Vol.2018-SLP-124 No.6
2018/10/10

フィラー検出機能を持つ自動字幕システムの開発

芦川 平[1,a)]　布目 光生[1]　藤村 浩司[1]

概要：近年，音声認識の性能は飛躍的に向上し，一般の日常生活にも利用されつつある．しかし，人が日常で話す話し言葉音声の認識性能は，人が機械などを意識した読上げ音声に比べて，大幅に劣化する．特に，「えーと」，「あのー」といった場つなぎのフィラーと呼ばれる挿入語による性能の劣化は大きい．そこで，我々は，音響イベント検出と音韻識別を同時に行う音響モデルを利用して，フィラーが検出できる音声認識エンジンを開発した．一方，音声認識技術の応用先として，情報保障ツールとしての自動字幕の期待は大きく，我々はこれまで音声認識を利用した自動字幕システムを開発し，運用してきた．今回，我々が開発した自動字幕システムに，フィラー検出機能を導入し，実際に情報処理学会のイベント・講演にて，情報保障ツールとして利用した．本報告では，フィラー検出の手法と，自動字幕システムへの導入事例について報告する．

Development of Realtime Captioning System using the Automatic Filler Detection

Taira Ashikawa[1,a)]　Kosei Fume[1]　Hiroshi Fujimura[1]

1. はじめに

近年，ディープラーニング技術の導入により，音声認識の性能は飛躍的に向上し，音声認識技術を活用した応用や用途が広がりつつある．最近では，音声アシスタント機能を持つスマートスピーカーが，一般家庭にも普及している．また，保守点検時の作業者支援や，コンタクトセンターでのオペレータ支援等の業務支援にも応用されつつある [1], [2].

しかし，人が日常のコミュニケーションのために話す話し言葉音声の認識性能は，人が機械などを意識した読上げ音声に比べて，大幅に劣化する．この原因の一つとして，フィラーの取り扱いがある．フィラーとは，「えーと」「あの」「まー」などの場繋ぎ的に発声される表現である．フィラーが他の単語として誤って認識されてしまうと音声認識精度が低下してしまう．実際に，フィラーが多い発話に対して認識精度が低下するという報告もある [3]．そこで，我々は，フィラーを検出し除去できる音声認識エンジンを開発してきた [4].

また，音声認識技術の応用先の一つに，情報保障ツールとしての自動字幕への期待は大きい．今まで，リアルタイムの自動字幕として，テレビの字幕放送 [5]，大学の講義 [6], [7]，1 対 1 の会議向け [8] 等の応用が検討されてきた．また，会議や講義後の議事録のために，オフラインでの自動字幕生成も検討されてきた [7], [9]．我々も，これまで，音声認識技術を活用した自動字幕システムを開発し，社内の聴覚障碍者の方向けにフィージビリティスタディを行っている [10], [11].

今回，我々は新たにフィラー検出機能を導入し，実際の社内会議や情報処理学会のイベント・講演にて，聴覚障碍者の方のための情報保障ツールとして，自動字幕システムを利用した．本報告では，フィラー検出の手法の概要と，その応用例として，講演での自動字幕にフィラー検出を利用した事例を報告する．

2. フィラー対応音声認識エンジンの開発

我々は，音響イベント検出と音韻識別を同時に行う音響モデルを利用して，フィラーを検出できる音声認識のデ

[1] (株) 東芝 研究開発センター
[a)] taira.ashikawa@toshiba.co.jp

© 2018 Information Processing Society of Japan

コーディングアルゴリズムを開発した [4]．ここでは，その手法の概要を説明する．

2.1 フィラー検出音響モデル

我々は，まず，音響モデルで単語辞書に存在するフィラーを検出するため，音響モデルの学習に用いるラベルとして，フィラーラベルを導入した [12]．フィラーラベルは，各フィラーの末尾に付与する．例えば，フィラーラベルをＦとすると，「(ええっと F)，今日の，(あの F)，課題は」のようにラベルを振る．ここで，ひらがな 1 文字に相当する音節を識別するようにモデルを構築すると，音響モデル学習に使用するラベルは「ええっとＦきょうのあのＦかだいわ」となる．

さらに，音響モデルの学習コストを下げるため，長時間の依存関係をモデル化できる LSTM(Long Short-Term Memory) と，時間軸上でのラベル位置を推定しながら学習を進める CTC(Connectionist Temporal Classification) [13] を導入し，通常の音声認識をしながら，フィラーラベルを検出する手法を確立した．

