How Does Label Noise Affect the Quality of Speaker Embeddings?

Minh Pham, Zeqian Li, and Jacob Whitehill

Worcester Polytechnic Institute, USA

{mnpham, zli14, jrwhitehill}@wpi.edu

Abstract

A common assumption when collecting speech datasets is that the accuracy of data labels strongly influences the accuracy of speaker embedding models and verification systems trained from these data. However, we show in experiments on the large and diverse VoxCeleb2 dataset that this is not always the case: Under four different labeling models (Split, Merge, Permute, and Corrupt), we find that the impact on trained speaker embedding models, as measured by the Equal Error Rate (EER) of speaker verification, is mild (just a few percent absolute error increase,) except with very large amounts of noise (i.e., every minibatch is almost completely corrupted). This suggests that efforts to collect speech datasets might benefit more from ensuring large size and diversity rather than meticulous labeling.

Index Terms: label noise, speaker verification, embeddings

1. Introduction

The state-of-the-art approach to speaker recognition, verification, and diarization is based on *speaker embeddings* that map utterances into an embedding space such that embedded utterances from the same speaker are close together and embedded utterances from different speakers are far apart. Embedding models can be trained using a loss such as triplet [1] or GE2E [2] that operates on *sets* of utterances spanning at least two speakers. Alternatively, the embedding model can be trained as a classification network on *single* utterances using a softmax/cross-entropy loss [3, 4] to predict the speaker ID from a fixed set of speakers; after training, the softmax layer is discarded and a previous layer is used as the embedding.

To train accurate embedding models, large datasets containing utterances from many different speakers are required. Prominent examples include VoxCeleb (versions 1 [5] and 2 [6]) and LibriSpeech [7]; we have also recently published the BookTubeSpeech dataset [8]. In order to ensure that the labels of these datasets accurately reflect *who spoke when*, dataset collectors have employed various methods, including targeted YouTube keywords for specific people, pre-trained speaker and face embeddings to detect when two speakers are the same or when a single video contains multiple people, and manual annotation. However, the significant effort involved in labeling raises the question: How much does the veracity of the data labels impact the accuracy of downstream speaker embedding models? If the impact is small, then speaker embedding models might be improved by training on much larger datasets with only approximate labeling.

The concrete motivation for this paper is the following: After we collected BookTubeSpeech [8] and used it to train a speaker embedding model, we found that we achieved a better speaker verification EER on the LibriSpeech test-clean dataset [7] (0.0437 vs. 0.0565) when we used the *entire* dataset (38,707 videos) rather than just the 8,450 video subset whose labels ensured that they all contained distinct speakers. In other words, when training with just the video IDs as noisy labels of speaker IDs, there seemed to be no penalty in terms of downstream speaker verification EER.

Contribution: In this paper we consider four different types of label noise, and we conduct experiments to measure their impact on speaker embedding models used to perform speaker verification. By varying the amount and type of label noise, we can better understand whether accurate labels are necessary for training embedding models.

2. Related Work

Mitigation: Most of the previous research on the impact of label noise has focused on *mitigating* label noise to improve *classification* accuracy on *images*. Xiao et al. [9] proposed a probabilistic model to infer the true labels from noisy labels for image classification. Li et al. [10] developed a distillation method to train convolutional neural network when a known subset of the training labels are noisy, but they did not investigate the impact of the label noise on test accuracy. Zhang et al. [11] proposed a generalized cross-entropy loss which can combat label noise. Han et al. [12] devised a deep learning paradigm consisting two networks to filter out label noise for each other.

For the *audio* domain, much less research on label noise has been conducted. Akiyama et al. [13] applied multitask learning, semi-supervised learning, and ensemble methods to an audio tagging task to overcome noisy data. Bekker et al. [14] proposed an Expectation-Maximization algorithm to train neural networks to be robust to noisy labels and tested it on TIMIT [15] for phoneme classification.

Measurement: Rather than mitigating label noise, a few papers systematically *measure* its impact on accuracy: Nettleton et al. [16] conducted experiments on how data noise and label noise could affect shallow machine learning models (SVMs, Naïve Bayes, etc.). Rolnick et al. [17] explored how label noise would affect deep neural networks, demonstrate that while deep networks are robust to label noise, using larger datasets and a larger minibatch size can help even more. However, while both survey were conducted on multiple datasets, speaker verification tasks were not included.

