A survey of speech enhancement algorithms

Wei Zhang[1], Zhongzheng Wang[2], Xin Shi[3], Xiaojing Lao[4], Huihui Zhang[5]
1.32370 troops of the Chinese People's Liberation Army, Beijing China 100042,
2.Central North University, Shanxi Provincial Key Laboratory of Information Detection and Processing, Taiyuan Shanxi 030000,
3.31620 troops of the Chinese People's Liberation Army,Xuzhou Jiangsu 221000,
4.Unit 50 of the 32302 unit of the Chinese People's Liberation Army, Beijing, China 100000,
5.Dancheng Vocational Secondary School, Dancheng Henan 477150,

Absrtact: speech is easy to be interfered by the external environment in real applications, resulting in the reduction of speech intelligibility and signal-to-noise ratio. In the past few decades, due to the wide application of speech based solutions in practical applications, speech enhancement of noisy speech signals has aroused considerable research interest. This paper classifies and introduces several main speech enhancement methods, summarizes the advantages and disadvantages of several main methods, and finally puts forward the next research direction of speech enhancement methods.

Keywords: speech enhancement; Intelligibility; machine learning

1. Introduction

Voice communication is one of the most natural ways of human communication. Voice is an important tool to exchange ideas or express needs and emotions. Thanks to the progress of technology, voice communication is no longer limited to face-to-face dialogue, but also can be carried out through remote communication, and even has become a way of human-computer interaction. However, in practical applications, the speech signal will be affected by various noises. The received signal may include not only the voice of the target speaker, but also the voice or background noise of other speakers. An appropriate amount of background noise may not affect the intelligibility, but will reduce the perceived quality of speech. With the decrease of signal-to-noise ratio, it will become more and more difficult to understand speech. Therefore, in order to reduce the interference of noise, researchers have proposed many speech enhancement algorithms after decades of research.

2. Theoretical basis of speech enhancement

Before introducing the speech enhancement algorithm, it is very important to understand the basic characteristics of speech signal, human ear perception characteristics, various noise characteristics and the differences between different noises, which is conducive to taking corresponding speech enhancement strategies according to the noise in different environments.

The characteristics of speech signal mainly refer to its acoustic characteristics, time domain waveform, spectrum characteristics and statistical characteristics. Here are three main features of voice signal:

1) Speech signal is a non-stationary random signal. Its generation process is closely related to the change process of human vocal cords, vocal tracts and other vocal organs. The change speed of these organs in the process of phonation is far less than that of speech signal. Therefore, the shape of vocal cords and vocal tracts is relatively stable over a period of time (10 ~ 30ms). It can be assumed that the speech signal is stable for a short time, That is, some of its physical characteristics and spectral characteristics can be approximately regarded as unchanged. In the process of speech signal processing, the speech signal can be divided into several frames, and each frame of speech can be regarded as stable.

2) In phonetics, the voice with vibrating vocal cords is called voiced, and the voice without vibrating vocal cords is called unvoiced. Voiced voice has obvious periodicity in a short time in the time domain and low zero crossing rate in the period. It has formant structure in the frequency domain. The energy is concentrated in the lower part of the frequency, which is the part of speech with large amplitude and high energy. In speech enhancement, the periodic characteristics of voiced speech can be used to extract speech components or suppress non speech signals using comb filters; Unvoiced tone has no obvious characteristics in time domain and frequency domain, which is difficult to distinguish from broadband noise, but it can provide more information at higher signal-to-noise ratio.

3) As a random process, speech signal can be analyzed with statistical analysis characteristics. Its statistical characteristics can be described by its waveform amplitude probability density function, mean value, autocorrelation function and other statistics. The estimation method of the probability density of the statistical characteristics of speech signals is to calculate the amplitude histogram based on the absolute value of the amplitude of a large number of sampled data of speech signals over a long period of time, and find the approximate probability density expression based on the statistical amplitude histogram.

3. Speech enhancement algorithm

The noise in the real environment is diverse and has different characteristics. It is difficult to find a general speech enhancement algorithm for different elimination. People must choose different countermeasures according to different noise sources. For decades, scholars have summarized many effective methods. According to the different principles, speech enhancement algorithms can be divided into parametric methods, non parametric methods, statistical methods and other digital signal processing speech enhancement methods, as well

as speech enhancement methods based on machine learning. The speech enhancement method of digital signal processing has a long history and has a deep technical foundation, which is the main idea of speech noise reduction in the engineering field. The speech enhancement method based on machine learning realizes speech enhancement through supervised training. Some thorny problems in the field of digital signal processing, such as the elimination of instantaneous noise, can be easily solved by this method. Next, these methods are introduced and analyzed.

