MULTISENSORY SPEECH ENHANCEMENT IN NOISY ENVIRONMENTS USING BONE-CONDUCTED AND AIR-CONDUCTED MICROPHONES

Mingzi Li ,Israel Cohen and Saman Mousazadeh

Department of Electrical Engineering, Technion - Israel Institute of Technology
Technion City, Haifa 32000, Israel
daming1225@gmail.com, icohen@ee.technion.ac.il, smzadeh@tx.technion.ac.il

ABSTRACT

In this paper, we propose a speech enhancement algorithm for estimating the clean speech using samples of air-conducted and bone-conducted speech signals. We introduce a model in a supervised learning framework by approximating a mapping from concatenation of noisy air-conducted and bone-conducted speech to clean speech in the short time Fourier transform domain. Two function extension schemes are utilized: geometric harmonics and Laplacian pyramid. Performances obtained from the two schemes are evaluated and compared in terms of spectrograms and log spectral distance measures.

Index Terms— Multisensory, bone-conducted microphone, geometric harmonics, Laplacian pyramid.

1. INTRODUCTION

Bone-conducted microphone, being less sensitive to surrounding noise, can work complementarily to a regular air-conducted microphone in noisy environments. However, high frequency components of bone-conducted microphone signals are attenuated significantly due to transmission loss. Hence the quality of speech signals acquired by a bone-conducted microphone is relatively low. We wish to enhance the speech signal by combining both air-conducted and bone-conducted microphones and producing high quality speech signal with low background noise.

Existing multisensory speech enhancement methods can be classified into two main categories, according to the role of the bone-conducted microphone [13]: In the first category, the bone-conducted microphone is used as a supplementary sensor, whereas in the the second one the bone-conducted microphone is used as a dominant sensor. Implementation in the first category relies on the accuracy of a voice activity detection or pitch extraction facilitated by the bone-conducted microphone. When the bone-conducted microphone is exploited as the main acquisition sensor, algorithms are related to either equalization, analysis-and-synthesize, or probabilistic approaches. Generally, equalization approaches reconstruct the clean speech through an finite impulse response (FIR) filter of the pre-enhanced air and bone conducted speech spectra ratio [7]. Similar to equalization approaches, analysis-and-synthesize methods require a speech enhancement procedure priorly, whereas the reconstruction filter is the ratio of the linear prediction model of both speech signals [15]. With Gaussian noise hypothesis in air and bone channels, probabilistic approaches can be conducted either in a maximum likelihood sense [8] or a minimum mean square error sense [12] . According to various assumptions of speech and noise models, more complicated algorithms have been proposed as well, such as non-linear information fusion [4], model-based fusion [6], and bandwidth extension [11].

In this paper, the clean speech is restored through a family of functions named geometric harmonics, i.e., eigenfunction extensions of a Gaussian kernel. Geometric harmonics can describe the geometry of high dimensional data and extend these descriptions to new data points, as well as the function defined on the data. In our case, the high dimensional data is defined by concatenation of air-conducted and bone-conducted speech in the short time Fourier transform (STFT) domain. A nonlinear mapping to the STFT of the clean speech defined on the new concatenation of speech signals can be obtained by a linear combination of geometric harmonics.

Application of geometric harmonics requires a careful adjustment of the correct extension scale and condition number. As a result, a multi-scale Laplacian pyramid extension is utilized to avoid this scale tuning. Based on the kernel regression scheme, Laplacian pyramid extension approximates the residual of the previous representation via a series of Gaussian kernels.

Experiments are conducted on simulated air-conducted and bone-conducted speech in interfering speaker and Gaussian noise environments. Geometric methods provide a consistent reconstruction of speech spectrograms in a variety of noise levels and categories. Log spectral distance results obtained using the proposed methods are compared to an existing probabilistic approach. We show that the Laplacian

This research was supported by the Israel Science Foundation (grant no. 1130/11).

pyramid method outperforms the other methods.

