Deep Cost-sensitive Cross-entropy for Gender Generalization in TIMIT Phoneme Classification

Reza Shokrzad
(1056369)

June 2021

1 Introduction

Converting a sound waveform into hypothesized phonetic units is one purpose of designing Automated Speech Recognition (ASR). This task is specially addressed in the phonetic classification part of ASR. In this regard, correctly distinguishing phonemes is still a big challenge even in the state-of-the-art methods [1]. Traditional models such as HMM[1] and GMM[2] were the initial phonetic-centered methodologies that tried to get trained by individual elements (pronunciation, acoustic, and language model). While deep neural networks like RNN[3] and CNN[4] have revolutionized the process of having automated speech recognition models. The most significant distinction between traditional and deep neural networks is the ability to extract features specially Mel-Frequency Cepstral Coefficients (MFCC) that is the most well-known in this regard [2]. This work is an attempt to answer two specific questions in terms of the TIMIT corpus. As it will be scrutinized in section 3.1, there is a considerable gap in the size of classes that made me answer this question 'How can establish a robust model to tackle the imbalanced TIMIT data?'. In this respect, section 4.2 is allocated to develop this idea. On the other hand, gender generalization was the first concern of doing this project. To find the reply to the objection 'is there a phone classification model to expand from a gender to the other one?' changing the combination of data (training and test set) was needed that is explained in section 4.3.

2 Method

As end-to-end speech recognition systems have been widely focused on the advantages of deep neural networks, in this work, we applied a state-of-the-art CNN model proposed by Zhang et al,[3]. In detail, I implemented a several-layer network with a specific architecture which can be seen in figure1.

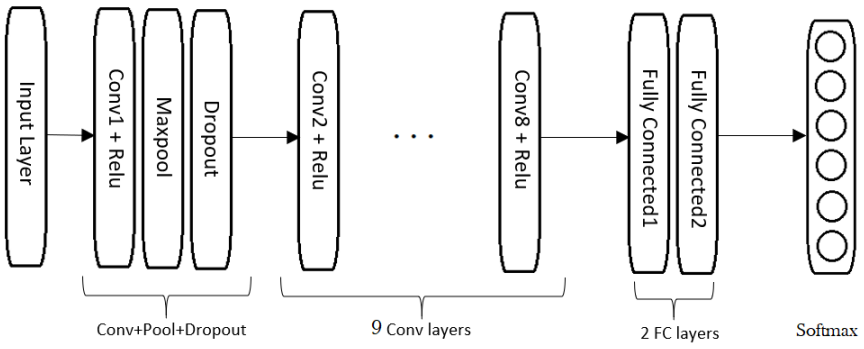


Figure 1: The proposed Convolutional architecture

[1]Hidden Markov Models
[2]Gaussian Mixture Models
[3]Recurrent Neural Networks
[4]Convolution Neural Networks

3 Setup

3.1 Description of the data

TIMIT is a widely-used speech corpus includes 16-kHz-wav voice files and phonemically transcribed speech from men and women with various dialects who speak American English. This dataset is appropriate for both speech and speaker recognition experiments. TIMIT has 6300 sentences overall that 630 narrators said 10 sentences, called utterances, from 8 general dialect regions of the United States. Furthermore, each utterance is phonetically hand-labelled from 61 unique label shown in table 1 [4]. As what was proposed in [7] 3590 labels of 'q' are also removed from the phonetics. The dataset includes 177080 training and 64145 test samples.

Phone Class	Numbers	TIMIT Labels
Vowel/Semivowel (VS)	25	aa ae ah ao aw ax axh axr ay eh er ey ih ix iy ow oy uh uw ux el l r w y
Nasal/Flap (NF)	8	em en eng m n ng nx dx
Strong Fricative (SF)	6	s z sh zh ch jh
Weak Fricative (WF)	6	v f dh th hh hv
Stop (ST)	6	b d g p t k
Closure (CL)	9	bcl dcl gcl pcl tcl kcl epi pau h#

Table 1: Classes of phonetics (Halberstadt, 1998)

The distribution of 61 phonemes initial phonemes is depicted in figure 2-(a). Furthermore, in figure 2-(b), we can see the suggested categorization to six classes.

