INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Glottal Model Based Speech Beamforming for Ad-Hoc Microphone Arrays

Yang Zhang[1], *Dinei Florêncio*[2], *Mark Hasegawa-Johnson*[1]
[1] University of Illinois, Urbana-Champaign, IL, USA
[2] Microsoft Research, Redmond, WA, USA
yzhan143@illinois.edu, dinei@microsoft.com, jhasegaw@illinois.edu

Abstract

We are interested in the task of speech beamforming in conference room meetings, with microphones built in the electronic devices brought and casually placed by meeting participants. This task is challenging because of the inaccuracy in position and interference calibration due to random microphone configuration, variance of microphone quality, reverberation etc. As a result, not many beamforming algorithms perform better than simply picking the closest microphone in this setting. We propose a beamforming called Glottal Residual Assisted Beamforming (GRAB). It does not rely on any position or interference calibration. Instead, it incorporates a source-filter speech model and minimizes the energy that cannot be accounted for by the model. Objective and subjective evaluations on both simulation and real-world data show that GRAB is able to suppress noise effectively while keeping the speech natural and dry. Further analyses reveal that GRAB can distinguish contaminated or reverberant channels and take appropriate action accordingly.
Index Terms: Beamforming, ad-hoc microphone array, speech enhancement, speech model, LPC residual

1. Introduction

Clean recordings of speech in conference rooms are useful in a number of scenarios. For instance, for remote participants, clear speech is vital for their understanding and participation. Currently, clean speech signals can be obtained via structured microphone arrays, if the conference room has any. However this is both inflexible and a waste of the resources available, because nowadays meeting participants tend to bring a lot of electronic devices, most of which carry microphones. These sensors are usually casually placed on or by the conference table, forming a large ad-hoc microphone array.

Beamforming with a heterogeneous ad-hoc microphone array is well known to be a challenging problem [1], because most beamforming algorithms rely heavily on calibration of source locations and interference characteristics, both of which can be quite inaccurate in this scenario. Without knowing the geometric configuration of the microphones, estimating the source location becomes a less constrained problem. What's worse, the sensors are heterogeneous, which adds to the errors when cross correlation is computed. Additionally, the interference characteristics vary drastically across channels, making it difficult to calibrate them specifically for each channel [2]. As a result, not many beamforming algorithms are robust in our intended scenario. MVDR, for example, is shown to deteriorate when distant microphones are included [3]. GSC will suffer from signal cancellation when position calibration is inaccurate [4].

Some previous works try to address these challenges. For example, some works [5–9] use external labels or audio events to synchronize channels. Some other works [10, 11] use information other than time delay to calibrate position. Himawan et. al. [3] proposed to select channels close enough for beamforming. These approaches address part of the challenges, but are either infeasible for the intended scenario, or yet to produce natural speech. Therefore, using the closest microphone has become a popular viable strategy.

In this paper we propose a beamforming algorithm, called Glottal Residual Assisted Beamforming (GRAB). It does not rely on position or interference calibration. Instead, it introduces a speech production model that locates the speech energy, and minimizes everything else that cannot be accounted for by the model. Experiments on both simulated and real-world data show that GRAB is able to produce clean and natural sounding speech even in very adverse conditions.

There have been past works on incorporating a speech model into beamforming. Gillespie et. al. [12] and Kumatani et. al. [13] proposed to maximize the kurtosis and negentropy. These works rest on the observation that the sample-wise distribution of speech has higher kurtosis and negentropy than corrupted speech. While such approaches leverage some information about speech, their speech models are still limited. Also, these approaches still rely on regular beamforming as initialization. Another class of methods, independent vector analyses (IVA) [14–16], introduces a prior distribution for speech and applies source independence as separation criteria, but is still vulnerable to reverberation and channel heterogeneity.

For the remainder of the paper, we will describe the algorithm in sections 2 and 3. Experimental results are analyzed in section 4. Final discussion is given in section 5.