The structure of this paper is as follows: In Section 2, we fomulate the model and describe the geometric harmonics and Laplacian pyramid. In Section 3, we present experimental results that demonstrate the advantage of the Laplacian pyramid method over the competing methods. Finally, in Section 4 we summarize the paper and present future directions.

2. GEOMETRIC METHODS

2.1. Model Fomulation

Based on an additive noise model, the noisy air-conducted and bone-conducted speech in the STFT domain can be represented as follows:

$$\begin{aligned} \boldsymbol{y}_a\left(k,l\right) &= \boldsymbol{x}\left(k,l\right) + \boldsymbol{n}\left(k,l\right) \\ \boldsymbol{y}_b\left(k,l\right) &= f\left(\boldsymbol{x}\left(k,l\right)\right). \end{aligned} \tag{1}$$

where $\boldsymbol{x}\left(k,l\right)$, $\boldsymbol{n}\left(k,l\right)$ are the clean speech and noise respectively, k and l are the frequency bin and time indices respectively, and f is the nonlinear mapping from the clean speech to the bone-conducted speech.

Our goal now is to learn the inverse of f and extend it to new bone-conducted speech signals. For a more reasonable experiment setup, we concatenate the STFT of noisy air-conducted (AC) and bone-conducted (BC) speech together and define them as the training set with m data points in R^n

$$S = \left[\boldsymbol{y}_a\left(k,l\right) \;\; \boldsymbol{y}_b\left(k,l\right)\right]_{k=1\ldots n, l=1\ldots m}. \tag{2}$$

where n is the total number of frequency bins and m is the total number of frames. As a result, we need to learn the high dimensional nonlinear function f and extend it in the test data for the clean speech reconstruction.

Geometric extension methods such as geometric harmonics and Laplacian pyramid can be used for high dimensional nonlinear empirical function extension. Generally, those methods rely on a Gassuian kernel to measure the distance between different data points.

The kernel $k : \bar{S} \times \bar{S} \rightarrow R$

• is symmetric, i.e., $k\left(\bar{x},\bar{y}\right) = k\left(\bar{y},\bar{x}\right)$ for all $\bar{x}$ and $\bar{y}$ in $\bar{S}$.

• is positive semi-definite, i.e., for any $m \geq 1$ and any choice of real numbers $\alpha_1, ..., \alpha_m$ and of points $\bar{x}_1, ..., \bar{x}_m$ in $\bar{S}$, we have

$$\sum_{i=1}^{m}\sum_{j=1}^{m} \alpha_i \alpha_j k\left(\bar{x}_i, \bar{x}_j\right) \geq 0. \tag{3}$$

• is bounded on $\bar{S} \times \bar{S}$ by a number $M > 0$, i.e., for any points in $\bar{S}$, we have

$$k\left(\bar{x}_i, \bar{x}_j\right) \geq M. \tag{4}$$

In the training procedure, we first concatenate the noisy AC speech $\boldsymbol{y}_a$ and BC speech $\boldsymbol{y}_b$ as $\boldsymbol{y}$ and build a Guassian kernel $\boldsymbol{K}$ between every two data points. This affinity matrix between all the training data pairs are then eigendecomposed, where the weights $\boldsymbol{w}$ of the eigenspace description are saved as the inner product of clean speech $\boldsymbol{x}$ and the eigenvectors $\boldsymbol{\varphi}_l\left(\boldsymbol{y}\right)$.

We concatenate a new speech captured by the air-conducted and bone-conducted microphone simultaneously in a noisy environment as the test data $\bar{\boldsymbol{y}}$. The extension of the function f mapping $\bar{\boldsymbol{y}}$ to $\bar{\boldsymbol{x}}$ is built on the formulation of the geometric harmonics $\boldsymbol{\varphi}_l\left(\bar{\boldsymbol{y}}\right)$ which average over the eigenfunctions of the new affinity matrix between the test speech and all the training speech signals. Finally, the clean speech $\bar{\boldsymbol{x}}$ is estimated by a linear combination of the weighted geometric harmonics.

