ANALYSIS OF DISTORTION IN AUDIO SIGNALS INTRODUCED BY MICROPHONE MOTION

Vladimir Tourbabin and Boaz Rafaely

Department of Electrical and Computer Engineering
Ben-Gurion University of the Negev, Beer-Sheva, Israel
{tourbabv,br}@bgu.ac.il

ABSTRACT

Signals recorded by microphones form the basis for a wide range of audio signal processing systems. In some applications, such as humanoid robots, the microphones may be moving while recording the audio signals. A common practice is to assume that the microphone is stationary within a short time frame. Although this assumption may be reasonable under some conditions, there is currently no theoretical framework that predicts the level of signal distortion due to motion as a function of system parameters. This paper presents such a framework, for linear and circular microphone motion, providing upper bounds on the motion-induced distortion, and showing that the dependence of this upper bound on motion speed, signal frequency, and time-frame duration, is linear. A simulation study of a humanoid robot rotating its head while recording a speech signal validates the theoretical results.

I. INTRODUCTION

Processing of signals from moving sensors is a topic with an increasing interest in the signal processing community. Moving sensors are encountered in many applications that naturally involve moving platforms, e.g. vehicle geolocation [1], towed underwater sonars [2], and mobile robots [3]. Furthermore, motion has been introduced to enhance performance in systems with stationary sensors, e.g. reduction of the side-lobe level in beamforming [4]–[6], improvement of spatial resolution in source localization [7]–[9], reduction of the reconstruction error in the sampling of spatial fields [10], and for rapid system identification [11]–[13]. The processing of signals from moving sensors is typically performed by dividing the signal into relatively short time frames, and assuming that the sensor is quasi-static, i.e. stationary within a single time frame [2], [3], [7], [9], [14]. However, if the effect of motion is significant, this assumption may introduce a significant error leading to poor performance.

In audio signal processing, with microphones as sensors, two types of motion are common (i) motion along a straight line (linear motion) [2], [4], [6], and (ii) circular motion [5], [9], [13]. The effect of sensor motion along a straight line on the measured signal is commonly studied in terms of the Doppler shift (see, for example, [15]). The motion generates a frequency-dependent shift in the signal spectrum by an amount that depends on the sensor velocity relative to the source. The effect of circular motion is similar, depending on the angular velocity [16].

Although the effect of microphone motion on the measured signal is generally known, a theoretical framework that expresses the magnitude of the mismatch between the signal acquired by a stationary and a moving sensor as a function of important system parameters, is not available. Moreover, for a given system, there are no guidelines assisting a system designer to predict whether the effect of microphone motion may be significant, or can be neglected in practice.

The validity of the quasi-static assumption for signals measured with moving microphones is studied in this paper. A theoretical framework is developed to provide bounds on the mismatch between signals recorded by stationary and moving microphones within a single time frame. The dependence of these bounds on frequency, microphone velocity, and time-frame length are derived for linear and circular microphone motion. The theoretical results are validated by a simulation study of a microphones on a mobile robot.

The remainder of the paper is organized as follows. Section II provides a review for signal models with linear and circular motion. Section III presents a theoretical analysis of the motion-induced distortion. A simulation-based experiment is presented in Section IV, and Section V concludes the paper.

II. SIGNAL MODEL - MOVING MICROPHONE

This section reviews representations for a signal recorded by a moving microphone, which form the basis for the results in the following section. Two different coordinate systems are used in this paper, related to the two types of motion. The first is the Cartesian coordinate system with a position in space indicated by $\mathbf{x} = (x,\, y,\, z)$, used to describe linear motion. The second is the spherical coordinate system [17] with a position indicated by $\mathbf{r} = (r, \theta, \phi)$, were r is the radial distance from the origin, θ is the elevation angle measured from the z axis, and ϕ is the azimuth measured from the x axis. The spherical coordinate system is used to describe circular motion. The two coordinate systems are related via $x = r \sin\theta \cos\phi$, $y = r \sin\theta \sin\phi$, and $z = r\cos\theta$.

