Computing Speed-of-Sound From Ultrasound: User-Agnostic Recovery and a New Benchmark

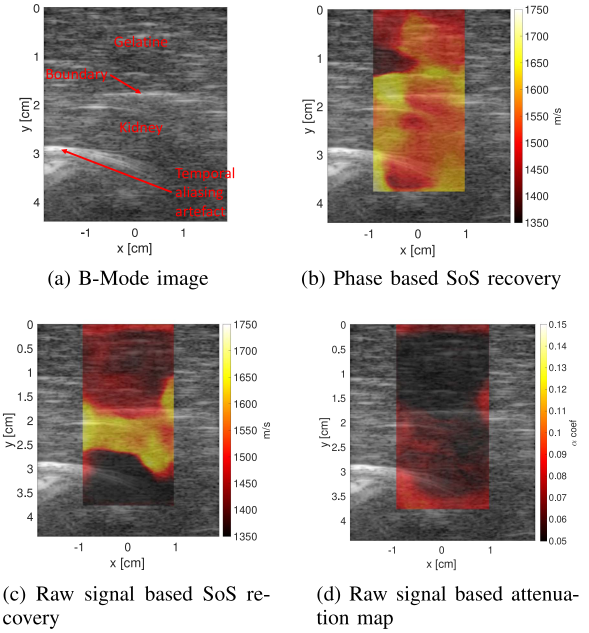
Micha Feigin, Daniel Freedman, *Member, IEEE*, and Brian W. Anthony

Abstract*—*Objective: **Medical ultrasound is one of the most accessible imaging modalities, but is a challenging modality for quantitative parameters comparison across vendors and sonographers. B-Mode imaging, with limited exceptions, provides a map of tissue boundaries; crucially, it does not provide diagnostically relevant physical quantities of the interior of organ domains.This can be remedied: the raw ultrasound signal carries significantly more information than is present in the B-Mode image. Specifically, the ability to recover speed-of-sound and attenuation maps from the raw ultrasound signal transforms the modality into a tissue-property modality. Deep learning was shown to be a viable tool for recovering speed-of-sound maps. A major hold-back towards deployment is the domain transfer problem, i.e., generalizing from simulations to real data. This is due in part to dependence on the (hard-to-calibrate) system response. *Methods:* We explore a remedy to the problem of operator-dependent effects on the system response by introducing a novel approach utilizing the phase information of the IQ demodulated signal. *Results:* We show that the IQ-phase information effectively decouples the operator-dependent system response from the data, significantly improving the stability of speed-of-sound recovery. We also introduce an improvement to the network topology providing faster and improved results to the state-of-the-art. We present the first publicly available benchmark for this problem: a simulated dataset for raw ultrasound plane wave processing. *Conclusion:* The consideration of the phase of the IQ-signals presents a promising appeal to traversing the transfer learning problem, advancing the goal of real-time speed-of-sound imaging.**

***Index Terms*—Deep learning, inverse problems, speed-of-sound inversion, ultrasound.**

I. Introduction

MEDICAL ultrasound imaging is one of the most widespread and accessible modalities due to its low cost and the availability of portable hardware. It does not use ionizing radiation, as is the case with X-ray imaging, nor does it use strong magnetic fields, requiring special precautions and shielding, as is the case with MRI imaging. It is also one of the hardest modalities with which to compare quantitative parameters across vendors or sonographers. The most common ultrasound modality is the B-Mode (brightness mode) image (see Figs. 1 and 2). The B-Mode image is constructed by transmitting an acoustic pulse and plotting the log amplitude (decibel domain) of the reflected signal, where the imaging plane is a 2D slice extending away from the probe into the body.



(a) B-Mode image

(b) Phase based SoS recovery

(c) Raw signal based SoS recovery

(d) Raw signal based attenuation map

Fig. 1. Sheep kidney in gelatin phantom: benefits of speed-of-sound and attenuation information. The background speed-of-sound is 1500 m/s and the speed-of-sound in the Kidney is 1560 m/s. Other than the bundary, the kidney cannot be differentiated from the background in the B-Mode image (a) Shows the annotated B-Mode image, top edge of the kidney is visible 1.5 cm from the top (b) show the Speed-of-sound map (m/s) using phased based inversion (c) speed-of-sound map using raw data inversion, and (d) the attenuation map (alpha coefficient). (a) B-Mode image. (b) Phase based SoS recovery. (c) Raw signal based SoS recovery. (d) Raw signal based attenuation map.

The information in the B-Mode ultrasound image consists of a combination of speckle density (echogenicity) and discontinuities in the domain [1]. The construction of the ultrasound image assumes a (constant) known speed of sound. An image is distorted due to variability in the speed of sound. Discontinuities between organs and bone surfaces are relatively specular. As such, these may not appear in the image (Fig 8) or produce artifacts and ghost reflections due to interreflections and refractions (Fig. 1(a)).

Manuscript received 24 January 2023; revised 11 May 2023 and 8 August 2023; accepted 9 October 2023. Date of publication 24 October 2023; date of current version 21 March 2024. *(Corresponding author: Micha Feigin.)*

Micha Feigin is with the Department of Mechanical Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139 USA (e-mail: michaf@mit.edu).

Daniel Freedman is with the Verily Research, Israel.

Brian W. Anthony is with the Department of Mechanical Engineering, Massachusetts Institute of Technology, USA.

Digital Object Identifier 10.1109/TBME.2023.3327147

Our goal is an ultrasound-based image reconstruction capable of imaging the spatially varying, diagnostically relevant, tissue properties throughout the domain, in real-time. Natural candidates are elastic coefficients [2] such as Young's modulus and the shear modulus. These are measures of resistance to deformation and, under the soft tissue regime (nearly incompressible elastic model), are closely correlated; where Young's modulus is roughly equal to three times the shear modulus. Viscoelastic properties [3] and scattering manifest as signal attenuation.

Ultrasound shear wave elastography (SWE) [4], [5] has been used to approximately measure the shear and Young's modulus. This is done by generating (mechanical) shear waves in the tissue and subsequently tracking their propagation using acoustic (pressure) waves. This approach does suffer from several serious drawbacks. SWE is highly sensitive to sonographer and subject motion, subject to low frame rates, has high power requirements, and requires more costly hardware.

To address these limitations, we focus instead on the local pressure wave (acoustic) speed of sound. While the shear wave speed depends only on the shear modulus (and density), the pressure wave speed depends on both the bulk and the shear modulus (and the density) [2]. As the bulk modulus also carries diagnostic merit [6], this can be thought of as both an alternative and a complementary method to SWE. Our approach enables significantly higher frame rates, potentially in the hundreds of frames per second, as opposed to the single-digit frame rates of SWE. On the downside, traditional methods for recovering speed-of-sound require large imaging apertures (large probes or tomographic setups), a considerable computational load, and often person-in-the-loop processing.

A deep learning-based approach combined with simulated data has been proposed to address this problem [7]. The authors have shown that given appropriate training data, a deep learning approach can provide a high frame rate approximator to the speed of sound inversion problem. As we show in our experimental results, similarly to the classical approaches, this is sensitive to good calibration for the system response, both the pulse shape and pulse amplitude. This is the domain transfer problem [8], i.e., the ability to generalize results from simulation-based training to deployment to real-world data. This implies that parts of the information pertaining to speed of sound lies in the way the pulse deforms during propagation.

In this paper, we propose a novel method to tackle the problem of sonographer and imaging-procedure independent learning, greatly simlifying the system calibration problem. We focus on operator-dependent parameters that cannot be calibrated for, mainly amplification and gain. Amplification (global amplitude multiplier) and attenuation (time-dependent multiplier) are highly dependent on the force applied by the sonographer, the angle of incidence, and the amount of acoustic gel used. Signal loss due to attenuation is also spatially dependent, and spatial scattering is difficult to correctly model with a 2D simulation. We present initial results for addressing variability in pulse shape through preprocessing by match filtering and Wiener filtering. We provide initial stability tests on real data; however, due to the current lack of calibrated real data, training and testing with real data are left for future work.

