情報処理学会研究報告
IPSJ SIG Technical Report

Vol.2014-SLP-102 No.5
2014/7/24

残響下音声認識のための音声強調・認識技術：REVERBチャレンジにおけるNTT提案システムについて

デルクロア マーク[1,a)]　木下慶介[1]　吉岡拓也[1]　小川厚徳[1]　久保陽太郎[1,†1]　藤本雅清[1]
伊藤信貴[1]　エスピ ミケル[1]　堀貴明[1]　中谷智広[1]　中村篤[1,†2]

概要：遠隔発話音声認識の精度は雑音や残響によって大きく劣化する。NTT では、以前から雑音・残響下での音声強調及び認識の研究を行っている。本発表では、2014 年 5 月に開催された残響下音声認識の評価イベント（REVERB チャレンジ）で我々が提案したシステムを紹介する。提案システムは、線形予測に基づく残響除去、ビームフォーマと音声モデルに基づく雑音抑圧、DNN 音響モデル、RNN 言語モデル、及び音響モデルの教師なし適応で構成され、本チャレンジでトップスコアを達成した。

Speech enhancement and recognition for reverberant speech: overview of the NTT REVERB challenge system

1. Introduction

Recently, automatic speech recognition (ASR) technologies are being deployed more and more in actual products. However, current applications still require the use of close-talking microphones to achieve reasonable speech recognition performance. To further expand the usage of ASR, there is a need to make systems work reliably in hands-free situations. In such scenarios, speech captured at a distant microphone is degraded by noise and reverberation.

The problem of noise robustness has attracted much attention and has been evaluated through several benchmarks [1], [2], [3], [4]. In contrast, robustness to reverberation has remained a challenging problem [5] and no evaluation benchmark was available until recently. The REVERB challenge 2014 [6], [7] was organized to resolve this situation by proposing a common reverberant speech database to evaluate recent progress in the field of reverberant speech enhancement and recognition.

In this paper, we briefly review the system we proposed for reverberant speech recognition that combines linear prediction based dereverberation, beamforming, model-based noise reduction and deep neural network (DNN) based ASR [8]. We then present summary results for the REVERB challenge task that attest the efficiency of the proposed recognition system. In this paper, we focus on the system and results obtained for the ASR task of the REVERB challenge, but our system also performed well on the speech enhancement task [8], [9].

2. Proposed system

Here we briefly describe the main parts of the system we developed for the REVERB challenge, details about the system can be found in [8]. Figure 1 shows a schematic diagram of the proposed system.

It consists of the following elements:

- **Dereverberation**: We use the weighted prediction error (WPE) dereverberation algorithm [16]. WPE modifies long-term linear prediction based dereverberation by introducing two main modifications, i.e. the introduction of a delay in the calculation of the linear prediction filter coefficients, and the modeling of speech with a short term Gaussian distribution with time varying variance. WPE can be derived for single and multi-channel cases. In the latter case,

[1] 日本電信電話株式会社 NTT コミュニケーション科学基礎研究所
[†1] 現在，amazon.com
[†2] 現在，名古屋市立大学大学院
[a)] marc.delcroix@lab.ntt.co.jp

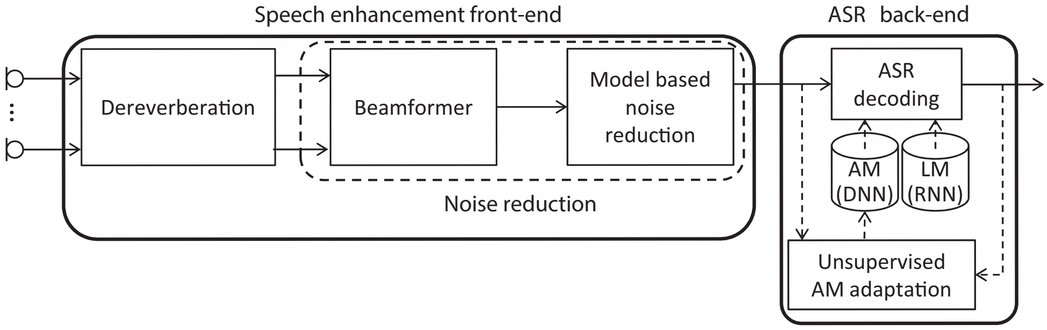
Speech enhancement front-end
ASR back-end
Dereverberation
Beamformer
Model based noise reduction
Noise reduction
ASR decoding
AM (DNN)
LM (RNN)
Unsupervised AM adaptation

