Fractal Nature of Chewing Sounds

Vasileios Papapanagiotou[1](✉), Christos Diou[1], Zhou Lingchuan[2], Janet van den Boer[3], Monica Mars[3], and Anastasios Delopoulos[1]

[1] Aristotle University of Thessaloniki, Thessaloniki, Greece
{vassilis,diou}@mug.ee.auth.gr
http://mug.ee.auth.gr
[2] CSEM SA, Landquart, Switzerland
lingchuan.zhou@csem.ch
[3] Wagenigen University, Wagenigen, Netherlands
monica.mars@wur.nl, adelo@eng.auth.gr

Abstract. In the battle against Obesity as well as Eating Disorders, non-intrusive dietary monitoring has been investigated by many researchers. For this purpose, one of the most promising modalities is the acoustic signal captured by a common microphone placed inside the outer ear canal. Various chewing detection algorithms for this type of signals exist in the literature. In this work, we perform a systematic analysis of the fractal nature of chewing sounds, and find that the Fractal Dimension is substantially different between chewing and talking. This holds even for severely down-sampled versions of the recordings. We derive chewing detectors based on the the fractal dimension of the recorded signals that can clearly discriminate chewing from non-chewing sounds. We experimentally evaluate snacking detection based on the proposed chewing detector, and we compare our approach against well known counterparts. Experimental results on a large dataset of 10 subjects and total recordings duration of more than 8 hours demonstrate the high effectiveness of our method. Furthermore, there exists indication that discrimination between different properties (such as crispness) is possible.

1 Introduction

Monitoring and managing dietary behaviour has received extensive focus during the last few years, since both Obesity (OB) and Eating Disorders (ED), such as Anorexia Nervosa (AN) and Bulimia Nervosa (BN), currently affect a very large portion of the population[1][2]. Recent advancements in the field of mobile computing have enabled the use of wearable sensors for monitoring the human behaviour in various aspects of everyday life. Vast development and enhancement of the capabilities of mobile phones, as well as networking, combined with various wearable sensors (e.g. smart watches) have practically transformed them into personal monitoring devices, that can be used to exploit data otherwise

[1] www.who.int/gho/ncd/risk_factors/overweight/en/
[2] www.anad.org/get-information/about-eating-disorders/eating-disorders-statistics/

V. Murino et al. (Eds.): ICIAP 2015 Workshops, LNCS 9281, pp. 401–408, 2015.
DOI: 10.1007/978-3-319-23222-5_49

unavailable even to clinician experts. This data can be used to detect risks (for example for the development of OB or ED, such as the case of the SPLENDID project[3]) and even help reduce those risks.

Regarding the monitoring of dietary activities and behaviour, one of the most commonly proposed sensors in bibliography is the microphone, either open-air or bone conduction. The microphone is usually placed in an unobtrusive location, such as housed in a set of ear phones, and continuously records audio throughout the day, or during long periods of a day. The streamed audio is analysed, usually in real-time, and chewing activity is detected. Most proposed algorithms for the processing of such signals employ well known methods from the field of digital signal processing, such as computation of various statistical features of buffered audio segments, and usually combine them with statistical machine learning methods. Other approaches try to model the distinct structure of chewing sounds, employing heuristically defined rules.

O. Amft was one of the first to systematically analyse chewing sounds, and develop an off-line algorithm to detect chews on continuous streaming audio data. In [2], various positions of a condenser microphone recording at 44.1 kHz are studied, to determine the optimal for automatic chewing detection. Placing the microphone at the inner ear, directed towards the ear drum was found to yield the best results, as in this position chewing sounds are recorded louder than speech sounds. Thus, recognition of chewing sounds is based on the amplitude of the recorded signal. The useful frequency content is determined from 0 to 10 kHz (requiring 20 kHz sampling rate). Furthermore, a speech recognition system is used to reject talking and further increase the precision of chewing detection. In a later work, a complex pipeline is proposed in [1] that (a) estimates various multi-resolution statistical features of audio segments, (b) performs feature selection, and (c) uses a feature similarity measure to detect chews. The detection system is able to discriminate between three food types of distinct texture qualities (crispiness and wetness). However, the computational burden is significantly high, increasing the required resources for a real-time implementation.

