International Journal of Current Engineering and Technology
E-ISSN 2277 – 4106, P-ISSN 2347 – 5161

Available at http://inpressco.com/category/ijcet

Research Article

Unsupervised Speech Separation using DNN

Mandar Diwakar and Dr. R.A. Satao

Department of Computer Engineering Smt. Kashibai Navale College of Engineering Pune, India

Received 10 Nov 2020, Accepted 10 Dec 2020, Available online 01 Feb 2021, **Special Issue-8 (Feb 2021)**

Abstract

We propose a relapse approach by means of Deep Neural Network (DNN) for solo discourse partition in a solitary channel setting. We depend on a key presumption that two speakers could be very much isolated in the event that they are not very like one another. A divergence measure between two speakers is then proposed to portray the partition capacity between contending speakers. We exhibit that the separation between speakers of various sexes is sufficiently enormous to warrant a potential detachment. We finally propose a DNN design with double yields, one speaking to the female speaker gathering and the other portraying the male speaker gathering. Prepared and tried on the Speech Separation Challenge corpus our trial results show that the proposed DNN approach accomplishes enormous execution increases over the best in class solo strategies without utilizing specific information about the blended objective and meddling speakers and even outflanks the directed speech based strategy.

Keywords: *Deep Neural Network , Speech , Channel Separation*

Introduction

unaided discourse detachment system for blends of two inconspicuous speakers in a single channel setting dependent on deep neural network systems (DNNs). We depend on a key suspicion that two speakers could be well isolated on the off chance that they are not very like one another. A difference measure between two speakers is first proposed to describe the partition capacity between contending speakers. We at that point appear that speakers with the equivalent or various sexes can regularly be isolated if two speaker bunches, with huge enough separations between them, for every sexual orientation gathering could be set up, bringing about four speaker groups. Next, a DNN-based sexual orientation blend identification calculation is proposed to decide if the two speakers in the blend are females, guys or from various sexual orientations. This locator depends on a recently proposed DNN design with four yields, two of them speaking to the female speaker groups and the other two describing the male gatherings. At long last we propose to develop three autonomous discourse division DNN frameworks, one for every one of the female-female, male-male furthermore, female-male blend circumstances. Each DNN gives double yields, one speaking to the objective speaker gathering and the other portraying the meddling speaker group. Prepared and tried on the Speech Separation Challenge corpus, our trial results demonstrate that the proposed DNN-based methodology accomplishes enormous execution increases over the best in class solo procedures without utilizing a particular information about the blended target and meddling speakers being isolated. Single-channel source detachment means to recoup at least one source sign of enthusiasm from a blend of sign. A significant application in sound sign preparing is to acquire clean discourse signals from single-channel chronicles with non-stationary commotions, in request to encourage human-human or human-machine correspondence in negative acoustic conditions. Well known calculations for this errand incorporate model-based methodologies, for example, nonnegative grid factorization (NMF) and, all the more as of late, regulated learning of time-recurrence covers for the uproarious range. Be that as it may, it is eminent that these techniques don't straight forwardly upgrade the real target of source partition, which is an ideal recreation of the ideal signal(s). Beginning investigations have as of late demonstrated the advantage of consolidating such criteria for NMF and profound neural system based discourse partition.The objectives of system are 1]. The comparing speech recognition accuracy of a target speech signal that was extracted from a mixture of two speakers. 2] To determine whether the two speakers in the mixture are females, males or from different genders.

The rest of this paper is organized as follows. Section II summaries the literature survey. Section III introduces the proposed methodology. Design in Section V. Result and discussion in Section IV. Section V focuses on the conclusion.

Literature Survey

In this section, we have discussed different papers referred, based on Separation of speech based using various techniques.

In [1], A deep ensemble method, named multi-context networks, to address monaural speech separation. The first multi-context network averages the outputs of multiple Deep Neural Network whose inputs employ different window lengths. Second was a stack of multiple Deep Neural Network. Each Deep Neural Network in a module of the stack takes the concatenation of original acoustic features and expansion of the soft output of the lower module as its input, and predicts the ratio mask of the target speaker; the Deep Neural Network in the same module employ different contexts. They also compared the two optimization objectives systematically and found that predicting the ideal time frequency mask is more efficient in utilizing clean training speech, while predicting clean speech is less sensitive to Signal-to-Noise Ratio variations.

