況を前提に，ユーザの参加の形態を「断片テキストの投稿」とした．これにより，一時停止せずにコンテンツを楽しみながら，貢献することを可能にした．しかし参加人数によっては，書き起こしの質を充分に高められないことが起きうる．そこで，ライブストリーミング終了後に，その時点での書き起こしを PodCastle 上で訂正できるようにする拡張が将来的には考えられる．

現在の簡易的なプロトタイプ実装では，まだ Web サーバや Web ブラウザ上で動作するクライアントとしては実装が完了しておらず，今後の課題である．実装はより困難となるが，原理的には，Yourscribe はテレビ等の既存の放送メディアにも有効と考えられる．既

にテレビ視聴中のテキスト入力は，ニコニコ実況[⋆1] や，torne[⋆2] でのライブ機能 (twitter 連動機能) 等のように一般化しつつあり，親和性は高い．

6. おわりに

本稿では，音声認識を活用し，ライブストリーミングの音声の書き起こしテキストを不特定多数のユーザが協調して作成することを可能にする「Yourscribe」という新たなソーシャルアノテーションシステムを提案した．長時間のライブストリーミングは後からの視聴が容易でないが，その場でリアルタイムに視聴している人達の一部が，楽しみながら片手間に聴き取ったテキストの断片を気軽に入力するだけで，書き起こしが作成可能になる点が優れている．本研究はまだ，Web サービスの形で公開することを目指した準備段階であるが，音声認識技術でサポートされたリアルタイム書き起こしサービスとして，エンドユーザの役に立つという社会的意義を持たせるべく，研究を進めていく予定である．

また，Yourscribe は，ユーザに対して音声認識の現状を開示し，ユーザの協力を得て音声認識技術を発展させていく研究アプローチ「音声認識研究 2.0」[7)8)9)] に基づいた新たな事例として，PodCastle に続くサービスとなることを目指している．今後は，より実装を進めていくことで，そうした貢献につなげていきたいと考えている．

参考文献

1) Chen, S.S., Eide, E.M., Gales, M.J., Gopinath, R.A., Kanevsky, D. and Olsen, P.A.: Recent Improvements to IBM's Speech Recognition System for Automatic Transcription of Broadcast News, *Proc. ICASSP'99*, Vol.1, pp.37–40 (1999).
2) Woodland, P.C., Gales, M.J., Pye, D. and Young, S.J.: Broadcast News Transcription Using HTK, *Proc. ICASSP'97*, Vol.2, pp.719–722 (1997).
3) Glass, J., Hazen, T.J., Cyphers, S., Malioutov, I., Huynh, D. and Barzilay, R.: Recent Progress in the MIT Spoken Lecture Processing Project, *Proc. of Interspeech 2007*, pp.2553–2556 (2007).
4) Janin, A., Baron, D., Edwards, J., Ellis, D., Gelbart, D., Morgan, N., Peskin, B., Pfau, T., Shriberg, E., Stolcke, A. and Wooters, C.: The ICSI Meeting Corpus, *Proc. ICASSP 2003*, Vol.1, pp.364–367 (2003).
5) Metze, F., Waibel, A., Bett, M., Ries, K., Schaaf, T., Schultz, T., Soltau, H., Yu, H. and Zechner, K.: Advances in Automatic Meeting Record Creation and Access, *Proc. ICASSP 2001*, Vol.1, pp.601–604 (2001).
6) Yu, H., Clark, C., Malkin, R. and Waibel, A.: Experiments in Automatic Meeting Transcription Using JRTk, *Proc. ICASSP'98*, Vol.2, pp.921–924 (1998).
7) 後藤真孝，緒方　淳，江渡浩一郎：PodCastle: ユーザ貢献により性能が向上する音声情報検索システム，人工知能学会論文誌，Vol.25, No.1, pp.104–113 (2010).
8) Goto, M., Ogata, J. and Eto, K.: PodCastle: A Web 2.0 Approach to Speech Recognition Research, *Proc. of Interspeech 2007*, pp.2397–2400 (2007).
9) Ogata, J. and Goto, M.: PodCastle: Collaborative Training of Acoustic Models on the Basis of Wisdom of Crowds for Podcast Transcription, *Proc. of Interspeech 2009*, pp.1491–1494 (2009).
10) Goto, M. and Ogata, J.: PodCastle: A Spoken Document Retrieval Service Improved by Anonymous User Contributions, *Proc. of PACLIC 24*, pp.3–11 (2010).
11) 緒方　淳，松原勇介，後藤真孝：PodCastle: 集合知に基づく Web キーワードを考慮した言語モデリング，日本音響学会講演論文集，pp.97–100 (2008).
12) Ogata, J., Goto, M. and Eto, K.: Automatic Transcription for a Web 2.0 Service to Search Podcasts, *Proc. of Interspeech 2007*, pp.2617–2620 (2007).
13) 緒方　淳，後藤真孝：PodCastle: 動的言語モデリングに基づくポッドキャスト音声認識，情処研報 音声言語情報処理 2010-SLP-84-2 (2010).
14) Gales, M. J.F.: Maximal likelihood linear transformations for HMM-Based speech recognition, *Computer Speech & Language*, Vol.12, pp.75–98 (1998).
15) 今井　亨，小林彰夫，佐藤庄衛，本間真一，奥　貴裕，都木　徹：放送用リアルタイム字幕制作のための音声認識技術の改善，第 2 回音声ドキュメントワークショップ講演論文集，pp.113–120 (2008).
16) 安藤彰男他：音声認識を利用した放送用ニュース字幕制作システム，信学論 (D-II), Vol.J84-D-II, pp.887–887 (2001).
17) Akita, Y., Mimura, M. and Kawahara, T.: Automatic Transcription System for Meetings of the Japanese National Congress, *Proc. of Interspeech 2009* (2009).
18) 秋田裕哉，三村正人，Neubig, G.，河原達也：国会音声認識システムの音響・言語モデルの半自動更新，情処研報音声言語情報処理 2010-SLP-84-3 (2010).
19) Kawahara, T., Nanjo, H., Shinozaki, T. and Furui, S.: Benchmark test for speech recognition using the corpus of spontaneous japanese, *Proc. SSPR 2003* (2003).
20) 小暮　悟，西崎博光，土屋雅稔，富樫慎吾，山本一公，中川聖一：日本語講義音声コンテンツコーパスの構築と講義音声認識手法の検討，第 2 回音声ドキュメントワークショップ講演論文集，pp.7–14 (2008).
21) Ogata, J. and Goto, M.: Speech Repair: Quick Error Correction Just by Using Selection Operation for Speech Input Interfaces, *Proc. of Interspeech 2005*, pp.133–136 (2005).

⋆1 http://jk.nicovideo.jp/
⋆2 http://www.jp.playstation.com/ps3/torne/