Data Valuation for Acoustic Models in Automatic Speech Recognition

Ali Raza Syed
Computer Science
The Graduate Center, CUNY
New York, NY, USA
asyed2@gradcenter.cuny.edu

Michael I. Mandel
Computer and Information Science
Brooklyn College, CUNY
Brooklyn, NY, USA
mim@sci.brooklyn.cuny.edu

Abstract

Data valuation is concerned with how to assign value to examples used for training a supervised learning algorithm. A data valuation scheme holds promise for selection and curation of high quality data, especially in automatic speech recognition (ASR) of low-resource languages. Shapley values are a recently proposed method for data valuation in a machine learning context. While Shapley values can be approximated for classification models, we investigate the feasibility of these schemes for ASRs which are performing a structured prediction task. In particular, we show that a proxy model can be learned for the acoustic model component of an ASR and used to estimate Shapley values for acoustic frames. We further demonstrate that the estimated Shapley values estimated from the proxy model provide a strong signal of example quality at the frame level. In a phonetic classification task on the AN4 dataset, we can identify a high quality subset, comprising 20% of the available data, which yields near-optimal model performance.

1 Introduction

Data Valuation in Machine Learning is concerned with how to assign value to examples used in training a given supervised learning algorithm. This is particularly useful for data markets where individuals or vendors may be compensated for their data (e.g. [9]). Quantifying the value of data can also be useful for data curation as a means for ranking and selecting a subset of the data for optimal model performance. This suggests that data valuation may also be used to guide data collection efforts or develop selection heuristics for an active learner [3, 10]. Automatic speech recognition (ASR) systems require large amounts of acoustic data which are laborious and expensive to annotate; for example, a word level transcription of a minute of speech can take as long as 10 minutes [17]. This is especially severe for low-resource languages and dialects with limited access to expert human annotators [12, 13]. More generally, data valuation is also important for identifying and selecting high quality examples to provide stronger signals for data-efficient training of speech processing systems [7, 14]. Thus we are interested in exploring the feasibility of data valuation for selection and curation of spoken language documents.

The Shapley value [11], arising from cooperative game theory, has been proposed as an equitable way to allocate value to individual examples based on a model's performance on a training set [4, 5]. Since the Shapley value is computationally prohibitive ($\mathcal{O}(2^N)$) to compute, computations usually employ Monte-Carlo algorithms for estimation [e.g., 4]. Monte-Carlo algorithms require retraining a model, several times per example, making such schemes prohibitive for ASR systems which are expensive to train. Jia et al. [6] have recently proposed a method, based on a proxy nearest neighbors model, to approximate Shapley values without re-training a model. However, their examples are limited to neural network classifiers with fixed-length examples. We investigate the feasibility of

NeurIPS 2020 Workshop on Dataset Curation and Security

approximating Shapley values for sequential models. Modern ASR systems typically employ an end-to-end framework with recurrent neural networks for the structured prediction task of mapping variable length speech to variable length text. The model is comprised of Encoder and Decoder RNNs. The encoder RNN is an acoustic model that maps a variable length sequence of acoustic frames to a vector. The decoder RNN maps this vector to a variable length sequence of text. We focus on the acoustic model component of an ASR to determine if a proxy model is feasible and whether the approximated Shapley values provide a signal of example utility for the model.

2 Method

2.1 Shapley Values

In cooperative game theory, a coalition of players cooperate toward a common goal to earn some reward. The Shapley valuation framework provides one method of allocating rewards to individual players [11] based on their contribution. In the supervised machine learning setting, we think of training examples as participants in a game. The learning algorithm uses these points to achieve a reward based on performance measured on a held-out set. Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ be the training data and $\mathcal{D}_{\text{eval}} = \{(x_j, y_j)\}_{j=1}^{N_{\text{eval}}}$ be the held-out data used for evaluation. A learning algorithm $\mathcal{A}$ may take as input any subset $S \subseteq \mathcal{D}$ of training examples. The performance of the algorithm $\mathcal{A}$ using examples S is measured by an evaluation function $U_{\mathcal{A}}(S)$. The Shapley value $\sigma(x_i)$ of training example x_i is defined as the expected marginal contribution of x_i to any subset of the other training points $S \subseteq D \setminus \{x_i\}$:

$$\sigma(x_i) = \frac{1}{N} \sum_{S \subseteq \mathcal{D} \setminus \{x_i\}} \frac{1}{\binom{N-1}{|S|}} \left[U_{\mathcal{A}}(S \cup \{x_i\}) - U_{\mathcal{A}}(S) \right] \qquad (1)$$

The Shapley value is an allocation scheme that uniquely satisfies some rudimentary fairness axioms [5, 11]. An advantage of Shapley valuation is that it makes no assumptions about the training data distribution $\mathcal{D}$ or whether the examples are independent or identically distributed.