J. Le Roux, J. R. Hershey *et al.* Propose "profound NonNegative Matrix Factorization ", a novel non-negative profound system engineering which comes about because of unfurling the Non-Negative Matrix Factorization cycles also, loosening its parameters. This design can be discriminatively prepared for ideal division execution. To enhance its non-negative parameters, They show how a new type of back-spread, in light of multiplicative updates, can be utilized to protect non negativity, without the requirement for compelled enhancement. They appear on a difficult discourse detachment task that profound Non-Negative Matrix Factorization improves regarding exactness upon Non-Negative Matrix Factorization and is focused with ordinary sigmoid profound neural systems, while requiring a tenth of the quantity of parameters.

Here author [4] To improve the deep neural networkbased discourse improvement framework, including worldwide fluctuation leveling to ease the over-smoothing issue of the relapse model, and the dropout and commotion mindful preparing procedures to further improve the speculation capacity of Deep Neural Network to concealed clamor conditions. Trial results exhibit that the proposed system can accomplish huge upgrades in both target and emotional measures over the ordinary MMSE based strategy. It is additionally intriguing to see that the proposed Deep Neural Network approach can well stifle profoundly nonstationary clamor, which is difficult to deal with when all is said in done. Moreover, the subsequent Deep Neural Network model, prepared with fake combined information, is moreover viable in managing loud discourse information recorded in genuine world situations without the age of the irritating melodic antiquity ordinarily saw in customary upgrade techniques.

In [5] A multilayer bootstrap sorts out based speaker batching estimation. It uses GMM-UBM or the novel UBSC as the comprehensive establishment model to expel a high-dimensional part from the first MFCC acoustic component, by then uses MBN to diminish the highdimensional component to a low-dimensional space, finally bunching the low dimensional data. We have differentiated it and GMM-UBM-, PCA-, and k-infers packing based techniques. Exploratory outcomes have shown that the proposed strategy beats the referenced strategies. What's more, it is hardhearted toward parameter settings, which energizes its utilitarian use.

L.-R. Dai, and C.-H. Lee et al. [6] A relapse based talk redesign framework using significant neural frameworks (DNNs) with a different layer significant structure. In the DNN learning process, a huge getting ready set ensures an astonishing showing capacity to check the tangled nonlinear mapping from watched riotous talk to needed clean banner. Acoustic setting was found to improve the lucidness of talk to be confined from the establishment commotions adequately without the bothering melodic doodad ordinarily found in customary talk improvement counts. A movement of pilot examinations was driven under multi-condition getting ready with more than 100 hours of repeated talk data, realizing a nice hypothesis limit even in perplexed testing conditions. Right when differentiated and the logarithmic least mean square bumble approach, the proposed DNN-based figuring will by and large achieve immense upgrades similar to various objective quality measures.

In another work, J. Le Roux, J. R. Hershey *et al.* [8] An all around assessment of planning criteria, mastermind structures and feature depictions for backslide based single-channel talk separation with significant neural frameworks (DNNs). We use a traditional discriminative getting ready premise identifying with perfect source generation from time-repeat cloak, and present its application to talk division in a diminished segment space (Mel zone). A close to appraisal of time-repeat cover estimation by DNNs, dreary DNNs and non-negative system factorization on the second Toll Speech Separation and Recognition Challenge shows unsurprising updates by discriminative getting ready, while long transient memory irregular DNNs get the general best results. Furthermore, our outcomes certify the essentialness of tweaking the part depiction for DNN getting ready.

In [7] Significant depiction learning for model-based single-channel source segment (SCSS) and phony exchange speed extension (ABE). The two tasks are not first rate likewise; source-express prior learning is required. What's more to comprehended generative models, for instance, restricted Boltzmann machines and higher solicitation contractive auto encoders two starting late introduced significant models, specifically

generative stochastic frameworks (GSNs) and total thing sorts out (SPNs), are used for learning spectrogram depictions. For SCSS we survey the significant structures on data of the 2 CHiME talk parcel challenge additionally, offer outcomes to a speaker destitute, a speaker free, an organized upheaval condition and an unparalleled racket condition task. GSNs secure the best PESQ and when all is said in done perceptual score on ordinary in all of the four assignments. Therefore, diagram quick GSNs can repeat the missing repeat bunches in ABE best, assessed in repeat space segmental SNR. They beat SPNs embedded in covered Markov models and the other depiction models basically.

