CS230

Learning Acoustic Properties of Human Skull from Medical Images - Final Report (Healthcare)

Ningrui Li
Department of Electrical Engineering
Stanford University
ningruil@stanford.edu
ningruil

Kris Quah
Department of Electrical Engineering
Stanford University
krisquah@stanford.edu
krisquah

Kasra Naftchi-Ardebili
Department of Bioengineering
Stanford University
knaftchi@stanford.edu
knaftchi

Abstract

Phase aberration due to variation in skull geometry and composition across patients is an immense impediment to consistently safe and effective transcranial focused ultrasound treatments. Aberrations can be corrected with knowledge of each skull's acoustic properties, including acoustic velocity and attenuation, which are currently estimated using linear models based on CT imaging. However, current models are limited; they only take into account Hounsfield unit values of skull voxels, whereas the pore structure of the skull majorly influences its acoustic properties. This project seeks to use neural networks to achieve superior predictions of the skull's acoustic properties by leveraging additional information about the skull's pore structure from neighboring skull voxels. While variance error remains a challenge, we demonstrate a $\sim$25% reduction in error of sound speed estimates when using a convolutional neural network over current linear models. Ultimately, this work will improve transcranial focused ultrasound treatment outcomes and enable a wider patient population to be treated.

1 Introduction

Therapeutic ultrasound is a noninvasive means for treating neurological disorders, including essential tremor, Alzheimer's disease, and depression [1, 2]. However, the skull remains a barrier to safe and effective treatments. It aberrates the focal spot's shape and shifts its location from the intended target, causing inconsistent treatment outcomes. These problems are exacerbated by the significant variability in skull geometry and composition across patients. To reduce aberration due to the skull, patient-specific corrections are applied to the ultrasound transducer that account for the differing sound speeds in each skull region. These corrections are computed by simulating ultrasound propagation through estimated acoustic models of the skull [3, 4]. While these advances have enabled a larger patient population to be treated, there remain challenges in obtaining accurate acoustic models of the skull.

Acoustic models of the skull are currently estimated using pre-operative computed tomography (CT) scans. Sound speed and attenuation are assumed to linearly increase with Hounsfield unit values [5]. However, these relationships are simplistic [4] and do not account for the skull's pore structure, which has been shown to affect its sound speed [6]. To address this problem, we aim to train a neural network that takes a skull CT as input and accurately outputs the sound speed (Figure 1B). We hypothesize that the additional information about pore structure provided by neighboring skull voxels will result in superior predictions of acoustic properties compared to the current gold-standard.

CS230: Deep Learning, Spring 2020, Stanford University, CA. (LateX template borrowed from NIPS 2017.)

2 Dataset and Features

Our dataset consists of CT scans and acoustic properties (density, sound speed, and acoustic attenuation) measured from 100 human skull fragments. Multiple CT scans of each fragment were acquired with a range of X-ray energies and reconstruction kernels using scanners from two different manufacturers, as detailed in [7]. To match Stanford's clinical protocols, we used CT scans acquired using the Discovery CT750 (GE, Waukesha, WI, USA) at an x-ray energy of 120 kVp with a bone mineral density reconstruction kernel as inputs to our neural network model. Sound speed was measured experimentally using a hydrophone set-up. This provides the actual sound speed through the skull, which cannot be obtained in-vivo as it is an invasive procedure.

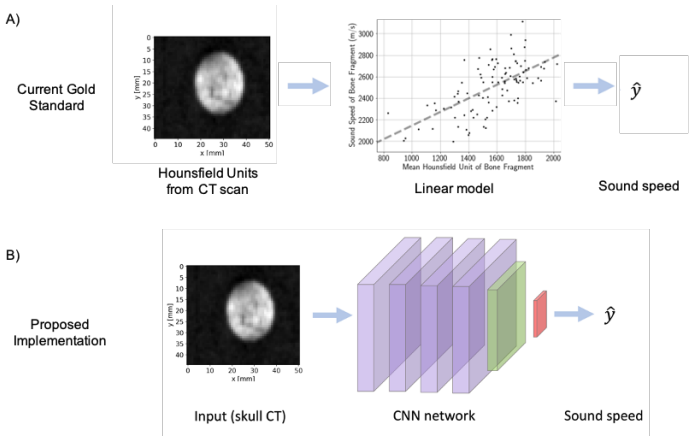


Figure 1: A) The current gold-standard is linear regression, predicting sound speed based the Hounsfield unit value of each voxel. B) We hypothesize that neural networks can leverage additional information about the skull's pore structure from neighboring voxels to achieve superior performance.