Approximating Shapley values. Jia et al. [5] proposed an algorithm for exact computation of Shapley values for K-Nearest Neighbor (KNN) models with quasi-linear complexity $\mathcal{O}(N_{\text{eval}} N \log N)$, where N_{eval} is the number of examples in the heldout set used for evaluating model performance. Jia et al. [6] proposed learning a proxy KNN model for a neural network by using the learned representation as inputs, then computing Shapley values for the KNN model. The estimated values were empirically shown to capture model-based utility of examples in a computer vision task.

2.2 Acoustic Model Proxy

An end-to-end ASR system consists of an encoder RNN, an attention layer, and a decoder RNN. The encoder RNN can be thought of as learning an acoustic model, while the decoder RNN learns a language and transition model for outputting sequences of text. We investigate Shapley valuation for ASR by restricting our attention to the encoder, i.e., the acoustic model component. Given a speech utterance, we map the constituent acoustic frames to the representation learned by the encoder network. These encoder embeddings are the inputs to our proxy KNN model. We use the ground truth phonetic labels for the frames found via forced alignment. Thus, the proxy KNN approximates an acoustic model learned in the end-to-end architecture, mapping acoustic frames to phonetic classifications. We evaluate the utility of the proxy acoustic model by determining the frame classification accuracy on held-out utterances. We evaluate the approximated Shapley values by examining the model's performance on the held out frames as we drop the lowest-valued or highest-valued examples from the training set, as is typical in the data valuation literature.

3 Experiments and Results

3.1 Dataset and Model

We use the AN4 dataset [1] which contains 948 training utterances and 130 test utterances from male and female speakers. Our inputs are 83-dimensional vectors per 25 ms frames using 80-dimensional

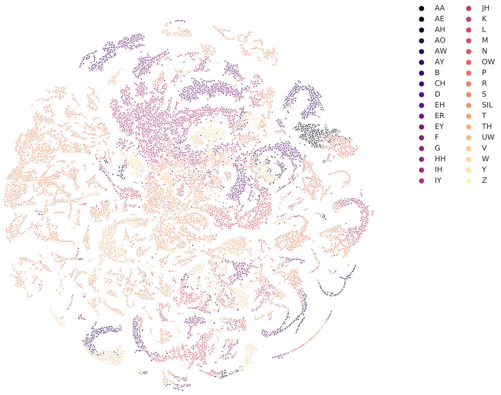


Figure 1: T-SNE based visualization of acoustic frames based on the representation learned by ASR encoder network.

log-Mel filterbank coefficients concatenated with a 3-dimensional pitch vector. We employ a state of the art end-to-end ASR model using hybrid CTC-Attention model [15] using the AN4 recipe from the open-source ESPnet ASR framework [16].

3.2 Proxy Acoustic Model

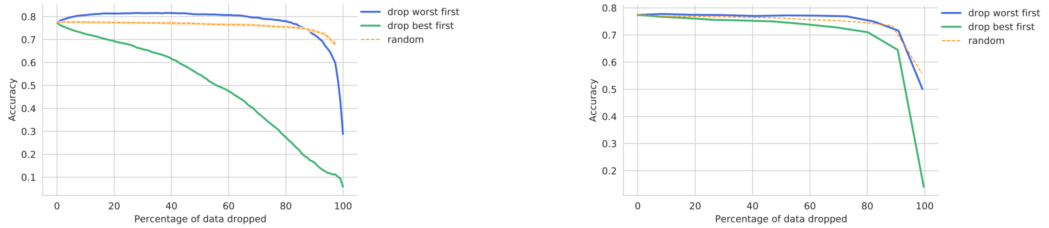
We begin by investigating the feasibility of learning a proxy KNN acoustic model for the AN4 dataset. The ground truth transcriptions for the acoustic frames were found using monophone based forced alignment from Kaldi [8]. The input features were the encoder representations of the acoustic frames. Figure 1 shows a T-SNE based visualization of the acoustic frames using these representations colored by their phonetic classification. The presence of structure in this visualization suggests that the encoder is learning a phonetic feature space.

