2019 Master's Thesis

Anchor Word based Deep Attractor Network for Multi-Speaker Separation

A Thesis Submitted to the Department of Computer Science and Communications Engineering, the Graduate School of Fundamental Science and Engineering of Waseda University in Partial Fulfillment of the Requirements for the Degree of Master of Engineering

Submission Date: July 22nd, 2019

Supervisor: Hayato Yamana
Research guidance: Research on Parallel and Distributed Architecture

Department of Computer Science and Communications Engineering,
Graduate School of Fundamental Science and Engineering,
Waseda University
Student ID: 5117FG22-1

QIAN JIAYI

Abstract

With the development of technology, speech input is taking the place of traditional input method for its high convenience and efficiency. However, due to the influence of various realistic environment, a general problem in speech recognition is how to isolate the target speaker's voice from other voice interference. To challenge this problem, in this research, we propose an anchor word based deep attractor network which is able to separate multiple speakers' mixed speech and without any limit to the number of speakers. In our work, we enhance Chen et al.'s original deep attractor network by utilizing anchor word from each speaker as the source of attractor points. Here, the conception of "anchor word" is derived from Wang et al.'s work, which is a short available utterance from target speaker. However, unlike Wang et al.'s work only extracts the target speaker from speech mixture, our task focuses on the separation of multiple speakers' speech and produce clear individual speech of each speaker for the future recognition work. In the training process, we do not adopt the method in Wang et al.'s work which analyzes the relevance between anchor word and speech mixture and build attractor points from both two sources of audio. Instead, in our work, the building of attractor points is only depending on anchor word by directly extracting the feature from anchor word audio. While this enhancement simplifies the training process, it also shows good performance in evaluation results. According to our experiment on CHiME-5 dataset, our anchor word based speech separation system outperforms the original deep attractor network by 2.7% at most.

Contents

1. Introduction

With the development of science and technology, people pursue the convenience of technology products, resulting in plenty of voice-controlled devices to come out such as Amazon Echo and Google Home. In the speech recognition field, an inviable problem is the interferences in a real environment which decreases the accuracy of speech recognition, such as the echo produced by the speaker, sound reflection between walls, various diffuse noise and other human voice. This turns into a classic problem in the speech recognition field called the "cocktail party problem". The "cocktail party problem" refers to when you are in a complex auditory environment like a cocktail party, how to trace and recognize the speech of a specific speaker when multiple speakers talk simultaneously. In this paper, we focus on this problem and attempt to separate multiple speakers' voice from a noisy environment.

In this research, we propose an anchor word based deep attractor network for English speech separation and recognition. Deep attractor network (DANet) refers to Chen's work [1]. In Chen's work, the authors built a deep learning neural network built for single-channel multiple speaker speech separation which creates several reference attractor points of acoustic signals to attract time-frequency bins from each speaker, and their experiment shows DANet could solve the label permutation and dimension unmatched problem in previous work. However, it still exists source number estimation error and compared with other deep-learning based techniques in speech separation, it has no large advantage in efficiency. Depending on this work, in our research, we utilize anchor word, which is a short available utterance from the target speaker such as "hello" from Wang et al.'s work [14]. In most real-world application which uses speech separation and recognition technology such as voice-controlled home devices Amazon echo, Google home and meeting recording applications. It is possible to input an anchor word from each speaker at the beginning of using. Wang et al.'s work [14] is also built on DANet [1], but they focus on tracing only one target speaker's voice from speech mixture, In the training process, they build a deep extractor network (DENet) which utilizes 100 target speakers' utterance as anchor word, incorporating with the feature extraction step from speech mixture in DANet [1] to build new attractor points from both two sources of audio: anchor word and mixture speech.

In our proposed deep attractor network, we directly use anchor word as the source of attractor points to attract time-frequency bins from speech mixture. We extract each

speaker's first time individual speech from dataset, and clip a three second utterance as anchor word after deleting blank pieces. By calculating the centroid of the anchor word speech in place of the speech mixture in the original deep attractor network, our model achieves higher efficiency than DANet [1]. In addition, as the system has known the number of source in advance, the estimation error will not arise in the training process. In the comparison of DENet [14], our proposed model could output all speakers' individual speech without interference. By reducing the amount of anchor word data and analysis process of speech mixture, our model is more flexible and efficient than DENet [14] in implementation.

To implement our idea, we built a deep attractor network as a baseline, and an anchor word based deep attractor network as our proposed method. We utilize ChiME-5 dataset [11] for training and testing process. Finally, signal energy ratio and perceptual evaluation of speech quality are calculated for evaluation.

The outline of this paper is as followed: In chapter 2 and 3, we introduce the background and related work of our research. Our proposed method is explained in chapter 4. Experimental process and results are shown in chapter 5 and 6, and we conclude our research in chapter 7.

2. Background

2.1 Automatic Speech Recognition(ASR)

Automatic speech recognition (ASR) [2] is a technology that uses computer to recognize and translate human's voice into a system. A basic automatic speech recognition system generally consists of two sections: frontend preprocessing and decoder (shown in Figure 2.1). In the frontend processing, the system extracts features of preprocessed rough speech signal, and decoder utilizes the extracted features to translate human voices into text. Section 2.1.1 and 2.1.2 describe the frontend processing and decoder in more detail, respectively.

