Vol.2016-SLP-114 No.25
2016/12/21

LSTMを用いた線形予測フィルタの推定に基づく残響下音声認識

木田 祐介[1,a)]　谷口 徹[1,b)]

概要：DNN (Deep Neural Network) によって音声認識の精度がめざましく向上したが，マイクから離れた位置から発せられた遠隔音声の認識は依然として大きな課題である．音圧の減衰による SNR (Signal-to-Noise Ratio) の低下と，床や壁，天井などによる音の反射が引き起こす残響が認識精度を劣化させる主な要因として知られており，これまでに様々な対策が提案されている．本稿では，DNN を用いた新たな残響抑圧技術を提案した．提案法は，残響による歪みを加えた特徴量からクリーンな特徴量へのマッピング関数を直接 DNN に学習させる従来の方式とは異なり，線形予測フィルタの係数を推定する DNN を学習し，DNN から出力されたフィルタを用いて残響の抑圧を行う．残響を精度よくモデル化するため，提案法は長時間の時系列パターンのモデル化に適した LSTM (Long-Short Term Memory) を用いてネットワークを構築する．2014 年に開催された国際コンペである REVERB challenge の単一マイクを用いたタスクにて提案法の評価を行った結果，処理にかかる遅延を 10 ミリ秒に抑えつつ実音声の単語認識誤りを 29.7% から 25.3% に削減できた．

キーワード：残響下音声認識，ディープラーニング，DNN，LSTM，線形予測

LSTM-based linear prediction filter estimation for reverberant speech recognition

Yusuke Kida[1,a)]　Toru Taniguchi[1,b)]

Abstract: Performances of automatic speech recognition (ASR) systems have been drastically improved by DNN (Deep Neural Network). However, distant ASR is still a challenging problem. The difficulty of the distant ASR is caused mainly by two factors; decrease of SNR (Signal-to-Noise Ratio) due to sound attenuation, and reverberation which is created when a sound reflects off the wall, floor and ceiling. In this paper, we propose a novel dereverberation method based on DNN. Different from conventional DNN-based approaches which train mapping functions from corrupted features to clean features directly, the proposed method trains DNN which estimates coefficients of a linear prediction filter, and then dereverberates using the filter outputed from the trained DNN. To model reverberation accurately, the proposed method utilizes LSTM (Long-Short Term Memory) which is appropriate for modeling time-series patterns. Experiments were performed on the REVERB challenge task which was an international competition held in 2014. The proposed method reduced WER (Word Error Rate) from 29.7% to 25.3% with short latency of 10 ms.

Keywords: Reverberant speech recognition, Deep learning, DNN, LSTM, Linear prediction

1. はじめに

DNN (Deep Neural Network) によって音声認識の精度がめざましく向上したが，マイクから離れた位置から発せ

[1] （株）東芝 研究開発センター
Corporate Research & Development Center, 1, Saiwai, Kawasaki, Kanagawa 212–8582, Japan
[a)] yusuke.kida@toshiba.co.jp
[b)] toru.taniguchi@toshiba.co.jp

られた遠隔音声の認識は依然として大きな課題である．遠隔音声の認識精度を劣化させる主な要因は，音圧の減衰による SNR (Signal-to-Noise Ratio) の低下と残響による発話の歪みであることが知られており，これらの問題に対処するために古くから様々な方法が検討されてきた．本稿では，このうち残響に着目する．

残響下における音声認識の頑健性を高めるための方法は，音声認識を行う前にあらかじめ残響成分を取り除くフロントエンド側のアプローチと，残響のかかった音声により音響モデルの学習や適応を行うバックエンド側のアプローチに分けられる．フロントエンド側のアプローチにおける典型的な手法は，観測信号自身から残響特性を推定して観測信号に含まれる残響成分を打ち消す方法である．例えば，室内のインパルス応答モデルを用いて残響信号を推定し，スペクトル・サブトラクションによって観測信号から残響信号を差し引く手法 [1] や，残響によるパワーの時間包絡を推定してそれを打ち消す逆フィルタを適用することでクリーンな信号を回復する手法 [2] などが挙げられる．また，複数のマイクを用いたアレー信号処理により残響信号を抑圧する方法も提案されている [3] [4] ．これらの方法は観測信号自身から残響特性を推定するため，事前に音声入力を行う環境を想定する必要がない．一方，精度よく推定を行うためには多くの信号を与える必要があるため，音声認識をリアルタイムで行う場合には遅延時間の低減が課題となる．