Four major aspects are involved in the system response 1) the acoustic pulse center frequency 2) pulse bandwidth 3) system efficiency, transmit power, and impedance matching (scalar multiplier) resulting in fixed amplitude scaling, and 4) attenuation, scattering, and time-dependent gain, producing time variable scaling. While unknown attenuation and gain factors can be somewhat trained for in simulations (Table II), the results do not translate well to real data (Fig 8). We propose to use the in-phase and quadrature (IQ) demodulated signal [9]. To decouple gain and attenuation, rather than using the complex IQ demodulated signal, we use only the phase, or argument, of this complex IQ signal. We show that this approach transfers significantly better to real data, providing both better and more stable results. to show the decoupling of attenuation information from the signal, we test the ability to train the network to recover attenuation information from both the raw and IQ-phase signals, showing that this endeavor fails on the IQ-phase signal. Deep learning paves the way for such new approaches that do not fit well in the classical inversion model that depends on amplitude-phase fitting.

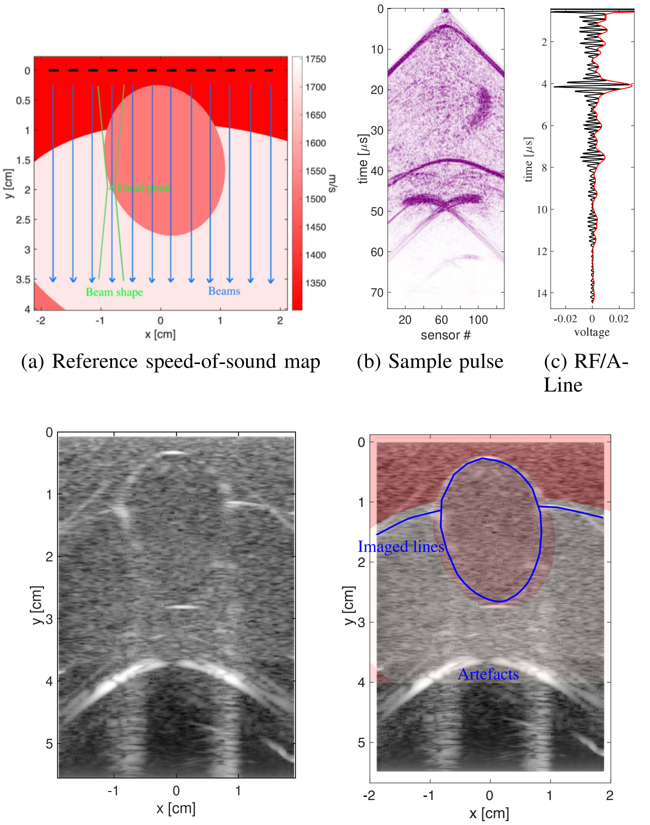
An interesting outcome of our experiments is that the network is highly dependent on both the center frequency and the bandwidth of the pulse. This is to the point that the network fails to train on a mixed-bandwidth data set even when the center frequency is kept fixed. This implies that critical information is obtained from the pulse shape. Applying a Wiener filter [10] to translate pulse shapes does allow the application of a given network signal produced by a system using a different pulse shape. This paves the path for further signal pre-processing and self-calibration based on spectral content, probe modeling, and matched filtering, but is beyond the scope of this work and is left for future research. Our speculation from the results is that while the reflection amplitude itself provides unstable information, the sign of the reflection (positive or negative reflected wave) and consistency of the pulse shape provide additional information on the sign and magnitude of the discontinuity in the speed of sound at the boundary.

We also further optimize the network topology presented in [7], [11] (Fig. 3). Our network maintains the U-net topology with an input geometry of 64 channels by 2048 time samples, and an output speed-of-sound map with a resolution of 128 by 256. In contrast to previous works, each network layer consists of three levels of three-by-three convolutions. We show a significant improvement over the existing results in both accuracy and processing speed.

Finally, we introduce the first publicly available simulated raw signal benchmark dataset for speed-of-sound and attenuation mapping. This dataset is described in Section III. Although using a simulated dataset has its limitations, the physics of acoustic image formation is well understood; it is hard to impossible with current technologies to acquire a significant real-data training set with ground truth.

A. Background and Related Work

a) Ultrasound image formation: An ultrasound image is generated by transmitting an acoustic pulse into the tissue [1], often modeled as a sine wave multiplied by a Gaussian or Hanning window. The most common imaging method is the B-Mode (brightness) image, shown in Fig. 2(c). This is an image (spatial map) of the (log) amplitude of the measured acoustic



(a) Reference speed-of-sound map

(b) Sample pulse

(c) RF/A-Line

(d) B-Mode image

(e) B-Mode with reference speed-of-sound

Fig. 2. B-Mode ultrasound imaging model. (a) shows the generating speed-of-sound map, with the scanning beam model (blue) and actual beam shape (green), (b) shows a sample raw "channel data" beam, (c) shows the matching (cropped) RF line (black) and A-Line (Red) (d) is the resulting B-Mode image. (e) Shows the overlaid speed-of-sound model and the imaged discontinuities (blue) showing the distortion due to incorrect imaging speed-of-sound. (a) Reference speed-of-sound map. (b) Sample pulse. (c) RF/A-Line. (d) B-Mode image. (e) B-Mode with reference speed-of-sound.

reflections. Acoustic reflection results from discontinuities in the acoustic impedance (speed-of-sound multiplied by the density in the acoustic case). For direction incidence (1D case), the reflection amplitude is roughly proportional to the difference in acoustic impedance and can be described by the Zoeppritz equations in the more general 3D elastic case [12].

The two main sources of information in the ultrasound image are due to discontinuities in the speed-of-sound of the tissue and speckle "noise". Speed-of-sound varies when transitions through different tissues such as organs, fat, tumors, and muscles, creating the edge image. Speckle "noise" results from multiple reflections from densely distributed (nearly) point scatterers, often referred to as the tissue echogenicity.

Significant work has been done on using speckle for quantitative diagnosis in various domains [13], [14], [15], [16], [17], [18], [19]. However, speckle structure is inherently sensitive to the choice of both hardware and imaging parameters. As a result, other than extreme cases such as cysts with no speckle and fibrosis that severely degrades image quality and clarity, the requirement for closely controlled capture severely limits the ability to translate results to real-world scenarios.

Discontinuities in speed-of-sound may distort the image due to refraction, and impact image quality due to the breakdown in the imaging assumptions (constant speed-of-sound). Their specular nature can occlude them in some scenarios. Interreflection can produce ghost images and other artifacts. Fig. 2 presents the B-Mode image formation model. Fig. 2(a) is the generating speed-of-sound (used for the simulation), and Fig. 2(d) shows the resulting B-Mode image. The far end of the central ellipse is almost completely missing, and other discontinuities are hard to spot. Edges are misplaced due to an incorrect choice of the speed of sound used in the delay-and-sum image formation model. Fig. 1(a) also shows an example of temporal aliasing artifacts, where reflections outside the field of view are displaced in time, and as a result, distance, as they arrive after the following pulse has already been transmitted.

