

# Blind Separation of Speech Convolutive Mixtures via Time-Frequency Masking

Gustavo Fernandes Rodrigues  
 Centro Universitário de Belo Horizonte - UNI-BH  
 Belo Horizonte - MG - Brazil  
 gfernandes@acad.unibh.br  
 Leonardo Carneiro de Araújo  
 Universidade Federal de Belo Horizonte - UFMG  
 Belo Horizonte - MG - Brazil  
 leoca@cefala.org

Ana Cláudia Silva de Souza  
 Universidade Federal de Belo Horizonte - UFMG  
 Belo Horizonte - MG - Brazil  
 ssouza@cefala.org  
 Hani Camille Yehia  
 Universidade Federal de Belo Horizonte - UFMG  
 Belo Horizonte - MG - Brazil  
 hani@cefala.org

**Abstract**—An ideal binary masking, which specifies regions in the time-frequency domain whose concerned signal energy is greater than the interference signals is analyzed. The performance of the signal separation when these ideal binary masks are applied is evaluated. In the tests, these ideal masks remove almost all the interference from the other source of convolutive mixtures using simulated room impulses. A method for blind signal separation in the time-frequency domain using only the relative amplitude information of each time-frequency cluster cells is proposed. In reverberant environment the proposed method can not identify the clusters, but we may find out that the referring attenuation values of each source are concentrated in the extremities of the curve of relative attenuation histogram. Experimental results show that our proposed method can separate signals with little interference from the other source even in a real reverberant environment.

**Index Terms**—Blind Source Separation, Independent Component Analysis.

## I. INTRODUCTION

The purpose of this work is to study the blind source separation (BSS) for convolutive mixtures. Blind source separation problem is to extract independent sources from observed mixtures. Since the mixture is instantaneous, with no noise, there is an invertible system what allows us to retrieve the source signals up to a scale and a permutation indeterminacy [1], [2]. However, dealing with convolutive mixtures, the problem becomes more difficult. Several methods have been proposed for blind source separation of convolutive mixtures [3], [4], [5]. Recently, many blind source separation methods based on the time-frequency masking have been proposed [6], [7]. One of the first algorithms in the time-frequency domain, based on the sparseness assumption, for separation of speech signals in environments without reverberation, was the DUET (Unmixing Estimation Technique) [8], [6]. This method considers the extraction of signals sources through the application of a binary mask in the time-frequency domain. In the Duet algorithm [6], two parameters called relative amplitude (attenuation) and relative delay between two microphones are used to achieve the clusters and the binary masks. However, this method causes distortion and loud musical noise in the

retrieved signals because it results in a too much discontinuous zero-padded signal [7]. Another disadvantage of the Duet algorithm is the low performance when convolutive mixtures are used. In the present work, the sources are speech signals and the number of sources is known. Both anechoic and reverberant environments are considered. The performance of signals separation for a ideal binary masks is evaluated. A method for blind source separation, in the time-frequency domain, using only the relative amplitude information, is proposed, instead of using the relative amplitude and the relative delay parameters simultaneously.

## II. IDEAL MASKING IN THE TIME-FREQUENCY DOMAIN

The ideal masking, considered in this work, specifies the regions, in the time-frequency domain, whose concerned signal energy is greater than the interference signals energy. The ideal binary mask is a binary matrix, which assumes value one if the concerned signal energy is greater than the interference signals energy for one frequency component at a specific instant of time, and assumes value zero otherwise. When we apply an ideal mask (for a specific speech signal) to an instantaneous mixture of two speech signals, we observe that the extracted signal is perfectly audible, with good quality and with little interference from the other signal. One of the purposes of this work is to evaluate the extraction of each speech signal, using an ideal binary mask, from a mixture of speech signals in the presence of delays and reverberation.

