Vol.2016-NL-226 No.5
Vol.2016-SLP-111 No.5
2016/5/16

音声中の検索語検出のための回帰結合ニューラルネットワークを用いた正解音素推定

澤田 直輝[1,a)]　西崎 博光[2,b)]

概要：本稿では，Recurrent Neural Network (RNN) を用いた複数の音声認識システムの結果から音素誤りパターンを学習した正解音素推定器と，この結果を利用した音声中の検索語検出について述べる．正解音素推定器は回帰結合ニューラルネットワークの一種である Long Short-Term Memory（LSTM) を用い，LSTM で複数の音声認識システムの音素出力系列パターンを学習させることで，正しい音素を予測する．この提案手法で正解音素を推定した結果，音素認識率が大語彙音声認識システムの N-best と比較して改善した．さらに，提案手法を STD タスクに適用した結果，我々が以前に提案した条件付き確率場を用いた triphone 検出器に基づく STD システムの性能を大きく改善することができた．

Phoneme Sequence Estimation using Recurrent Neural Network for Spoken Term Detection

Naoki SAWADA[1,a)]　Hiromitsu NISHIZAKI[2,b)]

1. はじめに

音声ドキュメント検索の 1 つである音声中の検索語検出（Spoken Term Detection：STD）の目的は，検索語（1 個以上の単語からなる言葉）が話されている箇所を音声ドキュメント中から特定することにある．一般的に，STD では音声認識システムとその出力を利用する．そのため，音声認識誤りを起こした語や音声認識辞書に登録していない語（未知語）を正しく検出することができない．このことから，音声認識性能を上げることで STD 性能を改善することが可能である．しかし，完全な音声認識を行うのは現在のところほぼ不可能である．

[1] 山梨大学大学院医工農学総合教育部
Department of Education, Interdisciplinary Graduate School of Medicine, Engineering, and Agricultural Sciences, University of Yamanashi Takeda 4–3–11, Kofu-shi, Yamanashi, 400–8511 Japan
[2] 山梨大学大学院総合研究部工学域
The Graduate School of Interdisciplinary Research, Faculty of Engineering, University of Yamanashi Takeda 4–3–11, Kofu-shi, Yamanashi, 400–8511 Japan
[a)] sawada@alps-lab.org
[b)] hnishi@yamanashi.ac.jp

そこで，音声認識誤りや未知語に対して頑健な検索処理を行うために，サブワードラティスを用いた STD 手法 [2], [3] が提案されている．ラティス表現を用いることにより，単一の認識結果と比較して，豊かな表現で検索を行うことができる．しかし，その一方で，誤りを起こした音素情報が増えることで，これらが検索性能を下げることもある．

そこで，本稿では複数の音声認識システムから出力された音声認識結果に対して，回帰結合ニューラルネットワーク（Recurrent Neural Network: RNN）である Long Term-Short Memory（LSTM[4]）を用いた正解音素推定器を提案する．

LSTM を用いたネットワークモデルは，音声認識システムの言語モデル [5], [6] や音響モデル [7], [8] のような様々なタスクに用いられている．そして，これらのタスクにおいて性能が改善されることが知られている．そこで，本研究では，LSTM を音声認識システムの出力結果から正解音素を推定する枠組みに適用する．また，LSTM を用いた正解音素推定から得られる推定音素系列で STD テストコレクションの STD 性能の改善を図る．

本研究は，RNN を利用して音素列の履歴情報を用いる

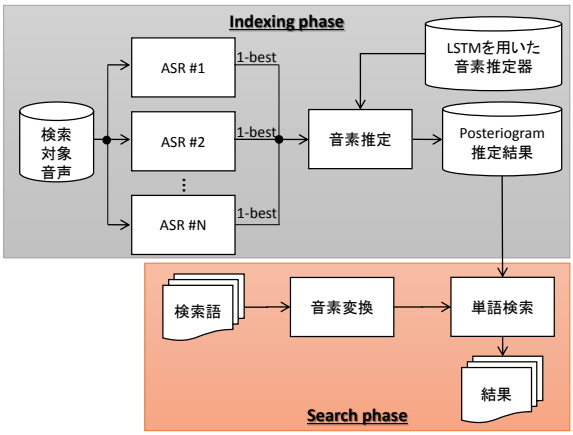

図 1 音素推定器を用いた STD の概要

ことで正解の音素を推定するものである．この枠組みは，RNN 言語モデル [5] と同様のものである．従来の RNN の言語モデルと異なる点は，音素推定として複数の音声認識システムの認識結果から音素誤りを推測する．言い換えると，提案手法である LSTM を用いた音素推定は，音素履歴における音素誤りパターンと認識システムが出力した音素の種類から，正しい音素を推定する．

