Survey of Speech Enhancement Supported by a Bone Conduction Microphone

Ho Seon Shin, Hong-Goo Kang*, Tim Fingscheidt°*

Department of Electrical and Electronic Engineering, Yonsei University, 120-749 Seoul, Republic of Korea*
Institute for Communications Technology, Technische Universität Braunschweig, 38106 Braunschweig, Germany°
Email: signal@dsp.yonsei.ac.kr, hgkang@yonsei.ac.kr, fingscheidt@ifn.ing.tu-bs.de
Web: ee.yonsei.ac.kr*, www.ifn.ing.tu-bs.de°

Abstract

This paper gives an overview to speech enhancement algorithms using a bone-conducted (BC) microphone. Unlike conventional air-conducted (AC) microphones picking up particle movement in the air, the transmission channel of the BC microphone is only related to the human body. Therefore, voice communication with the BC microphone is not affected by ambient noise. However, since the high frequency components of the BC microphone signal are attenuated significantly due to transmission loss, additional signal processing techniques are needed to provide comfortable communication. This paper summarizes the idea of BC-based speech enhancement algorithms. It also provides some pros and cons of the approaches, which might be helpful for determining the direction for further research in the BC microphone-based speech enhancement field.

1 Introduction

As mobile wireless communication technology progresses, people can communicate with others anytime and anywhere. In consequence, voice communication users may be in various types of background noise environments, in which it is very difficult to provide a comfortable communication. Many researchers have investigated methods to improve the quality and intelligibility of target speech signals using a single, dual, or multiple microphones [1]. Though these approaches are effective for voice communications in certain environments, e.g., stationary noises, the performance improvement is more difficult in non-stationary noise cases. Please note that all algorithms mentioned above adopt air-conducted (AC) microphones, which is one of the reasons for this performance limit.

To overcome the limit of AC microphone-based speech enhancement, a new device, in our case, a bone-conducted (BC) microphone, is introduced. As depicted in Fig. 1, the conducting path of the BC microphone is different from the AC one. Since the origin of the speech signal obtained by the BC microphone is the vibration of human skull, as opposed to air propagation, the BC signal is not contaminated by background noise [2, 3]. Though the essential characteristic of the BC microphone is noise robustness, researchers pointed out that the high-frequency components of the BC speech are attenuated significantly due to channel loss [2-5]. Other researchers proved that the frequency characteristic, intelligibility, and sound quality of BC speech varied depending on the microphone location (top of the head, temple, etc.) [6, 7].

The BC microphone-based speech enhancement algorithms are classified into two approaches; whether the signals acquired by the BC microphone play a dominant role in the enhancement algorithms or not. At an early stage of research, the BC microphone was just regarded as a supplementary device for improving noise reduction performance of AC speech. For example, BC speech was used for distinguishing non-speech segments from speech segments [4]. Since the BC microphone is immune to background noise, BC speech is good for voice activity detection in a low SNR environment [8]. The BC microphone has also been used for detecting glottal source information that is closely linked to pitch [9]. Many speech coders that use a source-filter model need accurate pitch information. By attaching the BC microphone to the glottal source location, the quality of the synthesized speech can be enhanced even in very harsh noisy environments [5, 10]. Other approaches use the signals acquired by the BC microphone for enhancing the low frequency components [11, 12].

Utilizing the fact that the combination of AC and BC microphones generates enhanced results, Zheng et al. developed a prototype hardware device with AC and BC microphones [4]. It may be a turning point that the BC microphone is used as a main acquisition device, which of course needs additional signal processing techniques. For example, typically an on-line or off-line training stage is required to calibrate the characteristical differences between AC and BC microphones [2, 13]. Algorithmic details differ depending on the chosen calibration method.

This paper summarizes some BC microphone-based enhancement algorithms and discusses possible research directions for the future. At first, the characteristical differences between AC and BC microphones are briefly explained. Then, algorithms are sketched by categorizing into three approaches; (1) equalization, (2) analysis-and-synthesis, and (3) probabilistic approach. Please note that the algorithms addressed in this paper all belong to the class, where BC speech plays a dominant role in the enhancement process.

The rest of the paper is organized as follows. The characteristics of AC and BC microphones and the basic concept of BC speech enhancement are introduced in Section 2. Section 3 describes the methods and the pros and cons of the enhancement algorithms. Section 4 presents some concluding remarks.

