Source Separation for Target Enhancement of Food Intake Acoustics from Noisy Recordings

Antoine Liutkus, Temiloluwa Olubanjo, Elliot Moore, Maysam Ghovanloo

SOURCE SEPARATION FOR TARGET ENHANCEMENT OF FOOD INTAKE ACOUSTICS FROM NOISY RECORDINGS

Antoine Liutkus[1] *Temiloluwa Olubanjo*[2] *Elliot Moore*[2] *Maysam Ghovanloo*[2]

[1]Inria, Speech Processing Team, Villers-lès-Nancy, France
[2]Georgia Institute of Technology, Atlanta, GA, USA

ABSTRACT

Automatic food intake monitoring can be significantly beneficial in the fight against obesity and weight management in our society today. Different sensing modalities have been used in several research efforts to accomplish automatic food intake monitoring with acoustic sensors being the most common. In this study, we explore the ability to learn spectral patterns of food intake acoustics from a clean signal and use this learned patterns for extracting the signal of interest from a noisy recording. Using standard metrics for evaluation of blind source separation, namely signal to distortion ratio and signal to interference ratio, we observed up to 20dB improvement of separation quality in very low signal to noise ratio conditions. For more practical performance evaluation of food intake monitoring, we compared the detection accuracy for chew events on the mixed/noisy signal versus on the estimated/separated target signal. We observed up to 60% improvement in chew event detection accuracy for low signal to noise ratio conditions when using the estimated target signal compared to when using the mixed/noisy signal.

***Index Terms*—food intake monitoring, audio source separation, nonnegative matrix factorization, harmonizable processes**

1. INTRODUCTION

Obesity is a prevalent chronic condition that affects 1 of 3 adults in the U.S. today [1]. It leads to an increased risk of heart disease, high blood pressure, type 2 diabetes, arthritis-related disabilities and even some cancers [1]. Weight gain (or loss) is often linked to an imbalance between energy intake through food consumption, and energy expenditure through physical activities. In recent years, a lot of research effort has been committed towards recognizing and quantifying physical activities for estimating energy expenditure [2], [3]. Such work has even progressed into development of commercially available wearable products such as Fitbit [4] and Samsung Gear Fit [5]. These products often work with smartphone applications that allow for manual food intake tracking by selecting meals from an extensive food database. Meanwhile, objective and automatic food intake monitoring using wearable systems is still a work in progress towards development of a realistic and reliable system. In previous research, different sensing modalities have been used towards objective food intake monitoring, including acoustic sensors [6], [7], [8], image sensors [9], [10], electromyography (EMG) sensors [11], and even eletroglottograph (EGG) sensors [12].

Acoustic sensors are the most common sensing modality used by researchers for food intake monitoring; either in a single-sensor system such as BodyScope [13] or multi-sensor system as in [14], [15]. According to [16], there are still many issues that motivate development of new techniques for wearable activity monitoring systems to improve feasibility in more realistic conditions. Two of these issues are 1) development of portable, unobtrusive, and inexpensive data acquisition systems, and 2) collection of data under realistic conditions. Collecting food intake acoustics in a realistic, noise-prone environment, and properly handling environmental noise that is bound to interfere with the target signal is an existing gap in published research. In [17], the authors use a two-microphone system, an in-ear microphone for recording in-body sounds and an outside-ear microphone for recording environmental sounds, then apply spectral subtraction for noise reduction. Another approach for environmental noise reduction in food intake acoustics shown in [18], is a microphone hardware design approach using soft and hard silicone for internal and external acoustic isolation respectively. Although [17], [18] made accommodations to reduce environmental noise that can contaminate food intake acoustic recordings, both datasets were collected in a quiet/laboratory environment therefore performance is unknown for recordings from a loud restaurant for example.