2.2 フィラー検出音声認識デコーダ

しかし，音響モデル側でこれらのラベルを検出できたとしても，通常の単語辞書を用いてデコードする場合，単語辞書側にラベルＦを付随した全てのフィラー語を用意しておく必要がある．例えば，「ええっと F」，「あの F」などを全て単語辞書として網羅しなければならない．また，音響的にはフィラーを検出しても前後の単語の関係によっては言語モデルのスコアにより他の単語に変換されてしまう．

そこで，我々は，単語辞書に特別なフィラー語を追加することなく，さらに信頼度を付与したフィラーを検出することができるアルゴリズムを開発した．

アルゴリズムの概要としては，単語辞書に存在する単語を構成する音節ネットワークの経路探索中に，音響モデルが出力するフィラーラベルを受け付けた場合にその単語をフィラーとみなす，といった挙動をする．具体的には図 1 のような，重み付き有限状態トランスデューサ（ＷＦＳＴ）を構築することで実現する．デコーダは，ＷＦＳＴ上の経路の入力記号をみることで各単語がフィラーであるか否かを判断する．通常の音声認識においては，探索したＷＦＳＴ上の経路上にある出力記号をつなげて認識結果とする．このとき，デコーダはその経路上の入力記号も参照することができる．経路上の入力記号には，図 1 のＷＦＳＴの入力記号が現れるので，単語 w に対応する経路上の入力記号列に<f>が含まれるとき，単語 w はフィラーであるとして認識結果を出力する．このようなＷＦＳＴを，通常認識に使用するＷＦＳＴと合成することにより，単語辞書にフィラーラベルが付与されたフィラー語彙を網羅していなくてもフィラー単語を検出することが可能となる．

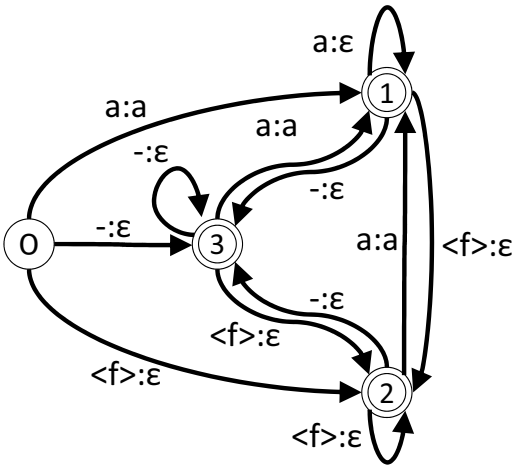

図 1 フィラー処理用ＷＦＳＴの例 ([4] から転載))

2.3 フィラー信頼度

さらに我々は，フィラーの信頼度を導入するために，「えＦえＦっＦとＦ」のようにフィラー単語の音節毎にフィラーラベルを挿入し，学習するようにした．このように学習すると，単語ｗに対応する経路上の入力記号列に複数の<f>が現れる．その語がフィラーであれば，学習時と同じように音節の数と同じ入力記号列<f>が現れるはずである．これを利用し，その単語の音節数とデコード中に現れる入力記号列<f>の数によってフィラー信頼度を算出する．例えば，「ええっと」のように 4 音節のフィラー単語に対して，<f>が 2 つ検出された場合，信頼度は 50%となり，<f>が 4 つ検出された場合，信頼度は 100 %となる．本手法を用いることにより，頻出するフィラー語を部分系列に持つ長い単語を，誤ってフィラー語と判断してしまうことを防ぐことができる．例えば，「グラデュエート」などに含まれる「エート」などに誤って<f>が検出されてしまっても，「グラデュ」部分には<f>が発生しにくいため，信頼度が低くなる．

3. アプリケーションへの応用

我々は，今までリーンスタートアップ形式で自動字幕システムを開発し，運用してきた．そこで，今まで得られた知見を活かし，フィラー検出機能を自動字幕システムへ導入した．本章では，自動字幕システムの概要，情報処理学会講演での利用実績，及び，本システムへのフィラー検出機能の導入について述べる．

3.1 自動字幕システムの開発

3.1.1 背景

前述の通り，音声認識技術を活用した応用先の一つとして，情報保障ツールとしての自動字幕への期待がある．これは社会的背景として，既に 2016 年 4 月からいわゆる「障

© 2018 Information Processing Society of Japan

害者差別解消法」が施行されており，その骨子である合理的配慮が求められていることが大きい．また，聴覚障碍を持つ当事者にとっては，日々の授業や講義，あるいは，日常での業務会議や連絡会など，必ずしも合理的配慮がなされてない場面でも，情報を入手する必要が常に生じている．

我々は，こうした状況を鑑み，日常業務で欠かす事ができない情報伝達の際や業務会議における，困りごとやニーズ把握を目的として，当社従業員の聴覚障碍者の方に対し，アンケート調査を実施し，その結果を踏まえた上で，自動字幕システムを開発してきた [10], [11].

