Vol.2015-SLP-108 No.2
2015/10/30

音声認識を用いた講義・講演の字幕作成・編集システム

秋田 祐哉[1,3]　三村 正人[2]　河原 達也[2,3]

概要：オープンコースウェアなどの講義・講演アーカイブにおいて，音声や映像に字幕を付与することは，専門的な内容の理解や障碍者・非ネイティブの利用者の視聴支援に有用である．しかし，このためには専門的な内容を書き起こした上で音声と同期を取る必要があり，作成に大きな手間がかかる．これに対して，我々は音声認識を用いた効率的な字幕付与の研究を進めている．この一環として，インターネット上のサーバで一般のユーザから講義・講演のコンテンツを受け付け，音声認識を用いて自動的に字幕の草稿を作成し，字幕の編集用に設計されたエディタでユーザがこの草稿を編集できるシステムを開発している．本稿では，この字幕作成・編集システムのあらましを述べ，実際の字幕の作成例について報告する．このシステムと関連して，我々は音声認識を用いて講義・講演の会場でリアルタイムに字幕を作成する枠組みについても検討しており，あわせて報告する．

An Online Caption Generation and Editing System using Automatic Speech Recognition

Yuya Akita[1,3]　Masato Mimura[2]　Tatsuya Kawahara[2,3]

Abstract: In an audio/video archive of academic and classroom lectures such as OpenCourseWare, captions (subtitles) are helpful for better understanding of its technical contents, and are essential aids for handicapped or non-native users. However, it is a time-consuming work to transcribe a technical lecture and align transcripts to its audio. Therefore, we have been engaging research of efficient creation of captions by using automatic speech recognition (ASR) technology. As a part of this research, we have developed a captioning system which consists of an ASR server and an editor. The server accepts audio/video materials from the public, and then generates caption drafts by ASR. Users can revise the drafts by using the editor which is designed to easily edit caption texts. Here, we report the proposed captioning system, together with trials of caption creation with this system. We also report real-time captioning of lectures using ASR.

1. はじめに

近年では，TED[*1]やオープンコースウェア（OCW）[*2]など，インターネットを通じて講義や講演の映像・音声を広く一般に公開する動きが活発になっている．これらの視聴に際して，難解な専門用語が少なくないこと，必ずしも音声の聴取ができる利用場面とは限らないこと，また聴覚障碍者や非ネイティブの視聴者は音声だけでは内容の理解が難しいことから，視聴の一助として字幕が有用である．一部の講義・講演では要約筆記などによるリアルタイムの字幕の提供も行われるようになってきており，作成した字幕をその場で聴衆に提供するだけでなく，事後に映像・音声のアーカイブとともに配信することも考えられる．

しかし，リアルタイムであっても事後であっても，手作業による書き起こし・字幕付与作業を広範な講義・講演を対象として行うのは人的資源の面で実質的に困難である．これは，書き起こしが時間を要する作業であることに加えて，講義・講演のような専門性の高い内容では，そもそも話題を理解できる作業者でないと正確な聞き取り，書き起こしが難しいことによる．

1 京都大学 経済学研究科
Graduate School of Economics, Kyoto University
2 京都大学 情報学研究科
Graduate School of Informatics, Kyoto University
3 京都大学 学術情報メディアセンター
Academic Center for Computing and Media Studies, Kyoto University

*1 www.ted.com
*2 www.ocwconsortium.org

これに対して，我々は字幕の草稿を音声認識技術を用いて自動的に作成することを検討している．近年の話し言葉音声認識は大きく進展しており，たとえば『日本語話し言葉コーパス』（CSJ）の講演音声認識では 80%を超える単語認識精度が報告されている [1]．大学の講義音声に関しても多くの大学でそれぞれ取り組みが進められており [2], [3], [4], [5], [6]，我々も OCW のコンテンツの音声認識に取り組んでいる [7], [8]．これらは主に事後的に行われる音声認識であるが，リアルタイムの字幕作成の取り組みも行われている [9], [10], [11]．音声認識では短時間に全て書き起こせるうえ，あらかじめ登録しておけば専門用語などの認識も可能である．したがって，十分な精度で音声認識を行うことができれば，それを編集することで効率的に字幕を作成することが期待できる．ただし，音声認識システムを対象に合わせて構成することは専門的なスキルが必要である．