The ideal mask for the signals ( $s_1$  and  $s_2$ ) can be specified in the time-frequency domain through the spectrograms:

$$s_1 \rightarrow S_1(w, t), \quad (1)$$

$$s_2 \rightarrow S_2(w, t), \quad (2)$$

where  $w$  is a frequency e  $t$  is the time instant of the speech frame. The ideal binary mask can be determined comparing the magnitude of the two spectrograms, as illustrated in Figure 1:

The ideal binary masks ( $M_1$  and  $M_2$ ) are defined as:

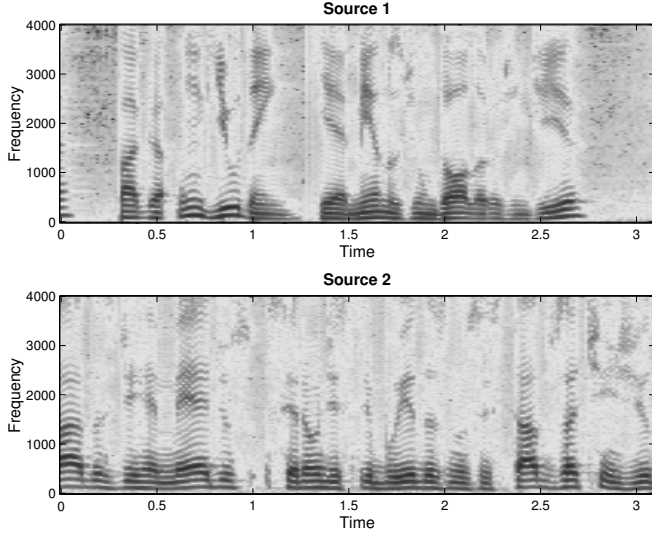


Fig. 1. Spectrograms of the two speech signals.

$$M_1 = 1, \quad \text{for } |S_1(w, t)| > |S_2(w, t)|, \quad (3)$$

$$M_2 = 1, \quad \text{for } |S_2(w, t)| > |S_1(w, t)|, \quad (4)$$

and zero otherwise. In these work, the ideal binary masks were obtained using two speech signals (with 16 kHz sampling rate) and a hanning window of 1024 samples length and 512 overlapping samples. These masks are used to extract the signals from mixtures in the time-frequency domain.

### III. RELATIVE AMPLITUDE AND RELATIVE DELAY ESTIMATION

Considering the measurements from the two sensors have got no reverberation, we may take the first mixture ( $x_1$ ) as the reference. In this case, we may consider the attenuation and delay parameters as null. The mixtures may be expressed as follows [6]:

$$x_1(t) = \sum_{j=1}^N s_j(t), \quad (5)$$

$$x_2(t) = \sum_{j=1}^N \alpha_j s_j(t - \delta_j), \quad (6)$$

where  $s_j$ ,  $j = 1, \dots, N$ , are the  $N$  sources;  $\delta_j$  their relative delays between the sensor and the  $j$ -th source; and  $\alpha_j$  is the relative attenuation for the  $j$ -th source.

Writing equations 5 and 6 in a matrix form, under the Fourier domain, we have:

$$\begin{bmatrix} \hat{x}_1(w, \tau) \\ \hat{x}_2(w, \tau) \end{bmatrix} = \begin{bmatrix} 1 & \dots & 1 \\ \alpha_1 e^{-iw\delta_1} & \dots & \alpha_N e^{-iw\delta_N} \end{bmatrix} \begin{bmatrix} \hat{s}_1(w, \tau) \\ \vdots \\ \hat{s}_1(w, \tau) \end{bmatrix} \quad (7)$$

Considering the sources are sparse and only one of them is active (a value different from zero) at a given frequency ( $w$ ) and time ( $\tau$ ) instant, Equation 7 can be simplified:

$$\begin{bmatrix} \hat{x}_1(w, \tau) \\ \hat{x}_2(w, \tau) \end{bmatrix} = \begin{bmatrix} 1 \\ \alpha_j e^{-iw\delta_j} \end{bmatrix} \hat{s}_j(w, \tau). \quad (8)$$

The relative attenuation and delay associated to a specific time-frequency bin may be expressed as:

$$(\alpha_j, \delta_j) = \left( \left| \frac{\hat{x}_2(w, \tau)}{\hat{x}_1(w, \tau)} \right|, \frac{1}{w} \angle \left( \frac{\hat{x}_2(w, \tau)}{\hat{x}_1(w, \tau)} \right) \right), \quad (9)$$

in which  $\angle(ae^{i\phi}) = \phi$ ,  $-\pi \leq \phi \leq \pi$ .

In the proposed method only the relative amplitude ( $\alpha_j$ ), obtained directly with the DFT components, is used to obtain the binary mask in the time-frequency domain.

