Extraction of a target speech signal from 3-channel signal mixtures

Jie Meng (7874703). Electrical and Computer Engineering,University of Ottawa
Xingyu Xiong (300033654). Electrical and Computer Engineering,University of Ottawa

***Abstract*—This project's objective is the extraction of a target speech signal from 3-channel signal mixtures simulating a linear microphone array in an acoustic environment.In order to make our derivation more concrete and detailed, we first introduce the concept of Time Differences Of Arrival (TDOA),and then analyze its working principle and algorithm implementation, mainly to explore the simple implementation of GCC-PHAT. Secondly, we introduce Beamforming and also introduces its principles and examples in detail. Then, we will introduce the basic concept of Minimum Variance Distortionless Response beamformer (MVDR).Afterwards, we will introduce the method of removing background noise during the post-processing stage, that is, the single channel speech enhancement de-noising stage. Finally, we will implement one by one the content involved in this project and give mathematical description of the proposed solution and simulation of our proposed methods**

***Index Terms*—TDOA, GCC-PHAT, MVDR, Post-Processing**

I. Introduction And Overview

The whole team project will be conducted in following 3 steps in order to extract the target speech component (English-speaking female speech).The first step is to determine the source of the signal,namely,estimation of the directions of arrivals/angles of arrivals (DOA/AOA) of the sources or the angles θ_1,θ_2,θ_3.Equivalently, the "time differences of arrival" (TDOA) of the sources can be estimated.

After determining the angles of three difference sources and we need to remove the two interferences(German-speaking female speech and English-speaking male speech),that is,to extract the target speech from the noisy environment.Beamforming filtering may be a practical way to implement.In this part,we use a Minimum Variance Distortionless Response beamformer(MVDR) to solve this problem,which is effective and efficient to produce a single-channel output where the other two directional interferences have been removed.

The single-channel output of the beamforming stage should have both the target English-speaking female signal and some strong background noise component (possibly even stronger than in the original signal). Since the background noise is stationary and since its statistics in the time-frequency domain differ from the statistics of the target speech signal, it is possible to use single channel time-frequency filtering methods to attenuate the noise (as opposed to spatial domain filtering like the beamforming of the previous step). This step of filtering is often called "post-filtering", and it can be achieved using single channel speech enhancement de-noising methods.

II. Estimation of Time Difference Of Arriveal

The TDOA principle is a signal propagation delay based method. The idea for TDOA is to measure signal propagation delay differences.Let us assume that two signals are transmitted from two sources i and j synchronously at time instance T_0. At the receiver we receive the signal from the sources i and j at time instances T_i and T_j respectively. The corresponding propagation distance difference[1]

$$\begin{aligned}\Delta d_{i,j} &= d_i - d_j \\ &= c(T_i - T_0) - c(T_j - T_0) \\ &= c \cdot \Delta T_{i,j}\end{aligned} \tag{1}$$

is calculated from the propagation delays. It is obvious that the propagation distance difference depends on the signal propagation delay difference $\Delta T_{i,j}$. Due to building differences, we get rid of the unknown time T_0 of signal transmission and we do not face problems due to different time scales at the sources and the receivers. Both time instances T_i and T_j that are used for calculating the signal propagation delay difference are measured at the receivers and thus result from the same time base.

In contrast to TOA,where a propagation delay defines a circle around a source,a signal propagation delay difference measurement of TDOA characterizes points of equal distance differences to the considered sources. This is the definition of a hyperbola in the two-dimensional case or a hyperboloid for the three-dimensional space. Hence, TDOA is often called hyperbolic positioning.

There are three general approaches of TDOA estimation.In this project,we focus on angular spectrum,which is a popular and effective method.In

this method,we recommend to use the generalized cross-correlation with phase transform(GCC-PHAT) and we will give a detailed explanation in the following.

A. Generalized cross-correlation

Refer to previous works[2], Ideally, when there is no reverberation in the environment,the signal received by the i_{th} microphone is

$$x_i(t) = \alpha_i s(t - \tau_i) + n_i(t) \tag{2}$$

where $s(t)$ is source signal and τ_i is the delay from the i_{th} microphone.α_i is the attenuation of the sound source to the i_{th} microphone and $n_i(t)$ is additive noise.$s(t-\tau_i)$ and $n_i(t)$ are both real, jointly stationary random processes. Signal $s(t-\tau_i)$ is assumed to be uncorrelated with noise $n_i(t)$.