Proposed Methodology

A] Architecture of Proposed Scheme

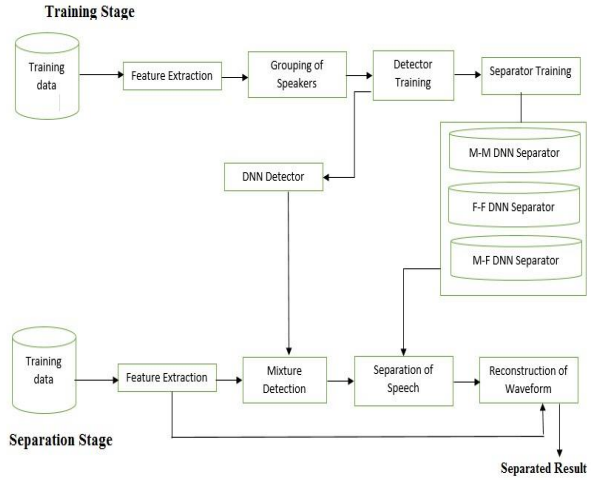

Fig. 3.1 Proposed Scheme

Gender Mixture Detection - To show the importance of the gender mixture detector and the effectiveness of the DNN-based approach, we first introduce a Gaussian mixture model - universal background model (GMMUBM) method widely used in the speaker recognition community as a comparison in experiments. With a UBM for the alternative speaker representation and a form of Bayesian adaptation to derive the speaker models from the UBM, two GMMs representing male speakers and female speakers are trained and then used to determine the gender identities of mixed speech.

Speech Separation - Speech separation or segregation is that the separation of a desired speech signal from a mixture of environmental signals. These can include ambient room noise, other talkers and the other nonstationary noise. The majority of speech separation techniques attempt to reduce noise by replicating the signal processing performed naturally by human auditory sensory system. Speech segregation can be separated into two categories. The first is monaural approaches, which incorporates speech enhancement techniques and computational auditory scene analysis (CASA).

Waveform Reconstruction - An array of voltage values, (y axis) and an array of time values (x axis) and I would like to reconstruct a wave from from these values. The time values are not evenly distributed, they range from 3 ms to 6 ms, the wave form frequency being reconstructed is about 125 hz. Per the shannon-nyquist theorem, the required frequency of samples is 8 ms or less. Using lab view 8.51. I would like to reconstruct the wave form, (currently I can plot it in the xy chart, but I want to do analysis on waveform) In general you can use sin x/ x to reconstruct a waveform mathematically.

B] Algorithm

Step 1 :- Training data set X, corresponding labels set L
Initial bias parameters b and a
Number of layers N, Number of epochs P
Weights between layers W
Momentum M and learning rate

Step 2: - The parameter W,b,a for i = 1 to N do for j = 1 to P do if i=1 then h=X else for i=I to L do

end end

Step 3:- Calculate the state of next layer

$$P(h_q^{i+1} = 1|h^i) = \sigma(b_q + \sum_p h_p^i w_{pq})$$

$$P(h_p^{i+1} = 1|h^{i+1}) = \sigma(a_p + \sum_q h_q^{i+1} w_{pq})$$

Step 4:- Update the weight and biases

$$w^i = \theta W^i + \varepsilon_w(< h_p^i h_q^{i+1} > data - < h_p^i h_q^{i+1} > recon)$$

$$a_p^1 = \theta a_p^i + \varepsilon_a(< h_p^i > data - < h_p^i > recon)$$

$$b_p^1 = \theta b_p^{i+1} + \varepsilon_b(< h_q^{i+1} > data - < h_p^{i+1} > recon)$$

Step 5:- Update the parameters using the gradient of the sparse regularization term

Step 6:- Repeat step 4 and step 5 until convergence

end

end

IV. RESULT AND DISCUSSION

We have taken a set which includes 50M-F, 50F-F, 50M-M mixture parameters to judge the performance of gender mixture detection system. For training of DNNs, all the utterances of the some speakers within the training set were used while the corresponding. The test set for speaker consisted of 25 male and 25 female, which are not included in the training stage. The separation performance was evaluated using three measures, namely output SNR, a short-time objective intelligibility (STOI). The number of epoch for each layer of pre-training was 30 while the learning rate of pre-training was 0.0010. For the fine-tuning, learning rate was set at 0.5 for the first 10 epochs, then decreased by 8% after every epoch. The total number of epoch was 50 and the mini-batch size was set to 125. Input features of DNNs were globally normalized. When the experiments were conducted in the semi-supervised mode, the number of interfering speakers in the training stage for predicting the unseen interferer in the separation stage should be determined. The test set with three gender combinations namely male and male (M+M), male and female (M+F), female and female (F+F), female and male (F+M). The number of interferers was set to 10, 30, and 60 while the corresponding size of training set was from 20 hours to 80 hours. It has been observed that using an adequate size of interferers the trained DNN can well predict an unseen interferer within the separation stage thanks to the powerful modeling capability of DNN. The best performance was achieved.