In [8], seven chewing detection algorithms are evaluated on a common dataset. The dataset includes recordings of 51 subjects, consuming 6 different food types, using a microphone recording at 11,025 Hz. One algorithm requires the use of a second microphone, placed behind the ear, and uses the difference of the signals' power between the two microphones to detect chews. Another algorithm associates chewing sounds with a particular shape of the signal energy (a local maximum followed by an interval of lower energy). Another one is based on the principle that the power spectrum of chewing sounds is centred around specific frequencies, and thus can be used to distinguish chewing from other sounds. Other algorithms detect chewing regions by identifying the dominant frequency at which chews occur, which is commonly around 1 and 3 Hz. Authors report accuracy from 50% to 60% and precision from 75% to 91% on average. However, it is important to note that the recording of the dataset was performed in

[3] splendid-program.eu/

laboratory conditions and the participants were instructed not to talk or make any other disturbing sound, which makes the recognition task significantly easier.

In this work, we explore the Fractal Dimension (FD) of chewing sounds, as recorded by such an open air microphone placed inside the outer ear canal, in comparison to the FD of other sounds recorded by this sensor, such as talking, coughing, ambient noise and silence. Section 2 presents the analysis and the method for computing the FD, whereas in Section 3 further analysis is performed to design a detection algorithm. In Section 4 various experiments are presented, including the application of the detection algorithm on a large dataset for the purpose of detecting snacking events. Results are compared to other state-of-the-art algorithms. Finally, Section 5 concludes this work.

2 Fractal Dimension of Chewing Sounds

Mandelbrot [4] defines the FD of a graph of a real valued function as its Hausdorff Dimension (HD). In the work of Maragos et al. [6], an algorithm for estimating the FD of such real valued functions is presented, based on a morphological covering of the function using the erosion and dilation operators [3].

Given a real valued function $x(t), 0 \leq t \leq T$, its graph can be defined formally as $F = \left\{(t, x(t)) \in \mathbf{R}^2 : t \in [0, T]\right\}$. The FD can then be defined as follows. Given a morphological element B and a scaling factor ϵ, the FD is estimated as

$$D = 2 - \lim_{\epsilon \to 0} \frac{\log\left(A_B(\epsilon)\right)}{\log\left(\epsilon\right)} \tag{1}$$

where $A_B(\epsilon)$ is the area resulting from dilating the graph by ϵB. In the case where B is a compact, single-connected, symmetric planar set, the two-dimensional processing of the signal can be avoided [5,7]. If we define the structuring function $G_\epsilon(t) = \sup\left\{y \in \mathbf{R} : (t, y) \in \epsilon B\right\}$, then the area $A_B(\epsilon)$ can be approximated by

$$A_B(\epsilon) \approx \int_0^T \left(\left[x \oplus G_\epsilon\right](t) - \left[x \ominus G_\epsilon\right](t)\right) dt \tag{2}$$

where $\left[x \oplus G_\epsilon\right](t)$ and $\left[x \ominus G_\epsilon\right](t)$ are the dilation and erosion of $x(t)$ by $G_\epsilon(t)$.

In the case of discrete signals, and for discrete structure elements, we can approximate $A_B(\epsilon)$ as

$$A_B(\epsilon) \approx \sum_{n=0}^{N-1} \left[x_k^d(n) - x_k^e(n)\right], \epsilon = \epsilon_0 k, k = 0, 1, 2, \ldots, M \tag{3}$$

where the discrete version of dilation $x_k^d(n)$ and erosion $x_k^e(n)$ at level k are computed recursively as

$$x_0^d(n) = x(n) \tag{4}$$
$$x_0^e(n) = x(n) \tag{5}$$
$$x_k^d(n) = [x_{k-1}^d \oplus v](n) \tag{6}$$
$$x_k^e(n) = [x_{k-1}^e \ominus v](n) \tag{7}$$

In practice, we choose a flat structure element v of length $L = \lceil f_s T \rceil$ where $T = 3$ msec, and thus

$$x_k^d(n) = \max\{x_{k-1}^d(n+i) : i = -\lfloor \frac{L}{2} \rfloor, \ldots, 0, \ldots, \lceil \frac{L}{2} \rceil - 1\} \tag{8}$$

For the erosion, the max operator is replaced with min.

