Frequency-dependent speaker detection using a microphone array

Jürgen Freudenberger, Simon Grimm
Institute for System Dynamics, HTWG Konstanz, Germany, Email: {jfreuden, sgrimm}@htwg-konstanz.de

Abstract

In this paper we propose a method to determine the active speaker for each time-frequency point in the noisy signals of a microphone array. This detection is based on a statistical model where the speech signals as well as noise signals are assumed to be multivariate Gaussian random variables in the Fourier domain. Based on this model we derive a maximum-likelihood detector for the active speaker. The decision is based on the a posteriori signal to noise ratio (SNR) of a speaker dependent max-SNR beamformer.

Introduction

Many speech processing systems require a voice activity detection (VAD), e.g. for noise reduction, speech coding or automatic speech recognition. However, many VAD algorithms assume that there is only a single speech source [1, 2, 3, 4]. In order to support multiple speakers it is desirable to distinguish different speakers if speech activity is detected [5, 6, 7]. In this work we propose a method to detect the active speaker for each time-frequency point, where each speaker is associated with a speaker dependent max-SNR beamformer. The proposed detector produces a binary time-frequency mask for each speaker that indicates speaker activity.

Binary time-frequency masking is useful for source separation of speech signals. Furthermore, time-frequency masking can be used to improve the parameter estimation, e.g. to estimate speaker dependent beamformer coefficients. For instance, in [8] the DUET algorithm was presented which can separate several speech sources based on two microphone signals provided the short-time Fourier transforms of the sources do not overlap. However, the DUET algorithm assumes a noise free and anechoic environment which is not appropriate in many practical situations. In [9] a two-stage algorithm based on independent component analysis (ICA) was proposed. The independent component analysis is first employed in each frequency bin and then time-frequency masking is used to improve the performance. This algorithm is suitable for reverberant environments provided that the sources are close to the microphones. A similar concept was presented in [10].

In this paper we present a speaker detector that is based on a statistical model of the signals in the frequency domain using the short-time Fourier transform. We assume that in the Fourier domain the speech signals as well as noise signals are multivariate Gaussian random variables [11, 12, 13, 4]. Based on this model we derive a maximum-likelihood detector for the active speaker. The decision is based on the a posteriori signal to noise ratio (SNR) of a max-SNR beamformer, where each speaker is associated with a speaker dependent beamformer. This method requires estimates of the a priori SNR at the output of the max-SNR beamformer, which can be obtained using a decision-directed approach.

The paper is organized as follows: In section , we introduce notation. The proposed detection method is presented in section . In section some simulation results are presented, where we use the speaker detection to suppress interference.

Notation

Throughout the paper we use the following notation. We denote the trace and the determinant of a matrix $\mathbf{A}$ by $\mathrm{tr}(\mathbf{A})$ and $\det(\mathbf{A})$, respectively. The transpose and Hermitian transpose are denoted by $\mathbf{A}^T$ and $\mathbf{A}^\dagger$. $\mathbf{I}$ denotes an $M \times M$ identity matrix.

We assume a microphone array with M microphones and two active speakers. The microphone signals can be expressed in the frequency domain as

$$Y_i(\kappa,\nu) = S_i^{(1)}(\kappa,\nu) + S_i^{(2)}(\kappa,\nu) + N_i(\kappa,\nu)\,, \quad (1)$$

where $Y_i(\kappa,\nu)$, $S_i^{(1)}(\kappa,\nu)$, $S_i^{(2)}(\kappa,\nu)$, and $N_i(\kappa,\nu)$ denote short-time spectra. $S_i^{(1)}(\kappa,\nu)$ is the speech component of the first speaker, $S_i^{(2)}(\kappa,\nu)$ the speech component of the second speaker, and $N_i(\kappa,\nu)$ the noise component of the i^{th} microphone signal. The subsampled time index and the frequency bin index are denoted by κ and ν, respectively. However, the dependencies on κ and ν are often omitted for simplicity. We can define the M-dimensional vectors $\mathbf{S}^{(1)}$, $\mathbf{S}^{(2)}$, $\mathbf{N}$, and $\mathbf{Y}$, in which the signals are stacked as follows:

$$\mathbf{S}_1 = [S_1^{(1)}\ S_2^{(1)}\ \cdots\ S_M^{(1)}]^T \quad (2)$$
$$\mathbf{S}_2 = [S_1^{(2)}\ S_2^{(2)}\ \cdots\ S_M^{(2)}]^T \quad (3)$$
$$\mathbf{N} = [N_1\ N_2\ \cdots\ N_M]^T \quad (4)$$
$$\mathbf{Y} = \mathbf{S}_1 + \mathbf{S}_2 + \mathbf{N}\,. \quad (5)$$