一方，DNN を用いて残響を抑圧する方法が近年さかんに提案されている [5] [6] [7] [8] [9] [10] ．これは，残響による歪みを加えた特徴量からクリーンな特徴量への非線形なマッピング関数を事前に DNN に学習させる方法であり，雑音の抑圧に対しても同様の方法が数多く提案されている [11] [12] ．最近では，別途推定した部屋の残響特性に関する情報を特徴量に加えて DNN に入力する方法 [7] [8] や，クリーンな特徴量の推定と音素の識別の二つのタスクを同時にネットワークに学習させるマルチタスク学習を用いる方法 [8] により精度が高まることが報告されている．また，残響は後続の特徴量に長く影響を及ぼすことから，長い時系列パターンのモデル化に適した LSTM (Long-Short Term Memory) を用いる方法も注目を集めている [9] [10] ．DNN を用いる方法は，事前に大量のデータを用意して学習を行う必要があるが，実際に処理を行う際に生じる遅延時間は少ない．しかし，学習時と評価時の残響環境のミスマッチが大きい場合には効果が得られにくい点が課題である．

本稿では，LSTM を用いた DNN に基づく新たな残響抑圧技術を提案する．残響のかかった特徴量からクリーンな特徴量への理想的なマッピング関数は非常に複雑であると考えられるため，DNN のよいパラエータを求めるためには膨大な学習データが必要だと考えられる．そこで，提案法は線形予測フィルタの係数を DNN に推定させ，推定したフィルタを用いて残響の抑圧を行う．この方法では，フィルタの形式をヒューリスティクスとして付与して DNN に解かせる問題を小さくすることで，限られた学習データでも効率的によいパラメータが求まることを期待している．また，DNN と線形予測を組み合わせることで，少ない遅延時間で残響環境のミスマッチに対する頑健性を得るねらいもある．提案法の評価実験は，2014 年に実施された残響抑圧技術の国際コンペである REVERB challenge [13] のタスクに基づいて実施した．

2. 提案法

2.1 マルチステップ線形予測

提案法の概要を図 1 に示す．提案法は，吉岡らが提案した WPE (Weighted Prediction Error) [14] に用いられているマルチステップ線形予測 [15] により残響を抑圧する．通常，マルチステップ線形予測は複素数領域の周波数スペクトルに対して用いられるが，提案法ではこれを実数領域のフィルタバンク特徴に適用する．すなわち，観測信号の特徴量を $y_n[k]$，残響抑圧後の特徴量を $x_n[k]$ とすると，提案法は以下の式 (1) で残響を抑圧する．

$$x_n[k] = y_n[k] - \sum_{\tau=T_\perp}^{T_\top} g_\tau[k] y_{n-\tau}[k] \qquad (1)$$

ここで，n と k $(1 \le k \le K)$ はフレーム及びフィルタバンク特徴のインデックスである．$G = \{g_\tau\}$ $(T_\perp \le \tau \le T_\top)$ はフィルタ係数であり，$T_\perp$ と $T_\top$ がフィルタリングに用いる特徴量の範囲を定めている．通常の線形予測では $T_\perp$ を 1 に設定するが，マルチステップ線形予測では $T_\perp$ を 3 程度に設定することで，音声の声道特性の歪みを防ぎつつ後部残響成分が抑圧できるとされている．一方，$T_\top$ は想定する残響時間の長さに応じて設定する．

2.2 DNN によるフィルタの推定

WPE は，残響抑圧後の音声が事前に定めた音声モデルに対するゆう度を最大化するよう，観測信号自体からフィルタを推定する．一方，提案法は大量のデータを用いて事前にフィルタの推定法を DNN に学習させる点で WPE と明確に異なる．図 1 に示すように，提案法で用いる DNN は 3 層の LSTM と後段に続く 2 層の MLP (Multi-Layer Perceptron) から構成される．先に述べたように，LSTM は長い時系列パターンのモデル化に適しているため，線形予測フィルタの推定にも有用だと考えられる．DNN への入力は着目するフレームのフィルタバンク特徴である．DNN からはフィルタ G が出力され，これを式 (1) に直接代入することで残響抑圧後の特徴量を得る．ここで，出力ノードの数は $(T_\top - T_\perp + 1) \times K$ である．DNN と式 (1) の計算に用いる特徴量は着目しているフレームより過去の