In this paper, we explore the performance of source separation techniques shown to be successful in the music information retrieval domain for separation of food intake acoustics from background noise in a restaurant recording. Food intake acoustic activities such as bites, chews, and swallows are known to be low energy signals especially when recorded non-invasively with a wearable system. In this work, a single microphone approach was employed to support usability and acceptability criteria of wearable technology that calls for portable and unobtrusive systems. To facilitate performance evaluation, acoustics from food intake activities were recorded in a quiet environment to have the clean and uncontaminated signal for comparison, while the restaurant noise was recorded separately. Both recordings were collected using a sampling frequency of 16 kHz with 16-bit resolution. The iASUS NT3 throat microphone [19], which has a frequency response of 20 Hz - 20 kHz, and a sensitivity of -46 dB +/- 3dB was used for data collection. In this preliminary study, an exemplary signal of one randomly picked subject from a larger dataset was used. This database includes tracheal recordings from 12 subjects (7 males, 5 females, age range: 24-33 years) as they ate five foods with varying textures (almonds, apple, chips, crackers and bread). The mixed signal containing food intake acoustics and the background noise from a loud restaurant was created by instantaneous addition for evaluation of the proposed source separation technique. The remainder of this paper is organized as follows, section 2 describes the proposed source separation method applied of food intake isolation, separation results are presented in section 3, finally the conclusion is presented in section 4.

2. MODEL AND METHOD

In this section, we present the method used for isolation of chewing sounds from background noise. The method used is semi-supervised non-negative matrix factorization (NMF), which we present in detail after introducing the particular probabilistic model we adopted for the waveforms.

2.1. Notations and source separation background

Let the tilde notations like, $\tilde{s}$, denote a regularly sampled audio signal in the time domain. Let then s be the corresponding Short-Term Fourier Transform (STFT). It is a $F \times T$ matrix, where F

is the number of frequency bands while T is the number of times frames. Its entries are written $s(f,t) \in \mathbb{C}$. When $\tilde{s}$ is a real waveform, its spectrum is Hermitian and we assume that the redundant information in its STFT has been discarded.

In this study, we will adopt the α-harmonizable model recently introduced in [20]. It assumes that all entries $s(f,t)$ of the STFT are independent, and distributed with respect to complex isotropic α-stable distributions (abbreviated $S\alpha S_c$):

$$s(f,t) \sim S\alpha S_c\left(\sigma^\alpha(f,t)\right), \tag{1}$$

where $\sigma^\alpha(f,t) \geq 0$ is called the *scale parameter*. In essence, it corresponds to the strength —or *power*— of the signal across time and frequency. As shown in [20], this model generalizes the well-known local Gaussian model (LGM) discussed in [21], [22], [23], that corresponds to $\alpha = 2$, and for which the scale parameter is amenable to a *variance* called the Power Spectral Density (PSD). In analogy to this Gaussian case, σ^α in (1) is called the fractional PSD, abbreviated α-PSD.

The α-harmonizable model may be understood the following way: First, the signal $\tilde{s}$ is split into frames, which are all assumed independent. Second, each frame is assumed harmonizable, which basically means that all the entries of its Fourier transform are independent if the frames are long enough. This leads to all $s(f,t)$ being independent. Third, an isotropic α-stable model is picked for each $s(f,t)$ as in (1). As demonstrated in [24, th. 6.5.1], this last step is equivalent to assuming that the waveform of each frame is both an α-stable and a stationary process, which generalizes the well-known Gaussian case. These two features can be used for modeling waveforms. Stability is important in our contex because it means that if J signals s_j are α-harmonizable, so will be their sum. More precisely, if $\forall j$, $s_j(f,t) \sim S\alpha S_c\left(\sigma_j^\alpha(f,t)\right)$ are J independent α-harmonizable processes called *sources*, then their sum x, called the *mixture*, is distributed as[1]:

$$x(f,t) \triangleq \sum_j \sim S\alpha S_c\left(\sum_j \sigma_j^\alpha(f,t)\right), \tag{2}$$

so that we have:

$$\sigma_x^\alpha(f,t) = \sum_j \sigma_j^\alpha(f,t). \tag{3}$$

Now, let us consider the case where we observe the mixture x. If we have estimates $\hat{\sigma}_j^\alpha$ for the α-PSDs σ_j^α of the sources, it is straightforward to estimate the actual source signals s_j, because we have [20]:

$$\mathbb{E}\left[s_j(f,t) \mid x(f,t), \left\{\sigma_j^\alpha\right\}_j\right] = \frac{\sigma_j^\alpha(f,t)}{\sum_{j'} \sigma_{j'}^\alpha(f,t)} x(f,t). \tag{4}$$

Thus, a practical estimate $\hat{s}_j$ of s_j is obtained by replacing the true α-PSD by their estimates $\hat{\sigma}_j$ in (4). This is coined in as α-Wiener filtering in [20]. It is the direct generalization of the classical Wiener filter to $\alpha < 2$.