そこで本研究では，専門家・技術者でない利用者でも字幕の作成・編集が行えるよう，入力された音声に対して自動的に音声認識をセットアップ・実行して草稿を作成するサーバと，字幕草稿を編集するために設計されたエディタからなる，字幕作成・編集システムを提案する[*3]．音声認識を用いた自動的な字幕作成は，たとえば YouTube[*4][12] や UD トーク[*5]でも行われているが，提案システムでは音声に対して具体的な資料を与えて言語モデルや単語辞書の適応処理を行うことができる．これと関連して，本研究ではリアルタイムに字幕を作成する枠組みについても検討しており，本稿ではあわせて報告する．

2. 字幕作成・編集システムの構成

本システムでは，ユーザにより収録された講義・講演や討論などの音声・映像に対して，事後的に字幕を付与することを想定している．**図 1** にシステムの利用の流れを示す．まず，ユーザがこれらのコンテンツを字幕サーバにアップロードする．音声・映像に加えて，言語モデルを話題に適応させるために，コンテンツの話題と関連するテキスト（たとえば講演予稿やスライド）もアップロードすることができる．字幕サーバではコンテンツからの音声の抽出および検査が行われ，ユーザの指定に応じて自動的に音声認識システムが構成された上で認識処理が実行される．認識処理はおおむね 1 日以内に終了し，音声と同期した字幕ファイルがサーバ上に出力されるとともに，これらにアクセスするためのアドレスがユーザに通知される．なお，ここで音声認識結果を用いる代わりに，あらかじめ人手で書き起こされた正しいテキストを与えて，音声への同期処理のみを行うこともできる．

[*3] caption.ar.media.kyoto-u.ac.jp
[*4] www.youtube.com
[*5] udtalk.jp

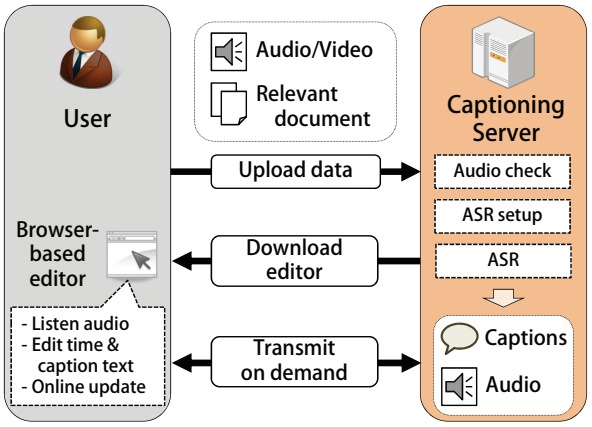

図 1 システムの利用の流れ

ユーザは通知されたアドレスから字幕をダウンロードできる．また，Web ブラウザ上で字幕エディタをダウンロード・起動して，サーバ上の音声を聴取しながら字幕のテキストや時刻を編集することもできる．編集した結果はサーバ上に保存され，更新された字幕ファイルとして取得可能である．

3. 字幕サーバ

本節では字幕サーバのあらましを述べる．字幕サーバは，ユーザからのコンテンツを処理するフロントエンド，音声認識，後処理・字幕生成の 3 つの機能で構成されている．

3.1 フロントエンド

字幕サーバには，コンテンツとして PCM（Microsoft WAV）や MP3 形式の音声のほか，MPEG 等の映像ファイルを入力することができる．入力がいずれの形式であっても，フロントエンドで 16kHz・16bit のモノラル音声に変換されて処理される．なお，音声に対してはビットレートや周波数の分布，SN 比のチェックが行われ，これらが一定の品質条件を満たさない場合は音声認識に進まず処理を中止する．