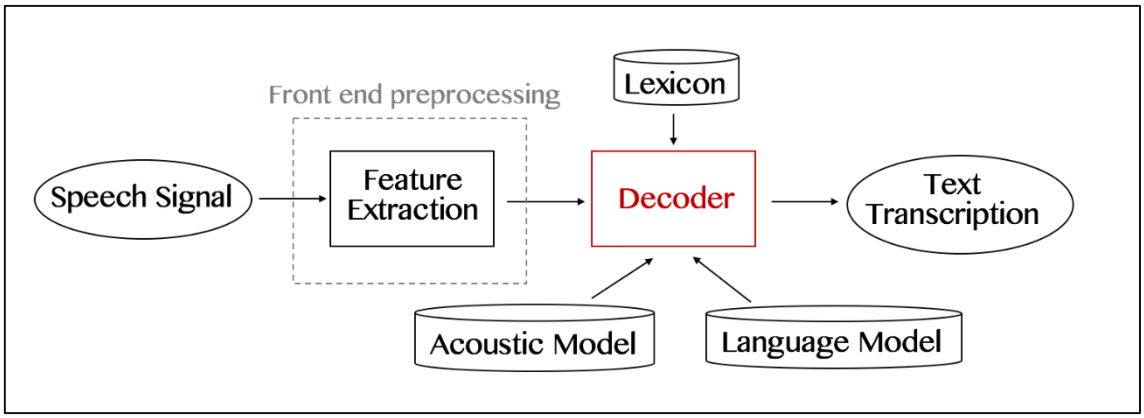

Figure 2.1: A basic ASR algorithm [2]

2.1.1 Frontend Processing

Due to the difference between human's vocal tract and change in prosody when different people speak, features inside the speech could be utilized to distinguish the identity of people. This is called a behavior feature in automatic speech recognition. After speech signals are input into recognition systems, signals will go through the preprocessing steps to improve the audio quality for better extraction of features. Preprocessing steps mainly include echo cancellation and noise suppression. This step is particularly important in far-field speech recognition.

Speech signals after preprocessing are divided into a series of overlapping window frames. Normally, the frame size is between 10 to 30 milliseconds and the overlapping region ranges from 0 to 75% of the frame size. [12] In common method, window frames are converted to a magnitude spectrum by discrete Fourier transform, and the powers of the spectrum are mapped onto the Mel scale, then take the logs of the powers. After

inverse of by another discrete Fourier transform, a Mel-Cepstrum consisting of features of input signals is built.

2.1.2 Decoder

Figure 2.1 shows how the decoder works. Acoustic model shows the relationship between the audio signal and phones, it calculated the probability of the acoustic signal that the given phoneme states. Lexicon could be also called as pronunciation model, it is a dictionary gives the phoneme of each word. Language model calculates which word combinations have the highest probability incorporating with the context.

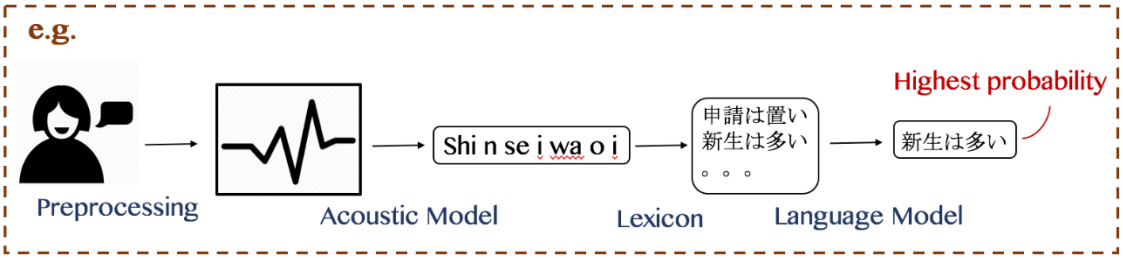

Figure 2.1: An example of decoder diagram

2.2 Deep Attractor Network (DANet)

In this research, an anchor word based deep attractor network is built to solve the speech separation problem of speech with multiple speakers. The deep attractor network is based on Chen's work [1].

Deep attractor network (DANet) is a deep learning neural network which utilizes "attractors" to solve the speech separation problem. This framework refers to Perceptual Magnet Effect [3] in human speech perception, which indicates that brain circuits produce perceptual attractors to attract the closest sound. Under this principle, DANet randomly generates several attractors in the high embedding place which attract time-frequency bins of the source speech. By calculating the difference between the attractors and time-frequency bins, an estimation mask is generated. Through the optimization of embedding space in network, the reconstruction error between estimation mask and clean source alignment will be reduced to least. More detailed models and functions of deep attractor network are described in chapter 4.

2.3 Main problems

There are two main problems in the speech separation field, label permutation problem and dimension unmatched problem.

2.3.1 Label permutation problem

Label permutation problem arises as more deep learning-based techniques are used in the speech separation field. Figure 2.3 shows an example of this problem. In most works using neural network, input signals are tagged with different labels depending on their features. Both sequences of source speeches are reasonable during the training step, but it causes a conflict gradient problem that results in wrong combinations of speech fragment. Certainly, for two labels training like two speakers separation we can fix the position of one source. However, in the situation we have three and more sources to separate, as we assign speaker A a fixed position in both speech mixture (A+B) and (A+C), it leads to a conflict problem in mixture (B+C), because both two sources want to be in the second position.