2 Fundamentals of BC Speech and Its Enhancement Processing

2.1 Characteristics of BC Speech

As mentioned above, BC speech frequency characteristic differs from the AC speech one. Fig. 2 shows spectrograms of recorded speech signals acquired by an AC microphone (top figure) and by an in-ear type BC microphone (bottom figure). These synchronized signals

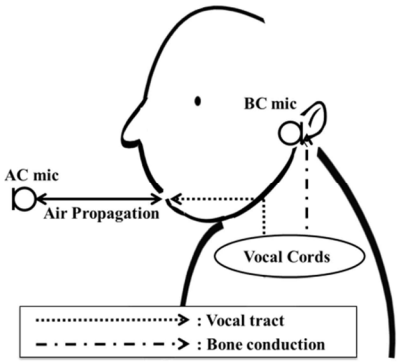

Figure 1: Conducting paths of AC and BC microphones (after [3])

are recorded in an office environment with interfering speaker near the talker. In the BC channel, we can observe strong low frequency components, but attenuated high frequency components, and a little amount of leakage noise. The pattern shows the general consensus of frequency characteristic of BC speech reported in state-of-the-art papers.

2.2 Fundamental Concept of BC Speech Enhancement Processing

Assuming an additive noise model, the AC and BC microphone-based system schematically depicted in Fig. 1 can be represented as follows [14]:

$$Y^{AC}(l,k) = S(l,k) + N(l,k) + U(l,k), \quad (1)$$

$$Y^{BC}(l,k) = H(l,k)S(l,k) + G(l,k)N(l,k) + W(l,k), \quad (2)$$

where l and k mean frame and frequency bin index, respectively. $S(l,k)$ denotes the spectrum of the clean speech signal at the AC microphone. $N(l,k)$ and $U(l,k)$ represent the spectra of background and sensor noise associated with the AC channel. $H(l,k)$ is the speech transfer function from the clean speech signal in the AC channel to the one in the BC channel. $G(l,k)$ and $W(l,k)$ denote the noise leakage function of the BC channel and the sensor noise associated with the BC channel, respectively.

As shown in Fig. 3, the AC clean speech $S(l,k)=S^{AC}(l,k)$ can be obtained by passing the BC clean speech, $S^{BC}(l,k)$ through an inverse speech transfer function, $H^{INV}(l,k)$.

$$S^{AC}(l,k) = H^{INV}(l,k)S^{BC}(l,k) = H^{-1}(l,k)S^{BC}(l,k). \quad (3)$$

Therefore, the inverse speech transfer function can be considered as a cascade of the BC inverse filter, $1/B(z)$, and the vocal tract filter with lip radiation, $V(z)$. $V(z)$ is commonly modeled as an autoregressive (AR) filter. However, it is very difficult to clearly determine the characteristic of the transfer function $B(z)$, because the system characteristic varies depending on the sensor types, locations, syllables, and users [3, 15]. In this paper, the inverse speech transfer function, $H^{INV}(l,k)$, is assumed to be stable relatively [16].

3 BC Speech Enhancement Overview

Fig. 4 illustrates the block diagram of a BC microphone-based speech enhancement algorithm. It consists of three processing steps; pre-processing, core enhancement module, and post-processing.

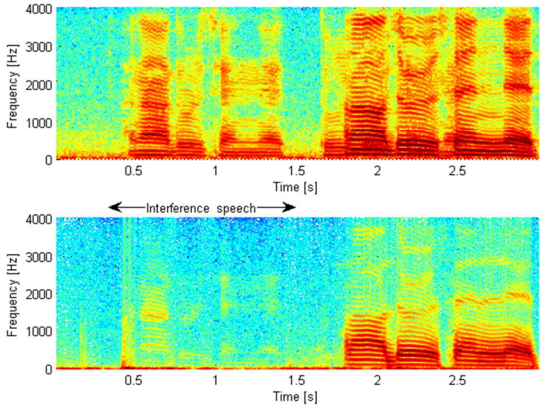

Figure 2: Spectrograms of AC speech (top) and BC speech (bottom) with interference speech (sampling frequency: 8 kHz)

To remove noise in the BC channel, general single-channel speech enhancement algorithms such as spectral subtraction and minimum mean square error (MMSE) estimation based methods were used for pre- or post-processing [2, 3, 17, 18].