コンテンツに関連する文書としては，プレーンテキストのほか PDF や Microsoft Word/PowerPoint などの文書も受け付ける．フロントエンドによりこれらの文書から自動的にテキスト部分が抽出され，次節で述べる音声認識のための言語モデル適応に用いられる．

3.2 音声認識

本システムでは，講演や討論など，コンテンツの種類に応じていくつかの音響モデル・言語モデルの組み合わせをプロファイルとして用意しており，ユーザは実際に音声認識に利用するプロファイルをコンテンツのアップロードの際に選択することができる．現時点でのプロファイルの一覧を**表 1** に示す．たとえば，講演には CSJ の学会講演デー

表 1　モデルのプロファイルの一覧

名称	講演	ビデオ講義	討論
学習データ	CSJ（学会講演）	CSJ（学会講演） ＋放送大学音声	国会音声
音響モデル	DNN-HMM	DNN-HMM	GMM-HMM
言語モデル	単語 Trigram		

タから学習したモデル [13], [14] を，討論には国会音声・会議録から学習したモデル [15] を用意している．

サーバで行われる処理の流れを図 2 に示す．前述した音声品質チェックの後，コンテンツとともに関連文書が与えられている場合は，選択されたプロファイルの言語モデルに対してテキスト混合に基づく適応が行われる．混合重みの推定は難しいため，適応用のテキストの総単語数が一定の程度になるよう頻度のスケーリングのみ行っている．なお，与えられたテキストに出現する単語は原則として全て単語辞書に登録される．

次に，音響モデルが GMM-HMM の場合は，音声の話者区間の推定・分割を行う．これにはベイズ情報量基準（BIC）に基づく手法を用いる．話者区間に分割するのは後段の声道長正規化（VTLN）のためであり，DNN-HMM モデルの場合は VTLN を行わないので，セグメンテーションは省略される．

1 回目の音声認識は，GMM-HMM モデルの場合は VTLN のワープ係数を推定することが目的のため，簡易なモデル・パラメータで行われる．この結果を用いて，音声の区間ごとに VTLN を適用し，2 回目の音声認識を行う．話者が 1 名の場合は[*6]，この認識結果を教師なし音素ラベルとして，さらに MLLR による話者適応を音響モデルに適用したのち 3 回目の音声認識を行う．DNN-HMM モデルでは VTLN および MLLR 話者適応を行わないため，1 回目の音声認識のみ行う．本システムで用いる音声認識エンジンは Julius[*7]である．

3.3　字幕生成

音声認識結果には各単語の推定時刻が付与されているので，これを表示のタイミングとして，認識文からなる字幕ファイルを作成する．この際，音声認識結果には文や節の境界は与えられていないため，句読点の自動推定 [16] を用いて字幕の行に分割する．また，フィラーや口語表現，文末表現などの冗長部分を削除・修正するため，自動整形手法を適用する [17].

一方，あらかじめ人手による字幕テキストが与えられた場合は，このテキストと音声認識結果との間で文字単位のアライメントを行って時刻を字幕テキストに付与し，字幕ファイルを作成する [7].