Sound fields are commonly represented using a superposition of plane waves [18]. In this work it is assumed, for simplicity, that the sound field is composed of a single unit-amplitude plane wave. Although the results developed using this simplifying assumption can be extended to more complex sound fields, the study of this extension is proposed for a future work. The wave vector of the plane wave is given by

$$\mathbf{k} = \frac{\omega}{c}(\sin\theta_0 \cos\phi_0\,, \sin\theta_0 \sin\phi_0,\, \cos\theta_0), \tag{1}$$

where ω is the angular frequency, c is the speed of sound, and (θ_0, ϕ_0) denote the propagation direction. Under free field conditions, the sound field at a time t and a position $\mathbf{x}$, is given by [18]

$$p(t, \mathbf{x}) = e^{j\omega t - \mathbf{k}\cdot\mathbf{x}}, \tag{2}$$

where $j = \sqrt{-1}$. Using (2), the time-sampled signal measured by a stationary microphone positioned at the origin of the coordinate system can be written as

$$s(n) = p(n/f_s, \mathbf{0}) = e^{j\omega n/f_s}, \tag{3}$$

where $n \in \mathbb{Z}$ is the time index, f_s denotes the sampling frequency and $\mathbf{0} = (0, 0, 0)$ is the origin.

The time-sampled signal measured by a moving microphone can be defined in a similar manner. Consider a microphone moving at a constant speed along a straight line through the origin of the coordinate system in the direction of the wave propagation. The position of this microphone at a time t is given by $\mathbf{x}(t) = \mathbf{v}t$, with $\mathbf{v}$ denoting the velocity of the microphone. The signal measured by the microphone in this case can be written as

$$s_l(n) = p(n/f_s, \mathbf{v}n/f_s) = e^{j\omega(1-\beta)n/f_s}, \tag{4}$$

where $\beta = \frac{\|\mathbf{v}\|}{c}$ is the Mach number. The motion-induced frequency shift obtained in (4) is frequently referred to as the Doppler shift [19]. The shift is dependent on the direction of motion. In the case of motion in the direction of the wave propagation, the shift is maximal. The signal measured by a microphone moving in other directions can be represented using (4) but with β given by

$$\beta = \frac{\|\mathbf{v}\|}{c}\cos(\gamma) \tag{5}$$

with γ denoting the angle between the directions of motion of the microphone and the wave propagation.

Another type of motion considered in this paper is circular, i.e. the microphone moves along a circle with a constant angular velocity denoted by α, measured in radians per second. It is assumed, without loss of generality, that the circle lies in the xy plane with its center at the origin of the coordinate system. In this case, the microphone position at a time t is given by $\mathbf{x}(t) = r_a(\cos\alpha t, \sin\alpha t, 0)$, where r_a is the radius of the circle. The signal measured by the microphone for the case of a circular motion can therefore be written as

$$\begin{aligned} s(n) &= p\,(\,n/f_s, [\cos(\alpha n/f_s), \sin(\alpha n/f_s), 0]\,) \\ &= e^{j\omega[n/f_s - u_x\cos(\alpha n/f_s) - u_y \sin(\alpha n/f_s)]}, \end{aligned} \tag{6}$$

where $u_x = \frac{r_a}{c}\sin\theta_0\cos\phi_0$ and $u_y = \frac{r_a}{c}\sin\theta_0\sin\phi_0$.

The expressions in Eqs. (3)-(6) representing the signals measured by stationary and moving microphones, will be exploited in the next section for developing bounds for the motion-induced mismatch.

III. THEORETICAL FORMULATION OF MOTION-INDUCED DISTORTION

This section presents a theoretical formulation of the motion-induced distortion in signals measured by moving microphones, for both linear and circular motion. The microphone signal is assumed to be divided into time-frames with a duration of L samples. Then, processing is applied to individual time frames. Using Eqs. (3) and (4), the magnitude of the difference between the signals measured by stationary and moving microphones within a single time frame can be expressed as

$$\begin{aligned} \Delta(n, \beta, \omega) &= \left| e^{j\omega n/f_s} - e^{j\omega(1-\beta)n/f_s} \right| \\ &= \left| 1 - e^{-j\beta\omega n/f_s} \right| \\ &= 2\left| \sin\left(\frac{\beta\omega n}{2f_s}\right) \right|,\ n = 1, 2, ..., L. \end{aligned} \tag{7}$$