If there are 2 microphones 1, 2 in a microphone array, the formula can be expressed like this

$$x_1(t) = \alpha_1 s(t - \tau_1) + n_1(t) \tag{3}$$

$$x_2(t) = \alpha_2 s(t - \tau_2) + n_2(t) \tag{4}$$

Assume that the time delay between the sound source and the two microphones is denoted by τ_{12}, that is, $\tau_{12} = \tau_1 - \tau_2$ .Because $n_1(t)$, $n_2(t)$ and $s_(t)$ are uncorrelated.If we take a different τ to examine the similarity between $x_1(t)$ and $x_2(t-\tau)$, we can analyze the relationship between $x_1(t)$ and $x_2(t)$ about $s(t)$.

The two speech signals received by the microphone can be represented by the multiplication of the auto-correlation function and attenuation of the sound source signal.

$$R_{x_1x_2}(\tau) = \alpha_1\alpha_2 R_{ss}(\tau - \tau_1 + \tau_2) \tag{5}$$

when $\tau - \tau_1 + \tau_2 = 0$, namely, $\tau = \tau_1 - \tau_2$,$R_{ss}(\tau - \tau_1 + \tau_2)$ has the maximum value.At the same time, $R_{x_1x_2}(\tau)$ also reaches the peaks.τ is the time delay.

The complexity of directly performing cross-correlation operations in time domain detection is high, so we can transform it into frequency domain by using FFT. The cross-power spectral density function of the signal is

$$G_{x_1x_2}(\omega) = \int_{-\infty}^{+\infty} R_{x_1x_2}(\tau) e^{-j2\pi\omega\tau} d\tau \tag{6}$$

which can be used by IFFT and then we can get the the cross-correlation operations in time domain

$$R_{x_1x_2}(\tau) = \int_{-\infty}^{+\infty} G_{x_1x_2}(\omega) e^{j2\pi\omega\tau} d\omega \tag{7}$$

B. Phase transform

Since the noise of the recording environment and the short-time stability of the speech signal all make the maximum value of the cross-correlation function insignificant and the accuracy of the time delay estimation will also reduced. Therefore, the PHAT weight function[3] is used to improve the cross-correlation power spectrum, its expression is

$$\psi_{x_1x_2}(w) = \frac{1}{|\Phi_{x_1x_2}(w)|} \tag{8}$$

and the $\psi^{\star}_{x_1x_2}(w)\Phi_{x_1x_2}(w)$ is

$$\begin{aligned} \psi^{\star}_{x_1x_2}(w)\Phi_{x_1x_2}(w) &= \frac{\Phi_{x_1x_2}(w)}{|\Phi_{x_1x_2}(w)|} \\ &= \frac{|\Phi_{x_1x_2}(w)|e^{-j\phi(w)}}{|\Phi_{x_1x_2}(w)|} \\ &= e^{-j\phi(w)} \end{aligned} \tag{9}$$

It can be seen that the improved cross-correlation power spectrum is a pure phase function. Its amplitude is 1, so the PHAT method uses the phase information to find the delay. One of its great advantages is that the form is completely simplified and the computational complexity is low.

In this project, the sampling rate of given mixture1.wav, mixture2.wav and mixture3.wav is 16KHz while the total duration is 12s for each.Using mixture1 and mixture3 to do correlation and then we have 160000 $\times$ 12 samples, which equals to 192000 samples as we can see in Figure 1: Total Sampling Points.