*6　現時点で話者クラスタリングが実装されていないため．
*7　julius.osdn.jp

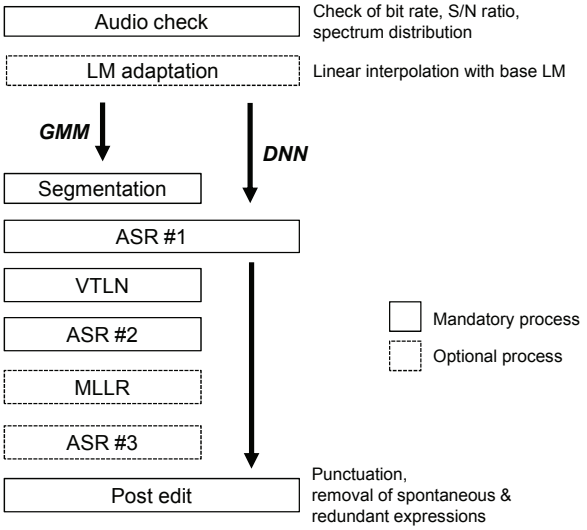

図 2　サーバ上の処理の流れ

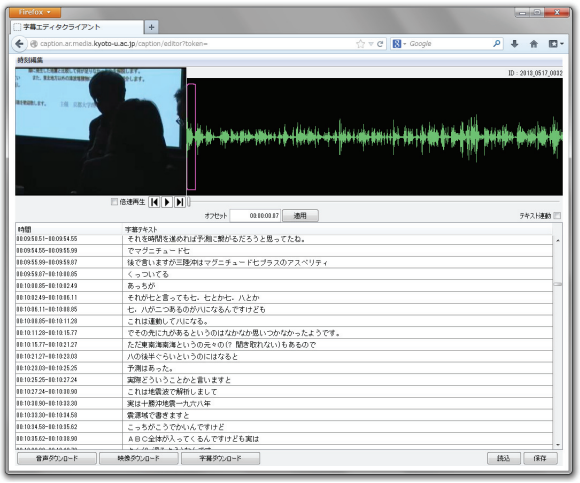

図 3　エディタの画面

字幕ファイルは SAMI・SRT など，複数の形式で出力される．これらの字幕はサーバに保存され，インターネットからアクセスが可能である．

4.　字幕エディタ

本システムでは生成された字幕を編集するためのエディタを提供する．音声認識結果の編集システムには様々な商用ソフトウェアがあり，またサーバ上で音声認識結果を編集するアプリケーションとしては PodCastle [18] が知られている．これらに対して，字幕に特化して行単位でテキストと時刻を編集するエディタであること，サーバ上のデータをオンラインで聴取・編集できることが本エディタの特徴といえる．オンラインの編集により，任意の場所や多様な環境で作業が可能である．

エディタは Java アプリケーションとして実装されている．字幕の編集は，語句の修正や文の長さ・タイミングの調整を主に行うフェーズと，最終的な字幕の出力を確認す

表 2　字幕作成対象の講義

科目名	講義数	文字正解率	平均作業時間			実時間比
			編集時間	確認時間	合計	
リスク社会のライフデザイン	12	88.5%	3 時間 46 分	46 分	4 時間 33 分	6.1
心理臨床の基礎	15	90.8%	3 時間 16 分	40 分	3 時間 56 分	5.2