$\mathbf{R}_S^{(1)} = \mathbb{E}\left\{\mathbf{S}_1\mathbf{S}_1^\dagger\right\}$ denotes the speech correlation matrix for the first speaker. $\mathbf{R}_S^{(2)} = \mathbb{E}\left\{\mathbf{S}_2\mathbf{S}_2^\dagger\right\}$ is the speech correlation matrix of the second speaker and $\mathbf{R}_N = \mathbb{E}\left\{\mathbf{N}\mathbf{N}^\dagger\right\}$ is the spatial noise correlation matrix, where $\mathbb{E}\{\cdot\}$ denotes the expectation operator. We assume that the speech signals are zero-mean random processes with PSD $\phi_X^{(1)}$ and $\phi_X^{(2)}$, respectively. For a time-invariant acoustic system, the correlation matrices of the

speech signals can be written as

$$\begin{aligned}\mathbf{R}_S^{(1)} &= \mathbb{E}\left\{\mathbf{S}_1\mathbf{S}_1^\dagger\right\} = \phi_X^{(1)}\mathbf{A}_1\mathbf{A}_1^\dagger \\ \mathbf{R}_S^{(2)} &= \mathbb{E}\left\{\mathbf{S}_2\mathbf{S}_2^\dagger\right\} = \phi_X^{(2)}\mathbf{A}_2\mathbf{A}_2^\dagger \end{aligned} \quad (6)$$

where $\mathbf{A}_1$ and $\mathbf{A}_2$ denote the vectors of channel coefficients.

Speaker detection

In this section we derive a detector for the active speaker. We assume that the speech signals are sparse, i.e. that only one speaker is active for each time-frequency point [8]. Speech activity can be detected with a time-frequency dependent VAD [3, 4]. According to the multichannel signal model we can distinguish between the two hypotheses H_1 (the first speaker is active)

$$\mathbf{Y} = \mathbf{S}_1 + \mathbf{N} \quad (7)$$

and H_2 (the second speaker is active)

$$\mathbf{Y} = \mathbf{S}_2 + \mathbf{N}\,. \quad (8)$$

The speech and noise components are assumed as multivariate Gaussian random variables, where the real and imaginary parts of all signals are uncorrelated and identically distributed. Hence the conditional probability density functions of our observed signal vector can be modelled as [13, 14]

$$\begin{aligned} f(\mathbf{Y} \mid H_i) &= \frac{1}{\pi^M \det(\mathbf{R}_S^{(i)} + \mathbf{R}_N)} \cdot \\ &\quad \exp\left(-\mathbf{Y}^\dagger(\mathbf{R}_S^{(i)} + \mathbf{R}_N)^{-1}\mathbf{Y}\right) \end{aligned} \quad (9)$$

where i is the index of the active speaker. In order to derive a maximum-likelihood detector for the active speaker, we define the likelihood ratio

$$\Lambda = \frac{\mathrm{P}(\mathbf{Y} \mid H_1)}{\mathrm{P}(\mathbf{Y} \mid H_2)}\,. \quad (10)$$

Based on this likelihood ratio, the first speaker is detected for $\Lambda > 1$ and otherwise the second speaker is detected. Using the conditional probability density functions, we obtain

$$\begin{aligned}\Lambda &= \frac{\det\left(\mathbf{R}_S^{(2)} + \mathbf{R}_N\right)}{\det\left(\mathbf{R}_S^{(1)} + \mathbf{R}_N\right)} \exp\Big(-\mathbf{Y}^\dagger\Big[(\mathbf{R}_S^{(1)} + \mathbf{R}_N)^{-1} \\ &\quad -(\mathbf{R}_S^{(2)} + \mathbf{R}_N)^{-1})\Big]\mathbf{Y}\Big)\,. \end{aligned} \quad (11)$$

