Speech Recognition Technology Based On Microphone Array

YANG Wen-shuai¹, ZHAO Hong-xu²

(College of Information Engineering, Engineering University of PAP, Xi’an 710086, P.R.China)
¹yangswjgd@163.com
²3302569257@qq.com

Abstract. In recent years, with the development of voice technology and the improvement of technology level, speech recognition technology has reached a large-scale commercial level. However, speech recognition technology still has many defects, and the existence of these defects restricts the accuracy of speech recognition. For example, when the multi speaker problem exists, the accuracy of speech recognition will drop dramatically, which will affect the effect of speech recognition. From the perspective of microphone array, this paper attempts to combine the positioning and noise reduction technology of microphone array with speech recognition technology in order to solve this problem. Experimental results show that microphone arrays can locate specific speakers, and design nulls to remove noises beyond the speaker's direction.

1. Introduction

Speech Recognition Technology, also known as Automatic Speech recognition (ASR), is a speech processing technique that converts the vocabulary of human speech into machine-readable input. Unlike speaker recognition based on voice recognition and speaker confirmation, the latter attempts to identify or confirm speakers rather than the vocabulary contained. With the development of speech recognition technology, this technology has been used in voice dialing and voice navigation, and has a very wide application prospect.

However, the maturity of speech recognition technology is far from the expectations of people. There are many problems to be solved in the application of this technology. KE Dengfeng and XU Bo pointed out in the literature ¹ that there are three unsolved problems of speech recognition technology——environmental noise problem, spoken language problem and multiple speaker problem. In order to improve the precision of speech recognition and promote the development of speech recognition technology, we must find a way to solve these problems.

This article focuses on the problem of multiple speakers, in order to solve this problem and improve the accuracy of speech recognition. In recent years, microphone array technology has been widely used in the field of sound source localization² and noise suppression³, and has received good results. The author tries to combine the microphone array technology with the speech recognition technology, and to solve the problem by positioning the speaker and designing the zero-trap method.

2. Problem description and signal model

2.1. Problem description

Suppose that in the indoor environment, there are G (G≥2) speakers who are talking at the same time.
When using speech recognition technology to identify a particular speaker, people expect the final computer's readable input include only the data that the speaker's voice converts to\(^4\). While speech recognition technology works in an environment where multiple speakers speak simultaneously, the voice signals of multiple speakers will overlap. At present, there is no good algorithm to separate the overlapping parts, and there is no speech recognition algorithm that supports simultaneous decoding of multiple targets. Therefore, the author from the source of speech recognition technology——information source to make the received voice signal of speech recognition technology contain the rest of the speaker's voice as little as possible so as to reduce the signal aliasing of speakers. Therefore, the received speech signal of speech recognition algorithm is as pure as possible.

2.2. Array signal model

Suppose a uniform linear array composed of 4 elements is arranged within the computer, and the array interval is less than half the wavelength of the received speech signal (\(d \leq \frac{\lambda}{2}\)). Select one of the microphones as the reference microphone array, and assume that the processing of the signal of microphone array accords with the conditions of far-field incident. Expect the speech signal is \(\theta_1\) to be relative to the microphone array, its angle range is (-10\(^\circ\), 10\(^\circ\)) relative to the reference array, and the other speaker's speech signal is in the range of the reference array \(\theta_2\), its angle range is (60\(^\circ\),80\(^\circ\)). Figure 1 is the reference array element.

![Figure 1: The signals’ model of this article.](image)

It is assumed that the frequency range of the expected speaker and the other speakers is \([f_a, f_b]\), which is divided into M sub band in the wideband signal processing. Suppose the \(m\) speaker's desired voice signal and the rest of the speaker signal received by the reference microphone can be \(S_{Q/f_m}(t)\) and \(\sum_{g=1}^{G} S_{g/f_m}(t)\) respectively. For the microphone array, the k-signal received by the k-array can be expressed as:

\[
x_k(t) = \sum_{m=1}^{M} a(f_m, \theta) S_{Q/f_m}(t) + \sum_{m=1}^{M} \sum_{g=1}^{G} a(f_m, \theta) S_{g/f_m}(t) \tag{1}
\]

In \(k = 1, 2, ..., 4\), \(f_m\) represents the central frequency corresponding to the sub band, and \(f_m \in [f_a, f_b]\).