Recall that $\sin(\psi)$ increases monotonically in the range $\psi \in [0\ \frac{\pi}{2})$. Therefore, assuming $\frac{\beta\omega L}{2f_s} < \frac{\pi}{2}$, an upper bound on the motion-induced distortion in a single time frame is derived:

$$\begin{aligned} \Delta(n, \beta, \omega) &\le 2\left| \sin\left(\frac{\beta\omega L}{2f_s}\right) \right| \\ &= \Delta(L, \beta, \omega),\ n = 1, ..., L, \end{aligned} \tag{8}$$

This assumption is expected to hold as long as the speed of the microphone is significantly lower than the speed of sound, i.e. $\beta << 1$. Furthermore, with $\frac{\beta\omega L}{2f_s} << \frac{\pi}{2}$, applying $\sin(\psi) \approx \psi$, it approximately holds that

$$\Delta(L, \beta, \omega) \approx \beta\omega L/f_s, \tag{9}$$

which implies that this upper bound is linearly proportional to the microphone speed, the signal frequency, and the time-frame duration. This relation will be demonstrated by means of a numerical simulation in Section IV.

Similar to the case of a linear motion, in the case that the microphone is moving along a circle, the motion-induced distortion is expected to depend strongly on the angle between the wave propagation and the instantaneous direction of microphone motion. When this angle is zero, i.e. the direction of wave propagation and microphone motion is the same, the distortion is largest. Following the derivation of Eq. (6), at time $n = 0$ the microphone position is given by $(r, \theta, \phi) = (r_a, \pi/2, 0)$. Assuming further that the wave propagation direction is given by $(\theta_0, \phi_0) = (\pi/2, \pi/2)$, and substituting in Eq. (6), the signal at the microphone can be written as

$$s(n) = e^{j\omega[n/f_s - \sin(\alpha n/f_s) r_a/c]},\ n = 1, ..., L. \tag{10}$$

Next, using Eqs. (3) and (10), the difference between the signals measured by the stationary and moving microphones is given by

$$\begin{aligned}\Delta(n, \alpha, \omega) &= \left| e^{j\omega n/f_s} - e^{j\omega[n/f_s - \sin(\alpha n/f_s) r_a/c]} \right| \\ &= \left| 1 - e^{-j\omega \sin(\alpha n/f_s) r_a/c} \right|,\ n = 1, ..., L. \end{aligned} \tag{11}$$

Recall that the duration L/f_s of a typical time frame in speech processing is in the range of tens of milliseconds, while the angular velocity of a physical microphone is not expected to exceed several radians per second. It can therefore be assumed that $\frac{\alpha n}{f_s} << 1$. Substituting this assumption into (11) and applying $\sin(\psi) \approx \psi$, leads to

$$\begin{aligned}\Delta(n, \alpha, \omega) &\approx \left| 1 - e^{-j\omega\alpha n r_a/(c f_s)} \right| \\ &= 2 \left| \sin\left(\frac{\omega \alpha n r_a}{2 c f_s}\right) \right|,\ n = 1, ..., L. \end{aligned} \tag{12}$$

Assuming further that $\frac{\omega\alpha n r_a}{c f_s} < \frac{\pi}{2}$, an upper bound on the distortion induced by a circular microphone motion is obtained:

$$\begin{aligned}\Delta(n, \alpha, \omega) &\le 2 \left| \sin\left(\frac{\omega \alpha L r_a}{2 c f_s}\right) \right| \\ &= \Delta(L, \alpha, \omega),\ n = 1, ..., L. \end{aligned} \tag{13}$$

Finally, similar to the linear motion case, in the range of parameters where it holds that $\frac{\omega\alpha n r_a}{c f_s} << \frac{\pi}{2}$, the upper bound approximately equals to

$$\Delta(L, \alpha, \omega) \approx \frac{\omega \alpha L r_a}{c f_s}. \tag{14}$$

It should be emphasized that the small angle approximation mean that the displacement of the microphone from the initial position is small, and so the analysis is relevant to the starting position, i.e. $(r_a, \pi/2, 0)$. At this position the direction of motion is the same as the direction of the wave propagation, which implies that the distortion is largest. For other starting positions the distortion is expected to be smaller, and so Eq. (14) can indeed be considered as an upper bound.