文字正解率は作成した字幕を正解と見なして算出しているため，厳密なものではない．

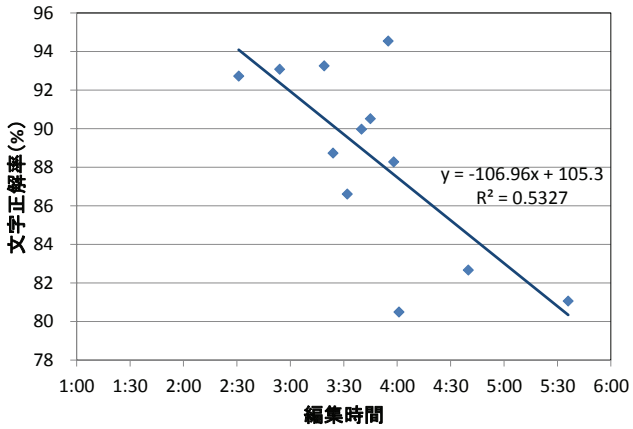

図 4　文字正解率と編集時間の相関（リスク社会のライフデザイン）

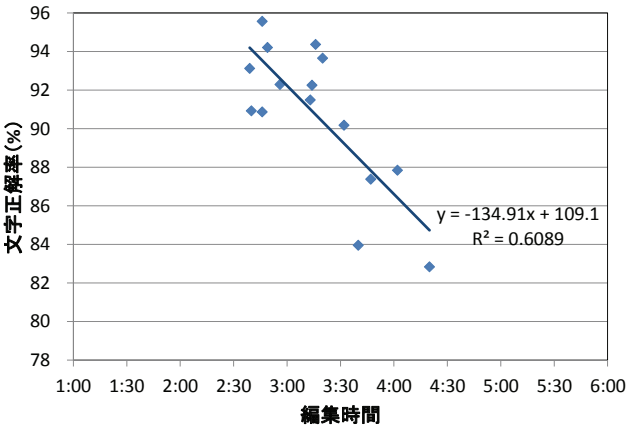

図 5　文字正解率と編集時間の相関（心理臨床の基礎）

るプレビューのフェーズに大別できることから，本システムではそれぞれに合わせた 2 つのエディタを設計した．

編集の際は，まず指定されたアドレスに Web ブラウザでアクセスすると，エディタがブラウザ上にダウンロードされたのちに起動し，サーバから字幕や音声などのデータを取得する．**図 3** は字幕エディタの実際の画面である．エディタ上では字幕の時刻やテキストを入力・編集できる．これに加えて音声波形も表示され，アップロードされた元のコンテンツが映像の場合は一定時間ごとのサムネイル画像もあわせて表示される．エディタは音声の再生も可能で，音声の再生にあわせて字幕テキスト・音声波形・サムネイル画像が連動して表示される．行の結合や分割も容易であり，行操作を行った場合は字幕の時刻がその都度自動調整される．エディタ上でなされた変更は，一定時間ごとにサーバに送信されてファイルが更新される．プレビュー用のエディタでは，音声波形に代わって，字幕が指定されたタイミングで表示されるインターフェースとなる．再生速度の調整も可能で，音声の早回しをしながら迅速にチェックすることができる．

なお，ユーザがローカルに保持する音声をサーバ上の音声の代わりに使用して，オフラインで編集することもできる．この場合は，作成された字幕ファイルをいったんダウンロードした上で，オフライン版のエディタ（スタンドアロンのアプリケーション）を起動して編集する．

5. 放送講義における字幕作成

本研究では，提案システムによる字幕作成の効率を測定するため，実際の講義音声を用いて字幕の作成を行った．ここで利用したのは放送大学で実施されたラジオ講義で，**表 2** に挙げた 2 科目，計 27 講義である．講義の長さは 1 件あたり 45 分となっている．なお，本研究ではこれらの講義に対して正確な書き起こしを作成していないため，表 2 に示した文字正解率は作成した字幕と音声認識結果を比較して算出したものであり，厳密なものではない．

これらの講義音声を字幕サーバに入力して，得られた音声認識結果を字幕エディタにより編集して字幕を作成した．言語モデルの適応用テキストとしては，講義の教科書に加えて，台本がある場合はこれも使用した．ただし台本は必ずしも全ての発話を網羅しているわけではなく，また正確とは限らない．字幕作成にあたった作業者は 1 名で，計算機の一般的な操作スキルはあるが，字幕作成の専門家ではなく，またこれらの講義内容の専門家でもない．作業にあたり，講義ごとに所要時間を計測し，あわせて前述した文字正解率を算出した．表 2 で挙げた編集時間は主に認識誤りの修正や文の長さ・タイミングの調整に要した時間で，確認時間はプレビュー用のエディタ上で行った出力チェックに要した時間である．実時間比は，合計の作業時間を講義の長さ（45 分）で除したものである．平均して実時間の 5.2〜6.1 倍の作業時間となっており，このうちほぼ実時間分は確認作業に費やされていることがわかる．

科目ごとに各講義の文字正解率と編集時間をプロットしたものを**図 4・図 5** に示す．相関の程度を測るため，これ

らの図では近似直線とその決定係数 R^2 の値も示している．これらの図より，文字正解率と編集時間には強い相関があることがわかる．同様に放送大学の講義を対象とした字幕作成の報告 [19] では，45 分の講義に対して約 4 時間の書き起こし時間を必要としていることから，図 4・図 5 中の近似直線より，86%〜87%以上の精度が得られれば，はじめから人手で書き起こすよりも効率的であるといえる．

6. リアルタイムの字幕作成システム

これまでに述べた字幕作成は配信を前提とした事後的な処理である．一方，講義・講演の会場で情報保障のためにリアルタイムに字幕を作成・表示するシステムの開発にも我々は取り組んでいる．

