Speech separation

By neural networks

University of Southampton

Individual project

Content

[1. Introduction 3](#_Toc6838840)

[1.1 Introduction of speech separation 3](#_Toc6838841)

[1.2 Single-channel and multi-channel contrast 4](#_Toc6838842)

[2. Neural networks 4](#_Toc6838843)

[2.1 Introduction of neural networks 4](#_Toc6838844)

[2.2 Basic components of neural networks 5](#_Toc6838845)

[2.3 Parameters of neural networks 5](#_Toc6838846)

[2.4 Different artificial neural networks 7](#_Toc6838847)

[2.5 Contrast of different artificial neural networks 10](#_Toc6838848)

[3. Pre-Processing Methodology 10](#_Toc6838849)

[3.1 Time-frequency Transformation 10](#_Toc6838850)

[3.2 Input-feature 11](#_Toc6838851)

[3.3 Separation Target 12](#_Toc6838852)

[3.4 Model Training 13](#_Toc6838853)

[4. Literature review 14](#_Toc6838854)

[5. Methodology Analysis 16](#_Toc6838855)

[6. Experiment results 16](#_Toc6838856)

[6.1 Generation of training data 17](#_Toc6838857)

[6.2 Results of DNN training 19](#_Toc6838858)

[6.3 Results of CNN training 19](#_Toc6838859)

[6.4 Results of RNN training 19](#_Toc6838860)

[6.5 Results of LSTM training 19](#_Toc6838861)

# Introduction

## 1.1 Introduction of speech separation

Speech separation is a classic, highly practical and valuable research topic. It also has been called “the cocktail party problem”, which was raised by XXX in the 1957s. The description of the problem: during a cocktail party, people talk with each other simultaneously. Everyone’s ears receive different voices and noises. It is relatively easy for a human-being to focus on a specific person’s voice and ignore others’. However, for a machine, this task is far more difficult. This is because, when different voices – or signals, as received by the machine - mingle together, the recognition of the voice will decrease dramatically as the machine cannot discriminate different signals coming from different sources.

The prelude of speech separation is the denoising problem, which describes the separation of useful signals from general noise. In speech separation, the goal is the to separate multiple useful signals. Hence, the speech separation problem is a higher-level version of the denoising problem, and also far more challenging.

Speech separation technology is widely used in many aspects of modern technology, such as voice communication, noise processing, intelligent speech recognition and so on. As all known, language recognition technology has developed very quickly over the last few years with the help of machine learning. As such, transforming the audio language signal into written language is not difficult for a computer anymore. However, in some cases, two or more audio signals mixed together, such as when multiple voices speak simultaneously, the computer would not be able to understand the signal. Some technology companies are already able to use the voice as an input to command smart facilities such as smartphone and robot. To achieve this, the audio signal is first transformed into written language, which includes words that can be programmed as commands. For example, ‘Siri’, a famous voice control system designed by Apple Co., can communicate with human easily. The problem is, if multiple people speak to Siri at the same time, the Siri will not be able to understand what are they saying, because the mingling of voices will ‘confuse’ Siri. Speech separation aims to separate different voices from different sources in order to allow the machine to understand each one separately.

In short term，speech separation technology is meant to distinguish between different sources of sound.

Due to its inherent complexity and difficulty, the speech separation technology in real life is far from satisfactory. With increasing study effort, many practical methods have been proposed and some breakthroughs have been made. What is most worth mentioning is that in recent years, along with the recent rise of powerful and practical neural networks, new prospects have been proposed for solving the speech separation problem.

The combination of traditional signal processing methods and modern machine learning through the use of neural networks has become a very viable solution for this problem.

## 1.2 Single-channel and multi-channel contrast

Speech separation can be based on single-channel and multi-channel signals. Here, the single-channel means the signal has been recorded by only one microphone; the multi-channel means more than one microphone has been used to record.

The main difference between single-channel and multi-channel signals is whether it contains the spatial orientation information. Because the multi-channel is a combination of multiple microphones in different positions microphones, the distances from the voice sources to different microphones will be different. Thus, the arriving time of the same voice source for different microphone will be also different. In addition, because of the damping effect of air, the voice will be weakened during the transmission in air. Hence, through the difference of arriving time and voice strength, multi-channel recording can achieve the spatial orientation information of the voice sources. However, for single-channel recording, only one microphone is used, so the spatial orientation information of the voice sources cannot be achieved. All in all, the conclusion is that the speech separation based on single-channel signals will be more challenging than multi-channel, because of the missing spatial orientation information.

This project is mainly focusing on single-channel speech separation. This is not only because the single-channel is more challenging, but also because, if the single-channel separation can achieve good performance, then the multi-channel separation must exceed it.

# Neural networks

## 2.1 Introduction of neural networks

Artificial neural networks are inspired by the biological neural networks which make up human and animal brains. As depicted in Figure???, where signals move through the axon, and are transmitted to the dendrites of connecting neurons, allowing the communication of information across the neurons system. Each neuron conducts calculations which influence the traveling signal, allowing for greater diversity in the transmitted information. 缺少衔接

Due to its outstanding performance in image recognition and speech recognition, it has received extensive attention from researchers in recent years, and is developing at a very fast pace.

图片包含 物体

已生成高可信度的说明

Fig ???

## 2.2 Basic components of neural networks

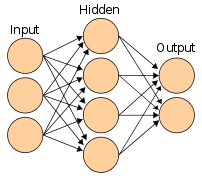


Fig ???

As the figure shown above, artificial neural networks have several different layers consisting of neurons. Each neuron can be considered as a memory node which contains weights and biases, which influence the strength of the signal. The input of neurons are the output of the upper layer neurons. If the neuron is in the first layer, the input will be the original input. When an input enters the neural node, its neural nodes perform calculations, and the result of the calculation is y=Wx+b. Then, pass the result of the calculation to the next layer, and so on until the last layer, and then output.

During the training process, each final output is compared to the target output (we call it ‘label’) and the difference between the predicted output and target output is calculated. This difference usually been called ‘loss’. In the next round of calculations, the neural networks will automatically adjust the value of the parameters, weights and biases. The way of adjustment is called backpropagation, that is, using the loss calculated in the previous round, calculate the gradient of the loss to the parameter. Then, the parameters are adjusted (increased or decreased) according to the direction in which the gradient falls, to achieve a smaller loss, which is closer to the desired prediction. That is roughly how neural networks work.

## 2.3 Parameters of neural networks

**2.3.1 Weight and bias**

Weight and bias are the most two basic and essential parameters in an artificial neural networks. Both independently, weight increases or decreases the input signal strength, and bias adjusts the value of the output. It can be simply expressed as the output equals the weight multiplied by the input plus the bias. This can be expressed in the following equation:

Where W is weight and b is bias.

The initial values of weight and bias ​​are either randomly generated or set manually. During training, the weights and biases are re-updated. The direction of the update is determined by their respective gradient directions. The size of the update changing is determined by the learning rate, so that the error between the predicted output value and the expected goal value – as known as label - becomes smaller and smaller. That is how training or learning work.