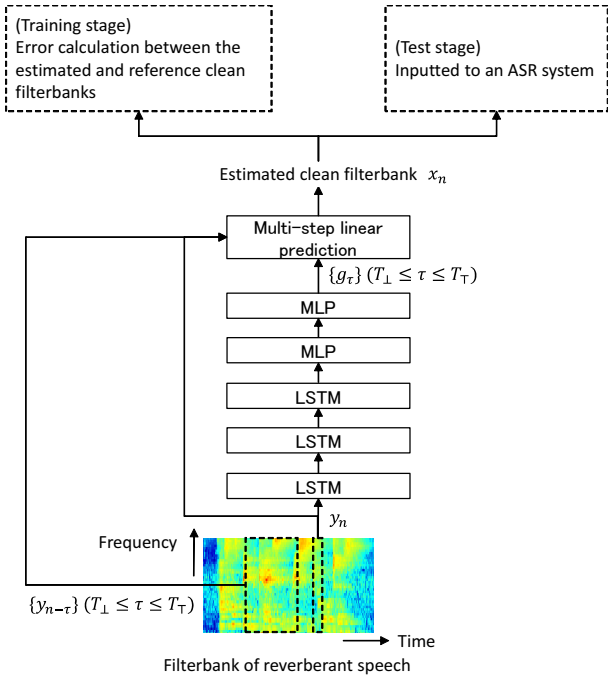

図 1　提案法の概要
Fig. 1　Overview of the proposed method.

ものであるため，提案法は 1 フレームの短い処理遅延で動作する点が特徴である．

DNN の学習は，残響による歪みを加えた特徴量と，対応するクリーンな特徴量のペアを用いて行う．すなわち，残響による歪みを加えた特徴量を DNN に入力し，出力されたフィルタから式 (1) を介して得られた特徴量と，教師信号として与えたクリーンな特徴量の誤差が小さくなるよう DNN のパラメータを学習させる．ここで，誤差関数には二乗誤差を用いる．

3. 評価実験

3.1 タスク

2014 年に開催された国際コンペである REVERB challenge [13] の単一マイクを用いた音声認識タスクにて評価実験を行った．学習用の音声データは，クリーンな環境でヘッドセットマイクを用いて収録された WSJCAM0 コーパス [16] の音声データに，様々な室内環境で測定した 24 種類のインパルス応答を畳み込み，さらに背景雑音を 20 dB の SNR で重畳することで生成されたシミュレーション音声である．測定されたインパルス応答の残響時間（T60）は概ね 0.2 ～ 0.8 秒である．評価データは，WSJCAM0 コーパスの音声データに学習時と異なるインパルス応答と背景雑音を用いて生成されたシミュレーションデータ（SimData）と，実環境の発話が収録された MC-WSJ-AV コーパス [17] の音声データ（RealData）の二種類であり，それぞれ開発用のセット（DevSet）と評価用のセット（EvalSet）に分けられる．SimData のインパルス応答は，3 つの部屋でそれぞれマイクとスピーカの間隔を 50 cm と 200 cm に設定して測定されたものである．一方，RealData の発話は，一つの部屋でマイクと話者の間隔を 100 cm と 250 cm に変えてそれぞれ収録された．評価データの残響条件を表 1 にまとめる．なお，音声のサンプリングレートは 16 kHz である．

3.2 提案法の動作条件

提案法で用いたフィルタバンク特徴は 40 次元（$K = 40$）の対数メルフィルタバンクである．発話単位で平均が 0，分散が 1 になるよう事前に正規化した上で特徴量を DNN に入力した．ここで，特徴抽出時の窓長は 25 ミリ秒であり，フレームシフトは 10 ミリ秒とした．LSTM のセルの数と後段の MLP のノード数は共に 300 とした．DNN から出力されるフィルタの範囲を定める二つのパラメータのうち，$T_\perp$ は 3 とした．$T_\top$ は実験的に求め，20 と設定した．この結果，出力層のノード数は $(20 - 3 + 1) * 40 = 720$ となった．MLP のアクティベーション関数は恒等写像とした．DNN の学習には BPTT (Back Propagation Through Time) アルゴリズムを用いた．学習を現実的な時間で行うため，30 フレームごとに LSTM の保有する計算履歴を消去した．学習時のミニバッチサイズは 128，学習率は 0.001，学習の繰り返し回数は 40 とした．学習データからランダムに選択した 10% を検証用データとして用い，検証用データに対する誤差が上昇する度に学習率を 1/10 に下げた．