The advantage of picking an α-harmonizable model and not simply an LGM was highlighted in [20]. It permits adequate modeling of signals that feature very high dynamic ranges through the use of heavy tails α-stable distributions, while keeping the computational benefit of separating them effectively in the Time-Frequency (TF) domain. It thus puts together robust signal filtering, that has mostly been achieved in the time domain for now [25], and the efficiency of Wiener filtering, that only holds for wide-sense stationary signals [23].

[1] $\triangleq$ stands for a definition.

2.2. Model and parameter estimation

The separation procedure (4) permits to recover good estimates of the sources if their α-PSD α_j^α are available, which corresponds to their strength in the TF domain. However, only the mixture x is available in practice, and the α-PSD need to be estimated from the mixture only. All that is available at this stage is equation (3), that provides us with a *data-fit* idea: whatever $\hat{\sigma}_j$we pick, they should sum up so as to correspond to the α-PSD $\sigma_x^\alpha(f,t)$ of the mixture. This appears as a natural idea, since it means that we want the estimated power of our sources to explain the energy of the actual signal we observe. However, two main issues still prevent us from readily applying this idea to estimate the α-PSDs σ_j.

First, our model is still heavily underdetermined, i.e. with more unknowns than available equations and thus leading to an infinite number of equally good solutions. Indeed, we have FTJ unknown parameters in σ_j to estimate, while (3) provides us with FT equations only, and $\sigma_j(f,t) \geq 0$ provides us with FTJ inequalities, resulting in an ill-posed problem. To address this issue, a common idea is to assume that there is some *structure* to be expected in the α-PSDs of the sources and to exploit it, e.g. by expressing each $F \times T$ matrix σ_j using a low-rank model:

$$\sigma_j^\alpha(f,t) = \sum_{k=1}^{K_j} W_j(f,k) H_j(k,t), \tag{5}$$

where $K_j \ll \min(F,T)$ is called the *number of components* for source j, while W_j and H_j are $F \times K_j$ and $K_j \times T$ nonnegative matrices, respectively. Equation (5) is called an NMF model of σ_j^α and only comprises $K_j(F+T) \ll FT$ parameters. It can be understood as assuming that the α-PSD of each source is well explained by the superposition of some spectral patterns (the columns of W_j) modulated over time by their respective *temporal activations* (the lines of H_j). Due to its success in capturing most features of audio spectrograms, it has been a very popular model in audio processing for more than 10 years (see, e.g. [26], [27], [28] and references therein). In matrix form, we see that picking an NMF model, we can replace (3) by the now well-posed problem:

$$\sigma_x^\alpha = \sum_{j=1}^{J} W_j H_j, \tag{6}$$

that provides FT equations for the $\sum_j K_j(F+T) \ll FT$ unknowns, gathered in the parameter space $\Theta = \{W_j, H_j\}_j$.

Our second issue stems from the fact that we do not really observe σ_x^α in (6), but rather the STFT x of the mixture, whose entries $x(f,t)$ are assumed independent, and distributed as:

$$x(f,t) \sim S\alpha S_c\left(\sigma_x^\alpha(f,t)\right), \tag{7}$$

which combines with (6) to yield:

$$x(f,t) \sim S\alpha S_c\left(\sum_{jk} W_j(f,k) H_j(k,t)\right).$$

On probabilistic grounds, a natural idea here would be to estimate the parameters Θ through a maximum likelihood approach. Since all TF bins $x(f,t)$ are independent, this would lead to:

$$\hat{\Theta} \leftarrow \operatorname*{argmin}_\Theta \sum_{f,t} -\log p\left(x(f,t) \mid \Theta\right), \tag{8}$$

which has been routinely achieved for many years in the Gaussian case [29], [30], [31]. Unfortunately, such an approach is not possible in general, because no analytical expression is available

for the $S\alpha S_c$ probability density function except for the $\alpha = 2$ (Gaussian) and $\alpha = 1$ (Cauchy) cases[2].