図 1 Schematic diagram of the proposed system for recognition of reverberant speech. Note that for the 1ch system we do not perform noise reduction before ASR.

WPE was shown to preserve spatial information in the output signals [17] and can thus be effectively interconnected with multi-channel speech enhancement processing such as beamformer. WPE is well suited for the REVERB challenge task because it has been shown perform well even in presence of noise. Moreover, the algorithm can be derived in the STFT domain, which allows a fast implementation.

- **Noise reduction**: The REVERB challenge data contains a non negligible amount of background noise. We reduce the noise using a conventional minimum variance distortionless response (MVDR) beamformer [18] followed by model-based noise reduction approaches [19], [20].
- **Speech recognition**: Recognition is performed using a DNN-HMM based recognizer, which was trained with multi-condition training data. We also employed recurrent neural network based language model with fast on-the-fly rescoring [21]. Finally, we performed unsupervised environmental adaptation of the acoustic model, by retraining the first layer of the DNN-HMM with a small learning rate, using labels obtained from a first recognition pass [8], [22]. This process is performed in full batch processing, i.e. using a set of test utterances from a same acoustic condition but from different speakers.

3. Experiments

In this section, we introduce the REVERB challenge task and present the experimental results obtained for the 1ch/8ch recognition tasks of the RealData set of the challenge.

3.1 REVERB challenge task

The REVERB challenge consists of speech enhancement and speech recognition tasks. Both tasks rely on the same database. The challenge data consists of the following data sets that are all based on the WSJ/WSJCAM0 text prompts [10], [11].

- **The Development set (Dev)** consists of reverberant speech data recorded in 4 different rooms. The reverberant speech signals for the first 3 rooms were generated through simulations (SimData) using clean speech test data obtained from the WSJCAM0 corpus, and room impulse responses and noise measured in actual rooms. The reverberation time (RT60) varies from 0.25 to 0.7 sec. All utterances include stationary noise at SNR of about 20 dBs. For the fourth room, speech consists of real recordings (RealData) in a meeting room with RT60 of about 0.7 sec obtained from the MC-WSJ corpus [12].
- **The Evaluation set (Eval)** consists of the same acoustic environments than the Dev set, but with different speakers and different speaker positions in the rooms.
- **The Training set (Train)** consists of the clean training data set of WSJCAM0 and several room impulse responses and noise signals measured in real rooms. A script to generate multi-condition training data is also available [13].

For all data sets, 1 microphone (1ch), 2 microphones (2ch) and 8 microphones (8ch) versions are available. All data sets are available through LDC [14], [15] and the REVERB challenge webpage [13]. In addition to the above data sets, the challenge webpage also provides evaluation scripts [13] and description of the challenge regulation.

3.2 Settings

The 1ch speech enhancement front-end consists only of dereverberation (no noise reduction was performed). The 8ch speech enhancement front-end includes both dereverberation and denoising as shown in Fig. 1. Our DNN-

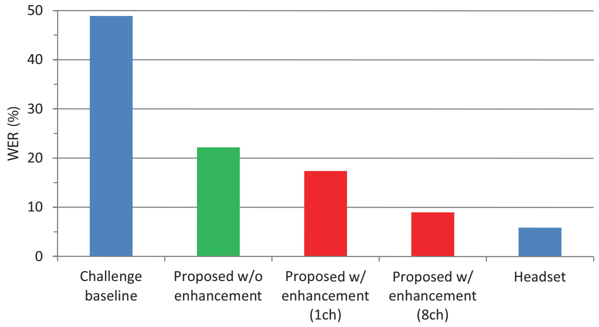

図 **2** Results for the evaluation set (RealData).