According to [5], the FD D can be estimated by linear fitting on $\log(A_B(\epsilon)) = (2-D)\log(\epsilon)$, for discrete scales of $\epsilon = k\epsilon_0$, $k = 1, 2, \ldots, M$. Instead, we estimate D as the mean of local gradients

$$D = \frac{1}{M} \sum_{\epsilon=1}^{M} \frac{\log\left(A_B\left((k+1)\,\epsilon_0\right)\right) - \log\left(A_B\left(k\epsilon_0\right)\right)}{\log(k+1) - \log(k)} \tag{9}$$

In order to examine the fractal properties of chewing sounds, in particular compared to other sounds commonly recorded by such a microphone as the one used in this work (e.g. talking, coughing, etc), we extract recordings of individual chews of six food types of various properties (such as crispness), as well as segments of approximately same duration of coughing, talking, and silence (and some ambient noise). The number of audio segments for each category are presented in Table 1. The recordings that contain these segments belong to a much larger dataset which is presented in Section 4, and used in the final experiment of snacking detection.

Table 1. The extracted audio segments of chewing and non-chewing segments.

Food Type	No.	Type	No.
Apple	156	Cough	15
Banana	63	Pause	1032
Bread	84	Talking	147
Candy bar	96		
Chewing gum	126		
Potato chips	149		
Total	674	Total	1194

Fig. 1 shows the data points for 20 chews of "apple" and audio segments of "talking", for $k = 1, 2, \ldots, 40$. We use both the audio segments, and their time-derivatives. The fact that these curves are approximately linear is a strong indication that these chew segments are highly fractal in nature. Note that for some audio segments, the curves' gradients tend to decrease for larger values of k. This is not accurate however, but rather a computational artifact, since for such values of k the length of the equivalent structuring element is comparable to the length of the audio segment, causing this inaccurate result.

Furthermore, only a few data points are required to estimate the FD. In the following, we have selected $M = 6$. Selecting such a low value for M, combined with the computationally lighter method for computing the dilation and erosion banks (by iterative application of the same structuring element), allows the implementation of a fast and computationally inexpensive detection algorithm. Finally, it also avoids the problem of the computational artifacts caused by excessively large structuring elements, as noted above.

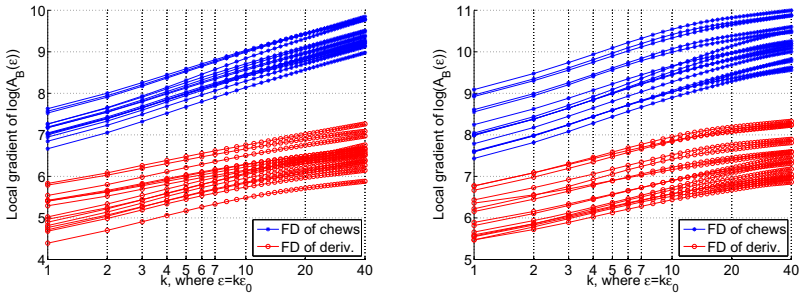

Fig. 1. Local gradient for $\log(A_B(\epsilon))$ versus $k = 1, 2, \ldots, 40$ (in log-scale), for 20 chew segments (blue) and their derivatives (red) of "apple" (left), and 15 segments of "coughing" (right).

3 Designing a Detection Algorithm

In order to examine the fractal nature of the chewing sounds, we compute the FD of the extracted audio segments and their derivatives, at various sampling rates, lower than the original. This is achieved by resampling the original recordings. Fig. 2 (*left*) shows the results for "apple" and "talking". The mean ($\pm$ standard deviation) curves of the FD of the signals are presented, sampling at frequencies $0.5, 1, 2, 4, 8, 16$ and 32 kHz, as well as for the original frequency of 44.1 kHz, in a log-scale plot. The statistics are not affected by the down-sampling, even for as low as 2 kHz, which corresponds to a narrow frequency content of only 1 kHz. Very similar results are obtained for all six food types. This observation reduces the detector requirements for the sampling frequency at just 2 kHz, significantly reducing the computational effort required to process the audio signals.

Using the down-sampled (at 2 kHz) segments, a three-dimension feature vector is computed for each, using the FD of the segment D_x, the FD of the derivative of the segment D_s, and the segment energy E. The features for all the extracted segments are shown in Fig. 2 (*right*). The six food types have been grouped into two clusters for visual clarity, whereas the non-chewing categories are presented separately. As it can be seen, the union of the two chewing clusters is almost linearly separable from "silence", based solely on the energy feature, which is expected. Furthermore, it as also separable (again almost linearly) from "talking" and "coughing".