Since we cannot readily perform maximum likelihood estimation of the parameters, another *optimization-based* route we take in this study consists of first obtaining an estimate $\hat{\sigma}_x^\alpha$ for σ_x^α, and then using it to obtain Θ, by minimizing some *cost-function* D_Θ:

$$\hat{\Theta} \leftarrow \underset{\Theta \geq 0}{\operatorname{argmin}} \left\{ D_\Theta \triangleq \sum_{f,t} d\left(\hat{\sigma}_x^\alpha(f,t) \mid \sum_{jk} W_j(f,k) H_j(k,t) \right) \right\}, \quad (9)$$

where $d(a \mid b) \geq 0$ is a user-defined *divergence* that is small whenever $a \approx b$ and high otherwise. A common choice for this purpose is to pick a β-divergence, which is a family of cost functions indexed by a parameter $\beta \in [0\,2]$. It comprises the Euclidean ($\beta = 2$), Kullback-Leibler ($\beta = 1$) and Itakura-Saito (IS, $\beta = 0$) as special cases, see e.g. [30], [32].

Now, we study how to estimate $\sigma_x^\alpha(f,t)$ based on $x(f,t)$ so as to apply the idea in (9). Intuitively, σ_x^α corresponds to the *power* of x. In the Gaussian case we can simply take $\hat{\sigma}_x^2(f,t) = |x(f,t)|^2$, which states that the PSD of a wide-sense stationary signal is straightforwardly estimated with the power spectrogram. However, this estimate is no more valid for $\alpha < 2$. Indeed, all moments $\mathbb{E}[|x(f,t)|^p]$ for $p \geq \alpha$ are undefined in that case: $\forall \alpha < 2, \forall p \geq \alpha, \mathbb{E}[|x(f,t)|^p] = \infty$. However, the p-moments of $x(f,t)$ for $p < \alpha$ are defined and we have [24, p. 19]:

$$\lim_{p \uparrow \alpha} (\alpha - p)\, \mathbb{E}[|x(f,t)|^p] = \alpha \lambda_\alpha \sigma_x^\alpha(f,t), \quad (10)$$

where λ_α is a constant only depending on α. Thus, if we pick a $p < \alpha$ that is sufficiently close to α and $\hat{\alpha}$ for notational convenience, we may assume that:

$$\mathbb{E}\left[|x(f,t)|^{\hat{\alpha}} \right] \approx \lambda(\alpha, \hat{\alpha})\, \sigma_x^\alpha(f,t),$$

where $\lambda(\alpha, \hat{\alpha})$ is now a constant that only depends on $\hat{\alpha}$ and α, but not on σ_x^α. Consequently, the empirical $\hat{\alpha}$-spectrogram of x:

$$v^{\hat{\alpha}}(f,t) \triangleq |x(f,t)|^{\hat{\alpha}} \quad (11)$$

is expected to match its α-PSD σ_x^α, up to a multiplicative constant, if $\hat{\alpha} < \alpha$ is close enough to α. We may hence use it to learn the parameters σ_j^α, or rather $\lambda(\alpha, \hat{\alpha})\, \sigma_j^\alpha$. Since this constant will cancel out when performing α-Wiener filtering (4), we do not need to compensate for it. The strategy we propose for estimating the model parameters thus amounts to picking:

$$D_\Theta = \sum_{f,t} d_\beta \left(|x(f,t)|^{\hat{\alpha}} \mid \sum_{jk} W_j(f,k) H_j(k,t) \mid \right), \quad (12)$$

for some $\hat{\alpha} \in [0\,2]$ and $\beta \in [0\,2]$ that are fixed beforehand. The choice of these parameters in our case is left to the user and is discussed in our evaluation. Minimization of (12) is achieved through standard NMF methodology, and included in algorithm 1 for completeness. In practice, only 10 iterations of this algorithm were sufficient in our experiments.

