Vol.2016-AAC-1 No.6
2016/7/29

音声自動認識による字幕情報保障トライアル

平賀瑠美[†1]　秋田祐哉[†2]

概要：　研究発表を含め，様々な場面で音声自動認識を用いた字幕情報保障への期待が高まっている．本セッションでは，健聴者の研究発表とそれに対する質疑で音声認識の結果を発表中の調整なしに字幕化する．これは，研究会や全国大会で音声認識による字幕情報保障を行う場合を模したものであり，このような字幕表示が実際の研究発表の場面で可能なのかどうかを参加者全員で考える．システムによる研究発表の字幕化が不調の場合も想定し，研究発表の読み上げ原稿を作成して必要な人には配り情報伝達の不備に備える．今後も引き続き，音声自動認識による字幕提示システムが研究発表の場に利用できるのかどうか，について，様々なシステムに検証に参加してもらいたいと考えている．

キーワード：情報保障，音声認識，研究会

Captioning by Speech Recognition for Research Presentation (Trial)

Rumi Hiraga[†1]　Yuya Akita[†2]

Abstract: We expect to use speech recognition systems for captioning at research presentations. In this session, a speech recognition system captions a presentation without any editing that is the mimic of the actual use of a speech recognition system at a research presentation. Participants of the session witness the current speech recognition system for captioning as it is. In case the caption is not well produced by the speech recognition system, reading manuscript is given to people that may difficult to get information from the presenter's speech. We hope that many other speech recognition systems participate this trial to see that systems are available for captioning research presentation.

Keywords: Information support, Speech recognition, Research presentation

1. はじめに

情報処理学会では 2015 年度より，情報保障の予算を計上し，必要とする研究会からの申請に応じて発生する費用を補助する仕組みを設けた．2016 年度には障害者差別解消法も施行され，アクセシビリティ研究グループは，合理的配慮をできるだけ行いたいと考えている．

本セッションは情報処理技術をアクセシビリティの現場に生かす音声自動認識に焦点を置き，音声自動認識による研究発表の字幕付与の現状を研究会参加者と検証して，問題点や期待を洗い出そうとするものである．

なお，本稿における字幕とは事前に音声内容を得て字幕をつけるものではなく，リアルタイムで付与を行う字幕を指し，要約筆記とほぼ同じ意味で用いる．また，聴覚障害者の情報保障について，字幕情報保障に限ったものと考えるものではまったくない．

1.1 研究発表における字幕情報保障

情報処理学会アクセシビリティ研究会は，研究グループとして 2015 年に始まり,障害を持つ当事者の多数の参加を期待することから，少なくとも字幕情報保障を行うことにしているa)．また，情報処理学会では，京都大学で行われた 2014 年度第 77 回全国大会からいくつかのセッションで字幕情報保障を行いb)，2015 年度も引き続き慶應義塾大学での全国大会で字幕情報保障を行った．

研究会や全国大会での字幕情報保障は，以下の役割を持つと考える．

1) 研究講演を聞きに来る聴覚障害者に対して発表内容を正確に伝える．
2) 研究講演を聞きにくる聴覚障害者に対して質疑応答を正確に伝える．
3) 研究講演を行う聴覚障害者の発表内容を正確に参加者に伝える．
4) 研究講演を行う聴覚障害者に対する質疑とその応答を正確に伝える．

このように研究発表会での字幕情報保障は，聴覚障害者に対して情報を提供するだけではなく，聴覚障害者からの情報を伝えるという双方向性を持つことと，正確な情報伝達が要求される．聴覚障害者が口頭で発表し，その内容を字幕表示する場合，字幕表示の担当者は事前に発表者から発表の読み上げ原稿を受け取ることがあり，そのことで、発音が不明瞭である場合でも，発表を正しく伝えられるようにしている．3) に記した聴覚障害者が口頭で発表する場

†1 筑波技術大学
Tsukuba University of Technology
†2 京都大学
Kyoto University

a) 2015 年度は、字幕情報保障以外にも，磁気ループと手話通訳の提供ならびに移動支援を行った．
b) 要望により，手話通訳も提供した．

ⓒ2016 Information Processing Society of Japan

合は，1) の健聴者の口頭発表よりも音声認識が難しくなることは必至であるが，聴覚障害者が研究発表会に参加するのは，他の人の研究発表を聞きに行くためだけではないことを忘れてはならない．実際，情報保障がついた情報処理学会第 77 回，第 78 回全国大会では，聴覚障害者が発表を行った．

