Open access database of raw ultrasonic signals acquired from malignant and benign breast lesions

Hanna Piotrzkowska-Wróblewska[a)]
Department of Ultrasound, Institute of Fundamental Technological Research, Polish Academy of Sciences, Pawińskiego 5B, Warsaw, 02-106, Poland

Katarzyna Dobruch-Sobczak
Department of Ultrasound, Institute of Fundamental Technological Research, Polish Academy of Sciences, Pawińskiego 5B, Warsaw, 02-106, Poland
Department of Radiology, Cancer Center and Institute of Oncology M. Skłodowska-Curie Memorial, Wawelska 15, 02-034 Warsaw, Poland

Michał Byra and Andrzej Nowicki
Department of Ultrasound, Institute of Fundamental Technological Research, Polish Academy of Sciences, Pawińskiego 5B, Warsaw, 02-106, Poland

(Received 27 September 2016; revised 15 August 2017; accepted for publication 21 August 2017; published xx xxxx xxxx)

Purpose: The aim of this paper is to provide access to a database consisting of the raw radio-frequency ultrasonic echoes acquired from malignant and benign breast lesions. The database is freely available for study and signal analysis.
Acquisition and validation methods: The ultrasonic radio-frequency echoes were recorded from breast focal lesions of patients of the Institute of Oncology in Warsaw. The data were collected between 11/2013 and 10/2015. Patients were examined by a radiologist with 18 yr' experience in the ultrasonic examination of breast lesions. The set of data includes scans from 52 malignant and 48 benign breast lesions recorded in a group of 78 women. For each lesion, two individual orthogonal scans from the pathological region were acquired with the Ultrasonix SonixTouch Research ultrasound scanner using the L14-5/38 linear array transducer. All malignant lesions were histologically assessed by core needle biopsy. In the case of benign lesions, part of them was histologically assessed and another part was observed over a 2-year period.
Data format and usage notes: The radio-frequency echoes were stored in Matlab file format. For each scan, the region of interest was provided to correctly indicate the lesion area. Moreover, for each lesion, the BI-RADS category and the lesion class were included. Two code examples of data manipulation are presented. The data can be downloaded via the Zenodo repository (https://doi.org/10.5281/zenodo.545928) or the website http://bluebox.ippt.gov.pl/~hpiotrzk.
Potential applications: The database can be used to test quantitative ultrasound techniques and ultrasound image processing algorithms, or to develop computer-aided diagnosis systems. © *2017 American Association of Physicists in Medicine* [https://doi.org/10.1002/mp.12538]

Key words: breast lesions, dataset, ultrasonic signals, ultrasonography

1. INTRODUCTION

Breast cancer is the most common cancer in women. According to a report by the World Health Organization (WHO), 0.5 million women die of breast cancer each year.[1] Independent of all other methods, ultrasound plays an important role in breast cancer diagnosis and monitoring, especially in differentiating between benign and malignant tumors.

To standardize the reporting process and make diagnosis more detailed, the breast imaging reporting and data system (BI-RADS) was introduced.[2] This classification takes into account several features of the lesion (shape, orientation, margin, lesion boundary, internal echo pattern, posterior acoustic features, vascularization of the lesion, and its elastic properties) and, based on them, indicates the probability of malignancy. Unfortunately, the BI-RADS scale is based only on ultrasound image analysis, and depends on operator experience and the quality of the aperture. Therefore, biopsy (which is an invasive procedure and, in many cases, is done unnecessarily due to misdiagnosis or ambiguity) is still a gold standard to confirm or exclude the presence of the breast cancer.

Nowadays, quantitative ultrasound (QUS) is becoming an important tool in breast lesion classification. It enables the estimation of tissue specific properties contained in the received echoes from the tissue under examination. However, the raw radio-frequency (RF) ultrasonic echoes (signals) are difficult to obtain in clinical practice, since the acquisition demands a dedicated ultrasound scanner. The Open Access Series of Breast Ultrasonic Data (OASBUD), presented in this paper, is a set of RF signals which were recorded from breast lesions and are now freely available for studying the specificity of the ultrasonic backscattered echoes from malignant and benign breast masses. Various methods have been

proposed in the literature to model the ultrasonic echoes.[3] In recent years, a number of articles have appeared on ultrasonic characterization of breast tissue using statistical properties of backscattered ultrasound echoes.[4–8] The OASBUD were previously used by our group to assess the statistical properties of ultrasound echoes and differentiate breast lesions. We have shown that the use of the homodyned K distribution may be helpful in differentiation of lesion type,[9] and that using the combination of the Nakagami distribution and the BI-RADS may improve the specificity of traditional ultrasonic examination.[10,11]