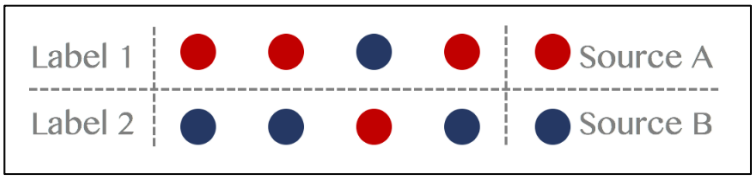

Figure 2.3: Label permutation problem

2.3.2 Dimension unmatched problem

Under the premise that the neural network does not know the number of sources in the mixture before results come out, theoretically there could be infinite combinations of speech that are treated as possible solutions by neural network. However, the layer output nodes are fixed. It will lead to a result that the neural network is not able to separate arbitrary number of sources and will cause the dimension unmatched problem.

2.4 Evaluation metrics

In this research, we evaluate the performance of the baseline and our proposed system from both alternative and objective aspect. From alternative aspect, perceptual evaluation of speech quality (PESQ) [5], an objective method which simulates human listening test

is utilized to determine the perceived quality of separation. And objectively calculation of numerical measures, signal-to-distortion ratio (SDR), signal-to-interference ratio (SIR), signal-to-artifacts ratio (SAR), and signal-to-noise ratio (SNR) are also used to show the results [4][6].

2.4.1 Energy ratios

Objective evaluation methods evaluate the performance of a system by objective data. Most objective numerical performance measures are devised for Blind source separation algorithms. In this research, we calculate four energy ratio parameters [4]: signal-to-distortion ratio (SDR), signal-to-interference ratio (SIR), signal-to-artifacts ratio (SAR), and signal-to-noise ratio (SNR). The definition of these parameters could be described as the ratio of total signal power (Signal + {interference, artifact, noise}) to the unwanted signal power ({interference, artifact, noise}). The unit of energy ratios is decibels(dB).

First, we define $\widehat{s_j}$ as the total signal power of the estimated speech signal source, where j is the number of sources. $\widehat{s_j}$ is defined as:

$$\widehat{s_j} = s_{target} + e_{interf} + e_{noise} + e_{artif} \tag{1}$$

where s_{target} is the clear speech signal source, and e_{interf}, e_{noise} and e_{artif} respectively define the interference, noise and artifact error terms, respectively. Interference refers to the other interference sources, noise is the diffuse noise like air conditioner, and any other errors not included in the previous two are referred to artifact error term, so artifact also be called as statistical artifact. When a time-invariant instantaneous matrix $\boldsymbol{W}$ is utilized to separate the mixture signal and $\boldsymbol{A}$ represents the mix system, $\widehat{s_j}$ is expressed as followed: (2)

$$\widehat{s_j} = (\boldsymbol{WA})_{jj} s_j + \sum_{j' \neq j} (\boldsymbol{WA})_{jj'} s_{j'} + \sum_{i=1}^{m} W_{ji} n_i$$

where $(\boldsymbol{WA})_{jj}$ is the time-invariant gain and $s_{j'}$ refers to the unwanted sources. The definition is based on the orthogonal projector which is generated by a $T \times T$ matrix, where T is the length of dimension vectors. We build three orthogonal projectors:

$$P_{sj} := \prod \{s_j\} \tag{3}$$

$$P_s \coloneqq \prod \{(s_{j\prime})1 \leq j' \leq n\} \tag{4}$$

$$P_{s,n} \coloneqq \prod \{(s_{j\prime})1 \leq j' \leq n, (n_i)1 \leq i \leq m\} \tag{5}$$

And four error terms are defined as:

$$s_{target} = P_{sj}\widehat{s_J} \tag{6}$$

$$e_{interf} = P_s\widehat{s_J} - P_{sj}\widehat{s_J} \tag{7}$$

$$e_{noise} = P_{s,n}\widehat{s_J} - P_s\widehat{s_J} \tag{8}$$

$$e_{artif} = \widehat{s_J} - P_{s,n}\widehat{s_J} \tag{9}$$

Through the equation 1 and error terms equations, we could calculate the following four energy ratio by:

$$SDR = 10log_{10}\frac{\left\|s_{target}\right\|^2}{\left\|e_{interf} + e_{noise} + e_{artif}\right\|^2} \tag{10}$$

$$SIR = 10log_{10}\frac{\left\|s_{target}\right\|^2}{\left\|e_{interf}\right\|^2} \tag{11}$$

$$SNR = 10log_{10}\frac{\left\|s_{target} + e_{interf}\right\|^2}{\left\|e_{noise}\right\|^2} \tag{12}$$

$$SAR = 10log_{10}\frac{\left\|s_{target} + e_{interf} + e_{noise}\right\|^2}{\left\|e_{artif}\right\|^2} \tag{13}$$

2.4.2 Perceptual evaluation of speech quality (PESQ)

Although objective evaluation measures are more straightforward to visualize the performance of algorithm. Subjective listening tests also show its efficiency in determining the quality of output signal source. Compared to the traditional method that using real human as judges to evaluate the quality of audio, PESQ is more impartiality. In this research, we use PESQ [5] to predict the subjective quality with good correction. PESQ is standardized by the ITU-T Recommendation P.862 for automated assessment of the speech quality.