In the Laplacian pyramid method, we first build a Gaussian kernel $\boldsymbol{K}_0$ between a test speech signal $\bar{\boldsymbol{y}}$ and all the training data $\boldsymbol{y}$ with an initial scale σ_0. Then a preliminary estimation $\boldsymbol{s}_0$ is obtained by the superposition of all the training data points $\boldsymbol{x}$ weighted by this initial kernel. We acquire a residual $\boldsymbol{d}_1$ by subtracting this first estimation $\boldsymbol{s}_0$ from the original function $\boldsymbol{x}$. The second step estimation $\boldsymbol{s}_1$ is obtained by a superposition of the residual $\boldsymbol{d}_1$ weighted by a finer kernel $\boldsymbol{K}_1$. The overall result is the sum of the estimations in a couple of iterations.

2.2. Geometric harmonics

Geometric Harmonics (GH) [3] is derived from the Nystr$\ddot{o}$m extension, which has been widely used in partial differential solvers, machine learning and spectral graph theory to subsample large data sets [1]. It can also be applied to extend functions from a training set to accommodate the arrival of new samples. The extension of function f defined on a set S to a larger set $\bar{S}$ is based on the construction of a specific family of functions termed as geometric harmonics.

The geometric harmonics are defined by the extended eigenfunction ψ_j (when $\lambda_j \neq 0$)

$$\psi_j\left(\bar{x}\right) = \frac{1}{\lambda_j} \int_S k\left(\bar{x}, y\right) \psi_j\left(y\right) d\mu\left(y\right). \tag{5}$$

These functions constitute a generalization of the prolate spheroidal wave functions of Slepian in the sense that they are optimally concentrated on S. GH algorithm is described in Table 1.

The drawbacks of geometric harmonics may come from three pespectives: 1) It needs parameters tuning, i.e., kernel scale and condition number of remaining eigenvectors; 2) The function is extended in the eigenvector space rather than the original space, which may not decribe the high dimensional data well; 3) The complexity of the function may influence the extesion range, thus will lead to an inaccurate extension result for complicated function.

Laplacian pyramid method can circumvent these problems via iterating the kernel on the residual of previous result

Table 1: GH algorithm

```
Algorithm
Input: $X \in R^{n \times m}, f(x) \in R^{n/2}, \varepsilon, l, y$
Output: Extended function $f(y)$
1. Build a Gaussian kernel $k = e^{-\frac{\|x_i - x_j\|^2}{2\varepsilon}}$.
2. Compute the set of eigenvalues $\varphi_l(x)$ for this kernel
   $\lambda_l \varphi_l(x_i) = \sum_{x_j \in S} e^{-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\varepsilon}} \varphi_l(x_j),\ x_i \in S.$
3. Extend the eigenvalues $\varphi_l(y)$ via geometric harmonics
   $\varphi_l(y) = \frac{1}{\lambda_l} \sum_{x_j \in S} e^{-\frac{\|y - \mathbf{x}_j\|^2}{2\varepsilon}} \varphi_l(x_j),\ y \in \bar{S}.$
4. Extend function via linear combination of the basis
   $f(y) = \sum_{\lambda_l \geq \delta\lambda_0} \langle \varphi_l, f \rangle \tilde{\varphi}_l(y).$
```