3.1.2 システム構成

情報保障ツールとしてのリアルタイム字幕機能は，日常業務の中で無理なく自然に使える事が重要である．そこで，業務利用の PC の延長で簡単に普段使いができるように，Web アプリケーションによる提供を基本構成とした (図 2).

エンドユーザは，Web ブラウザを用意すれば字幕を閲覧する事ができ，また，PC にマイク音声を入力することで，話し手の音声の取り込みが可能となる．さらに，一般的なサーバ／クライアント構成とすることで，音声認識に必要な音響モデルや言語モデルの更新，また社内特有の専門用語や略称，固有名詞などをサーバ側で一括管理することもできる．

また，音声認識エンジンはサーバ化されており，Web アプリケーションと疎結合で構成されている．これによって，各モジュールの可搬性や独立性が維持され，それぞれの改良を行った場合に，迅速にそれらの更新をシステムに反映できるようになっている．

3.1.3 機能とユーザインタフェース

本システムの主な機能を以下に示す．

(1) リアルタイム字幕機能
　Ｗｅｂブラウザを利用して，PC に接続されたマイクからの音声を録音しながら，リアルタイムに音声認識結果（一時結果と確定結果）が表示される．

(2) 書き起こし機能
　リアルタイム字幕機能で録音した音声を，後から音声認識結果を利用して書き起こす（音声を聞きながら，認識結果を修正する）ことができる (図 4).

(3) ユーザ単語辞書管理機能
　上記のリアルタイム字幕機能，及び，書き起こし機能において，音声認識時に利用するユーザ単語の辞書を作成・管理できる．

字幕の提示については，エンドユーザにとって直感的で混乱が無く，見やすい字幕提示にすることが重要である．そこで，講演型の字幕提示では一般的なハイコントラストで提供されている黒背景白文字のベーシックな画面 (図 3) を踏襲した．また，聴覚障碍者側からの意思表示，または発言者以外からの注釈表示の要求を考慮し，対話型の画面

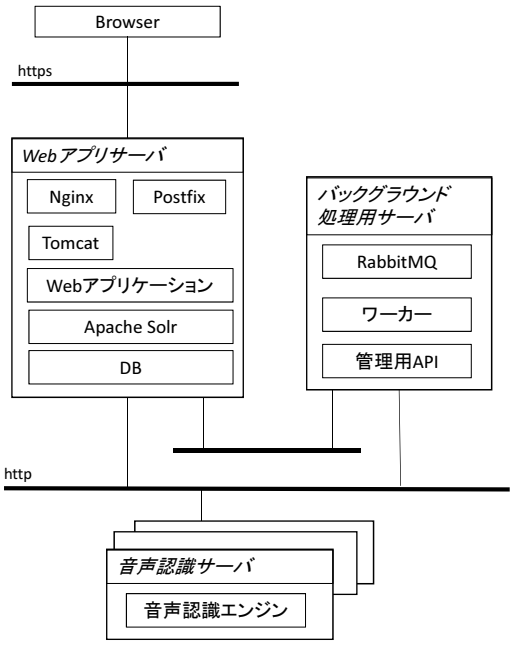

図 2　システムの基本構成

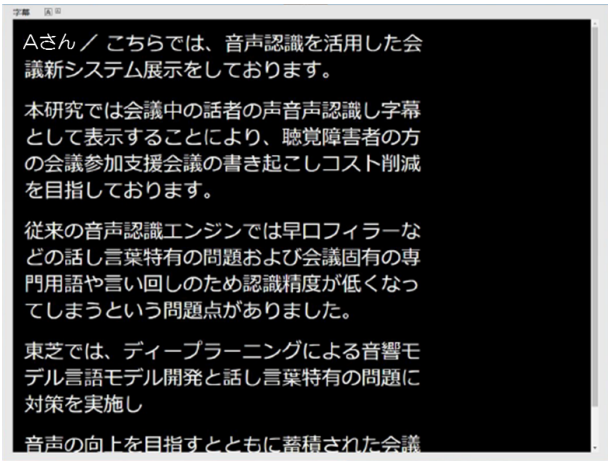

図 3　講演型の字幕表示 ([10] から転載)

