RESEARCH

Multi-Microphones Reverberation Cancellation for Distant Speech Recognition

Xuewei Zhang

Correspondence:
zxw@cslt.riit.tsinghua.edu.cn
Center for Speech and Language Technology, Research Institute of Information Technology, Tsinghua University, ROOM 1-303, BLDG FIT, 100084 Beijing, China
Full list of author information is available at the end of the article

Abstract

In distant speech recognition(DSR), a target signal is disrupted by reverberation and noise effects so that the target signal is reduced and distorted. Recognising distant speech robustly, however, still remains a challenge. The contributions are that a global overview of multi-microphones reverberation cancellation for distant speech recognition is provided, the problem of errors in microphones array processing is analysed, and above all fundamental and recent reverberation cancellation techniques are illustrated, such as multi-input multi-output inverse filtering of room acoustics, spectral subtraction multi-channel linear prediction based on short time Fourier transform representation, the proposed linear-predictive multi-input equalization algorithm, the method using neural network front-ends and beamforming, along with theoretical analysis and experimental results verifying the effectiveness of the various algorithms.

Keywords: distant speech recognition; microphone arrays; reverberation cancellation; beamforming

1 Basic problems in microphone array processing

Several problems in microphone array processing appear as is shown in Fig. 1, the first problem is that the same type microphones have different amplitudes and phases, the second problem is that the beam is steered in wrong direction, the last problem is that the positions of the microphones are different from their original positions.

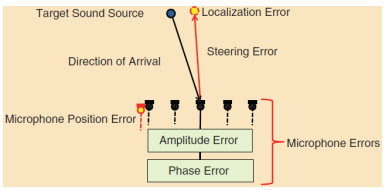

Figure 1 errors in microphones array processing [1]

2 Conventional Reverberation Cancellation Techniques

2.1 Multi-input multi-output inverse filtering of room acoustics

The diagram of the conventional inverse-filtering system based on the least squares error(LSE) is shown in Fig. 2, $g(k)$ is the impulse response of the system, $h(k)$ is

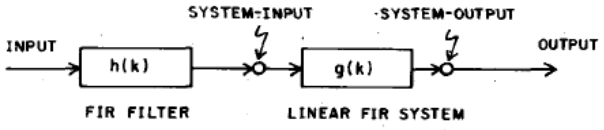

Figure 2 single-input single-output linear FIR system [2]

the coefficients of FIR filter,the relationship is as follows

$$d(k) = g(k) * h(k) \tag{1}$$

where

$$d(k) = \begin{cases} 1 & when \quad k = 0 \\ 0 & when \quad k = 1, 2, ... \end{cases}$$

The conventional inverse-filtering system based on LSE is only applied to single-input single-output system, however, multi-microphones reverberation is suitable for multi-input/output system, the n+1 input n-output system is shown in Fig. 3, S is sound source, M_1 and M_2 are microphones, $G_{i,j}(z^{-1})$ and $G_{i,j}(z^{-1})$ are transformation channels from sound source to microphones, $F_{i,j}(z^{-1})$ and $F_{i,j}(z^{-1})$ are FIR filters. Considering n=1, the equation is written as

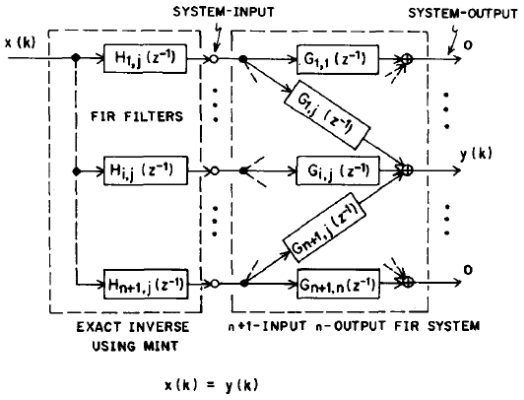

Figure 3 reverberation cancellation system using MINT [2]

$$d(k) = g_1(k) * h_1(k) + g_2(k) * h_2(k) \tag{2}$$

the z transformation is written as

$$D = G_1 H_1 + G_2 H_2 = \begin{bmatrix} G_1 & G_2 \end{bmatrix} \begin{bmatrix} H_1 \\ H_2 \end{bmatrix} \tag{3}$$

The error comparison of the LSE method and the proposed method is show in Fig. 4, from which the power spectral of errors of the LSE method the proposed

method performs better than that of desired response, the proposed inverse filtering performs best.

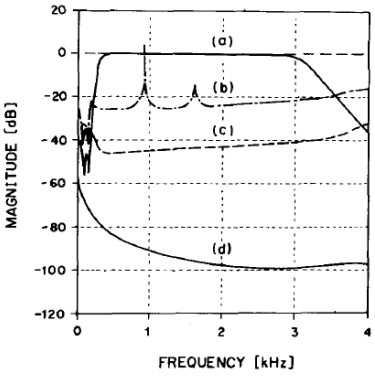

Figure 4 Power spectra of errors (a)desired response (b)and (c) error in the LSE method (d) error in the proposed method [2]

2.2 Spectral subtraction

The spectral subtraction algorithm is shown in Fig. 5. The reverberated signal x(n) is transformed by STFT, extracting phase and amplitude. By power spectral density(PSD), reverberation is estimated. The spectral amplitude of x(n) subtracts the estimated reverberation, receiving the spectral amplitude of dereverberated signal. The estimated s(n) is obtained from phase and the amplitude of dereverberation signal. The equation of spectral subtraction is expressed as