In the following the likelihood ratio is simplified. We first consider the ratio $\frac{\det\left(\mathbf{R}_S^{(2)}+\mathbf{R}_N\right)}{\det\left(\mathbf{R}_S^{(1)}+\mathbf{R}_N\right)}$ and obtain

$$\begin{aligned}\frac{\det\left(\mathbf{R}_S^{(2)} + \mathbf{R}_N\right)}{\det\left(\mathbf{R}_S^{(1)} + \mathbf{R}_N\right)} &= \frac{\det\left(\mathbf{R}_N(\mathbf{R}_N^{-1}\mathbf{R}_S^{(2)} + \mathbf{I})\right)}{\det\left(\mathbf{R}_N(\mathbf{R}_N^{-1}\mathbf{R}_S^{(1)} + \mathbf{I})\right)} \\ &= \frac{\det\left(\mathbf{R}_N\right)\det\left(\mathbf{R}_N^{-1}\mathbf{R}_S^{(2)} + \mathbf{I})\right)}{\det\left(\mathbf{R}_N\right)\det\left(\mathbf{R}_N^{-1}\mathbf{R}_S^{(1)} + \mathbf{I})\right)} \\ &= \frac{1 + \mathrm{tr}(\mathbf{R}_N^{-1}\mathbf{R}_S^{(2)})}{1 + \mathrm{tr}(\mathbf{R}_N^{-1}\mathbf{R}_S^{(1)})}\,. \end{aligned} \quad (12)$$

Defining the a priori SNR as

$$\xi_i = \mathrm{tr}(\mathbf{R}_N^{-1}\mathbf{R}_S^{(i)}) = \phi_X^{(i)}\mathbf{A}_i^\dagger\mathbf{R}_N^{-1}\mathbf{A}_i\,, \quad (13)$$

we obtain

$$\frac{\det\left(\mathbf{R}_S^{(2)} + \mathbf{R}_N\right)}{\det\left(\mathbf{R}_S^{(1)} + \mathbf{R}_N\right)} = \frac{1 + \xi_2}{1 + \xi_1}\,. \quad (14)$$

It is worth to note that ξ_i is the narrow-band output SNR of a max-SNR beamformer for the i^{th} speaker [15].

Now consider the exponential function in equation (11). Using the matrix inversion lemma [16], we can rewrite the term $(\mathbf{R}_S^{(i)} + \mathbf{R}_N)^{-1}$ in equation (11) as follows

$$\begin{aligned}(\mathbf{R}_N + \mathbf{R}_S^{(i)})^{-1} &= \mathbf{R}_N^{-1} - \frac{\mathbf{R}_N^{-1}\mathbf{R}_S^{(i)}\mathbf{R}_N^{-1}}{1 + \mathrm{tr}(\mathbf{R}_N^{-1}\mathbf{R}_S^{(i)})} \\ &= \mathbf{R}_N^{-1} - \frac{\mathbf{R}_N^{-1}\mathbf{R}_S^{(i)}\mathbf{R}_N^{-1}}{1 + \xi_i}\,. \end{aligned} \quad (15)$$

Applying (14) and (15), we get the likelihood ratio

$$\begin{aligned}\Lambda &= \frac{1 + \xi_2}{1 + \xi_1} \cdot \\ &\quad \exp\left(\mathbf{Y}^\dagger\mathbf{R}_N^{-1}\left[\frac{\mathbf{R}_S^{(1)}}{1 + \xi_1} - \frac{\mathbf{R}_S^{(2)}}{1 + \xi_2}\right]\mathbf{R}_N^{-1}\mathbf{Y}\right)\,. \end{aligned} \quad (16)$$

