ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-7605-5/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICASSP39728.2021.9414288

ASYNCHRONOUS ACOUSTIC ECHO CANCELLATION OVER WIRELESS CHANNELS

Robert Ayrapetian, Philip Hilmes, Mohamed Mansour, Trausti Kristjansson, Carlo Murgia

Amazon Inc., USA

ABSTRACT

We introduce a novel acoustic echo cancellation framework for systems where the loudspeaker and the microphone array are not synchronized. We consider the problem in the most general form where the loss of synchronization is time-varying. The proposed system is linear and it utilizes microphone array beamforming for echo cancellation. It is shown to provide significant improvement over standard echo cancellation and noise suppression techniques in both noise suppression and speech recognition.

Index Terms— Beamforming, Microphone Arrays, Acoustics, smart speakers, wireless channel, synchronization, Acoustic Echo Cancellation.

1. INTRODUCTION

The recent success of smart speakers in the consumer electronics market has created new opportunities and challenges for array signal processing [1]. This success is primarily attributed to advances in digital signal processing and machine learning systems [2]. In general, a smart speaker provides hands-free spoken language interface that supports the following functionalities: *Play* audio content, *Listen* to voice queries, *understand* the voice queries, and *respond* either with a vocal response or an action. A key requirement for smart speaker systems is the ability to listen to voice commands in the presence audio playback. This requires an effective Acoustic Echo Cancellation (AEC) that suppresses the interfering playback signal without distorting the target speech.

In this work, we consider a general arrangement for audio playback where the smart speaker has both internal and external playbacks. External loudspeakers might be connected wirelessly to the smart speaker, e.g., through WiFi or Bluetooth. This wireless channel introduces time-varying delay between the reference signal (at the smart speaker) and the playback signal (at the external loudspeaker). Further, the external loudspeaker has its own system clock that is different from the system clock of the smart speaker; which introduces time/frequency desynchronization. These challenging problems hinder conventional AEC systems, and significantly reduce level of echo suppression. With loss of synchronization between playback path and microphone, there are two choices to handle the AEC problem:

1. Synchronize the reference and target signals prior to AEC processing [3–7]. For example, sample rate offset was estimated and corrected with proper interpolation in [4, 6, 7]. In general, this body of work is effective for the desynchronization problem, when it is not time-varying. However, the wireless channel that is considered here has a more general desynchronization model, where the time/frequency offset might be time-varying due to frame drops and other factors (as will be discussed in section 2.2). The above solutions cannot handle this model within reasonable latency and/or complexity constraints. Moreover, these solutions are restricted to single playback channel and do not generalize to the multichannel case when multiple loudspeakers are connected wirelessly to the smart speaker. Further, this body of work does not handle possible loudspeaker nonlinearity due to high playback volume, and a separate nonlinear AEC is needed.
2. Ignore the reference signal, and treat the echo cancellation problem as a noise suppression problem [8–11]. In general, these techniques work relatively well under certain assumptions and at relatively high SER, e.g., > -5 dB. In our case, the target speech is usually buried in the playback audio, and the input SER can be well below -30 dB, and these approaches usually fail.

In our setup, we use a general form of multichannel playback, where the time and frequency offsets between the playback signals and the microphone signal are time-varying due to the clock offset and possible packet errors in the wireless channel. We utilize microphone array beamforming to segment the acoustic surrounding to two disjoint sets: one set is dominated by the playback audio, and the other set contains the far-field target speech with a lower level of the playback audio. This segmentation is then utilized in a multichannel AEC (MCAEC) setting to cancel the playback audio component in the segment that contains the target speech. We describe many implementation details to mitigate distortion to the target speech and significantly improve the output SER. The proposed approach has low complexity that is suited for embedded implementation. It is shown to be effective for asynchronous AEC in terms of both Echo Return Loss Enhancement (ERLE) and False Rejection Rate (FRR) of keywords in smart speaker systems.