2.3. Exploiting learning data

In our particular food intake monitoring application, we have $J = 2$ sources: the throat signal s_1 and background signal s_2. Even if our objective is to separate them in test conditions when they are both unknown, we can exploit the fact that in a controlled silent situation, *learning examples* s_l of throat signals may be observed without any superimposed background. Even if these signals differ from those we want to separate later, they ought to "sound the same". In our model (5), this can be translated as stating that their spectral patterns W_1 should be the same in clean and noisy conditions, even if their activation times are different.

[2]In the general case, the $S\alpha S_c$ distribution is indeed rather defined through its characteristic function $\mathbb{E}[\exp i\theta x] = \exp(-|\theta|^\alpha |x|^\alpha)$ [24].

Algorithm 1 Fitting NMF parameters of an $\hat{\alpha}$-spectrogram $v^{\hat{\alpha}}$.

Always using the latest parameters available for computing $\hat{\sigma}_x^\alpha = \sum_{j=1}^J W_j H_j$, and for all W_j or H_j not fixed, iterate:

$$W_j \leftarrow W_j \frac{\left(v^{\hat{\alpha}} \cdot [\hat{\sigma}_x^\alpha]^{\cdot(\beta-2)} \right) H_j^\top}{[\hat{\sigma}_x^\alpha]^{\cdot(\beta-1)} H_j^\top}$$

$$H_j \leftarrow H_j \frac{W_j^\top \left(v^{\hat{\alpha}} \cdot [\hat{\sigma}_x^\alpha]^{\cdot(\beta-2)} \right)}{W_j^\top [\hat{\sigma}_x^\alpha]_j^{\cdot(\beta-1)}},$$

where $a \cdot b$, $\frac{a}{b}$ and $a^{\cdot c}$ correspond to element-wise multiplication, division and exponentiation, respectively.

Hence, a natural approach inspired by a previous similar study [33] is to learn W_1 using the $\hat{\alpha}$-spectrogram $|s_l|^{\hat{\alpha}}$ of the clean signal s_l, while assuming that only one source $J = 1$ is present at that time, and then fix it during test conditions, for which we set $J = 2$ and only estimate H_1, W_2 and H_2 when fitting the $\hat{\alpha}$-spectrogram (11) of the mixture. This method proved to yield very satisfying results in practice. A complete implementation in Matlab is available on the webpage dedicated to this paper[3]. Even if implemented as a batch method now, the computational complexity of the separation step is small in practice and can be implemented for online processing.

3. EVALUATION

3.1. Separation performance

To evaluate performance of the proposed method, we first considered classic metrics from BSSeval toolbox [34], that notably feature Signal to Distortion Ratio (SDR) and Signal to Interference Ratio (SIR). While SDR gives an overall score for separation of each source, SIR provides a metric for interference reduction between sources. Both are expressed in dB and are higher for better separations. In order to quantify improvement brought by the method, corresponding dSDR and dSIR metrics are computed, these metrics give the difference between the score obtained after separation as compared when trivially picking the mixture as an estimate for both sources, i.e. not doing any separation at all.

The $K_1 = 10$ target spectral patterns W_1 were learned on 1 minute of diverse food intake acoustics recorded in silent laboratory setting. Then, these sounds were mixed with real-world background recorded from a restaurant environment, so as to form the mixture signals to separate. A gain was applied to the background noise, to achieve any desired Signal to Noise Ratio (SNR) in the mixture. The same microphone was used to record both the clean throat sounds and the background. After applying the proposed separation procedure to the mixture (using $K_2 = 5$), separation quality was then assessed by comparing the estimated throat sounds with their true values.

First, for a fixed $SNR = -17$dB, the performance of the method was computed for 200 $(\hat{\alpha}, \beta)$ values in $[0.2\,2] \times [0\,2]$. Results are displayed in figure 1.

Considering figure 1, we see that the benefit of using the method is not the same for all $(\hat{\alpha}, \beta)$. On the contrary, we notice that the value $\hat{\alpha} = 0.5$, i.e. modelling square-root magnitude STFTs can yield a near 20dB improvement, depending on the β-divergence considered, over using the classical power spectrogram.