2. Glottal Residual Assisted Beamforming

In this section, the proposed algorithm will be introduced. Denote the signal recorded by the lth channel as $y_l[t]$ within a single analysis frame of length T, and total number of channels as L. t denotes the discrete time. Each channel records the single clean speech source, denoted as $s[t]$, corrupted by reverberation and additive noise sources.

2.1. The Algorithm Framework

Our task is to determine a set of k-tap beamforming filter coefficients $\{h_1[t], \cdots, h_L[t]\}$ to obtain an estimate of the clean speech:

$$x[t] = \sum_{l=1}^{L} y_l[t] * h_l[t] \tag{1}$$

where $*$ denotes discrete time convolution.

The target function to be minimized is the L2 distance between the LPC residual of $x[t]$ and the estimated LPC residual of $s[t]$. Formally, denote the operator $\mathcal{R}_k\{x\}[t]$ as the LPC residual signal of $x[t]$ of order k. Then the optimization problem can be divided into two steps.

Step 1: Use a nonlinear speech production model to estimate $\mathcal{R}_k\{s\}[t]$, i.e. the LPC residual of the clean speech. Denote the estimate as $\hat{\mathcal{R}}_k\{s\}[t]$. The LPC order k is set to 13, which is common in speech analysis.

http://dx.doi.org/10.21437/Interspeech.2017-1659

Step 2: Obtain the beamforming filter coefficients by solving the following optimization problem:

$$\min_{\{h_1[t],\cdots,h_L[t]\}} \mathbb{E}\left(\mathcal{R}_k\{x\}[t] - \hat{\mathcal{R}}_k\{s\}[t]\right)^2 \quad (2)$$

such that eq. (1) is satisfied. $\mathbb{E}$ denotes sample mean.

The intuitions behind this formulation are twofold. First, the LPC residual of clean speech is highly structured and well studied, and therefore can be estimated from noisy observations with adequate accuracy. Second, rather than resynthesizing the clean speech directly from the estimated LPC residual, we apply a beamforming filter to retain the estimated clean speech energy. This step eliminates the artifacts and is very robust against the minor errors produced in step 1. In short, with the regularization of a strong speech model and the beamforming filter as a failsafe, the proposed algorithm is expected to perform reliably even in very adverse scenarios.

Since step 2 is simpler, it will be discussed first in section 2.2. Step 1 is solved by leveraging the relation between the clean speech LPC residual and the glottal pressure wave, which will be discussed in detail in section 3.

2.2. Iterative Wiener Filtering

The goal of this subsection is to solve the optimization problem in eq. (2). For brevity, denote a supervector $\boldsymbol{h}$ as

$$\boldsymbol{h} = [h_1[0],\cdots,h_1[B],\cdots,h_L[0],\cdots,h_L[B]]^T \quad (3)$$

Define $b_k[t;\boldsymbol{h}]$ as the LPC inverse filter impulse response of $x[t]$ of order k, i.e.

$$\mathcal{R}_k\{x\}[t] = b_k[t;\boldsymbol{h}] * x[t] = \sum_{l=1}^{L} b_k[t;\boldsymbol{h}] * y_l[t] * h_l[t] \quad (4)$$

Note that $b_k[t;\boldsymbol{h}]$ is a function of $\boldsymbol{h}$ because it is the LPC coefficients of $x[t]$, which is a function of $\boldsymbol{h}$ from eq. (1).