我々の提案手法は音声認識システムの出力だけを利用しており，音声認識システムと提案手法は独立している．これは提案手法の有用な点である．本稿では，言語モデルと音響モデルが異なる 10 種類の認識システムを使用した．

一方で，Fiscus[9] は複数の音声認識システムの出力結果の多数決を用いて音声認識誤りを削減する手法である ROVER 法を提案した．本稿の提案手法もこれに類似するものであるが，ROVER の枠組みを深層学習を用いてさらに改良しようとするものであり，STD のための検索対象音声の N-best の認識精度の改善を図ることを目的としている．

本稿では，LSTM を用いた正解音素推定器を用いて推定した音素 N-best 表現が，単独の大語彙連続音声認識システムから得た音素 N-best 表現よりも音素誤り率が改善していることを示す．

さらに，この音素推定器を STD システムに適用することで，我々が以前提案した条件付き確率場（Conditional Random Fields: CRF）を用いた triphone 推定器に基づく STD システムと比較して，STD 性能が大きく改善したことについても述べる．STD の評価実験では，2 つのテストコレクションに対して，STD 性能の評価尺度の一つである Mean Average Precision（MAP）において，それぞれ 8.8% と 9.9% の改善を得た．

2. LSTM を用いた正解音素推定と STD

図 1 に正解音素推定器とそこからの STD 処理の概要を示す．本稿では，複数の音声認識システムとして，異なる

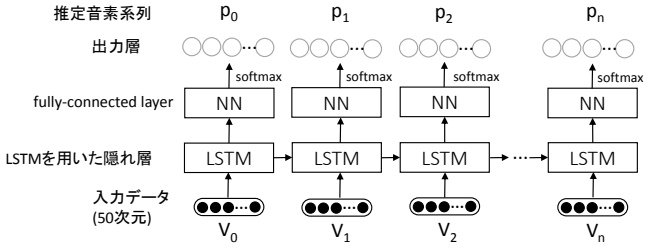

図 2 LSTM を用いた音素推定器

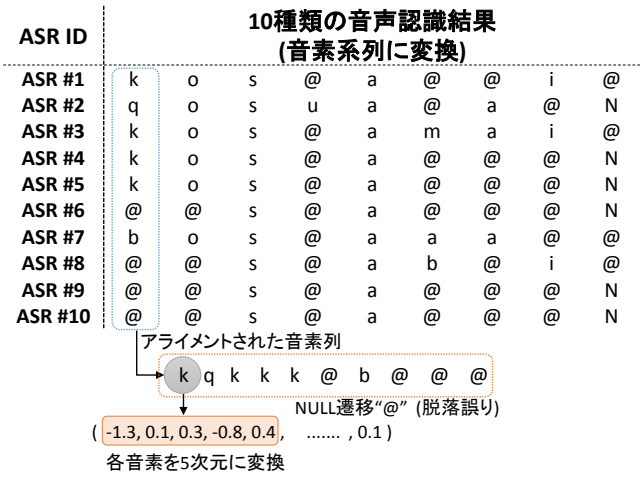

図 3 LSTM を用いた音素推定器のための特徴量変換

音響モデル，言語モデルを組み合わせた 10 種類の音声認識システムを用いる．

2.1 LSTM を用いた正解音素推定器

図 2 に LSTM を用いた正解音素推定器の概要図を示す．V_n は n 番目の入力データ系列であり，p_n は n 番目における softmax 関数により決められた正解推定音素である．提案手法での LSTM を用いた音素推定器は，1 つの LSTM と 4 つの fully-connected layer で構成されている．活性化関数は，Rectified Linear Unit（ReLU [11]）を使用し，最適化手法として Stochastic Gradient Descent（SGD）を使用した．隠れ層のノード数は 512 個であり，出力層は，各音素に対応させた 35 個となっている．

図 3 に，複数の音声認識結果から入力データに変換する例を示す．まず，複数の音声認識結果を音素列に変換する．この認識結果の音素列に対して動的計画法（Dynamic Programing: DP）でアライメントを行う．