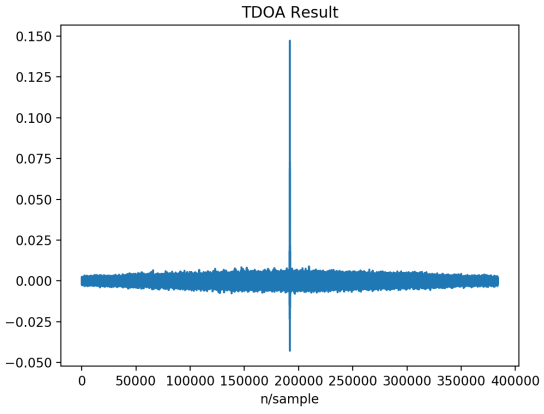

Figure 1: Total Sampling Points

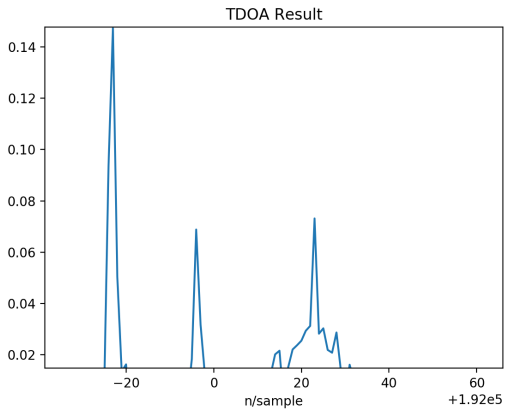

Figure 2: TDOA PEAKS

$\tau_{max} = \frac{d}{c}$, where d is the distance between microphone1 and microphone3 and c is the speed of sound of 343 m/s.Thus we can compute $\tau_{max} \approx 1ms$, which means the TDOAs of these 3 speeches are in 50 sampling points.Throughout the experiment, we calculate the sampling points of τ_1, τ_2, τ_3 which located at -23, -4, 23 as we shown in Figure 2: TDOA PEAKS and then we divide it by 16000hz and receive $\tau_1 = -0.0014375s, \tau_2 = 0.00025s, \tau_3 = 0.0014375s$.

III. Beamformer part

It is convenient for us to use a minimum variance distortionless response (MVDR) to deal with the directional interferences in this project.MVDR is an adaptive beamforming algorithm based on the maximum signal to interference plus noise ratio (SINR) criterion. The MVDR algorithm can adaptively minimize the power of the array output in the desired direction and maximize the SINR, which can greatly improve the noise suppression performance.

As a real environment, the signal model can be written as

$$\begin{aligned} y_1(t) &= x_1(t) + v_1(t) \\ y_2(t) &= x_2(t) + v_2(t) \\ y_3(t) &= x_3(t) + v_3(t) \end{aligned} \tag{10}$$

where $y_m(t)$,$x_m(t)$,$v_m(t)$,m=(1,2,3)are the noisy, clean speech and noise signals.In this project, we are going to process it in frequency domain because the signals are fixed and we can use constant beamformer coefficients so that we won't to process the signals in STFT frame by frame.

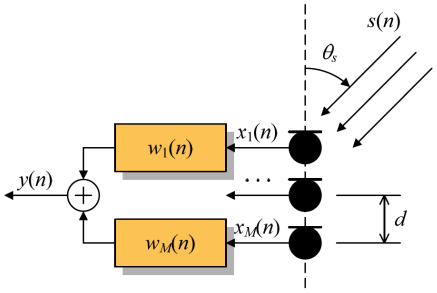

Figure 3: Microphone array system,which M is the number of microphones while in this project,M=3,d is the microphone spacing and $\frac{\pi}{2} - \theta_s$ is the incidence angle of the desired source.

In Figure 3, we can see the model briefly. To make the processing efficient, we work in the frequency domain. In this domain, the signal model is written as[4]

$$\begin{aligned} Y_1(w) &= X_1(t) + V_1(w) \\ Y_2(w) &= X_2(t) + V_2(w) \\ &= e^{-jwt_2 cos(\frac{\pi}{2} - \theta_2)} X(w) + V_2(w) \\ &= e^{-jwt_2 sin\theta_2} X(w) + V_2(w) \\ Y_3(w) &= X_3(t) + V_3(w) \\ &= e^{-2jwt_3 cos(\frac{\pi}{2} - \theta_3)} X(w) + V_3(w) \\ &= e^{-2jwt_3 sin\theta_3} X(w) + V_3(w) \end{aligned} \tag{11}$$

where $Y_m(w), X_m(w), V_m(w)$ and $X(w)$ are the Fourier transform of $y_m(t), x_m(t), v_m(t)$ and $x(t)$ m=(1,2,3).

In practice, our goal is to compute τ_m, and τ_m equals to $t_m cos\theta_m$ so we don't need to calculate θ directly.