The work most similar to ours is by Zheng et al. [18]. It looked at how label noise affected trained x-vector embeddings [3] for speaker verification and how effective were different regularization methods. They examined just a single type of noise – permutation of the dataset labels. In contrast, our work considers four different label noise models and both an embedding loss function (GE2E loss) and softmax/cross-entropy loss.

3. Dataset Labeling and Noise Models

A standard approach to collecting a large and diverse dataset (such as VoxCeleb and BookTubeSpeech) for speaker embeddings is to harvest audio and video files from repositories such as YouTube, Vimeo, etc., and to identify either manually or automatically (a) which speaker(s) appear in which files, and (b) when each person speaks within each file. What kinds of errors are likely to arise in this process? We consider four types – Split, Merge, Permute, and Corrupt – described below.

Assumptions and definitions: Except in the Corrupt noise model, we assume that the dataset contains no simultaneous speech and that each utterance is short enough to contain speech from only one speaker. We distinguish between the **true labels** of the dataset that accurately identify who spoke when, and the **noisy labels** that may contain some labeling errors and that are used to train a speaker embedding model. In the context of noisy labels, we define a **group** as a set of utterances that ostensibly belong to the same speaker (according to the noisy labels) but that might actually come from multiple real speakers.

Label Noise Model 1 (Split): In this noise model, the utterances from a randomly selected speaker are split into two distinct groups (Fig. 1 left). This can occur if a dataset collector believes that two audio files contained two distinct speakers, whereas in fact they both contained the same speaker.

Label Noise Model 2 (Merge): Utterances from two distinct speakers are merged into a single group (Fig. 1 center). This could happen if a dataset collector believes that an audio file belongs to a single speaker but in reality it belonged to two different speakers.

Label Noise Model 3 (Permute): The speaker IDs attributed to a *set* of utterances are randomly permuted (Fig. 1 right). This could occur (at least approximately) in a dataset consisting of just a few but very long audio files with many speakers where the speaker identities of many utterances were mislabeled. Alternatively, it could occur if a dataset collector blindly trusted a YouTube keyword search for a large set of specific people but did not check the search results for correctness.

Label Noise Model 4 (Corrupt): Finally, we consider a noise model whereby each individual utterance may contain speech from two speakers but is instead attributed to just a single speaker. (This contrasts with our other noise models that assume that each utterance contains speech from only a single speaker.) Another way of looking at it is that, even though a single utterance is attributed to just a single speaker, in fact it contains speech from multiple people (and thus the utterance's label for the second speaker is missing). To simulate this condition, for each utterance u from speaker s, we randomly select an utterance v from a random speaker t. Then, for a randomly selected segment within the utterance, we either (a) superpose v onto u or (b) replace u with v during that segment. Strategy (a) or (b) is chosen randomly with equal probability.

4. Embedding Architecture

We conducted our experiments using an embedding network trained with the GE2E embedding loss [2]. We also tried training the embedding network as a classifier using a softmax/cross-entropy loss and then discarding the softmax layer, similar to x-vectors [3]. We obtained generally better results with GE2E and thus focus on these in our paper to save space. Regarding the impact of label noise, we observed mostly similar trends for both training approaches; when significant differences exist, we note them when presenting results.

Training and testing data: We used the VoxCeleb2 development set (5,994 speakers) for training and the VoxCeleb2 test set (118 speakers) for testing. From each audio file, we used the `librosa.effects.split` function from the Librosa library [19] to identify non-silent intervals, from which we extract the MFCCs if the segment is sufficiently long. All models were trained for 400 epochs on the VoxCeleb2 development set (5,994 speakers) and evaluated on the VoxCeleb2 test set (118 speakers). Evaluation for speaker verification was conducted in minibatches consisting of 10 different speakers: For each speaker, we selected 3 random utterances for enrollment and 3 random utterances for verification. Based on the similarity scores obtained from the embedding model between pairs of utterances, we computed the Equal Error Rate (EER). We iterate over all speakers in the test set 300x to compute EER.