本研究では各音素は Bhattacharyya 距離 [12] に基いてベクトル化される．これには，各音素の Gussian Mixture Model (GMM) から Bhattacharyya 距離を計算し，全ての音素同士の距離マトリックスを用意しておく．そして，このマトリックスに対して主成分分析を行い，5 次元に圧縮することでベクトル化を実現した．音素ベクトルの例として，第 1 主成分と第 2 主成分のみを図示したものを図 4 に示す．本稿では，10 種類の音声認識結果から得られた 10 の音素全てを 5 次元に変化しているため，入力データは合

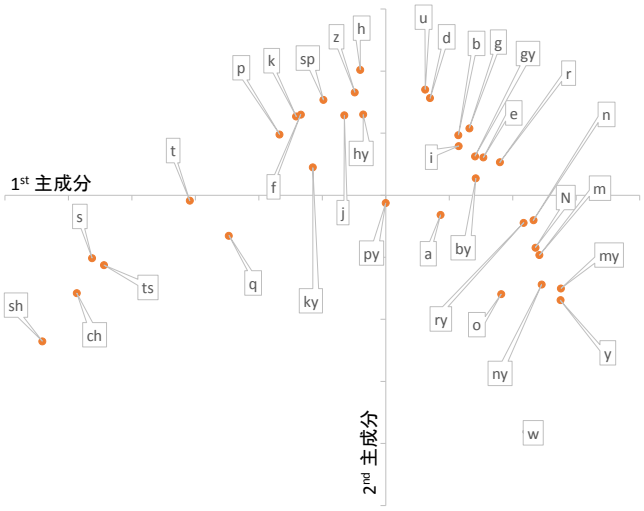

図 4 音素の 2 次元表現の例（第 1 主成分と第 2 主成分）

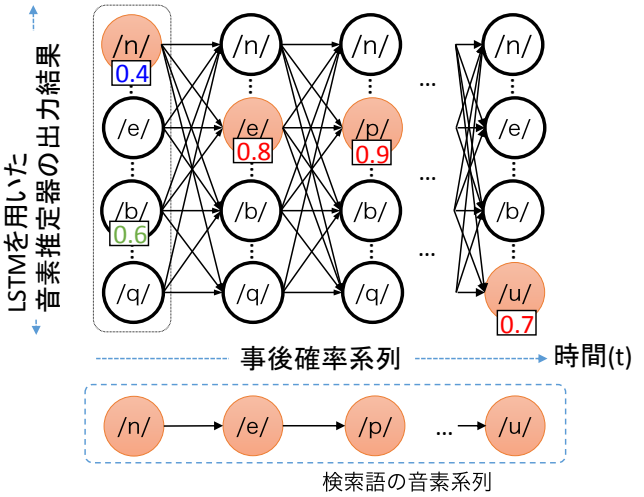

図 5 posterior-gram 系列からの単語検出率の計算例

計で 50 次元となる．なお，NULL 遷移は後方の音素に置き換えている．

2.2 STD エンジン

図 5 に，検索対象音声から 7 音素の単語を検索する例を示す．検索処理には，検索語の音素と音素推定結果の事後確率から DP マッチングを用いて検索する．

図 5 の例では，検索語の検出確率は 0.6 になる．同時に，検索語の検出区間での事後確率の最大確率を計算する．この場合では，最大確率は 0.7 となる．最終的な STD スコアは同じ発話内の検出確率を最大確率で割ることで得られる．この例での最終的な STD スコアは 0.85 となる．

3. CRF を用いた triphone 推定器

我々は以前に，CRF の枠組みを用いて発話中の正解 triphone 検出に基づく STD 手法 [10] を提案した．本稿の提案手法では，RNN の枠組みを用い過去の履歴と認識システムの誤りパターンから 1 つの正解音素を推定しているが，以前に提案した CRF 手法では発話に含まれているであろう正解 triphone を検出し，その検出確率を用いて検索語を検出していた．これら 2 つの手法は，正解音素を検出するという点においては同じ考え方であるが，利用する機械学習の枠組みが異なっている．なお，CRF に基づく手法については，文献 [10] で説明しているため，本稿では詳細を省略する．