HMM recognizer was trained using a conventional procedure [23], i.e. RBM pre-training followed with SGD fine tuning. The input features of the DNN acoustic model consists of 40 log mel filterbank coefficients with delta and delta-delta, augmented by 5 left and right context window. The DNN acoustic model consists of 7 hidden layers, each with 2048 units. The output layer corresponds to 3129 HMM states. We used about 85 hours of multi-condition training data to train our recognition system. Please refer to [8] for further details about the experimental settings.

3.3 Results

Figure 2 plots the word error rate (WER) for the RealData set for the challenge baseline system, our DNN-based recognizer without and with speech enhancement pre-processing for 1ch and 8ch, and the results obtained by recognizing speech recorded with a headset microphone with our DNN-based recognizer.

Figure 2 shows a large performance improvement brought by our DNN-based recognizer over the challenge baseline. We observe significant additional performance improvement on top of this strong baseline by using 1ch and 8ch speech enhancement front-end. With 8ch, the performance becomes close to that obtained with a headset microphone. This was the lowest WER achieved on this task.

Note that detailed results and comparison with the systems from the other participants can be found in [9]. Other techniques that achieved high performance on the task includes i-vector based speaker compensation [24], [25] and system combination [25], [26], [27]. Such approaches could be included into our system to further improve performance.

4. Conclusion

In this paper, we described the system we proposed for the REVERB challenge task. The proposed system demonstrated high recognition performance even for speech recorded in severe reverberant conditions. When using 8 microphones, WER close to that obtained with a headset microphone could be achieved. However, for the single microphone case, there remains much room for improvement. Moreover, future work will include testing the proposed system in more severe conditions, with more noise and spontaneous speech.

参考文献

[1] D. Pearce and H.-G. Hirsh, “The Aurora Experimental Framework for the Performance Evaluation of Speech Recognition Systems under Noisy Conditions,” in Proc. ISCA ITRW ASR2000, pp. 29–32, 2000.

[2] N. Parihar, J. Picone, D. Pearce and H.G. Hirsch, “Performance analysis of the Aurora large vocabulary baseline system,” in Proc. European Signal Processing Conference, pp. 553–556, 2004.

[3] J. Barker, E. Vincent, N. Ma, C. Christensen and P. Green, “The PASCAL CHiME speech separation and recognition challenge,” Computer Speech and Language, vol 27(3), pp. 621–633, 2013.

[4] E. Vincent, J. Barker, S. Watanabe, J. Le Roux, F. Nesta and M. Matassoni, “The second CHiME Speech Separation and Recognition Challenge: Datasets, tasks and baselines,” in Proc. ICASSP, pp. 126–130 , 2013.

[5] T. Yoshioka, A. Sehr, M. Delcroix, K. Kinoshita, R. Maas, T. Nakatani and W. Kellermann, “Making machines understand us in reverberant rooms: robustness against reverberation for automatic speech recognition,” IEEE Signal Process. Mag., vol. 29, no. 6, pp. 114126, 2012.

[6] K. Kinoshita, M. Delcroix, T. Yoshioka, T. Nakatani, E. Habets, R. Haeb-Umbach, V. Leutnant, A. Sehr, W. Kellermann, R. Maas, S. Gannot and B. Raj, “The REVERB Challenge: A Common Evaluation Framework for Dereverberation and Recognition of Reverberant Speech,” in Proc. WASPAA, 2013.

[7] `http://reverb2014.dereverberation.com`

[8] M. Delcroix, T. Yoshioka, A. Ogawa, Y. Kubo, M. Fujimoto, I. Nobutaka, K. Kinoshita, M. Espi, T. Hori, T. Nakatani and A. Nakamura, “Linear prediction-based dereverberation with advanced speech enhancement and recognition technologies for the REVERB challenge,” in Proc. REVERB Workshop, 2014.