The authors hope that free access to the unique set of ultrasonic data will contribute to the development of new methods which will be helpful in distinguishing the type of lesion. As a consequence, we hope that this will contribute to a decrease in the number of unnecessary biopsies, which are often performed in a group of benign lesions. The dataset can be used to test QUS techniques and image processing algorithms, or to develop computer-aided diagnosis systems.

2. ACQUISITION AND VALIDATION METHODS

2.A. Subjects

The ultrasonic RF echoes contained in the OASBUD were recorded in the Department of Ultrasound, Institute of Fundamental Technological Research Polish Academy of Sciences, from 100 breast focal lesions of patients of the Oncology Institute in Warsaw. Seventy-eight women, aged from 24 to 75 yr (mean age of 49.5 yr) participated in this study. The data were collected between 11/2013 and 10/2015 and all patients provided informed consent before the examination. The study protocol was approved by The Institutional Review Board. Patients were examined by a radiologist with 18 years' experience in the ultrasonic examination of breast lesions, with numerous publications on this topic. American College of Radiology BI-RADS guidelines, and Polish Ultrasound Society standards, were employed during the examination.[2,12] All lesions which were examined and included in the OASBUD database were classified using the BI-RADS scale, which describes the probability of lesion malignancy. In a set of 100 lesions, 52 were malignant and 48 were benign. In the group of malignant lesions, all were histologically assessed by core needle biopsy. Thirty-seven benign lesions were also histologically assessed. According to current medical standards, the remaining 13 benign lesions did not qualify for a biopsy. Instead, they were observed by the radiologist over a 2-year period.

For all patients who participated in the project, detailed protocols were followed. They included information about age, type and number of lesions, result of the BI-RADS US classification, and history of the diseases which occurred in the family, especially breast cancer and ovarian cancer. However, in the case of the OASBUD, all patient identifying information was removed. Data which are presented in the database were marked using random identification numbers.

2.B. BI-RADS classification

The BI-RADS system takes into account several features of the lesion and describes the probability of malignancy. Attributes which are analyzed concern six morphological features (shape, orientation, margin, lesion boundary, internal echo pattern, and posterior acoustic features), vascularization of the lesion, and its elastic properties. Depending on the amount of features which characterize the lesion, it is classified to one of the BI-RADS categories by the radiologist.

In the analyzed set of masses, lesions categorized to class BI-RADS 4 were divided into three subgroups — 4a, 4b, and 4c — using the following guidelines. Lesions with only one suspicious feature were assigned to the group 4a indicating a likelihood of malignancy of less than 10%. Two-feature lesions were assigned to BI-RADS category 4b (probability of malignancy between 10% and 50%). Finally, when three or more features were found lesions were assigned to BI-RADS 4c (moderate-suspicious, expected to have a rate of malignancy between 50% and 95%). Lesions that were highly suggestive of malignancy were assigned to BI-RADS category 5.

2.C. Data acquisition

All sets of signals were acquired in such a way as to minimize the amount of shadowing and number of artifacts affecting the quality of the recorded data. However, for large lesions exceeding 40 mm, it was not always possible to expose clearly the lower edge of the lesion. For each lesion, two individual longitudinal and transverse scan planes, radial scanning around the nipple, and antiradial scan planes were done. Finally, two perpendicular scans (longitudinal and transverse) were recorded for every breast lesion using the Ultrasonix SonixTouch Research ultrasound scanner with an L14-5/38 linear array transducer (centre frequency 10 MHz). Single focusing beamforming was used with the focusing region always placed at the depth of the lesion. Each scan consisted of 510 RF echo lines. Signals were digitized with 40 MHz sampling frequency. The number of samples in every RF signal depended on the chosen penetration depth. For example, for a 40 MHz sampling rate the distance between adjacent samples is 0.0192 mm (assuming the speed of sound in tissue is 1540 m/s). Therefore, for a 5 cm penetration depth 2596 samples were collected, while for a 3 cm depth the number of samples decreased to 1558. Settings which were used by the operator had no influence on the characteristics of the raw ultrasonic echoes. Figure 1 shows an example of the B-mode image of a benign lesion displayed by the ultrasound scanner.