In summary, similar to the case of a linear motion, the upper bound on the distortion induced by the circular motion is also linearly proportional to the angular velocity, the signal frequency, and the time-frame duration. These relations are further investigated using numerical simulations in the next section.

IV. SIMULATION STUDY - MICROPHONE ON A ROBOT

This section studies the distortion in speech signals measured by microphones mounted on the head of a humanoid robot. A far-field sound source was placed in free field, generating speech signals composed of 10 different sentences from the TIMIT speech database [20], sampled at $f_s = 10\,\text{kHz}$. The robot NAO [21] was positioned at the origin of the coordinate system, having two microphones mounted on its head, as illustrated in Fig. 1. The radial distance of the two microphones from the origin of the coordinate system is 3.7 cm and 6.1 cm, respectively. The signals recorded at the two microphones were computed by filtering the speech signals with the appropriate Head-Related transfer Function (HRTF) of the robot, calculated from the geometry of the head using the Boundary Element Method (BEM) [22].

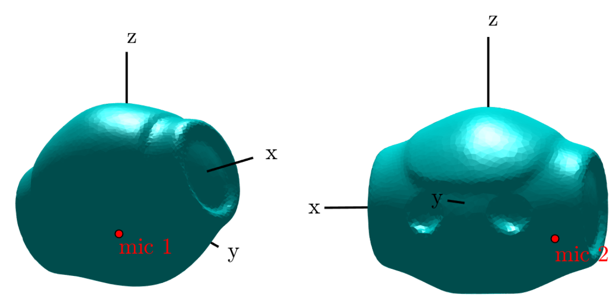

Fig. 1. The robot head with the two microphone positions at $r_a = 3.7\,\text{cm}$ (left) and $6.1\,\text{cm}$ (right).

The motion of the microphones was generated by rotation of the robot head. The filtering of the speech signal with the HRTFs was realized using the overlap-save method to allow modification of the filters in time, to account for head rotation. Two signals types were generated for each microphone: (i) the first signal simulating head rotation accurately by changing the overlap-save filters each sample according to the new head orientation, (ii) the second signal simulating a quasi-static head motion by changing the filter only once per time frame of L samples. The simulations were repeated with different time-frame lengths and angular velocities. The motion-induced distortion was computed from the normalized difference between the Short-Time Fourier Transform (STFT) of the two signals. This difference was then averaged over all time frames and for all 10 speech signals. Values of $128 \le L \le 2014$ samples and $45 \le \alpha \le 360$ deg/s were used for the time-frame length and the head rotation, respectively.

The resulting normalized distortion is plotted in Figs. 2, 3, and 4 as a function of the time-frame length, frequency, and angular velocity, respectively. The figures clearly show that the dependence of the distortion on all three parameters,

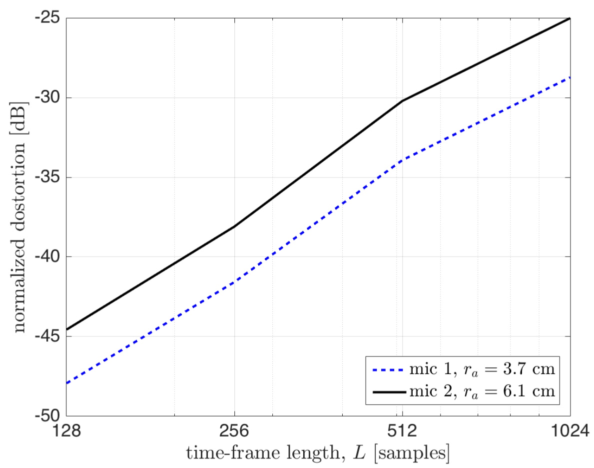

Fig. 2. Normalized motion-induced distortion versus time-frame length, L, at a frequency of 1 kHz and for an angular velocity of 90 deg/s.