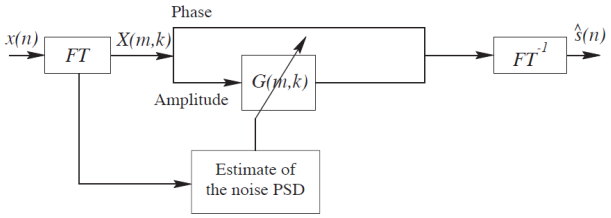

Figure 5 the spectral subtraction algorithm [3]

$$\begin{aligned} |\hat{S}(m,k)| &= |X(m,k)| - \hat{\gamma}^{1/2}(m,k) \\ &= G(m,k)|X(m,k)| \end{aligned} \tag{4}$$

where m is the time index, k is the frequency index. The estimation equation of reverberation PSD is expressed as

$$\hat{\gamma}_{rr}(m,k) = e^{-2\Delta T}\hat{\gamma}_{xx}(m-T,k) \tag{5}$$

In order to make the negative value of $|\hat{S}(m,k)|$ zero, set a threshold for $\hat{\gamma}_{rr}(m,k)$. The corresponding equation of $|\hat{S}(m,k)|$ is expressed as

$$|\hat{S}(m,k)| = \begin{cases} G(m,k)|X(m,k)| & when \geq \lambda\sqrt{\hat{\gamma}_{rr}(m,k)} \\ \lambda\sqrt{\hat{\gamma}_{rr}(m,k)} & otherwise \end{cases} \tag{6}$$

The algorithm of spectral subtraction reduces the frequency bands whose energy is larger than a threshold, and set other frequency band zero. The performance assessment methods have input to output SNR gain(G_{SNR}), noise reduction(NR), cepstral distance(CD) and speech recognition scores. Compared with Bloom algorithm, the spectral subtraction has more signal distortion, however, G_{SNR} and NR work better, speech recognition scores are increased by 13%.

2.3 Multi-channel linear prediction based on short time Fourier transform representation

In order to solve the problem of the large computing cost and the recognition accuracy based on multi-channel linear prediction(MCLP), the method with window effect reduction(MWER) and the method with window effect compensation(MWEC) are represented [4]. MCLP is expressed as

$$x_t^{(1)} = \sum_{l=1}^{2}\sum_{\tau=1}^{K} c_\tau^{(l)} x_{t-\tau}^{(l)} + s_t \tag{7}$$

where $c_t^{(l)}$ is MCLP coefficients,$x_t^{(1)}$ is the observed signal,s_t is source signal.Consider that $y_t^{(1)}$ is the convolving of $x_t^{(1)}$ and $c_t^{(l)}$, in STFT domain windowed function is expressed as $W_N(Y^L(Z))$. MWER is computed approximately in STFT domain, however, MWEC is computed precisely, compensating the error of MWER, MWEC uses the conjugate gradient method instead of the covariance matrix. The comparison with MWER and MWEC is shown in Fig. 6, the average cepstral distoritions of the two methods is smaller than baseline, in particular MWER performs better.

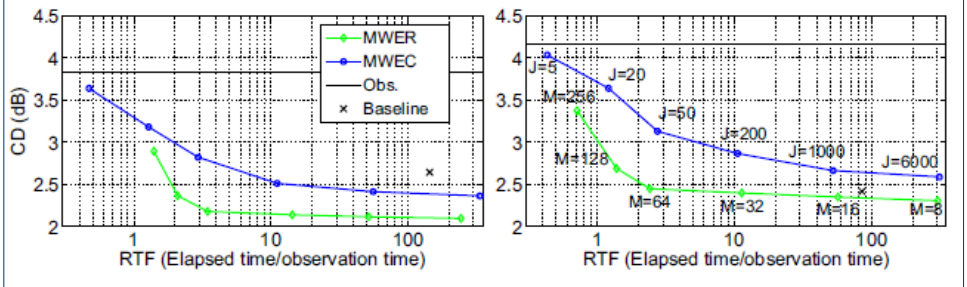

Figure 6 Average cepstral distortions (CD) of the signals dereverberated by MWER and MWEC, average CDs of the observed signals (Obs.) and average CDs of the signals dereverberated by the timedomain algorithm (Baseline) when using u1 (left panel) and u5 (right panel) and controlling the real time factors (RTF) [4].

2.4 Linear-predictive multi-input equalization algorithm

2.4.1 Adaptive LIME-ZF

In the conventional linear-predictive multi-input equalization algorithm(LIME), the estimated MLP filter and AR coefficients are expressed as

$$\hat{w} = \hat{Q}(1 : PL_w, 1), \tag{8}$$

$$\hat{a} = \lambda \hat{Q}. \tag{9}$$

The diagram of LIME algorithm is shown in Fig. 7. The estimated source signal is obtained by MLP filter and AR coefficients. For reducing computing cost, by the

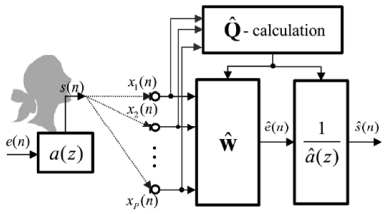

Figure 7 the diagram of LIME algorithm [5]

analysis Q matrix can be substituted with Q_w, Q_w satisfies the all characteristic of Q.

$$Q_w = Q(I \leftarrow \hat{I}) \tag{10}$$

For reducing high leakage level between Q and $\hat{Q}$, the zero-forcing(ZF) is applied. The experiment result is shown in Fig. 8. LIME is applied to real-time adaptive

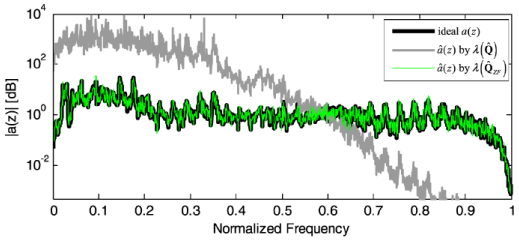

Figure 8 In ISM case, AR model using $\hat{Q}$ and improved accuracy by the zero-forcing modification [5]

equalization according to minimum mean squared error(MMSE) criterion, the equation is expressed as

$$e^{(i)}(n) = x_i(n) - w^{(i)T}x(n-1) \tag{11}$$

where i represents the i-th filter.The adaptive equalization method is shown in Fig. 9. The spectral comparison is shown in Fig. 10, the ZF-LIME cancels most of

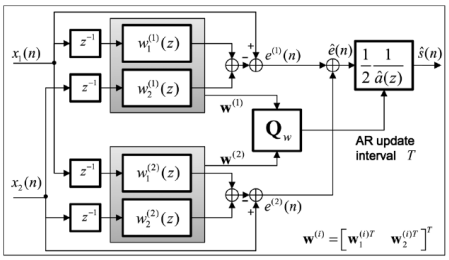

Figure 9 the diagram of real-time adaptive equalization method [5]

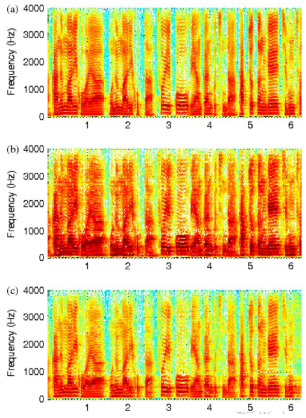

Figure 10 spectrogram(a) clean signal (b) observed signal (c) ZF-LIME output [5]

reveberation. Future work can both adopt the fast version of ADFs and improve the performance of the adaptive multichannel prediction filter in an environment of backnoise.