**2.3.2 Loss**

Loss is the error between the predicted value and the expected goal value. There are many ways to define and calculate it. Generally speaking, the mean squared error is the most common one, as shown in Formula 1. Cross-entropy is particular useful for image recognition, as shown in Formula 2. When the error is obtained, the partial derivative of the error to the weight or deviation can be calculated, thereby obtaining the gradient. Then, adjusting the values of the weight and bias according to the corresponding gradient direction. That is the basic idea of back propagation.

**2.3.3 Layer**

The structure of a normal artificial neural network can be divided into an input layer, hidden layers, and an output layer. The input layer is used to accept input data and conduct it to the hidden layer. The hidden layer passes the data to the output layer, which generates the output i.e. the prediction, to calculate the loss. In rare cases, networks may include multiple input or output layers. In general, the output layer and the input layer tend to have only one layer, and the hidden layer can have multiple layers. More hidden layer can significantly improve the perception ability of the neural network, and explore the deeper connotation of the data, thereby improving its learning ability and predicting more accurate results. It is worth noting that more hidden layers are not necessarily better, as too many hidden layers will lead to too many parameters, leading to gradient explosion, gradient disappearance and other issues, resulting in a decline in learning effects. Specific experiments showing this effect will be described in detail later. This is the basic idea of the structure of neural networks.

**2.3.4 Learning rate**

The magnitude of the learning rate is a parameter that needs to be set by people. The learning rate means the magnitude of each weight and deviation update adjustment. If the learning rate is too big, the learning might be not accurate enough because it will miss the best optimum point. For example, in an experiment the best value of a weight maybe 0.121011. If the learning rate is only 0.1, then the weight can only update by a 0.1 distance, from 0.1 to 0.2, or from 0.2 to 0.1. Thus it can never hit the best value 0.121011. On the other hand, if the learning rate is very small, the training process will be too slow. For example, the initial value of a certain weight is 0.1, the optimal value is 0.121011, and if the learning rate is 0.000001, then at least 21011 steps are required to achieve the best value. Therefore, the choice of learning rate is a hyperparameter that requires empirical judgment and practice to adjust.

**2.3.5 Activation function**

The activation function is usually placed between layers. For example, Suppose a network utilises the activation function ‘Relu’. If the result of the upper layer network is -0.5, then after experiencing the effect of the activation function, the input of the next layer would be 0; if the outgoing result of the upper layer network is 0.5, then after the activation function is applied, it will still be 0.5 input to the next layer. This is very effective in transforming a neural network from a single linear structure to a nonlinear structure. (maybe an example of sigmoid will be better?)

## 2.4 Different artificial neural networks

**2.4.1 Deep neural network**

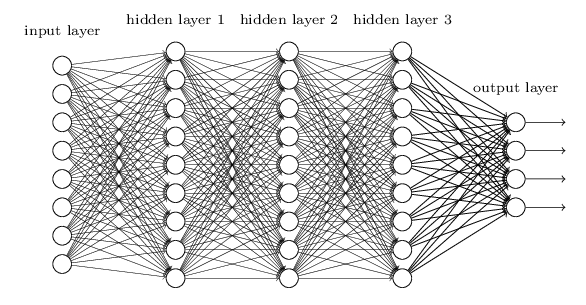
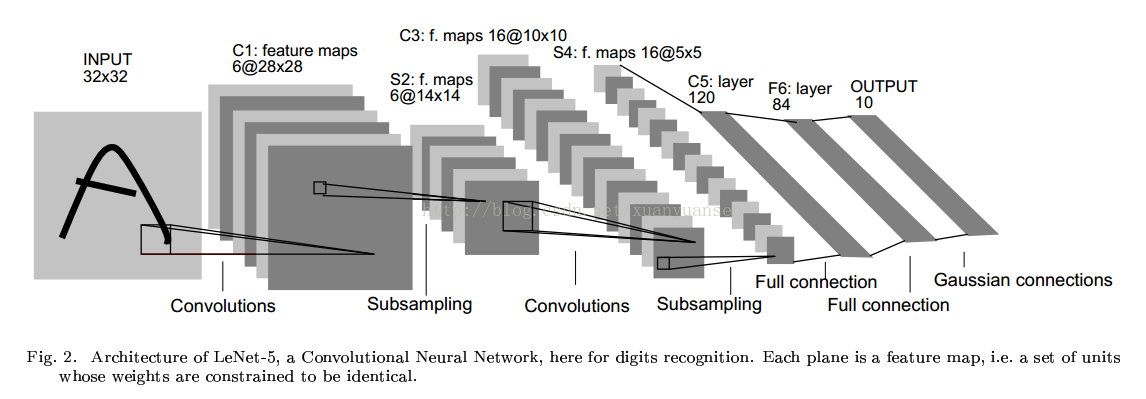


Figure ???

DNN (deep neural network) is the most common and simple neural network, and its basic structure is shown in the Figure???. DNN is an artificial neural network with a multi-layer network structure between the input layer and the output layer, hence the name "deep" networks. . Although simple and ordinary, DNN is very practical at the same time, and it is also the most widely used artificial neural network. Most of the issues, such as image recognition and speech recognition, DNN have shown excellent performance. This is because DNN can model both linear and non-linear situation. However, DNN also has some shortcomings, such as over-fitting, too much parameters, too long training time, and so on. Methods such as dropout, weight reduction, sparseness, etc. were invented and created to solve the over-fitting problem. These will be explained in detail in later experimental results and comparisons.

**2.4.2 Convolutional neural network**

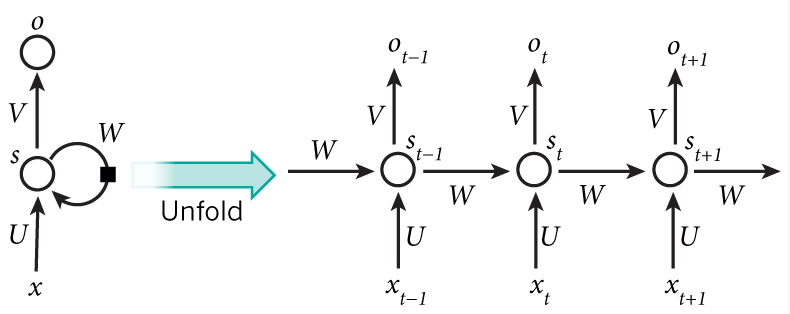
CNN (Convolutional neural network) is a very classic artificial neural network, especially in the field of image processing. Convolutional networks were inspired by biological processes in that the connectivity pattern between neurons resembles the organization of the animal visual cortex. Let's take a look at some of the concepts in convolutional neural networks: convolution, pooling, full connected, receptive field.