Define channel LPC residuals and its supervector form as

$$\begin{aligned} \rho_l[t;\boldsymbol{h}] &= b_k[t;\boldsymbol{h}] * y_l[t] \\ \boldsymbol{\rho}[t;\boldsymbol{h}] &= \big[\rho_1[t;\boldsymbol{h}],\cdots,\rho_1[t-k;\boldsymbol{h}], \\ &\quad \cdots,\rho_L[t;\boldsymbol{h}],\cdots,\rho_L[t-k;\boldsymbol{h}]\big]^T \end{aligned} \quad (5)$$

Combining eqs. (3)-(5), eq. (2) is reduced to

$$\min_{\boldsymbol{h}} \mathbb{E}\left[\left(\hat{\mathcal{R}}_k\{s\}[t] - \boldsymbol{h}^T\boldsymbol{\rho}[t;\boldsymbol{h}]\right)^2\right] \quad (6)$$

The problem in eq. (6) is non-linear in $\boldsymbol{h}$, and bears no closed-form solution. Yet, it can be solved iteratively, fixing $\boldsymbol{h}$ and $\boldsymbol{\rho}(t;\boldsymbol{h})$ alternatively. Denote the $\boldsymbol{h}$ obtained in the mth iteration as $\boldsymbol{h}^{(m)}$. Then each iteration essentially solves

$$\boldsymbol{h}^{(m)} = \underset{\boldsymbol{h}}{\text{argmin}}\mathbb{E}\left[\left(\hat{\mathcal{R}}_k\{s\}[t] - \boldsymbol{h}^T\boldsymbol{\rho}[t;\boldsymbol{h}^{(m-1)}]\right)^2\right] \quad (7)$$

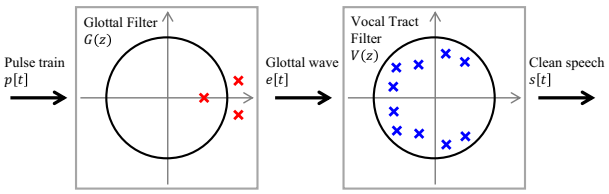
Eq. (7) is a Wiener filtering problem, whose solution is

$$\boldsymbol{h}^{(m)} = \left(\boldsymbol{R}^{(m-1)}\right)^{-1}\boldsymbol{\gamma}^{(m-1)} \quad (8)$$

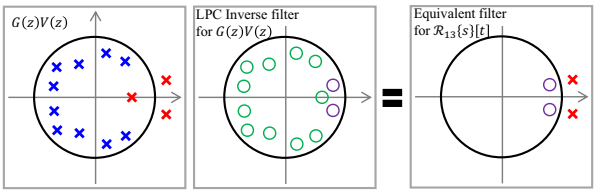
where

$$\begin{aligned} \boldsymbol{R}^{(m-1)} &= \mathbb{E}\left[\boldsymbol{\rho}(t;\boldsymbol{h}^{(m-1)})\boldsymbol{\rho}(t;\boldsymbol{h}^{(m-1)})^T\right] \\ \boldsymbol{\gamma}^{(m-1)} &= \mathbb{E}\left[\boldsymbol{\rho}(t;\boldsymbol{h}^{(m-1)})\hat{\mathcal{R}}_k\{s\}[t]\right] \end{aligned} \quad (9)$$

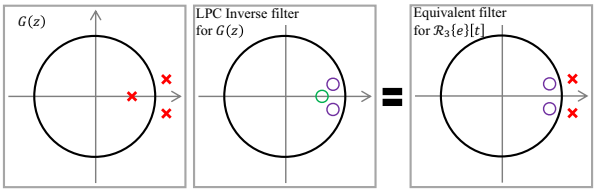
$\boldsymbol{h}^{(0)}$ is set to a delta function for the channel with the lowest 0.4 quantile in squared signal samples (usually among the cleanest channels), and 0 for the rest. 3 iterations suffice to converge.



(a) *The source-filter model for speech generation*



(b) *LPC inverse filter for clean speech.*



(c) *LPC inverse filter for glottal wave.*

Figure 1: *The source-filter model and LPC inverse filter. The green zeros in the middle plots exactly cancel the poles; the purple zeros are placed at the conjugate positions of their corresponding anti-causal poles.*

3. Estimating Clean Speech LPC Residual

This section introduces the theory and procedure of estimating the LPC residual of clean speech (step 1 mentioned in section 2.1). Unless specified otherwise, the following discussion focuses on voiced speech only. Unvoiced speech will be estimated as 0. The beamforming filter in step 2 would still retain the unvoiced speech, because it has to turn its beam towards the voiced speech source to retain voiced energy, and the unvoiced speech source is at the same of location of the voiced speech source.