Fragments have higher HU values are assumed to be denser and, as expected, have correspondingly higher sound speeds (Figure 2). However, linear regression is too simplistic, as fragments with the same mean HU values can have vastly different sound speeds, and there can be up to a 800 m/s discrepancy. Local variations in HU of each fragment could provide information about its underlying pore structure, which could explain some of these differences in sound speed between fragments with similar mean HU values.

3 Methods

3.1 Initial Model

As a first basic model, we built a fully-connected 7-layer neural network with the following architecture:

256 units {linear/Relu} $\rightarrow$ 512 units {linear/Relu} $\rightarrow$ 512 units {linear/Relu} $\rightarrow$ 512 units {linear/Relu}$\rightarrow$ 512 units {linear/Relu} $\rightarrow$ 256 units {linear/Relu} $\rightarrow$ 1 unit {linear}

Each CT scan was cropped to be $24 \times 24 \times 3$, then flattened to a 1728-element vector as an input to the neural network. Due to the small size of our dataset, we trained our model using 10-fold cross validation. The model was trained with a learning rate of 0.001, *L2* regularization was applied to the last two 512-layers to reduce overfitting, and batch normalization was applied following every activation layer.

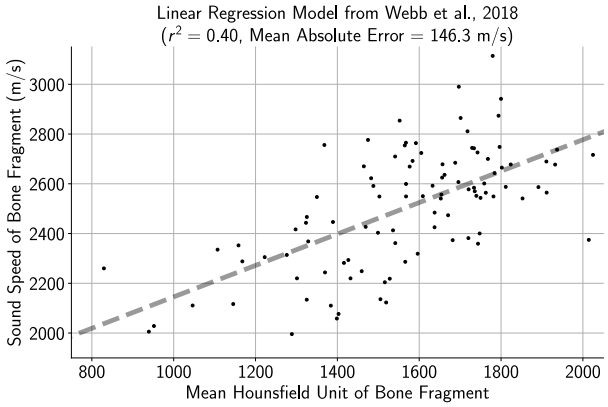


Figure 2: Sound speed is weakly associated with HU value. However, linear regression models clearly underexplain the data, as fragments with the same mean HU value can have vastly different sound speeds.

The mean squared error (MSE) was used as our loss function, but the mean absolute percent error (MAPE) was also computed to assess model performance compared to the current gold-standard (linear regression).

$$\text{MAPE} = \left(\frac{1}{90} \sum_{j=1}^{90} \left| \frac{y_j - \hat{y}_j}{y_j} \right| \right) \times 100 \tag{1}$$

3.2 Convolutional Neural Network (CNN)

For the next iteration of our model, we built a CNN with 4 convolution layers, 4 max pool layers and 3 fully connected layers. We continued to use MSE as our loss function and MAPE to monitor how well we are performing. In an attempt to reduce overfitting, we reduced the model to only contain 1 fully connected layer (Figure 3).

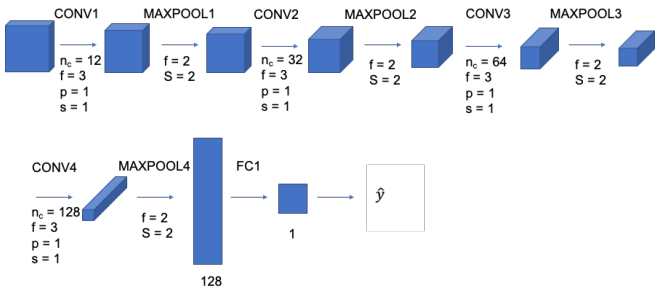


Figure 3: CNN model architecture, which consisted of 4 convolution layers, 4 max pool layers and 1 fully connected layer. A Relu activation was used in every layer except for the final output layer, which used a linear activation.