The specific individual region of interest (ROI) was indicated by the radiologist for each lesion scan.

2.D. Data validation

The RF data presented in this paper were acquired simultaneously with standard B-mode images during

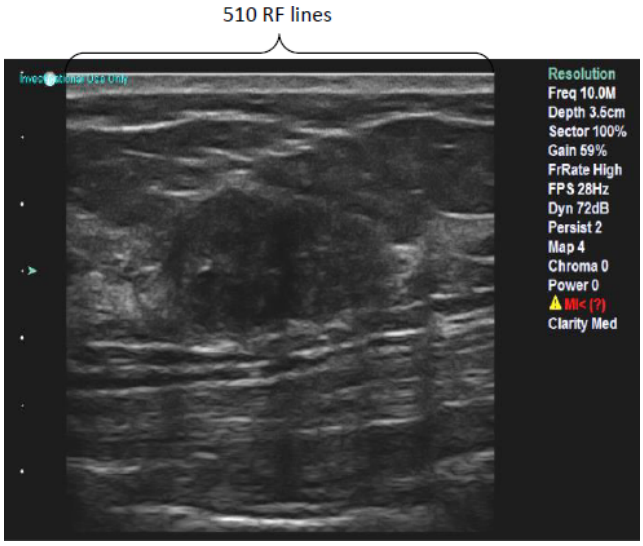

FIG. 1. The B-mode image of a benign lesion, localized at a depth of 10–20 mm, obtained using the Ultrasonix SonixTouch Research scanner with the L14-5/38 linear array transducer.

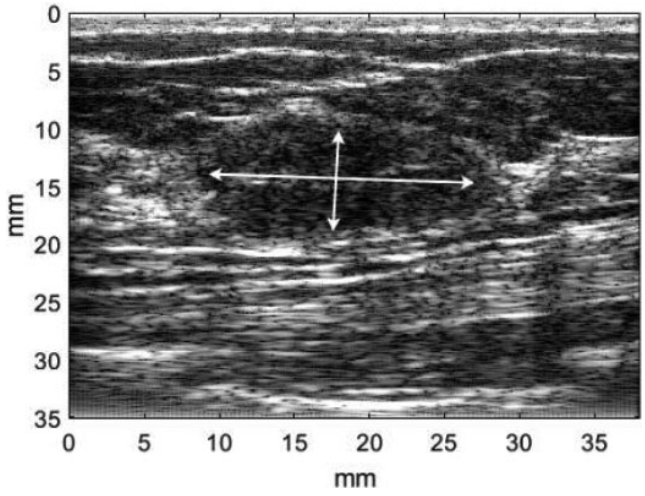

FIG. 2. The B-mode image of the lesion shown in Fig. 1, which was reconstructed from the set of RF signals (510 RF lines, 1824 samples in the each line, penetration depth equals 35.76 mm). White arrows indicate the lesion.

routine US breast examination scanning. The radiologist followed the American College of Radiology guidelines on the performance of US breast examination.[2] These guidelines describe how to position the breast during the examination or how to expose the lesion. According to current standards, the B-mode images obtained by one radiologist during the routine diagnostic procedures should be clearly interpretable by another radiologist or by a physician who is looking after the patient. The employed US medical scanner is widely used in hospitals. Our dataset includes raw data which are implemented into the scanner's algorithms for US image reconstruction. Figure 2 shows an image of a lesion reconstructed from RF data corresponding to the B-mode from Fig. 1. The image was reconstructed based on the set of 510 RF echo lines, however, image enhancement and noise removal were not applied. All procedures were performed in Matlab (MathWorks Inc, Natick, MA, USA).

The main limitations and uncertainties associated with the dataset are related to the acquisition method. We used the single focusing beamforming technique to obtain the RF data. As suggested in one study, the assessment of tissue scattering properties in the case of single focusing may be less sensitive in the region outside of the focal zone.[13] However, in our case, the focusing region was always placed at the lesion depth so this effect should not have a large impact on the estimation of breast lesion scattering properties. Other uncertainties are related to the BI-RADS categories and the ROIs which were chosen by the radiologist. These selections rely on the radiologist's experience and, theoretically, could be chosen differently by another radiologist.