Then we get the steering vectors for the source from angle θ, and we can see that it also includes the information of the TDOA.Due to the measurement from microphone1 and microphone3, we need divide it by 2.

$$\begin{aligned} d_{\theta_1}(w) &= [1, e^{-jw\frac{t_1}{2}}, e^{-jwt_1}]^T \\ d_{\theta_2}(w) &= [1, e^{-jw\frac{t_2}{2}}, e^{-jwt_2}]^T \\ d_{\theta_3}(w) &= [1, e^{-jw\frac{t_3}{2}}, e^{-jwt_3}]^T \end{aligned} \tag{12}$$

where $w = \frac{2\pi}{N} \cdot k$

With the frequency-domain signal model given above, the objective of beamforming is to recover the clean speech signal,$X(w)$, given the observation signals,$Y_m(w), m = 1, 2, 3$. This can be achieved by applying a complex weight to $Y_m(w)$ and then summing all the 3 weighted signals together.

$$\begin{aligned} Z(w) &= \sum_{m=1}^{3} Y_m(w) H_m^{\star}(w) \\ &= h^H(w) y(w) \\ h(w) &= [H_1(w), H_2(w), H_3(w)]^T \end{aligned} \tag{13}$$

where $Z(w)$ is an estimate of $X(w)$ and $h(w)$ is the beamforming filter.

we consider the case where there are 2 point noise sources in addition to the spatially white noise.Assuming that the incidence angle of the point noise source is θ_n, the corresponding pseudo-coherence matrix can be written as

$$\Gamma_{psn1}(w) = d_{\theta_{n1}}(w) d_{\theta_{n1}}^H(w) \tag{14}$$

$$\Gamma_{psn2}(w) = d_{\theta_{n2}}(w) d_{\theta_{n2}}^H(w) \tag{15}$$

where d_{θ_n} is the steering vector of the point noise source and ^{H} operator is Hermitian operator, which is both a transpose operator and a complex conjugate operator.Then, the pseudo-coherence matrix of the point-source-plus-white noise is

$$\begin{aligned} \Gamma_{pswn}(w) = &(1 - \alpha_{psn}) I_M \\ &+ \alpha_{psn} [\Gamma_{psn1}(w) + \Gamma_{psn2}(w)] \end{aligned} \tag{16}$$

where α_{psn} is a parameter that controls the level of the point source noise relative to that of the spatially white noise. As we compute them all, we can calculate the $h_{\theta_{d(w)}}$, the MVDR beamformer.

$$h_{\theta_{d(w)}} = \frac{\Gamma_v^{-1}(w) d_{\theta_d}(w)}{d_{\theta_d}^H(w) \Gamma_v^{-1}(w) d_{\theta_d}(w)} \tag{17}$$

IV. Post-Processing

Let $x[t]$ and $d[t]$ denote the speech and the noise processes,respectively.The observed signal $y[t]$ is given by

$$y[t] = x[t] + d[t] \tag{18}$$

In order to avoid the cut-off effect caused by frame division, a Hamming window w(n) is added before the Short Time Fourier Transform (STFT) to perform frame division.In other words, use a window function w(n) which is multiplied by s(n) to form a windowed speech function $s_w(n)$.By the way,the frame length is N and the frame shift is M (M=N/ 2).Perform STFT transformation of the above equation yields

$$X[k,l] = S[k,l] + D[k,l] \tag{19}$$

X[k,l] and S[k,l] respectively indicate

$$\begin{aligned} X[k,l] &= R[k,l] e^{j\theta_x[k,l]} \\ S[k,l] &= A[k,l] e^{j\theta_s[k,l]} \end{aligned} \tag{20}$$

where X[k,l], S[k,l] and D[k,l] are the noise speech signal, the speech signal and the k spectral components of the l-th frame of the noisy speech signal;R[k,l] and A[k,l] are the amplitudes of X[k,l] and S[k,l] respectively; θ_x and θ_y are the phase of X[k,l] and S[k,l] respectively.

Since the human ear is insensitive to the phase of the speech, only the logarithm of the spectral amplitude is considered. The short-term spectrum of the noisy speech signal can be obtained through the STFT. After the extracted phase is stored, using MMSE on the short-term logarithmic spectrum of the noisy speech. The enhanced speech uses the phase of the noisy speech signal to modify the spectrum amplitude, and the processed speech can be reconstructed by the estimated amplitude spectrum and phase.