For this signal model, this signal is composed of the desired speech signal and the voice signal of the G-1 speaker. Based on the model, the author will study the speech recognition technology of microphone array with a view to starting from the point of view of speech recognition technology, and suppressing the \(\sum_{m=1}^{M} \sum_{g=1}^{G} a(f_m, \theta) S_{g/f_m}(t)\) mixed speech signal (which is not expected speaker) so as to solve the
problem of speech overlap when multiple speakers are co-existence in speech recognition.

2.3 The model of speech recognition technology
The main principle of speech recognition technology is based on statistical pattern recognition. The goal of speech recognition is to convert the input speech eigenvector sequences $S = s_1, s_2, ..., s_I$ into word sequences $W = W_1, W_2, ..., W_J$ by using linguistic and phonetic information. The speech recognition model based on this technique can be expressed as:

$$\tilde{W} = \arg \max \{W \mid S\} = \arg \max \{\frac{P(W \mid S)P(S)}{P(S)}\}$$

$$= \arg \max \{\log P(W \mid S) + \lambda \log P(W)\}$$

This formula shows that in order to get the best output, the numerator of $(P(W \mid S)P(W))$ must be the largest.

3. The algorithm proposed in this article

3.1. Algorithm introduction
The basic idea of speech recognition algorithm (PL2) in this article is as follows: joint angle estimation based on PM and beam forming based on LCMV is proposed, which combines the speech recognition algorithm of the filter microphone array. First, we estimate the direction of the expected speaker speech signal (the desired signal) to the reference microphone array through the PM technique, then divide the speech signal by the voice separation technology based on music method, and finally use the LCMV beam forming technology to process the separated signal so that the expected signal can be all passed and the unexpected signal to be suppressed.

3.2. Method of DOA estimation based on PM
First, the direction of the wave is obtained by DOA estimation to determine the angle. The PM algorithm mentioned in Liu Chengcheng has good DOA estimation performance and can be used for DOA estimation under this model. The core idea is as follows: the signal model of the problem is:

$$x_k(t) = \sum_{m=1}^{M} a(f_m, \theta_s) S_{Qf_m}(t) + \sum_{m=1}^{M} \sum_{g=1}^{M-1} a(f_m, \theta_g) S_{Qg}(t)$$

The total array output can be expressed as:

$$X = (x_1, x_2, ..., x_M)$$

The PM algorithm does not need a subspace based method for EVD or SVD of the high-dimensional covariance matrix, and the PM algorithm uses the covariance of the output data of the element to calculate the noise subspace and its projection operator. The core of the PM algorithm is to remove the propagation operator by least squares. The specific steps are as follows:

1. The covariance matrix of the output data of the microphone array is computed and its expression is as follows:

$$R = E\{X, X^H\}$$

2. Partitioning the covariance matrix of output data:

$$R = [G, H]$$

3. Extract propagation operator:

$$P = (G^H G)^{-1} G^H H$$

4. The zero space of R is obtained by means of propagation operator, that is, noise subspace:

$$Q_H = [P_H, -I_{M-N_E}]$$

5. To determine the direction of the wave, the maximum spectral peaks of the power spectrum are searched:
\[
F(\theta) = \frac{1}{\|Q^{\theta} a(\theta)\|} = \frac{1}{a(\theta)^{Q}\| Q^{\theta} a(\theta) \|}
\] (9)

3.3. Voice separation

In the process of speech separation, two FIR filters are selected to separate the mixed speech signals. Because the discussion of the wideband signal model is based on the frequency domain, the author generalizes the LMS algorithm to the frequency domain so as to realize the speech separation under this model. The basic principles are as follows:

The mixed speech signal received by the reference microphone is known as:

\[
x_d(t) = \sum_{m=1}^{M} a(f_m, \theta) S_{g f_m}(t) + \sum_{m=1}^{M} \sum_{g=1}^{G-1} a(f_m, \theta_N) S_{g f_m}(t)
\] (10)

At the k-frequency point of t-time, the power coefficients of the filter can be expressed as:

\[
W(t, k) = FFT[w(t, n), 0_M]
\] (11)

\(w(t, n)\) is the time domain representation of the weight factor. According to the same time-frequency conversion method, we can get the cross-correlation function of time t, and the final frequency domain expression of the algorithm can be calculated as follows:

\[
W(t + 1, k) = W(t, k) + \mu FFT[\phi(t, n), 0_M]
\] (12)

\(\mu\) is the step parameter of the algorithm.

3.4. Wideband signal filter based on LCMV

After the speech separation, the mixed speech signal can be separated into two signals, which are the speech signals of the expected speaker and the mixed signal of the voice signal of the non-desired speaker. Therefore, we can use the idea of design filter, by designing the proper zero-trap in the direction of the unwanted speaker and making the speech signal of the speaker's direction complete, we can eliminate the interference of the separated non-speaker speech signal so as to make the speech signal received by the speech recognition algorithm purer. Based on the design principle of minimum variance criterion (LCMV) filter, combined with the model described in this article, the idea can be expressed by the mathematical model shown in the following formula.

\[
\begin{align*}
\min w(k)^H R_s(k) w(k) \\
\text{s.t.} C_1^{\theta} w(k) = h \\
\text{s.t.} C_2^{\theta} w(k) = 0
\end{align*}
\] (13)

In this formula, \(C_1\) and \(C_2\) is the constraint matrix; \(h\) is the response vector in the direction of constraint. The sub band signal of \(m\) is received from each array of microphone array, and the DFT is used to obtain:

\[
X(k, m) = [X_1(k, m), ..., X_N(k, m)]
\] (14)

In \(k = 0, 1, ..., N - 1\), \(N\) is the number of sampling points that are included in the sub band.

\[
x_k(t) = \sum_{m=1}^{M} a(f_m, \theta) S_{g f_m}(t) + \sum_{m=1}^{M} \sum_{g=1}^{G-1} a(f_m, \theta_N) S_{g f_m}(t)
\] (16)

For the pure LCMV method, combined with the Lagrange function operator, it can be solved:

\[
w_{out}(k) = R_s^{-1}(k) C^{(k)} R_s^{-1}(k) C^{-1} g^H
\] (17)

Suppose:

\[
w(k)^H a(f_m, \theta) = 1 \quad g^H = 1_M^H
\] (18)

The optimal weight vector of the mathematical model can be:

\[
w(k) = R_s^{-1}(k) a(f_m, \theta) a(f_m, \theta) a^H(f_m, \theta) a^H(f_m, \theta) R_s^{-1}(k) a(f_m, \theta)^{-1}
\] (19)
4. Experimental simulation and analysis
In order to prove the validity of the above algorithm, the author simulates the working process of the algorithm based on the MATLAB tool. Suppose in a $7 \times 4 \times 3$ space (indoor environment), there are three speakers. The expected speaker's direction is $10^\circ$ to the reference microphone array so that the outer two non-expected speaker is in the $70^\circ$ direction of the reference array. The microphone array is a uniform linear array composed of 4 elements, and the array spacing is half wavelength. Based on this model, the author launches the experiment and obtains the following experimental results.

4.1. DOA estimation based on PM

![Figure 2: DOA estimated results.](image)

It can be seen from this graph that there is a voice signal in the $10^\circ$ and $70^\circ$ directions of the reference array. According to the existing hypothesis, it can be concluded that the former is the voice signal of the expected speaker, and the latter is a mixed signal of two non-expected speaker speech signals.

4.2. Voice separation

![Figure 3: Signal schematic of mixed speech.](image)
Figure 3, Figure 4 and Figure 5 give the mixed speech signal spectrum of the microphone array, the spectrum of the recovery signal in the 10° direction and the spectrum of the recovery signal in the direction of 70°. From these three graphs, we can see that after the speech separation, the signal is divided into two kinds of signals roughly. The former is the voice signal of the expected speaker, and the latter is the mixed signal of the voice signal of the non-desired speaker. After the speech separation algorithm, the speech separation of mixed signal has been realized, which indicates the direction of the next filter design.