2. SYSTEM SETUP

2.1. System Configuration

The configuration of the system under study is shown in Fig. 1, where one or more external loudspeakers are paired with a smart speaker via wireless channels, e.g., WiFi or bluetooth. The smart speaker is equipped with a microphone array and the external loudspeakers are in the far-field of the microphone array. Each external loudspeaker might preprocess the streamed audio data from the smart speaker prior to playback (as in commercial bluetooth loudspeakers). The internal loudspeakers might simultaneously play one or more playback channels while the smart speaker is paired with external loudspeakers.

In addition to the ambient noise, other directional noise sources might simultaneously exist during audio playback as in Fig. 1. At any instance during audio playback, a user might talk to the smart speaker with voice queries starting with a keyword, e.g., *Alexa* or *Hey Siri*. The input SER at the microphone array can be as low as -30 dB. The positions of the external loudspeakers and the target user are not known a priori to the smart speaker, and these positions could change over time.

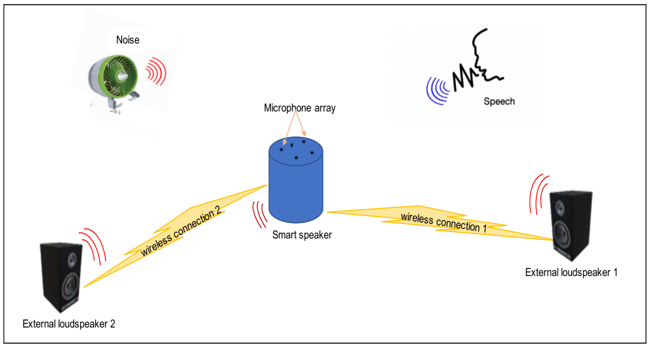

Fig. 1: The smart speaker system configuration with external loudspeakers over wireless connections

The system objective is to suppress the playback echo and interference from other noise sources at the microphone array such that the output SER is high enough to properly detect the keyword and recognize the target speech. Therefore, it is critical to minimize the distortion to the target speech while suppressing playback echo and other interferences. Note that, the voice command event is a relatively sparse event as compared to audio playback. In most smart speakers, the playback volume is reduced significantly when a keyword is detected such that subsequent voice commands have a much higher SER. Therefore, the SER can fluctuate by as much as 30 dB within a single phrase, and the system has to properly react to maintain seamless user experience.

2.2. Channel Model

Each external loudspeaker has its own system clock that is different from the smart speaker clock, and there is no time nor frequency synchronization of the different clocks. Packet errors that result in packet drops might frequently occur in the wireless connection. Assume there are P external playback channels and Q internal playback channels. Let $\{x_i(\omega,t)\}_{i=1}^{P}$ denote the time-frequency representation of the i-th external playback channel. Similarly, let $\{u_i(\omega,t)\}_{i=1}^{Q}$ denote the internal playback channels. Let $\{y_m(\omega,t)\}_{m=1}^{M}$ denote the time-frequency representation of the m-th microphone. Let $s_m(\omega,t)$ and $n_m(\omega,t)$ denote respectively the speech and noise components at the m-th microphone. Then, the model for the system configuration in Fig. 1 has the form

$$\begin{aligned} y_m(\omega,t) = &\sum_{i=1}^{P} x_i(\omega+\delta_i(\omega), t+\epsilon_i(t)).h_{m,i}(\omega,t).b_i(t) \\ &+ \sum_{i=1}^{Q} u_i(\omega,t).g_{m,i}(\omega,t) + s_m(\omega,t) + n_m(\omega,t). \end{aligned} \tag{1}$$

where $b_i(t)$ is an indicator function to account for possible packet drops at time t, $\{h_{m,i}, g_{m,i}\}$ are the time-frequency representation of the room impulse response for the i-th external and internal playback signals at the m-th microphone respectively. The time and frequency offsets of the i-th external playback component are denoted ϵ_i and δ_i respectively. Note that, these offsets could be time-varying. Note that, in a regular AEC setup the playback signals are used to minimize the estimation error (in mean square sense) of the microphone signal, where the adaptive filters approximate the room impulse responses $\{h_{m,i}, g_{m,i}\}$.