A general-purpose ultrasound probe is an array of individual transducer elements, most often constructed of 128 to 192 piezoelectric elements that act as both transmitters (speaker) and receivers (microphone). The B-Mode image is generated by interrogating the domain with an acoustic beam (Fig. 2(a)). The actual beam width (transverse resolution) is controlled by the transmit frequency, and the transmit and receive apertures. Individual beams are generated by transmitting using a (usually small) number of elements at a time. The resulting per-element signal is called the channel data (Fig. 2(b)). This in turn is converted to an RF signal line via delay and sum focusing (digital lens, a technique known as Kirchoff migration in the seismic domain [12]). Exponential time-gain correction (TGC) is applied to compensate for tissue attenuation and scattering. To this, in turn, envelope detection is applied to produce the A-Line (amplitude line). The result is then filtered to reduce speckle noise and rasterized (interpolated) to produce the B-Mode (brightness) image.

b) Shear-wave elastography: B-mode provides approximate boundaries and structure, not information about the properties of the tissue. While images depend on physical properties, the inverse relationship, determining properties from the received signal, is non-trivial. Such quantitative inversion is beneficial in determining tissue health, e.g. benign vs malignant tumors [20]. The current state of the art for quantitative imaging is Shear-Wave Elastography (SWE) [4], [5]. Elastography is a method to create a tissue property map. It measures tissue stiffness (shear modulus, and under some assumptions, Young's modulus), and can be thought of as quantitative remote palpation. SWE is performed by transmitting a slowly traveling mechanical shear wave (on the order of 1–10 m/s in healthy tissue [21], [22]), and tracking the wavefront using the much faster ultrasound pressure wave (having a mean speed-of-sound commonly in the range of 1450–1650 m/s [1]); a stroboscopy approach of sorts. SWE has been shown to have significant diagnostic abilities [23], [24], [25], [26], [27], [28], [29], this method is extremely sensitive to subject and sonographer motion [30], suffers from very low frame rates, and has difficulty imaging deep tissue. Due to high power requirements, it is also limited to more expensive systems with specialized electronics.

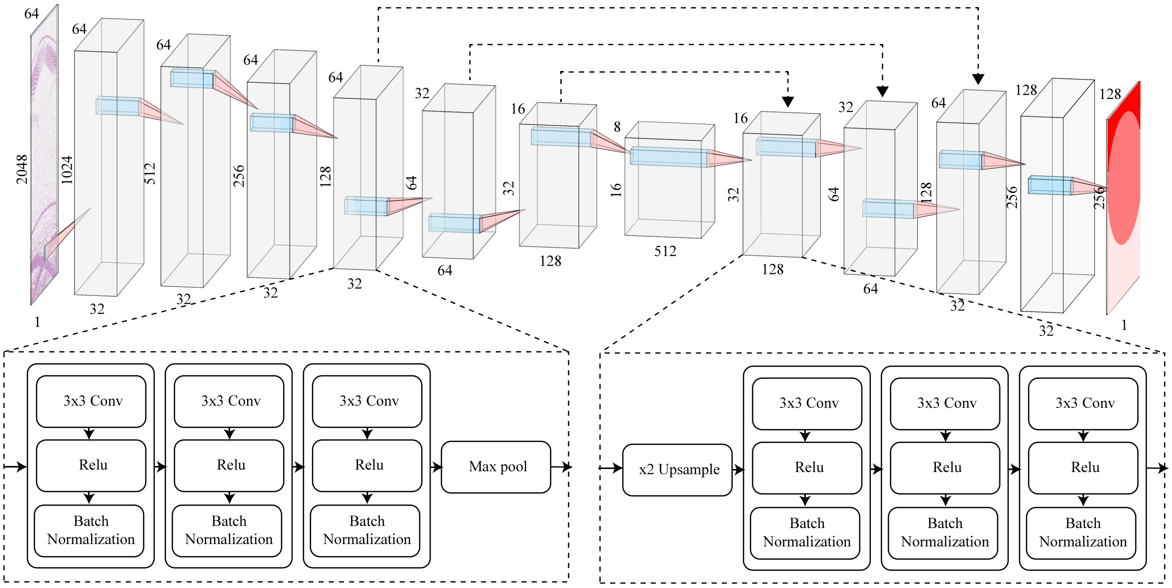


Fig. 3. We utilize a U-net encode-decoder type topology. Input is a 64 × 2048 raw channel data or phase image, and output is a 128 × 256 speed-of-sound map. Each up/down sampling block consists of three consecutive convolutions, Relu activation, and batch normalization.

c) Speed-of-sound inversion: The speed of sound in soft tissue generally varies in the range between 1450 m/s and 1650 m/s. Where in soft tissue, the shear-wave speed depends mostly on the shear modulus and that of the pressure waves depends on both the shear and bulk modulus. As such, the speed of sound in fat is closer to the lower end of this range, while that of muscle and tumors is on the higher end [31]. Recent research has shown that speed-of-sound values have a diagnostic ability similar to that of shear wave elastography [20], [31], [32], [33].

Ultrasound image formation is most commonly done under the assumption of a constant speed of sound of 1540 m/s. Deviation of the actual speed of sound in tissue from this value affects image quality and results in a distorted image. The implication is incorrect physical measurements when assessing things such as organ dimensions. While longitudinal distortion due to deviation from the correct speed of sound is minimal, as the variability is averaged, transverse distortion can be significant when multiple refractions are present. This requires significant experience on the part of the sonographer in choosing an appropriate imaging orientation, which is not always physically possible.

Using the speed-of-sound maps has a significant advantage over SWE in that the speed of sound is recovered directly from the pressure wave signal. As a result, speed-of-sound techniques are capable of significantly higher frame rates. Current inversion techniques however are extremely computationally intensive, require multiple frames and/or large amounts of data, and depend on a good system response model, calling for novel methods both in the medical and seismic domains.

A correct, or at least improved, speed of sound map, can also be used to correct image distortion and improve image quality.

Longitudinal speed-of-sound inversion has a long history in the seismic domain, both for imaging and for inferring ground characteristics. These methods can be roughly categorized into two groups. The first, full waveform inversion (FWI), is a differential inversion technique capable of high resolution but is extremely sensitive to the initial conditions. The second, travel time tomography, is an integral first arrival method, i.e., an inverse Radon transform without the straight ray assumption of x-ray computerized tomography (CT); as such, it produces much smoother results but struggles with recovering discontinuities. In imaging terminology, FWI is sensitive to the blank wall problem, i.e., cannot recover information where there is no strong reflector. Travel-time tomography is sensitive to limited aperture reconstruction. It requires a large aperture, and cannot recover variability in orientations where there are no cross beams. Both have been investigated in the context of ultrasound imaging, significantly in breast imaging [32], and in limb imaging [34]. These systems require significant computations, special capture topology, and custom hardware. Results have also been presented using a standard probe topology based on various travel time methods [35], [36], [37], [38], [39]

Recently, deep learning methods for speed-of-sound inversion, capable of real-time single-sided imaging, have been introduced in the ultrasound domain [7], [11], [40], and have been replicated in the seismic domain using mostly the same approach [41], [42]. Generalization to real data however is still limited, due, among others, to dependence on system parameters and limitations of simulation-based methods.

B. Contributions

Our contributions presented in this paper are as follows:

- We present a new signal preprocessing approach utilizing the phase information of the IQ demodulated signal. This decouples the attenuation and gain due to the sonographer interaction and the unknowns in the tissue scattering and the viscoelastic response by removing the amplitude information. This, in turn, improves the generalization of simulation-based results to real data providing more procedure-agnostic results.

- We present an optimized speed-of-sound inversion network topology in medical ultrasound imaging, significantly improving on the state-of-the-art in both performance and processing speed.
- We present the first large-scale public benchmark (synthetic) dataset to test and compare plane-wave methods in medical ultrasound.