3.1. The Source-Filter Model

The well-known source-filter model provides a useful signal processing perspective on speech production [17]. According to the source-filter model, as shown in figure 1(a), speech signal $s[t]$ is generated by passing a (quasi) periodic pulse train, denoted as $p[t]$, through two successive filters. The first filter, $G(z)$, is called the glottal filter, the output of which models the acoustic pressure immediately above the glottis (the so-called glottal wave), denoted as $e[t]$; the second filter, $V(z)$, is the vocal tract filter.

The impulse response of $G(z)$, denoted as $g[t]$, is essentially the glottal wave within one cycle. The LF model [18] provides an analytical approximation of its form:

$$g[t] = \begin{cases} E_0 e^{\alpha(t+t_e)}\sin\omega_g(t+t_e) & \text{if } t < 0 \\ -\frac{E_0}{\varepsilon t_\alpha}\cdot\left[e^{-\varepsilon t} - e^{-\varepsilon(t_c-t_e)}\right] & \text{if } t \geq 0 \end{cases} \quad (10)$$

It was shown that the parameters in eq. (10) (t_e, ω_g, t_α, ε and t_c) can be empirically reduced to a single parameter R_d [19].

Accordingly, in z-domain, as shown in figure 1(a), $G(z)$ can be modeled by three poles [20]: a pair of anti-causal poles

that corresponds to the $t < 0$ part in eq. (10), and a real causal pole that corresponds to the $t \geq 0$ part.

On the other hand, as shown in figure 1(a), $V(z)$ can also be modeled as an all-pole filter [17], with poles depicting resonant frequencies of the vocal tract. As a result, the combined system $G(z)V(z)$ is all-pole in nature, as shown in the left plot in figure 1(b). The number of poles is usually assumed to be 13.

3.2. LPC Analysis

The all-pole nature of $G(z)$ and $V(z)$ justifies LPC analysis on speech. The LPC residual is produced by passing the signal through a minimum-phase all-zero LPC inverse filter. In z-domain, the LPC inverse filter uses a zero to cancel every causal pole in the system. For anti-causal poles, however, it puts zeros at their conjugate positions. The conjugate position of z is z^{-1}. Figure 1(b) shows LPC analysis on speech system. As discussed, all the poles of $G(z)V(z)$ are canceled, except for the two anti-causal poles of $G(z)$. Therefore, the LPC residual of speech, $\mathcal{R}_{13}\{s\}[t]$, is equivalently generated by passing $p[t]$ through an all-pass filter.

Similarly, if we perform the order-3 LPC analysis on the glottal wave $e[t]$, which is the output of $G(z)$, we will get the same all pass filter, as shown in figure 1(c). Therefore,

$$\mathcal{R}_{13}\{s\}[t] \approx \mathcal{R}_3\{e\}[t] \tag{11}$$

3.3. Estimating $\mathcal{R}_{13}\{s\}[t]$

Eq. (11) implies the estimation of $\mathcal{R}_{13}\{s\}[t]$ can be approximated by that of $\mathcal{R}_3\{e\}[t]$. Notice from figure 1(a) that $e[t] = p[t] * g[t]$, so the task is further simplified as estimating $p[t]$ and $g[t]$. Denote the estimates as $\hat{p}[t]$ and $\hat{g}[t]$. Then

$$\hat{\mathcal{R}}_{13}\{s\}[t] = \mathcal{R}_3\{\hat{p} * \hat{g}\}[t] \tag{12}$$

The estimation of $p[t]$ and $g[t]$ is based on the cleanest channel, $y^*[t]$, which is the one with the lowest 0.4 quantile in squared signal samples.