2.3. Beamformer Design

We use the Minimum-Variance-Distortionless-Response (MVDR) procedure [12, 13], which minimizes the beamformer output energy with no distortion to the target signal. If the steering vector in the target direction is $\mathbf{d}(\omega)$, and $\mathbf{\Psi}(\omega)$ is the spatial coherence matrix, then the MVDR beamformer has the form [14]

$$\widehat{\mathbf{w}}(\omega) = \frac{\mathbf{\Psi}^{-1}(\omega)\mathbf{d}(\omega)}{\mathbf{d}^H(\omega)\mathbf{\Psi}^{-1}(\omega)\mathbf{d}(\omega)} \tag{2}$$

The steering vector in the above equation utilizes the acoustic model described in [15] to account for the scattering due to the device surface. To cover the whole 3D space around the microphone array, a set of K fixed beamformers at different nonoverlappling directions are designed as above. The beams are parameterized by different steering vectors $\{\mathbf{d}^{(k)}(\omega)\}_{k=1}^{K}$ and the corresponding beamformer filters are $\{\mathbf{w}^{(k)}(\omega)\}_{k=1}^{K}$. The output of each beamformer is

$$v_k(\omega,t) = \sum_{m=1}^{M} y_m(\omega,t).w_m^{(k)}(\omega) \tag{3}$$

One of the disadvantages of fixed beamformers is the poor white noise gain (WNG), especially in the low frequency range. In [16], the WNG was constrained in the fixed beamformer design. The WNG constraint sets a minimum value on $A(\omega) \geq \gamma > 0$, which enables the problem to be formulated as a convex problem.

3. ADAPTIVE REFERENCE ADAPTATION

3.1. Architecture

In the presence of clock offset, the reference signals $\{x_i(\omega,t)\}_{i=1}^{P}$ cannot be used directly in a typical synchronous MCAEC configuration unless the time and frequency offsets are properly estimated and corrected. The joint estimation of the time-varying offsets $\{\delta_i(\omega), \epsilon_i(t)\}_{i=1}^{P}$ and the gating functions $\{b_i(t)\}_{i=1}^{P}$ is prohibitively complex and quite challenging to implement within the memory/latency/computation constraints of embedded systems. If these offsets are not estimated accurately, the resulting ERLE is typically below 10 dB, which is not sufficient for satisfactory performance.

Rather than using $\{x_i(\omega,t)\}_{i=1}^{P}$ directly in AEC, these playback references are utilized to create new references that are properly aligned in time and frequency with the microphone signals $\{y_m(t,\omega)\}_{m=1}^{M}$. This is illustrated in Fig. 2, where these new references are computed from the beamformed signals $\{v_k(\omega,t)\}_{k=1}^{K}$, and they are denoted by $\{z_l(\omega,t)\}_{l=1}^{L}$. We refer to this arrangement as the Adaptive Reference Adaptation (ARA) algorithm, which operates as follows:

1. The internal playback components $\{u_i(\omega,t)\}_{i=1}^{Q}$ are cancelled using a conventional MCAEC arrangement.
2. The external playback components are cancelled in the second MCAEC stage, where the target signals are $\{v_{i_k}(\omega,t)\}_{k=1}^{R}$; which is a subset of the fixed beamformer outputs $\{v_k(\omega,t)\}_{k=1}^{K}$.
3. The reference signals for the second MCAEC stage $\{z_l(\omega,t)\}_{l=1}^{L}$ form another disjoint subset of $\{v_k(\omega,t)\}_{k=1}^{K}$.
4. For each target signal $v_{i_k}(\omega,t)$, one or more inputs from $\{z_l(\omega,t)\}_{l=1}^{L}$ are used as references in the second MCAEC stage.