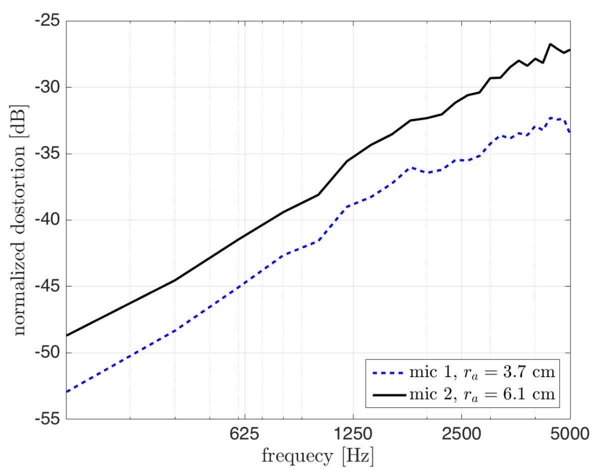

Fig. 3. Normalized motion-induced distortion versus frequency, for a rotating with an angular velocity of 90 deg/s, using $L = 256$ samples.

the time-frame length, frequency, and angular velocity is approximately linear, i.e. the distortion increases by about 6 dB per octave. This observation is in complete agreement with the results of Section III. A slight deviation from linearity can be observed in Fig. 3 at higher frequencies. This is believed to be due to violation of the assumption that led to the result in (14), which may be expected at high frequencies. In addition, note that in all three figures the distortion in the second microphone is larger by about 4 dB, which is approximately equal to the ratio between the radii of the two microphones, i.e. $20 \log_{10}(6.1/3.7) \approx 4.3$ dB. This observation is also in agreement with Eq. (14), that predicts

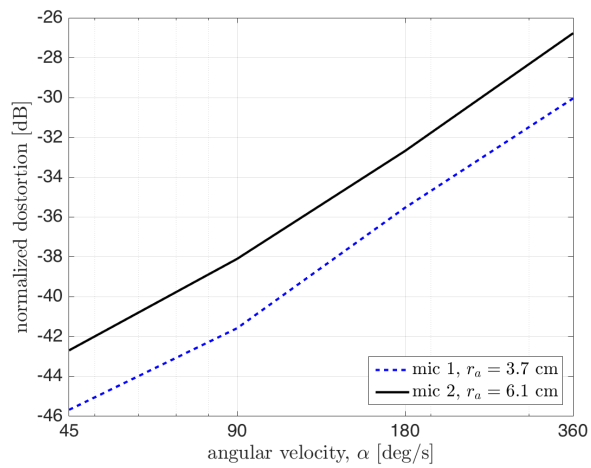

Fig. 4. Normalized motion-induced distortion versus angular velocity of the robot head, α, at a frequency of 1 kHz and for a time-frame length of $L = 256$ samples.

a linear dependence on the rotation radius r_a.

Figures 2-4 demonstrate that for a typical scenario of a moving humanoid robot recording a speech signal, the distortion is lower than -25 dB. This provides a support for the common practice in the literature, in which the effect of motion within a time frame is ignored. Nevertheless, as demonstrated above, the distortion is expected to increase linearly with frame length, frequency and speed of motion. For example, for some dereverberation algorithms, the required time-frame length may exceed 1s [23]. In this case the 6 dB/octave rule may predict an increase of 30 dB or more in the distortion level, and the motion-induced distortion within a time frame can no longer be ignored.

V. CONCLUSION

A theoretical model for the motion-induced distortion for a signal recorded by moving microphone has been presented. A moving humanoid robot has been studied as an example. It has been shown that motion within a single frame can be ignored when assuming typical values for the robot motion and time-frame duration employed for the audio signal processing. However, significant deviation from these typical values may lead to significant distortion, in which case modeling the motion within a time frame may be necessary.

ACKNOWLEDGEMENT

The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 609465.

VI. REFERENCES

[1] E. Tzoreff, B. Z. Bobrovsky, and A. J. Weiss, "Single receiver emitter geolocation based on signal periodicity

with oscillator instability," *IEEE Transactions on Signal Processing*, vol. 62, no. 6, pp. 1377–1385, March 2014.

[2] J. A. Fawcett, "Synthetic aperture processing for a towed array and a moving source," *The Journal of the Acoustical Society of America*, vol. 94, no. 5, pp. 2832–2837, 1993.

[3] J. M. Valin, F. Michaud, B. Hadjou, and J. Rouat, "Localization of simultaneous moving sound sources for mobile robot using a frequency-domain steered beamformer approach," in *2004 IEEE International Conference on Robotics and Automation, 2004. Proceedings. ICRA '04.*, April 2004, vol. 1, pp. 1033–1038.