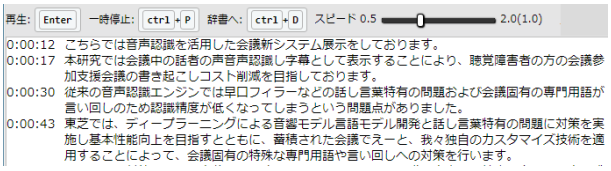

図 4　書き起こし画面

を用意した．これらの表示スタイルは，字幕を閲覧するエンドユーザが，必要に応じてタブを切り替えることでいつでも変更することができる．

3.2 講演でのシステム利用実績

本システムは，当社社内の聴覚障碍者の方が情報保障

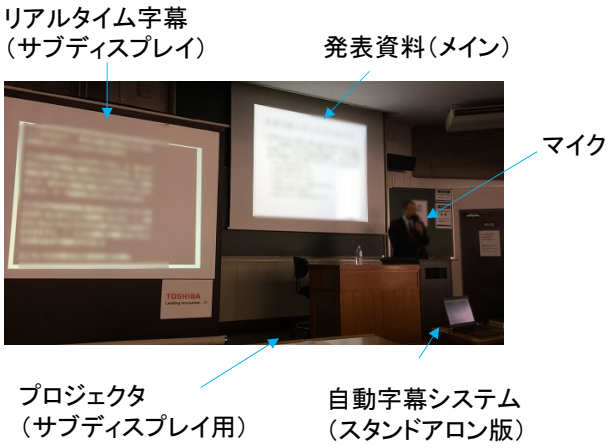

図 5　イベントでの利用例

ツールとして日々利用されているが，情報処理学会主催のイベントにおいても，本システムの講演型の字幕表示を用いて，聴覚障碍者の方が講演の内容を把握する支援を行っている．これまでのイベントでの利用実績を，**表 1** に示す．また，実際に自動字幕システムを利用した様子を**図 5** に示す．右側のメインディスプレイに発表資料が提示され，左側のサブディスプレイには，本システムを用いた音声認識結果が，リアルタイムで字幕モードで表示されている．なお，会場が，地下等の原因で，ネットワークが繋がらない，または不安定な会場も想定されたため，Web アプリケーションと音声認識サーバを仮想環境として用意したスタンドアロン版を利用した．また，多くの会場では，マイク音声が，バックヤードのミキサーに入力され，会場内に流されるため，ミキサーから音声を分岐出力して，本システムに入力した．

表 1　自動字幕システムの IPSJ イベントの利用実績

イベント名	開催日	実施数
ソフトウェアジャパン 2017	2017.2.3	4 セッション
情報処理学会第 79 回全国大会	2017.3.16-18	5 セッション
FIT2017	2017.9.12-14	8 セッション
情報処理学会第 80 回全国大会	2018.3.13-15	11 セッション

なお，イベント前後には以下のことを実施しており，本システムの字幕モード以外の単語登録機能や書き起こし機能を，イベント前後で活用している．

(1) イベント前
音響設備やサブディスプレイの場所など，会場設備を確認する．また，発表者の資料から専門用語等の未知語を辞書登録する．

(2) イベント当日
会場にてシステムを設置し，リアルタイム字幕を実施する．必要であれば，イベントごとに辞書を切り替える，または，その場で未知語の登録を行う．

(3) イベント後
事務局から議事録作成の依頼があれば，音声データと認識結果を回収し，本システムの書き起こし機能を用いて，必要に応じて認識結果を修正し，書き起こした発言記録を議事録としてまとめる．

表 2　講演時のフィラーの数

セッション数	全発話数	フィラー発話数	フィラー出現回数	フィラー種類
20	12,489	9,102	38,172	200

3.3　フィラー検出機能の導入

イベントでの書き起こしを利用して，講演発表でのフィラーの出現回数とフィラーの種類を調査した．その結果を**表 2** に示す．なお，フィラー発話数は，フィラーが含まれた発話数であり，フィラー種類は，書き起こしたテキストがユニークなフィラーの数である．書き起こしの部分には，質疑応答の部分も含まれるが，フィラーの出現する発話の割合は，72.3%と，実際の講演においても比較的出現頻度が高いことが確認でき，講演時でのリアルタイム字幕においても，フィラー対応が重要であることがわかる．