****

As shown above, the most important concept in a convolutional network is convolution. Here, there is some difference between the concept of convolution and the convolution in mathematics. More appropriately, the convolution referred to here should be the cross-correlation referred to in mathematics. The biggest benefit of convolution is that the number of parameters is greatly reduced. For example, a small image of 100\*100, if you use the DNN fully connected layer, then each neuron in the second layer will have 10,000 weights, which will make training very difficult and will cause the gradient to disappear or explode. . However, the invention of the convolutional neural network brought a solution to this problem because it drastically reduced the number of free parameters and made the network deeper. For example, in the classification problem, regardless of the size of the image, using 3\*3 pixels, combined with max-pooling, can reduce the number of parameters that need to be learned to one-ninth. In this way, not only the number of training parameters required is greatly reduced, but also the network becomes deeper and a deeper potential law can be learned.

**2.4.3 Recurrent neural network**

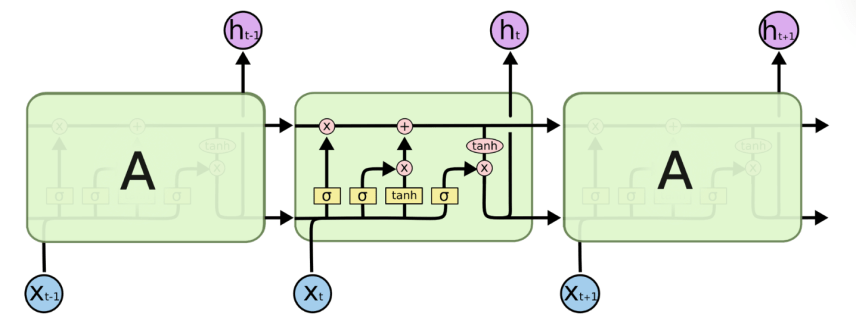
The most unique feature of the RNN network is its ability to handle timing issues such as speech recognition. From the above description, it is generally understood that the modelling process of DNN and CNN can be roughly expressed as inputting a bunch of numbers or matrices into a neural network with a particular characteristic structure, and then outputting a stack of numbers as a prediction. In this case, we can easily find that during this procedure, it obviously ignores the chronological structure of the data. In particular, one of the biggest differences between speech signals processing and image recognition is that the speech signal is a time series that changes with time. For example, if you ignore the chronological structure of the voice signal, it is equivalent to losing an important clue to detect an criminal case. Let's take a closer look at how RNN utilises the chronological structure of the data



As shown in the Figure, it is a schematic diagram of the transmission procedure of an RNN. The left side of the figure shows one of the RNN neurons, and the right side of the figure expands the RNN neurons in chronological order. That is, the right side is the state of the neuron at three different times, not three neurons.

The input X is sequentially passed into the neuron according to the internal chronological order of its data, and the neuron "remembers" the weight generated by the input and the corresponding output after each input. This weight is special. It is not preserved in its entirety, but is gradually 'dilute' and 'forgotten' over time, because as the newer input received, new weights will be generated, thus more memory is occupied by the new memory unit and the old memory will be gradually ‘forgotten’. This ‘forgotten’ mechanism is very similar to the processing mechanism in which the human brain receives sound signals. It is also because of this unique structure that RNN can achieve a good perception of the chronological order of the data, hence has a superior performance.

**2.4.4 Long Short Term Memory neural network**



LSTM is one of the RNNs, but improvements have been made on the basis of RNN, which can be called a progressive version of RNN. LSTM introduced the concept of 'forgotten gate' into the basic idea of ​​the original RNN. The memory of the original RNN for weight update is completely dependent on the passage of time. If the time is longer, then the less memory about the weight, or the weight will be weakened; on the contrary, if the time is closer, then the more memory about the weight, or the weight will be strengthened. In reality, however, the speech signal does not exhibit an absolute association with time. For example, the speech signals of human speech are not absolutely continuous, because people have pauses between words and words between sentences and sentences. These pauses are blank on the recording of the voice signal. This blank input and output will cause the weight to fail, resulting in a decrease in training effectiveness. The introduction of LSTM has solved this problem very well. The presence of the Forgotten Gate enables neurons to focus on remembering the weight of certain inputs (such as useful pronunciation), while giving up the weight of memory—that is, forgetting—some segments (such as gaps and pauses). For example, the pause in people's speech is in the blank time period corresponding to the spectrogram. In the ordinary RNN, because it can't be distinguished from the real pronunciation, it is memorized without any difference. For LSTM, the blank segment can be selectively forgotten. This is why LSTM has made further improvements in processing speech recognition than RNN and has demonstrated better performance.

From the above expressions, you may be aware that RNN and LSTM not only deal well with time series problems, but also greatly reduce the number of parameters that need to be trained. If it is a traditional DNN, assuming that every second of speech contains 10 points (in fact, much larger than this number), then ten seconds of speech need to be received with 100 neurons, and to the second layer there will be 100 For neurons, the resulting weight will be 10,000. For RNN or LSTM, only ten neural nodes are needed, and the same neuron can be input and output 10 times in time order. Therefore, the number of training parameters required is greatly reduced. The real experiment also proves that LSTM's performance in speech processing far exceeds other network structures, which we will detail in the results section.

## 2.5 Contrast of different artificial neural networks

The neural network comparison here is limited to a shallow macroscopic contrast. Specific and more detailed comparisons, including the effects of various parameters on training, will be made in the results section.

DNN

Advantages: simple structure and wide universality

Disadvantages: parameters are cumbersome, training time is long, over-fitting often occurs

CNN

Advantages: The convolution process cuts the parameters and makes the network deeper, making it more capable of adapting to nonlinear problems. Suitable for image processing such as image recognition and image classification.

Disadvantages: Pooling during convolution may result in loss of information.

RNN&LSTM

Advantages: a unique advantage for timing issues

Disadvantages: the network structure is more complicated

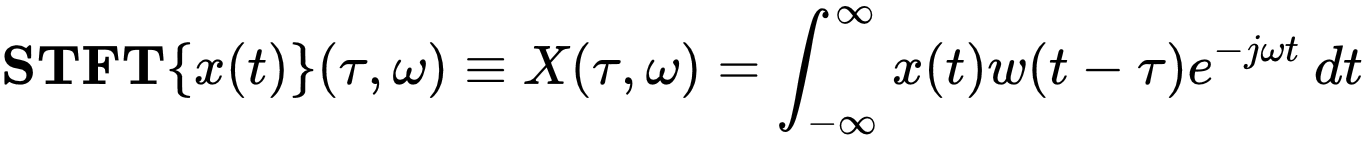
# Pre-Processing Methodology

The figure below shows the general structural composition and operational flow of speech separation. It can be divided into five modules: 1) Time-frequency transformation, 2) Input-feature, 3) separation target, 4) model training, and 5) signal synthesis, which are elaborated separately as following.