Figure 2.4 shows a general structure of PESQ. In the auditory transform, Fourier transformer is used to estimate the spectrum with preprocessed separated window frames. In the error parameter extraction, the reference and degraded signals are compared frame by frame. By calculating the average regression, we get the prediction value of perceived quality. PESQ score is in the range from -0.5 to 4.5, and higher score shows better speech quality.

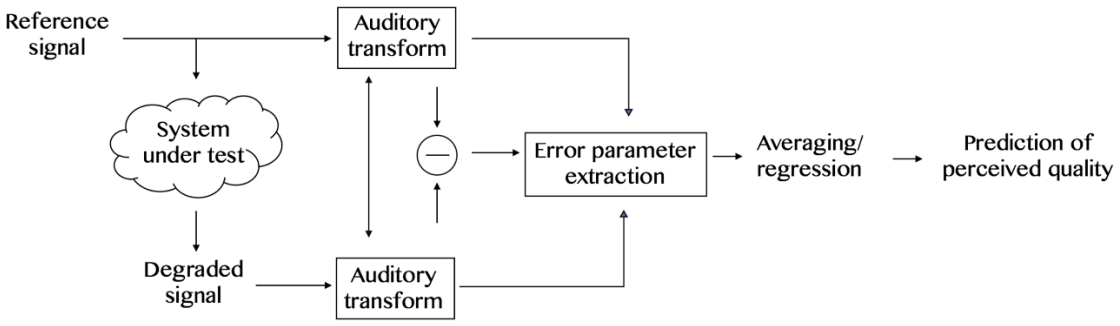

Figure 2.4: General structure of PESQ [5]

3. Related work

Many efforts have been made in the automatic speech separation field to address the problem. Before neural network came into use in this area, most two famous methods were factorial GMM-HMM [6] and computational auditory scene analysis (CASA) [7].

Computational auditory scene analysis (CASA) [7] is a segregation system which is able to separate speech from various sounds also including other speakers' voice. It simulates the way that human listener separating mixture speech.

Factorial GMM-HMM [6] is the most efficient algorithm which showed better performance than human in 2006 monaural speech separation and recognition challenge [8]. Factorial GMM-HMM models the interaction between the target source signals and other interference, and the joint inference is used to calculate the similarity between the target and other signals.

Recently, several deep learning-based techniques are used to address speech separation problem. The major difficulties in deep learning-based works are arbitrary source permutation problem and the problem of unknown number of source. Although Deep Attractor Network [1] is used in our research, there are also proposed methods trying to solve the difficulties.

Weng et al. [9] built a deep neural network containing five key ingredients:

- a multi-type training strategy for mixed speech mixture
- a separate deep neural network to estimate the probability of poster speaker's volume frame by frame
- a weighted finite-state transducer-based two-speaker decoder to jointly estimate and correlate
- a speaker changing punishment estimated from the energy pattern in the speech mixture
- a combined strategy with confidence based system

And in 2006 monaural speech separation and recognition challenge [8], the best approach of this proposed system achieved average word error of 18.8% and outperforms the state-of-the-art IBM superhuman system by 2.8%.

In 2017, Yu et al. [10] proposed a method called permutation invariant training (PIT) to address the problem of speech separation. PIT supervised trains two deep neural networks with senone (phoneme with context) labels from clean speech alignment, the target

sources are trained as a set without sequence, and based on the average energy power of utterance, the frames from the same speaker are trained under different layers.

Kine et al.[13] and Wang et al.[14] both propose to adopt anchor word for speech separation and recognition problem. Kine et al.[13] focus on the recognition difficulties of voice-controlled home applications. Anchor word is utilized as a desired word to distinguish the target speaker's voice from all of the other interference including background noise and other speech. They proposed an anchored mean subtraction (AMS) model which extracts the mean of anchor word which is defined as the representation of desired speech, to normalize the speech features in the utterance. They compared their proposed model with casual mean subtraction (CMS) which conducts the normalization to every frame in the speech. The evaluation result shows the AMS model outperforms CMS which is with up 35% relative WER improvement.

Similar as paper [13], Wang et al. [14] aims to follow only one speaker's voice in the mixture speech. In this paper, they define anchor word as a short clear utterance with one second length of the target speaker, and they prepared 100 anchor words and 500 mixture speech for each speaker. They proposed a deep extractor network based on Chen et al. [1] by creating extractor point in place of the previous attractor point. In this model, anchor word and mixture speech are separately constructed in a primary embedding space, and then combined as an input to a deep neural network to transform to a high dimension embedding space. In the high dimension embedding space, extractor point is created by the combination of anchor word and mixture speech and utilized to attract time-frequency bins from original mixture speech. Their result shows the deep extractor network model outperforms the previous work DANet [1] with 5.2% and 6.6% relative improvements in SDR and PESQ respectively.