In this section, BC speech enhancement approaches as a core processing are separately explained depending on the type of filtering methods: equalization, analysis-and-synthesis, and probabilistic approaches. More details on how to estimate the filter are described in the following subsections.

3.1 Equalization Approach

Shimamura and Tamiya designed a linear-phase impulse response filter in a training stage, which is calculated by taking an inverse discrete Fourier transform (IDFT) of the ratio of long-term AC and BC speech spectra [2]:

$$\hat{h}^{INV}(n) = \mathrm{IDFT}\left[\left|\overline{S}^{AC}(k)\right| / \left|\overline{S}^{BC}(k)\right|\right]. \quad (4)$$

Over-bars denote long-term spectra. Three types of AC and BC speech were utilized for training the filter such as five vowels, combined long sentences, and normal speech sentences of the same phrases. Although the quality of the reconstructed speech was improved overall, there was no consistency w.r.t. speakers and phrases. Since the filter order was set to be shorter than the frame size of long-term spectra, the output performance obtained by the convolution varied depending on the filter length.

Kondo et al. proposed the short-term DFT magnitude ratio-based method [3], which needs a training process for each user. The method estimated the equalization filter with a frame-by-frame basis approach, and then each filter was averaged to obtain a mean estimate:

$$\hat{H}^{INV}(k) = \mathrm{E}\left\{\left|S^{AC}(l,k)\right| / \left|S^{BC}(l,k)\right|\right\}, \quad (5)$$

with $\mathrm{E}\{\}$ being the expectation operator to carry out the averaging over the frames. A smoothing process was also applied to remove fluctuations across adjacent frequency bins.

Please note that the equalization filter only considers the magnitude ratio, thus the phase of the signal is kept the same as in the input signal. Experimental results also showed that the quality of the enhanced speech depended on the filter length and the smoothing length.

A least mean square (LMS) self-adaptive filter that does not need a training stage was introduced in [18]. With a

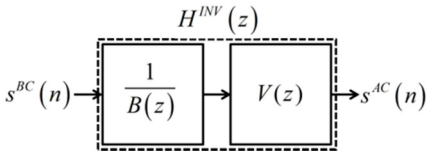

Figure 3: Block diagram of transforming BC speech into AC speech (after [2])

constraint to minimize the mean square error, $\mathrm{E}\{e^2(n)\}$ between AC and BC speech, the weighting vector, $\hat{h}^{INV}(n)$, is recursively updated as follows:

$$\hat{h}^{INV}(n) = \hat{h}^{INV}(n-1) + 2\mu e(n-1) y^{BC}(n-1), \quad (6)$$

$$\hat{H}^{INV}(z) = \sum_{n=0}^{M} \hat{h}^{INV}(n) z^{-n}. \quad (7)$$

where M and μ denote filter order and step size, respectively. The output weighting vector of the LMS algorithm results in the equalized system after being converged. Since $y^{AC}(n)$ and $y^{BC}(n)$ are almost the same as $s^{AC}(n)$ and $s^{BC}(n)$ in a high SNR environment, typically $e(n) = y^{AC}(n) - \hat{h}^{INV}(n) \otimes y^{BC}(n)$ is chosen, with $\otimes$ being the convolution operator.

Though such equalization approach is simple, the main limitation is that it is not able to deal with leakage noise of BC speech. In addition, although speaker-dependent algorithms with pre-training are appropriate to find optimized filters or parameters for each speaker [2, 3], they are not suitable for practical situations with unknown speakers. Besides, further analysis is required for determining the step size and convergence speed.

3.2 Analysis-and-Synthesis Approach

The AC and BC speech are considered to originate from the same source as shown in Fig. 1. Assuming that the BC filter, $B(z)$ shown in Fig. 3, can be modeled as an AR filter, a linear prediction (LP) filter has been introduced. The LP-based approach estimates $H^{INV}(l, k)$ using the LP coefficients of AC and BC speech [15, 19] or only of BC speech [5].