II. Method

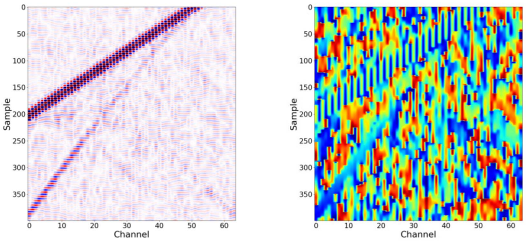
Toward our goal of generating speed-of-sound and attenuation maps (inversion), we utilize a U-net-type encoder-decoder network [43]. We present our network setup in Fig. 3. The input is the 64 raw channel data traces by 2048 samples or the corresponding 64×2048 IQ-phase data traces. The output is a 128×256 speed-of-sound map. Each encoder block layer consists of three layers. Each of these, in turn, chains a 3×3 convolution, a Relu activation function, and batch normalization. In the first four-layer blocks, reduction in the time axis is performed using a strided convolution in the last layer (without downsampling in the channel, or spatial, axis). In subsequent layers, we use 2×2 max pooling. The decoder path consists of a factor of two upsampling, followed again by three layers consisting of a 3×3 convolution, a Relu activation function, and batch normalization. The final stage is a 1×1 convolution to produce the output image. Skip connections are applied to the inner three layers.

We use a sum of squared differences (MSE) for the loss function for training. To assess the results, however, we utilize a mean absolute error measure, as this provides a significantly more meaningful result. The root mean square error (RMSE) is sensitive to outliers resulting from small inconsistencies in the boundary location. Small discrepancies are less important considering the variation in scale between the simulation and reference resolution as well as discrepancies that are caused by small errors in the recovered speed-of-sound map.

Training is performed on a synthetic dataset of 9216 (9×1024) training samples, 1024 testing samples (see Section III for more details). The results are then validated on a 1024 sample validation set. The data are produced by randomly generating the speed-of-sound and attenuation maps, numerically simulating the resulting channel data signals, cropping and scaling the speed-of-sound map and simulated signals to the expected input and output resolutions, and switching roles between input and output for training.

The signal is preprocessed during the training step as follows to minimize over-training and improve transfer learning:

1) (Optional) matched filter
2) Gaussian (white) noise, random amplitude normally distributed between 1% and 100% of the signal standard deviation
3) Quantization noise (multiply the signal by 4096 and round to digitize the signal)
4) Random channel drop of 0–2 channels



(a) Channel data (b) Phase data

Fig. 4. Comparison of (a) raw channel data and (b) post-processed phase data.

We found that matched filtering is too efficient at denoising during training and induces over-training if applied post-noise. The signal is then IQ-demodulated for processing.

A. Phase Based Inversion

Sonographer interaction, physics, and electronics dictate four main sources of scan-to-scan variability in the ultrasound signal, even when using the same hardware 1) fixed gain (constant scaling), 2) time-dependent attenuation, scattering, and gain correction, 3) pulse center frequency, and 4) pulse bandwidth. While we know the center frequency and bandwidth, attenuation and gain are difficult, or impossible, to fully calibrate in the clinical setting, making deployment significantly harder.

As shown by the testing results, speed-of-sound recovery from raw ultrasound data is sensitive to correct, or at least consistent, gain correction and scaling to a point that is nearly impossible to achieve in real-world conditions. This also means that recovery is sensitive to the quality of probe-skin contact.

Training the network for resilience to a constant gain factor by randomly scaling input signals, somewhat works, although not well (Table II. Attempting to do the same for dealing with variability in the time-gain correction performs even worse. Neither results translate to real data.

We propose a different approach for removing most of the system response from the signal. We perform IQ demodulation of the signal, i.e generating the in-phase (I) and quadrature (Q) components by multiplying the signal by a cosine and sine respectively, and applying a low pass filter [9]. We next use the complex notation $s = I + iQ$, taking the argument of the results, i.e. $\theta = \tan^{-1}(I, Q)$, and train the network on this phase component. This almost completely decouples the fixed gain, attenuation, time gain factor, and center frequency from the data. The effect is shown in Fig. 4. We presume that most of the information pertaining to the existence of the pulse lies in the signal-to-noise level of the resulting signal. Interestingly, the resulting network is sensitive to the actual pulse frequency and shape (bandwidth), where we found it impossible to train the network effectively to deal with different pulse shapes at the same time. This can be mitigated by applying a Wiener filter to shape the date as expected.

Worth noting, while we expected phased wrapping artifacts to cause issues, applying phase unwrapping to the phase signal deteriorated the results significantly. This is due to a phase drift effect, i.e. the phase signal is not a zero-mean process, with a significant drift in mean value over time. As a result phase unwrapping drowns out the important signal. Our experiments have shown that applying the network to the raw phase signal worked significantly better than any manual prepossessing of the unwrapped signal that we tested.

B. Joint SoS and Attenuation Recovery

As a means to test the hypothesis that most of the information pertaining to attenuation and time-gain correction is removed from the signal using our proposed approach, we look at the network's ability to recover the attenuation coefficient. To this end, we compare attenuation recovery and joint speed-of-sound and attenuation recovery from the raw signal, to attenuation recovery from the phased data.

When performing joint speed-of-sound and attenuation inversion, to compensate for the fact that the losses differ by several orders of magnitude, with the speed-of-sound MSE factor on the order of 1000 on a trained network, while the α attenuation MSE factor is on the order of 0.05. We thus use a weighted sum of the squared difference loss function, multiplying the difference in attenuation fields $\lambda = 4 \cdot 10^4$.

$$L = \|C - C_0\|_2^2 + \lambda \|\alpha - \alpha_0\|_2^2 \qquad (1)$$

where C and α are the speed of sound and attenuation values recovered by the network, C_0 and α_0 are the known target values, and λ is the normalizing factor.

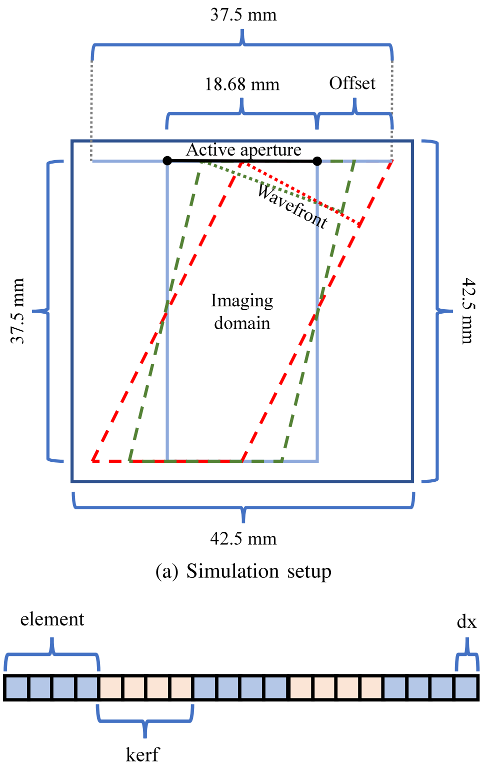
III. Dataset and Code

The code and model [44] and the data [45] are publicly available for download. The full dataset consists of 112640 simulations split into 9216 simulations in the training set, 1024 in the validation set, and 1024 in the test set. The measured signal is simulated using the k-wave [46] MATLAB toolbox. Simulations were performed for nine plane waves at 0, ±8, ±16, ±24, and ±32 element offsets, with corresponding wavefront angles of 0, ±6.7, ±13.7, ±20.2, and ±26.3 (the time delay is calculated based on 1540 m/s so the actual angle will differ per sample), set to pass through the center of the domain. See Fig. 5 for details (three of the 9 plane waves are shown to reduce clutter). Each simulation was performed with two center frequencies, 2.5 MHz and 5 MHz, with a Gaussian window (pulse width) of 5 oscillations.

A. Simulation Setup

Each simulation comprised of 1152×1152 random speed-of-sound and α (attenuation) coefficient maps following power law attenuation [$\mathrm{dB}/\mathrm{cm}/\mathrm{MHz}^2$] in a domain 42.35×42.35 mm in size (see Fig. 5(a)).