本稿では，LSTM を用いた STD 手法と CRF を用いた STD 手法を STD テストコレクションにおける検索性能で比較する．

4. 評価実験

4.1 実験条件

4.1.1 テストコレクション

2 種類の STD テストコレクションに対して適用した．まず 1 つめとして，日本語話しコーパス（CSJ）[14] の未知語テストセット [13] を用いた．このテストコレクションは，CSJ 講演 177 講演（39 時間）を対象としている．発話数は合計で 53,892 発話となる．未知語セットの検索語は 50 個で，177 講演中に 233 箇所の出現箇所のテストコレクションになっている．2 つめのテストセットは，NTCIR-10 SpokenDoc-2 moderate-size task [15] である．これは，音声ドキュメント処理ワークショップ（SDPWS）講演の 104 講演（28.6 時間）を対象としたものである．このテストセットは，100 個の検索語で構成されており，その内訳は既知語が 47 個，未知語が 53 個である．それぞれの発話数は，既知語が 444 発話，未知語が 456 発話である．

4.1.2 音声認識

本研究では 2 種類の音響モデルを用いた．1 つは音節単位でモデル化した GMM-Hidden Markov Model (HMM) で，もう 1 つは triphone ベースの GMM-HMM である．これらのモデルの学習は，認識対象音声がオープンになるように CSJ から学習を行っている [16].

言語モデルは，単語単位，文字単位などの単位を変えた 5 種類の言語モデルを用意した [16].

音響モデル 2 種類，言語モデル 5 種類の計 10 種類の音声認識システムを用意した．これら 10 種類の音声認識システムの中で最も認識性能が高い音声認識システムは音響モデルが triphone ベース，言語モデルは形態素単位の trigram であり，これを認識性能の比較対象として用いる．

4.1.3 評価尺度

本稿では，提案手法を 2 タスクで評価した．1 つは，正解音素推定タスク，もう 1 つは STD タスクである．

正解音素推定タスクでは，音素認識率で評価を行った．

STD タスクでは，recall，precision，recall-precision カー

表 1 CSJ 講演における N-best 音素推定性能 [%].

N-best	1-best	2-best	3-best	4-best	5-best
The best ASR	91.4	92.4	92.9	93.2	93.3
LSTM (10 ASRs)	**92.8**	**96.5**	**97.5**	**98.9**	**98.4**
LSTM (10-best)	91.5	94.5	95.5	96.1	96.5

表 2 SDPWS 講演における N-best 音素推定性能 [%].

N-best	1-best	2-best	3-best	4-best	5-best
The best ASR	83.5	84.6	85.0	85.3	85.5
LSTM (10 ASRs)	**86.9**	**91.7**	**93.5**	**94.8**	**95.8**
LSTM (10-best)	83.7	87.6	89.6	90.9	92.0

ブの最適点である F 値, MAP[17] の 5 種類で評価を行った.

4.2 正解音素推定

表 1 と表 2 に 2 つのテストコレクションにおける N-best の音素認識率を示す．ベースライン（最適な）音声認識システムは，前述したように音響モデル／言語モデルがそれぞれ triphone ／形態素単位 trigram である．LSTM を用いた正解音素推定器では，出力層の posterior-gram の確率が高い順に N 個の音素を取り出した音素列を，N-best とした.

本稿では，音響・言語モデルが異なる複数の音声認識システムの利用の効果を確かめるため，LSTM の学習素性の作成方法として 2 種類の方法を比較した．1 つは，10 種類の音声認識システムの 1-best の出力結果を用いる場合（“LSTM (10 ASRs)” と表記），もう 1 つは，最適な音声認識システムの 10-best 出力結果を用いた場合（“LSTM (10-best)” と表記）である.

表 1，表 2 から，2 つのテストコレクションにおいて “LSTM (10 ASRs)” が “LSTM (10-best)” と比べて，高い音素推定性能を持っていることが示された．10 個の音声認識システムから得られた音素列から学習した音素推定器は，最良の音声認識システムの 10-best から学習した音素推定器よりも，高い推定性能が得られていることが分かる．また，LSTM を用いた正解音素推定器は，CSJ の認識結果から学習しているが，学習コーパスと異なる SDPWS 講演においても高い推定性能が得られていることが分かる.

以上の結果から，正解音素推定タスクにおいて異なる音声認識システムの認識結果から学習することが有効であり，LSTM を用いることで複数の音声認識システムの出力結果から正しい音素推定をすることができることが示された.

4.3 STD 実験

図 6 と図 7 に，CSJ 未知語セットと moderate-size task に対する recall-precision カーブを示す．また，表 3 と表 4 に各テストコレクションの F 値と MAP を示す.