Using (6) and (13), we can decompose the numerator of the exponent of the likelihood ratio as follows

$$\begin{aligned}\mathbf{Y}^\dagger\mathbf{R}_N^{-1}\mathbf{R}_S^{(i)}\mathbf{R}_N^{-1}\mathbf{Y} &= \frac{\mathbf{Y}^\dagger\mathbf{R}_N^{-1}\mathbf{R}_S^{(i)}\mathbf{R}_N^{-1}\mathbf{Y}}{\xi_i}\xi_i \\ &= \frac{\phi_X^{(i)}\mathbf{Y}^\dagger\mathbf{R}_N^{-1}\mathbf{A}_i\mathbf{A}_i^\dagger\mathbf{R}_N^{-1}\mathbf{Y}}{\phi_X^{(i)}\mathbf{A}_i^\dagger\mathbf{R}_N^{-1}\mathbf{A}_i}\xi_i \\ &= \frac{\mathbf{A}_i^\dagger\mathbf{R}_N^{-1}\mathbf{Y}\mathbf{Y}^\dagger\mathbf{R}_N^{-1}\mathbf{A}_i}{\mathbf{A}_i^\dagger\mathbf{R}_N^{-1}\mathbf{A}_i}\xi_i \end{aligned} \quad (17)$$

Note that

$$\mathbf{G}_i^{\mathrm{MAX}} = \mathbf{R}_N^{-1}\mathbf{A}_i \quad (18)$$

is the filter vector of a max-SNR beamformer for the i^{th} speaker [15]. In practise, a max-SNR beamformer might also be implemented as relative transfer function (RTF) beamformer or multichannel Wiener filter.

$$\hat{X}_i = \mathbf{G}_i^{\mathrm{MAX}\dagger}\mathbf{Y} \quad (19)$$

is the estimated speech signal at the output of the max-SNR beamformer. Thus, we obtain

$$\begin{aligned}\mathbf{Y}^\dagger\mathbf{R}_N^{-1}\mathbf{R}_S^{(i)}\mathbf{R}_N^{-1}\mathbf{Y} &= \frac{\mathbf{G}_i^{\mathrm{MAX}\dagger}\mathbf{Y}\mathbf{Y}^\dagger\mathbf{G}_i^{\mathrm{MAX}}}{\mathbf{G}_i^{\mathrm{MAX}\dagger}\mathbf{R}_N\mathbf{G}_i^{\mathrm{MAX}}}\xi_i \\ &= \frac{|\hat{X}_i|^2}{\phi_N^{(i)}}\xi_i\,, \end{aligned} \quad (20)$$

$|\hat{X}_i|^2$ and $\phi_N^{(i)}$ can be interpreted as the instantaneous signal magnitude and the noise PSD at the output of a

max-SNR beamformer for speaker i, respectively. Defining the a posteriori SNR for the multichannel scenario as

$$\gamma_i = \frac{|\hat{X}_i|^2}{\phi_N^{(i)}} \tag{21}$$

we have

$$\mathbf{Y}^\dagger \mathbf{R}_N^{-1} \mathbf{R}_S^{(i)} \mathbf{R}_N^{-1} \mathbf{Y} \quad = \quad \gamma_i \; \xi_i \, . \tag{22}$$

This results in the following expression for the likelihood ratio

$$\Lambda \quad = \quad \frac{1+\xi_2}{1+\xi_1} \cdot \exp\left(\frac{\gamma_1 \; \xi_1}{1+\xi_1} - \frac{\gamma_2 \; \xi_2}{1+\xi_2} \right) . \tag{23}$$

Based on this likelihood ratio, we conclude that for $\Lambda > 1$ the hypotheses H_1 is more likely than hypotheses H_2. Consequently, the first speaker is detected for $\ln(\Lambda) > 0$ and otherwise the second speaker is detected. Taking the logarithm of the likelihood ratio we obtain the following maximum-likelihood speaker detector

$$\beta(\kappa,\nu) = \begin{cases} 1, \; \frac{\gamma_1 \; \xi_1}{1+\xi_1} > \frac{\gamma_2 \; \xi_2}{1+\xi_2} + \ln\left(\frac{1+\xi_1}{1+\xi_2}\right) \\ 2, \text{ otherwise} \, . \end{cases} \tag{24}$$

Note that the weights

$$G_i^{WF} = \frac{\xi_i}{1+\xi_i} \tag{25}$$

are equivalent to the filter coefficients of a single channel Wiener filter at the output of the max-SNR beamformer $\mathbf{G}_i^{\mathrm{MAX}} = \mathbf{R}_N^{-1}\mathbf{A}_i$.