### IV. PERFORMANCE MEASUREMENT

In order to analyze the results of sources separation, a method for performance evaluation based on distortion measures, is used and proposed in [9]. These distortion measures take into account interference from other sources, signal to noise ratio (SNR) and artifacts introduced by the algorithm of source separation, and define a performance criterion for each one of these contributions separately [1], [10]. The measurement criterion used in this work, considers that the only allowed distortions are time invariant gains.

The measures SDR (Source to Distortion Ratio), SAR (Source to Artifacts Ratio) and SIR (Sources to Interference Ratio) can be expressed by equations 10, 11 and 12:

$$SDR = 10 \log_{10} \frac{\|s_{\text{target}}\|^2}{\|e_{\text{interf}} + e_{\text{artef}}\|^2}, \quad (10)$$

$$SAR = 10 \log_{10} \frac{\|s_{\text{target}} + e_{\text{interf}}\|^2}{\|e_{\text{artef}}\|^2}, \quad (11)$$

$$SIR = 10 \log_{10} \frac{\|s_{\text{target}}\|^2}{\|e_{\text{interf}}\|^2}. \quad (12)$$

### V. ARTIFICIAL DATA

#### A. Instantaneous mixtures considering only different amplitudes between sources and sensors

In this item an instantaneous mixture of two speech signals was analyzed (with 16 kHz sampling rate), with different gains in relation to each source and sensor position. An ideal binary mask was obtained for the speech signals. Our objective is to analyze the signals separation using an ideal mask and using only information of the relative amplitude, without the use of phase information (relative delay). The performance of some blind source separation algorithms was evaluated, such as: Infomax, Fastica, Jade, Duet (using a hanning window of 1024 samples length and 512 overlapping samples) and an ICA algorithm (in the frequency domain) proposed by [11] (using the same hanning window). Table I shows the values of SDR, SIR and SAR for the analyzed mixture.

The mixture matrix used was:

$$A = \begin{bmatrix} 0.9 & 0.8 \\ 0.7 & 0.9 \end{bmatrix}. \quad (13)$$

TABLE I  
DISTORTION MEASUREMENTS FOR INSTANTANEOUS MIXTURES  
CONSIDERING DIFFERENT AMPLITUDES BETWEEN SOURCES AND  
SENSORS.

Method		SDR (dB)	SIR (dB)	SAR (dB)
Ideal Mask	$\tilde{y}_1$	11	38	11
	$\tilde{y}_2$	14	34	14
	Average	13	36	13
Relative Amplitude (Proposed Method)	$\tilde{y}_1$	7	33	7
	$\tilde{y}_2$	12	44	12
	Average	10	34	10
Duet	$\tilde{y}_1$	10	36	10
	$\tilde{y}_2$	13	32	13
	Average	12	34	12
ICA Frequency Domain	$\tilde{y}_1$	2	20	2
	$\tilde{y}_2$	6	15	6
	Average	4	18	4
Infomax	$\tilde{y}_1$	45	63	45
	$\tilde{y}_2$	32	57	32
	Average	39	60	39
FastICA	$\tilde{y}_1$	45	45	280
	$\tilde{y}_2$	46	46	275
	Average	46	46	278
Jade	$\tilde{y}_1$	48	48	275
	$\tilde{y}_2$	47	47	275
	Average	48	48	275

In Table I, we observe that Infomax, Fastica and Jade algorithms present better results than the other methods. The retrieved signals for these algorithms present smaller interference between the sources, that is, larger source to interference ratio (SIR). Figure 2 shows that the relative amplitude values of the source  $s_1$  are next to the value  $-0.2$  and the values of the source  $s_2$  are next to the value  $0.5$ . The DUET algorithm presents an average source to distortion ratio better than the proposed method ( $\overline{SDR} = 12$  dB) and average source to interference ratio similar to the proposed method ( $\overline{SIR} = 34$  dB). When the ideal mask is used it results in smaller distortion ( $\overline{SDR} = 13$  dB) and interference from other sources ( $\overline{SIR} = 36$  dB) than the other methods shown in Table I.