4. Proposed method

4.1 Overview

To address the single-channel multi-speaker speech separation problem, in this research, we propose to build an anchor word based deep attractor network shown in Figure 4.1. Here, anchor word refers to a short available utterance from each speech source. In this anchor word based deep attractor network, we hypothesize that each time-frequency bin belongs to only one source speaker. We use anchor word utterance in place of mixture speech to create attractor points in high embedding space, and attractor points draw time-frequency bins from mixture speech to do clustering. This improvement decreases both estimation error and computation complexity.

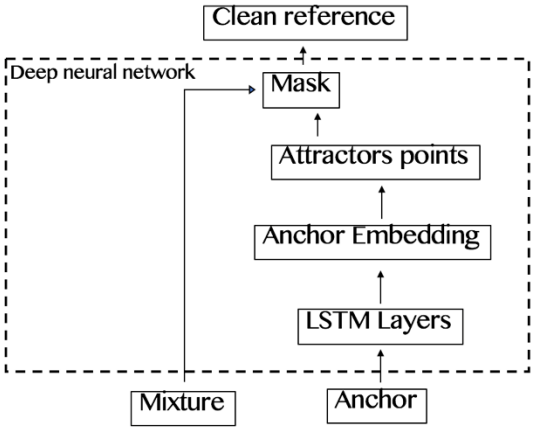

Figure 4.1: Proposed algorithm

4.2 Model

4.2.1 Deep attractor network (DANet)

The Deep attractor network I implemented as baseline in this research is based on Chen's work [1]. Firstly, a mixture speech X is trained by a neural network to map on a k dimension embedding place with the following function:

$$\mathcal{L} = \sum_{ft,s} \left\| C_{ft,s} - X_{ft} \times M_{ft,s} \right\|_2^2 \tag{14}$$

where C is the clean speech signal spectrogram and X is the mixture speech signal spectrogram with frequency f and time t, s refers to the source. M is the estimated mask of the target source that forms to extract features, it is defined as follows:

$$M_{f,t,c} = Sigmoid\left(\sum_{f,t,s} \mathrm{A}_{s,k} \times V_{ft,k}\right) \tag{15}$$

The reconstruction mask M is formed in a k dimension embedding space V, and $\mathrm{A}_{s,k}$ are the attractors for sources s defined with the following equation:

$$\mathrm{A}_{s,k} = \frac{\sum_{f,t} \mathrm{V}_{k,ft} \times Y_{\mathrm{s},ft}}{\sum_{f,t} Y_{\mathrm{s},ft}} \tag{16}$$

where Y represents the energy ranking of the target source. For example, when source s has the highest energy among all sources at frequency f and time t, $Y_{\mathrm{c},ft}$=1.

During the training process, we first generate an embedding place V for the speech mixture, and the attractor points are formed by Equation (16) in a higher dimension embedding place for each source. By calculating the similarity of each time-frequency bin in embedding place V and attractor points A by Equation (15), we estimate the reconstruction mask M through a sigmoid function which weights the mask from 0 to 1. In Equation (16), more similar the time-frequency bin and target attractor point are, the larger the score produced by reconstruction mask for this time-frequency bin. Then, the error between the source signal and clean alignment which is calculated by Equation (14) will be also smaller. Through the optimization of embedding space in network, the reconstruction error between estimation mask and clean source alignment will be reduced to least. Figure 4.2 shows the architecture of this deep attractor network.

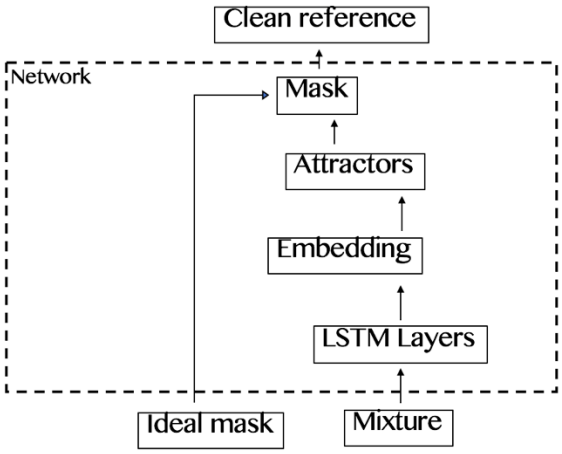

Figure 4.2: Deep attractor network architecture [1]

4.2.2 Deep extractor network (DENet)

Depending on Chen et al.'s work [1], a deep extractor network (DENet) is proposed in Wang et al. [14]. In this DENet, they proposed anchor word, a short available utterance, and utilize it incorporating with speech mixture to build the attractor points. Figure 4.3 shows the architecture of DENet. In this model, anchor word and mixture speech are separately constructed in a primary embedding space through LSTM layers, and then combined as an input to a deep neural network to transform to a high dimension embedding space. In the high dimension embedding space, extractor point is created by extracting the features in time-frequency bins from both anchor word and mixture speech. By using these new extractor points, similar time-frequency bins in speech mixture will be attracted.