3.2. Adaptive Reference Generation

The creation of the adaptive references $\{z_l(\omega,t)\}_{l=1}^{L}$ from the beamformed outputs $\{v_k(\omega,t)\}_{k=1}^{K}$ is a key component in the ARA algorithm. For each target beam, $v_{i_k}(\omega,t)$, a set of beams are selected as MCAEC references. There are two important use cases:

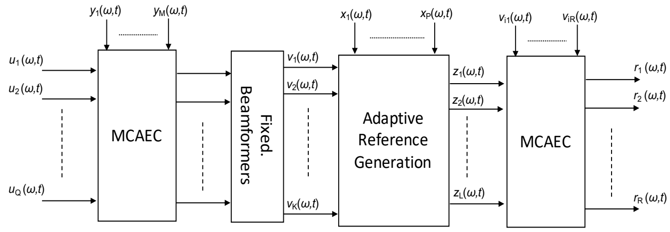

Fig. 2: MCAEC with Adaptive Reference Adaptation

- *External loudspeaker(s)*: the target and reference beams are chosen adaptively in run-time based on the average power of the fixed beamformer output $v_k(\omega, t)$ over a frequency range of interest $[\omega_1, \omega_2]$. Beams towards the loudspeaker direction have higher average power as compared to other beams. If more than one external loudspeaker are active (i.e., $P > 1$), then the high-power beams are assigned to particular input by measuring the coherence with $\{x_i(\omega, t)\}_{i=1}^{P}$. The remaining beams are assigned as target beams. The reference beams $\{z_l(\omega, t)\}_{l=1}^{L}$ could be either used for all target beams or only a subset is used for each target beam.
- *External noise*: In the absence of a reference signal, the selection of the reference beams to cancel external interference is based on beam energy.
- *Diffuse noise*: a predefined look-up table for the reference beams is used. In this case, the canceling beams for each target beam are fixed.

Two different WNG thresholds γ_{target} and $\gamma_{reference}$ are used to create two separate sets of fixed beamformers for the target and reference beams respectively. The number of beams for these sets could be different as well. Since the ARA algorithm operates on the local microphone signals, they do not depend on frequency offset, delay variation, or distortion. Therefore, it is robust to these errors (without explicitly estimating corrections), and it outperforms traditional AEC algorithms under these conditions.

3.3. AEC with leakage

The risk in the ARA algorithm is the possible distortion of the target speech signal that might leak to the reference beams $\{z_l(\omega, t)\}_{l=1}^{L}$ due to the limited directivity of the microphone array. The performance of the adaptive reference algorithms is also expected to depend on the reverberation and noise characteristics of the room.

To mitigate these issues, we use a variable step-size to control the adaptation rate of the MCAEC. The step size μ is computed using ARA MCAEC output; where for the q-th output channel we have:

$$\mu_q(\omega, t) = \frac{P_{slow}^{(q)}(\omega, t)}{P_{slow}^{(q)}(\omega, t) + \alpha . P_{fast}^{(q)}(\omega, t)} \tag{4}$$

where, $P_{slow}^{(q)}$ and $P_{fast}^{(q)}$ are slow and fast moving average power of $r_q(\omega, t)$ and $\alpha \geq 0$. The adaptation is frozen when double-talk is detected.

The above procedure introduces some distortion to the target speech. To quantify this distortion consider, without loss of generality, the case with a single external playback signal $x(\omega, t)$, a single reference beam, $z(\omega, t)$, and a single target beam $v(\omega, t)$. Let $s(\omega, t)$ denote the far-end target speech. we have

$$\begin{aligned} z(\omega, t) &= a_z(\omega)\, x(\omega, t) + b_z(\omega)\, s(\omega, t) + n_z(\omega, t) && (5) \\ v(\omega, t) &= a_v(\omega)\, x(\omega, t) + b_v(\omega)\, s(\omega, t) + n_v(\omega, t) && (6) \end{aligned}$$

where the weights $\{a_z, a_v, b_z, b_v\}$ are due to the convolution of the room impulse response and the fixed beamformer; and $\{n_z(\omega, t), n_v(\omega, t)\}$ are the residual ambient noise after beamformers. These weights are slowly varying with time as compared to $x(\omega, t)$ and $s(\omega, t)$; and in practice the dependency on time can be omitted. By applying the selection rules of reference and target beams as described in section 3.2, we get

$$\frac{|a_z(\omega)|}{|a_v(\omega)|} > \frac{|b_z(\omega)|}{|b_v(\omega)|} \tag{7}$$