The domain is constructed by layering a randomly selected set of ellipses and half-planes. For each of the resulting domains (organs), we randomly selected the speed of sound, the attenuation coefficient, the speckle density, and the speckle amplitude. Domains were verified to not slice the probe face; i.e. the resulting maps are verified not to have a discontinuity at the probe face. A sample speed-of-sound map and simulation geometry are shown in Fig. 2(a). Cropped sample speed-of-sound maps used as the recovery targets are shown in Fig. 6(a) and (b).



(a) Simulation setup

(b) Array

Fig. 5. Image (a) shows the k-wave simulation setup. The US array is placed at line 60 of the numerical grid. Due to kerf, slightly less than half of the array (64 elements) is excited to generate the outgoing plane wave. To better match the actual signal and avoid artifacts, a continuous section is excited. The angle is set based on an assumed 1540 m/s speed of sound so that the plane wave overlaps the center of the domain. Image (b) shows the array structure, with four active elements and four kerf elements interleaved. The recorded signal is the average of the four receiving cells for each element.

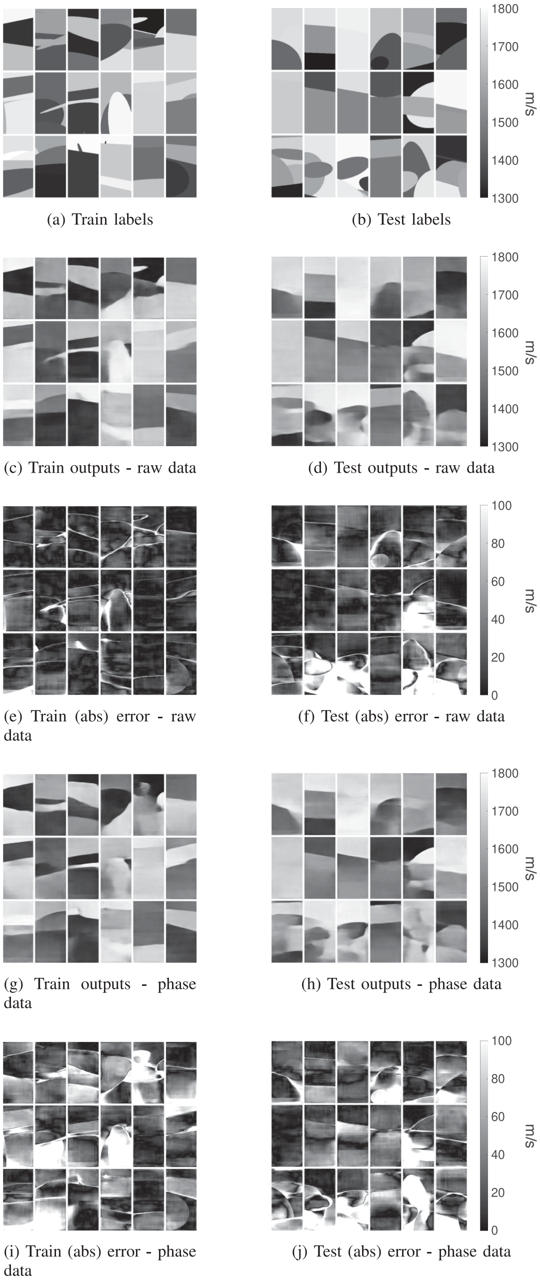
The speed of sound range is 1300 {m/s} to 1800 {m/s}. The α coefficient range is 0.05 to 0.15 $\mathrm{dB}/\mathrm{cm}/\mathrm{MHz}^2$. The background density is set to 0.9 g/cm^3 (density of fat).

Speckle noise is randomly generated in the density domain so as not to affect the wavefront propagation speed (uniformly distributed point sources with 2–10 points per wavelength and uniformly distributed amplitude at $\pm 10\%$).

B. Probe

To match our physical hardware, we simulated a 128-element array with 64 active transmit elements. The simulation was carried out with two pulse center frequencies, 2.5 MHz and 5 MHz with a Gaussian window of 5 oscillations.

The central plane wave (zero degrees) is centered at elements 33 to 96. The probe face is placed at $y = 60$ (outside the perfectly matched layer) and centered on the x axis. The numerical receive array is 4 elements per sensor element, with a matching kerf (spacing) value, i.e., 4 on 4 off (see Fig. 5(b)). The signal for each receiver is summed across the 4 receiver elements to generate



(a) Train labels (b) Test labels

(c) Train outputs - raw data (d) Test outputs - raw data

(e) Train (abs) error - raw data (f) Test (abs) error - raw data

(g) Train outputs - phase data (h) Test outputs - phase data

(i) Train (abs) error - phase data (j) Test (abs) error - phase data

Fig. 6. Sample results: SoS recovery from raw data. Each grid shows the matching results on 7 (H) by 6 (W) speed-of-sound maps. Signal in simulation travels top to bottom (probe is at the top pointing down) (a) and (b) show the true speed-of-sound map (cropped and scaled from the original) for the train and test sets respectively. (c) and (d) show the recovered speed-of-sound maps using raw input. (g) and (h) show the same for IQ-phase input. (e) and (f) show the absolute error for the recovery on the raw signal and (i) and (j) show the error on the IQ-phase signal. Speed-of-sound is capped at 1300–1800 m/s and the error at 0–50 m/s.

the 128 receive channels, and the signal is down-sampled to a 40 MHz sampling rate (ADC rate). For the transmit signal, we use a continuous array, as we found that it better matches real-world signals, so for the centered plane wave, a source is placed on all pixels with $y = 60$ and $322 \leq x \leq 830$ with a zero time delay on all elements.

IV. Experimental Results

Based on results from [7], within the current framework combining multiple plane waves is best done by averaging results over all views. As such, there is no added value to address multiple plane waves for this assessment, so in the interest of reducing clutter and maximizing clarity in the experimental results we focus on assessing reconstruction using a single plane wave at 0 degrees.

We start by looking at the performance of both our new optimized network topology (Fig. 3) and the proposed IQ-phase preprocessing. The results for the training and testing synthetic datasets are presented in Fig. 6 and Table I. Results are compared to previous works [7] (denoted as TBME) and [11] (denoted as EMBC). All implementations are for a single view plane wave at direct incidence (propagation direction longitudinal to the probe face). To isolate the two modifications, we compare speed-of-sound recovery from raw channel data as was done in prior works (effects of network topology), and recovery from the IQ phase data. All error values are mean absolute errors and are measured in m/s. We use this error measure as it is less sensitive to small displacements in the location of the discontinuities in the images than the root mean square error and thus more indicative of the quality of the results in this context as it gives a better indication of the expected error in local speed-of-sound. To assess sensitivity to depth (both decreased aperture size and increased additive error), in addition to global error, we break down error values by depth. We split the domain into three non-overlapping distance bands (a third of the distance range for each), denoted as near, mid, and far. All networks were trained on the same dataset. It should be noted that for this work we employ a more difficult dataset than previous works. The dataset uses both elliptical and linear reflectors, spatially varying random attenuation (α coefficient), and spatially varying random speckle noise levels. Training and testing were performed on the data with a fixed time-gain profile.

As can be seen in the results, the optimized network topology is significantly better than the first results in the field [7] (TBME), and also better than the improved results presented in [11] (EMBC). Using raw data does perform better than IQ-phase based recovery in this case where there is a consistent system response (i.e. controlled and fixed gain profile), however, not significantly so. Run times are roughly half those of [11] and about equivalent to [7]. Results do degrade with distance, suggesting that as expected, effective depth will depend on array size. This suggests that replacing the raw signal with the IQ-phase signal can produce at least comparable results.