We learn a KNN model mapping the encoder based acoustic features to the phonetic labels. Tuning on a held out set, we determine an optimal value of $k = 8$ for the KNN model with phonetic classification accuracy of 77.7% on a test set with unseen speakers. The high performance validates our approach of learning a proxy KNN for the ASR's acoustic model.

3.3 Evaluation of Shapley Values

We apply the Proxy KNN-Shapley algorithm [6] to estimate Shapley values for the training data. To evaluate whether Shapley values measure the utility of an example to the acoustic model, we following the approach used by the data valuation community. We rank the examples in ascending order of Shapley value (i.e. from lowest-valued or "worst" examples to highest-valued or "best" examples). We then evaluate the performance of the model by dropping examples in two different orders: dropping worst values first and dropping best values first. These curves are compared to model performance when dropping examples at random.

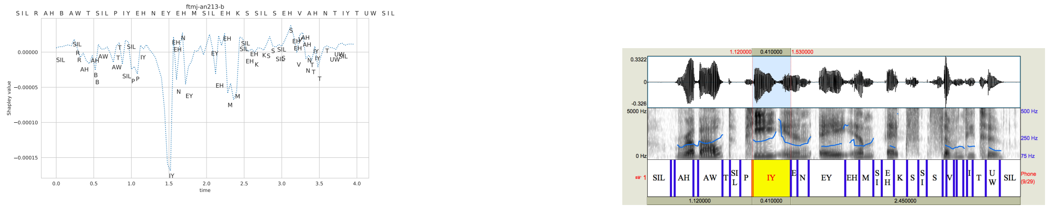
Dropping acoustic frames. Since our acoustic model classifies frames as phones, we evaluate our Shapley values by dropping acoustic frames, "re-training" the KNN proxy acoustic model and measuring its performance on a test set. The performance curves are shown in Figure 2a. The drop-worst-first curve shows that dropping $\sim$5% of lowest-valued data yields an improvement in performance. This suggests that Shapley values are able to identify examples that may be misleading or difficult to learn, and thus not of utility to the model. It is interesting that we can drop roughly 80% of the data before we begin see a significant drop in model performance. This suggests that many frames are redundant and convey similar patterns to the model, thus even randomly dropping frames does not produce a significant drop in performance. This concurs with observations made for neural net learning by Arpit et al. [2]. About 20% of the available data is sufficient to build an optimal acoustic model. Moreover, this subset of the data is identified by the (lowest) 80th percentile of Shapley values. Thus, we find that Shapley valuation of data can be used to curate data effectively.



(a) We measure proxy test accuracy as we drop data used for training the KNN model. In drop-worst-first, we drop data with lowest Shapley values first, while in drop-best-first, we drop data with highest Shapley values first. The random baseline shows results from 5 runs of of dropping random examples, with the mean performance as a dashed line.

(b) We compute duration-normalized Shapley values for utterances, and train and evaluate a KNN proxy model by dropping entire utterances. Summing and normalizing Shapley values of frames to value an utterance does not appear to produce a meaningful ranking of data.

Figure 2: Model performance curves for evaluating Shapley values.



(a) Shapley values of frames in an utterance. The lowest Shapley value occurs for frames labeled as phoneme “IY”.

(b) Praat grid showing annotated waveform and spectrogram for the same utterance, with the low-Shapley “IY” highlighted in yellow.

Figure 3: Examining Shapley values for frames in a selected utterance.

For further analysis, we investigated why some frames received particularly low values. An example is shown in Figure 3a where one frame with phone label “IY” has a particularly low Shapley value. We notice from the corresponding annotations in Figure 3b that the frame appears to be mis-labeled. The waveform suggests that the forced alignment is confused toward the end of the “IY” phone when it merges into the “EH” phone. This illustrates how Shapley values can be used effectively to identify and investigate annotations while curating data.

Dropping utterances. The acoustic frames used above are extracted from variable length spoken utterances. We are interested in learning whether Shapley values for the constituent frames can be used to value the utterances themselves, since in ASR it is almost always utterances that are labeled and not isolated acoustic frames. We compute values for utterances by summing up the Shapley values of the constituent frames and dividing by the number of acoustic frames in the utterance, thus obtaining a duration-normalized value. Performance curves from dropping utterances are shown in Figure 2b. Here, the performance of drop-worst-first is similar to that of drop-random. This suggests that our method of valuing utterances from Shapley values of frames does not yield an effective ranking. Figure 3a shows very high and very low Shapley values occurring within the same utterance, making it more difficult to value the entire utterance, thus further work is needed in this regard.