The output of the MCAEC filter has the form

$$r(\omega, t) = v(\omega, t) - h(\omega, t) z(\omega, t) \tag{8}$$

As described earlier, the existence of speech signal in a smart speaker system is a sparse event. Therefore, by employing the double-talk detector to control the adaptation rate, the adaptive filter weights converge to (ignoring the misalignment error)

$$h(\omega, t) \rightarrow \frac{a_v(\omega)}{a_z(\omega)} \tag{9}$$

Hence the output becomes

$$r(\omega, t) \approx b_v(\omega)\, s(\omega, t) - \beta(\omega) s(\omega, t) + n(\omega, t) \tag{10}$$

where $n(\omega, t)$ is the residual noise, and

$$\beta(\omega) = \frac{a_v(\omega)}{a_z(\omega)} b_z(\omega) \tag{11}$$

Note that, the playback interference is eliminated but the target speech is distorted by the factor $\beta(\omega)$. If $|a_z(\omega, t)| \gg |a_v(\omega, t)|$, then $\beta(\omega) \rightarrow 0$, i.e., no distortion. In general, the ARA procedure provides significant SER improvement at the cost of small speech distortion. The FRR improvement due to improved SER outweighs the distortion problem at low input SER, e.g., less than -10 dB, whereas speech distortion becomes an issue at relatively high input SER, e.g., greater than -5 dB.

3.4. Implementation Issues

Referring to the system configuration in Fig. 2, there are few relevant use cases:

- The choice of the system parameters, e.g., the number of fixed beams and the number of reference beams depends on the available computational resources in the system.
- In the absence of external loudspeakers, the system is reduced to the standard beamformer after echo cancellation configuration [17].
- It is possible to have the internal MCAEC along with external MCAEC after beamforming. Both configurations yield similar performance. However, for a large number of beams the first configuration has lower complexity.
- In the absence of internal playback signal, the beamformer is applied directly to the input microphones.
- For each target beam, one or more reference beams per external loudspeaker are selected for MCAEC.
- A robust double-talk detector is crucial for proper operation of the ARA algorithm.
- In the absence of reference signal, the selection of the reference beams is based on beam energy and look-up table.

The ARA algorithm is linear and has low complexity because it uses only a set of fixed beamformers. This set of fixed beamformers are needed to cover all possible locations in the 3D space, and it distinguishes the ARA algorithm from other adaptive beamforming algorithms, e.g., Generalized Sidelobe Canceller (GSC) [18], that uses blocking matrices to generate the interference reference. In the ARA algorithms, the same set of fixed beamformers are used to generate both the target and interference reference, and the adaptive selection of references in the ARA algorithm provides flexibility to track interference source with minor computational overhead.

In practice, the target source position is unknown, and the final output is computed from $\{r_k(\omega, t)\}_{k=1}^{R}$. A low-complexity solution combines the output beams with weights proportional to the corresponding SNR, and the final output has the form

$$\begin{aligned} U(\omega, t) &= \sum_{k=1}^{R} a_k . r_k(\omega, t), \\ a_k &= \frac{SNR_k}{\sum_{k=1}^{R} SNR_k} \end{aligned} \tag{12}$$

4. EXPERIMENTAL RESULTS

The ARA algorithm was validated using a large corpus of data in real-life use case scenarios especially with paired bluetooth/WiFi loudspeakers. It has also been validated for a wide variety of microphone array geometry, device form factor, and room environment.

In the following, we provide analysis for a cylindrical smart speaker with a microphone array of size 7 at the top surface. The smart speaker is paired with bluetooth loudspeaker via BT A2DP protocol and playing stereo. We evaluated a large set of bluetooth loudspeakers with wide range of clock and latency variations. For good BT speakers, the conventional MCAEC with clock offset compensation provides ERLE around 30 dB; whereas for low-cost BT speakers the ERLE of conventional AEC is almost 0 dB. In Fig. 3, we give an example of the achieved ERLE with external BT loudspeaker in the presence of time varying clock jitter, and an input SER of -30 dB. Note that, the minima points in the ERLE curve for ARA correspond to double-talk events. As evident from the figure, the ARA provides significantly higher echo cancellation than conventional MCAEC.