Features and architectures: We used MFCC features as inputs (window size of 0.025s, window step of 0.01s, and 40 filter-band banks), consisting of 160 frames (1.6sec of audio) fed to a 3-layer LSTM with 768 units, and then fed to a dense layer of 256 units. This layer is either normalized to have unit L_2 norm (for GE2E loss) or fed to a softmax classification layer (for softmax/cross-entropy loss).

GE2E: We implemented the "softmax" variant of GE2E [2]; note, however, that this loss function still operates on sets, not individual utterances, and seeks to achieve high cosine-similarity for embeddings from the same speaker and low similarity for embeddings from different speakers. Each minibatch consists of 5 utterances each from 4 groups (putative speakers according to noisy labels).

Softmax/cross-entropy: The network is optimized using cross-entropy over all speakers in the training set. Each minibatch consists of 64 randomly selected utterances from the entire training set. (We also tried training with minibatches of 4 groups of 5 utterances each, similar as for GE2E, but found that it worked poorly.) After training, the softmax layer is discarded, and the penultimate layer is used as the embedding vector.

Baseline performance: Training with GE2E loss and no label noise, we obtain an EER of 10.9% on VoxCeleb2 test and 6.66% on the VoxCeleb1 test. Though not quite state-of-the-art, the latter result is better than the VGG-based (7.8%) embedding model reported in the original VoxCeleb1 paper [5].

5. Experiments

Using the embedding architecture described above, we conducted experiments to measure the impact of label noise on speaker verification for each type of label noise.

5.1. Split

We initially implemented Split noise at the dataset level: We split 180 VoxCeleb2 speakers into 10 groups each, and also added 200 more groups from real speakers; we then selected 5 utterances from each group. When we trained a speaker embedding model using these noisy labels compared to the dataset's true labels, however, we found that the EER with noisy labels was no worse than with true labels (for both GE2E and softmax). The reason is likely that, within each *minibatch*, the amount of label noise is very small despite the large amount of label noise at the *dataset* level. Since the performance of a trained embedding is ultimately determined by the minibatches, we thus implemented a more aggressive form of Split noise directly on minibatches: With probability p, we generated a minibatch so that k groups were split from the same speaker. We then compared the EER of different embedding models trained with the noisy labels (for different combinations of $p \in \{0, 0.25, 0.5, 0.75\}, k \in \{2, 3, 4\}$) and also with true labels. Note that these values of p and k are very high. To put them into perspective, suppose a dataset of 6000 videos contains only 60 unique speakers with 100 videos/speaker, but that the

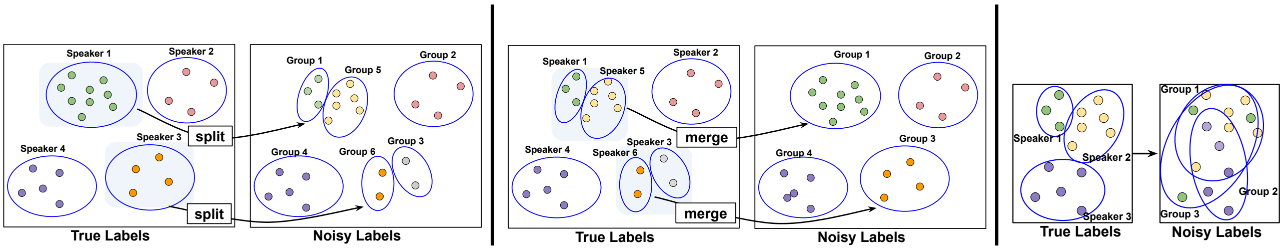

Figure 1: *Label noise models 1 (Split), 2 (Merge), and 3 (Permute), respectively. Best viewed in color.*