### B. Convolutional mixtures using simulated room impulse responses

The reverberation and sound absorption characteristics of a room can be simulated through the convolution of the room's impulse response and the original source signals. An acoustic impulse response can be precisely and efficiently simulated by the Image Method [12][13]. Figure 3 shows the room impulse response simulated from each source to each sensor.

The results using the ideal masks are better than the other methods in the time-frequency domain shown in Table II. In

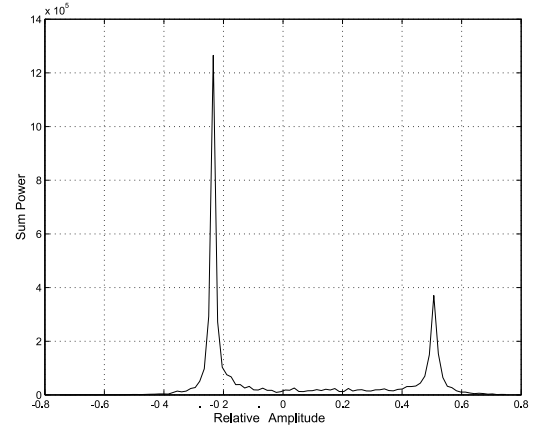


Fig. 2. Histogram of the relative attenuation values from the signal sources for instantaneous mixtures considering only different values of gains between sources and sensors.

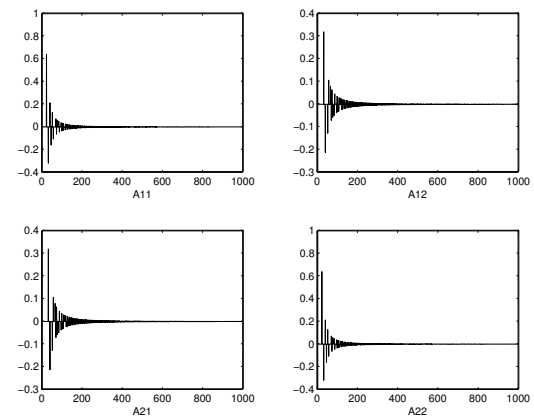


Fig. 3. Impulse Response simulated by the image method from each source to each sensor.

Figure 4 we do not perceive the presence of clusters, but we can verify through experimental tests that the referring attenuation values of  $s_1$  and  $s_2$  are concentrated in the extremities of the curve. The results obtained by the proposed method present a smaller distortion ( $\overline{SDR} = -17$  dB) and a smaller interference from the other source ( $\overline{SIR} = 24$  dB) compared to the Duet algorithm. The ideal mask presents better results than the methods shown in Table II.

## VI. OPEN FIELD RECORDED DATA

We evaluated the BSS algorithms in the time-frequency domain through experiments with real recorded speech signals. The speech signals were recorded in an open field, with low noise level and very low reverberation, as shown in the layout in Figure 5.

In Figure 6, we can perceive the presence of clusters around the relative amplitude values referring to sources  $s_1$  and  $s_2$ . The proposed method results in a smaller interference from the other source ( $\overline{SIR} = 12$  dB) than the other methods shown in Table III. In terms of distortion, the proposed method results

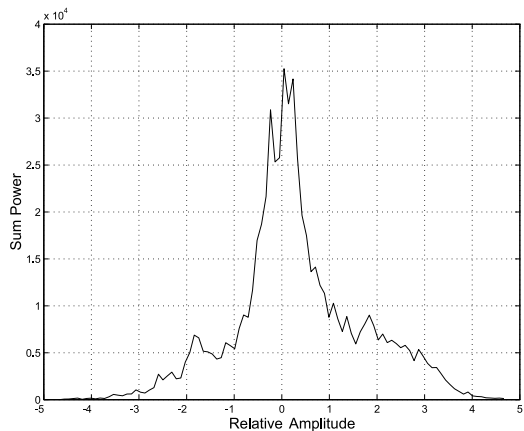


Fig. 4. Histogram of the relative attenuation values from the signal sources for convolutive mixtures using artificial room impulse responses.