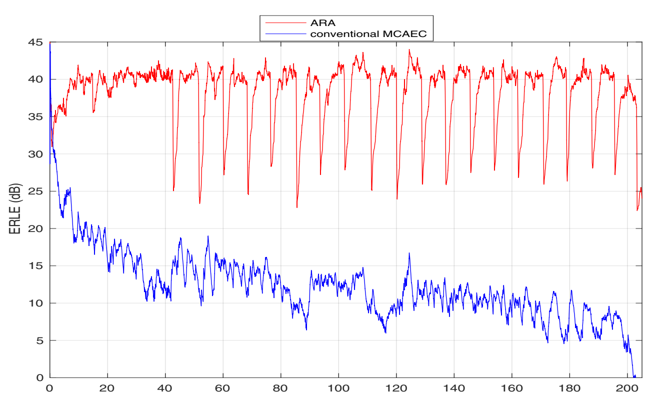

Fig. 3: Average ERLE with ARA and conventional MCAEC with external BT loudspeaker and a time-varying clock jitter of 100 ppm

In Fig. 4, we compare the false rejection rate (FRR) of the keyword, *Alexa*, using ARA and conventional MCAEC with an external BT loudspeaker paired with the smart speaker. The input SER corresponds to different playback volumes. As evident from the figure, the conventional MCAEC performs better than ARA at high SER, but its performance degrades rapidly with decreasing SER. In contrast, the ARA is quite robust to decreasing SER, due to the better selection of the reference beams and high coherency between reference and target beams. This is an experimental validation of the performance analysis in section 3.3.

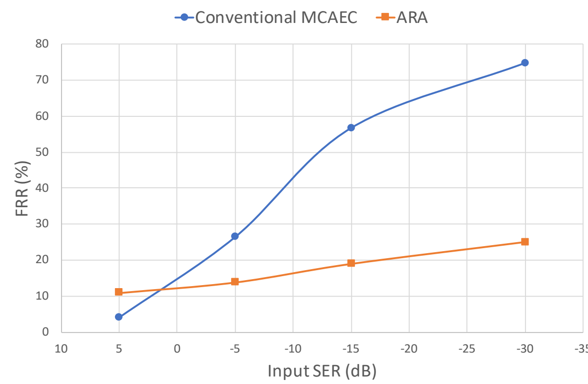

Fig. 4: FRR performance with ARA and conventional MCAEC procedures with external BT loudspeaker at different playback volumes

5. DISCUSSION

We introduced a novel framework for multichannel echo cancellation in the presence of time/frequency desynchronization between the reference signal and the playback signal. We used a general channel model that accommodates most real-life scenarios with loudspeakers that are paired through wireless channels. The framework utilizes a combination of multichannel echo cancellation and microphone array beamforming to achieve acoustic echo suppression in excess of 40 dB. The framework has the following merits:

- It is linear with low complexity and latency. It automatically corrects time/frequency desynchronization without explicit estimation algorithms that are usually too complex for embedded implementation.
- It is effective in removing directional interference even in the absence of reference signals.
- It provides high ERLE due to the increased coherence between the reference and target signals.
- It is flexible and scalable to accommodate a variable number of playback channels, with any combination of internal and external playback.

Future work investigates DNN architectures for ARA. In recent years a great deal of attention has been devoted to applying DNNs to the spatial filtering problem [19]. Most of these efforts rely on the presence of speech to adapt the coeficients of the spatial filter. For keyword detection applications, it is impractical to use the speech signal itself to adapt the spatial filter due to the low latency requirement as well as brevity of the keyword. The ARA structure however is ideal since it hypothesizes a set of look-directions that are fixed and not adapted, however it adapts the noise cancellation of each look direction. DNN based methods are known to perform well for speech and noise activity classification. Therefore, DNN based classification could improve reference and target beam selection, control the filter adaptation as well as SNR estimation for use in the fusion of the final output.

