JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Speech Separation using Python

First Author Name - Shubham Agale, Sinhgad Academy of Engineering, Pune, Maharashtra, India
Second Author Name - Vishal Mane, Sinhgad Academy of Engineering, Pune, Maharashtra, India
Third Author Name - Adinath Gonjare, Sinhgad Academy of Engineering, Pune, Maharashtra, India
Fourth Author Name - Ashutosh Rode, Sinhgad Academy of Engineering, Pune, Maharashtra, India

Abstract - In this paper, a novel deep neural network (DNN) architecture is proposed to generate the speech features of both the target speaker and interfere for speech separation. DNN is adopted here to directly model the highly nonlinear relationship between speech features of the mixed signals and the two competing speakers. With the modified output speech features for learning the parameters of the DNN, the generalization capacity to unseen interferes is improved for separating the target speech. Meanwhile, without any prior information from the interfere, the interfering speech can also be separated. Experimental results show that the proposed new DNN enhances the separation performance in terms of different objective measures under the semi-supervised mode where the training data of the target speaker is provided while the unseen interfere in the separation stage is predicted by using multiple interfering speakers mixed with the target speakers mixed with the target speaker in the training state.

Index Terms - Speech separation, deep neural network

1. INTRODUCTION

Speech separation aims at separating the voice of each speaker when multiple speakers talk simultaneously. It is important for many applications such as speech communication and automatic speech recognition. In this study, we focus on the separation of two voices from a single mixture, namely single-channel (or co-channel) speech separation. Based on the information used the algorithms can be classified into two categories: unsupervised and supervised modes.

2. SYSTEM OVERVIEW AND OVERALL DEVELOPEMENT FLOW AND ARCHITECTURE

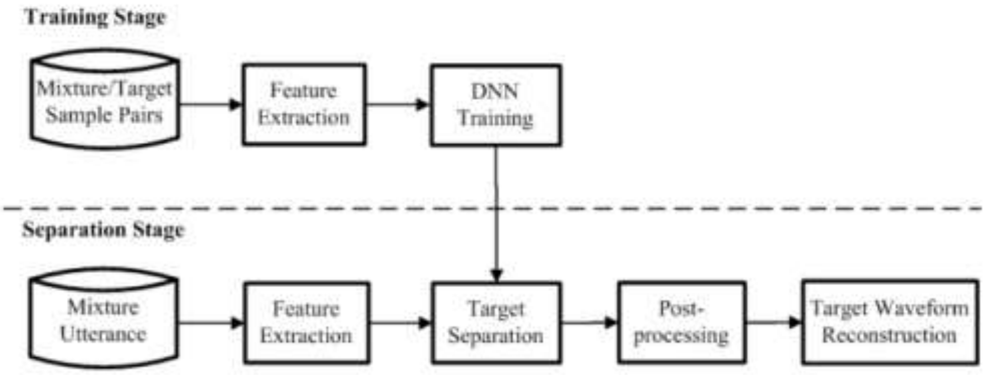

Figure 1: Overall Development flow and Architecture

An overall flowchart of our proposed speech separation system is illustrated in above Fig. In the training stage, the DNN as a regression model is trained by using log-power spectral features from pairs of mixed signal and the sources. in this work there are only two speakers in the mixed signal, namely the target speaker and the interfering speaker. In the separation stage, the log-power spectral features of the mixture utterance are processed by the well-trained DNN model to predict the speech feature of the target speaker. Then the reconstructed spectra could be obtained using the estimated log-power spectra from DNN and the original phase of mixed speech.