As the first step toward investigating the possibility of reconstructing BC speech, Thang et al. utilized an LP residual (related to the source) and LP coefficients (related to the vocal tract) [19]:

$$\hat{H}^{INV}(z) = \frac{Z\left[e^{AC}(n)\right] \cdot \sum_{i=0}^{Q} a^{BC}(i) z^{-i}}{Z\left[e^{BC}(n)\right] \cdot \sum_{i=0}^{P} a^{AC}(i) z^{-i}}, \quad (8)$$

where $a^{AC}(i)$ and $a^{BC}(i)$ represent the i-th LP coefficient of AC and BC speech, respectively. Each channel's LP orders are denoted by P and Q. Z[] denotes the Z transform. $e^{AC}(n)$ and $e^{BC}(n)$ represent the excitation from AC and BC speech, respectively. An extended algorithm reconstructs BC speech without utilizing AC speech information by predicting line spectral frequencies (LSF) via a recurrent neural network [15]. Since LSFs are relatively insensitive to quantization noise, the LSF-based approach is more appropriate than using LP coefficients for modeling.

Recently, Rahman and Shimamura proposed a method that did not exploit the spectral characteristic of AC speech but only utilized modified LP coefficients of BC

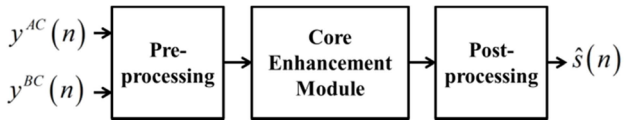

Figure 4: Block diagram of general BC speech enhancement system

speech [5]. They modified the first and third poles to have standard bandwidths, which eliminates an irregular broadening of the spectral peaks. By recomputed poles, the modified LP coefficients $\hat{a}^{BC}(i)$ are estimated back. In the synthesis step, the excitation from BC speech is used unmodified.

$$\hat{H}^{INV}(z) = \sum_{i=0}^{Q} a^{BC}(i) z^{-i} \Big/ \sum_{i=0}^{Q} \hat{a}^{BC}(i) z^{-i}. \quad (9)$$

The LP-based filtering converts muffled BC sound into much enhanced one. Please note that the approaches mentioned above do not concern BC channel noise. In fact, it is very hard to use the filters in a practical application. Although BC speech is robust to ambient noise, BC channel noise or physical noise, e.g., teeth clack, can occur. For blind BC restoration (i.e., no AC channel), the LP filter designed from mismatched LSF coefficients causes speech distortion.

3.3 Probabilistic Approach

The probabilistic approaches try to combine AC and BC signals by considering noise components as well. The approaches focus on how to reliably estimate the transfer function while removing the noises that have Gaussian assumptions. Liu et al. proposed the maximum likelihood estimation that minimizes the following cost function [17]:

$$R = \sum_{l=1}^{N} \left\{ \begin{array}{c} \left|Y^{AC}(l,k) - S(l,k)\right|^2 \Big/ \left(2\sigma_N{}^2\right) \\ + \left|Y^{BC}(l,k) - H(l,k)S(l,k)\right|^2 \Big/ \left(2\sigma_W{}^2\right) \end{array} \right\}. \quad (10)$$

N denotes the number of frames being utilized for estimation. The spectra of the background noise, $N(l,k)$, and the BC sensor noise, $W(l, k)$, were assumed as being distributed according to $\mathcal{N}(0,\sigma^2_N)$ and $\mathcal{N}(0,\sigma^2_W)$. By taking a partial derivative of R with respect to $H(l,k)$ and $S(l,k)$, the estimated transfer function and the estimated clean speech were derived. As a refinement to previous work [17], the explicit model including the noise leakage term was proposed [14]. $N(l,k)$, $U(l,k)$, and $W(l,k)$ are assumed to have a zero-mean Gaussian distribution. The noise leakage function, $G(l,k)$, and transfer function, $H(l,k)$, were also assumed as being Gaussian: $\mathcal{N}(\mu_G,\sigma^2_G)$ and $\mathcal{N}(\mu_H,\sigma^2_H)$. As a result, the enhanced speech was expressed as the weighted sum of the AC speech and the leakage-removed BC speech [14]. The advantage of these algorithms is that no prior training process is needed. In addition, these algorithms do not need to optimize estimators in a speaker-dependent manner. However, a weak point of these algorithms is that they do not have a speech model.