2.4.2 Combination with LIME and least squares filtering

From Fig. 11, in most segments log spectral distortion(LSD) and energy of the estimated segment are lower than source signal, segmental SNR of the estimated segment is always higher than source signal, and number of unstable poles of the estimated segment is shown. When estimated segment has unstable poles LIME does not perform well. In some segments LIME does not performance well, the least squares(LS) filters are applied to those segments. LIME algorithm is used to dereverberate all the segments, LS filters are constructed by using the segments which LIME dereverberates successfully, then the filters are used to derevereberate those segments which LIME fails to dereverberate successfully, the system diagram of combination with LIME and Least squares filtering is shown in Fig. 12. After the combination with LIME and Least squares filtering, the LSD of the estimated segment is always lower than source signal,and the segSNR of the estimated segment is always higher than source signal.

2.5 Using neural network front-ends

Speech enhancement with beamforming is popular, however, the DNN front-end has better results. The DNN front-end structure is shown in Fig. 13. The struc-

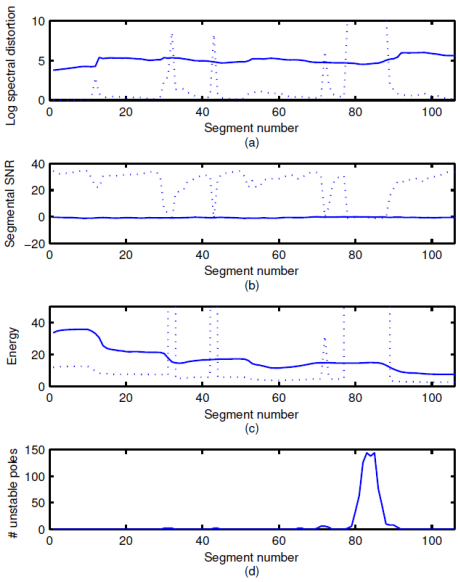

Figure 11 (a)the LSD comparison of source signal(dotted) and the LIME estimated segment (b)the segSNR comparison of source signal(dotted) and the LIME estimated segment (c)the energy comparison of source signal(dotted) and the LIME estimated segment (d)number of unstable poles of each segment's estimated AR process [6]

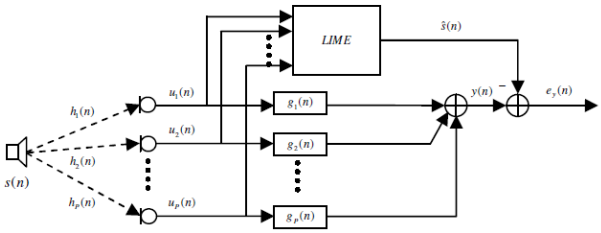

Figure 12 The system diagram of combination with LIME and Least squares filtering [6]

ture is composed of one input layer, one output layer, two hidden layers and one bottleneck layer, which realize direct channel concatenation. The error rates using PLP features and BN+PLP feature are shown in Fig. 14, the BN features perform significant improvement compared with PLP features. The error rate comparison of direct channel concatenation and beamforming is shown in Fig. 15. Direct channel concatenation works better than multi-channels beamforming.

2.6 Beamforming

Beamforming is a versatile approach to spatial filtering and the goal is to abstract source signal from signal with noise and reverberation. A diagram of a ditant speech recognition(DSR) system is shown in Fig. 16. Speaker tracker evaluates speaker's position, beamformer estimates source signal, the postfilter enhances beamformed signal, finally the filtered signal is put into speech recognizer. The DSR system can have several problems that speaker tracker causes wrong direction, microphones have different amplitudes and phases, and the position of microphones deviates from their original position. A signal f(t) from a plane wave arrives at different

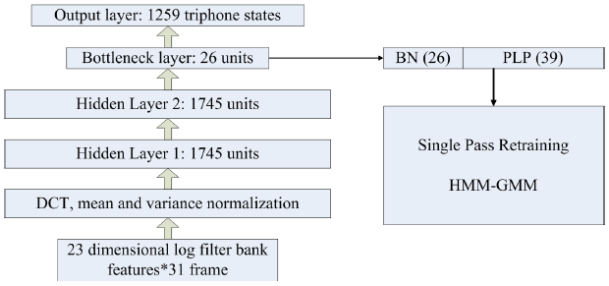

Figure 13 the DNN front-end structure [7]

Feature	Conf	#o	2cct	4cct	8cct
PLP	-	0	62.1	62.2	-
PLP	-	4	67.9	68.3	-
BN+PLP	*2TS*	0	46.8	46.5	47.4
BN+PLP	*2TS*	4	54.1	54.7	55.6

Figure 14 error rate using direct channel concatenation on PLP fetures and BN+PLP features with non-overlapping speech(cc is direct channel concatenation,#0 is overlapping,2TS is BN features from the DNN front-end structure) [7]

microphones at different time, the delayed signal is expressed as

$$\mathbf{f(t)} = [f(t-\tau_0)f(t-\tau_1)...f(t-\tau_{s-1})]^T \tag{12}$$

The corresponding vector is expressed as

$$\mathbf{F}(\omega) = F(\omega)v(k,\omega) \tag{13}$$

where $F(\omega)$ is the transform of f(t) and

$$\mathbf{v(k,}\omega\mathbf{)} \triangleq [e^{-i\omega\tau_0}e^{-i\omega\tau_1}...e^{-i\omega\tau_{s-1}}] \tag{14}$$

A BF is expressed as

$$Y(\omega) = \mathbf{w}^H(\omega)X(\omega) \tag{15}$$

where $w(\omega)$ is channel weight, $X(\omega)$ is input, $Y(\omega)$ is output. The difference of kinds of BP methods depend on $w(\omega)$.