3.3 音声認識システム

音声認識には，KALDI ツールキット [18] を用いて学習した DNN-HMM (Deep Neural Network–Hidden Markov Model) 型の音響モデルを用いた．DNN に入力する特徴量は，40 次元の対数メルフィルタバンクに一次及び二次の動的特徴量（Δ，ΔΔ）を加えた 120 次元の特徴量を 11 フレーム分（着目フレームの前後 5 フレーム）連結させた計 1,320 次元とした．特徴量の平均と分散は発話ごとに正規化した．DNN は 5 層の隠れ層をもつ MLP であり，隠れ層のノード数は 2,048 とした．出力は 2,063 の状態共有トライフォンに対する事後確率とした．事前に学習した GMM-HMM (Gaussian Mixture Model–Hidden Markov Model) 型の音響モデルを用いて音素アライメントを取得し，プリトレーニングは行わずクロスエントロピー基準にて DNN の学習を行った．言語モデルは Wall Street Journal コーパスで学習した 5,000 語の 3-gram モデルを用い，KALDI のデコーダを用いて認識を行った．なお，言語モデルの重みは事前に DevSet を用いてモデルごとに調整した．

表 1　評価データの残響条件
Table 1　Reverberation condition of evaluation data

	SimData						RealData	
	Room1		Room2		Room3		Room1	
	Near	Far	Near	Far	Near	Far	Near	Far
Disntance (cm)	50	200	50	200	50	200	100	250
T60 (s)	0.3		0.6		0.7		0.7	

表 2　評価結果
Table 2　Evaluation results

	Latency (ms)	SimData							RealData		
		Room1		Room2		Room3		Ave.	Room1		Ave.
		Near	Far	Near	Far	Near	Far		Near	Far	
No Frontend	N/A	6.9	8.1	8.4	14.1	10.0	17.6	10.8	28.7	30.7	29.7
MLP	60	10.8	11.5	10.3	15.2	11.9	17.6	12.9	26.6	27.1	26.8
Proposed	10	7.7	8.9	8.8	11.8	10.6	14.9	10.5	24.5	26.1	25.3
WPE	(offline)	6.9	7.9	7.7	10.9	8.8	12.6	9.1	24.4	25.5	24.9

3.4　処理例に基づく提案法の検証

図 2 に，提案法の処理結果の例を示す．図の（A）は入力信号のフィルタバンク特徴である．なお，入力信号の残響時間は約 0.7 秒であり，図は平均と分散を正規化した上で表示している．(B) は，別途学習した MLP により（A）の信号からクリーンな特徴量を直接推定させて得られた特徴である．ここで，MLP へ入力する特徴量は，40 次元の対数メルフィルタバンクを 11 フレーム分（着目フレームの前後 5 フレーム）連結させた計 440 次元とした（処理遅延は 60 ミリ秒）．隠れ層は 2,048 のノードをもつ 4 層から成り，アクティベーション関数はシグモイド関数を用いた．学習率は 0.1 とし，プリトレーニングを行わずに 20 回学習を繰り返した．(C) は提案法の処理結果であり，(D) は雑音と残響を付加する前のクリーンな信号から得た特徴である．

はじめに，MLP の結果である（B）に着目すると，入力信号に含まれる雑音と残響が大きく抑圧され，クリーン信号に近い特徴系列が得られたことがわかる．ただし，クリーン信号に比べて特徴系列全体にぼやけが見られ，細部の情報が失われたようにも感じられる．一方，提案法の結果である（C）に着目すると，こちらも残響成分が抑圧される効果が見られた（例えば，30 フレーム・30 番目付近のフィルタバンクや，120 フレーム・35 番目付近のフィルタバンク）が，特徴系列自体はクリーン信号よりも入力信号のものに近く，雑音を抑圧する効果もほとんど見られなかった．(B) と (C) の間に見られたこのような違いは，クリーンな特徴への直接のマッピング関数を推定した MLP と，音声成分の欠損を防ぎつつ残響のみを抑圧するマルチステップ線形予測の性質から生じたものと考えられる．以上より，提案法が意図通り動作していることが確認された．

3.5　評価結果

評価結果を表 2 に示す．ここでは，提案法（Proposed）を含む 3 種類の残響抑圧手法を比較した．比較手法の一つは前節で述べた MLP である．もう一つは 2.1 で述べた WPE であり，公開されている Matlab のツール [19] を用いて評価した．動作時のパラメータは REVERB challenge の単一マイクのタスクで使用されたもの [19] と同様とし，$T_{\perp} = 3$, $T_{\top} = 40$ とした．使用したツールでは入力信号全体からバッチ的にフィルタが推定された[*1]．