To solve this issue, Subramanya et al. incorporated a graphical speech model that has two states [20]. The speech model was assumed to be $P(S(l,k)|T(l)) \sim \mathcal{N}(S(l,k); 0,\sigma^2_S)$ where the state $T(l)=0$ means no speech, and $T(l)=1$

means speech in frame *l*. The graphical model is a probabilistic model for which a diagrammatic representation denotes the structure with nodes and links. It enables to solve complicated probabilistic models purely by the decomposition of joint distribution into a product of factors which depends only on a subset of the variables [21]. The joint distribution over all the variables was factorized on the subset by the graphical model, and then the posteriors were inferred. The MMSE estimate of the clean speech was obtained by the posterior on the state and the conditional expectation:

$$\hat{S}(l,k) = \sum_{t=0}^{1} \left\{ \begin{array}{l} P\left(T(l)=t \middle| Y^{AC}(l,k), Y^{BC}(l,k)\right) \\ \cdot \mathrm{E}\left\{S(l,k) | Y^{AC}(l,k), Y^{BC}(l,k), T(l)=t\right\} \end{array} \right\}, \quad (11)$$

where *t* is 0 (non-speech) or 1 (speech). With the help of the speech model, the speech distortion is reduced. But this algorithm using a single Gaussian model is not robust when training and test conditions are mismatched. In addition, this model does not guarantee good performance when any unexpected transition occurs.

Later on, Subramanya et al. proposed magnitude-normalized complex spectra for speech modeling with the mixture of Gaussians to reliably estimate the clean speech [13]. They considered temporal changes by introducing a constraint towards the state of the previous frame. The constraint enabled the transition between frames smoothly. The algorithm showed that the performance of the speech model trained by multiple speakers was fairly comparable to that by a single speaker. To minimize speech distortion, the probabilistic approaches needed to specify more than two states with the use of huge databases.

4 Concluding Remarks

BC microphone based speech enhancement algorithms have seen growing interest due to insensitivity to background noise. Though numerous algorithms have been proposed to improve the intelligibility and perceptual quality, it is still a challenging task. Much consideration needs to be taken for the issues on minimizing speaker dependency, reconstructing harmonic components, and reducing complexity.

References

[1] Philipos C. Loizou, *Speech Enhancement: Theory and Practice*, CRC Press, 1 ed., Jun. 2007.

[2] Tetsuya Shimamura and Toshiki Tamiya, "A Reconstruction Filter for Bone-Conducted Speech," *Proc. of the Midwest Symposium on Circuits and Systems (MWSCAS)*, *Covington*, *Kentucky*, *USA*, vol. 2, pp. 1847-1850, Aug. 2005.

[3] Kazuhiro Kondo, Tomoe Fujita, and Kiyoshi Nakagawa, "On Equalization of Bone Conducted Speech for Improved Speech Quality," *Proc. of the IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*, *Vancouver*, *British Columbia*, *Canada*, pp. 426-431, Aug. 2006.

[4] Yanli Zheng, Zicheng Liu, Zhengyou Zhang, Mike Sinclair, Jasha Droppo, Li Deng, Alex Acero, and Xuedong Huang, "Air- and Bone-Conductive Integrated Microphones for Robust Speech Detection and Enhancement," *Proc. of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, *St. Thomas*, *United States Virgin Islands*, pp. 249-254, Nov. 2003.

[5] M. Shahidur Rahman and Tetsuya Shimamura, "Intelligibility Enhancement of Bone Conducted Speech by an Analysis-Synthesis Method," *Proc. of the IEEE International Midwest Symposium on Circuits and Systems (MWSCAS)*, *Seoul*, *Korea*, pp. 1-4, Aug. 2011.

[6] Raymond M. Stanley and Bruce N. Walker, "Intelligibility of Bone-Conducted Speech at Different Locations Compared to Air-Conducted Speech," *Proc. of the Human Factors and Ergonomics Society (HFES) Annual Meeting*, *San Antonio*, *Texas*, *USA*, pp. 1086-1090, Oct. 2009.

[7] Maranda McBride, Phuong Tran, Tomasz Letowski, and Rafael Patrick, "The Effect of Bone Conduction Microphone Locations on Speech Intelligibility and Sound Quality," *Journal of Applied Ergonomics*, vol. 42, issue. 3, pp. 495-502, Mar. 2011.

[8] Mingzhe Zhu, Hongbing Ji, Falong Luo, and Wei Chen, "A Robust Speech Enhancement Scheme on the Basis of Bone-Conductive Microphones," *Proc. of the International Workshop on Signal Design and Its Applications in Communications (IWSDA)*, *Chengdu*, *China*, pp. 353-355, Sept. 2007.