	IHM	SDM	2bmit	4bmit	8bmit	2cct	4cct
M+	22.4	49.4	48.3	49.0	47.0	47.0	46.1
M-	24.6	47.4	45.3	45.0	43.2	45.5	45.6
H+	23.3	45.8	43.9	43.7	41.5	43.7	43.5
H-	25.7	51.3	49.5	49.5	48.4	49.5	49.3

Figure 15 error rate using BN + PLP (2TS) features (IHM is individual headset microphones, SDM is single distant microphone,body movement is M^+,no body movement is M^-,head movement is H^+,no head movement is H^-) [7]

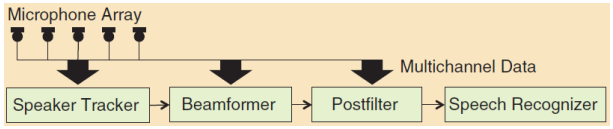

Figure 16 The diagram of a typical DSR system [1]

2.6.1 Delay-and-sum beamforming [1]

The delay-and-sum beamforming(DS BF) is expressed as

$$Y(\omega) = \mathbf{w}_{DS}^{H}(\omega)\mathbf{v}(k,\omega)F(\omega) = F(\omega) \tag{16}$$

That is expected output so that the equation is obtained

$$\mathbf{w}^{H}(\omega)\mathbf{v}(k,\omega) = 1 \tag{17}$$

The directivity of the linear DS BF at low frequencies is poor due to the fact that the wavelength is much longer than the aperture of the array. The beam pattern for very low frequencies is nearly flat, indicating that the directivity is effectively zero.

2.6.2 Minimum variance distortionless response beamforming [8]

In order to suppress noise and reverberation $N(\omega)$ by adjusting weights, minimizing the variance of the noise and reverberation are needed, the equation is expressed as

$$argmin_{\omega} w^{H}(\omega) \sum_{N}(\omega) w(\omega) \tag{18}$$

where $\sum_{N} \triangleq \varepsilon\{N(\omega)N^{H}(\omega)\}$ and $\varepsilon\{.\}$ is the expectation operator. The minimum variance distortionless response beamforming(MVDR BF) is expressed as

$$w_{MVDR}^{H}(\omega) = \frac{v^{H}(k,\omega)\sum_{N}^{-1}(\omega)}{v^{H}(k,\omega)\sum_{N}^{-1}(\omega)v(k,\omega)} \tag{19}$$

2.6.3 Super-directive beamforming [8]

The MVDR BF with the diffuse noise model is called the super-directive BF(SD BF). SD BF uses $\Gamma(\omega)$ to replace $\sum_{N}(\omega)$, $\Gamma(\omega)$ is expressed as

$$\Gamma_{m,n}(\omega) = sinc\frac{\omega d_{m,n}}{c} \tag{20}$$

where $d_{m,n}$ is the distance between the mth and nth elements of the array. The environment exists not only diffuse noise but also more sources of discrete interference, the equation is expressed as

$$\sum\nolimits_{N}(\omega) = \sum\nolimits_{N}(\omega)v(k_I)v^{H}(k_I) + \sigma_{SI}^{2}\Gamma(\omega) \tag{21}$$

where $\sigma_{SI}^{2}\Gamma(\omega)$ is the power spectral density of the diffuse noise.

2.6.4 Statistically optimum beamforming [9]

In statistically optimum beamforming, the weights are chosen based on different criteria which is shown in Fig. 17.

Type	MSC	Reference Signal	Max SNR	LCMV
Definitions	$\mathbf{x}_a$—auxiliary data y_m—primary data $\mathbf{r}_{ma}=E\{\mathbf{x}_a y_m^*\}$ $\mathbf{R}_a=E\{\mathbf{x}_a\mathbf{x}_a^H\}$ output: $y=y_m-\mathbf{w}_a^H\mathbf{x}_a$	$\mathbf{x}$—array data y_d—desired signal $\mathbf{r}_{xd}=E\{\mathbf{x}y_d^*\}$ $\mathbf{R}_x=E\{\mathbf{x}\mathbf{x}^H\}$ output: $y=\mathbf{w}^H\mathbf{x}$	$\mathbf{x}=\mathbf{s}+\mathbf{n}$—array data $\mathbf{s}$—signal component $\mathbf{n}$—noise component $\mathbf{R}_s=E\{\mathbf{s}\mathbf{s}^H\}$ $\mathbf{R}_n=E\{\mathbf{n}\mathbf{n}^H\}$ output: $y=\mathbf{w}^H\mathbf{x}$	$\mathbf{x}$—array data $\mathbf{C}$—constraint matrix $\mathbf{f}$—response vector $\mathbf{R}_x=E\{\mathbf{x}\mathbf{x}^H\}$ output: $y=\mathbf{w}^H\mathbf{x}$
Criterion	$\min_{\mathbf{w}_a} E\{\lvert y_m-\mathbf{w}_a^H\mathbf{x}_a\rvert^2\}$	$\min_{\mathbf{w}} E\{\lvert y-y_d\rvert^2\}$	$\max_{\mathbf{w}} \frac{\mathbf{w}^H\mathbf{R}_s\mathbf{w}}{\mathbf{w}^H\mathbf{R}_n\mathbf{w}}$	$\min_{\mathbf{w}}\{\mathbf{w}^H\mathbf{R}_x\mathbf{w}\}$ s.t. $\mathbf{C}^H\mathbf{w}=\mathbf{f}$
Optimum Weights	$\mathbf{w}_a=\mathbf{R}_a^{-1}\mathbf{r}_{ma}$	$\mathbf{w}=\mathbf{R}_x^{-1}\mathbf{r}_{xd}$	$\mathbf{R}_n^{-1}\mathbf{R}_s\mathbf{w}=\lambda_{max}\mathbf{w}$	$\mathbf{w}=\mathbf{R}_x^{-1}\mathbf{C}[\mathbf{C}^H\mathbf{R}_x^{-1}\mathbf{C}]^{-1}\mathbf{f}$
Advantages	Simple	Direction of desired signal can be unknown	True maximization of SNR	Flexible and general constraints
Disadvantages	Requires absence of desired signal from auxiliary channels for weight determination	Must generate reference signal	Must know $\mathbf{R}_s$ and $\mathbf{R}_n$, Solve generalized eigenproblem for weights	Computation of constrained weight vector
References	Applebaum [1976]	Widrow [1967]	Monzingo and Miller [1980]	Frost [1972]