はじめに，MLP の結果に着目すると，RealData では残響抑圧処理を行わなかった場合の結果（No Frontend）に対する誤りの改善が見られたが，SimData では誤りが逆に増大した．SimData で効果が得られなかった原因の一つは，図 2 (B) で見られた音声成分のぼやけにあると考えられる．また，SimData は音響モデルの学習データと同じ音声コーパスから生成され，雑音を重畳した際の SNR も同じであったため音響的なミスマッチが元々少なかった．そのため，残響抑圧の導入により増大したミスマッチによる悪影響が残響抑圧の効果よりも大きかったことも精度劣化の一因だと考えられる．次に，Proposed の結果に着目すると，RealData での No Frontend の誤りを 29.7% から 25.3% に削減し，提案法が MLP を上回る効果を得たことがわかる．また，SimData に対しても MLP で見られた精度の劣化を大きく軽減し，残響の大きな条件（Room2, Room3 の Far）では No Frontend に対して誤りを大きく削減できた．最後に，提案法と WPE を比較すると，提案法は 1 フレーム（10 ミリ秒）の短い遅延で動作するにもかかわらず，WPE に近い精度が得られたことがわかる．以上より，提案法の有効性を示すことができた．

[*1] 実際には 30 秒の入力の度にフィルタの推定と残響抑圧処理が実行されるが，評価データの長さはいずれも 30 秒以下であった．

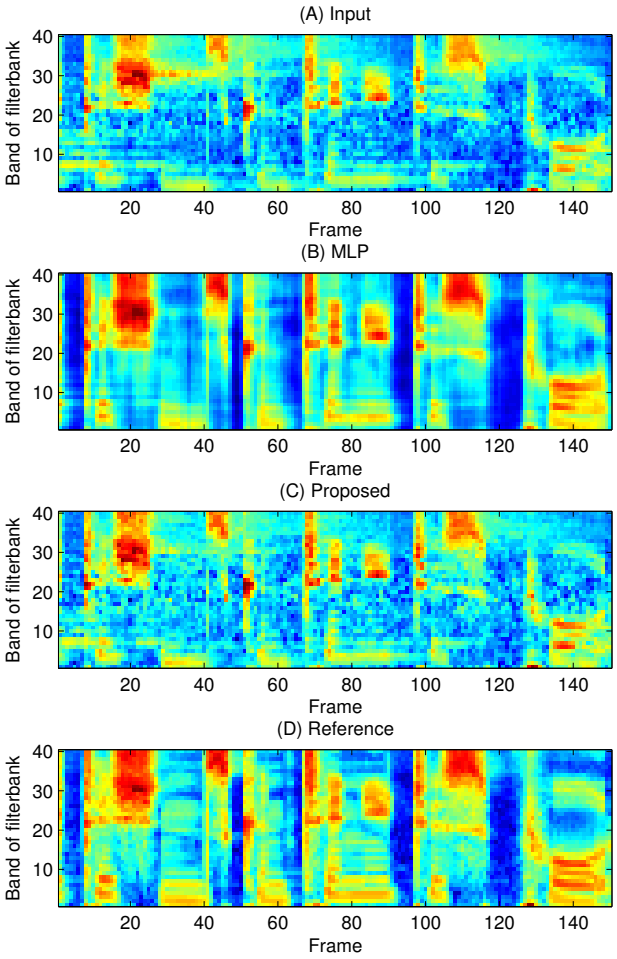

図 2 従来の MLP と提案法の適用により得られたフィルタバンク特徴の比較

Fig. 2 Comparison of processed filterbank features between conventional MLP and proposed method.

4. おわりに

事前に学習した DNN に推定させた線形予測フィルタを用いて残響抑圧を行う新たな手法を提案した．提案法は，限られた学習データで精度よく残響を抑圧すると共に，少ない遅延時間で学習時と評価時の残響環境のミスマッチに対する頑健性を得る効果をねらった．REVERB challenge のタスクにて実験を行った結果，提案法の導入により実音声に対する単語認識誤りを 29.7% から 25.3% に削減し，従来の DNN を用いた方法を上回る効果が得られた．