聴覚障害者が研究講演においてパワーポイントなどの発表資料と同時に字幕を見て発表内容を知り，考え，質疑応答に参加するためには，発表者が提示している発表資料と字幕の内容の提示の時間にずれがあると難しくなることが想定される．つまり，双方向性と正確さという上の二つの要求に加え，字幕情報保障には実時間性が望まれる．

これらのことは，研究発表においての字幕情報保障に限り必要とされることではく，状況によってはさらに場の雰囲気を伝えるといったことも重要となろう．先に記した研究発表を聞く，研究発表を行う，という場面で必要とされる三つは，聴覚障害者の字幕情報保障として基本的な事項である．

研究発表会においての字幕情報保障は，現在，訓練を経た筆記者が発表場所，あるいは，遠隔地で行っているが，いずれの場合も，依頼者は情報保障を行う法人などに対し支払いを行う．また，人手による入力のため，どうしても，発表との時間差が生じてしまう．音声自動認識を用いた字幕付与システムについては，報道で見ることも多くなってきており，研究発表会でのこのようなシステムの利用に対する期待も高まっている．

1.2 研究発表における合理的配慮

昨年度，アクセシビリティ研究グループc)が行ったシンポジウムにおける情報保障の費用は，情報処理学会の補助金の制度を利用してきた．この制度は，年度始めに決まった金額を準備し，必要とする研究会の申請をその都度委員会が審査し，許諾を出すというものである．年度途中でも予算が底をついたら，情報保障補助は打ち切りとなる．

先にも記したが，アクセシビリティ研究会は障害を持つ当事者になるべく多く参加してもらいたいと考えているため，情報保障に掛かる経費を補助してもらえることに非常に感謝している．昨年度この補助は字幕情報保障の依頼に対して使用し，手話通訳は聴覚障害者の所属大学の提供で行った．昨年度の手話通訳に関する費用は学会にお願いせずに済んだが，字幕，さらに手話通訳の二つの依頼補助をお願いできるのだろうか．

今年度から障害者差別解消法が施行となり，日本中で合理的配慮が進むことが期待されているが，合理的配慮実現のためには新たな“コスト”が関係者に要求される．福島智東京大学先端科学技術研究センター教授が危惧しているような「コストのかかることは無理しなくていいと認めているから，結果的に，サービスや支援を低いレベルで平準化させる恐れ」d)がないようアクセシビリティ研究会はコストを抑えつつ，より高いレベルでの合理的配慮の実現のための情報処理理技術の活用を進めたいと考える．

1.3 音声自動認識による字幕付与トライアル

字幕付与の実時間性と費用，また，モビリティの点から，音声自動認識による字幕付与への期待は大きい．また，「音声自動認識による字幕が使われました」という報道を目にすることも多くなってきている．しかし，少なくとも研究発表会でどの程度現状の音声自動認識による字幕付与が可能かどうかについては，当事者を交えて調べられたこと，あるいはその結果を学会で公表されたことはあったのだろうか．2016 年 10 月に Reno で行われる ACM ASSETS (The 18th International ACM SIGACCESS Conference on Computers and Accessibility) では，Captioning Challenge が予定されているe)．”competition”とは言っても，評価の方法もこれから決めましょう，ということで，ICT 利用の字幕付与研究を盛んにしたいという思いも込めたチャレンジのようである．使用される言語という制限があるため，日本の研究がそのチャレンジに参加することが必ずしも容易ではないとは思うが，このように字幕付与に関する情報処理研究が盛んになることは望ましいことである．

本発表では，研究会では音声自動認識による字幕付与がどのように行われ，どのような課題があるのか，について情報共有をする．その後，実際に研究発表に対して，音声自動認識による字幕情報付与を行う．発表者は普段どおりの速度，音量で話す．発表者の音声はマイクからシステムへ入力されるが，発表中のシステム調整はない．発表後の質疑応答もマイクへの入力を行い，音声認識の結果を字幕として提示する．発表については，予め準備した読み上げ原稿を研究会参加者のうち必要な人に配布し，一部の人に対し情報伝達内容が欠けることを防ぐ．

この試みは，ある一つのシステム，一人の発表者で行うためうまく字幕が提示されるにせよ，されないにせよ，そのような場合がある，という一例でしかない．もしも，うまく提示されなければ，このシステムはまだ研究発表会に使うことは難しいということになり，質疑応答も含めてうまく提示される場合でも，他の発表者の場合について確認する必要がある．