Figure 17 Summary optimum beamforming [9]

A. Multiple sidelobe canceller The object of multiple sidelobe canceller(MSD) is to select channel weights to suppress main channel interference. However, the weights to minimize output power can cause cancellation of the desired signal, so when the optimum weights are relative small the MSD is effective. The equation can be expressed as

$$Y(t) = [w_q(t) - B(t)w_a(t)]^H X(t) \tag{22}$$

where w_q is the quiescent weight vector lying in the constraint space, B is the blocking matrix and where w_a is the active weight vector that is adapted during the execution of the adaptive beamforming algorithms.

B. Multiple sidelobe canceller In some cases, A signal which is close to the desired signal is obtained, and that is called a reference signal. The covariance of the reference signal estimates approximately the covariance of the desired signal.

C. Maximization of signal to noise ratio The weights are optimized by maximation signal to noise ratio. The desired signal, noise signal and their covariance need to be given.

D. Linearly constrained minimum variance beamforming and generalize sidelobe canceler In some cases, the above methods are not ideal, such as the unkown desired signal, noise signal and their covariance. The linear constrained minimum variance (LCMV) beamforming uses linear constrains which solves the above limits,

whose controls the response of beamforming and weights are selected to minimum output variance. LCMV has more general constrain equation, so the generalized sidelobe canceller(GSC) is proposed, the constrained problem is changed into unconstrained form and decomposition is obtained. The equation of GSC is expressed as

$$Y(k,m) = [w_q(k,m) - B(k,m)w_a(k,m)]^H X(k,m) \tag{23}$$

where w_q is the quiescent weight vector, B is the signal blocking matrix, w_a is the active weight vector.B is orthogonal to w_q, that is $B^H w_q = 0$. GSC structure is shown in Fig. 18.

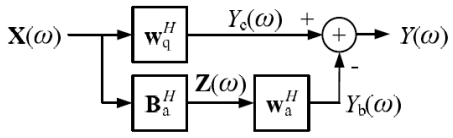

Figure 18 GSC structure [9]

E. Maximizing non-Gaussianity Speech signals are highly non-Gaussian, the desired signal with noise is closer to Gaissian than clean signal, reverberant speech is close to Gaissian than anechoic speech, so maximizing non-gaussianity optimization criteria is needed. The criteria of meature a degree of non-Gaussianis kurtosis and negentropy. We maximize kurtosis(MK) of the beamforming's output is to find $w_a(k,m)$, then use $w_a(k,m)$ to minimize the variance of the beamforming's output. The equation of the kurtosis meature is expressed as

$$kurt(Y) \triangleq \varepsilon\{|Y|^4\} - \beta_K(\varepsilon\{|Y|^2\})^2 \tag{24}$$

where β is typical set to 3.
MK BF requires large vector to compute. To improving efficiency and accuracy subspace filter is come up with. The subspace filter decomposes the output into directional signals and ambient noise, then the output of GSC beamforming subtract ambient noise, obtaining relative accurate estimation and less computing. The proposed MK beamformer is expressed as

$$Y(k) = [w_q(k) - B(k)U(k)w_a(k)]^H X(k) \tag{25}$$

where $U(k)$ is subspace filter. The diagram of maximum kurtosis beamformer with the subspace filter is shown in Fig. 19. Word error rate of every decoding pass is shown in Fig. 20 using single distant microphone(SDM), super-directive beamforming(SD BF), conventional maximum kurtosis beamforming(MK BF) and maximum kurtosis beamforming with the subspace filter(MK BF w SF). The proposed maximum kurtosis beamforming with the subspace filter is superior to MK BF.
The value of kurtosis can be influenced by a few samples with a low observation

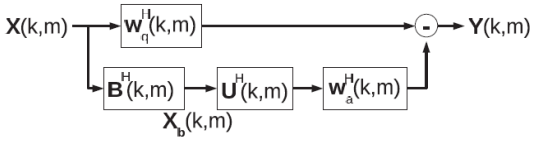

Figure 19 maximum kurtosis beamformer with the subspace filter [10, 11]

Algorithm	Pass 1 Exp.	Pass 1 Child	Pass 2 Exp.	Pass 2 Child	Pass 3 Exp.	Pass 3 Child
SDM	9.2%	31.0%	3.8%	17.8%	3.4%	14.2%
SD BF	5.4%	24.4%	2.5%	9.6%	2.2%	7.6%
MK BF	5.4%	25.1%	2.5%	9.0%	2.1%	6.5%
MK BF w SF	6.3%	25.4%	1.2%	7.4%	0.6%	5.3%
CTM	3.0%	12.5%	2.0%	5.7%	1.9%	4.2%

Figure 20 Word error rate of every decoding pass [10]

probability. However, negentropy is more robust than kurtosis, which is based on super-Gaussianity. Both different crieria for the detection of the subspace dimension such as AIC and MDL meatures and applying the online subplace learning algorithm will need to do in future work. The negentropy equation is expressed as

$$J_d(Y) = H_{gauss}(Y) - \beta H_{sg(Y)} \tag{26}$$

where $H_{gauss(Y)}$ is the entropy of the Gaussian PDF, $H_{sg(Y)}$ is the entropy of the super-Gaussian PDF and β is unity. For $H_{gauss(Y)}$ is small and influences other value, β is multiplied with $H_{sg(Y)}$, which is set to 0.5. The object of maximizing $J_d(Y)$ to find w_a, obtaining minimum variance of the beamforming's output. The generalized Gaussian probability density function(GG-PDF) is applied to maximum negentripy beamforming. The word error rate with different beamforming algorithms is shown in Fig. 21.