3.3 Data Augmentation

To reduce overfitting our model to our small dataset, we augmented our training set with additional examples obtained by rotating or flipping image volumes. Skull fragments were radially symmetric, and sound speed measurements were obtained with each fragment in an arbitrary orientation, so we were confident that any flips or rotations along the longitudinal axis (the direction of acoustic propagation) would not significantly change the sound speed label of each fragment.

Flipping augmentations were implemented as each image volume having a 50% probability of flipping in each dimension. Rotation angles were randomly chosen to be either 10°, 20°, ..., or 350°, and rotations were applied to the entire image volume along the longitudinal axis. To account for the additional augmentations, each model was trained for at least several thousand epochs.

3.4 Modified ResNet50: Convolution vs. Training Time

We further wanted to assess the effectiveness of convolutional layer versus training duration. For this experiment, we trained a modified version of ResNet50 from the course where we added a linear layer in the end for a singular regression output. We trained this model over 1,000 epochs with 0.001 learning rate. In parallel, we ran our initial fully-connected 7-layer neural network with similar learning rate, but over 10,000 epochs. The results were comparable, which suggested convolutional filters did not confer a substantial advantage. Our analysis of this comparison was that convolutional neural networks were best in classifying data where correlated patterns are fundamental. For instance, eyes of a cat are always below the ears and on either sides of the nose. One never finds a cat with an eye above an ear. Our data however, does not have any such correlated patterns and compared to a real image, looks more like noise than anything else. Therefore, a deep fully-connected neural network should do just as well as a CNN, if trained for long enough. To validate our line of thinking, we looked into a pre-trained VGG16 model.

3.5 Transfer Learning via VGG16

The last architecture we explored was a VGG16 model that was pre-trained on the ImageNet dataset. We froze the weights of all the layers except the last three, which included two 2D convolution filters followed by a max-pooling filer. We added a batch-normalized, 128-unit dense layer, connected to a final, single-unit output layer. Training this model with a learning rate of 0.001 and L2 regularization over 1,000 epochs resulted in a test MAPE of 5.53%. A convolutional layer that is only trained on our dataset does not show considerable advantage over a deeply trained dense neural network. A convolutional neural network with transfer learning however, may help our model accuracy.

4 Results

4.1 Performance Comparison and Hyperparameter Optimization

Table 1 summarizes performance using a series of model architectures and hyperparameters.

Model	LR	Regularization	Epochs	Augmentation	train MAPE	val MAPE	test MAPE
Lin. Reg.	N/A	N/A	N/A	none	5.782%	6.081%	N/A
CNN	0.05	N/A	2,000	none	0.0%	5.792%	N/A
CNN	0.05	N/A	2,000	flipping + rotation	0.594%	5.999%	N/A
CNN	0.05	N/A	5,000	flipping + rotation	0.403%	5.061%	N/A
CNN	0.00025	N/A	10,000	flipping + rotation	9.3%	8.907%	N/A
CNN	0.45	N/A	3,000	flipping + rotation	1.073%	4.431%	N/A
CNN	0.0992	L2	10,000	flipping + rotation	0.577%	5.175%	N/A
CNN	0.0479	L2	3,000	flipping + rotation	0.804%	5.473%	N/A
FCNN	0.001	L2	10,000	none	0.97%	8.3%	N/A
ResNet50	0.001	L2	1,000	none	3.53%	8.32%	N/A
VGG16	0.001	L2	1,000	90°-rotations	0.94%	8.94%	5.53%

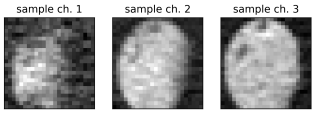
Table 1: Summary of the experimental models. With linear regression, the MAPE on the validation set was 6.081%. Factors such as number of epochs, augmentation, learning rate, and transfer learning appear to have the strongest effects on the model performance.

As expected, there was a significant mismatch between training MAPE (0.0%) and validation MAPE (∼6%) when no augmentations were applied, indicating that the model was strongly overfitting to the orientation of each fragment. When flipping and rotation augmentations were applied during training, this discrepancy was lessened, however, it was noted that more epochs were needed (>3,000) during training to achieve better performance on the validation set compared to the no-augmentation case.