2. 音声認識を用いた字幕作成

本節では，音声認識システムの構成について説明し，字幕作成への応用における問題，および実際の事例について述べる．

c) 今年度よりアクセシビリティ研究グループはアクセシビリティ研究会として活動を始めた．

d) 朝日新聞 2016 年 4 月 6 日朝刊，オピニオン & フォーラム「障害者とともに」の中の記事「コスト引き受ける覚悟」より．

e) http://assets16.sigaccess.org/captioning_challenge.html

ⓒ2016 Information Processing Society of Japan

2.1 音声認識の枠組み

音声認識システムの一般的な構成を図 1 に示す．音声認識システムは，音声を構成する音韻単位のスペクトル特徴を表す音響モデル，単語の発音を音韻単位で表す発音モデル（辞書），および単語の並びの言語的制約を記述する言語モデルの 3 つのモデルをもつ．入力音声から抽出した特徴量をx，発音をp，単語列をwとすると，これらはそれぞれ$P(x|p)$・$P(p|w)$・$P(w)$を表す確率的モデルである[f]．これらのモデルを用いて，次式に基づき，入力音声の特徴量xに対して最ももっともらしい単語列$\hat{w}$，すなわち音声認識結果を求めるのがデコーダである．

$$\hat{w} = \arg\max_{w} P(x|p)P(p|w)P(w)$$

したがって，音声認識システムの性能（何を書き起こせるか）は，これら 3 つのモデルによって規定される．入力音声に対してモデルにより展開される単語列の仮説は膨大であり，この仮説空間の中から，最も大きな確率を与える仮説を見落とすことなく，かつ効率的に探索することがデコーダの役割である．なお，これらのモデルは言語に依存しており，たとえば日本語用のシステムで英語を認識することはできない．

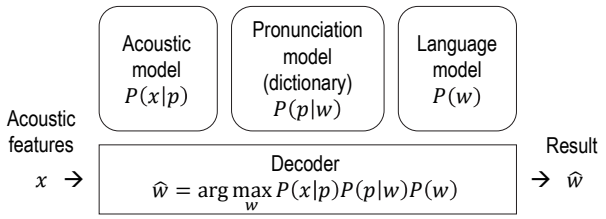

図 1　音声認識システムの構成

音響モデルでは，音韻単位として音素や音節が用いられる．音韻単位ごとに音声特徴量に対する確率を定めることになるが，音声のスペクトル特徴は 1 つの音韻単位内でも時間により変動するため，これを表現するために隠れマルコフモデル（HMM）が一般的に用いられる．近年では，HMM の確率のモデル化にディープニューラルネットワーク（DNN）を用いた DNN-HMM が利用されるようになってきている．いずれの場合も，音響モデルの確率は，数十～数千時間という大量の音声データから最尤または識別的な基準で統計的に学習される．音声には個人差があり，また収録された環境や機器によるひずみ等の影響も受けるため，認識対象とできるだけ合致した音声でモデルを構築することが望ましい[g]．特徴量の正規化や，モデルのパラメータの適応によりこれらに対処する手法も知られている．

言語モデルは，直前の N-1 単語（文脈）から次の単語の確率を定める，単語 N-gram モデルが広く用いられている．このような単語の連接関係は，大量のテキストデータベースをもとに統計的に学習される．また言語モデルの学習に際して語彙が定められ，これに基づいて発音辞書が構成される．音響モデルと同じように，言語モデルも認識対象の内容とできるだけ合致したテキストで学習することが望ましく，逆に言えば言語モデル・発音辞書でカバーされていない単語は認識することができない．文脈の長さ（つまり N）としては 3～5 が一般的である．文脈が長い方がより精密なモデルとなるが，より多くの学習データが必要となる．なお，言語モデルにおいても，近年ではニューラルネットワークによるモデルが用いられるようになってきているが，いったん出力した仮説を再評価（リスコアリング）して適用することが一般的で，リアルタイムの処理に適用することは現時点では難しい．