Beamforming Algorithm	Pass (%WER) 1	2	3	4
D&S BF	79.0	38.1	20.2	16.5
MVDR BF	78.6	35.4	18.8	14.8
SD BF	71.4	31.9	16.6	14.1
GEV BF	78.7	35.5	18.6	14.5
Conventional MN BF	75.1	32.7	16.5	13.2
SD-MN BF with GG-PDF	74.9	32.1	15.4	12.1
SD-MN BF with CGGD-PDF	75.3	30.9	15.5	12.2
SDM	87.0	57.1	32.8	28.0
CTM	52.9	21.5	9.8	6.7

Figure 21 The word error rate with different beamforming algorithms [12, 13]

2.6.5 Adaptive beamforming

In some cases, statistics are unkown, data is known over time, so adaptive beamforming is proposed. The structure of adaptive beamforming is shown in Fig. 22. The equation of adaptive beamforming is expressed as

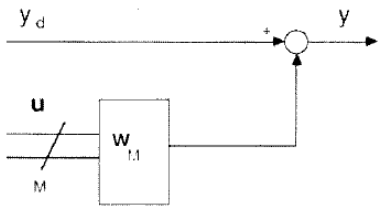

Figure 22 The structure of adaptive beamforming [9]

$$\begin{aligned} J(w_M) &= E|y_d - w_M^H u|^2 \\ &= \sigma_d^2 - w_M^H r_{ud} - r_{ud}^H + w_M^H R_u w_M \end{aligned} \tag{27}$$

where $\sigma_d^2 = E\{|y_d|^2\}$,$r_{ud} = E\{uy_d^*\}$ and $R_u = E\{uu^H\}$.
Two popular adaptive beamforming algorithms are leasr-t-mean-square(LMS) algorithm and exponentially weighted recursive least square(RLS) algorithm shown in Fig. 23. The equation of LMS is expressed as

Algorithm	*LMS*	*RLS*
Initialization	$\mathbf{w}_M(0) = 0$ $y(0) = y_d(0)$ $0. < \mu < \frac{1}{Trace\ [\mathbf{R}_u]}$	$\mathbf{w}_M(0) = 0$ $\mathbf{P}(0) = \delta^{-1}\mathbf{I}$ δ small, I identity matrix
Update Equations	$\mathbf{w}_M(k) = \mathbf{w}_M(k-1) + \mu\mathbf{u}(k-1)y^*(k-1)$ $y(k) = y_d(k) - \mathbf{w}_M^H(k)\mathbf{u}(k)$	$\mathbf{v}(k) = \mathbf{P}(k-1)\mathbf{u}(k)$ $\mathbf{k}(k) = \frac{\lambda^{-1}\mathbf{v}(k)}{1 + \lambda^{-1}\mathbf{u}^H(k)\mathbf{v}(k)}$ $\alpha(k) = y_d(k) - \mathbf{w}_M^H(k-1)\mathbf{u}(k)$ $\mathbf{w}_M(k) = \mathbf{w}_M(k-1) + \mathbf{k}(k)\alpha^*(k)$ $\mathbf{P}(k) = \lambda^{-1}\mathbf{P}(k-1) - \lambda^{-1}\mathbf{k}(k)\mathbf{v}^H(k)$
Multiplies per update	$2M$	$4M^2 + 4M + 2$
Performance Characteristics	Under certain conditions, convergence of $\mathbf{w}_M(k)$ to the statistically optimum weight vector $\mathbf{w}_{opt}$ in the mean-square sense is guaranteed if μ is chosen as indicated above. The convergence rate is governed by the eigenvalue spread of $\mathbf{R}_u$. For large eigenvalue spread, convergence can be very slow.	The $\mathbf{w}_M(k)$ represents the least-squares solution at each instant k and are optimum in a deterministic sense. Convergence to the statistically optimum weight vector $\mathbf{w}_{opt}$ is often faster than that obtained using the *LMS* algorithm because it is independent of the eigenvalue spread of $\mathbf{R}_u$.

Figure 23 Two popular adaptive beamforming algorithms

$$w_M(k+1) = w_M(k) + \mu(k)y^*(k) \tag{28}$$

where μ controls convergence. When too many eigenvalues cause slow convergence, exponentionally weighted RLS is proposed,the equation is expressed as

$$min_{w_M(k)} \sum_{k=0}^{K} \lambda^{K-k}|y_d(k) - w_M^H u(k)|^2 \tag{29}$$

where K is time step, λ is a positive constrain.
Multi channel speech dereverberation using LP residual cepstrum in an adaptive beamforming is proposed, the diagram is shown in Fig. 24. The reverberated signals

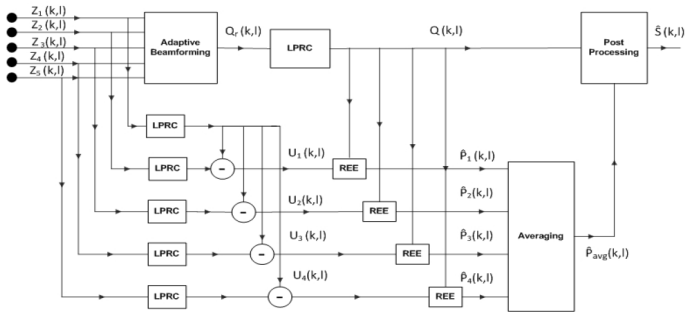

Figure 24 The diagram of multi channel speech dereverberation using LP residual cepstrum in an adaptive beamforming [14]

$Z_i(k,l)$ of frequency domain first carry out AD BF to remove early and late reverberation, obtaining the beamformer output $Q_r(k,l)$. $Z_i(k,l)$ and $Q_r(k,l)$ perform LP residual to cancel early reverberation. The power envelope(PE) is used to estimate remaining reverberation components in $Q_r(k,l)$. The postfilter is applied to canceling late reverberation [15]. The word error rates of the proposed method compared with excitation source information(ESI), close talking microphone(CTM), spatio temporal processing(STP), spectral enhancement(SE) and dual micro-phone(DM) are shown in Fig. 25. The proposed method is superior to ESI, STP, SE, DM. The proposed method with noise can be considered in next future. The future work

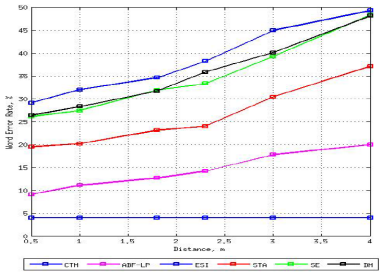

Figure 25 Comparison of the word error rate for various methods as a function of the distance between source and microphone [14]

needs to check the performance of the proposed method under noisy environment.