The pulse positions of $\hat{p}[t]$ are referred to as the glottal closure instants (GCIs). It has been shown [21] that GCIs correspond to peaks of the instant energy of speech, which turns out to be quite noise robust. Therefore, we apply a simple peak-picking rule on the instant energy of $y^*[t]$, picking peaks above a threshold τ as the pulse positions of $\hat{p}[t]$.

For $\hat{g}[t]$, recall that it is parameterized by a single parameter R_d. It was shown that R_d typically falls in the range $[0.3, 3]$ [19]. Therefore, we first quantize $[0.3, 3]$ into a candidate set $\mathcal{C}$. Then, R_d is estimated by optimizing the following problem via grid search:

$$\min_{R_d \in \mathcal{C}} \mathbb{E}\left[\mathcal{R}_3\{\hat{p} * \hat{g}\}[t] - \mathcal{R}_{13}\{y^*\}[t]\right]^2 \tag{13}$$

such that $\hat{g}[t]$ satisfies eq. (10) parameterized by R_d.

4. Experiments

Experiments are performed on both simulated data and real-world data, which shows that GRAB is able to produce clean and natural sounding speech even in very adverse conditions. Readers are encourage to access the code and sample audios available in `http://tiny.cc/2rgzjy`

Table 1: *Signal-to-Noise Ratio (SNR) and Direct-to-Reverberant Ratio (DRR) on the simulated data. E_r is energy ratio of speech source over noise source in dB.*

Metric	E_r	GRAB	closest	IVA	MVDR
SNR (dB)	20	**32.8**	22.1	24.8	31.9
	10	**28.1**	11.9	22.5	27.7
	0	19.6	1.97	19.4	**22.4**
DRR (dB)	20	**10.1**	7.91	-9.07	-1.84
	10	**9.23**	7.77	-8.49	-2.64
	0	**9.03**	7.72	-9.07	-3.76

4.1. Simulated Data

Simulated cubic rooms are generated with length, width and height uniformly drawn from $[2.5, 10]$, $[2.5, 10]$, $[2.5, 5]$ meters respectively. Within each room, eight microphones and two sources are uniformly randomly scattered with the same height, which mimics conference room scenario. Source 1 is speech randomly drawn from the TIMIT corpus [22]. Source 2 is noise randomly drawn from [23–25]. The energy ratio of speech over noise, E_r, is set to three levels, 20dB, 10dB and 0dB. The transfer function from each source to each microphone is computed using the image-source method [26, 27]. The reverberation time parameter is set to 0.1s, 0.2s and 0.3s equiprobably. Each E_r setting is run 300 times, and following metrics are evaluated:

- **Signal-to-Noise Ratio (SNR)**: The energy ratio of processed clean speech over processed noise in dB.
- **Direct-to-Reverberant Ratio (DRR)**: the ratio of the energy of direct path speech in the processed output over that of its reverberation in dB. Direct path and reverberation are defined as clean dry speech convolved with the peak portion and tail portion of processed room impulse response. The peak portion is defined as $\pm$6ms within the highest peak; the tail portion is defined as $\pm$6ms beyond.

Three baselines are compared with GRAB: closest mic strategy, time-domain MVDR with non-speech segment labels given, and IVA with Laplacian prior [14]. Specifically, the MVDR is told which segments are non-speech and calibrates noise characteristics using only these segments. For the IVA method, to resolve the channel ambiguity, the channel with the highest SNR is chosen. All the beamformers are 400-tap.

Table 1 shows the objective results. In terms of noise suppression, as measured by SNR, GRAB, MVDR and IVA have significant advantage over the closest mic strategy. GRAB and MVDR are almost the same, which is quite encouraging, because the target of MVDR is specifically noise reduction and side information about voice activity is given, whereas our algorithm achieves a similar performance without explicitly measuring noise or oracle information.