まず，異なる 2 つの音声認識結果から学習した LSTM を用いた音素推定器を比較する．結果から，10 種類の認識結果から学習することで，1 つの認識システムの 10-best から学習した場合と比較して，2 つのテストコレクションの全ての結果において高い性能が得られた．また，CRF 手法においても複数の音声認識システムから学習したほうが高い性能であった．これらの結果から，複数の音声認識システムを用いることが有効であることが分かる．複数の音声認識システムの音声認識結果は，単一の認識システムの N-best よりも，多くの認識誤りパターンを含んでいる．この認識誤りパターンが，機械学習での正解音素推定に有効であると考えられる.

次に，LSTM 手法と CRF 手法を比較する．これら双方ともに，10 種類の音声認識システムの出力結果から得られた素性から学習しており，正しい音素を推定するモデル

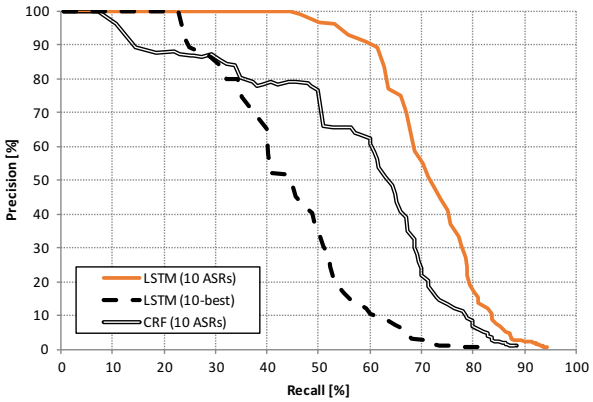

図 6 CSJ 未知語セットにおける recall-precision カーブ

表 3 CSJ 未知語セットにおける F 値と MAP

システム名	最大 F 値 [%]	MAP
LSTM (10 ASRs)	**72.8**	**0.847**
LSTM (10-best)	49.5	0.584
CRF (10 ASRs) [10]	61.1	0.759

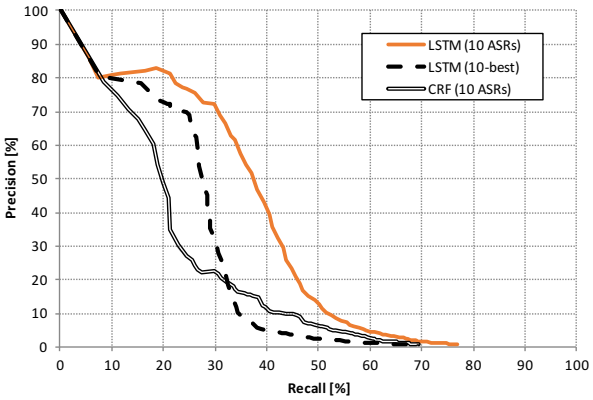

図 7 moderate-size task における recall-precision カーブ

表 4 moderate-size task における F 値と MAP

システム名	最大 F 値 [%]	MAP
LSTM (10 ASRs)	**43.7**	**0.559**
LSTM (10-best)	37.0	0.390
CRF (10 ASRs) [10]	28.6	0.460

である．そして音素推定結果を用いて入力クエリを検出する．2 つのテストコレクションにおける STD 評価結果から，LSTM 手法の方が CRF 手法よりも高い検索性能を得ており，LSTM 手法の方が音素推定性能が高いことが分かる．ニューラルネットワークに基づく手法のほうが，多様な音声認識パターンに適応する能力が高いことが示された結果となった．

5. まとめ

本稿では，STD のための LSTM を用いた正解音素推定器を提案した．提案手法である複数の音声認識システムから学習した LSTM を用いた音素推定器が，音声認識システムの後処理として音素認識の精度を改善することができるかを検証した．

LSTM を用いた音素推定器を，正解音素推定と STD の 2 つのタスクで評価した．音素推定の実験結果では，単一の音声認識システムと比較して，より精度が高い N-best 音素認識列が得られることが分かった．また，2 つのテストコレクションにおいて STD の評価実験から，LSTM を用いた STD システムは，我々の既存手法である CRF を用いた STD システムと比較して大幅な性能の改善が得られた．

本研究では，LSTM を用いた音素推定器のパラメータの最適化ができていない．そのため，今後の課題として，最適化を行う必要がある．また，Kaldi [18] のような他の音声認識システムを用いたときに，本手法が有効であるか検証する必要がある．

参考文献

[1] S. Meng, J. Shao, R. P. Yu, J. Liu, and F. Seide, “Addressing the out-of-vocabulary problem for large-scale chinese spoken term detection,” in *Proceedings of the 9th Annual Conference of the International Speech Communication Association (INTERSPEECH2008)*, pp. 2146–2149, 2008.