[9] `http://reverb2014.dereverberation.com/result_asr.html`

[10] D. B. Paul and J. M. Baker, “The design for the Wall Street Journal-based CSR corpus,” in Proc. HLT, pp. 357-362, 1992.

[11] T. Robinson, J. Fransen, D. Pye, J. Foote and S. Renals, “WSJCAMO: a British English speech corpus for large vocabulary continuous speech recognition,” in Proc. ICASSP-95, vol.1, pp. 81-84, 1995.

[12] M. Lincoln, I. McCowan, J. Vepa and H.K. Maganti, “The multi-channel Wall Street Journal audio visual corpus (MC-WSJ-AV): specification and initial experiments,” in Proc. ASRU-05, pp. 357-362, 2005.

[13] `http://reverb2014.dereverberation.com/download.`

html

[14] M. Lincoln, E. Zwyssig and I. McCowan, “Multi-Channel WSJ Audio LDC2014S03,” Web Download `http://catalog.ldc.upenn.edu/LDC2014S03`, Philadelphia: Linguistic Data Consortium, 2014.

[15] Robinson, Tony, et al. “WSJCAM0 Cambridge Read News LDC95S24,” Web Download `http://catalog.ldc.upenn.edu/LDC95S24`, Philadelphia: Linguistic Data Consortium, 1995.

[16] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi and B.-H. Juang, “Blind speech dereverberation with multichannel linear prediction based on short time Fourier transform representation,” in Proc. of ICASSP’08, 2008, pp. 8588.

[17] T. Yoshioka and T. Nakatani, “Generalization of multichannel linear prediction methods for blind MIMO impulse response shortening,” IEEE Trans. Audio, Speech, Language Process., vol. 20, no. 10, pp. 27072720, 2012.

[18] M. Souden, J. Benesty and S. Affes, “On optimal frequency-domain multichannel linear filtering for noise reduction,” IEEE Trans. Audio, Speech, Language Process., vol. 18, no. 2, pp. 260276, 2010.

[19] M. Fujimoto, S. Watanabe and T. Nakatani, “Noise suppression with unsupervised joint speaker adaptation and noise mixture model estimation,” in Proc. of ICASSP ’12, 2012, pp. 47134716.

[20] T. Nakatani, T. Yoshioka, S. Araki, M. Delcroix and M. Fujimoto, “Dominance based integration of spatial and spectral features for speech enhancement,” IEEE Trans. Audio, Speech, Language Process., vol. 21, no. 12, pp. 25162531, 2013.

[21] T. Hori, Y. Kubo and A. Nakamura, “Real-time one-pass decoding with recurrent neural network language model for speech recognition,” in Proc. of ICASSP ’14, 2014.

[22] H. Liao, “Speaker adaptation of context dependent deep neural networks,” in Proc. of ICASSP’13, pp. 79477951, 2013.

[23] A. Mohamed, G.E. Dahl and G. Hinton, “Acoustic modeling using deep belief networks,” IEEE Trans. Audio, Speech, Language Process., vol. 20, no. 1, pp. 1422, 2012.

[24] X. Feng, K. Kumatani and J. McDonough “The CMU-MIT REVERB Challenge 2014 System: Description and Results,” in Proc. REVERB Workshop, 2014.

[25] Md. J. Alam, V. Gupta, P. Kenny and P. Dumouchel “Use Of Multiple Front-Ends And I-Vector-Based Speaker Adaptation For Robust Speech Recognition,” in Proc. REVERB Workshop, 2014.

[26] Y. Tachioka, T. Narita, F. J. Weninger and S. Watanabe “Dual system combination approach for various reverberant environments with dereverberation techniques,” in Proc. REVERB Workshop, 2014.

[27] F. J. Weninger, S. Watanabe, J. Le Roux, J. Hershey, Y. Tachioka, J. T. Geiger, B. W. Schuller and G. Rigoll “The MERL/MELCO/TUM system for the REVERB Challenge using Deep Recurrent Neural Network Feature Enhancement,” in Proc. REVERB Workshop, 2014.