2.6.6 Generalized eigenvalue beamforming [16, 17]

The generalized eigenvalue beamforming(GEV BF) is obtained based on maximizing SNR, the maximum SNR of frequency domain is expressed as

$$F_{SNR}(k) := v_{max}(k) \tag{30}$$

where F(k) is filter coefficients,$v_{max}(k)$ is eigenvector. $F_{SNR}(k)$ is expressed as

$$F_{SNR}(k) = \zeta \Phi_{NN}^{-1}(k) H(l_t, k) \tag{31}$$

where $F_{SNR}(k)$ is the principal eigenvector of $\Phi_{NN}^{-1}(k)\Phi_{XX}(k)$, ζ is an arbitry complex scalar. The distortionless response (the target source signal) is expressed as

$$F_{GMVDR}^{H}(k)H(l_t,k) = 1 \tag{32}$$

where $H(l_t,k)$ is the transfer function from the source to microphones, which is known. The performance improvement is about noise reduction, the noise distortion needs to do in future.

2.6.7 Optical modal beamforming for spherical microphone arrays

The above multi-microphone arrays are planar, spherical microphone arrays are considered here. The standard Cartesian (x,y,z) is transformed into spherical $(r,/theta,/phi)$ coordinate. Beamforming design has many limited problems, so the optical modal beamforming is trade-off on these problems which influence the performance of beamforming. The optimization problem of beamforming is expressed as

$$\begin{aligned} \min_{w} w^{H}(k)R(\omega)w(k) \\ \text{subject to} B(ka,\Omega_0) = \frac{4\pi}{M} \\ |H(KA,\Omega)| \leq \varepsilon\frac{4\pi}{M}, \forall\Omega \in \Omega_{SL} \\ WNG(k) \geq \zeta(k) \end{aligned} \tag{33}$$

where Ω_{SL} is the sidelobe region, ε and ζ are parameters to control the sidelobes and the white noise gain(WNG). Special casess of the optimization are maximum out SINR, no WNG or sidelobe control, maximum directivity, no WNG or side-lobe control, maximum WNG,no directivity or side-lobe control and maximum output SINR with WNG control, no sidelobe control. The power estimation of delay-and-sum(DAS), a spherical-harmonics domain maximum directivity index(HMDI), and a spherical-harmonics domain white noise gain constrained(HWNC) methods is show in Fig. 26. The result proves the HWNC method performs higher resolution and better anti-interference. The accuracy of spatial samping and aliasing are

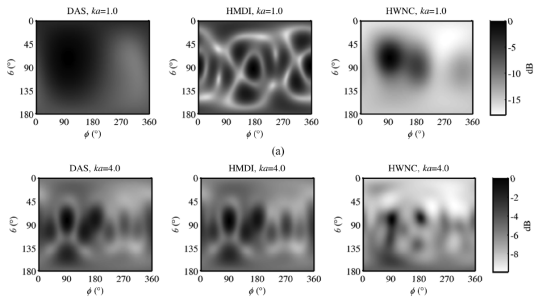

Figure 26 The power estimation of DAS, HMDI and HWNC methods [18]

neglected, those will be analysed in future.

2.6.8 Cluster blind beamforming [19]

In the case of unknown microphones position, beamforming is used to dereverberate blindly. Micrphones are grouped according to noise coherence. Microphone clustering is implemented base on TDOA. Then closet cluster(CC) beamfoming and weighted cluster combination(WCC) beamforming are proposed. The word error rate comparison of clusted array and a full array is shown in Fig. 27.Clustering performs better than no-clustering. The method of CC and WCC are superior to signal-channel input and the full microphone arrays by meaturing SNR and perceptual evaluation of speech quality(PESQ).

Other meatures for ranking clusters and automatically determining cluster weights

Speaker Position	Technique	Closest clust.	2^{nd} closest clust.	All microphones
S1	Single Ch.	43.2	64.0	-
	Delay-sum	35.7	56.5	44.4
	MVDR	36.0	59.8	48.1
	Superdirective	34.8	61.4	55.5
S2	Single Ch.	45.7	69.8	-
	Delay-sum	36.8	67.8	49.6
	MVDR	36.2	76.4	54.8
	Superdirective	35.5	83.3	64.3
S3	Single Ch.	53.4	62.6	-
	Delay-sum	46.3	56.5	49.9
	MVDR	45.8	64.2	54.4
	Superdirective	46.0	67.4	61.2

Figure 27 The word error rate comparison of clustered array and a full array [19]

for combination will be investigated in future.

2.6.9 Feature mapping of multiple beamformed sources [20, 21, 22]

The diagram of the feature mapping based on speech recognition is shown in Fig. 28. A nonlinear mapping of features from the target and interfering distant sound sources to the clean target features is proposed. Two or three beamformers are directed at the target and interfering speakers, and a frequency domain binary mask post-filter is followed for obtaining the target and interfering speech more accurately. We can demonstrate that better quality of the estimated target and interfering speech sa the inputs are helpful when using our non-linear feature mapping approach.