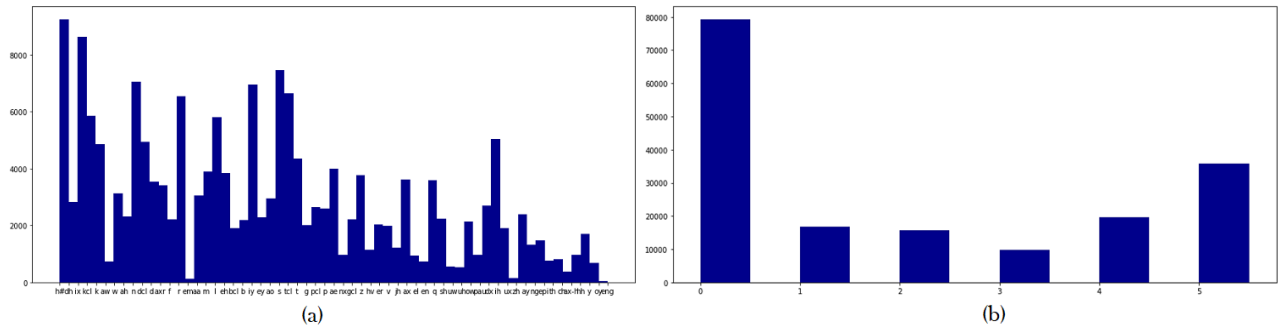


Figure 2: Distribution of phoneme numbers (a) before categorization with 61 distinct phonemes (b) after merging classes to have six recommended groups

As it is clear in figure 2-(b), the most significant issue related to this dataset is imbalanced data. This visualization proves the challenge of asymmetric classes. Here is a dictionary includes the number of phones in each class {class0: 79199, class1: 16866, class2: 15668, class3: 9898, class4: 19572, class5: 35877}.

Before diving into preprocessing phase that is explained precisely in section 4.1, here there is an example of one wav file from TIMIT dataset. This length of file is 64493 with sample rate 16000 that means the sound takes long 4.03 second. Figure 3 is a visualization of this file which has default parameters in (20, 126) that would change to (13, 404) after our process.

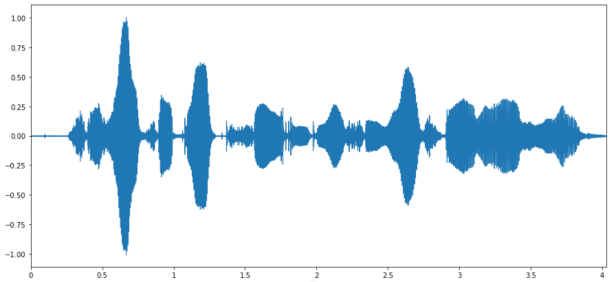


Figure 3: An example of a visualized wav file in TIMIT DataSet

3.2 Tuning/adaptation model parameters

Related the task of gender generalization, out of 630 speakers, 438 are men, and 192 are women in dataset. This combination is 70% men and 30% women that is an appropriate portion to make a new train and test set for the task of gender generalization. So, to address this task, we merge all old train and test content to generate a new version of the dataset. Afterward, we split the set to men's voices as train and women ones for the test. The frame length that is considered for this work is 10 ms. In addition, a standard recipe to convert TIMIT's audio files to MFCCs is Kaldi's 's5' that generates 13 MFCCs as input features [6].

4 Experiment(s)

4.1 Preprocessing

A Python package called **librosa** is applied to load wav files and their sample rate. Librosa used a default sample rate equals 22050 kHz, but it is changeable. As the frequency of dataset wav files is 16000 kHz, I altered the sample rate (sr) to this value. Furthermore, in our setup hop_length is considered 160. To extract features from wav files, **mfcc** is implemented that creates a matrix M of size (13 $\times$ N) in which N equals 100 times (duration of wav files in seconds). So, there are N vectors of 13 dimensions per file. In parallel, there is the PHN file information in the dataset that helps to identify all vectors. In the training process, I considered M[t-3:t+3] includes seven vectors as the input of the model and its corresponding label as the output. Besides, to present the matrix M[:, t-3:t+3] as input for the network, I reshaped it into one *supervector*, by stacking. For instance, if a phone has a length of 15, we would have 3 frames for it with size (13 $\times$ 7 = 91). Although, I set a threshold to remove those frames containing zero values more than four. Additionally, each wav file is divided by its energy to be normalized. To calculate the file's energy $wav.wav$ in which '.' stands an element-wise inner, would apply (equation 1). So, each element of the wav file would be divided into a specific value in this approach. In this formula wav_N indicates normalized wav file.

$$wav_N = \frac{wav}{\langle wav, wav \rangle} \tag{1}$$

Besides, *StandardScaler* from package *sklearn* is applied to standardize data that this module uses mean and variance (standard deviation) to standardize data. Altering the loss function by making it sensitive about missing data for some minority classes is the next step that is explained in detail in the following section. Figure 4 illustrates the whole procedure of the current task.