Next, we test how well the system can be trained to handle a varying scaling and gain profile, that is, address the uncontrolled system response we are aiming to address with this work. We do this by randomly varying the scaling factor and time-gain profiles. This is where the advantage of the proposed IQ-phase input becomes apparent. Results are shown in Table II. To simulate the variability in real data, we compare the previous network training results (raw), to the results when scaling each input signal by a random constant, uniformly sampled in the range 10^{-2} to 10^{2} (raw w/ scale), and applying both a randomly

TABLE I
COMPARING THE SPEED OF SOUND RECOVERY FOR THE TRAINING AND TESTING DATASETS

Method	Mean	Near	Mid	Far
Train				
Ours - raw	9.2	8.7	9.4	9.3
Ours - phase	12.3	11.2	12.5	13
TBME	24.2	23	26.2	23.4
EMBC	13.6	13.2	14	13.6
Test				
Ours - raw	32.5	21.8	32.9	42.7
Ours - phase	40.7	29.5	39.5	52.8
TBME	68.5	45.6	77.4	82.2
EMBC	43.8	27.9	47.6	55.8

Results shown for prior works, [7] (TBME) and [11] (EMBC). These are compared to the modified network Using raw channel data input (ours - raw), as with prior works, and using the proposed IQ-phase input (ours - phase). All error Values are in mean absolute errors in m/s, broken down by Global mean, and near, mid, and far thirds of the distance.

TABLE II
EFFECTS OF APPLYING RANDOM SCALING AND RANDOM GAIN COEFFICIENTS TO THE INPUT RAW CHANNEL DATA SIGNAL

Method	Mean	Near	Mid	Far
Train				
Phase	12.3	11.2	12.5	13
Raw	9.2	8.7	9.4	9.3
Raw w/ scale	6.6	6.4	6.6	6.7
Raw w/ gain	25.3	32	24.4	19.4
Test				
Phase	40.7	29.5	39.5	52.8
Raw	32.5	21.8	32.9	42.7
Raw w/ scale	48.3	38.2	47.8	58.8
Raw w/ gain	58.5	51.3	58.4	65.7

Errors are absolute mean errors measured in m/s.

TABLE III
RESULTS FOR RECOVERING ATTENUATION FROM RAW DATA VS. PHASE

Method	Mean	Near	Mid	Far
Train				
Raw	0.003	0.003	0.003	0.003
Phase	0.003	0.004	0.003	0.003
Test				
Raw	0.009	0.009	0.009	0.01
Phase	0.023	0.025	0.021	0.022

Values are the mean absolute error in the alpha coefficient.

varying exponential time-gain profile uniformly sampled in the range ±5dB/cm, with the same random scaling coefficient (raw w/ gain). While errors on the training data remain manageable, results on the test data, when using the raw input signal as before, degrade significantly. This implies over-training with no generalization to unseen data. The results when using the phase input remain unchanged in both cases, suggesting that both effects have been removed from the data by this preprocessing.

As another measure of how well the IQ-phase preprocessing decouples the gain and scaling system response, we look at attenuation recovery. Recovering attenuation maps can be thought of as the dual of recovering gain. To test this, we train the network to recover the attenuation coefficient (α coefficient), rather than the speed-of-sound, as presented in Table III. The α coefficient is randomly set with uniform distribution in the range of 0.05 to 0.15 $\mathrm{dB/cm/MHz}^2$. When using the raw signal as input, the α coefficient can be mostly recovered. For the IQ-phase information, however, there is significant over-training, with no generalization to the test data. The mean absolute error using IQ-phase information with regard to the test data is 0.023. For reference, the mean error with respect to the center value is 0.025. This implies that little to no information on attenuation is left in the signal.

TABLE IV
COMPARING RECOVERY RESULTS FOR A MISS-MATCHED PULSE SHAPE

Method	Mean	Near	Mid	Far
5 MHz				
5 cyc. (ref)	40	30	40	52
5 cyc. w/matched	43	31	43	55
4.4 cyc.	77	74	88	70
4.4 w/Weiner	42	31	42	51
4.4 w/matched	69	62	79	67
2.5 MHz				
5 cyc.	124	140	131	102

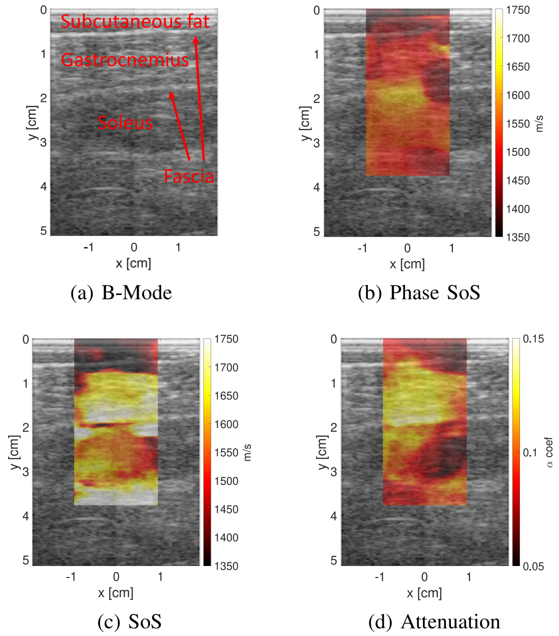
The reference pulse is at 5 MHz and 5 cycles. The first comparison is for training the network with a matched. The next three compare applying the same network weights to a 5 MHz 4.4 cycle pulse, as-is, after applying a wiener filter, and with a matched filter network. Finally, compared to a 2.5 MHz pulse with 5 cycles.

To test whether training results transfer to pulses with a different center frequency and bandwidth (sensitivity to pulse shape), we use the network weights trained on a 5 MHz center frequency pulse with a 5-cycle Gaussian window and apply the network to a test set generated with a 5 MHz pulse with a 4.4 cycle window and to a test set generated with a 2.5 MHz pulse with a 5 cycle window. Applying a matched filter and a wiener filter is only applicable verbatim with the 4.4 cycle pulse at 5 MHz, as linear filters cannot shift frequency.

The results for the test error for the 5 MHz pulse at 4.4 cycles are presented in Table IV and the results for the 2.5 MHz pulse in Table IV.

To assess the stability of the approach, we turn our attention to real data. Fig. 7 shows inversion results for the calf of a human participant from a posterior view. Experiments were carried out using a protocol approved by the MIT Committee on the Use of Humans as Experimental Subjects (COUHES). Fig. 7(a) shows the reference b-mode image, with a layer of subcutaneous fat at the top, the gastrocnemius muscle next, and the soleus muscle underneath. Fig. 7(b) shows the inversion results using IQ-phase data, compared to speed-of-sound and attenuation from the raw channel data in Fig. 7(c) and (d) respectively. While all samples differentiate fat from muscle tissue the speed-of-sound recovered from raw channel data overshoots the expected values significantly. The value for fat is far below expected (1440 m/s, standard deviation 21.9), and above expected for muscle (1588 m/s, standard deviation 21.6) [47]. Attenuation is close to the expected values of 0.08 for fat and 0.16 for muscle. The results based on the IQ-phase input are significantly more stable and close to the expected values throughout, also differentiating the higher velocity fascia tissue from the muscle.

Fig. 8 shows results for an agar-agar inclusion in a gelatin phantom. This provides a more controlled assessment of stability to changes in the input signal, including amplitude and gain changes. The results show that using the IQ-phase input provides significantly more stable results that are agnostic to change in input.



(a) B-Mode (b) Phase SoS

(c) SoS (d) Attenuation

Fig. 7. Human participant, calf image (Gastrocnemius and Soleus muscles): (a) Shows the B-Mode image, showing the probe's acoustic lens (thin bright line on top), subcutaneous fat layer (slightly darker bright section just below), down to the fascia (bright line) differentiating the discontinuity with the Gastrocnemius muscle, then the fascia between the Gastrocnemius and soleus muscles. (b) Speed-of-sound (m/s) map recovered using phase-based inversion. This shows the differentiation between the lower speed of sound in the fat layer and the higher speed of sound in the muscle. (c) and (d) Recovered speed-of-sound and attenuation maps from the raw signal. We see a significantly improved transfer learning from the phase-based inversion.