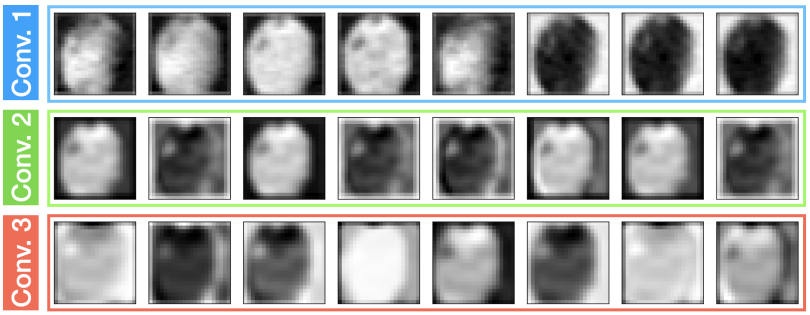
To optimize for the learning rate, different rates were randomly sampled over a log-scale from 10^{-4} to 1. Model performance did not vary greatly for learning rates above 0.01. However, model performance was poor with lower learning rates ($< 10^{-4}$), even when it was trained for 10,000 epochs. Rotation and flipping augmentations applied during training likely exacerbated this problem. To further reduce the variance issue, we also explored adding $L2$ regularization to each layer. Thus far, the best CNN model performance (train MAPE of 1.073% and validation MAPE of 4.341%) was achieved using a learning rate of 0.45 with 3,000 epochs and no regularization. The CNN model did not result in substantial improvement over the fully-connected neural network (FCNN). Similarly, neither did deeper networks (such as ResNet50). The small size of each image volume likely meant that the later layers of the deeper networks were essentially fully-connected layers.

4.2 Neural Net Feature Maps

To better understand features of the fragments that were deemed important by the neural networks, we investigated the feature maps for some of the convolutional layers. Note that some skull fragments were thinner than others, such as the one depicted in Figure 4 (a), where part of the skull fragment in channel/slice 1 is only partially visible.



(a) Channels 1 and 3 represent the scan slices immediately before and after the layer with highest signal intensity.



(b) Feature maps produced by only the first three convolutional layers of our CNN model. Here we are showing only 8 units per layer to showcase the different features learned by the network.

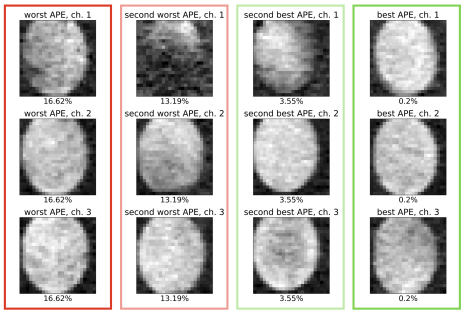
Figure 4: A sample input and its features extracted by the first three convolutional layers. Prior to processing, each skull CT was of different dimensions due to variations in skull fragment thickness. To standardize our input, we cropped all the scans to $24 \times 24 \times 3$.

The choice of cropping input width and height is informed by the feature maps produced by the convolutional layers. We noticed that if the cropped image only captured the skull fragment (e.g. $12 \times 12 \times 3$), with the exception of the first

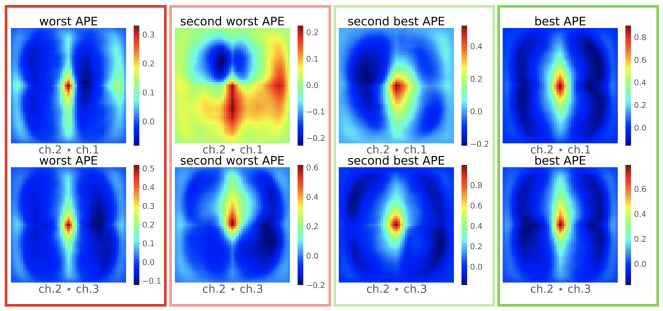
layer the remaining convolutional layers learned only blank weights with random wide frames around every output. A better choice was $24 \times 24 \times 3$, where both skull features and shapes were picked up by the convolutional layers (Figure 4 (b)). These feature maps suggested that the model was able to learn the boundaries of every fragment as well as strong features within each fragment. One concern is that the model might also be fitting to the background.