6. REFERENCES

[1] A. Chhetri, P. Hilmes, T. Kristjansson, W. Chu, M. Mansour, X. Li, and X. Zhang, "Multichannel Audio Front-End for Far-Field Automatic Speech Recognition," in *2018 European Signal Processing Conference (EUSIPCO)*, 2018, pp. 1527–1531.

[2] R. Haeb-Umbach, S. Watanabe, T. Nakatani, M. Bacchiani, B. Hoffmeister, M. Seltzer, H. Zen, and M. Souden, "Speech processing for digital home assistants: Combining signal processing with deep-learning techniques," *IEEE Signal Processing Magazine*, vol. 36, no. 6, pp. 111–124, 2019.

[3] M. Abe and M. Nishiguchi, "Frequency domain acoustic echo canceller that handles asynchronous a/d and d/a clocks," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 5924–5928.

[4] M. Pawig, G. Enzner, and P. Vary, "Adaptive sampling rate correction for acoustic echo control in voice-over-ip," *IEEE Transactions on Signal Processing*, vol. 58, no. 1, pp. 189–199, 2009.

[5] N. Ono, H. Kohno, N. Ito, and S. Sagayama, "Blind alignment of asynchronously recorded signals for distributed microphone array," in *2009 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. IEEE, 2009, pp. 161–164.

[6] E. Robledo-Arnuncio, T. S Wada, and B. Juang, "On dealing with sampling rate mismatches in blind source separation and acoustic echo cancellation," in *2007 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. IEEE, 2007, pp. 34–37.

[7] D. Cherkassky, S. Markovich-Golan, and S. Gannot, "Performance analysis of mvdr beamformer in WASN with sampling rate offsets and blind synchronization," in *2015 23rd European Signal Processing Conference (EUSIPCO)*, 2015, pp. 245–249.

[8] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on acoustics, speech, and signal processing*, vol. 27, no. 2, pp. 113–120, 1979.

[9] Y. Ephraim, "A bayesian estimation approach for speech enhancement using hidden markov models," *IEEE Transactions on Signal Processing*, vol. 40, no. 4, pp. 725–735, 1992.

[10] G. Enzner, H. Buchner, A. Favrot, and F. Kuech, "Acoustic echo control," in *Academic Press Library in Signal Processing*, vol. 4, pp. 807–877. Elsevier, 2014.

[11] Y. Luo and N. Mesgarani, "Tasnet: time-domain audio separation network for real-time, single-channel speech separation," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 696–700.

[12] J. Capon, "High-resolution frequency-wavenumber spectrum analysis," *Proceedings of the IEEE*, vol. 57, no. 8, pp. 1408–1418, 1969.

[13] H. Cox, R. Zeskind, and M. Owen, "Robust adaptive beamforming," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 35, no. 10, pp. 1365–1376, 1987.

[14] M. Brandstein and D. Ward, *Microphone arrays: signal processing techniques and applications*, Springer Science & Business Media, 2013.

[15] A. Chhetri, M. Mansour, W. Kim, and G. Pan, "On acoustic modeling for broadband beamforming," in *2019 27th European Signal Processing Conference (EUSIPCO)*, Sep. 2019, pp. 1–5.

[16] E. Mabande, A. Schad, and W. Kellermann, "Design of robust superdirective beamformers as a convex optimization problem," pp. 77–80, 2009.

[17] W. Kellermann, "Acoustic echo cancellation for beamforming microphone arrays," in *Microphone Arrays*, pp. 281–306. Springer, 2001.

[18] L. Griffiths and C. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Transactions on antennas and propagation*, vol. 30, no. 1, pp. 27–34, 1982.

[19] Y. Liu, A. Ganguly, K. Kamath, and T. Kristjansson, "Neural network based time-frequency masking and steering vector estimation for two-channel mvdr beamforming," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 6717–6721.