3. DATA FORMAT AND USAGE NOTES

3.A. Data format

The RF scan files were recorded by the ultrasound scanner in the binary *.rf format. These *.rf files were converted into matrices of 510 columns and the number of rows which matched the penetration depth using Matlab. Next, the matrices containing the RF signals and the ROIs, IDs, type, and BI-RADS category of each lesion were stored in Matlab structure arrays and saved to a file called OASBUD.mat. For each scan, the ROI logical matrix has exactly the same size as the matrix which contains the RF signals. The array multiplication of the RF file and the ROI file (value one within the ROI and zero outside) removes from the resulting array any data corresponding to the surrounding tissues and leaves only the echoes related to the pathological region (Fig. 3). Finally, the structure with the data consists of the fields which are presented in Table I.

3.B. Usage notes

The database is freely available for viewing and downloading via the Zenodo repository (https://doi.org/10.5281/zenodo.545928) or the website http://bluebox.ippt.gov.pl/~hpiotrzk.

TABLE I. Structure of the OASBUD file.

Field	Description
Id	Patient unique ID
rf1	The first scan plane
rf2	The second scan plane
roi1	ROI of the first scan plane
roi2	ROI of the second scan plane
Birads	BI-RADS category
Class	Lesion class, 0 – Benign, 1 – Malignant

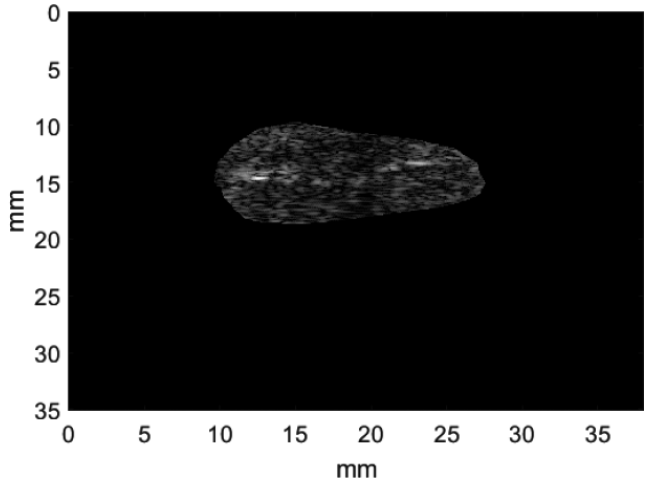

FIG. 3. B-mode image of pathologically changed fragment of breast tissue, which was reconstructed from the set of RF signals (510 RF lines, 1824 samples in each line) and multiplied by the ROI file.

It can be freely used by research laboratories, universities, students etc. In cases of presentation or publication of the results obtained with these data, a reference to this paper should be included. However, any reaping of financial benefits from the distribution of the data is prohibited.

Examples of methods of quantitative analysis of raw ultrasonic signals have been described in detail by several authors in previous papers.[9–11,14] Here, we provide two simple Matlab data-processing examples which may help other users to manipulate the data. The first example describes the reconstruction of the image in Fig. 3. The amplitude of the RF signal is calculated with the Hilbert transform and the amplitude is log-compressed and displayed. The Matlab code is as follows:

```
clc; close all; clear;
load('OASBUD.mat');
c = 1540; % speed of sound 1540 m/s
width = 38; % aperture width 38 mm
fs = 40e6; % sampling frequency 40 MHz
rf = data(1).rf1;
z_axis = 1000*linspace(0, size(rf, 1)*.5*c/fs, size(rf, 1)); % in mm
y_axis = linspace(0, width, size(rf, 2)); % in mm
envelope_image = 20 * log10(abs(hilbert(rf)));
envelope_image = envelope_image .* data(1).roi1;
figure;
imagesc(y_axis, z_axis, envelope_image);
colormap(gray); xlabel('mm'); ylabel('mm');
set(gca,'CLim',[40 80]);
set(gca, 'FontSize', 14).
```