The estimation criteria is to calculate the minimum value of

$$A = E[[[logA[k,l] - log\hat{A}[k,l]]^2] \tag{21}$$

where $\hat{A}_k$ is the estimation value of speech spectrum A_k.

Then we will focus on the real-value positive gain function $G(k,l)$. Assuming two function expressions are as follows

$$\begin{aligned} H_0(k,l) &: X(k,l) = D(k,l) \\ H_1(k,l) &: X(k,l) = S(k,l) + D(k,l) \end{aligned} \tag{22}$$

where $H_0(k,l)$ represents the non-speech signal, which is corresponded to the first two seconds in the beginning in this project and $H_1(k,l)$ represents the speech signal.So the Probability density function(PDF) is defined as followed

$$P(X(k,l)|H_0(k,l)) = \frac{e^{-\frac{|X(k,l)|^2}{\lambda_d(k,l)}}}{\pi \lambda_d(k,l)} \tag{23}$$

$$P(X(k,l)|H_1(k,l)) = \frac{e^{-\frac{|X(k,l)|^2}{(\lambda_s(k,l)+\lambda_d(k,l))}}}{\pi(\lambda_s(k,l) + \lambda_d(k,l))} \tag{24}$$

where $\lambda_s(k,l) = E[S(k,l)|^2 H_1(k,l)|]$ and $\lambda_d(k,l) = E[|D(k,l)|^2]$ respectively indicate the variance of voice and noise spectrum.Based on binary assumption model and probability density function,we can achieve $\hat{A}_k$[5]

$$\begin{aligned} \hat{A}_k(k,l) =& e^{E[lnA(k,l)|X(k,l)|]} \\ =& e^{E[lnA(k,l)|X(k,l),H_1(k,l)]p(k,l)} \\ &+ e^{E[lnA(k,l)|X(k,l),H_0(k,l)](1-p(k,l))} \\ =& G_{H_1}(k,l)^{p(k,l)} \cdot G_{min}^{1-p(k,l)} \cdot |X(k,l)| \end{aligned} \tag{25}$$

where $G_{H_1}(k,l)$ is the gain of the speech signal,which is

$$G_{H_1}(k,l) = \frac{\xi(k,l)}{1+\xi(k,l)} e^{(\frac{1}{2}\int_{v(k,l)}^{\infty} \frac{e^{-t}}{t} dt)} \quad (26)$$

where $G_{min}$is the gain of the non-speech signal,it is the subjective variable.$p(k,l)$ is the speech probability of speech estimation.As shown below

$$p(k,l) = [1 + \frac{q(k,l)}{1-q(k,l)} \cdot (1+\xi(k,l) \cdot e^{-v(k,l)})]^{-1} \quad (27)$$

where $q(k,l)$is the prior probability of non-speech signal,and $v(k,l)$ is below

$$v(k,l) = \frac{\xi(k,l)}{1+\xi(k,l)} \cdot \gamma(k,l) \quad (28)$$

$\xi(k,l)$ and $\gamma(k,l)$ are the prior SNR and posterior SNR,which is mentioned in the hints.

$$\begin{aligned} \xi(k,l) &= \frac{\lambda_s(k,l)}{\lambda_d(k,l)} \\ \gamma(k,l) &= \frac{[X(k,l)]^2}{\lambda_d(k,l)} \end{aligned} \quad (29)$$

According to the equation we our previously derived formula, the estimated speech signal can be represented by the gain function

$$\hat{S}(k,l) = G(k,l) \cdot X(k,l) \quad (30)$$

where the gain function can be expressed by

$$G(k,l) = [G_{H_1}(k,l)]^{p(k,l)} \cdot G_{min}^{1-p(k,l)} \quad (31)$$

By using IFFT,we can get the enhanced signal in time domain $\hat{s(n)}$,which is

$$\hat{s}(n) = \sum_{l} \sum_{k=0}^{n-1} \hat{S}(k,l) \cdot w(n-lM) e^{j(\frac{2\pi}{N})k(n-lM)} \quad (32)$$

V. Implementation

Python implementation procedure is listed as below.

We use the whole 12s data to debug the code. The code can split into 3 files according to the 3 steps in the project requirement.