一般的なフィラー対応として，例えば，「えー」，「えーと」等の想定される単語を事前に登録しておき，認識結果からそれらの単語を削除した上で，字幕として表示するということが考えられる．

ただし，イベントによっては，音響状況が悪かったり，発表者も様々なため，1 つのセッション内に限っても，認識精度のばらつきは大きい．そういった状況下で，認識結果から指定の語を単純に完全に削除してしまうと，誤認識があった場合に，内容把握に重要な箇所が抜けてしまい，読み手に正しく意味が通じない可能性がある．また，「あの」，「その」等は指示語である可能性もあり，単純に削除してしまうと，読み手に意図が通じない可能性もある．

これらの状況を踏まえて，自動字幕システムに，前章で紹介したフィラー検出機能を導入し，以下のフィラー対応を行った．

リアルタイム字幕では，字幕表示の即時性を上げるため，まず，発話区間が確定する前は，認識結果の一時結果を逐次的に表示し，発話区間が確定した後に，確定した認識結果 (確定結果) を表示するという形態を取っている．こういった場合，一時結果と確定結果の差分が大きくなってしまうと，読み手は字幕を追随することが難しくなり，好ましくない．さらに，発話区間によっては，フィラーだけの場合もあり，フィラーを全て削除してしまうと発話区間自体の字幕が削除されることになってしまう．

そこで，一時結果表示時には，フィラーを含めて認識結果を加工せずに表示し，確定結果の表示時には，フィラーと判定された部分のみをグレーにして表示した．表示例を，図 6 に示す．

フィラーの判定に関しては，前述のフィラーの信頼度を

© 2018 Information Processing Society of Japan

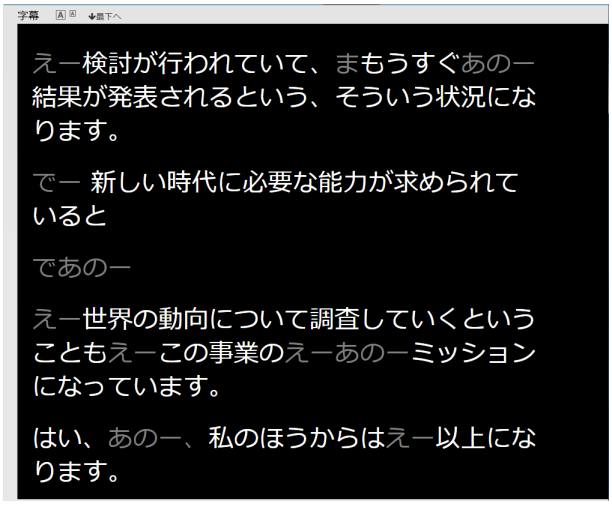

図 6 講演型の字幕表示 (フィラー表示) の例
実際のイベントの発話内容を一部加工

表 3 講演時のフィラーの内容 (上位 20 種類)

発言	出現回数	発言	出現回数
えー	6,591	こう	778
あの	3,864	あ	769
で	3,438	はい	487
ま	3,413	あー	469
えーと	3,158	おー	466
まあ	3,074	えっと	332
あのー	3,022	この	321
ですね	2,257	でー	216
その	1,955	ね	180
え	1,135	ちょっと	180

用いて，フィラーの信頼度が閾値を超えている単語を，フィラーと判定した．ただし，音響情報のみを用いて，フィラー検出を行ってしまうと，イベント会場や話者の音声データと，学習データとが一致しない場合に，検出を誤ってしまう可能性がある．

また，Berke らの報告 [8] によれば，リアルタイムの自動字幕において，音声認識の信頼度を用いて文字装飾（例えば，信頼度が低い場合にグレー表示にする等）を行い，ユーザ評価を行ったところ，多くの利用者に好まれない結果となった．これは，利用者にとって，文字装飾される原因がわかりにくいため，多様に文字装飾されてしまうと利用者は字幕に集中できなくなってしまったことが原因だと考えられる．

そこで，実際の利用場面では，出現頻度が高いフィラーのみを，フィラーとして判定することにした．具体的には，フィラーの信頼度が一定以上あり，かつ，実際の講演中に出現したフィラーの上位 20 種類 (**表 3**) と一致する語（全フィラー出現回数の 94.58%）を，フィラーとして判定した．