Combination	Detector	M-M	M-F	F-F
M-F	P-DNN	25	450	4
	CDNN	33	380	9
M-M-D	P-DNN	280	5	2
	CDNN	270	15	1
M-M-S	P-DNN	275	8	1
	CDNN	260	20	2
F-F-D	P-DNN	1	15	275
	CDNN	1	20	265
F-F-S	P-DNN	2	30	250
	CDNN	2	34	240

Table: - Confusion Matrix of testing 450 testing for utterances P-DNN (proposed), CDNN Detector for all SNR.

Input SNR (dB)	-6	0	6
DNN - M	0.72	0.9	0.95
DNN - F	0.78	0.93	0.97

Table: - Separation Performance Comparison of DNN for Male and Female

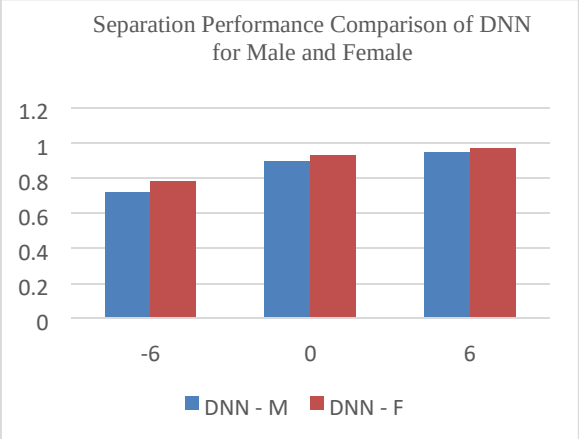

Fig: - Graph

Conclusion

A novel DNN-based gender mixture recognition also, discourse partition system for solo single channel discourse partition inspired by the investigation of the speaker dissimilarities. An extensive arrangement of trials also, examinations, including the significance of DNNbased finder also, the correlations among various blend mixes, are led. The proposed DNN structure could reliably beat the cutting edge CASA approach in wording of various target measures. This investigation is an effective show of applying the profound learning innovation to unaided discourse detachment in a solitary channel setting which is as yet a difficult open issue. Later on, we target refining the proposed system by structuring better speaker gathering calculations and improving the exhibition of both locator and separators. Besides, we intend to further build up our framework on bigger data sets and even some other dialects. The other neural system structures are likewise going to be investigated later on, for example, intermittent neural system for our framework. Another intriguing course is to consolidate the uniqueness measure with cost-capacities for DNN-based finder and separator.

References

[1] X.-L. Zhang and D. Wang, "A deep ensemble learning method for monaural speech separation," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 24, no. 5, pp. 967–977, 2016.

[2] J. Le Roux, J. R. Hershey, and F. Weninger, "Deep NMF for speech separation," in Proc. ICASSP, 2015.
[3] J. Du, Y. Tu, L. Dai, and C. Lee, "A regression approach to single channel speech separation via highresolution deep neural networks," IEEE Trans. Audio, Speech, and Language Processing, vol. 24, no. 8, pp. 1424–1437, 2016.
[4] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," IEEE/ACM Transactions. Audio, Speech, and Language Processing, vol. 23, no. 1, pp. 7–19, Jan 2015.
[5] X.-L. Zhang, "Universal background sparse coding and multilayer bootstrap network for speaker clustering," Proc. Inter speech, pp. 1858–1862, 2016.
[6] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "An experimental study on speech enhancement based on deep neural networks," IEEE Signal Processing Letters, vol. 21, no. 1, pp. 65–68, 2014.
[7] T. May and T. Dau, "Computational speech segregation based on an auditory-inspired modulation analysis," J. Acoust. Soc. Amer., vol. 136, no. 6, pp. 3350–3359, 2014.
[8] F. Weninger, J. Le Roux, J. R. Hershey, and B. Schuller, "Discriminatively trained recurrent neural networks for single channel speech separation," in IEEE Global SIP Symposium on Machine Learning Applications in Speech Processing, 2014.
[9] M. Zohrer and F. Pernkopf, "Representation models in single channel source separation," Proc. ICASSP, 2015, pp. 713-717.
[10] A. Narayanan and D. L. Wang, "Investigation of speech separation as a front-end for noise robust speech recognition," IEEE/ACM Trans. Audio, Speech, Lang. Process., vol. 22, no. 4, pp. 826–835, Apr. 2014.