3.1 Parameter method

Parameter method is a method of estimating population with sample data when the population distribution is known. The parameter method mainly depends on the speech generation model used (such as AR model). It is necessary to extract the model parameters (such as pitch period, LPC coefficient, etc.) and generally uses the iterative method. The biggest disadvantage of parameter methods is that if the actual noise or speech is greatly different from the model, or it is difficult to extract speech parameters due to related reasons, such methods are more likely to fail. Such methods often use some filters, such as Wiener filter, comb filter, Kalman filter and so on.

Because the voiced segment of speech signal has obvious periodicity, comb filter can be used to extract speech components to suppress noise. The average value of the weighted sum of the delay of the input signal is used as the output signal of this filter. When the pitch period and delay of the signal are the same, the average process makes the periodic component strengthen, while the non periodic component or other periodic components with a period different from the signal will be suppressed or eliminated. Obviously, the focus of this method is to need the correct pitch period parameters. When the signal-to-noise ratio is not high, it is very difficult to estimate the period parameters, which will lead to inaccurate parameters, and the enhancement effect will be significantly reduced.

Wiener filtering uses an error criterion that is easy to deal with mathematically, that is, the optimal mean square error to calculate the enhanced signal. The noisy signal is transformed by a linear filter to approximate the original signal, and the linear filter parameters with the minimum mean square error are obtained. The biggest advantage of Wiener filter is that the enhanced residual noise is similar to white noise, rather than the rhythmic music noise. However, because it is the optimal estimation of the stationary process, and the noise in speech and the actual environment is non-stationary, Wiener filtering is not widely used in practical problems.

Both Kalman filter and Wiener filter belong to the optimal estimation in the sense of minimum mean square error. The difference is that Wiener filter has the constraint of stationary conditions, while Kalman filter does not have this limitation. It can also ensure the minimum mean square error estimation under non-stationary conditions, and can make up for the error caused by Wiener filter to a certain extent. Under the condition that the state equation and the statistical characteristics of noise are known, the Kalman filtering method realizes the minimum mean square error of waveform by analyzing the parameters through linear prediction (LPC). The Kalman filtering method is recursive and can be used in both stationary and non-stationary noise cases. Its disadvantage is that it needs a lot of calculation, and it needs to assume that the excitation source of LPC generation model is white noise source and only holds true for voiceless segment.

3.2 Non parametric method

The nonparametric method does not need to use the population information (population distribution, some parameter characteristics of the population) to infer the population distribution based on the sample information, so it does not need to estimate the parameters of the speech model from the noisy speech signal, so the nonparametric method has a wide range of applications compared with the parametric method. However, due to the lack of available statistical information, the results are usually not optimal. Such methods include spectral subtraction, adaptive filtering and so on.

Spectral subtraction algorithm is one of the earliest noise reduction algorithms. This method does not use the reference noise source, but assumes that the expected value of the noise amplitude spectrum during speech is equal to the expected value of the amplitude spectrum without speech gap noise, and uses the estimated value of the noise spectrum calculated from the gap measurement without speech to replace the spectrum of the noise during speech, The estimated value of pure speech signal is obtained by subtracting it from the noisy speech audio spectrum, and is set to zero when the difference is negative. The advantage of spectral subtraction is that the total amount of calculation is small and it is easy to implement in real time. One disadvantage of spectral subtraction is that it gives up the assumption of analyzing the speech spectrum. For human ears, the perception of speech is mainly obtained by the amplitude of each spectral component in the speech signal. Moreover, if there is a large noise component on the frequency point of a frame, there will be a large noise residue after spectral subtraction, and the enhanced speech will be mixed with rhythmic “music noise”.