The second example describes the steps used to plot the probability density estimate of sample amplitudes which were collected in the lesion region specified by the ROI. The following Matlab code can be used to reproduce the plot from Fig. 4:

```
clc; close all; clear;
load('OASBUD.mat');
rf = data(1).rf1;
rf_amplitude = abs(hilbert(rf));
rf_amplitude = rf_amplitude(data(1).roi1 == 1);
rf_amplitude = rf_amplitude / max(rf_amplitude);
f = histogram(rf_amplitude, 200, 'Normalization','pdf');
f.FaceColor = [0 0 0];
axis([0 0.2 0 20]);
xlabel('Amplitude');
ylabel('Probability density');
set(gca, 'FontSize', 14).
```

4. DISCUSSION

The presented dataset should prove valuable for all researchers interested in QUS and breast lesion classification. In this section, we will list and discuss several potential applications of the OASBUD. The usefulness of the database mainly concerns two activities. The first is the development of novel methods for QUS and breast lesion classification. The second is the comparison of different approaches already proposed in the literature.

The dataset was originally used to assess the statistical properties of ultrasound echoes in breast tissue with the Nakagami distribution[10,11] and the homodyned K distribution.[9] It would be interesting to compare these distributions with other models proposed in the literature,[7,15,16] and to select the one most suitable for modeling of breast tissue scattering properties. Moreover, researchers can use the RF signals to estimate other QUS parameters, for example, the attenuation coefficients.[17] In fact, various QUS parameters may be investigated to find the best one for the breast lesion classification.

The OASBUD can be also used to develop computer-aided diagnosis (CAD) systems for breast lesion classification. Usually, the CAD system involves four steps: image processing, lesion segmentation, feature extraction, and classification.[18] The database can be used to develop and test algorithms dedicated to each of these steps. The first code example is a good starting point for researchers who are

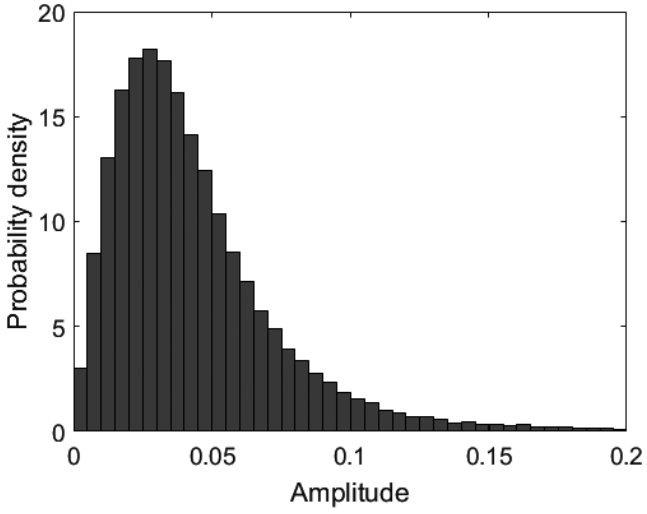

FIG. 4. The probability density estimate of the sample amplitudes for the lesion from Fig. 3.

interested in the B-mode image reconstruction and processing. The OASBUD can be used to test and compare algorithms used for ultrasound image enhancement,[19] for example, to determine efficient algorithm parameters. Researchers may incorporate the OASBUD into the development of their CAD systems or use it for validation. Results obtained on the data from various sources are usually more reliable. Moreover, as in the case of QUS methods, the OASBUD can be used to compare breast lesion CAD systems which have already been proposed.[20]

5. CONCLUSIONS

In this paper, the OASBUD database, which is a set of the raw RF ultrasonic echoes recorded from breast lesions, is presented. The signals were recorded from 78 women with 100 clinically diagnosed breast lesions. For each lesion, two orthogonal US scans were recorded. The data were stored in a single file which is freely available for all users via the Zenodo repository (https://doi.org/10.5281/zenodo.545928) or the website http://bluebox.ippt.gov.pl/~hpiotrzk. The collection of 100 breast lesions is sufficient for a wide range of computational analyses and the CAD system development. We hope that the publication of this dataset will be useful for everyone who is interested in QUS techniques, and that it will contribute to the development of new ultrasonic methods helpful in medical diagnosis.

ACKNOWLEDGMENTS

This work was supported by National Science Center, Grant Number UMO-2014/13/B/ST7/01271.

CONFLICT OF INTEREST

The authors do not have any conflict of interest.