The inclusion was made using 1 part agar-agar to 10 parts water with 5 grams of graphite to 300 grams of phantom material for speckle. The background was made from 1 part beef gelatin to 12 parts water, with 5 grams of graphite to 550 grams mixture. Speed-of-sound of sound in the inclusion was measured at approximately 1650 m/s, and in the background at approximately 1500 m/s.

The top left image shows the b-mode image of the phantom. Top right, top image, shows the inversion results for the material used for inclusion (independent of the phantom), and bottom images show the inversion results for background material. These show results consistent with the expected value when measured independently of a full phantom. The grid compares inversion results for our network trained on IQ-Phase input (first column), raw channel data (second column), and raw channel data trained with random scalar multiplier and exponential gain factor. The rows show inversion results for the raw signal without gain or amplification correction (first row) and two amplification factors, $2 \cdot 10^{-4}$ and $5 \cdot 10^{-4}$ (second and third row). These were tested to be close to true values.

The IQ-phase-based inversion is shown to be effectively agnostic to both amplification and gain while recovering speed-of-sound that are both close to the correct results as well as to the values recovered when imaging the background and inclusion materials independently. There are however still some artifacts inside the phantom and some bleeding from the higher speed inclusion onto the background. These are probably due to a combination of the limitations of transfer learning, which require training with real data and limited SNR in the shadow areas. In contrast, the inversion with the network trained on raw data is unstable, with results on uncorrected input being completely corrupt. The recovered speed-of-sound is fully dependent on amplification for the gain-corrected signal. While the inversion results when the network is trained on the raw signal by varying gain and amplification are more stable, they still vary significantly with amplification and gain.

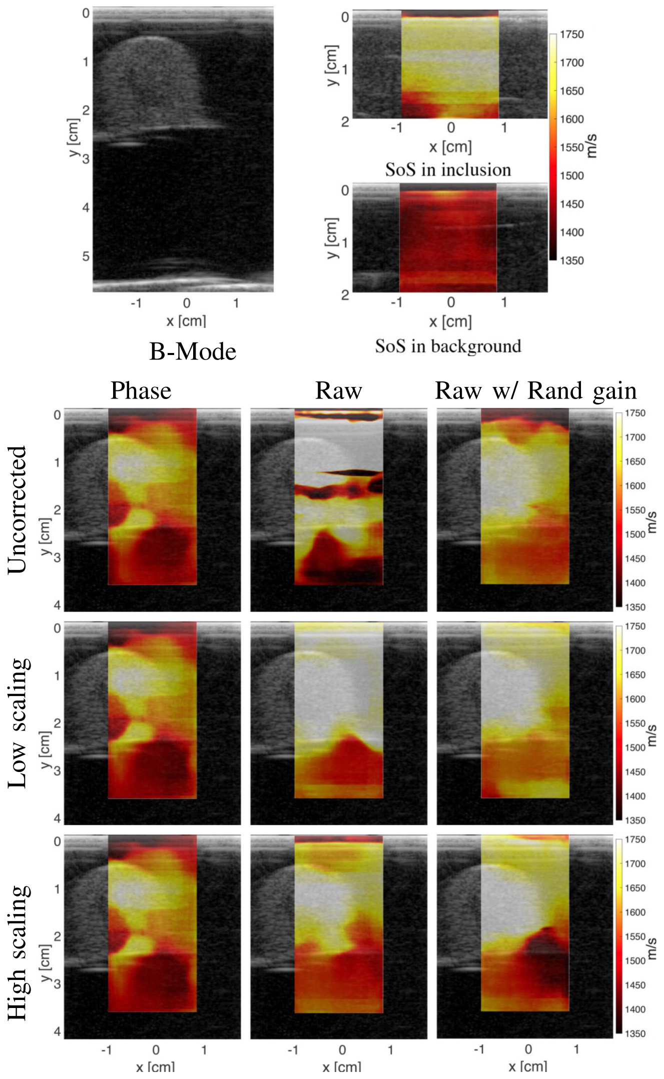


Fig. 8. Results for physical phantom (agar-agar inclusion, speed-of-sound ~1650 m/s, gelatin medium, speed-of-sound ~1500 m/s). Reference B-Mode image is shown top-left. Speed-of-sound inversion for inclusion material and background material samples shown top right. The columns show the results for inverting for speed-of-sound using the IQ-phase information, raw channel data, and raw channel data trained by randomly varying the gain and scaling of the data. The rows show results for inverting using the uncorrected input signal, scaling the signal by a factor of 2×10^{-4}, and a factor of 5×10^{-4}.

V. Conclusion

In this paper, we present a novel approach to train and run the neural network in an operator-agnostic manner, utilizing the complex phase of the IQ signal rather than raw input values. The results show that inversion from IQ-phase is viable and improves the stability to variability in the input while at the same time reducing the load of system calibration. The overall results strengthen the claim that, given calibrated real data, transfer learning can be employed to adapt the network to real data. That is however the subject of future work due to the current lack of real training data.

We also present the first publicly available synthetic benchmark for speed-of-sound inversion in medical ultrasound, to promote further contributions to the advancement of the field.

REFERENCES

[1] T. L. Szabo, *Diagnostic Ultrasound Imaging: Inside Out*. Amsterdam, The Netherlands: Elsevier, 2004.

[2] J. Pujol, *Elastic Wave Propagation and Generation in Seismology*. Cambridge, U.K.: Cambridge Univ. Press, 2003.

[3] W. Flugge, *Viscoelasticity*. Berlin, Germany: Springer, 1975.

[4] K. Nightingale, "Acoustic radiation force impulse (ARFI) imaging: A review," *Curr. Med. Imag. Rev.*, vol. 7, no. 4, pp. 328–339, Jul. 2011.

[5] A. Nowicki and K. Dobruch-Sobczak, "Wprowadzenie do ultradźwiękowej elastografii," *J. Ultrasonography*, vol. 16, no. 65, pp. 113–124, Jun. 2016.

[6] J. Fenner et al., "Macroscopic stiffness of breast tumors predicts metastasis," *Sci. Rep.*, vol. 4, 2014, Art. no. 5512.

[7] M. Feigin, D. Freedman, and B. W. Anthony, "A deep learning framework for single-sided sound speed inversion in medical ultrasound," *IEEE Trans. Biomed. Eng.*, vol. 67, no. 4, pp. 1142–1151, Apr. 2020.

[8] J. Yosinski et al., "How transferable are features in deep neural networks?," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 3320–3328.

[9] N. Levanon and E. Mozenson, *Radar Signals*. Hoboken, NJ, USA: Wiley, 2004.

[10] A. V. Oppenheim and G. C. Verghese, *Signals, Systems & Inference*. London, U.K.: Pearson, 2016.

[11] M. Feigin et al., "Detecting muscle activation using ultrasound speed of sound inversion with deep learning," in *Proc. IEEE 42nd Annu. Int. Conf. Eng. Med. Biol. Soc.*, 2020, pp. 2092–2095.

[12] Ö. Yilmaz, *Seismic Data Analysis*. Thousand Oaks, CA, USA: SAGE, Jan. 2001, doi: 10.1190/1.9781560801580.

[13] J. Krämer et al., "Two-dimensional speckle tracking as a non-invasive tool for identification of myocardial fibrosis in Fabry disease," *Eur. Heart J.*, vol. 34, no. 21, pp. 1587–1596, Jun. 2013, doi: 10.1093/eurheartj/eht098.