[9] M. Shahidur Rahman and Tetsuya Shimamura, "Pitch Characteristics of Bone Conducted Speech," *Proc. of the European Signal Processing Conference (EUSIPCO)*, *Aalborg*, *Denmark*, pp. 795-799, Aug. 2010.

[10] Eiji Uchino, Kazuaki Yano, and Tadahiro Azetsu, "A Self-Organizing Map With Twin Units Capable of Describing a Nonlinear Input-Output Relation Applied to Speech Code Vector Mapping," *Journal of Information Sciences*, vol 177, issue 21, pp. 4634-4644, Nov. 2007.

[11] Thomas F. Quatieri, Kevin Brady, Dave Messing, Joseph P. Campbell, William M. Campbell, Michael S. Brandstein, Clifford J. Weinstein, John D. Tardelli, and Paul D. Gatewood, "Exploiting Nonacoustic Sensors for Speech Coding," *IEEE Trans. on Audio, Speech, and Language Processing (ASLP)*, vol. 14, no. 2, pp. 533-544, Mar. 2006.

[12] M. Shahidur Rahman, Atanu Saha, and Tetsuya Shimamura, "Low-Frequency Band Noise Suppression Using Bone Conducted Speech," *Proc. of the IEEE Pacific Rim Conference on Communications, Computers and Signal Processing (PacRim)*, *Victoria*, *British Columbia*, *Canada*, pp. 520-525, Aug. 2011.

[13] Amarnag Subramanya, Zhengyou Zhang, Zicheng Liu, and Alex Acero, "Multisensory Processing for Speech Enhancement and Magnitude-Normalized Spectra for Speech Modeling," *Journal of Speech Communication*, vol. 50, issue 3, pp. 228-243, Mar. 2008.

[14] Zicheng Liu, Amarnag Subramanya, Zhengyou Zhang, Jasha Droppo, and Alex Acero, "Leakage Model and Teeth Clack Removal for Air- and Bone-Conductive Integrated Microphones," *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, *Philadelphia*, *Pennsylvania*, *USA*, pp. 1093-1096, Mar. 2005.

[15] Tat Vu Thang, Masashi Unoki, and Masato Akagi, "An LP-Based Blind Model for Restoring Bone-Conducted Speech," *Proc. of the International Conference on Communications and Electronics (ICCE)*, *Hoi An*, *Vietnam*, pp. 212-217, Jun. 2008.

[16] Jingna Yu, Luyong Zhang, and Zheng Zhou, "A Novel Voice Collection Scheme Based on Bone-Conduction," *Proc. of the IEEE International Symposium on Communications and Information Technology (ISCIT)*, *Beijing*, *China*, pp. 1164-1168, Oct. 2005.

[17] Zicheng Liu, Zhengyou Zhang, Alex Acero, Jasha Droppo, and Xuedong Huang, "Direct Filtering for Air and Bone Conductive Microphones," *Proc. of the IEEE Workshop on Multimedia Signal Processing (MMSP)*, *Siena*, *Tuscany*, *Italy*, pp. 363-366, Sept. 2004.

[18] Tetsuya Shimamura, Jun'ichiro Mamiya, and Toshiki Tamiya, "Improving Bone-Conducted Speech Quality via Neural Network," *Proc. of the IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*, *Vancouver*, *British Columbia*, *Canada*, pp. 628-632, Aug. 2006.

[19] Tat Vu Thang, Kenji Kimura, Masashi Unoki, and Masato Akagi, "A Study on Restoration of Bone-Conducted Speech With MTF-Based and LP-Based Models," *Journal of Signal Processing*, vol. 10, no. 6, pp. 407-417, Nov. 2006.

[20] Amarnag Subramanya, Zhengyou Zhang, Zicheng Liu, Jasha Droppo, and Alex Acero, "A Graphical Model for Multi-Sensory Speech Processing in Air-and-Bone Conductive Microphones," *Proc. of the Annual Conference of the International Speech Communication Association (Interspeech)*, *Lisbon*, *Portugal*, pp. 2361-2364, Sept. 2005.

[21] Christopher M. Bishop, *Pattern Recognition and Machine Learning*, Springer, 1 ed., Aug. 2006