These results are particularly encouraging. First, the fact that these five classes form separable clusters is strong evidence of the fractal nature of chewing sounds, and enables the detection of chewing sounds based on their fractal dimension. At second, they are also promising in discriminating between different food type properties. For example, the first cluster (as presented in Fig. 2 (*right*)) includes chews of "banana" and "potato chips". "Banana" is not crispy, which results in a relatively lower FD. "Potato chips" are crispy at first, but quickly transform into a wet bolus after the very few first chews. On the other

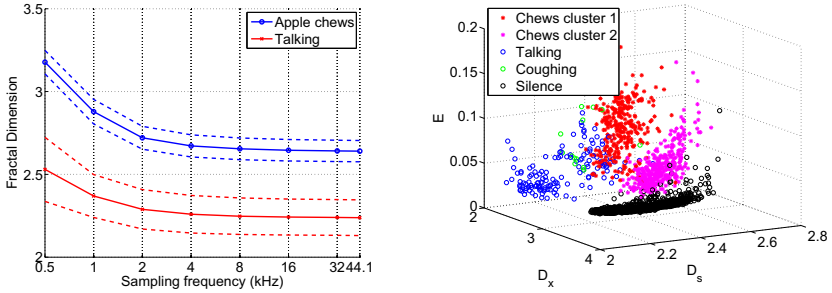

Fig. 2. *Left:* Mean (± standard deviation) of FD of apple chews, and talking segments, across various sampling frequencies (log-scale), showing (a) the linear separability of the two classes, and (b) that down-sampling up to 2 kHz does not significantly alter the actual value of FD. *Right:* Feature vectors for the entire dataset, at 2 kHz. Chews cluster 1 includes "banana" and "potato chips", cluster 2 the remaining 4 food types.

hand, the food types of the second cluster are consistently crispy throughout an entire bite (e.g. "chewing gum").

Finally, in order to enable processing of streamed audio data, we experiment with various lengths of sliding windows along each audio segment, in order to determine the minimum length that captures its fractal properties. We have found that a window length of 300 msec is sufficient to detect parts of chews, and thus enable robust chewing detection.

4 Experiments

A dataset was recorded at the Wageningen University, Netherlands, in the framework of EU funded program SPLENDID. It contains recordings of 10 individuals wearing a prototype sensor consisting of an FG-23329 microphone housed in an ear bud, and connected using audio cable to recording apparatus. Various activities were performed by each subject in randomised sequences, and include pauses, talking, listening to another person speaking, coughing, and consuming a variety of different foods and liquids, such as apples, lettuce, potato chips, toffee, water, milk, etc. The recording for each subject lasted approximately 30 minutes. It is important to notice that during the recordings there was no request for absolute silence. In contrast, some chewing activities were performed under non-silent conditions. For example, the subject was asked to consume a specific type of food while listening to the supervisor talking. The extracted chews of Table 1 belong to the recordings of two subjects of the dataset.

In order to validate our findings, we perform a classification experiment. We form a classification problem with three classes: "chew", "talk/cough" and "silence". The classification method is a two step process. First, the energy of the segment is compared to a threshold; this essentially removes all segments of "silence". Second, an optimal straight line on the $D_x \times D_s$ plane discriminates between "chew" and non-chew. Table 2 presents the results, however we show

each food type and non-chewing activity separately, to gain a better understanding of the misclassification cases. Out of the 6 different food types, only 7 potato chip chews have been misclassified as talking or coughing, which indicates a clear discrimination between chewing and talking/coughing. On the other hand, only 9 chews have been misclassified as silence; most probably due to the lower energy of those segments. Classification accuracy in the 3-class problem is 95.4%, whereas for the binary "chew" vs. non-chew problem is 96.5% (using all of the extracted segments).

Table 2. Confusion matrix for the classification experiment with linear kernel and three classes: "chew", "talk/cough" and "silence". Energy threshold is 0.0202, and the separating line in the $D_x \times D_s$ plane is $y = -2.62x + 8.73$.