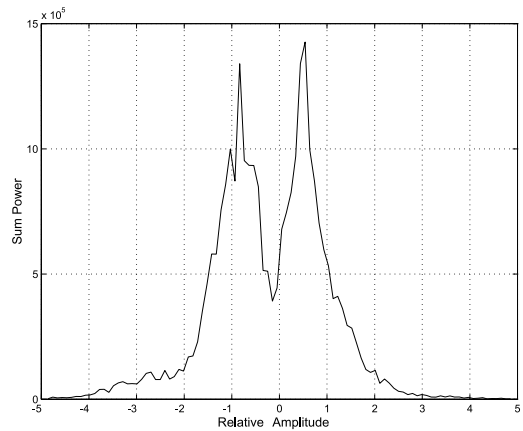


Fig. 6. Histogram of the relative attenuation values from the signal sources for real mixture recorded in an open field.

TABLE II

DISTORTION MEASUREMENTS FOR CONVOLUTIVE MIXTURES USING ARTIFICIAL ROOM IMPULSE RESPONSES.

Method		SDR (dB)	SIR (dB)	SAR (dB)
Ideal Mask	$\tilde{y}_1$	-18	32	-18
	$\tilde{y}_2$	-13	39	-13
	Average	-16	36	-16
Relative Amplitude (Proposed Method)	$\tilde{y}_1$	-18	23	-18
	$\tilde{y}_2$	-16	24	-16
	Average	-17	24	-17
Duet	$\tilde{y}_1$	-24	13	-24
	$\tilde{y}_2$	-50	-6	-42
	Average	-37	4	-33

TABLE III

DISTORTION MEASUREMENTS FOR REAL MIXTURE RECORDED IN AN OPEN FIELD

Method		SDR (dB)	SIR (dB)	SAR (dB)
Ideal Mask	$\tilde{y}_1$	-51	-6	-43
	$\tilde{y}_2$	-33	-23	-33
	Average	-42	-14	-38
Relative Amplitude (Proposed Method)	$\tilde{y}_1$	-51	16	-51
	$\tilde{y}_2$	-40	8	-40
	Average	-46	12	-46
Duet	$\tilde{y}_1$	-50	-4	-53
	$\tilde{y}_2$	-43	0	-39
	Average	-47	-2	-46

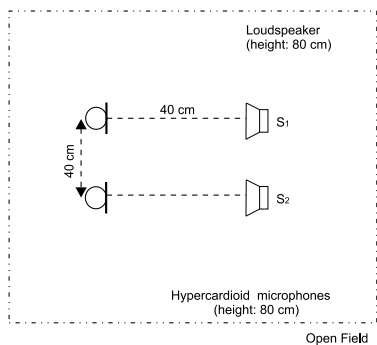


Fig. 5. Open field used in the experiments.

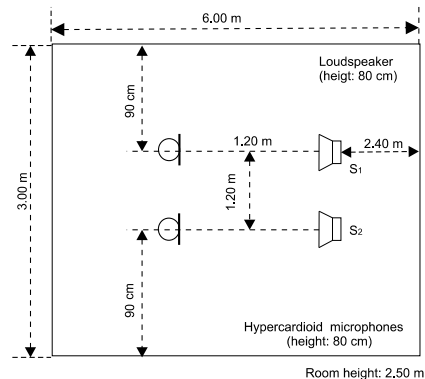


Fig. 7. Reverberant room used in experiments.

in similar average source to distortion ratio ( $\overline{SDR} = -46dB$ ) than the Duet algorithm.

VII. REAL ROOM RECORDED DATA

In this experiment, the speech signals were recorded in an real room. The sensor and source localizations are shown in Figure 7.

The relative amplitude peaks of the sources may not be determined in the histogram, illustrated in Figure 8. However, it is possible to identify that the extremities of the curve (in Figure 8) contains the amplitude values referring to sources  $s_1$  and  $s_2$ , with very low interference between them. The proposed method results in a better average source to interference

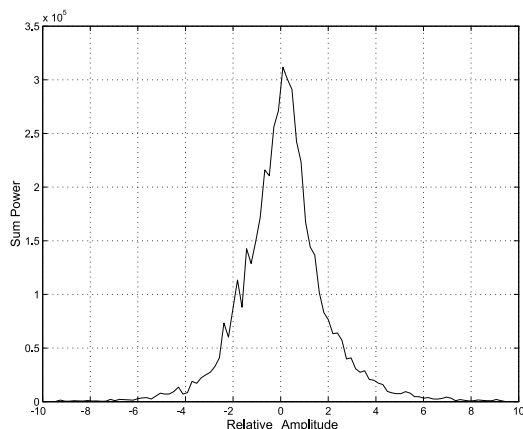


Fig. 8. Histogram of the relative attenuation values from the signal sources for real mixture recorded in an reverberant room.