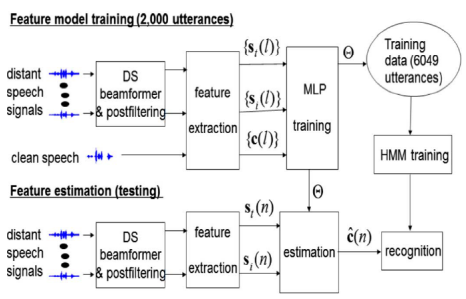

Figure 28 The diagram of the feature mapping based on speech recognition [20]

3 Conclusions

The contribution provided a comprehensive overview of multi-microphones cancellation for DSR. Multi-input multi-output inverse filtering of room acoustics and spectral subtraction cancel reverberation are proposed. In multi-channel linear prediction based on short time Fourier transform representation, two approaches of window effect reduction and window effect compensation are come up with, MWER approximately estimates reverberation , however, MWEC precisely estimates reverberation. The conventional LIME combines zero-forcing and adaptive filter, obtaining better performance. The neural network including one input layer, two or three hidden layers, another bottleneck layer of DNN is used to train, leading to better results than beamforming. Several beamforming algorithms are proposed, such as the delay-and-sum beamforming, minimum variance distortionless response beamforming, super-directive beamforming, statistically optimum beamforming, adaptive beamforming, generalized eigenvalue beamforming, optical modal beamforming for spherical microphone arrays, cluster blind beamforming, feature mapping of multiple beamformed sources. According to different given conditions, different dereverberation algorithms are chosen.

References

1. Kenichi Kumatani, John McDonough, and Bhiksha Raj, "microphone array processing for distant speech recognition," *IEEE SIGNAL PROCESSING MAGAZINE*, pp. 127–140, 2012.
2. MASATO MIYOSHI and YUTAKA KANEDA, "Inverse filtering of room acoustics," *IEEE TRANSACTIONS ON ACOUSTICS. SPEECH, AND SIGNAL PROCESSING*, vol. 36, no. 2, pp. 145–152, 1988.
3. K.Lebart, J.M.oucher, and P.N.Denbigh, "A new method based on spectral subtraction for speech dereverberation," *in Proc.Int.Workshop Acoustic Echo and Noise Control,Tel Aviv,Israel*, 2010.
4. Nakatani, T. T.Yoshioka, K. Kinoshita, M. Miyoshi, and Biing-Hwang Juang, "Blind speech dereverberation with multi-channel linear prediction based on short time fourier transform representation," *Acoustics, Speech and Signal Processing,ICASSP 2008*, pp. 85–88, 2008.
5. Jae-Mo Yang and Hong-Goo Kang, "Online speech dereverberation algorithm based on adaptive multichannel linear prediction," *IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING*, vol. 22, no. 3, pp. 608–619, 2014.
6. I. Ram, E. Habets, Y. Avargel, and I. Cohen, "Multi-microphone speech dereverberation using lime and least squares filtering," *Proc.Eur.Signal Process.Conf.(EUSIPXO'08)*, 2008.
7. Yulan Liu, Pengyuan Zhang, and Thomas Hain, "Using neural network front-ends on far field multiple microphones based speech recognition," *ICASSP 2014*, pp. 5579–5583, 2014.
8. Dr Matthias Woelfel and Dr. John McDonough, ," .
9. B.D. Van Veen and K.M. Buckley, "Beamforming: a versatile approach to spatial filtering," *ASSP Magazine, IEEE*, vol. 5, no. 2, pp. 4–24, 1988.
10. Kenichi Kumatani, John McDonough, and Bhiksha Raj, "Maximum kurtosis beamforming with a subspace filter for distant speech recognition," *in Proc.ASRU*, pp. 1–6, 2011.
11. Kenichi Kumatani, John McDonough, Barbara Rauch, Philip N. Garner, Weifeng Li, and John Dines, "Maximum kurtosis beamforming with the generalized sidelobe canceller," *in Pro.Interspeech,Brisbane,Austrilia*, pp. 423–426, 2008.
12. Kenichi Kumatani, Barbara Rauch, John McDonough, and Pittsburgh Dietrich Klakow, "Maximum negentropy beamforming using complex generalized gaussian distribution model," *in Pro.ASILOMAR,Pacific Grove,CA*, pp. 1420–1424, 2010.
13. Kenichi Kumatani, Liang Lu, John McDonough, Arnab Ghoshal, and Dietrich Klakow, "Maximum negentropy beamforming with superdirectivity," *in Pro.European Signal Processing Conf.(EUSIPCO),Alborg,Denmark*, pp. 2067–2071, 2010.
14. K. Nathwani, S. Khunteta, P. Nathwani, and R.M. Hegde, "Multi channel speech dereverberation using lp residual cepstrum in an adaptive beamforming framework," *Communications (NCC), 2014 Twentieth National Conference on*, pp. 1–6, 2014.
15. Tobias Wolff and Markus Buck, "A generalized view on microphone array postfilters," *Acta Acoust*, pp. 359–366, 2001.
16. Ernst Warsitz and Ernst Warsitz, "Blind acoustic beamforming based on generalized eigenvalue decomposition," *IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING*, vol. 15, no. 5, pp. 1529–1539, 2007.
17. Ernst Warsitz, Alexander Krueger, and Reinhold Haeb-Umbach, "Speech enhancement with a new generalized eigenvector blocking matrix for application in a generalized sidelobe canceller," *ICASSP 2008*, pp. 73–76, 2008.
18. Shefeng Yan, Haohai Sun, Xiaochuan Ma, and Jens M. Hovem, "Optimal modal beamforming for spherical microphone arrays," *IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING*, vol. 19, no. 2, pp. 361–371, 2011.
19. Ivan Himawan, Member Iain McCowan, and Sridha Sridharan, "Clustered blind beamforming from ad-hoc microphone arrays," *IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING*, vol. 19, no. 4, pp. 661–675, 2011.
20. Weifeng Li, Longbiao Wang, Yicong Zhou, Mathew Magimai, and Herve Bourlard, "Feature mapping of multiple beamformed sources for robust overlapping speech recognition using a microphone array," *IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING*, vol. 22, no. 12, pp. 2244–2255, 2014.
21. Pawel Swietojanski, Arnab Ghoshal, and Steve Renals, "Hybrid acoustic models for distant and multichannel large vocabulary speech recognition," *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*, pp. 285–290, 2013.
22. D. Marino and T. Hain, "An analysis of automatic speech recognition with multiple microphones," *INTERSPEECH*, pp. 1281–1284, 2011.