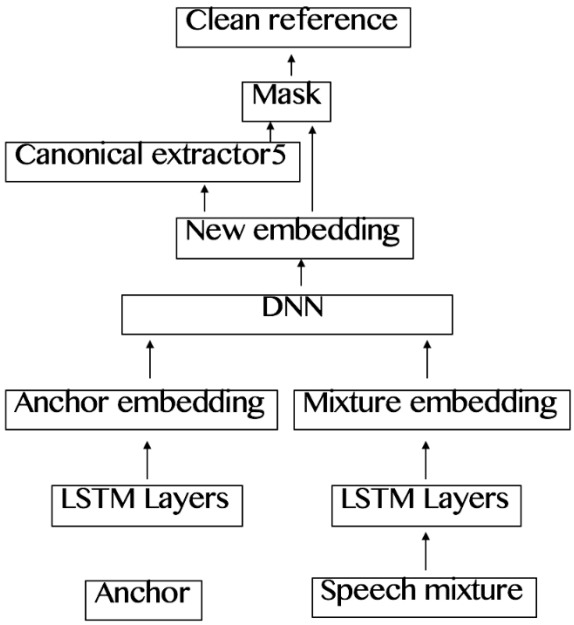

Figure 4.3: Deep extractor network architecture [14]

4.2.3 Anchor word based deep attractor network

Different from the deep attractor network stated in section 4.2.1 and deep extractor network stated in section 4.2.2, our proposed method directly utilizes anchor word as the source of attractor points to reduce the data volume and simplify the training process. Table 4.1 shows the difference between DANet, our proposed method and DENet.

Table 4.1 The comparison of DANet, our proposed method and DENet

Method	Input	Output	Basic model	Source of attractor points
DANet [1]	Speech mixture	All speakers' individual speech		Speech mixture
Proposed method	Speech mixture, anchor word of all speakers		DANet [1]	anchor word
DENet [14]	Speech mixture, anchor word of target speaker	Target speaker's individual speech		Speech mixture and anchor word

Our anchor word based deep attractor network adopts the anchor word from each speaker in the mixture to build the attractor points. First, anchor word audio and mixture speech audio are separately constructed in two k dimension spaces. For mixture speech S, dimension space is $X \in \mathbb{R}^{f_1 t_1 \times k}$, where f_1 represents frequency and t_1 represents time of time-frequency bins from mixture speech audio. Speech mixture is prepared for separation later. For anchor word w, dimension space is $Y \in \mathbb{R}^{f_2 t_2 \times k}$, where f_2 represents frequency and t_2 represents time of time-frequency bins from anchor word audio. Refer to Equation (16), attractor points A are created by anchor word in a high embedding space.

$$\mathrm{A}_{a,k} = \frac{\sum_{f_1 t_1} \mathrm{X}_{a,f_1 t_1} \times Z_{\mathrm{a},f_1 t_1}}{\sum_{f_1,t_1} X_{\mathrm{a},f_1 t_1}} \tag{17}$$

where a represents the anchor word time-frequency bins, and Z represents the energy ranking of the target source mentioned in Section 4.2.1. Equation (17) estimates the attractor vectors by calculating the source centroid of attractor points. Then a reconstruction mask is estimated by finding the similarity between time-frequency bins of anchor word based attractors and mixture speech using the following equation:

$$M_{f_2,t_2,a} = Sigmoid(\sum_{k} \mathrm{A}_{a,k} \times x_{f_2,t_2,k}) \tag{18}$$

Finally, the reconstruction error which represents the difference between mask and clean reference is calculated by Equation (1). Through the same optimization process mentioned in section 4.2.1, the deep neural network is forced to enhance the error and

achieving better separation results. In this network, as we utilize the anchor word utterance as the source of attractor points, it is more efficient for the system to find out the similarity of features between attractor points and time-frequency bins in the speech mixture. Future more, as the system knows the number of sources in advance, it could also solve the estimation error and dimension mismatch problem.

5. Evaluations

5.1 Dataset

5.1.1 Overview

In this research, we use ChiME-5 dataset [11] provided by the 5th Chi-ME Speech Separation and Recognition Challenge which challenges the problem of multiple speakers' speech separation and recognition in real home environments. Speech data are recorded in real home dinner scene from twenty families. Each dinner party has four participants. There are three recording locations in total: kitchen, living room and dining room.

The recording device used is Microsoft Kinect which consists of a camera and four microphones.

5.1.2 Data

In total, CHiME-5 dataset has twenty sets of different audio recording in different family, each has four participants, the distribution of data is shown as follows.

Table 5.1: Structure of dataset [11]

Dataset	Parties	Speakers	Hours	Utterances
Train	16	32	40:33	79,980
Dev	2	8	4:27	7,440
Eval	2	8	5:12	11,028

The total data size is 10.4GB. Each session of data consists of audio and transcription. Data format is WAV files and sampling rate is 16 kHz. Here shows a portion of data in training session:

Table 5.2: Detail of training data [11]

Session ID	Participants (Bold=Male)	Duration	#Utts	Notes
S03	P09, P10, P11, P12	2:11:22	4,090	P11 dropped from min ~15 to ~30
S04	P09, P10, P11, P12	2:29:36	5,563	
S05	P13, p14, p15, P16	2:31:44	4,939	U03 missing (crashed)
S06	P13, p14, p15, P16	2:30:06	5,097	

And the format of audio transcriptions is JSON, transcriptions include following information:

- *Session ID*
- *Location*
- *Speaker ID*
- *Transcription*
- *Start time*
- *End time*
- *Reference microphone array ID*

Here shows a transcription example in session 1:


```
{
    "end_time": {
        "original": "0:00:14.42",
        "U01": "0:00:14.42",
        "U02": "0:00:14.43",
        "U03": "0:00:14.43",
        "U04": "0:00:14.42",
        "U05": "0:00:14.42",
        "U06": "0:00:14.44",
        "P05": "0:00:14.42",
        "P06": "0:00:14.42",
        "P07": "0:00:14.44",
        "P08": "0:00:14.44"
    },
    "start_time": {
        "original": "0:00:13.10",
        "U01": "0:00:13.10",
        "U02": "0:00:13.11",
        "U03": "0:00:13.11",
        "U04": "0:00:13.10",
        "U05": "0:00:13.10",
        "U06": "0:00:13.12",
        "P05": "0:00:13.10",
        "P06": "0:00:13.10",
        "P07": "0:00:13.12",
        "P08": "0:00:13.12"
    },
    "words": "It seems delicious.",
    "speaker": "P01",
    "ref": "U01",
    "location": "dining",
    "session_id": "S01"
},
```


Figure 5.1: transcription annotation [11]

5.1.3 Anchor word

The original ChiME-5 dataset [11] consists of audio and transcription data. To create the anchor word data of each participant in the speech, we extract the individual first-time speech of each speaker in the transcription and clip the audio depending on the labeled start and end talking time. To avoid invalid audio fragment, we delete all blank pieces in each audio and randomly cut a three second utterance from it. Table 5.3 shows our data volume in comparison of DANet [1] and DENet [14].

Table 5.3 Data volume comparison

Model	Anchor word audio		Speech mixture audio		Total audio		Speakers per speech mixture
	Training data (h)	Testing data (h)	Training data (h)	Testing data (h)	Training data (h)	Testing data (h)	
Anchor DANet (Proposed)	0.49	0.01	40.56	5.10	41.05	5.11	4
DANet [1]	None	None	30.00	10.00	30.00	10.00	3
DENet [14]	8.97	1.39	103.18	15.97	112.15	17.36	3

From Table 5.3, we can see that our dataset is 5 hours longer than DANet [1], but we have more speakers to separate in the speech mixture. The data volume of DENet [14] is very huge in both anchor word and speech mixture. In our model, we are trying to realize speech separation with small data volume of anchor word audio: three seconds per speakers in average.

5.2 Experimental setup

The experiment is performed under Windows 10 system and implemented in Java and Keras. Table 5.4 shows the disposition of our experimental desktop. We construct both the baseline model [1] and anchor word based model with the same configuration as [1] which consists of 4 Long short-term memory layers and 20 embedding dimension space. Each layer has 600 hidden units in it. Also, in the feature extraction section, the length of window frames is set to 32ms and hop size is 8ms.

Table 5.4: Experimental computer disposition

Processor	Intel Core i7-7700K @ 4.20GHz
Memory	64 GB
Mainboard	GIGABYTE Z270X-Ultra Gaming-C

5.3 Results

In this research, we evaluate the performance of our algorithm with four parameters: perceptual evaluation of speech quality (PESQ) [5], signal-to-distortion ratio (SDR), signal-to-interference ratio (SIR), signal-to-artifacts ratio (SAR), and signal-to-noise ratio (SNR) [4][6]. Table 5.5 and Figure 5.2 show the evaluation results.

Table 5.5: Evaluation results

Model	Baseline DANet	Anchor DANet(Proposed)
SDR(dB)	10.95	12.30
SIR(dB)	11.47	14.53
SNR(dB)	17.35	19.69
SAR(dB)	19.62	20.57
PESQ	1.93	2.01

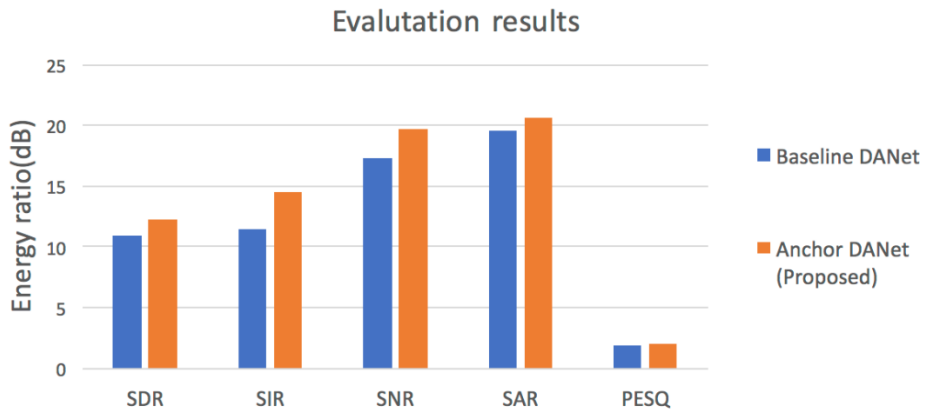

Figure 5.2: Evaluation results

From Table 5.4 and Figure 5.2 we can see that for all of the parameters our proposed anchor word based model achieved better results than the baseline model, especially SIR. SIR gains a 2.7% relative improvement in comparison with baseline, which testifies the largest interference in our dataset are other speakers' voice. And the improvement in SAR confirmed that by using the anchor word based attractor points in the model, it could successfully decrease the artifact which represents the estimation error of different source. Although all the results of SDR, SIR and SNR have better performance than Chen et al.'s DANet model [1], because our dataset is different, it is difficult to make direct comparison. I assume the reason of the improvement is because the dataset in [1] contains more interference.