The adaptive filter takes the minimum mean square error or variance as the criterion to estimate the noise in the noisy signal, and subtracts the estimated value from the noisy signal to achieve the purpose of speech enhancement. Its most important characteristic is that it can effectively track the time-varying input signal in the unknown environment, so that the output signal can reach the optimal. The adaptive filter usually uses FIR filter and LMS algorithm for iterative estimation. The key of this method is how to get the noise in noisy speech. There is a certain distance between the two microphones in the multi-channel acquisition system, so the noise of the two signals collected in real time is different, and it is also affected by echo and other variable attenuation characteristics. When using a mono system to collect noisy speech, the collected noise must be used for estimation during the speech interval. If the noise is non-stationary, it will seriously affect the speech enhancement effect; Another disadvantage is that the enhanced speech contains obvious “music noise”.

3.3 Statistical methods

Statistical methods make full use of the statistical characteristics of speech and noise. Such methods generally obtain the initial statistical parameters through a training process after the model base is established, and then update these statistical parameters according to the actual data after the subsequent processing process, so as to make the model better conform to the actual situation. Such methods mainly

include minimum mean square error estimation (MMSE), minimum mean square error of logarithmic spectrum estimation (mmse-lsa), auditory masking effect, etc.

Since the distortion criterion in the sense of hearing and the posterior distribution of speech spectrum under given noise can not be determined in speech enhancement, it is important to estimate the method insensitive to the specific distortion criterion and posterior probability, such as the short-term amplitude spectrum speech enhancement algorithm based on minimum mean square error (MMSE). This method uses the known noise power spectrum information to estimate the short-time spectrum of pure speech from the short-time spectrum of noisy speech, so as to achieve the purpose of speech enhancement. Since the human ear's perception of the sound intensity is proportional to the logarithm of the spectral amplitude, it is more reasonable to use the logarithmic distortion criterion when processing the speech amplitude spectrum. The MMSE estimation formula is extended to obtain mmse-lsa. MMSE makes a trade-off between noise reduction and speech intelligibility improvement. It is applicable to a wide range of signal-to-noise ratio, but it has a large amount of calculation and poor real-time performance. Moreover, the prior distribution of speech spectrum depends on the representativeness and reproducibility of statistical results to a large extent.

Auditory masking is an enhancement algorithm based on human auditory characteristics. The human ear can mask the noise signal with less energy in the speech signal, so that this part of the noise is not perceived by people. Auditory masking model is often combined with speech enhancement algorithm to achieve denoising. The implementation process is as follows: first, the speech signal is estimated based on a speech enhancement method, and then the auditory masking threshold is calculated from the speech signal rough estimation. According to the auditory masking threshold and the estimation of noise parameters, the gain is calculated combined with the corresponding enhancement algorithm, and the pure speech is estimated according to this. This method can reduce unnecessary speech distortion while eliminating noise. However, because the noise masking threshold is obtained on the basis of pure speech, in practical applications, only noisy speech can be used to estimate the masking threshold, which has a large error.

Compared with the nonparametric method, the method based on statistical model can not only obtain better enhancement effect, but also effectively remove the influence of music noise.

4. Summary

This paper expounds the parametric method, non parametric method and statistical method respectively, and analyzes the advantages and disadvantages of various methods. Speech enhancement involves not only signal processing theory, but also speech characteristics, noise characteristics and human ear perception characteristics; Moreover, the sources and types of noise are different, so the processing methods should also have diversity. In practical applications, through the continuous improvement of the algorithm or the combination of several algorithms can get better enhancement effect, but there are still shortcomings. The traditional digital signal processing method is integrated with the idea of machine learning, and the digital signal processing method is used to preprocess the speech signal, so as to achieve the goal of further improving the speech enhancement method. Such a combined algorithm may become the mainstream direction of speech enhancement in the future era of artificial intelligence.

References:

[1] Linjing Cao A review of speech enhancement technology [j]Journal of Hebei Academy of Sciences, 2020, 37 (2): 7

[2] Chengshan Fu Overview of adaptive speech enhancement algorithms [j]Information and communication, 2016 (10): 3

[3] Dong Yin,Shequan Jiang,Baoguang Liu, etal Overview and performance analysis of speech enhancement algorithms [j]Electroacoustic technology, 2015 (5): 5

[4] Yongjun He,Maoguo Fu,Guanglu Sun A review of speech feature enhancement methods [j]Journal of Harbin University of technology, 2014 (02): 23-29

About the author: Zhang Wei (1997 -), male, Han, Shanxi Yuanping, undergraduate, assistant engineer, mainly engaged in signal processing and speech recognition.