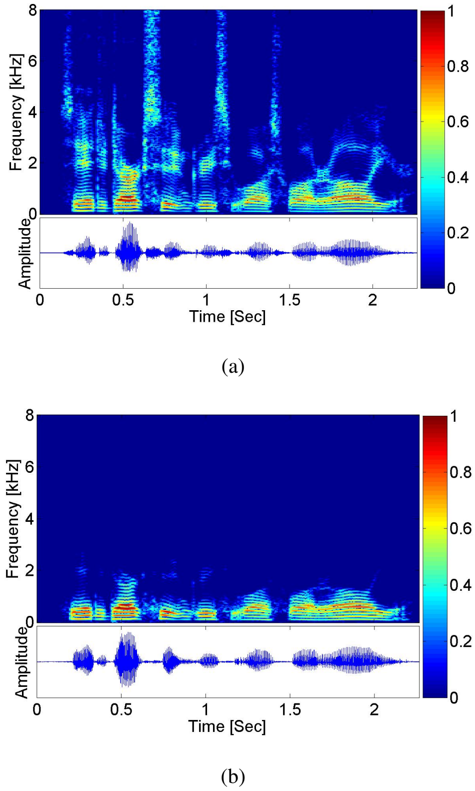
Table 2: LP algorithm

```
Algorithm
Input: $X \in R^{n \times m}, f(x) \in R^{n/2}, \sigma_0, y$
Output: Extended function $f(y)$
Initialize: Build a normalized Gaussian kernel $K_0$.
Approximate the function $s_0(y) = \sum_{i=1}^{n} k_0(x_i, y) f(x_i)$
Repeat
    1. Build the Gaussian kernel $W_l$ with decreasing scale
       $W_l = w_l(x_i, x_j) = e^{-\|x_i - x_j\|^2 / \left(\frac{\sigma_0}{2^l}\right)}$.
    2. Normalize the kernel by the row sum
       $K_l = k_l(x_i, x_j) = q_l^{-1}(x_i)\, w_l(x_i, x_j)$
         where $q_l(x_i) = \sum_j w_l(x_i, x_j)$.
    2. Compute the residual $d_l = f - \sum_{i=0}^{l-1} s_i$.
    3. Approximate the function $s_l(y) = \sum_{i=1}^{n} k_l(x_i, y) d_l(x_i)$.
End: Extend function via summation $f(y) = \sum_{k \leq l} s_k(y)$.
```

in the original space.

2.3. Laplacian pyramid

Laplacian pyramid (LP) is originally an image encoding technique where local operators of distinct scales, but of identical shapes, serve as the basis functions [9]. It can describe data iteratively via Gaussian kernels of decreasing widths. At a given scale l, the Laplacian pyramid is used for constructing a coarse representation of a function f . The algorithm is iterated by approximating the residual using a Gaussian kernel of a finer scale. The approximation function at a given scale can be extended to new data points. The overall extension is the sum of approximations in different scales. The LP algorithm is described in Table 2.



(a)

(b)

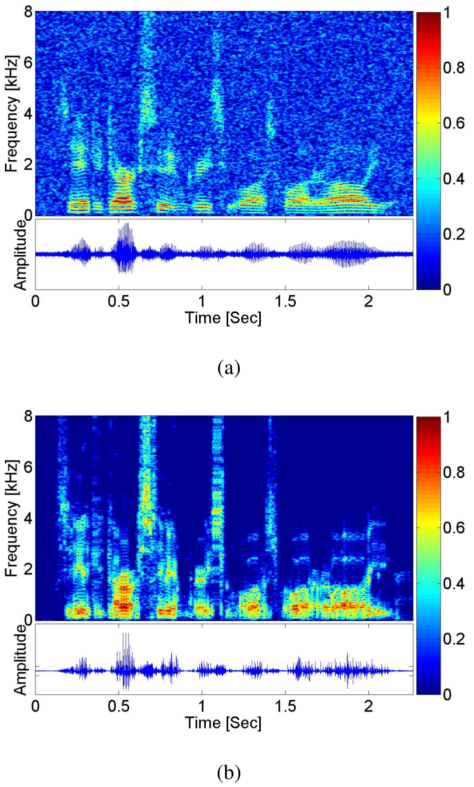
Fig. 1: Speech spectrograms and waveforms: (a) Clean signal; (b) bone-conducted signal (LSD=2.1317).

3. EXPERIMENTAL RESULTS

In this section, we present simulation results which demonstrate the performances of geometric harmonics and Laplacian pyramid compared to an existing probabilistic approach. Ten utterances of speech signals are taken from the TIMIT database [5]. The sampling frequency is f_s = 16 kHz. The STFT window is a Hanning window of length N = 256 and the overlap between two successive STFT frames is 50 percent. The BC speech signals are obtained by low pass filtering the AC speech signals, where the pass–band cutoff frequency is 300 Hz and stop–band cutoff frequency is 3 kHz. Noisy AC speech signals are generated by adding Gaussian noise and interfering speaker. We choose $\varepsilon = 10$ and $\sigma_0 = 100$.