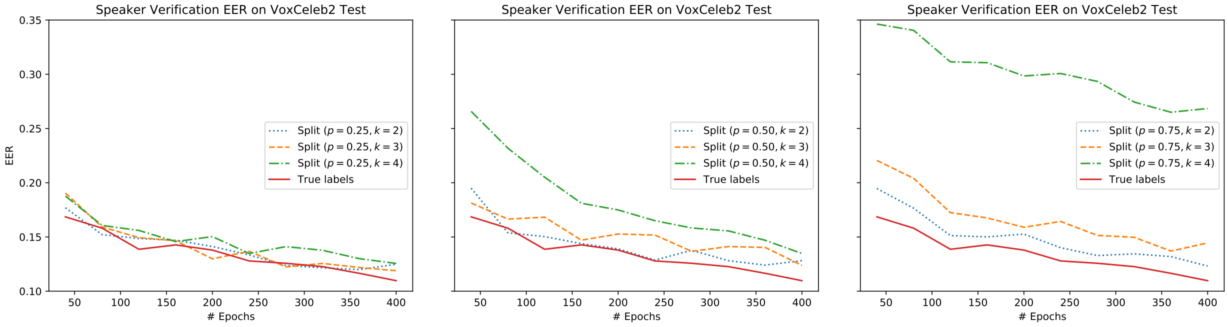

Figure 2: **Split label noise***: Effect on speaker verification EER, after training for 40 to 400 epochs, of splitting k out of 5 groups from a single speaker within each minibatch with probability p.*

noisy labels indicate that each video contains a *unique* speaker. Even under this amount of mislabeling, the probability over 400 epochs that a random minibatch contains $k \geq 3$ groups all from the same speaker is very small ($\frac{100}{100}\frac{99}{100}\frac{98}{100} = 0.0002695$).

Results are shown in Fig. 2. When 25% or even 50% of all minibatches are corrupted, then even for $k = 4$ (all 4 groups split from a single speaker), the impact on EER is mild by the end of training (just a few absolute percent difference) compared to training on true labels. Only for $p = 0.75$ do we find a large increase in EER (27.42% for $p = 0.75, k = 4$ compared to 10.9% for true labels). (Note: We did not conduct the experiment with softmax/cross-entropy since there was no obvious analog at the minibatch level.)

Discussion: One possible reason why Split noise has only a mild effect is that utterances belonging to the same *real* speakers can still be well-separated, even though they are erroneously subdivided into multiple *groups* due to noisy labels. For instance, in Fig. 1 (left), an embedding model is trained to separate groups 1 and 5 from each other. However, as long as the distance between groups 1 and 5 is relatively small compared to the distance between (for example) groups 1 and 4 or between groups 5 and 2, then the embedding can still be effective.

Also, the EER increased substantially only when almost all minibatches consisted of utterances from just one speaker ($p = 0.75, k = 4$). In this case, the GE2E loss that tries to separate utterances from different speakers must necessarily fail because *none* of the utterances come from different speakers. In a follow-up study with minibatches consisting of 6 groups, we found a similar trend: unless all 6 groups came from a single speaker, then the impact of Split noise was mild. This suggests a possible mitigation strategy to Split label noise: increase the minibatch size. As the minibatch size grows, the probability that all groups come from one speaker shrinks. This strategy has been previously suggested for image classification [17].

5.2. Merge

Like in Section 5.1, we varied both the probability and degree of label noise in each minibatch, as quantified by the number of corrupted groups. If a minibatch was randomly selected with probability p to be corrupted, then n of its groups were generated by merging m actual speakers. We varied $p \in \{0.25, 0.5, 0.75\}, n \in \{2, 3, 4\}$, and $m \in \{2, 5\}$. To put this into perspective, consider a hypothetical dataset of 6000 videos such that each video contains 8 unique speakers in equal proportion and with no speaker appearing in two videos. If the noisy labels indicated that every video contained just a *single* unique speaker (i.e., 8 real speakers per group), then over the course of 400 epochs, the probability p that a random minibatch would contain $k \geq 3$ groups each merged from $m = 5$ real speakers is less than 0.05 (estimated via simulation).

Results are in Fig. 3: The Merge label noise had a very small impact on the speaker verification EER except when (a) the probability of corruption within each minibatch was 0.75, (b) at least 3 out of 4 groups in the minibatch were merged from multiple speakers, and (c) each merged group consisted of a large number speakers ($m = 5$ in our experiments). Otherwise, the increase in EER was mild (around 1% or less). With softmax/cross-entropy, the trend was similar.

Discussion: The mild effect when training with GE2E loss might be due to a similar reason as for Split noise: As long as the embedding function can preserve the distinction *between* speakers within the same group while also separating (merged) groups, then it can still be effective for speaker verification.