## 3.1 Time-frequency Transformation

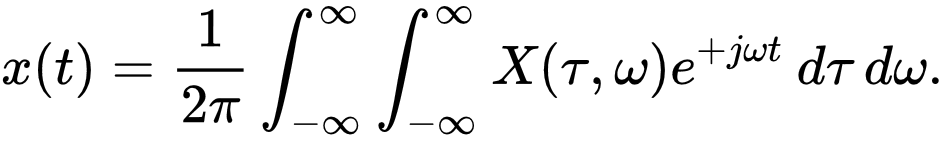
This is mainly intended to convert a monaural speech signal into a two-dimensional spectrogram. The commonly used methods are STFT and Gammatone filter. Due to professional background reasons, only STFT is briefly introduced, and Gammatone filter is not yet fully understood.

Simply, in the continuous-time case, the function to be transformed is multiplied by a [window function](https://en.wikipedia.org/wiki/Window_function) which is nonzero for only a short period of time. The [Fourier transform](https://en.wikipedia.org/wiki/Fourier_transform) (a one-dimensional function) of the resulting signal is taken as the window is slid along the time axis, resulting in a two-dimensional representation of the signal. Mathematically, this is written as:

{\displaystyle \mathbf {STFT} \{x(t)\}(\tau ,\omega )\equiv X(\tau ,\omega )=\int \_{-\infty }^{\infty }x(t)w(t-\tau )e^{-j\omega t}\,dt}

where w(t) is the [window function](https://en.wikipedia.org/wiki/Window_function), commonly a [hanning window](https://en.wikipedia.org/wiki/Window_function#Hann_window) or [Gaussian window](https://en.wikipedia.org/wiki/Window_function#Gaussian_window) centred around zero, and x(t) is the signal to be transformed. X(τ,ω) is essentially the Fourier Transform of x(t)w(t-τ), a [complex function](https://en.wikipedia.org/wiki/Complex_function) representing the phase and magnitude of the signal over time and frequency. Often [phase unwrapping](https://en.wikipedia.org/wiki/Phase_unwrapping) is employed along either or both the time axis, τ, and frequency axis, ω, to suppress any [jump discontinuity](https://en.wikipedia.org/wiki/Jump_discontinuity) of the phase result of the STFT. The time index τ is normally considered to be "slow" time and usually not expressed in as high resolution as time t.

The STFT is [invertible](https://en.wikipedia.org/wiki/Invertible_function), that is, the original signal can be recovered from the transform by the Inverse STFT. Mathematically, this is written as:



The most widely accepted way of inverting the STFT is by using the [overlap-add (OLA) method](https://en.wikipedia.org/wiki/Overlap%E2%80%93add_method), which also allows for modifications to the STFT complex spectrum. This makes for a versatile signal processing method, referred to as the overlap and add with modifications method.

## 3.2 Input-feature

For machine learning problem, input-feature extraction is always a crucial step. Extracting good input-features can greatly improve the performance of speech separation. From the basic unit of feature extraction, it is mainly divided into the time-frequency unit level and frame level.

The feature of the time-frequency unit level is to extract features from the signals of a time-frequency unit. The level of features is finer and can focus on even smaller details, but lacks the global and holistic description. Therefore, it cannot obtain the spatial-temporal distribution structure and inner temporal correlation of the voice. In addition, the signal of a time-frequency unit is difficult to characterize the perceptible speech characteristics (for example, phoneme). The time-frequency unit level input-feature was used as modeling input units in early time’s speech separation systems as reference[2,17]. These systems treat each time-frequency unit as an isolated unit and train binary values classifier ​​in each frequency band to determine whether a time-frequency unit on each frequency band is dominated by speech or is dominated by noise.

The feature of the frame unit level is to extract features from a frame spectrogram, like extracting a small image. It has a larger granularity and can capture the spatial-temporal distribution structure of speech spectrogram, especially the inner temporal correlation of the voice. It has better globality and integrity, and has obvious ‘voice-aware’ characteristics. The frame-level features are mainly used for frame-based modeling units. In speech separation systems, these systems generally input several frames of context frame level features to directly predict the separation target of the entire frame, for reference, [8,9,10,11,18]. In recent years, with the deepening of speech separation research, there are many auditory features that have been proposed and applied to speech separation, and have achieved good separation performance. These features include Mel-frequency cepstral coefficient(MFCC), Perceptual linear prediction(PLP), Relative spectral transform PLP(RASTA-PLP), Gammatone frequency cepstral coefficient(GFCC), Gammatone feature(GF), Amplitude modulation spectrogram(AMS), Pitch-based feature, Multi-resolution cochlea gram(MRCG), FFT-magnitude, FFT-log-magnitude, etc.[28,29,30,31].

In the Fourier transform domain, FFT-magnitude or FFT-log-magnitude is the most commonly used speech separation feature. Due to the low-frequency energy, FFT-log-magnitude can highlight high-frequency components compared to FFT-magnitude. However, some studies have shown that in speech separation, FFT-magnitude is slightly better than FFT-log-magnitude [32].

## 3.3 Separation Target

There are many important applications for speech separation. There are two main aspects to sum up: 1) Using the human ear as a target receptor to improve the intelligibility and perceived quality of the human ear to noisy speech, such as voice communication; 2) The machine acts as a target receptor, improving the accuracy of the machine's recognition of noisy speech, for example, for speech recognition.

For these two main speech separation applications, they have many close connections. For example, to improve the understanding of noisy speech, the speech separation system aiming at the degree of perception and perceived quality can usually be used as a pre-processing module for speech recognition, which can significantly improve the performance of speech recognition [33].

For these two main speech separation applications, many specific learning targets have been proposed. The commonly used separation targets can be roughly divided into three categories: time-frequency masking, speech amplitude spectrum estimation, and implicit time-frequency masking. The targets of time-frequency masking and speech amplitude spectrum estimation have been proven to be good at suppress noise and improve speech intelligibility and perceived quality [8,9]. Implicit time-frequency masking usually incorporates masking techniques into real-world application models, and time-frequency masking acts as an intermediate process to improve the performance of other targets, such as speech recognition [34], estimation of the target speech waveform.

Time-frequency masking is a common target for speech separation. Common time-frequency masking has ideal binary masking and ideal ratio masking, which can significantly improve the intelligibility and perceived quality of separated speech. Once the time-frequency masking target is estimated(if phase information is not considered), the time domain waveform of the target speech can be synthesized by inverse transform technique. However, recent studies have shown that phase information plays an important role in improving the perceived quality of speech [24]. The time-frequency masking targets of information are successively proposed, such as Complex ideal ratio mask (CIRM) [27]. The following are the brief introduction of those masking.

* + 1. Ideal binary mask, IBM

There are two methods to generate the IBM. The difference between two methods is the setting of the threshold( also called local criterion, LC).

or

As shown above, the first method use the signal noise ratio(SNR) as the criteria, but the second method use the absolute value of the spectrogram to compare with the value of threshold. No matter which method is using, the aim is to keep as much as voice information and filter out as much as noise.