[a)]Author to whom correspondence should be addressed. Electronic mail: hpiotrzk@ippt.pan.pl.

REFERENCES

1. Stewart BW, Wild CP. *World cancer report 2014*; 2014. doi: 9283204298.
2. Mendelson E, Bohm-Velez M, Berg W. *ARC BI-RADS Ultrasound. ACR BI-RADS® Atlas, Breast Imaging Report Data System*. Reston, VA: Am Coll Radiol; 2013.
3. Destrempes F, Cloutier G. A critical review and uniformized representation of statistical distributions modeling the ultrasound echo envelope. *Ultrasound Med Biol*. 2010;36:1037–1051.
4. Tsui PH, Yeh CK, Liao YY, et al. Ultrasonic Nakagami Imaging: a strategy to visualize the scatterer properties of benign and malignant breast tumors. *Ultrasound Med Biol*. 2010;36:209–217.
5. Liao Y-Y, Tsui P-H, Li C-H, et al. Classification of scattering media within benign and malignant breast tumors based on ultrasound texture-feature-based and Nakagami-parameter images. *Med Phys*. 2011; 38:2198–2207.
6. Tsui P-H, Yeh C-K, Chang C-C, Liao Y-Y. Classification of breast masses by ultrasonic Nakagami imaging: a feasibility study. *Phys Med Biol*. 2008;53:6027.
7. Shankar PM, Dumane VA, George T, et al. Classification of breast masses in ultrasonic B scans using Nakagami and K distributions. *Phys Med Biol*. 2003;48:2229.
8. Trop I, Destrempes F, El Khoury M, et al. The added value of statistical modeling of backscatter properties in the management of breast lesions at US. *Radiology*. 2015;275:666–674.
9. Byra M, Nowicki A, Wróblewska-Piotrzkowska H, Dobruch-Sobczak K. Classification of breast lesions using segmented quantitative ultrasound maps of homodyned K distribution parameters. *Med Phys*. 2016;43:5561–5569.
10. Dobruch-Sobczak K, Piotrzkowska-Wróblewska H, Roszkowska-Purska K, Nowicki A, Jakubowski W. Usefulness of combined BI-RADS analysis and Nakagami statistics of ultrasound echoes in the diagnosis of breast lesions. *Clin Radiol*. 2017;72:339.
11. Nowicki A, Piotrzkowska-Wroblewska H, Litniewski J, et al. Differentiation of normal tissue and tissue lesions using statistical properties of backscattered ultrasound in breast. In: *Ultrasonics Symposium (IUS), 2015 IEEE International*. Taipei: IEEE; 2015: 1–4.
12. Jakubowski W, Dobruch-Sobczak K, Migda B. Standards of the polish ultrasound society – Update. *Sonomammography examination. J Ultrason*. 2012;50:245–261.
13. Yu X, Guo Y, Huang S-M, Li M-L, Lee W-N. Beamforming effects on generalized Nakagami imaging. *Phys Med Biol*. 2015;60:7513.
14. Piotrzkowska-Wroblewska H, Litniewski J, Szymanska E, Nowicki A. Quantitative sonography of basal cell carcinoma. *Ultrasound Med Biol*. 2017;41:748–759.
15. Shankar PM. A statistical model for the ultrasonic backscattered echo from tissue containing microcalcifications. *IEEE Trans Ultrason Ferroelectr Freq Control*. 2013;60:932–942.
16. Shankar PM. A general statistical model for ultrasonic backscattering from tissues. *IEEE Trans Ultrason Ferroelectr Freq Control*. 2000;47:727–736.
17. Mamou J, Oelze ML. *Quantitative Ultrasound in Soft Tissues*. Netherlands: Springer; 2013.
18. Cheng HD, Shan J, Ju W, Guo Y, Zhang L. Automated breast cancer detection and classification using ultrasound images: a survey. *Pattern Recognit*. 2010;43:299–317.
19. Contreras Ortiz SH, Chiu T, Fox MD. Ultrasound image enhancement: a review. *Biomed Signal Process Control*. 2012;7:419–428.
20. Gomez Flores W, Pereira WCDA, Infantosi AFC. Improving classification performance of breast lesions on ultrasonography. *Pattern Recognit*. 2015;48:1121–1132.