講義や講演において，文字情報による情報保障の手段としては，手書きのノートテイクや PC を用いた要約筆記などが一般的に用いられている．手書きにより書き起こせる分量は限られており，遅延をともなう．また，いずれの方法であっても作業者が長時間作業し続けることができないため，複数の作業者を用意する必要があり，情報保障を提供する上での支障となっている．これに対して，音声認識により精度よく書き起こすことができれば作業の負担が軽減でき，より少ない人数（たとえば 1 人）で作業できることから，情報保障の機会を拡大することができる．

我々が検討している，音声認識を用いたリアルタイム字幕作成システムの構成を図 **6** に示す．本システムは，講義・講演会場で作業者が入出力・編集に使用する PC と，音声認識を行うサーバから構成される．PC とサーバはネットワークを通じて通信するため，十分な通信帯域が確保できれば，サーバは遠隔地に設置することができる．

講師の音声は PC に入力され，Julius 付属の音声入力ツールである Adintool によって発話検出・セグメンテーションを行ったのち，サーバ側の Julius にネットワーク経由で送信される．サーバではあらかじめ対象の講義・講演用に構成された音響モデル・言語モデルを使用して音声認識を行い，この結果をネットワーク経由で作業者の PC に送信する．なお，本研究では音響モデル・言語モデルとして表 1 に挙げた講演プロファイルに相当するものを主に使用している．GPU を用いてデコードすることにより，DNN-HMM 音響モデルでも実時間以下の処理時間で音声認識を行っている．

作業者の PC では，PC 要約筆記で一般的に用いられている IPtalk[*8] を字幕の編集ツールとして使用する．IPtalk では複数の要約筆記者がネットワーク経由で連携して入力することができるが，本システムでは Julius の出力をブリッジするツール Julius2IPtalk[*9] を使用して，IPtalk の「確認・編集パレット」に認識結果を入力する．作業者が

[*8] www.geocities.jp/shigeaki_kurita/
[*9] www.ar.media.kyoto-u.ac.jp/jimaku/julius2iptalk.html

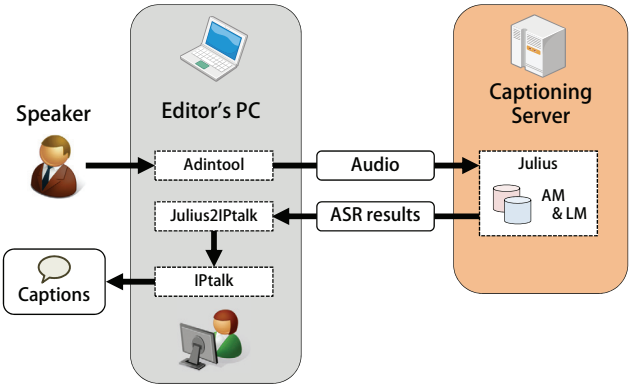

図 6　リアルタイム字幕作成システムの構成

認識結果をチェック・修正した上で送出すると，字幕として表示される．字幕は，PC に接続したプロジェクタや，ネットワークを通じた Web 配信などで表示可能である．

7. おわりに

本稿では，効率的な字幕作成を目的として我々が開発を進めている字幕作成・編集システムについて報告した．本システムでは，字幕サーバにアップロードされたコンテンツに対して音声認識を行って字幕の草稿を自動的に作成する．この草稿を字幕エディタを用いて編集することにより最終的な字幕とする．実際の講義音声を用いて本システムにより字幕作成を行ったところ，認識精度と作業時間には相関があり，86%〜87%の精度であれば実用的であることがわかった．また本稿ではリアルタイムに字幕を作成するシステムについても報告した．今後はそれぞれのシステムでより多くの検証を重ねていきたい．