TABLE IV

DISTORTION MEASUREMENTS FOR REAL MIXTURE RECORDED IN REAL ENVIRONMENT

Method		SDR (dB)	SIR (dB)	SAR (dB)
Ideal Mask	$\tilde{y}_1$	-46	34	-46
	$\tilde{y}_2$	-30	-13	-30
	Average	-38	11	-38
Relative Amplitude (Proposed Method)	$\tilde{y}_1$	-49	13	-49
	$\tilde{y}_2$	-44	12	-44
	Average	-47	13	-47
Duet	$\tilde{y}_1$	-48	5	-47
	$\tilde{y}_2$	-49	2	-47
	Average	-49	4	-47

ratio ( $\overline{STR} = 13dB$ ) than the other methods shown in Table IV.

## VIII. CONCLUSION

In this work a method for blind source separation based on the DUET algorithm was presented. The proposed method uses only the relative amplitude parameters to obtain the masks and to separate the mixtures in the time-frequency domain. This method reveals better performance in separating convolutive mixtures than the DUET algorithm. The experiments in real environments had shown the efficiency of the proposed method for convolutive mixtures. The proposed method causes distortion in the retrieved signals but it reduces significantly the interference from the source not desired. The efficiency of using an ideal binary mask for separation of instantaneous and convolutive mixtures was analyzed. The ideal mask suggested in this work presents good results for synthetic mixtures. The performance of these ideal masks in real recorded signals was not satisfactory. For future works we intend to improve the efficiency of the proposed method retrieving some information from the binary masks. Another suggestion consists on applying, in the separated signals, some

technique of processing, to reduce the distortions caused by the algorithm.

## REFERENCES

- [1] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Speech and Audio Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.
- [2] M. Babaie-Zadeh, C. Jutten, and A. Mansour, "Sparse ica via cluster-wise pca," *Neurocomputing*, vol. 69, pp. 1458–1466, 2006.
- [3] A. J. Bell and T. J. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural Computation*, vol. 7, no. 6, pp. 1129–1159, 1995.
- [4] S. Araki, R. Mukai, S. Makino, T. Nishikawa, and H. Saruwatari, "The fundamental limitation of frequency domain blind source separation for convolutive mixtures of speech," *IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 5, pp. 2737–2740, 2001.
- [5] K. Torkkola, "Blind separation of convolved sources based on information maximization," *IEEE Workshop on Neural Networks for Signal Processing, Kyoto*, pp. 423–432, september 1996.
- [6] 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, July 2004.
- [7] S. Araki, S. Makino, H. Sawada, and R. Mukai, "Underdetermined blind separation of convolutive mixtures of speech with directivity pattern based mask and ica," *Fifth International Conference on Independent Component Analysis and Blind Signal Separation*, pp. 898–905, 2004.
- [8] A. Jourjine, S. Rickard, and O. Yilmaz, "Blind separation of disjoint orthogonal signals: Demixing n sources from 2 mixtures," *In IEEE Conference on Acoustics, Speech, and Signal Processing (ICASSP2000)*, vol. 5, pp. 2985–2988, June 2000.
- [9] C. Févotte, R. Gribonval, and E. Vincent, "Bss eval toolbox user guide," IRISA, Rennes, França, Tech. Rep. 1706, 2005.
- [10] R. Gribonval, L. Benaroya, E. Vincent, and C. Févotte, "Proposals for performance measurement in source separation," *Proc. 4th Int. Symp. on Independent Component Anal. and Blind Signal Separation (ICA2003), Nara, Japan*, pp. 763–768, april 2003.
- [11] S. Ikeda and N. Murata, "A method of ica in time-frequency domain," *International conference on independent component analysis and signal separation*, pp. 365–371, 1999.
- [12] J. H. Rindel, "The use of computer modeling in room acoustics," *Journal of Vobroengineering*, vol. 3, no. 4, pp. 41–72, 2000.
- [13] L. Savioja, "Modeling techniques for virtual acoustics," Ph.D. dissertation, Helsinki University of Technology, august 2000.