5.3. Permute

In contrast to Split and Merge, we implemented permutation noise on the entire dataset, similarly to [18]. We randomly selected q percent of all utterances in the training set and permuted their speaker IDs, for $q \in \{10, 20, 50\}$.

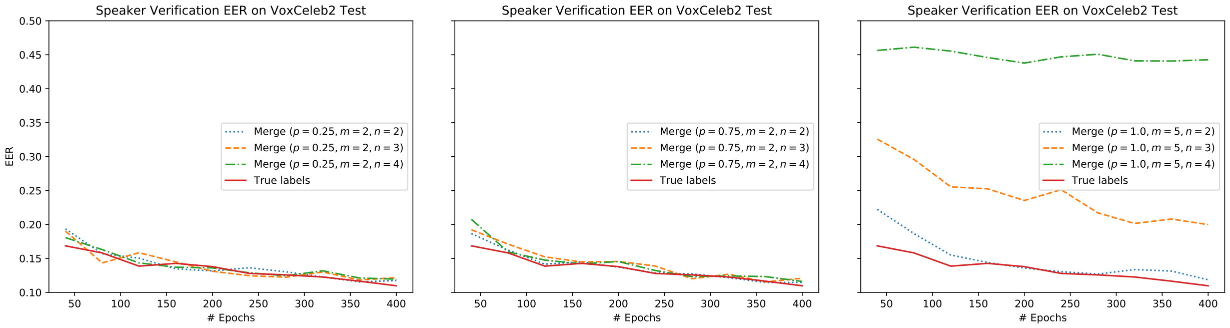

Figure 3: **Merge label noise**: *Effect on speaker verification EER, after training for 40 to 400 epochs, of merging m different speakers to form a single group for each of n (out of 5) different groups within each minibatch with probability p.*

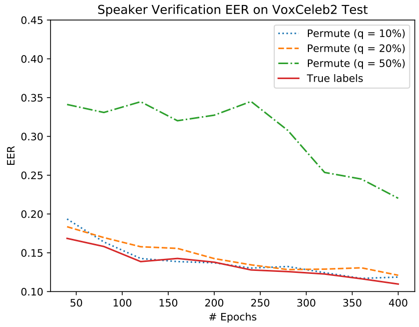

Figure 4: **Permute label noise**: *Effect on speaker verification EER, after training for 40 to 400 epochs, of permuting (once before training) the speaker labels of random utterances selected with probability p.*

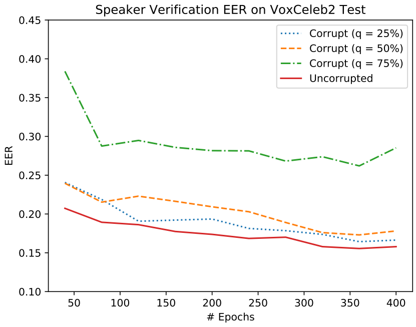

Figure 5: **Utterance corruption**: *Effect on speaker verification EER, after training for 40 to 400 epochs, of corrupting q percent of the speech signal of every utterance within every minibatch.*

Results are in Fig. 4: Even with the speaker IDs of 10% to 20% of the training utterances permuted, the speaker verification EER was hardly affected. We only saw a large increase in EER when we increase the q to 50%. The baseline EER was higher for softmax/cross-entropy, but the relative impact of permutation noise was actually even smaller when $q = 50\%$.

Discussion: Our findings contrast with those by [18], who saw a major increase in EER for 10-20% permutation noise on the NIST SRE dataset. Possible reasons include: (1) VoxCeleb2 might contain a greater number and diversity of recordings per speaker compared to NIST SRE04-10, and this might provide some robustness to noisy labels. (2) During enrollment, they used a single utterance embedding whereas we used the mean embedding from 3 utterances.

5.4. Corrupt

For this experiment we used only 3,000 speakers from the VoxCeleb2 development set for training and used the rest (2,994) for corruption noise; hence the baseline EER with true labels was higher. In each minibatch, all 5 utterances from all 4 groups are corrupted. In particular, each utterance has q percent of its raw waveform corrupted (at a randomly selected timepoint) by either replacing or superposing the utterance from a randomly selected speaker, for $q \in \{25, 50, 75\}$. This generates utterances that contain speech from two people instead of one.