3. DNN-BASED SPEECH SEPARATION

3.1 DNN-1 for predicting the target

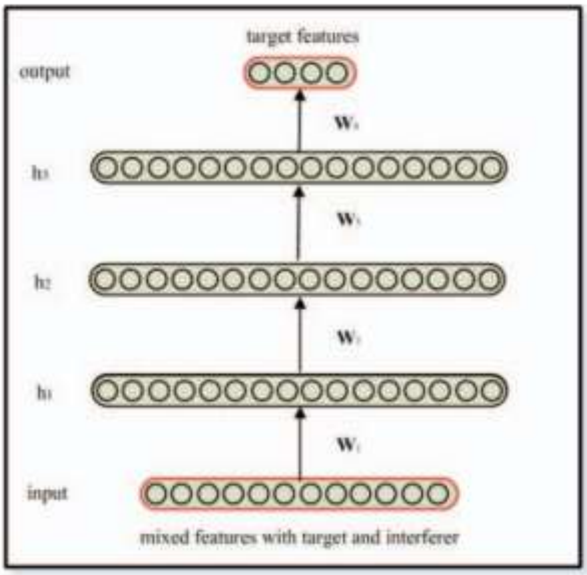

Figure 2: DNN-1 architecture

DNN is adopted as a regression model to predict the log-power spectral features of the target speaker given the input log-power spectral features of mixed speech with acoustic context as shown in above Fig. These spectral features provide perceptually relevant parameters. The acoustic context information along both time axis (with multiple neighboring frames) and frequency axis (with full frequency bins) can be fully utilized by DNN to improve the continuity of estimated clean speech while the conventional GMM-based approach do not model the temporal dynamics of speech. As the training of this regression DNN requires a large amount of time-synchronized stereo-data with target and mixed speech pairs, the mixed speech utterances are synthesized by corrupting the clean speech utterances of the target speaker with interferers at different signal-to-noise (SNR) levels (here we consider interfering speech as noise) based on Eq. (1). Note that the generalization to different SNR levels in the separation stage can be well addressed by the full coverage of SNR levels in the training stage inherently

Mathematical Model for DNN-1

Training of DNN consists of unsupervised pre-training and supervised fine-tuning. The pre-training treats each consecutive pair of layers. For the supervised fine-tuning, we aim at minimizing mean squared error between the DNN output and the reference clean features of the target speaker:

$$E_1 = \frac{1}{N}\sum_{n=1}^{N} \|\hat{\boldsymbol{x}}^{\mathrm{t}}_n(\boldsymbol{x}^{\mathrm{m}}_{n\pm\tau}, \boldsymbol{W}, \boldsymbol{b}) - \boldsymbol{x}^{\mathrm{t}}_n\|_2^2$$

Where, $x^{\wedge t}_n$ and $x^{\wedge t}_n$ are the n^{th} D-dimensional vectors of estimated and reference clean features of the target speaker, respectively. $x^m_{n\pm r}$ is a $D(2\tau + 1)$-dimensional vector of input mixed features with neighboring left and right τ frames as the acoustic context. W and b denote all the weight and bias parameters. The objective function is optimized using back-propagation with a stochastic gradient descent method in mini-batch mode of N sample frames. As this DNN only predicts the target speech features in the output layer, we denote it as DNN-1.

3.2. DNN-2 for predicting both the target and interference

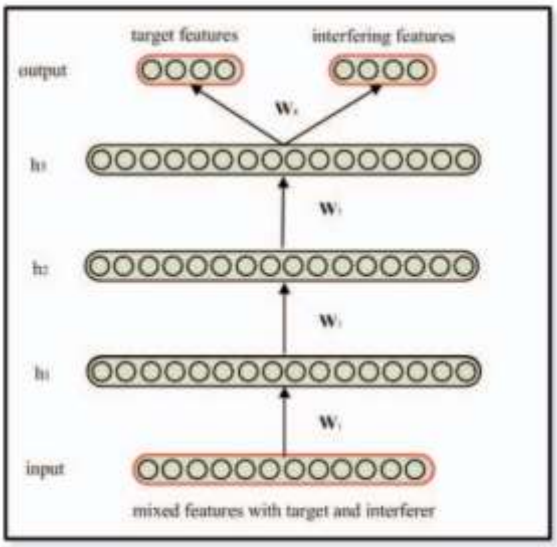

Figure 3: DNN-2 architecture

DNN-2 for predicting both the target and interference - In this work, we design a new DNN architecture for speech separation which is illustrated in above fig. The main difference from previous fig. is that the new DNN can predict both the target and interference in the output layer which is denoted as DNN-2.