[3]www.loria.fr/~aliutkus/fimWASPAA2015/

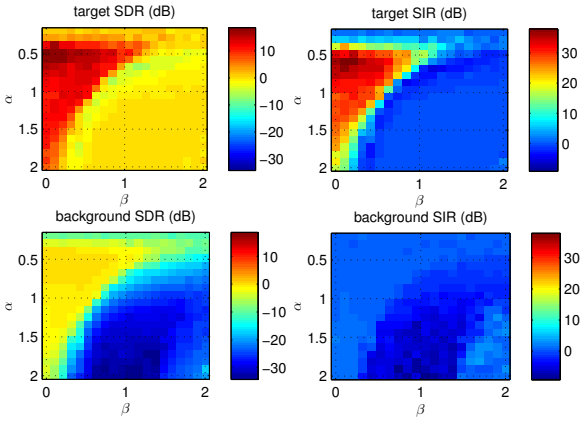

Figure 1. Separation performance for various $\hat{\alpha}$ and β with a signal to noise ratio of -17dB. Higher is better.

Figure 2. Separation performance for $\hat{\alpha} = 0.5$ and $\beta = 0$, for various signal to noise ratios (SNR).

This is a strong empirical confirmation of the superiority of the α-harmonizable model over LGM ($\alpha = 2$) in some cases. In our setting, this can be explained by the fact that throat sounds are extremely impulsive in nature, thus strongly benefiting from a model that explicitly handles large dynamic ranges and impulsive data, as does the α-harmonizable model. Then, fixing $\hat{\alpha} = 0.5$, we notice that $\beta = 0$ (IS) achieves the best results. These results bear similarities with those presented in [35], where $\alpha < 2$ showed better performance.

Now picking the identified $(\hat{\alpha} = 0.5, \beta = 0)$ optimal parameters, we performed a second set of evaluation of the separation performance, to study the benefit of using the method for various SNRs, from -20dB to 10dB. Results are displayed in figure 2.

From this figure, we see that the gain in using the proposed technique compared to using the original mixture is very high for small SNR, which is our use-case in practice. When the SNR gets very high (above 5dB), we notice only a marginal increase of the separation quality (dSNR), while interferences are still well reduced ($dSIR \approx 10$dB).

3.2. Counts of chewing events

Detecting and counting of chew events in a food intake cycle is an objective metric that can be used to evaluate an automatic food intake monitoring system [6], [17]. Päßler and Fischer, in [17], presented and evaluated eight different algorithms for automated chew event detection on food intake sounds from consumption of six types of food. In this study, we apply the most successful and efficient algorithm from [17], maximum sound energy algorithm, for evaluation of the proposed source separation method. As with

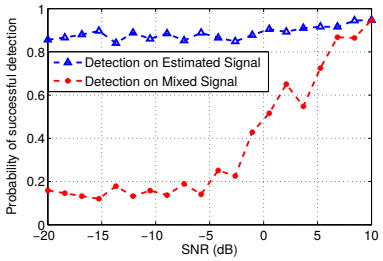

Figure 3. Chew event detection on mixed signal and estimated target signal relative to performance on clean signal, for various signal to noise ratios.

the maximum sound energy algorithm in [17], chew events were detected from a food intake cycle when the signal energy in a 23 ms frame segment and the following 12 frames exceeded a minimum threshold. Our minimum threshold value was found by comparing results of the chew event detection algorithm to a manually annotated ground truth of the test signal to obtain the best possible performance. See [17] for specific details on this chew event detection algorithm. Performance of the maximum sound energy algorithm for chew event detection on the mixed signal and the estimated target signal, relative the clean signal, was then computed for various SNR values.

Figure 3 shows the results achieved from comparing chew event detection on the estimated signal with chew event detection on the clean signal. We observe that in negative SNR cases, when the noise signal completely overpowers the target signal, for example: $[-20\ -5]$ dB, there is $\approx 60\%$ increase in chew event detection accuracy achieved from using the estimated signal. On the other hand, there is a little-to-no notable difference in detection accuracy when the SNR is $\geq$ 7dB. This result shows that in food intake monitoring applications, where the target is a low energy signal compared to the surrounding noise, in a loud restuarant for example, a huge benefit can be achieved from applying an intelligent source separation technique to estimate the clean signal compared to simply using the mixed signal for processing.