今後は，音響モデルの DNN を提案法の後段に結合させて一つのネットワークとして学習する joint training [20] の効果を確認する．

参考文献

[1] Lebart, K., Boucher, J. and Denbigh, P.: A new method based on spectral subtraction for speech dereverberation, *Acta Acustica United with Acustic*, Vol. 87, No. 3, pp. 359–366 (2001).
[2] Hirobayashi, S., Nomura, H., Koike, T. and Tohyama, M.: Speech waveform recovery from a reverberant speech signal using inverse filtering of power envelope transfer function, *IEICE Trans. A*, Vol. J81-A, No. 10, pp. 1323–1330 (1998).
[3] Flanagan, J., Berkley, A., Elko, G., West, J. and Soudhi, M.: Autodirective microphone systems, *Acoustica*, Vol. 73, pp. 58–71 (1991).
[4] Miyoshi, M. and Kaneda, Y.: Inverse filtering of room acoustics, *IEEE Trans. Speech Audio Process.*, Vol. 36, No. 2, pp. 145–152 (1988).
[5] Ishii, T., Komiyama, H., Shinozaki, T., Horiuchi, Y. and Kuroiwa, S.: Reverberant speech recognition based on denoising autoencoder, *Proc. INTERSPEECH* (2013).
[6] Mimura, M., Sakai, S. and Kawahara, T.: Exploiting deep neural networks and deep autoencoders in reverberant speech recognition, *Proc. HSCMA* (2014).
[7] Ueda, Y., Wang, L., Kai, A. and Ren, B.: Environment-dependent denoising autoencoder for distant-talking speech recognition, *EURASIP Journal on Advances in Signal Processing*, No. 1, pp. 1–11 (2015).
[8] Giri, R., Seltzer, M., Droppo, J. and Yu, D.: Improving speech recognition in reverberation using a room-aware deep neural network and multi-task learning, *Proc. ICASSP* (2015).
[9] Weninger, F., Watanabe, S., Rour, J., Hershey, J., Tachioka, Y., Geiger, J., Schuller, B. and Rigoll, G.: The MERL/MELCO/TUM system for the REVERB challenge using deep recurrent neural network feature enhancement, *Proc. REVERB Challenge Workshop* (2014).
[10] Mimura, M., Sakai, S. and Kawahara, T.: Speech dereverberation using long short-term memory, *Proc. INTERSPEECH* (2015).
[11] Lu, X., Matsuda, S., Hori, C. and Kashioka, H.: Speech restoration based on deep learning autoencoder with layer-wised pretraining, *Proc. INTERSPEECH* (2012).
[12] Seltzer, M., Yu, D. and Wang, Y.: An investigation of deep neural networks for noise robust speech recognition, *Proc. ICASSP* (2013).
[13] Kinoshita, K., Delcroix, M., Gannot, S., Habets, E., Haeb-Umbach, R., Kellermann, W., Leutnant, V., Maas, R., Nakatani, T., Raj, B., Sehr, A. and Yoshioka, T.: A summary of the REVERB challenge: state-of-the-art and remaining challenges in reverberant speech processing research, *EURASIP Journal on Advances in Signal Processing*, No. 1, pp. 1–19 (2016).
[14] Yoshioka, T., Sehr, A., Delcroix, M., Kinoshita, K., Maas, R., Nakatani, T. and Kellermann, W.: Making machines understand us in reverberant rooms: robustness against reverberation for automatic speech recognition, *IEEE Signal Process. Mag.*, Vol. 29, No. 6, pp. 114–126 (2012).
[15] Kinoshita, K., Delcroix, M., Nakatani, T. and Miyoshi, M.: Suppression of late reverberation effect on speech signal using long-term multiple-step linear prediction, *IEEE Trans. Audio, Speech and Language processing*, Vol. 17, No. 4, pp. 534–545 (2009).
[16] Robinson, T., Fransen, J., Pye, D., Foote, J. and Renals, S.: WSJCAM0: A British english speech corpus

for large vocabulary continuous speech recognition, *Proc. ICASSP* (1995).
[17] Lincoln, M., McCowan, I., Vepa, I. and Maganthi, H.: The multi-channel wall street journal audio visual corpus (MC-WSJ-AV): specification and initial experiments, *Proc. ASRU* (2005).
[18] Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G. and Vesely, K.: The Kaldi speech recognition toolkit, *Proc. ASRU* (2011).
[19] : http://www.kecl.ntt.co.jp/icl/signal/wpe/.
[20] Gao, T., Du, J., Dai, L. and Lee, C.: Joint training of front-end and back-end deep neural networks for robust speech recognition, *Proc. ICASSP* (2015).