4.3. LCMV filter

After speech separation, the signal is passed through the LCMV filter. The aim is to make the desired speaker's voice signal complete through the filter and contain as few unexpected speakers as possible. Compared to Figure 6 and Figure 4, it can be concluded that after filtering the LCMV algorithm, the output waveform of output signal is close to the speech signal waveform of the desired speaker, it is explained that the LCMV filter retains the speech signal of the speaker in 10° direction, and it effectively suppresses the mixed signal of the 70° direction.

4.4. Speech recognition

In the speech recognition experiment, the author uses the control experiment method. The first group (A Group) was not processed by the PL2 algorithm, but the speech voice of the hypothesis model was identified directly, the second group (B Group) was processed by the PL2 algorithm, and then the speech recognition was made using the processed speech signal. Among them, three speakers randomly speak, compare their error rate and get the following table:
Table.1. Error rate comparison table for two groups of experimental recognition algorithms.

| Group | 1    | 2    | 3    | 4    |
|-------|------|------|------|------|
| A     | 7.47 | 6.45 | 8.51 | 7.49 |
| B     | 5.12 | 5.09 | 3.91 | 4.27 |

It can be seen from this table that after the processing of this algorithm, the accuracy of speech recognition algorithm can be improved under the hypothetical model, which shows that this algorithm has a good solution to the problem of multiple speakers in speech recognition.

5. Conclusion

In this article, a PL2 speech recognition algorithm based on microphone array is proposed to explore the problem of multiple speakers in speech recognition. By estimating the direction of the wave and the separation of speech, the filtering of the voice signal of the non-expected speaker and the reservation of the desired speaker’s voice signal are realized through the LCMV filter. Experimental results show that the algorithm improves the accuracy of speech recognition when multiple speakers are present. However, the algorithm is complicated by introducing many algorithms of microphone array. And it has not solved the problem that the expected speaker and the unwanted speaker will bring to the speech recognition accuracy at the same angle. This is also the direction that the author needs to work hard.

References

[1] KE Dengfeng, XU Bo. The Basic Question of Speech Recognition in the Internet Age [J]. China Science, 2013,43(12):1578-1597.

[2] ZHAO Xiaoyan, ZHOU Lin, WU Zhenyang. Microphone Array Acoustic Source Localization Algorithm based on Compressed Censing [J]. Journal of Southeast University, 2015,45(2):203-207.

[3] Chiang Kai Shek. Research on Noise Reduction Method of Microphone Array Speech Signal Based on Tensor Analysis [D]. Beijing: Beijing Institute of Technology, 2016.

[4] YU Linlin. Speech Recognition Technology and Application Overview [J]. Modern Electronic Technology, 2013,36 (13):43-45.

[5] LU Zhao. Speech Recognition Algorithm in Noisy Environment [D]. An Hui: Anhui University, 2011.

[6] LIU Chengcheng, ZHAO Yongjun, HU Dexiu. High Resolution DOA Estimation Algorithm Based On Modified Propagator Algorithm [J]. Computer Application, 2010,30(5):1418-1424.

[7] WANG Wei, CAO Ning, LIU Weiwei. MIMO radar direction of arrival estimation based on PM and Root-MUSIC [J]. Science, Technology and Engineering, 2012,12(3):539-542.

[8] PANG Yu. Research and implementation of voice separation technology [D]. HuNan: National Defense University of science and technology, 2016.

[9] HE Meixin, SHU Kun. Application of Improved LCMV Algorithm in Multi Beam-forming [J]. Modern Electronic Technology, 2014,37(17):23-25.

[10] LI Yaxiong, ZHANG Jianqiang, PAN Deng. Speech Recognition Research Based on RNN - RBM Language Model [J]. Computer Research and Development, 2014,51(9):1936-1944.