* + 1. FFT ideal ratio mask, IRM FFT

Mathematically, this is written as:

Where Ys­­(t,f) is the pure voice signal STFT, and Yn­­(t,f) is the pure noise STFT.

* + 1. Complex ideal ratio mask, CIRM

The complex ideal ratio mask is very similar to the IRM above. The only difference is that traditional IRM is defined in the amplitude domain, while CIRM is defined in the complex domain. The benefit of CIRM is the influence of phase is considered which has been proven is related to the perceived quality.

## 3.4 Model Training

A typical supervised speech separation system uses a learning model to learn a mapping function from a noisy feature to a separate target. Many learning models have been applied to speech separation. The commonly used models can be roughly divided into two categories: shallow models and Deep model. In early supervised speech separation, shallow models usually directly model the probability distribution or differential modelling of the input noisy time-frequency unit, such as GMM[2] and SVM [35], or directly decompose the input noisy feature data to infer the components of speech and noise in the mixed data, such as NMF [36]. Since the shallow model does not automatically extract useful features from the data, Therefore, they rely heavily on artificially designed features. In addition, the ability of shallow models to process high-dimensional data is usually limited.

The deep-layer model has received great attention, has achieved great success in both fields of speech and image in recent years. Due to the nonlinear processing of the deep model layer, it can dynamically extracting the most powerful feature by learning the target from the input data. Compared with the shallow model, the deep model can process more primitive high-dimensional data, and the knowledge requirement for feature design is relatively low, and the deep model is good at learning data spatial-temporal structured features and output such structured prediction. Due to the mechanism of speech generation, the input features and output targets of speech separation exhibit obvious spatial-temporal structures. These features are very suitable for modelling with deep models. Many deep models are widely used in speech. In the separation, including DNN [9], CNN [13], RNN [10,11], Deep NMF [36] and LSTM [37]. For the consideration of words limits, the detail of those neural will be demonstrate in the final report.

# Literature review

In the real world, the voice signal is often disturbed by noise, which seriously impairs the intelligibility of the voice. As a result, the recognition of speech becomes infeasible. For noise, pre-processed speech separation technology is one of the most commonly used methods. A good speech pre-processing module can greatly improve the intelligibility of speech and the recognition performance of automatic speech recognition systems [1, 2, 3]. However, the scope of use of a particular voice pre-processing module is often limited. For example, the pre-processing module can do a good job in one task, but when switching to another task may be completely useless. Especially for monaural speech, the problem has not been well resolved.

In recent years, speech separation problems for monaural conditions have been highlighted and studied. From the perspective of signal processing, some methods have been proposed, such as Wiener filter [4] and spectral subtraction [5].

Wiener filter is a filter based on the separation of pure speech in the sense of minimum mean square error [6]. The assumption of Wiener filter is that the noise is a stationary or a slowly varying process, and that the noise spectrum does not change significantly in-between the update periods [7]. These methods can achieve good separation performance when these conditions are met. However, in reality, these assumptions are difficult to satisfy, and the separation performance is severely degraded, especially in the case of low signal-to-noise ratio, wherein the method is almost completely ineffective [6].

Speech separation is designed to separate useful signals from disturbed speech signals, this process that can be naturally expressed as a supervised learning problem [8, 9, 10, 11]. A specific supervised speech separation system typically learns a mapping function, which can result from a speech signal with noise to the desired separation target, the pure speech, through a supervised learning algorithm, such as a deep neural network [8]. Recently, supervised speech separation has received extensive attention from researchers and has achieved great breakthroughs and results. This new research trend, compared to traditional speech enhancement technology, does not require spatial orientation information of the sound source [12, 13, 14]. For monaural conditions, non-stationary noise and low signal-to-noise ratio, the deep neural network research have shown obvious advantages and promising prospects.

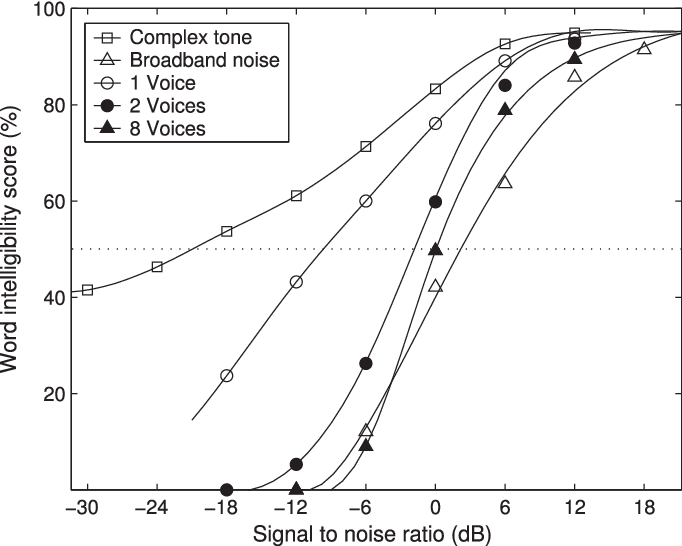
****

Figure ？？？. Word intelligibility score with respect to SNR for different kinds of interference (from [], redrawn from []). The dashed line indicates 50% intelligibility. For speech interference, scores are shown for 1, 2, and 8 interfering speakers.

From the perspective of supervisory learning, speech separation mainly involves three aspects: input-features, models and targets. Speech separation systems usually use time-frequency decomposition techniques to extract time-frequency spectrogram features from noisy speech. Commonly used time-frequency decomposition techniques are Short-time Fourier transform (STFT) and Gammatone auditory filtering models [15, 16]. Correspondingly, the speech separation features can be divided into Fourier transform spectrogram features and Gammatone filter transform features. Xu, Huang, Le Roux et al. used the Fourier amplitude spectrogram as an input feature for speech separation [9, 10, 14], while Chen, Wang et al. detailed and analysed some combined features and multi-resolution features under Gammatone filter transform features [17, 18]. The input-feature unit can be further divided into features at the time-frequency unit level and features at the frame level. The features of the time-frequency unit level are extracted from the signals of one or several columns in time-frequency unit, and the features of the frame level are extracted from one frame of the signal spectrogram. In the early days, due to the limitation of computational power, which limited machine learning ability, the supervised speech separation method usually used the time-frequency unit [2]. At this stage, supervised speech separation mainly uses frame-level features [8, 9, 10, 12, 14].

The supervised speech separation system can be further divided into a shallow model and a deep model. Early shallow models, included such as GMM, SVM, NMF. However, speech signals have spatial-temporal distribution structures and nonlinear relationships, and shallow structures are very limited in their ability to learn these nonlinear structural information. The deep model is very good at learning structural information in data due to its multi-layers nonlinear processing structure, and can automatically extract abstract feature representations. Therefore, in recent years, the deep model has been widely applied to the processing of images and has achieved excellent results.