The speaker detector according to (24) can be generalized for multiple speakers

$$\beta(\kappa,\nu) = \underset{i}{\operatorname{argmax}} \left\{ \frac{\gamma_i \; \xi_i}{1+\xi_i} - \ln(1+\xi_i) \right\} . \tag{26}$$

The speaker detector according to (26) requires estimates for the a priori SNR ξ_i and the noise power $\phi_N^{(i)}$. These parameters can be estimated using the decision-directed estimation approach proposed by Ephraim and Malah [11, 17]. Let $\alpha \in (0,1)$ be a smoothing parameter. If the i^{th} speaker is detected, we obtain the estimates

$$\xi_i(\kappa,\nu) = \alpha \xi_i(\kappa-1,\nu) + (1-\alpha)\gamma_i(\kappa,\nu) \tag{27}$$

and

$$\phi_N^{(i)}(\kappa,\nu) = \phi_N^{(i)}(\kappa-1,\nu). \tag{28}$$

Whereas we use

$$\phi_N^{(i)}(\kappa,\nu) = \alpha \phi_N^{(i)}(\kappa-1,\nu) + (1-\alpha)|\hat{X}_i(\kappa,\nu)|^2 \tag{29}$$

and

$$\xi_i(\kappa,\nu) = \xi_i(\kappa-1,\nu) \tag{30}$$

if the i^{th} speaker is not active.

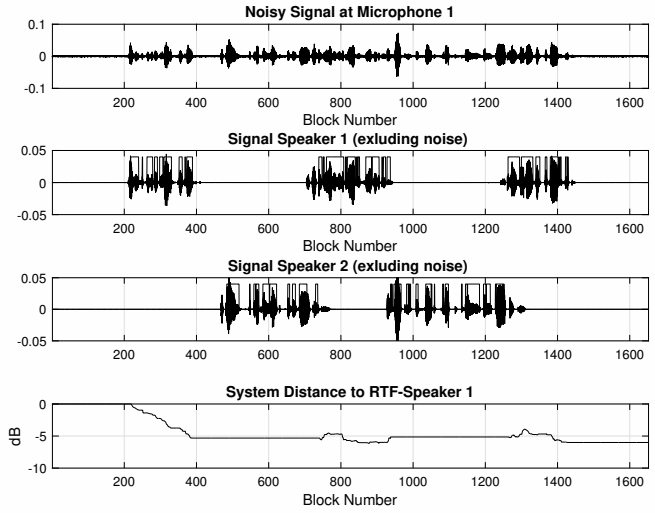

Figure 1: Microphone and speech signals.

Interference suppression

If the speech signal of a speaker that interferes with the target speaker is considered as noise, we can use the Wiener filter according to (25) to suppress this interference at the output of the max-SNR beamformer. In the following we present some simulation results for such an interference suppression approach.

We used the following set-up with two microphones in a car. The cardioid microphones were mounted close to the rearview mirror with a microphone distance of 10cm. Room impulse responses were measured with an artificial head in the position of the driver and co-driver, respectively. Noise signals were recorded at a car speed of 100km/h.

The max-SNR beamformer was implemented as a relative transfer function (RTF) beamformer as proposed in [18]. The relative transfer functions were estimated according to [19]. In order to adjust these estimation approach to two speakers, we calculated the time difference of arrival (TDOA) for each time frame. Each speaker was associated with a TDOA value and the corresponding RTF was only adapted if the TDOA was equal for two consecutive time frames.

Figure 1 depicts some of the signals of this simulation set-up. The top figure is the input signal of one microphone. The next two figures are the clean speech signals for driver and co-driver. Additionally the time frames for the RTF estimation are indicated. The lower figure is the distance between the estimated RTF and the actual RTF for the driver.

With this set-up the signal to interference ratio (SIR) for the driver signals is -1.48 dB for the first microphone and -1.63 dB for the second, respectively. The SIR at the output of the beamformer is -0.16 dB, where the beamformer is not used to suppress the interference. The SIR after the interference suppression is 4.77 dB.

Conclusions

In this paper we have proposed a method to determine the active speaker for each time-frequency point in the noisy signals of a microphone array. The main contribution of this work is the derivation of the maximum-likelihood detector. The simulation results are only the first trial of testing the theoretical model in a practical set-up. The time-frequency dependent detection might be useful for source separation of speech signals [20], to improve the parameter estimation, or for speaker localization.