The spectrograms of AC speech, BC speech, noisy clean speech and reconstructed speech via the Laplacian pyramid are demonstrated in Figures 1–3. Figure 1 illustrates the clean and bone-conducted speech, and Figures 2 and 3 illustrates the result for Gaussian noise and an interfering speaker. The figures demonstrate that the LP method facilitates enhancement of speech signals not only in stationary noise environments such as white Gaussian noise, but also in nonstationary noise environments such as an interfering speaker.

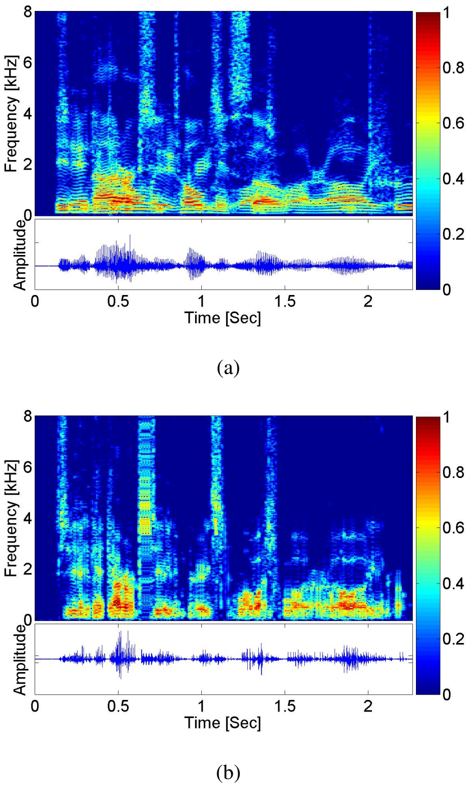
The log spectral distortion (LSD) measure is used to evaluate the quality of the reconstructed speech. The results have



(a)

(b)

Fig. 2: Speech spectrograms and waveforms: (a) noisy air-conducted signal in Gaussian noise environment (SNR=10, LSD=2.4526); (b) Speech enhanced using the LP method (LSD=1.1028).



(a)

(b)

Fig. 3: Speech spectrograms and waveforms: (a) noisy air-conducted signal in an interfering speaker environment (SNR=-1.7821, LSD=1.2804); (b) Speech enhanced using the LP method (LSD=1.1768).

been evaluated for GH, LP, optimally modified log spectral amplitude (OM-LSA) [2] and an existing probabilistic approach (PA) [14].

Table 3: LSD results for Gaussian noise with different SNR levels and interfering speech, obtained by using four different speech enhancement methods: GH, LP, OM-LSA and PA.

LSD	GH	LP	OM-LSA	PA
SNR =0	1.5726	1.1003	2.0901	1.9613
SNR = 10	1.5564	1.1028	1.4979	2.1592
SNR = 20	1.5755	1.1021	1.1085	2.2531
Interfering speech	1.5660	1.1768	1.3604	2.2672

Table 3 presents the LSD results for Gaussian noise with different SNR levels and interfering speech, obtained by using four different speech enhancement methods: GH, LP, OM-LSA and PA. The table demonstrates that the LP method consistently provides the lowest distortion for all tested SNR levels and noise types. The noise level has little influence on the geometric extension methods, which may result from the fact that the nonlinear mapping learned via geometric methods implicitly involves the noise model.

4. CONCLUSIONS

We have presented two function extension techniques for speech reconstruction under the knowledge of samples of air-conducted and bone-conducted speech. Experiments have been conducted on simulated air-conducted and bone-conducted speech in Gaussian noise environments and an interfering speaker. Although involving some distortion, geometric methods enable further noise reduction for a wide range of noise levels and categories.

A possibility for future research is conducting geometric harmonics in a multi-scale manner in accordance with the Laplacian pyramid mechanism, where the extension can be viewed as approximations of residuals in a series of decreasing scales. Furthermore, the relation between the iteration number in the Laplacian pyramid method and noise level in the observation can be derived to pre-determine the iteration number, rather than choosing it via trials of experiments.