In terms of reverberation reduction, as measured by DRR, GRAB achieves significantly better performance. Although MVDR and IVA can suppress noise effectively, it comes at the cost of increasing reverberation. GRAB, without measuring noise or reverberation information, strikes a good balance between noise suppression, which matches MVDR, and reverberation reduction, which outperforms the closest channel.

4.2. Real-world Data

To verify GRAB works in the intended scenario, we recorded a realistic dataset. The data were collected with eight different microphones - four wireless electret mics (numbered 1-4), three wired electret mics (numbered 5-7), and one wired dy-

Table 2: *SNR and Crowd MOS results on real-world data. Paper is short for paper shuffle.*

Metric	Noise	GRAB	closest	IVA	MVDR
SNR (dB)	Cell Phone	**18.9**	10.0	11.7	10.8
	CombBind	**17.4**	10.0	9.74	16.5
	Paper	**12.4**	10.0	6.38	7.72
	Door Slide	**18.5**	10.0	12.4	14.0
	Footstep	**17.4**	10.0	15.9	13.4
	Overall	**16.9**	10.0	11.2	12.5
MOS	Cell Phone	**3.12**	3.00	1.38	1.70
	CombBind	**3.35**	3.18	1.68	2.36
	Paper	3.21	**3.23**	1.59	2.04
	Door Slide	**3.88**	3.63	1.97	2.80
	Footstep	**3.78**	3.59	1.72	2.64
	Overall	**3.47**	3.33	1.66	2.31

Table 3: *Gain (norm of the filter coefficients) of each channel in speaker 1 + door slide scenario.*

Mic	1	2	3	4	5	6	7	8
Gain	0	0.17	0.55	0.26	0.32	0.52	0.43	0.15

namic mic (numbered 8), which mimicked the heterogeneity of recording devices. These mics were casually placed on the table of a conference room. There are two speakers, reading *My Grandfather* [28] and *The Rainbow* [29] respectively. Speaker 1 was beside mics 3 and 6; speaker 2 was beside mic 5.

To make the problem even more challenging, we deliberately introduced two special channels. Mic 1 suffered from strong hissing noise probably due to wireless interference. Mic 8 was placed right next to a noisy fan at the corner. Furthermore, five different types of noise were recorded separately, which are cell phone, CombBind machine, paper shuffle, door slide and footstep. Each was then mixed with the speech such that the SNR of the closest channel is 10dB.

Table 2 shows the objective measures. The metrics and baselines are the same as in section 4.1. The SNR of the closest channel is 10dB by construction. As can be seen, GRAB still suppresses noise more effectively than the MVDR and IVA, although all performances are worse than the simulated data. The paper shuffle case, in particular, presents challenge to all these algorithms, in part because it is a moving source. DRR cannot be evaluated on real-world data, so it is not included.

To assess the perceptual quality of the output speech, we performed a subjective evaluation via Amazon Mechanical Turk using crowdMOS [30]. The speech signal is divided into 12 short sentences of length 3-7 seconds, each combined with the five types of noise, so the total number of test sentences is 60. The subjects are asked to rate from a scale of 1-5 the quality of the speech. Each test unit, called a HIT, consists of one sentence processed by the four approaches with randomized order. Each HIT is assigned to 10 participants. Before the test, the subjects are presented with three anchor sentences, which are speaker 1's utterance with fan noise recorded by the closest mic (mic 6, with suggested score of 4 or 5), closest mic with 10dB cell phone noise (with suggested score of 2 or 3), and the bad mic (mic 1, with suggested score of 1). The anchor examples are excluded from the test set. To resolve the ambiguity of the true speech signal, which results from microphone heterogeneity, the spectral characteristics of all the test speech are normalized to match those of the TIMIT corpus via the filterbank approach.