References

[1] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Processing Letters*, vol. 16, no. 1, pp. 1–3, 1999.

[2] J. Ramirez, J. Segura, C. Benirez, A. de la Torre, and A. Rubio, "A new voice activity detector using subband order-statistics filters for robust speech recognition," in *Proc. IEEE Int. Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, May. 2004, pp. 849–852.

[3] J. Freudenberger and S. Stenzel, "Time-frequency dependent voice activity detection based on a simple threshold test," in *IEEE Workshop on Statistical Sig. Proc. (SSP), Nice*, 2011.

[4] S. Stenzel and J. Freudenberger, "Time-frequency dependent multichannel voice activity detection," in *Speech Communication; 11. ITG Symposium; Proceedings of*, Sept 2014, pp. 1–4.

[5] A. Bertrand and M. Moonen, "Energy-based multi-speaker voice activity detection with an ad hoc microphone array," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, March 2010, pp. 85–88.

[6] T. Matheja, M. Buck, and T. Wolff, "Enhanced speaker activity detection for distributed microphones by exploitation of signal power ratio patterns," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012, pp. 2501–2504.

[7] T. Matheja and M. Buck, "Detection of local disturbances and simultaneously active speakers for distributed speaker-dedicated microphones in cars," in *Proceedings of 11th ITG Symposium Speech Communication;*, Sept 2014, pp. 1–4.

[8] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Transactions on Signal Processing*, vol. 52, no. 7, pp. 1830 – 1847, 2004.

[9] H. Sawada, S. Araki, R. Mukai, and S. Makino, "Blind extraction of dominant target sources using ICA and time-frequency masking," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 6, pp. 2165 –2173, 2006.

[10] X. Ma, W. Liu, F. Yin, and X. Liu, "A speech separation method combining time-frequency masking and independent component analysis," in *Innovative Computing Information and Control, 2008. ICICIC '08. 3rd International Conference on*, 2008, p. 359.

[11] Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Log-Spectral Amplitude Estimator," *Signal Processing*, vol. 1, pp. 443–445, 1985.

[12] I. Potamitis and E. Fishler, "Speech activity detection of moving speaker using microphone arrays," *Electronics Letters*, vol. 39, no. 16, pp. 1223–1225, Aug 2003.

[13] M. Souden, J. Chen, J. Benesty, and S. Affes, "Gaussian model-based multichannel speech presence probability," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 18, no. 5, pp. 1072–1077, 2010.

[14] A. van den Bos, "The Multivariate Complex Normal Distribution-A Generalization," *IEEE Transactions on Information Theory*, vol. 41, no. 2, pp. 537–539, 1995.

[15] S. Stenzel and J. Freudenberger, "Blind matched filtering for speech enhancement with distributed microphones," *Journal of Electrical and Computer Engineering*, 2012, Article ID 169853, 15 pages.

[16] K. S. Miller, "On the inverse of the sum of matrices," *Mathematics Magazine*, vol. 54, no. 2, pp. 67 – 72, 1981.

[17] I. Cohen, "On the decision-directed estimation approach of Ephraim and Malah," in *Acoustics, Speech, and Signal Processing, 2004. Proceedings. (ICASSP '04). IEEE International Conference on*, vol. 1, May 2004, pp. I–293–6 vol.1.

[18] S. Stenzel, J. Freudenberger, and G. Schmidt, "A minimum variance beamformer for spatially distributed microphones using a soft reference selection," in *Hands-free Speech Communication and Microphone Arrays (HSCMA), 2014 4th Joint Workshop on*, May 2014, pp. 127–131.

[19] M. Schwab, P. Noll, and T. Sikora, "Noise robust relative transfer function estimation," in *European Signal Processing Conference (EUSIPCO)*, vol. 2, 2006, pp. 1–5.

[20] S. Markovich-Golan, S. Gannot, and I. Cohen, "A weighted multichannel Wiener filter for multiple sources scenarios," in *27th Convention of Electrical Electronics Engineers in Israel (IEEEI)*, 2012, pp. 1–5.