Results are in Fig. 5: For corruption of 25% and 50% of each training utterance, the speaker verification EER suffered less than 5%. Only for 75% corruption of each training utterance do we see a strong increase in EER. Only for this kind of label noise did we see a relatively higher impact for softmax/cross-entropy compared to with GE2E.

Discussion: As long as at least half of each utterance is from the correct speaker, then the embedding model does not suffer too much. It is possible that the corruption noise provides a regularization effect.

6. Conclusions

We have conducted experiments on one of the largest publicly available speech datasets (VoxCeleb2) to measure the impact on speaker verification Equal Error Rate (EER) of different kinds of label noise (Split, Merge, Permute, Corrupt) on trained speaker embeddings. Our results suggest that, contrary to some prior results [18], highly accurate speaker labels may not be necessary. In our experiments, even high levels of label noise had only a slight impact on downstream speaker verification EER, using either GE2E or softmax/cross-entropy loss functions. This suggests that new datasets to train speaker embeddings might benefit most from having very large numbers of distinct speakers and recording conditions, at the expense of highly accurate labels.

7. Acknowledgements

This material is based on work supported by the National Science Foundation Cyberlearning grant #1822768.

8. References

[1] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 815–823.

[2] L. Wan, Q. Wang, A. Papir, and I. L. Moreno, "Generalized end-to-end loss for speaker verification," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4879–4883.

[3] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5329–5333.

[4] R. Doddipatla, N. Braunschweiler, and R. Maia, "Speaker adaptation in dnn-based speech synthesis using d-vectors." in *INTERSPEECH*, 2017, pp. 3404–3408.

[5] A. Nagrani, J. S. Chung, W. Xie, and A. Zisserman, "Voxceleb: Large-scale speaker verification in the wild," *Computer Speech & Language*, vol. 60, p. 101027, 2020.

[6] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," *arXiv preprint arXiv:1806.05622*, 2018.

[7] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[8] M. Pham, Z. Li, and J. Whitehill, "Toward better speaker embeddings: Automated collection of speech samples from unknown distinct speakers," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7089–7093.

[9] T. Xiao, T. Xia, Y. Yang, C. Huang, and X. Wang, "Learning from massive noisy labeled data for image classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 2691–2699.

[10] Y. Li, J. Yang, Y. Song, L. Cao, J. Luo, and L.-J. Li, "Learning from noisy labels with distillation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1910–1918.

[11] Z. Zhang and M. Sabuncu, "Generalized cross entropy loss for training deep neural networks with noisy labels," in *Advances in neural information processing systems*, 2018, pp. 8778–8788.

[12] B. Han, Q. Yao, X. Yu, G. Niu, M. Xu, W. Hu, I. Tsang, and M. Sugiyama, "Co-teaching: Robust training of deep neural networks with extremely noisy labels," in *Advances in neural information processing systems*, 2018, pp. 8527–8537.

[13] O. Akiyama and J. Sato, "Multitask learning and semisupervised learning with noisy data for audio tagging," DCASE2019 Challenge, Tech. Rep, Tech. Rep., 2019.

[14] A. J. Bekker and J. Goldberger, "Training deep neural-networks based on unreliable labels," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 2682–2686.

[15] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1," *NASA STI/Recon technical report n*, vol. 93, 1993.

[16] D. F. Nettleton, A. Orriols-Puig, and A. Fornells, "A study of the effect of different types of noise on the precision of supervised learning techniques," *Artificial intelligence review*, vol. 33, no. 4, pp. 275–306, 2010.

[17] D. Rolnick, A. Veit, S. Belongie, and N. Shavit, "Deep learning is robust to massive label noise," *arXiv preprint arXiv:1705.10694*, 2017.

[18] S. Zheng, G. Liu, H. Suo, and Y. Lei, "Towards a fault-tolerant speaker verification system: A regularization approach to reduce the condition number," *Proc. Interspeech 2019*, pp. 4065–4069, 2019.

[19] B. McFee, C. Raffel, D. Liang, D. P. W. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," 2015.