謝辞 本研究の一部は科学研究費補助金 25730112 によって行われた．講義データをご提供いただいた，放送大学教授広瀬洋子先生に感謝いたします．

参考文献

[1] Masumura, R., Asami, T., Oba, T., Masataki, H., Sakauchi, S. and Ito, A.: Latent Words Recurrent Neural Network Language Models, *Proc. Interspeech*, pp. 2380–2384 (2015).
[2] Trancoso, I., Nunes, R., Neves, L., Viana, C., Moniz, H., Caseiro, D. and Mata, A.: Recognition of Classroom Lectures in European Portuguese, *Proc. Interspeech*, pp. 281–284 (2006).
[3] Glass, J., Hazen, T., Cyphers, S., Malioutov, I., Huynh, D. and Barzilay, R.: Recent Progress in the MIT Spoken Lecture Processing Project, *Proc. Interspeech*, pp. 2553–2556 (2007).
[4] Yamazaki, H., Iwano, K., Shinoda, K., Furui, S. and Yokota, H.: Dynamic Language Model Adaptation Using Presentation Slides for Lecture Speech Recognition, *Proc. Interspeech*, pp. 2349–2352 (2007).
[5] Togashi, S. and Nakagawa, S.: A Browsing System for Classroom Lecture Speech, *Proc. Interspeech*, pp. 2803–2806 (2008).

[6] Wiesler, S., Irie, K., Tüske, Z., Schlüter, R. and Ney, H.: The RWTH English Lecture Recognition System, *Proc. ICASSP*, pp. 3310–3314 (2014).
[7] 秋田祐哉，河原達也：オープンコースウェアを対象とした音声認識に基づく字幕付与，日本音響学会春季研究発表会講演論文集，2-9-9 (2013).
[8] Mimura, M. and Kawahara, T.: Unsupervised Speaker Adaptation of DNN-HMM by Selecting Similar Speakers for Lecture Transcription, *Proc. APSIPA ASC* (2014).
[9] Cerva, P., Silovsky, J., Zdansky, J., Nouza, J. and Malek, J.: Real-Time Lecture Transcription using ASR for Czech Hearing Impaired or Deaf Students, *Proc. Interspeech*, pp. 3343–3344 (2012).
[10] Ranchal, R., Taber-Doughty, T., Guo, Y., Bain, K., Martin, H., Robinson, J. and Duerstock, B.: Using Speech Recognition for Real-Time Captioning and Lecture Transcription in the Classroom, *IEEE Trans. Learning Technologies*, Vol. 6, No. 4, pp. 299–311 (2013).
[11] 桑原暢弘，秋田祐哉，河原達也：音声認識結果の有用性の自動判定に基づく講義のリアルタイム字幕付与システム，日本音響学会春季研究発表会講演論文集，2-4-5 (2014).
[12] Liao, H., McDermott, E. and Senior, A.: Large Scale Deep Neural Network Acoustic Modeling with Semi-supervised Training Data for YouTube Video Transcription, *Proc. ASRU* (2013).
[13] 三村正人，河原達也：話し言葉音声認識タスクにおける音素誤り最小化学習（MPE）の効果，日本音響学会秋季研究発表会講演論文集，3-Q-8 (2007).
[14] 三村正人，河原達也：CSJ を用いた日本語講演音声認識用 DNN-HMM の構築，日本音響学会秋季研究発表会講演論文集，1-P-42b (2013).
[15] 秋田祐哉，三村正人，河原達也：会議録作成支援のための国会審議の音声認識システム，電子情報学会論文誌，Vol. J93-D, No. 9, pp. 1736–1744 (2010).
[16] 秋田祐哉，河原達也：講演に対する読点の複数アノテーションに基づく自動挿入，情報処理学会論文誌， Vol. 54, No. 2 (2013).
[17] Neubig, G., Akita, Y., Mori, S. and Kawahara, T.: A Monotonic Statistical Machine Translation Approach to Speaking Style Transformation, *Computer Speech and Language*, Vol. 26, No. 5, pp. 349–370 (2012).
[18] 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，情報処理学会研究報告，2010-SLP-84-2 (2010).
[19] 長妻令子，福田健太郎，柳沼良知，広瀬洋子：クラウドソーシングを活用した効率良い字幕作成手法，電子情報通信学会技術研究報告，WIT-65-2 (2012).