2.2 リアルタイム字幕への応用

音声認識をリアルタイムの字幕作成に用いる取り組みとしては，NHK による放送番組への字幕付与が挙げられる[1]．放送ニュースではアナウンサーの音声を直接認識し，スポーツ中継などでは別のアナウンサーにより復唱（リスピーク）した音声を認識して字幕の草稿としている．いずれの場合も音声認識の対象は訓練を受けたプロのアナウンサーであり，精度の高い音声認識結果が得られるが，放送番組では完全な字幕が求められるから，草稿を修正・確認した上で字幕として送出される．一方，学会講演と類似したものとして，大学講義で講師の音声を認識してノートテイク・字幕作成を行う取り組みがある[2][3][4]．ただし講義は講演以上に音声認識が難しく，精度の改善が講義のリアルタイム字幕の重要な課題である．

講演の音声認識は，おおむね講義と同じアプローチである．このような音声認識をリアルタイム字幕のために用いる場合，大きく分けて（1）音声入力，（2）音声認識用のモデル，（3）後処理，の 3 つの面で課題がある．

まず音声入力については，音声認識システムにはある程度の品質の音声を入力する必要がある．すなわち，雑音やひずみが小さく，十分な S/N（信号対雑音）比がある音声を，適切な音量制御のもとに入力する必要があるということである．学会講演は毎回異なる会場で開催されることが一般的であるから，それぞれの回ごとに検討が必要となる．多くの講演会場では講師が会場のマイクを使用するため，会場の音響設備（PA）から分配できれば良質な音声を得られる可能性は大きいが，それでも配線や出力信号レベルの調整などに注意を払う必要がある．PA を使用していない，あるいは分配や品質に難がある場合は音声認識用に独自のマイクを使用することになり，マイクの設置や調整，またマイクの追加により講師の負担が増えることが問題である．通常，このような場でのオペレータは音響の専門家ではないから，実施上の負荷は小さくない．

f) ただし発音モデルについては，確率をすべて 1 として扱うことも多い．
g) なお，1.1 節で述べた聴覚障害者の音声の認識は，一般的なモデル・データベースでは特徴が十分にカバーされていないため，現状では難しいと考えられる．

ⓒ2016 Information Processing Society of Japan

講演の音声はいわゆる「話し言葉」で，発話の音響的特徴や言語表現がディクテーションの場合とは大きく異なる．このため，音声認識には話し言葉用に構築された音響モデル・言語モデル・発音辞書を用いる必要がある．音響モデルは大規模な音声データベースから学習されるが，講演会場や機器の音響的特性，講師の個人性は学習用音声とは異なるため，その都度適応することが望ましい．しかしこのためには事前に会場や講師の音声を入手する必要があり，実際に行うことは難しい．また，学術的な講演では専門的な用語・表現が発話されるが，大規模テキストデータベースから構築した言語モデルや発音辞書であってもこれらをカバーすることは難しく，そのままでは認識できないことから，講演に適応することが事実上必須である．対象の講演に関連した資料で言語モデルを学習・適応することが望ましいが，少なくとも専門用語とその発音を発音辞書に登録する必要がある．さらに，自然科学系の講演ではしばしば数式や記号類が読み上げられるが，これらは言語制約として表現することが容易ではなく，認識が困難である．

音声認識結果には誤りが含まれる．これに加えて，話し言葉の音声では間をつなぐためのフィラーや冗長な文末表現などが多くみられる．人手の要約筆記では作業者がこれらを適宜省略して書き起こすことができるが，音声認識は全て書き起こしてそのまま出力するため，たとえ完全に正しい認識結果であっても出力は読みにくいものとなる．また，音声認識自体は句読点を挿入する枠組みを持たない．音声認識におけるそれぞれの出力文は単に発声の都合で区切られたもので，必ずしも句読点には対応しない．話し言葉の自動整形や文境界推定[5][6][7][8]も提案されているが，いずれにしても正確な字幕のためには音声認識結果の編集作業が必要となり，このために字幕の遅れが発生する．

2.3 実施事例

実際の学会講演に対する，音声認識を用いたリアルタイム字幕付与の試みを，情報処理学会アクセシビリティ研究グループシンポジウム（AAC，2015 年 8 月・2016 年 2 月）および音声言語情報処理研究会（SLP，2015 年 10 月）の一部の講演で実施した．ここでは音声認識結果を作業者ができる限り編集したものを字幕として提示した．また，情報処理学会の研究会では講演の映像をインターネットで配信しており，作成した字幕（AAC は編集したもの，SLP は音声認識結果そのもの）は映像と合わせて配信した．表 1 に字幕を付与した講演の件数と時間を示す．