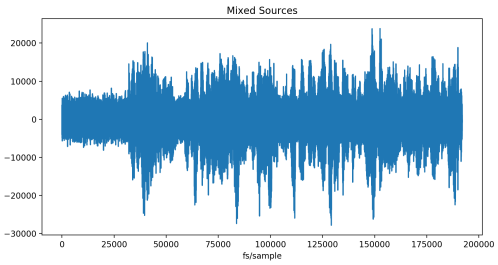

Figure 4: Mixed Sources

Figure 4 shows the mixed sources from mixture1.wav, which indicates the the 3-channel mixture speech in time domain and we can see that the first part of this speech is pure noise which is useful to do post-processing by modelling it.

1. TDOA :

tdoa.py

We can use two method to implement this part. One is *gcc_phat*; the other just use $numpy.correlationfunction$. Both can give the same result. From the correlation result of mixture1 and mixture3, we can find peeks in 50 samples, we got 3 τ values, -23, -4, 23 respectively, which are used in the Step2. You can see the result in Figure 2.

Pseudo-code:

Step 1: TDOA
$correlated\_data = np.correlation(m1, m3)$
or
$correlated\_data = GCC - PHAT(m1, m3)$
$[\tau_1, \tau_2, \tau_3] = find\_peaks(correlated\_data)$

---

2. Beamformer:

mvdr_beamfromer.py

The main difficult part in this file is to calculate the matrix of the filter H.

The function $steeringV(axisShape, \tau, M = 3)$ just takes the output shape and τ value as input, and return the steering vector. Notice in this function, we only need to calculate half N-point FFT length Steering vector.

The function $gammaInv(D\_n1, D\_n2, M = 3, \alpha_{psn} = 0.95)$ takes the 2 noise steering vectors and the α as input, and using Eq.(14) (15) (16) for computing the pseudo-coherence matrix $\Gamma(w)$; α value should be in 0 to 1, we tried 0.9, 0.95, 0.99 etc., but the difference is tiny.

Then we call the function $mvdr(D\_s, D\_n1, D\_n2)$, which takes the source steering vector and noises steering vectors as input. By using the return value of function 'gammaInv' and Eq.(17), we can get the filter system function H matrix.

We compute n-point FFT of the input 3 sound files as Y, when we got H matrix, we can get Z by Eq.(13). Since here we only calculated $n/2$

points steering vector, the Z also is an $N/2$ points frequency vector. For getting N-points Z result, we should concatenate the Z and the conjugate of inverse order Z as the Z_output. Then do the n-point IFFT of Z_output, we can get the time domain output z, the sound file which filtered the other 2 sources.

In this section, we use 20 percent overlap, which means when do the concatenating, we do overlap add. In this way, we can avoid the silent edge effect.

By treated the 3 τ as source τ respectively, we can find the τ_2 is the English Female voice. And the output wav file is "$mvdr\_output.wav$".

Pseudo-code:

Step 2:
MVDR

$$n = sample = 192000$$
$$Y[w] = n - point.fft \begin{pmatrix} y\_m1 \\ y\_m2 \\ y\_m3 \end{pmatrix}$$
$$w = \frac{2\pi}{N} \cdot K, K \in [0, \frac{n}{2}]$$
$$N = \frac{n}{2}$$
$$D_s = [1, e^{-j\frac{2\pi}{N}K\frac{\tau_2}{2}}, e^{-j\frac{2\pi}{N}K\tau_2}]^T$$
$$D_{n1} = [1, e^{-j\frac{2\pi}{N}K\frac{\tau_1}{2}}, e^{-j\frac{2\pi}{N}K\tau_1}]^T$$
$$D_{n2} = [1, e^{-j\frac{2\pi}{N}K\frac{\tau_3}{2}}, e^{-j\frac{2\pi}{N}K\tau_3}]^T$$
$$\Gamma_{n1}(K) = D_{n1} \cdot D_{n1}^H$$
$$\Gamma_{n2}(K) = D_{n2} \cdot D_{n2}^H$$
$$\Gamma(K) = (1-\alpha)I_M + \alpha(\Gamma_{n1} + \Gamma_{n2})$$
$$\Gamma_v^{-1}(K) = inverse(\Gamma(K))$$
$$H = \frac{\Gamma_v^{-1}(w)d_{\theta_d}(w)}{d_{\theta_d}^H(w)\Gamma_v^{-1}(w)d_{\theta_d}(w)}$$
$$Z(K) = \sum_{i=0}^{2} Y_m(w)H_m^\star(w)$$
$$Z\_output = Z[K] + Z\_inverse[K] \cdot conj\ ()$$
$$overlap : 20\%$$
$$z = ifft(Z\_output)$$
$$(wav.file) = write(z.real)$$