5. REFERENCES

[1] Y. Bengio, J.-F. Paiement, P. Vincent, O. Delalleau, N. Le Roux, and M. Ouimet, “Out-of-sample extensions for LLE, isomap, MDS, eigenmaps, and spectral clustering,”

in *Advances in neural information processing systems*, vol. 16, pp. 177–184, 2004.

[2] I. Cohen and B. Berdugo, “Speech enhancement for nonstationary noise environments,” in *Signal processing*, vol. 81, no. 11, pp. 2403–2418, 2001.

[3] R. R. Coifman and S. Lafon, “Geometric harmonics: a novel tool for multiscale out–of–sample extension of empirical functions”, in *Applied and Computational Harmonic Analysis: Special issue on Diffusion Maps and Wavelets*, vol. 21 (2006), pp. 31–52.

[4] L. Deng, Z. Liu, Z. Zhang, and A. Acero, “Nonlinear information fusion in multi-sensor processing-extracting and exploiting hidden dynamics of speech captured by a bone-conductive microphone,” in *Multimedia Signal Processing, 2004 IEEE 6th Workshop on. IEEE*, 2004, pp. 19–22.

[5] W. Fisher, G. Doddington, and Goudie K. Marshall, “The DARPA Speech Recognition Research Database: Specifications and Status,” *Proceedings of DARPA Workshop on Speech Recognition*, 1986, pp. 93–99.

[6] J. Hershey, T. Kristjansson, and Z. Zhang, “Model-based fusion of bone and air sensors for speech enhancement and robust speech recognition,” in *ISCA Tutorial and Research Workshop (ITRW) on Statistical and Perceptual Audio Processing*, 2004.

[7] K. Kondo, T. Fujita, and K. Nakagawa, “On equalization of bone conducted speech for improved speech quality,” in Signal Processing and Information Technology, *2006 IEEE International Symposium on. IEEE*, 2006, pp. 426–431.

[8] Z. Liu, Z. Zhang, A. Acero, J. Droppo, and X. Huang, “Direct Filtering for air-and bone-conductive microphones,” in *Multimedia Signal Processing, 2004 IEEE 6th Workshop on. IEEE*, 2004, pp. 363–366.

[9] N. Rabin and R. R. Coifman, “Heterogeneous datasets representation and learning using diffusion maps and laplacian pyramids,” in *Proc. 12th SIAM Int. Conf. Data Mining*, 2012.

[10] M. S. Rahman and T. Shimamura, “Pitch characteristics of bone conducted speech,” in *Proc. of European Signal Processing Conference (EUSIPCO)*, pp. 795–798, 2010.

[11] M. L. Seltzer, A. Acero, and J. Droppo, “Robust bandwidth extension of noise corrupted narrowband speech,” in *Proc. INTERSPEECH*, Lisbon, Portugal, 2005, pp. 1509–1512.

[12] A. Subramanya, Z. Zhang, Z. Liu, J. Droppo, and A. Acero, “A graphical model for multi-sensory speech processing in air-and-bone conductive microphones.” in *INTERSPEECH*, 2005, pp. 2361–2364.

[13] H. S. Shin, H.-G. Kang, and T. Fingscheidt, “Survey of speech enhancement supported by a bone conduction microphone,” in *Speech Communication; 10. ITG Symposium; Proceedings of VDE*, 2012, pp. 1–4.

[14] A. Subramanya, Z. Zhang, Z. Liu, and A. Acero, “Multisensory processing for speech enhancement and magnitude-normalized spectra for speech modeling,” in *Speech Communication*, vol. 50, no. 3, pp. 228–243, 2008.

[15] T. tat Vu, M. Unoki, and M. Akagi, “An LP-based blind model for restoring bone-conducted speech,” in *Communications and Electronics, 2008. ICCE 2008. Second International Conference on. IEEE*, 2008, pp. 212–217.