表 1　字幕を付与した講演

	講演数	講演時間（合計）
AAC・2015 年 8 月	4	1 時間 52 分
AAC・2016 年 2 月	3	2 時間 8 分
SLP・2015 年 10 月	3	1 時間 15 分

図 2 にこの字幕付与システムの構成を示す．音声の収録は，SLP では独自マイクを使用し，AAC では会場の PA を用いた．音声認識には Julius デコーダh)を使用し，音響モデルは『日本語話し言葉コーパス』（CSJ）から学習した DNN-HMM モデルを使用した[9]．CSJ は講演の音声・テキストデータベースであり，日本語の話し言葉音声認識で一般的に用いられている．言語モデルも CSJ を用いて学習し，SLP では講演予稿のテキストを用いて話題への適応を行った．AAC では，情報保障のために読み上げ原稿が用意された講演はそれを使用しi)，それ以外の講演は予稿および必要に応じてインターネットから資料の収集を行って言語モデルを適応した．なお，音声認識は会場の編集端末では実行せず，音声が端末に入力されるたびにそのデータを京都大学に設置された字幕作成サーバj)[10]にインターネット（LAN または 4G モバイル回線）経由で送信して認識処理を行い，その結果を逐次的に取得した．これによる通信のオーバーヘッドは作業上の問題とはならなかった．音声の送信には Julius 付属の Adintool を，認識結果の取得には Julius2IPtalk k)を使用し，音声認識結果は PC 要約筆記で一般的に用いられているソフトウェア IPtalk l)に入力して編集を行って字幕として提示した．このとき，フィラーは自動的に削除し，また典型的な文末表現に限って句点を自動挿入したうえで IPtalk に入力している．

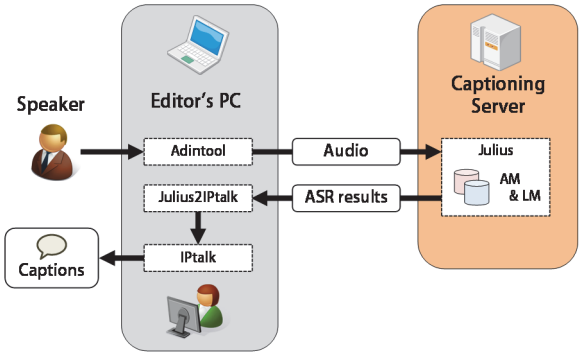

図 2　リアルタイム字幕作成システムの構成

これらの実施事例のうち，2016 年 2 月の AAC シンポジウムにおいて作成した字幕のあらましを表 2 に示す．このシンポジウムでは，講演のうち 3 件について，音声認識を用いてリアルタイム字幕を提供した．合計で 2 時間 8 分の講演時間に対して 1 名の作業者で編集作業を行い，32,601 文字の字幕を送出した．編集した文字数は 5,116 文字で，このうち半数は音声認識結果の削除であった．講演時間 1 分あたりでは 40 文字の編集である．ただし，編集作業には IPtalk の「確認・編集パレット」を使用しており，不要な

h) http://julius.osdn.jp/
i) 読み上げ原稿通りに発話する場合は，ほぼ正確に認識することができる．
j) http://caption.ist.i.kyoto-u.ac.jp/
k) http://sap.ist.i.kyoto-u.ac.jp/jimaku/julius2iptalk.html
l) http://www.geocities.jp/shigeaki_kurita/

ⓒ2016 Information Processing Society of Japan

認識文は一括削除できる．編集の半数が削除であるから，実際の作業負荷は 40 文字よりも小さかったといえる．なお，完全に正確な書き起こしを作成していないため，実際の音声認識の精度は明らかではないが，字幕の 92%は音声認識結果をそのまま用いている（残りの 8%は表 2 の置換・挿入にあたる）．

表 2　音声認識を用いて作成した字幕のあらまし
（AAC・2016 年 2 月）

作業者	1 名
字幕送出文字数	32,601
音声認識文字数	33,697
編集文字数	5,116（16%） 内訳：置換 1,190（3.7%） 挿入 1,415（4.3%） 削除 2,511（7.7%） （字幕送出文字数に対する割合）

3. トライアル

トライアルでは音声認識による字幕表示システムを用いて、研究発表と質疑応答の字幕付与を行う．システムに対し，読み上げ原稿の提供は行わず，認識結果に対する編集作業も行わない．