4. おわりに

本報告では，話し言葉に対する音声認識性能向上と可読性向上を目的として，音響イベント検出と音韻識別を同時に行う音響モデルを利用したフィラー検出の手法について報告した．また，その導入事例として，自動字幕システムにフィラー検出機能を導入し，実際に情報処理学会でのイベント・講演での情報保障ツールとして活用した事例を報告した．

今後は，実際の利用者からアンケート評価等を行い本機能の導入効果を確認し，継続的なシステム改善を行っていく．

参考文献

[1] 渡辺奈夕子，藤村浩司：端末型音声認識を用いた現場作業支援システムの実用性検討，研究報告ヒューマンコンピュータインタラクション（HCI），Vol. 2018-HCI-179, No. 12, pp. 1–8 (2018).
[2] 長野 徹，壁谷佳典，岡原勇郎，吉田一星，倉田岳人，立花隆輝：音声認識技術を用いたコンタクトセンターオペレータ支援，研究報告音声言語情報処理（SLP），Vol. 2017-SLP-118, No. 9, pp. 1–5 (2017).
[3] 篠崎隆宏，斎藤洋平，堀 智織，古井貞熙：話し言葉音声の認識を目指して，電子情報通信学会技術研究報告. SP, 音声， Vol. 100, No. 523, pp. 7–12（オンライン），入手先 ⟨https://ci.nii.ac.jp/naid/110003297814/⟩ (2000).
[4] FUJIMURA, H., NAGAO, M. and MASUKO, T.: SIMULTANEOUS SPEECH RECOGNITION AND ACOUSTIC EVENT DETECTION USING AN LSTM-CTC ACOUSTIC MODEL AND A WFST DECODER, *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2018).
[5] 今井 亨：リアルタイム字幕放送のための音声認識，NHK 技研 R&D， Vol. 2012/1, No. 131, pp. 4–13（オンライン），入手先 ⟨https://www.nhk.or.jp/strl/publica/rd/rd131/PDF/P04-13.pdf⟩ (2012).
[6] 秋田祐哉，塩野目剛亮，白石優旗：音声自動認識による字幕情報保障トライアル (2)，研究報告アクセシビリティ（AAC），Vol. 2016-AAC-002, No. 6, pp. 1–3 (2016).
[7] Ranchal, R., Taber-Doughty, T., Guo, Y., Bain, K., Martin, H., Robinson, J. P. and Duerstock, B. S.: Using Speech Recognition for Real-Time Captioning and Lecture Transcription in the Classroom, *IEEE Transactions on learning technologies*, Vol. 6, No. 4, pp. 299–311 (2013).
[8] Berke, L., Caulfield, C. and Huenerfauth, M.: Deaf and Hard-of-Hearing Perspectives on Imperfect Automatic Speech Recognition for Captioning One-on-One Meetings, *Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '17, New York, NY, USA, ACM, pp. 155–164 (online), DOI: 10.1145/3132525.3132541 (2017).
[9] 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，電子情報通信学会論文誌 D，Vol. J93-D, No. 9, pp. 1736–1744 (2010).
[10] 布目光生，渡辺奈夕子，芦川 平，藤村浩司：アクセシブルな講演を実現する自動字幕提示手法の検討，研究報告アクセシビリティ（AAC），Vol. 2016-AAC-2, No. 4, pp.

© 2018 Information Processing Society of Japan

1–6 (2016).

[11] Fume, K., Ashikawa, T., Watanabe, N. and Fujimura, H.: Implementation of Automatic Captioning System to Enhance the Accessibility of Meetings, *Computers Helping People with Special Needs - 16th International Conference, ICCHP 2018, Linz, Austria, July 11-13, 2018, Proceedings, Part I*, pp. 187–194 (online), DOI: 10.1007/978-3-319-94277-3_31 (2018).

[12] Nasu, Y. and Fujimura, H.: Acoustic event detection and removal using LSTM-CTC for speech recognition(In Japanese), *IEICE technical report*, Vol. 116, No. 208, pp. 121–126 (2016).

[13] Graves, A., Fernández, S., Gomez, F. and Schmidhuber, J.: Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks, *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, New York, NY, USA, ACM, pp. 369–376 (online), DOI: 10.1145/1143844.1143891 (2006).

© 2018 Information Processing Society of Japan