4.3 Test Performance: Best and Worst Case Scenarios

Next, we looked at the best two and the worst two performances on the test set to gain better intuition about what features the network was learning most effectively. Figure 5 (a) suggested that samples with most similar channels tend to perform better. In other words, sudden changes from one channel to another deteriorated the absolute percent error for that sample. To better assess this hypothesis, we computed the cross-correlation between channel 2 and either of the outer channels for these samples. If all three channels show very similar features embedded in an oval-shaped fragment, we anticipated cross-correlation maps where the highest values were in the center of the fragment, tapering off in a semi-oval geometry. Moreover, we expected lowest cross-correlation values to correspond to the edges and the backgrounds of the channels. Figure 5 (b) does indeed support our hypothesis. Compared to an RGB input where every channel represent the same 2D feature, a 3D array with different channels is more challenging to learn; hence the high APE for the two left most samples in Figure 5 (a).



(a) Three channels of the worst APE of 16.62% on the left compared to the three channels of the best APE of 0.2% on the right .



(b) Cross-correlation of the middle layer, ch. 2, with channels 1 and 3 for each of the samples in (a).

Figure 5: Cross-correlations between the outermost channels and the inner channel are highest for samples with the lowest absolute percent error (APE). Moreover, fragments with best APEs demonstrate a general oval cross-correlation with its peak located near the center of the skull fragment.

5 Conclusions and Future Work

We have demonstrated that improved sound speed predictions of human skull fragments could be made using CNNs, resulting in a $\sim$25% reduction in MAPE on the test set when compared to the gold-standard linear regression model. Unfortunately, variance error still appears to be the primary impediment to our model. While we have found that rotation and flipping augmentations significantly reduced the difference between training and validation errors, there remains a substantial discrepancy. We next plan to simulate ultrasound propagation through separate hold-out test sets, including imaging data from 9 treatments on human patients and measurements of acoustic transmission through 3 human skulls and 20 sheep skull caps. Simulations will be run using acoustic models estimated with the CNN and compared to ground-truth acoustic transmission values for each patient/skull cap. Fidelity of these simulation results will be compared to results obtained using current gold-standard acoustic models, ensuring that our CNN models are generalizable and achieve improved performance in practice. Finally, we plan to also train additional models using higher resolution model-based iterative reconstruction CT scans, micro-CT scans, or magnetic resonance imaging as inputs to determine if additional structural information is helpful for improving our model. We hope that improved patient-specific corrections derived from this work will enhance the efficacy of tFUS treatments while reducing risk to the subject.

6 Contributions

NL worked on data pre-processing and exploring data augmentation methods. KNA worked on transfer learning and data analysis. KQ worked on regularization and hyperparameter tuning. All authors contributed equally to the writing of the final report and video presentation.

References

[1] MR-guided Focused Ultrasound to Treat Essential Tremor. `https://med.stanford.edu/neurosurgery/divisions/EssentialTremor.html`, 2019.

[2] Focused Ultrasound Foundation. `https://www.fusfoundation.org/diseases-and-conditions/overview`, 2019.

[3] Mathieu Pernot, Jean-François Aubry, Mickaël Tanter, Jean-Louis Thomas, and Mathias Fink. High power transcranial beam steering for ultrasonic brain therapy. *Physics in Medicine & Biology*, 48(16):2577, 2003.

[4] Steven A. Leung, Taylor D. Webb, Rachelle R. Bitton, Pejman Ghanouni, and Kim Butts Pauly. A rapid beam simulation framework for transcranial focused ultrasound. *Scientific Reports*, 9(1):1–11, 2019.

[5] Hounsfield Scale. `https://en.wikipedia.org/wiki/Hounsfield_scale`, 2019.

[6] T.D. Webb. *Predicting the transmission of ultrasound through the skull: estimation of the acoustic properties of bone using computed tomography and magnetic resonance imaging*. PhD thesis, Stanford University, Dec. 2018.

[7] T. D. Webb, S. A. Leung, J. Rosenberg, P. Ghanouni, J. J. Dahl, N. J. Pelc, and K. B. Pauly. Measurements of the Relationship Between CT Hounsfield Units and Acoustic Velocity and How It Changes With Photon Energy and Reconstruction Method. *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, 65(7):1111–1124, 2018.