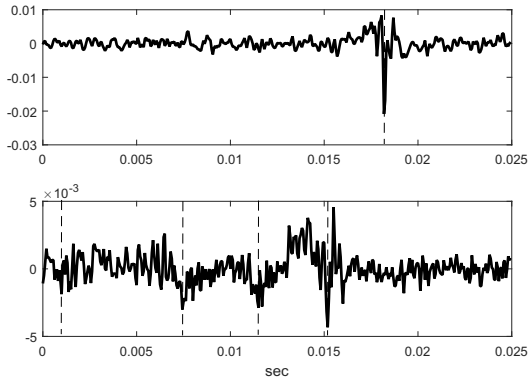


Figure 2: *Beamforming filter coefficients. Upper: channel 6, a dry channel. Lower: channel 4, a reverberant channel. Dashed lines mark the instances of impulses.*

Table 2 shows the results. Both GRAB and closest channel significantly outperform MVDR and IVA, which suggests that the heavier reverberation introduced by MVDR and IVA is perceptually unpleasant. On the other hand, GRAB is able to produce dry and clean results that are preferred over even the closest channel, except for the paper shuffle case, where the noise suppression is not so successful.

4.3. Beamforming Filter Coefficients Analyses

To demonstrate how GRAB process channels with different qualities, table 3 displays the gain of each channel, defined as the norm of the beamforming coefficients, in speaker 1 with door slide noise scenario. Recall that mic 1 is problematic and mic 8 is placed close to a noisy fan. From table 3, the gain of these two channels are very low, especially for channel 1, whose gain is very close to 0. Meanwhile, the close channels, channels 3 and 6, have the highest gains. This result shows that GRAB can automatically distinguish good channels from bad, even without explicit position or noise information.

Furthermore, to see how GRAB deals with reverberation, figure 2 shows the beamforming filter coefficients of channel 6, a dry channel, and channel 4, a reverberant channel. As can be seen, for the dry channel, the impulse response contains 1 major impulse, indicating the algorithm lets it pass distortionlessly. On the other hand, the impulse response of the reverberant channel consists of several major impulses of decreasing height from right to left, which resembles an inverse filter of the reverberation. More intuitively, rather than canceling the reverberation as proposed in many beamforming algorithms, GRAB adds reverberation back to the direct path signal. This result, again, indicates that GRAB is able to detect reverberant channels and automatically figure out a good way to process it, without any explicit reverberation measurement.

5. Conclusion and Future Directions

We have proposed GRAB, which does not rely on position and interference calibration, but locates speech energy guided by a speech model and minimize the non-speech energy. Experiments have shown that it can suppress both noise and reverberation. One of our next steps is to adapt the algorithm to be real-time, after which many standing problems with ad-hoc microphone arrays can potentially be solved, including clock drift and moving speaker.

6. References

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

[2] S. Markovich-Golan, A. Bertrand, M. Moonen, and S. Gannot, "Optimal distributed minimum-variance beamforming approaches for speech enhancement in wireless acoustic sensor networks," *Signal Processing*, vol. 107, pp. 4–20, 2015.

[3] I. Himawan, I. McCowan, and S. Sridharan, "Clustered blind beamforming from ad-hoc microphone arrays," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 661–676, 2011.

[4] J. Bitzer, K. U. Simmer, and K.-D. Kammeyer, "Theoretical noise reduction limits of the generalized sidelobe canceller (GSC) for speech enhancement," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 5. IEEE, 1999, pp. 2965–2968.

[5] R. Sakanashi, N. Ono, S. Miyabe, T. Yamada, and S. Makino, "Speech enhancement with ad-hoc microphone array using single source activity," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*. IEEE, 2013, pp. 1–6.

[6] N. D. Gaubitch, W. B. Kleijn, and R. Heusdens, "Auto-localization in ad-hoc microphone arrays," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2013, pp. 106–110.