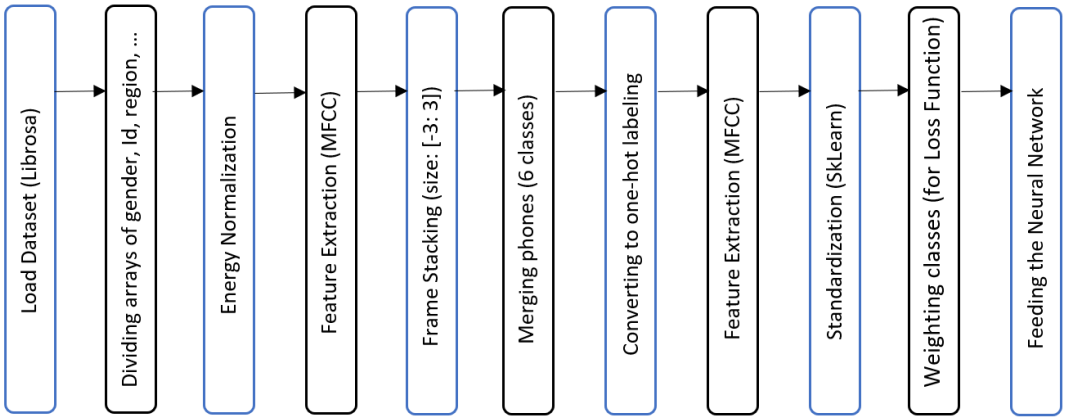


Figure 4: A big picture of the entire process

4.2 Data Balancing Strategy

Imbalanced class division in classification tasks is one of the most significant obstacles in the learning process. In this case, missing even one true-positive instance for the minority class would be much worse than incorrectly classifying an example in majority one[8]. Since scarce instances appear infrequently, classifiers predict the small classes undiscovered or completely ignored; consequently, samples belonging to the minorities would be misclassified more probable than those from the prevalent groups [9]. From the other perspective, the network prefers to minimize the loss for the high-occurrence classes, while all groups have the same importance in our task. Three strategies to

tackle this difficulty are *sampling* techniques, *cost-sensitive* techniques, *one-class learning*. Over-sampling (figure 5-(a))the minority, under-sampling the majority (figure 5-(b)), and a hybrid model, are three kinds of sampling methods.

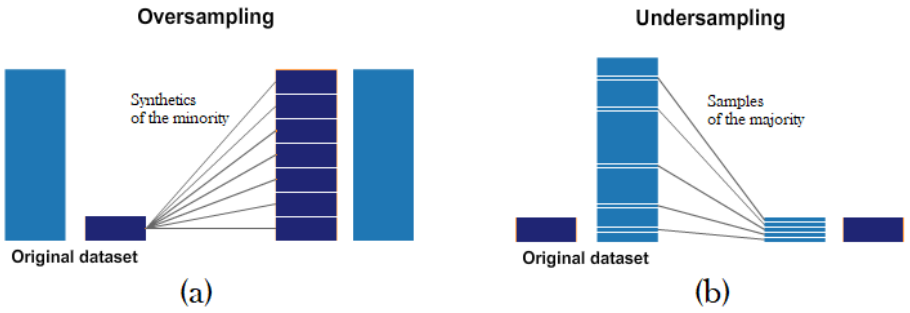


Figure 5: Two approaches (over and under sampling) in the sampling strategy

Regarding one-class learning models, only target-class instances are applied in the absence of counter-class samples. This method calculates the degree of similarity between a query object and the target class, with a threshold applied to the similarity value. In the case of cost-sensitive classification, misclassification loss is totally oriented on the size of each class, as the smaller class, the more cost for wrongly classifying. In this work, I tried to change the loss function to penalize the missing information of smaller classes. As we know, Cross-Entropy (CE) loss function is elicited from the formula which can be seen in equation 2 in which τ represents the target probability distribution.