4 Conclusions and Future Work

We have shown the feasibility of using Shapley values for valuation of spoken language data for automatic speech recognition. We verified that a proxy KNN scheme can be employed to approximate Shapley values for that KNN classifier, and that dropping frames with the lowest Shapley values improves model performance. In future work, we will verify that these Shapley values also hold for the original acoustic model itself and will apply these findings to value variable length utterances for an end-to-end ASR model.

Broader Impact

Speech recognition is a key component of voice driven systems which are quickly becoming ubiquitous (e.g., Siri, Alexa). Communities with low-resource languages are under-served by these products and systems because of the scarcity of statistical and linguistic information as well as lack of expert annotators. Schemes for valuing data can aid in collection and curation of high quality data for faster development of ASR systems for low-resource languages. Our method is designed to identify a subset of data that is most valuable for a given model. It is possible that any bias arising from the original dataset may also propagate to a model trained on the subset. We have not investigated whether the subset improves or worsens any resulting model bias.

References

[1] A. Acero. *Acoustical and environmental robustness in automatic speech recognition*, volume 201. Springer Science & Business Media, 2012.

[2] D. Arpit, S. K. Jastrzebski, N. Ballas, D. Krueger, E. Bengio, M. S. Kanwal, T. Maharaj, A. Fischer, A. C. Courville, Y. Bengio, et al. A closer look at memorization in deep networks. In *ICML*, 2017.

[3] D. Cohn, L. Atlas, and R. Ladner. Improving generalization with active learning. *Machine Learning*, 15:201–221, 1994.

[4] A. Ghorbani and J. Zou. Data shapley: Equitable valuation of data for machine learning. In *International Conference on Machine Learning*, pages 2242–2251, 2019.

[5] R. Jia, D. Dao, B. Wang, F. A. Hubis, N. Hynes, N. M. Gürel, B. Li, C. Zhang, D. Song, and C. J. Spanos. Towards efficient data valuation based on the shapley value. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1167–1176, 2019.

[6] R. Jia, X. Sun, J. Xu, C. Zhang, B. Li, and D. Song. An empirical and comparative analysis of data valuation with scalable algorithms. *arXiv preprint arXiv:1911.07128*, 2019.

[7] Y. Liu, R. Iyer, K. Kirchhoff, and J. Bilmes. Svitchboard ii and fisver i: High-quality limited-complexity corpora of conversational english speech. In *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[8] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, et al. The kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*, number CONF. IEEE Signal Processing Society, 2011.

[9] R. Raskar, P. Vepakomma, T. Swedish, and A. Sharan. Data markets to support ai for all: Pricing, valuation and governance. *arXiv preprint arXiv:1905.06462*, 2019.

[10] B. Settles. Active learning literature survey. Technical report, University of Wisconsin-Madison Department of Computer Sciences, 2009.

[11] L. S. Shapley. A value for n-person games. *Contributions to the Theory of Games*, 2(28): 307–317, 1953.

[12] A. R. Syed, A. Rosenberg, and E. Kislal. Supervised and unsupervised active learning for automatic speech recognition of low-resource languages. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5320–5324. IEEE, 2016.

[13] A. R. Syed, A. Rosenberg, and M. Mandel. Active learning for low-resource speech recognition: Impact of selection size and language modeling data. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5315–5319, 2017. doi: 10.1109/ICASSP.2017.7953171.

[14] A. R. Syed, V. A. Trinh, and M. Mandel. Concatenative resynthesis with improved training signals for speech enhancement. In *Proc. Interspeech 2018*, pages 1195–1199, 2018. doi: 10.21437/Interspeech.2018-2439. URL `http://dx.doi.org/10.21437/Interspeech.2018-2439`.

[15] S. Watanabe, T. Hori, S. Kim, J. R. Hershey, and T. Hayashi. Hybrid ctc/attention architecture for end-to-end speech recognition. *IEEE Journal of Selected Topics in Signal Processing*, 11(8): 1240–1253, 2017.

[16] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N.-E. Y. Soplin, J. Heymann, M. Wiesner, N. Chen, et al. Espnet: End-to-end speech processing toolkit. *Proc. Interspeech 2018*, pages 2207–2211, 2018.

[17] X. Zhu. *Semi-supervised learning with graphs*. PhD thesis, 2005.