Mathematical Model for DNN-2

$$E_2 = \frac{1}{N}\sum_{n=1}^{N}(\|\hat{\boldsymbol{x}}_n^{\mathrm{t}} - \boldsymbol{x}_n^{\mathrm{t}}\|_2^2 + \|\hat{\boldsymbol{x}}_n^{\mathrm{i}} - \boldsymbol{x}_n^{\mathrm{i}}\|_2^2)$$

where xˆtn and x^tn are the nth D-dimensional vectors of estimated and reference clean features of the interference, respectively. The second term of Eq. can be considered as a regularization term for previous euqation of DNN, which leads to better generalization capacity for separating the target speaker. Another benefit from DNN-2 is the inference can also be separated as a by-product for developing new algorithms and other applications.

4. EXPERMENTS

We have two target speakers, namely one male and one female. For each target speaker, 200 utterances were used for training with 30 utterances for testing. The interfering speakers were randomly selected from a large set with thousands of speakers. For training of DNNs, all the utterances of the target speakers in the training set were used while the corresponding mixtures were generated by adding randomly selected interferers to the target speech at SNR levels ranging from -15 dB to 15 dB with an increment of 5 dB The test set for each target speaker consisted of 25 male and 25 female interferers, which are not included in the training stage. Then the mixtures are generated by the target speaker and each interferer at SNRs from -9 dB to 6 dB with an increment of 3 dB for evaluation.

5. CONCLUSION

we have presented a novel architecture of DNN for separating speech of both the target and the interfering speaker. With the additional requirements of predicting the speech feature of the interesting speaker we believe the proposed DNN-2 is more powerful than the baseline DNN-1 in speech separation.

6. REFERENCES

[1] D. L. Wang and G. J. Brown, Computational, Auditory Scene Analysis: Principles, Algorithms and Applications, Wiley-IEEE Press, Hoboken, 2006.

[2] D. L. Wang and G. J. Brown, "Separation of speech from interfering sounds based on oscillatory correlation," IEEE Trans. Neural Netw., Vol. 10, No. 3, pp. 684-697, 1999.

[3] M. Wu, D. L. Wang, and G. J. Brown, "A multi-pitch tracking algorithm for noisy speech," IEEE Trans. Audio Speech Processing, Vol. 11, No. 3, pp. 229-241, 2003.

[4] Y. Shao and D. L. Wang, "Model-based sequential organization in co channel speech," IEEE Trans. Audio, Speech, and Language Processing, Vol. 14, No. 1, pp. 289-298, 2006.

[5] G. Hu and D. L. Wang, "A tandem algorithm for pitch estimation and voiced speech segregation," IEEE Trans. Audio, Speech, and Language Processing, Vol. 18, No. 8, pp. 2067-2079, 2010.

[6] K. Hu and D. L. Wang, "An unsupervised approach to co channel speech separation," IEEE Trans. Audio, Speech, and Language Processing, Vol. 21, No. 1, pp. 120-129, 2013.

[7] J. Ming, R. Srinivasan, D. Crookes, and A. Jafari, "Close data-driven approach to speech separation," IEEE Trans. Audio, Speech, and Language Processing, Vol. 21, No. 7, pp. 1355-1368, 2013.

[8] S. Roweis, "One microphone source separation" Adv. Neural Inf. Process. Syst. 13, 2000, pp. 793-799.

[9] R. Weiss and D. Ellis, "Speech separation using speaker-adapted eigen voice speech models," Computer. Speech Lang., Vol. 24, pp. 16-29, 2010.

[10] S. J. Rennie, J. R. Hershey, and P. A. Olsen, "Single-channel multitalker speech recognition," IEEE Signal Process. Mag., Vol. 27, No. 6, pp. 66-80, 2010.

[11] M. Stark, M. Wohlmayr, and F. Pernkopf, "Source-filter-based single-channel speech separation using pitch information," IEEE Trans. Audio, Speech, and Language Processing, Vol. 19, No. 2, pp. 242-255, 2011.

[12] A. M. Reddy and B. Raj, "Soft mask methods for single-channel speaker separation," IEEE Trans. Audio, Speech, and Language Processing, Vol. 15, No. 6, pp. 1766-1776, 2007.