3. Post-filter:

$logmmse.py$

Read the $mvdr\_output.wav$ and call the function $logmmse$ which refers to [6], we finally get the output file '$final\_output.wav$', which filtered the noise. In this part, we should try different parameter values, such as noise threshold and window size, etc. to find the best set for filtering the noise. Finally, we use 0.3 as noise threshold, 10 as noise frame length. Fig.5 shows the result of our code.

Pseudo-code:

Step 3: Post-filter

final_output = log-mmse[6](args)

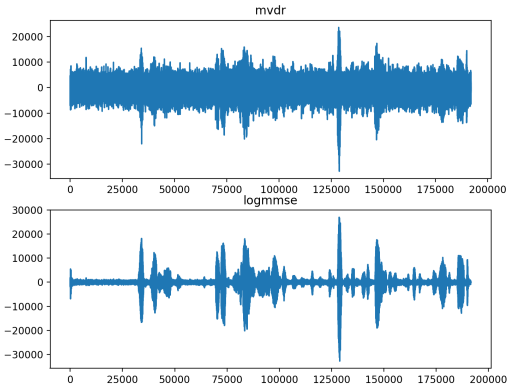

Figure 5: After MVDR and After log-MMSE

The first image in Figure 5 represents the output after MVDR processing, where only an English-speaking female's speech left.

The second image demonstrates the filtered speech has greatly reduced the noise, which we can easily compare with the first image.

Running rule:
1. run $tdoa.py$, you can get the figure of correlation between mixture1 and mixture3;
2. run $mvdr\_beamfromer.py$, output is the only one source wav file;
3. run $logmmse.py$, output is the wav file which the noise is filtered out. Also, you can get the comparing figure between input and output.

VI. Conclusion

GCC-PHAT is an effective and efficient way to deal with TDOA problems in this project. For beamformer part, MVDR is practical spacial filter to divide the desirable speech from mixtures while LOG-MMSE is a quite useful try-out to receive the expected output.

VII. Future work And Discussion

The English-speaking female's speech has successfully extracted from the 3-channel mixtures by using GCC-PHAT, MVDR, LOG-MMSE step by step where we put our full efforts on it.

The result seems good and the English-speaking female's speech is clear to identified. But if we divide the n points into smaller frames, and multiple kaiser window as paper[2] mentioned, use overlap add or STFT algorithm to do the step2, the performance will be better.

We do implement the split frames version, but the current result seems not good enough. We still need to modify the window parameter, frame size, α value, overlap percentage etc. to find a better result.

Also, we can use other algorithms for post-filter to filter noise better. These are the future work we can optimize.

References

[1] Stephan Sand, Armin Dammann and Christian Mensing, *"Positioning In Wireless Communications System"*, German Aerospace Center (DLR), Germany, pp.26-27,2014.

[2] Charles H.Knapp and G.Clifford Caeter, *"The Generalized Correlation Method for Estimation of Time Delay"*, IEEE Trans. Acoustics Speech and signal Process vol.24,no.4,pp.320–321,Apr.1976.

[3] XIA Yang and ZHANG Yuan-yuan, *"A rectangular microphone array based improved GCC-PHAT voice localization algorithm"* SHANGDONG SCIENCE,vol.24,no.6,pp.76–77,Dec.2011.

[4] Chao Pan, Jingdong Chen and Jacob Benesty, *"Performance Study of the MVDR Beamformer as a Function of the Source Incidence Angle"* IEEE Trans. Audio language Process vol.22,no.1,pp.67-79, 2014.

[5] YARIV EPHRAIM and DAVID MALAH, *"Speech Enhancement Using a- Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator"* IEEE Trans. Acoustics speech and signal Process vol.32,no.6,pp.1109-1121,Dec.1984.

[6] https://github.com/braindead/logmmse