[14] J. C. D'Souza et al., "B-mode ultrasound for the assessment of hepatic fibrosis: A quantitative multiparametric analysis for a radiomics approach," *Sci. Rep.*, vol. 9, no. 1, Dec. 2019, Art. no. 8708. [Online]. Available: http://www.nature.com/articles/s41598-019-45043-z

[15] N. Kagiyama et al., "A low-cost texture-based pipeline for predicting myocardial tissue remodeling and fibrosis using cardiac ultrasound," *EBioMedicine*, vol. 54, Apr. 2020, Art. no. 102726. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S2352396420301018

[16] G. Song et al., "Usefulness of speckle-tracking echocardiography for early detection in children with duchenne muscular dystrophy: A meta-analysis and trial sequential analysis," *Cardiovasc. Ultrasound*, vol. 18, no. 1, Dec. 2020, Art. no. 26, doi: 10.1186/s12947-020-00209-y.

[17] A.-H. Liao et al., "Deep learning of ultrasound imaging for evaluating ambulatory function of individuals with duchenne muscular dystrophy," *Diagnostics*, vol. 11, no. 6, May 2021, Art. no. 963. [Online]. Available: https://www.mdpi.com/2075-4418/11/6/963

[18] J. Civale, J. Bamber, and E. Harris, "Amplitude based segmentation of ultrasound echoes for attenuation coefficient estimation," *Ultrasonics*, vol. 111, Mar. 2021, Art. no. 106302, doi: 10.1016/j.ultras.2020.106302. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0041624X20302390

[19] J. Wang et al., "Assessment of myocardial fibrosis using two-dimensional and three-dimensional speckle tracking echocardiography in dilated cardiomyopathy with advanced heart failure," *J. Cardiac Failure*, vol. 27, no. 6, pp. 651–661, Jun. 2021, doi: 10.1016/j.cardfail.2021.01.003. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S107191642100004X

[20] M. Sak et al., "Using speed of sound imaging to characterize breast density," *Ultrasound Med. Biol.*, vol. 43, no. 1, pp. 91–103, Jan. 2017.

[21] T. Shiina et al., "WFUMB guidelines and recommendations for clinical use of ultrasound elastography: Part 1: Basic principles and terminology," *Ultrasound Med. Biol.*, vol. 41, no. 5, pp. 1126–1147, May 2015. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0301562915002227

[22] X. Zheng et al., "Shear-wave elastography of the breast: Added value of a quality map in diagnosis and prediction of the biological characteristics of breast cancer," *Korean J. Radiol.*, vol. 21, no. 2, pp. 172–180, 2020, doi: 10.3348/kjr.2019.0453.

[23] K. Arda et al., "Quantitative assessment of normal soft-tissue elasticity using shear-wave ultrasound elastography," *Amer. J. Roentgenol.*, vol. 197, no. 3, pp. 532–536, Feb. 2011.

[24] J. M. Chang et al., "Comparison of shear-wave and strain ultrasound elastography in the differentiation of benign and malignant breast lesions," *Amer. J. Roentgenol.*, vol. 201, no. 2, pp. W347–W356, Dec. 2013.

[25] J. Correas et al., "Prostate cancer: Diagnostic performance of real-time shear-wave elastography," *Radiology*, vol. 275, no. 1, pp. 280–289, Apr. 2015.

[26] G. Ferraioli et al., "Shear wave elastography for evaluation of liver fibrosis," *J. Ultrasound Med.*, vol. 33, no. 2, pp. 197–203, Feb. 2014.

[27] M. S. Taljanovic et al., "Shear-wave elastography: Basic physics and musculoskeletal applications," *RadioGraphics*, vol. 37, no. 3, pp. 855–870, May 2017.

[28] J. Gandhi et al., "The evolving role of shear wave elastography in the diagnosis and treatment of prostate cancer," *Ultrasound Quart.*, vol. 34, no. 4, pp. 245–249, 2018.

[29] Y. L. Chen et al., "Ultrasound shear wave elastography of breast lesions: Correlation of anisotropy with clinical and histopathological findings," *Cancer Imag.*, vol. 18, no. 1, pp. 1–11, 2018.

[30] H. Naganuma et al., "Diagnostic problems in two-dimensional shear wave elastography of the liver," *World J. Radiol.*, vol. 12, no. 5, pp. 76–86, May 2020. [Online]. Available: https://www.wjgnet.com/1949-8470/full/v12/i5/76.htm

[31] F. A. Duck, *Physical Properties of Tissue: A Comprehensive Reference Book*. Cambridge, MA, USA: Academic Press, 1990.

[32] N. Duric et al., "Breast imaging with the SoftVue imaging system: First results," *Proc. SPIE*, vol. 8675, 2013, Art. no. 86750K.

[33] C. Li et al., "In vivo breast sound-speed imaging with ultrasound tomography," *Ultrasound Med. Biol.*, vol. 35, no. 10, pp. 1615–1628, 2014.

[34] J. R. Fincke et al., "Towards ultrasound travel time tomography for quantifying human limb geometry and material properties," *Proc. SPIE*, vol. 9790, Apr. 2016, Art. no. 97901S.

[35] M. Jaeger et al., "Computed ultrasound tomography in echo mode (CUTE) of speed of sound for diagnosis and for aberration correction in pulse-echo sonography," *Proc. SPIE*, vol. 9040, Mar. 2014, Art. no. 90400A.

[36] P. Stähli et al., "Improved forward model for quantitative pulse-echo speed-of-sound imaging," *Ultrasonics*, vol. 108, Dec. 2020, Art. no. 106168.

[37] S. J. Sanabria, M. B. Rominger, and O. Goksel, "Speed-of-sound imaging based on reflector delineation," *IEEE Trans. Biomed. Eng.*, vol. 66, no. 7, pp. 1949–1962, Jul. 2019.

[38] V. Vishnevskiy, S. J. Sanabria, and O. Goksel, "Image reconstruction via variational network for real-time hand-held sound-speed imaging," in *Proc. Int. Workshop Mach. Learn. Med. Image Reconstruction*, 2018, pp. 120–128.

[39] R. Rau et al., "Speed-of-sound imaging using diverging waves," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 16, no. 7, pp. 1201–1211, Jul. 2021.

[40] F. K. Jush et al., "DNN-based speed-of-sound reconstruction for automated breast ultrasound," in *Proc. IEEE Int. Ultrason. Symp.*, 2020, pp. 1–7.

[41] F. Yang and J. Ma, "Deep-learning inversion: A next-generation seismic velocity model building method," *Geophysics*, vol. 84, no. 4, pp. R583–R599, 2019.

[42] M. J. Park and M. D. Sacchi, "Automatic velocity analysis using convolutional neural network and transfer learning," *Geophysics*, vol. 85, no. 1, pp. V33–V43, 2020.

[43] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Med. Image Comput. Comput.-Assist. Interv.*, 2015, pp. 234–241.

[44] M. Feigin, D. Freedman, and B. W. Anthony, "dl_us_sos_inversion (Revision 6ce7d82)," 2023. [Online]. Available: https://huggingface.co/laughingrice/dl_us_sos_inversion

[45] M. Feigin, D. Freeman, and B. W. Anthony, "Ultrasound_planewave_sos_inversion (revision 046c9a7)," 2023. [Online]. Available: https://huggingface.co/datasets/laughingrice/Ultrasound_planewave_sos_inversion

[46] B. E. Treeby and B. T. Cox, "K-Wave: MATLAB toolbox for the simulation and reconstruction of photoacoustic wave fields," *Proc. SPIE*, vol. 15, no. 2, 2010, Art. no. 021314.

[47] P. Hasgall et al., "IT'IS database for thermal and electromagnetic parameters of biological tissue," Feb. 2022. [Online]. Available: https://www.scienceopen.com/document?vid=a95fbaa4-efd8-429a-a59e-5e208fea2e45