[2] D. Can and M. Saraclar, “Lattice indexing for spoken term detection,” *IEEE Trans. on Audio, Speech and Language Processing*, vol. 19, no. 8, pp. 2338–2347, 2011.

[3] G. Chen, S. Khudanpur, D. Povey, J. Trmal, D. Yarowski, and O. Yilmaz, “Quantifying The Value Of Pronunciation Lexicons For Keyword Search In Low Resource Languages,” in *Proceedings of the 2013 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP 2013)*, pp. 8560–8564, 2013.

[4] S. Hochreiter and J. Schmidhuber, “Long Short-Time Memory,” *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[5] T. Mikolov, M. Karafiát, L. Burget, J. Černocký, and S. Khudanpur, “Recurrent neural network based language model,” in *Proceedings of the 11th Annual Conference of the International Speech Communication Association (INTERSPEECH2010)*, pp. 1045–1048, 2010.

[6] M. Sundermeyer, R. Schlüter, and H. Ney, “LSTM neural networks for language modeling,” in *Proceedings of the 13th Annual Conference of the International Speech Communication Association (INTERSPEECH2012)*, pp. 194–197, 2012.

[7] H. Sak, A. Senior, and F. Beaufays, “Long Short-Term Memory Recurrent Neural Network Architectures for Large Scale Acoustic Modeling,” in *Proceedings of the 15th Annual Conference of the International Speech Communication Association (INTERSPEECH2014)*, pp. 338–342, 2014.

[8] A. W. Senior, H. Sak, F. de Chaumont Quitry, T. N. Sainath, and K. Rao, “Acoustic modelling with CD-CTC-SMBR LSTM RNNS,” in *Proceedings of the 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU 2015)*, pp. 604–609, 2015.

[9] J. G. Fiscus, “A Post-processing System to Yield Reduced Word Error Rates: Recognizer Output Voting Error Reduction (ROVER),” in *Proceedings of the 1997 IEEE Workshop on Automatic Speech Recognition adn Understanding (ASRU’97)*, pp. 347–354, 1997.

[10] N. Sawada, S. Natori, and H. Nishizaki, “Re-Ranking of Spoken Term Detections Using CRF-based Triphone Detection Models,” in *Proceedings of the 6th Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA2014)*, pp. 1–4, 2014.

[11] X. Glorot, A. Bordes, and Y. Bengio, “Deep Sparse Rectier Neural Networks,” in *Proceedings of the 14th International Conference on Articial Intelligence and Statistics (AISTATS) 2011*, pp. 315–323, 2011.

[12] B. Mak and E. Barnard, “Phone clustering using the Bhattacharyya distance,” in *Proceedings of the fourth International Conference on Spoken Lanugage Processing (ICSLP’96)*, pp. 2005–2008, 1996.

[13] Y. Itoh, H. Nishizaki, X. Hu, H. Nanjo, T. Akiba, T. Kawahara, S. Nakagawa, T. Matsui, Y. Yamashita, and K. Aikawa, “Constructing japanese test collections for spoken term detection,” in *Proceedings of the 11th Annual Conference of the International Speech Communication Association (INTERSPEECH2010)*, pp. 677–680, 2010.

[14] K. Maekawa, “Corpus of Spontaneous Japanese: Its design and evaluation,” in *Proceedings of the ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition (SSPR2003)*, pp. 7–12, 2003.

[15] T. Akiba, H. Nishizaki, K. Aikawa, X. Hu, Y. Itoh, T. Kawahara, S. Nakagawa, H. Nanjo, and Y. Yamanashita, “Overview of the NTCIR-10 SpokenDoc-2 Task,” in *Proceedings of the 10th NTCIR Conference*, pp.573–587, 2013.

[16] S. Natori, Y. Furuya, H. Nishizaki, and Y. Sekiguchi, “Spoken Term Detection Using Phoneme Transition Network from Multiple Speech Recognizers’ Outputs,” Journal of Information Processing, vol.21, no.2, pp.176–185, 2013.

[17] T. Akiba, H. Nishizaki, K. Aikawa, T. Kawahara, and T. Matsui, “Overview of the IR for Spoken Documents Task in NTCIR-9 workshop,” in *Proceedings of the 9th NTCIR Workshop Meeting*, pp. 223–235, 2011.

[18] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, “The Kaldi Speech Recognition Toolkit,” in *Proceedings of the IEEE 2011 Workshop on Automatic Speech Recognition and Understanding (ASRU 2011)*, 2011.