[4] N. C. Yen and W. Carey, "Application of syntheticaperture processing to towedarray data," *The Journal of the Acoustical Society of America*, vol. 86, no. 2, pp. 754–765, 1989.

[5] A. Cigada, M. Lurati, F. Ripamonti, and M. Vanali, "Moving microphone arrays to reduce spatial aliasing in the beamforming technique: Theoretical background and numerical investigation," *The Journal of the Acoustical Society of America*, vol. 124, no. 6, pp. 3648–3658, 2008.

[6] E. Chang, "Irregular array motion and extended integration for the suppression of spatial aliasing in passive sonar," *The Journal of the Acoustical Society of America*, vol. 129, no. 2, pp. 765–773, 2011.

[7] F. Haber and M. Zoltowski, "Spatial spectrum estimation in a coherent signal environment using an array in motion," *IEEE Transactions on Antennas and Propagation*, vol. 34, no. 3, pp. 301–310, Mar 1986.

[8] S. Kim, D. H. Youn, and C. Lee, "Temporal domain processing for a synthetic aperture array," *IEEE Journal of Oceanic Engineering*, vol. 27, no. 2, pp. 322–327, Apr 2002.

[9] V. Tourbabin and B. Rafaely, "Direction of arrival estimation using microphone array processing for moving humanoid robots," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 11, pp. 2046–2058, Nov 2015.

[10] J. Unnikrishnan and M. Vetterli, "Sampling and reconstruction of spatial fields using mobile sensors," *IEEE Transactions on Signal Processing*, vol. 61, no. 9, pp. 2328–2340, May 2013.

[11] T. Ajdler, L. Sbaiz, and M. Vetterli, "Dynamic measurement of room impulse responses using a moving microphone," *The Journal of the Acoustical Society of America*, vol. 122, no. 3, pp. 1636–1645, 2007.

[12] C. Antweiler. and G. Enzner, "Perfect sequence lms for rapid acquisition of continuous-azimuth head related impulse responses," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, WASPAA '09.*, Oct 2009, pp. 281–284.

[13] N. Hahn and S. Spors, "Identification of dynamic acoustic systems by orthogonal expansion of time-variant impulse responses," in *Communications, Control and Signal Processing (ISCCSP), 2014 6th International Symposium on*, May 2014, pp. 161–164.

[14] J. Sheinvald, M. Wax, and A. J. Weiss, "Localization of multiple sources with moving arrays," in *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on*, Apr 1997, vol. 5, pp. 3521–3524.

[15] K.W. Lo and B.G. Ferguson, "Flight path estimation using frequency measurements from a wide aperture acoustic array," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 37, no. 2, pp. 685–694, Apr 2001.

[16] M. A. Poletti, "Series expansions of rotating two and three dimensional sound fields," *The Journal of the Acoustical Society of America*, vol. 128, no. 6, pp. 3363–3374, 2010.

[17] George B. Arfken, Hans J. Weber, and Frank E. Harris, "Chapter 16 - angular momentum," in *Mathematical Methods for Physicists (Seventh Edition)*, G. B. Arfken, H. J. Weber, and F. E. Harris, Eds., pp. 773 – 814. Academic Press, Boston, seventh edition edition, 2013.

[18] E. G. Williams, *Fourier Acoustics*, Academic Press, Cambridge, UK, 1999.

[19] F. Dunn, T. Rossing, W.M. Hartmann, D.M. Campbell, and N.H. Fletcher, *Springer Handbook of Acoustics*, EBL-Schweitzer. Springer New York, 2015.

[20] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallet, and N. S. Dahlgren, "DARPA TIMIT acoustic-phonetic continuous speech corpus," CD-ROM, 1993.

[21] Aldebaran Robotics, *NAO NEXT Gen H25 Datasheet*, December 2011.

[22] V. Tourbabin and B. Rafaely, "Theoretical framework for the optimization of microphone array configuration for humanoid robot audition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1803–1814, Dec 2014.

[23] H. Hacihabiboglu and Z. Cvetkovic, "Multichannel dereverberation theorems and robustness issues," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 2, pp. 676–689, Feb 2012.