トライアルで発表する内容は、2015 年 4 月 18 日に京都大学で行われた「聴覚障害者のための字幕付与技術」シンポジウムで著者の一人が発表したものと同じである[m]．シンポジウム時は，（ヒトによる）要約筆記と手話通訳による情報保障が行われた．タイトルは「聴覚障害者と音楽」で，以下，発表内容を簡単に記す．

聴覚障害を持っていても音楽を好きで，よく聞く，カラオケに行く，ダンスサークルに属しているという人は多い．音楽に接する機会を増やすことで聴覚障害を持っていても能動的聴取の力を増すことができるのではないか，という仮説を持ち，音楽ゲームの多用が音楽に接する機会を増やす一つの方法であると考えた．この発表では，音楽ゲーム Music Puzzle（図 3）を開発し，それを用いた実験の結果を述べる．この結果から，耳鼻科での検査結果と音楽を理解する力は関係がないと考えられる[11]．ビデオや音楽も使用する発表である．

上記について 20 分の口頭発表，5 分程度の質疑応答を行う．これらについて，音声認識システムによる字幕表示を行う．その後，音声自動認識による研究発表に対する字幕付与に関するディスカッションは（ヒトによる）字幕をつけて行い，参加者全員で考えたい．

m) http://sap.ist.i.kyoto-u.ac.jp/jimaku/jimaku15.html

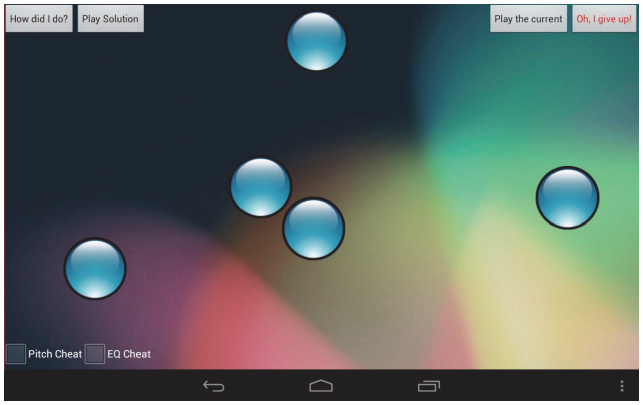

図 3 Music Puzzle のインタフェース

このトライアルは初回としてアクセシビリティ研究会運営委員が関わるシステムを用いる．研究発表会や情報処理学会などの全国大会で音声自動認識による字幕表示を提供しようとするシステムは，実際の場面で使用する前にこのようなトライアルで試してみてはいかがでしょう．

謝辞　音声認識については京都大学大学院情報学研究科教授　河原達也先生にご支援とご助言をいただいており，深く感謝いたします．音声認識の研究の一部は科学研究費補助金 16H02847 による．トライアルで発表する研究は科学研究費補助金 26282001 による．

参考文献

[1] 今井亨，奥貴裕，小林彰夫．音声認識によるリアルタイム字幕放送の進展．情報処理学会研究報告，SLP-88-4，2011.
[2] P.Cerva, J.Silovsky, J.Zdansky, J.Nouza and J. Malek. Real-Time Lecture Transcription using ASR for Czech Hearing Impaired or Deaf Students. In Proc. Interspeech, 2012.
[3] R.Ranchal, T.Taber-Doughty, Y.Guo, K.Bain, H.Martin, J.Robinson and B.Duerstock. 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.
[4] 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.
[5] G.Neubig, Y.Akita, S.Mori and T.Kawahara. A Monotonic Statistical Machine Translation Approach to Speaking Style Transformation. Computer Speech and Language, Vol.26, No.5, pp.349-370, 2012.
[6] Y.Fujii, K.Yamamoto and S.Nakagawa. Improving the Readability of ASR Results for Lectures using Multiple Hypotheses and Sentence-Level Knowledge. IEICE Trans. Inf. & Syst., Vol.E95-D, No.4, pp.1101-1111, 2012.
[7] 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, pp.463-470, 2013.
[8] 大野誠寛，村田匡輝，松原茂樹．講演のリアルタイム字幕生成のための逐次的な改行挿入．電気学会論文誌, Vol.133, No.2, pp.418-426, 2013.
[9] 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.
[10] 秋田祐哉，三村正人，河原達也．音声認識を用いた講義・講演の字幕作成・編集システム．情報処理学会研究報告，SLP-108-2，2015.
[11] R. Hiraga, et al., Music perception of hearing-impaired persons with focus on one test subject, IEEE SMC 2015.

ⓒ2016 Information Processing Society of Japan