Deep neural network (DNN) has also been widely used in speech separation, as shown on these documents [9, 11, 13, 20]. Recently, Le Roux, et al. extended NMF into a deep structure and applied it to speech separation, which has achieved great performance improvement [14]. The ideal time-frequency spectrogram mask and the amplitude spectrogram of the target speech are common targets for supervised speech separation [19]. If the phase of spectrogram is not considered, the target speech signal can be synthesized by using the estimated mask and amplitude spectrum. Experiments show that speech separated using this method can significantly suppress the noise [21, 22], improving the intelligibility of speech and the performance of speech recognition systems [23]. However, recent studies have shown that phase information is important for the perceived quality of speech [24]. Some speech separation methods begin to focus on the estimation of phase and improve the separation performance [25, 26]. In order to consider the phase information of speech during speech separation, Williamson et al. extended the ratio mask - different from the binary mask - to the complex domain, which be discussed in further detail later, and proposed the complex domain masking target. This has been found to significantly improve the perceived quality of the separated speech in a speech separation system based on deep neural networks [27].

As an important research field, speech separation has received extensive attention from researchers all over the world in recent decades. Recently, supervised speech separation technology has made important research progress, especially because of the application of deep learning. In the following, we will introduce and compare the input-features, models and targets used by aforementioned researchers.

# Methodology Analysis

# Experiment results

The experimental results are mainly divided into two parts. The first part is to compare the influence of various important parameters on the training results (i.e. good or bad); the second part is to compare the quality and training results of different neural networks.

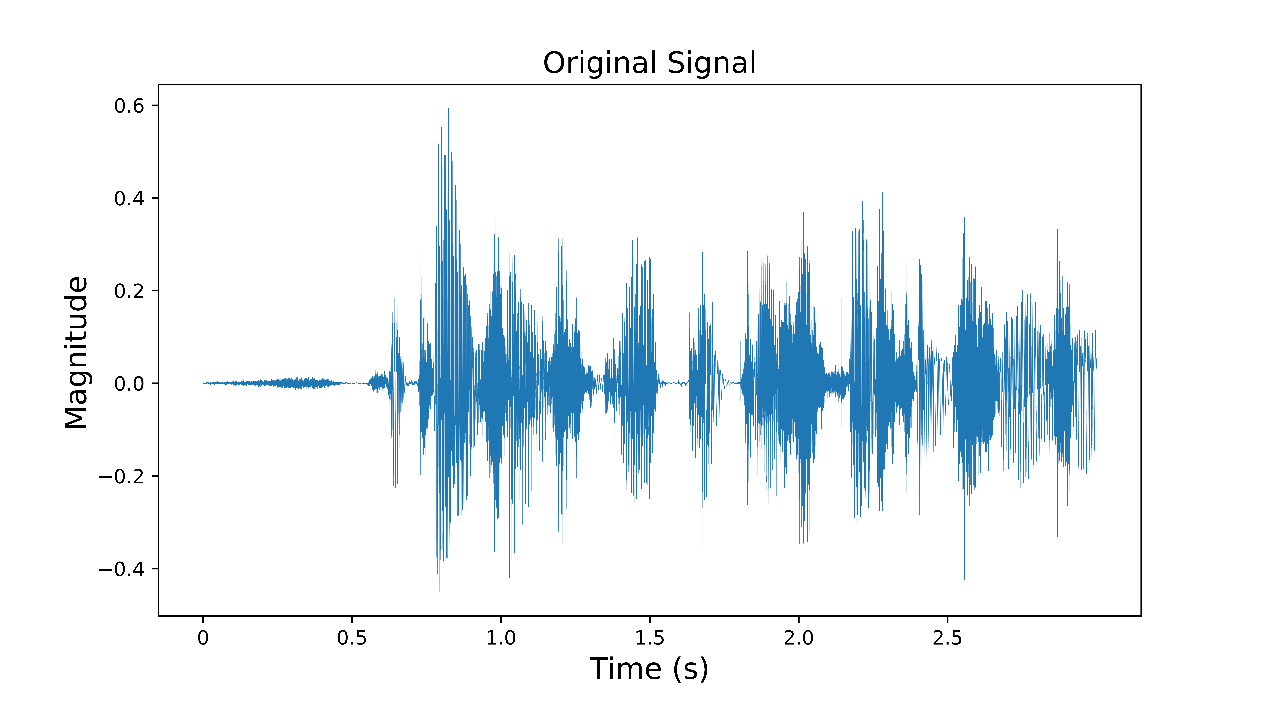
For the first part, although the network structure of DNN, RNN and CNN is different, many of their parameters are common, such as learning rate, activation function and so on. So most of the hyperparameters shared by each neural network will be compared when using DNN training. Can all the code for training the nerves be on my Github? ? ? See it? ? ? . The code is mainly based on Python, and the neural network is built on the Keras with TensorFlow as backend.

Because of computer power and time constraints, plus considerations for balancing unrelated variables, in the next 6.1??? , 6.2??? The data set used in the parameter comparison section remains unchanged. In the end, we will explore the impact of analysing dataset size and data generation methods on experimental results.

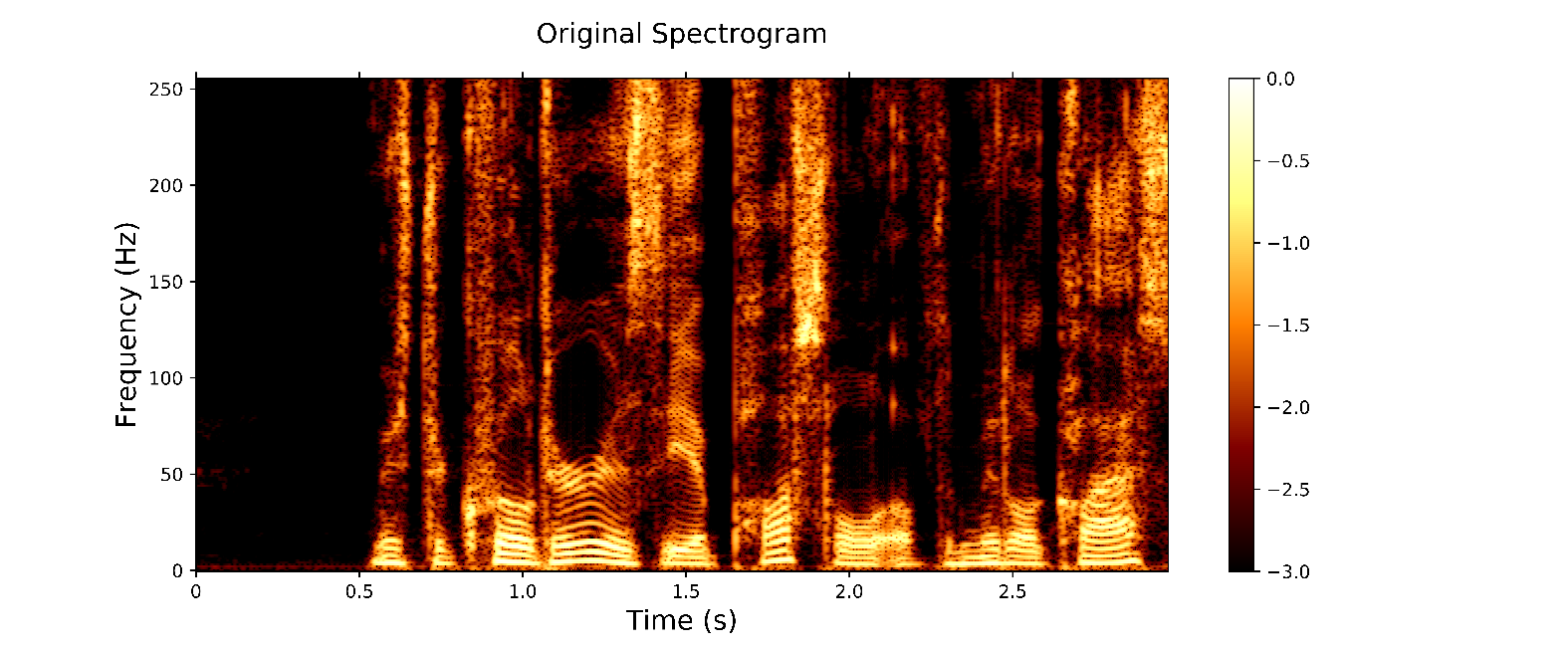
The above experiments are based on the separation of speech and noise. The separation between speech and speech will be in the next big chapter.