4. CONCLUSION

In this study, we have demonstrated how the recent source separation models and methods can be used to denoise signal of interest in real-world single-sensor food intake acoustic data. Using only a limited recording, 1 minute, of the target signal, obtained in a silent laboratory setting, we showed that we can learn an adequate signal model for use in isolating the food intake acoustics from adverse background noise. We also showed the benefit of using this technique to exploit the denoised data for automatic monitoring applications is very high, compared to using the original mixture data. Additionally, in the case of automatic food intake recognition, we observed that using the proposed method to obtain an estimated target signal provided up to 60% improvement in chew event detection compared to the detection accuracy achieved on the mixed signal.

On practical grounds, using the α-harmonizable model for denoising real-world food intake acoustics recorded in a noisy environment proved to be very beneficial due to its performance and low computational requirement. This makes it feasible to be embedded in a small wearable system. On theoretical grounds, the results obtained in this work shows that the recently proposed α-harmonizable model can achieve excellent separation in cases where the classical Gaussian model fails. Since food intake acoustics are very impulsive in nature, we interpret this result as a strong claim in favor of the α-harmonizable model, when the dynamic range of the signals to separated is very high.

5. REFERENCES

[1] National Center for Chronic Disease Prevention and Health Promotion. The power of prevention. Chronic disease... the public health challenge of the 21st century. http://www.cdc.gov/chronicdisease/pdf/2009-power-of-prevention.pdf, 2009.

[2] C. Yang and Y. Hsu. A review of accelerometry-based wearable motion detectors for physical activity monitoring. *Sensors*, 10(8):7772–7788, 2010.

[3] J. Altini, M.and Penders and O. Amft. Energy expenditure estimation using wearable sensors: a new methodology for activity-specific models. In *Proceedings of the conference on Wireless Health*, page 1. ACM, 2012.

[4] Fitbit. http://www.fitbit.com.

[5] Samsung Gear Fit. http://www.samsung.com/us/mobile/wearable-tech.

[6] M. Shuzo, S. Komori, T. Takashima, G. Lopez, S. Tatsuta, S. Yanagimoto, S. Warisawa, J. Delaunay, and I. Yamada. Wearable eating habit sensing system using internal body sound. *Journal of Advanced Mechanical Design, Systems, and Manufacturing*, 4(1):158–166, 2010.

[7] S. Päßler, M. Wolff, and W. Fischer. Food intake monitoring: an acoustical approach to automated food intake activity detection and classification of consumed food. *Physiological Measurement*, 33(6):1073, 2012.

[8] W. Walker and D. Bhatia. Automated ingestion detection for a health monitoring system. *Biomedical and Health Informatics, IEEE Journal of*, 18(2):682–692, 2014.

[9] G. Shroff, A. Smailagic, and D. Siewiorek. Wearable context-aware food recognition for calorie monitoring. In *Int. Symp. on Wearable Computers*, pages 119–120. IEEE, 2008.

[10] Tatsuya Miyazaki, Gamhewage C de Silva, and Kiyoharu Aizawa. Image-based calorie content estimation for dietary assessment. In *IEEE International Symposium on Multimedia (ISM)*, pages 363–368. IEEE, 2011.

[11] O. Amft and G. Troster. Methods for detection and classification of normal swallowing from muscle activation and sound. In *Pervasive Health Conference and Workshops*, pages 1–10. IEEE, 2006.

[12] M. Farooq, J. Fontana, and E. Sazonov. A novel approach for food intake detection using electroglottography. *Physiological Measurement*, 35(5):739, 2014.

[13] K. Yatani and K. Truong. Bodyscope: a wearable acoustic sensor for activity recognition. In *Ubiquitous Computing*, pages 341–350. ACM, 2012.

[14] Oliver Amft and Gerhard TrÃ¶ster. Recognition of dietary activity events using on-body sensors. *Artificial Intelligence in Medicine*, 42(2):121 – 136, 2008. Wearable Computing and Artificial Intelligence for Healthcare Applications.

[15] Jing Liu, Edward Johns, Louis Atallah, Claire Pettitt, Benny Lo, Gary Frost, and Guang-Zhong Yang. An intelligent food-intake monitoring system using wearable sensors. In *Wearable and Implantable Body Sensor Networks (BSN), 2012 Ninth International Conference on*, pages 154–160. IEEE, 2012.