Class	Chew	T/C	Sil.
Apple	156	0	0
Banana	62	0	3
Bread	83	0	1
Candy bar	95	0	1
Chewing gum	120	0	6
Cough	2	13	0
Pause	27	0	1005
Potato chips	142	7	0
Talking	21	106	20

In order to examine the efficiency of the proposed algorithm in real time conditions, we apply our algorithm on the large dataset presented in this Section, so as to detect individual chews. This is achieved by thresholding the energy against an adaptively computed mean energy, and using the optimal separating line from the previous experiment on the $D_x \times D_s$ feature space. A median filter is then applied as a post processing step. Finally, chews are created from subsequent windows that are classified as chewing. We then apply an aggregation method to obtain chewing bouts (each chewing bout contains multiple chews) and evaluate this result, as a binary classification problem, based on duration of predicted intervals. To compare our algorithm with other known algorithms of the literature, we also apply some algorithms of [8] so as to detect individual chews, and use the same aggregation method to obtain the corresponding chewing regions. The aggregation algorithm assigns chews to the same bout if they are no more than 5 seconds apart. This relatively relaxed condition allows the chewing detection to "miss" a chew (or two) without fragmenting the bout. This yields consecutive intervals of chewing and non-chewing activity. We present the prediction precision and recall of each algorithm in Table 3. The proposed algorithm maintains a balance between high precision and recall, compared to other algorithms such as Ch. Band Power, that achieves higher precision (by 1%) at the cost of much lower recall.

Furthermore, we subsequently aggregate chewing bouts to snacks, by assigning to the same snack all bouts that are no more than 45 seconds apart. This interval seems realistic in real time application. However, in the dataset, subjects performed activities based on a schedule, and recordings of different snacking events are sometimes recorded much closer than 45 seconds. In these cases, we explicitely split the chewing bouts properly into different snacks. We then use a one-to-one method to assign predicted snacks to ground truth snacks. Table 3 presents the precision and accuracy at the snack classification level.

5 Conclusions

Table 3. Precision and recall for chew bouts and s1nacks

Algorithm	Chew bout		Snack	
	Prec	Rec	Prec	Rec
Max. Sound En.	0.85	0.75	0.77	0.90
Max. Spec. B. En.	0.89	0.76	0.81	0.89
L. P. Filtering	0.86	0.78	0.79	0.94
Ch. Band Power	0.92	0.61	0.92	0.87
Fractal Dim.	0.91	0.87	0.86	0.98

In this work, we explored the FD of chewing sounds, as a means to automatic monitoring of dietary activity, using a wearable microphone sensor. We have performed a systematic analysis of the fractal nature of chewing sounds, which indicates that chewing sounds are highly fractal. Thus, the FD can be used to discriminate chewing from non-chewing sounds, such as talking, coughing, or silence. This property persists even after significant down-sampling of the audio into very narrow spectral bandwidth. Furthermore, promising evidence was found that FD can be used to discriminate between different food properties, such as crispness. Based on these findings, we then proposed a chewing detection algorithm, and tested it on a large, realistic dataset. Results show an improvement in both precision and recall compared to other literature algorithms.

Acknowledgments. The work leading to these results has received funding from the European Community's ICT Programme under Grant Agreement No. 610746, 01/10/2013 - 30/09/2016.

References

1. Amft, O., Kusserow, M., Troster, G.: Bite weight prediction from acoustic recognition of chewing. IEEE Transactions on Biomedical Engineering **56**(6), 1663–1672 (2009)
2. Amft, O., Stäger, M., Lukowicz, P., Tröster, G.: Analysis of chewing sounds for dietary monitoring. In: Beigl, M., Intille, S.S., Rekimoto, J., Tokuda, H. (eds.) UbiComp 2005. LNCS, vol. 3660, pp. 56–72. Springer, Heidelberg (2005)
3. Gonzalez, R.C., Woods, R.E.: Digital image processing (2002)
4. Mandelbrot, B.B.: The fractal geometry of nature/revised and enlarged ed., 495p., 1. WH Freeman and Co., New York (1983)
5. Maragos, P.: Fractal signal analysis using mathematical morphology. Advances in electronics and electron physics **88**, 199–246 (1994)
6. Maragos, P., Potamianos, A.: Fractal dimensions of speech sounds: Computation and application to automatic speech recognition. The Journal of the Acoustical Society of America **105**(3), 1925–1932 (1999)
7. Maragos, P., Sun, F.-K.: Measuring the fractal dimension of signals: morphological covers and iterative optimization. IEEE Transactions on signal Processing **41**(1), 108–121 (1993)
8. Päßler, S., Fischer, W.-J.: Evaluation of algorithms for chew event detection. In: Proceedings of the 7th International Conference on Body Area Networks, ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), pp. 20–26 (2012)