## 6.1 Generation of training data

**6.1.1 Short-time Fourier transform (STFT)**



The above picture is a time domain signal of a mono signal. First, we convert a speech signal into a time-frequency domain signal using the STFT method introduced in the methodology, as shown in the following figure.



This picture is wrong vertical axis.

The horizontal axis represents time, the vertical axis represents frequency, and the shade of colour indicates the magnitude of energy. When the colour is bright, the energy is higher.

At this point, the sound signal becomes an image. As shown in the figure, the bright areas, that is, the concentrated parts of the energy are mainly concentrated in the low frequency area, which is determined by the characteristics of human pronunciation. In general, vowel syllables, such as a, e, i, o, u are low frequencies and full of energy, while clear consonant syllables such as b, p, k, h are high frequency sounds and have lower energy. In fact, another particularly well-known field, speech recognition, is achieved by relying on these spectral features to classify human pronunciation.

**6.1.2 The parameter choice for STFT**

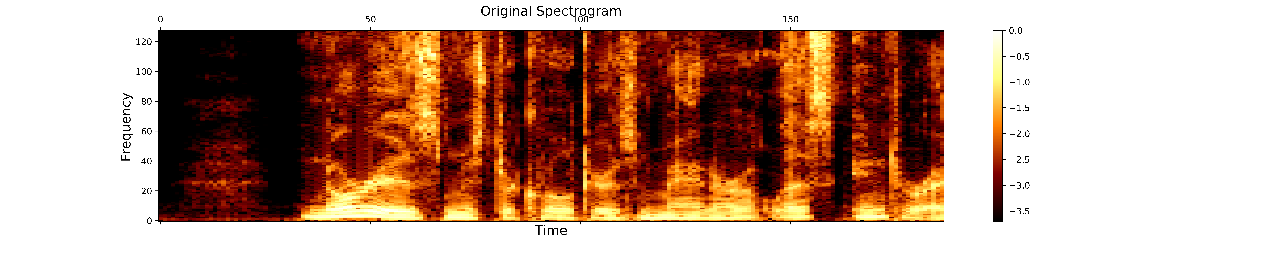
The mathematical principle and foundation of STFT have been introduced in detail in the previous section. However, when actually selecting the parameters of STFT, it is not only necessary to rely on the experience of the predecessors, but also many attempts. The choice of these parameters is important and will directly affect the results of the training (this will also be discussed as an argument in heading 8) because the use of different parameters will produce spectrograms of different sizes. That is to say, for the same segment of speech signal, different STFT parameters will be used to generate spectrograms of different sizes. Among them, the larger the spectrum map, the more detailed, it can be understood that the higher the resolution of the image, the richer and more complete the information it covers; the higher the resolution does not mean the better. Especially for neural networks. As we said in the principle section of the neural network we introduced earlier, too much input can cause the neural network parameters to be very cumbersome, resulting in a series of problems, such as the gradient explosion, gradient disappearing. However, if the resolution of the spectrogram is very low, it will be very rough, and a lot of information will be lost, and the distortion of the sound will occur when the ISTFT is performed, resulting in the failure of the task. Therefore, when actually selecting the parameters of the STFT, it is necessary to try according to the specific situation, and then select according to the result. Next, we will show the experiment of the actual selection of the parameters.

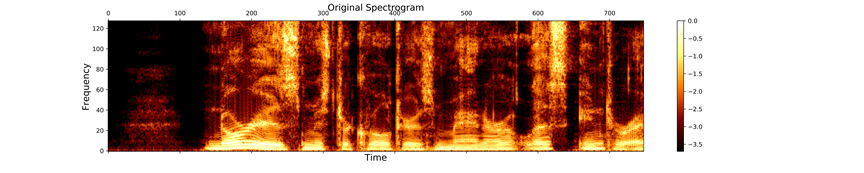
The two most important controllable parameters for the generation of the spectrogram are the window length (fft) and step size.

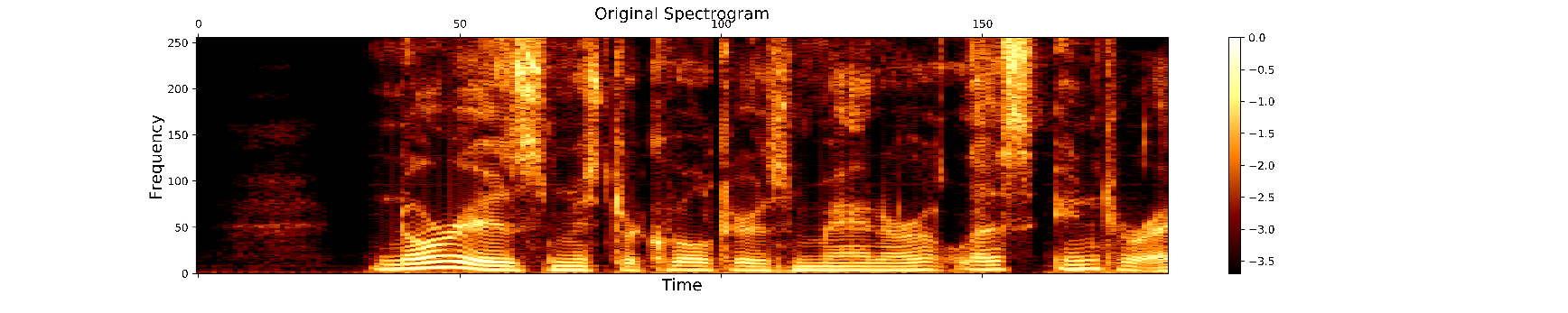
The step size is the size of the sliding window: the larger the step size, the smaller the density of the horizontal axis data points of the spectrum obtained by the Fourier transform; on the contrary, the smaller the step size, the denser the horizontal axis data points. An ideal example (without considering the boundary condition and overlap), if the signal of 1 second has 16000 points and the step size is 10, then 1600 windows are needed to complete the STFT; if the step size is 100, It only needs to be completed with 160 windows.