[7] M. H. Hennecke and G. A. Fink, "Towards acoustic self-localization of ad hoc smartphone arrays," in *Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA)*. IEEE, 2011, pp. 127–132.

[8] R. Lienhart, I. Kozintsev, S. Wehr, and M. Yeung, "On the importance of exact synchronization for distributed audio signal processing," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 4. IEEE, 2003, pp. IV–840.

[9] V. C. Raykar, I. V. Kozintsev, and R. Lienhart, "Position calibration of microphones and loudspeakers in distributed computing platforms," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 1, pp. 70–83, 2005.

[10] Z. Liu, Z. Zhang, L.-W. He, and P. Chou, "Energy-based sound source localization and gain normalization for ad hoc microphone arrays," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 2. IEEE, 2007, pp. II–761.

[11] M. Chen, Z. Liu, L.-W. He, P. Chou, and Z. Zhang, "Energy-based position estimation of microphones and speakers for ad hoc microphone arrays," in *2007 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. IEEE, 2007, pp. 22–25.

[12] B. W. Gillespie, H. S. Malvar, and D. A. Florêncio, "Speech dereverberation via maximum-kurtosis subband adaptive filtering," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 6. IEEE, 2001, pp. 3701–3704.

[13] K. Kumatani, J. McDonough, B. Rauch, D. Klakow, P. N. Garner, and W. Li, "Beamforming with a maximum negentropy criterion," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 5, pp. 994–1008, 2009.

[14] T. Kim, H. T. Attias, S.-Y. Lee, and T.-W. Lee, "Blind source separation exploiting higher-order frequency dependencies," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 70–79, 2007.

[15] Y.-O. Li, T. Adali, W. Wang, and V. D. Calhoun, "Joint blind source separation by multiset canonical correlation analysis," *IEEE Transactions on Signal Processing*, vol. 57, no. 10, pp. 3918–3929, 2009.

[16] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, "Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 9, pp. 1626–1641, 2016.

[17] T. F. Quatieri, *Discrete-time speech signal processing: principles and practice*. Pearson Education India, 2006.

[18] G. Fant, J. Liljencrants, and Q.-G. Lin, "A four-parameter model of glottal flow," *STL-QPSR*, vol. 4, no. 1985, pp. 1–13, 1985.

[19] G. Fant, "The LF-model revisited. transformations and frequency domain analysis," *Speech Trans. Lab. Q. Rep., Royal Inst. of Tech. Stockholm*, vol. 2, no. 3, p. 40, 1995.

[20] W. R. Gardner and B. D. Rao, "Noncausal all-pole modeling of voiced speech," *IEEE Transactions on Speech and Audio Processing*, vol. 5, no. 1, pp. 1–10, 1997.

[21] Y. M. Cheng and D. O'Shaughnessy, "Automatic and reliable estimation of glottal closure instant and period," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 12, pp. 1805–1815, 1989.

[22] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1," *NASA STI/Recon technical report n*, vol. 93, 1993.

[23] A. Kumar and D. Florêncio, "Speech enhancement in multiple-noise conditions using deep neural networks," *INTERSPEECH*, 2016.

[24] "Freesound," https://freesound.org/, 2015.

[25] G. Hu, "100 nonspeech sounds," http://web.cse.ohio-state.edu/pnl/corpus/HuNonspeech/HuCorpus.html, 2015.

[26] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[27] E. A. Lehmann and A. M. Johansson, "Diffuse reverberation model for efficient image-source simulation of room impulse responses," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 6, pp. 1429–1439, 2010.

[28] A. E. Aronson and J. R. Brown, *Motor speech disorders*. WB Saunders Company, 1975.

[29] G. Fairbanks, *Voice and articulation: drillbook*. Harper & Brothers, 1940.

[30] F. Ribeiro, D. Florêncio, C. Zhang, and M. Seltzer, "CrowdMOS: An approach for crowdsourcing mean opinion score studies," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2011, pp. 2416–2419.