[16] O. Lara and M. Labrador. A survey on human activity recognition using wearable sensors. *Communications Surveys & Tutorials, IEEE*, 15(3):1192–1209, 2013.

[17] S. Päßler and W. Fischer. Food intake monitoring: Automated chew event detection in chewing sounds. *Biomedical and Health Informatics, IEEE Journal of*, 18(1):278–289, 2014.

[18] T. Rahman, A. Adams, M. Zhang, E. Cherry, B. Zhou, H. Peng, and T. Choudhury. Bodybeat: A mobile system for sensing non-speech body sounds. In *International Conference on Mobile systems, Applications, and Services*, pages 2–13. ACM, 2014.

[19] iASUS NT3 throat microphone. http://www.iasus-concepts.com/product?pn=NT3-Throat-Mic-System.

[20] A. Liutkus and R. Badeau. Generalized Wiener filtering with fractional power spectrograms. In *Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, April 2015.

[21] A.T. Cemgil, P. Peeling, O. Dikmen, and S. Godsill. Prior structures for time-frequency energy distributions. In *IEEE Workshop on App. of Sig. Proc. to Audio and Acoustics (WASPAA)*, pages 151–154, October 2007.

[22] N.Q.K. Duong, E. Vincent, and R. Gribonval. Under-determined reverberant audio source separation using a full-rank spatial covariance model. *Audio, Speech, and Language Processing, IEEE Transactions on*, 18(7):1830 –1840, sept. 2010.

[23] A. Liutkus, R. Badeau, and G. Richard. Gaussian processes for underdetermined source separation. *IEEE Transactions on Signal Processing*, 59(7):3155 –3167, July 2011.

[24] G. Samoradnitsky and M. Taqqu. *Stable non-Gaussian random processes: stochastic models with infinite variance*, volume 1. CRC Press, 1994.

[25] P. Kidmose. *Blind separation of heavy tail signals*. PhD thesis, Technical University of Denmark, Lyngby, Denmark, 2001.

[26] A. Cichocki, R. Zdunek, A. H. Phan, and S. Amari. *Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-way Data Analysis and Blind Source Separation*. Wiley Publishing, September 2009.

[27] P. Smaragdis, C. Févotte, G.J. Mysore, N. Mohammadiha, and M. Hoffman. Static and dynamic source separation using nonnegative factorizations: A unified view. *IEEE Signal Processing Magazine*, 31(3):66–75, May 2014.

[28] A. Ozerov, E. Vincent, and F. Bimbot. A general flexible framework for the handling of prior information in audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(4):1118–1133, May 2012.

[29] C. Févotte, N. Bertin, and J.-L. Durrieu. Nonnegative matrix factorization with the Itakura-Saito divergence. With application to music analysis. *Neural Computation*, 21(3):793–830, March 2009.

[30] C. Févotte and J. Idier. Algorithms for nonnegative matrix factorization with the beta-divergence. *Neural Computation*, 23(9):2421–2456, Sep. 2011.

[31] C. Févotte and A. Ozerov. Notes on nonnegative tensor factorization of the spectrogram for audio source separation: statistical insights and towards self-clustering of the spatial cues. In *Int. Sym. on Computer Music Modeling and Retrieval (CMMR)*, volume 6684, pages 102–115. Springer, 2010.

[32] D. FitzGerald, M. Cranitch, and E. Coyle. On the use of the beta divergence for musical source separation. In *Irish Signals and Systems Conference (ISSC)*, June 2008.

[33] C. Damon, A. Liutkus, A. Gramfort, and S. Essid. Non-negative matrix factorization for single-channel EEG artifact rejection. In *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, May 2013.

[34] E. Vincent, R. Gribonval, and C. Févotte. Performance measurement in blind audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(4):1462 –1469, July 2006.

[35] B. King, C. Févotte, and P. Smaragdis. Optimal cost function and magnitude power for nmf-based speech separation and music interpolation. In *Machine Learning for Signal Processing (MLSP), 2012 IEEE International Workshop on*, pages 1–6. IEEE, 2012.