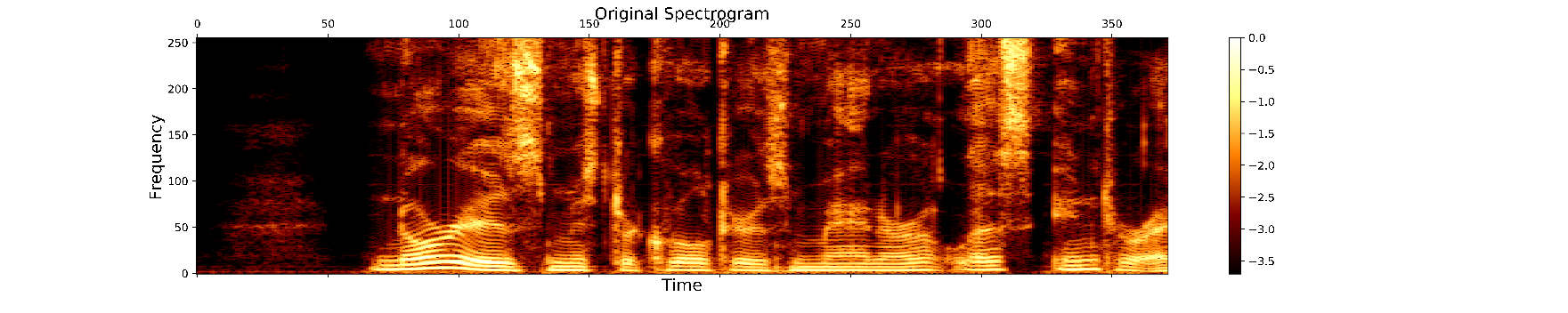
Where fft is related to the density of the data points of the vertical axis: the larger the value of fft, the denser the data points of the number axis; otherwise, the more sparse.

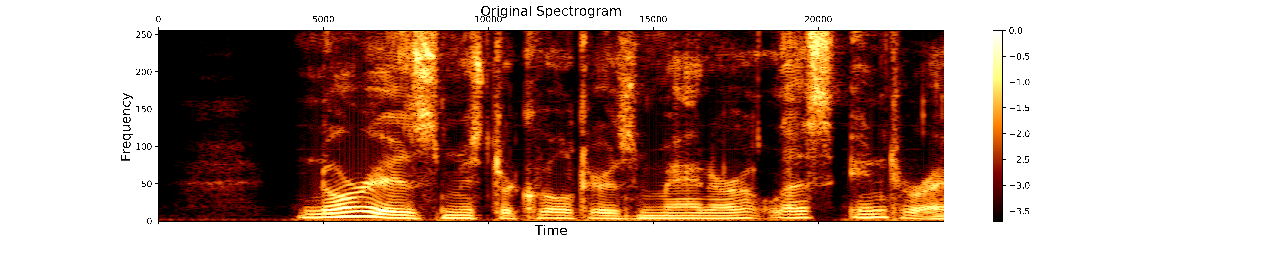
The following is a spectrum diagram generated by the variation of STFTs with five different parameter combinations for intuitive sensing.











The above five figures show the effect of these two parameters on the resolution of the spectrogram. But this does not tell us which one should be used as input for training. What we need to do is to inverse transform these spectrograms into a speech signal, judge the recognizability through the human ear, and finally make a decision (the generated demo can be found in my github, please click on it) Hyperlink------).

Finally, we chose the fft window size of 512, the step size of 128, 3 seconds of speech (48000 points) produced a spectrum map size of 372X256. The reason for choosing such a size is because this is the minimum spectrum under the premise that its inverse transform can be recognized by the human ear. The reason for choosing the smallest is to make training easy, so that as few data points as possible can cover more comprehensive information.

**6.1.3 The threshold choice for generating masks**

When generating a mask, the only determining factor is the magnitude of the threshold value. In simple terms, the size of the threshold determines how much information is retained in the original spectrogram. The selection of this value is also a matter of experience and skill.

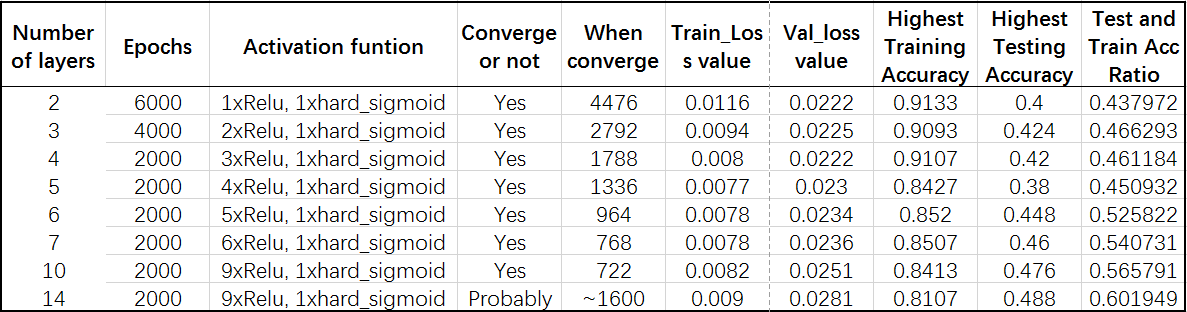
The principle of choosing the value of threshold is to include all the voice information as much as possible, and to filter out a considerable part of the background white noise. However, these two points are contradictory. The smaller the value is, the more it can filter out the noise. However, due to the high threshold, some useful low-energy sound signals such as unvoiced consonants will be filtered out at the same time. If the value is larger, the more likely it is to include all the sound signal information. However, because the threshold is too low, the white noise in the background cannot be filtered out, which becomes a kind of interference.

放三张图

It can be intuitively learned that the most comprehensive spectrum information is included in Figure 1, but it is also the largest; while Figure 3 is very concise, but it also obviously misses the information of the high frequency segment. Therefore, in order to select the appropriate threshold value, we add white noise to the original speech signal, and then let it be masked, then inverse Fourier transform into a speech signal, and select the denoising effect through the recognition of the human ear. The best one. In the end, we chose the result of taking a threshold of 4.

## 6.2 Results of DNN training

**6.2.1 The influence of number of layers**



**6.2.2 The influence of learning rate**

**6.2.3 The influence of number of columns**

**6.2.4 The influence of batch size**

**6.2.5 The influence of activation function**

## 6.3 Results of CNN training

## 6.4 Results of RNN training

## 6.5 Results of LSTM training