In the comparison of training time, due to the replacement of attractor point source, the reduce of data size makes our proposed model efficient. It took 1.7 hours to finish the experiment including the time of data prepared preprocessing and the baseline model took 2.3 hours. Our proposed model is practical for specific speaker recognition and could be enabled under most scenarios like meeting record and house voice-controlled devices.

Table 5.6: Running time

Model	Baseline DANet	Anchor DANet(Proposed)
Time(h)	2.3	1.7

6. Conclusion

In this paper, we proposed a method that utilized the anchor word of source speakers to build an anchor word based speech separation algorithm. Here, we define anchor word as a short available utterance from the target speaker. In our proposed deep attractor network, anchor word is utilized as the source of attractor points, and by calculating the centroid of the anchor word speech in place of speech source in original deep attractor network, the experiment results show our new model not merely reserve the advantage of DANet [1], but also achieves higher efficiency and solves the estimation error problem. In comparison of DENet [14], our model uses less data especially anchor word audio to build attractor points, and simplify training process by reducing the analysis steps of speech mixture. Due to the difference of objective, we did not implement DENet [14] and make a direct comparison. However, we believe our model shows good flexibility and efficiency in speech separation task.

This model for speech separation is usable under most speech recognition scenarios as house voice-controlled devices, meeting recording and hearing impairment assistant. Although our proposed method achieves good separation, there are still some limitations. Currently, this model only supports dataset which announced the number of source speakers in advanced. For those with unknown source, anchor word based attractor point is impossible to build.

Acknowledgement

Upon the completion of my thesis, I would like to express my sincere thanks to some people. First, I extend my heartfelt gratitude to my supervisor, Prof Yamana for his valuable advices and selfless help on my research and thesis. Under his guidance and comments, I completed my study of research. Second, I would like to express my acknowledgement to my parents, friends and those people who care and love me. Thanks for your company all these year.

Reference

[1] Chen, Z., Yi, L., Nima, M.: Deep attractor network for single-microphone speaker separation. In Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 246-250 (2017).

[2] Narang, S., Gupta, M.D.: Speech feature extraction techniques: a review. International Journal of Computer Science and Mobile Computing, 4(3), pp.107-114 (2015).

[3] Kuhl, P.K.: Human adults and human infants show a "perceptual magnet effect" for the prototypes of speech categories, monkeys do not. Perception & psychophysics, 50(2), pp.93-107 (1991).

[4] Vincent, E., Gribonval, R., Févotte, C.: Performance measurement in blind audio source separation. IEEE transactions on audio, speech, and language processing, 14(4), pp.1462-1469 (2006).

[5] Rix, A.W., Hollier, M.P., Hekstra, A.P., Beerends, J.G.: Perceptual Evaluation of Speech Quality (PESQ) The New ITU Standard for End-to-End Speech Quality Assessment Part I--Time-Delay Compensation. Journal of the Audio Engineering Society, 50(10), pp.755-764 (2002).

[6] Yu, D., Weng, C., Seltzer, M.L., Droppo, J.: Microsoft Technology Licensing LLC. Mixed speech recognition. U.S. Patent 9,390,712 (2016).

[7] Shao, Y., Srinivasan, S., Jin, Z., Wang, D.: A computational auditory scene analysis system for speech segregation and robust speech recognition. Computer Speech & Language, 24(1), pp.77-93 (2010).

[8] Cooke, M., Hershey, J.R., Rennie, S.J.: Monaural speech separation and recognition challenge. Computer Speech & Language, 24(1), pp.1-15 (2010).

[9] Weng, C., Yu, D., Seltzer, M.L., Droppo, J.: Deep neural networks for single-channel multi-talker speech recognition. IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP), 23(10), pp.1670-1679 (2015).

[10] Yu, D., Chang, X., Qian, Y.: Recognizing multi-talker speech with permutation invariant training. arXiv preprint arXiv:1704.01985 (2017).

[11] Jon, B., Shinji, W., Emmanuel, V., Jan, T.: The fifth `CHiME' Speech Separation and Recognition Challenge: Dataset, task and baselines. In Proc. of Interspeech (2018).

[12] Ma, J., Hu, Y., Loizou, P.C.: Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions. The Journal of the Acoustical Society of America, 125(5), pp.3387-3405 (2009).

[13] King, B., Chen, I.F., Vaizman, Y., Liu, Y., Maas, R., Parthasarathi, S.H.K., Hoffmeister, B.: August. Robust Speech Recognition via Anchor Word Representations. In Proc. of Interspeech, pp. 2471-2475 (2017).

[14] Wang, J., Chen, J., Su, D., Chen, L., Yu, M., Qian, Y., Yu, D.: Deep extractor network for target speaker recovery from single channel speech mixtures. In Proc. of Interspeech 2018, pp.307-311 (2018).