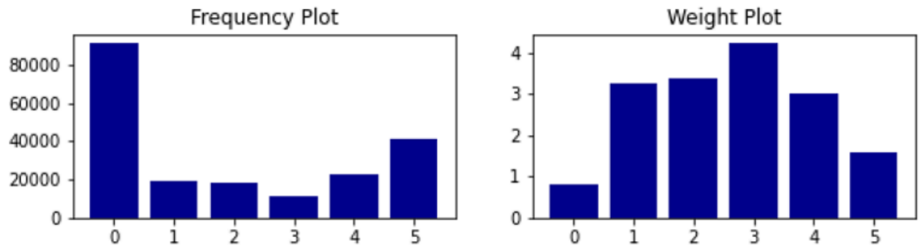
$$E^{CE} = -\sum_{i=1}^{c} \tau_i log(y_i) \tag{2}$$

To be more detailed, I calculated the partial ratio of each group and their shares in total. I considered a penalty by dividing the total by the number of each class, followingly smoothing by log for two majorities and square for minority classes. The equation 2 will be updated to equation 3 in which λ_i is a weight of class i and will be calculated with equaiotn 4.

$$E^{CE} = -\sum_{i=1}^{c} \lambda_i \tau_i log(y_i) \tag{3}$$

$$\lambda_i = \begin{cases} log(\frac{totalfreq}{freq_i}) & i == 0 and 5 \\ \sqrt{\frac{totalfreq}{freq_i}} & otherclasses \end{cases} \tag{4}$$

Figure 6 shows the distribution of how classes get weighted in the loss function.



Share of Class 1	91122
Share of Class 2	19235
Share of Class 3	18036
Share of Class 4	11453
Share of Class 5	22445
Share of Class 6	41464

(a)

Penalty of Class 1	0.8
Penalty of Class 2	3.25
Penalty of Class 3	3.36
Penalty of Class 4	4.21
Penalty of Class 5	3.01
Penalty of Class 6	1.59

(b)

Figure 6: (a) The distribution of each class's share (b) The distribution of the amount of penalty for each class using in the loss function

4.3 Gender Generalization

TIMIT has been one of the most familiar speech datasets for a long time. Therefore, there are several attempts to address any possible question related to it. For instance, Yao, et al proposed an integrated system for robust gender classification with convolutional restricted Boltzmann machine (CRBM) and spiking neural network. They tried to use CRBM as a feature extractor, tailed by a spike-latency encoding layer. Followingly, spiking neural networks with the tempotron learning instruction will process the obtained spikes[10]. Sulaiman et al. proposed a numerical integration method based on Simpson's rule. Focusing on the fact that males and females can be classified based on the intensity of their utterances was their strategy. They highlighted measuring the area under the normalized curve[11]. To keep the scope of the course, I assumed that it is possible to classify the gender by using the same architecture for the main-six classification task. Thus, first, I merged both train and test TIMIT data, then split it again to 70% men as train data and 30% women as the test instances. After this step, I exactly implemented the same procedure explained in the preprocessing part for this section.

5 Analysis and Results

In the model, the first convolution layer includes 32 filters, also the following ones have from 64 to 1024 filters. Both Fully Connected layers include 1024 neurons. In all layers, the filter size is 3 and the model is one-dimensional. Input data shape is (91, 1) in which 91 stands for 13×7. Additionally, in this setup, the loss function is cross-entropy, and the optimizer is Adam. I ran the model several times with batch-size 32 and the number of epochs 64 and 128. The training (a) and validation (b) accuracy of the model is displayed in figure 7 which indicates that the accuracies are 83.4, 83.3, respectively, in the task of gender classification.

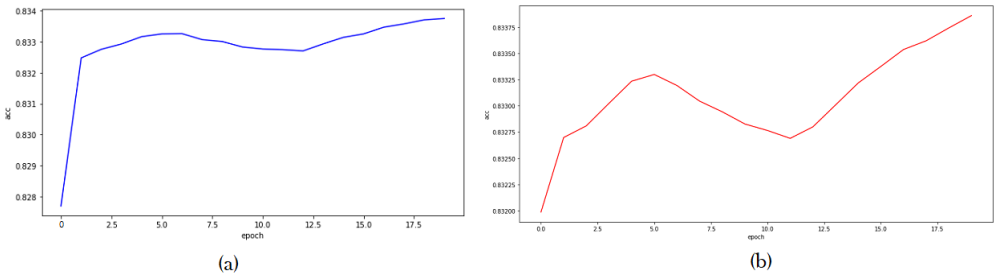


Figure 7: (a) Training accuracy trend (b) Validation accuracy trend

6 Discussion and Conclusion

The complexity of imbalanced data in the well-known six-level classification is a significant obstacle that we could partially deliver by a cost-sensitive loss strategy. Although we could prove that a cost-sensitive cross-entropy classification can be helpful for sex generalization task with accuracy 83.3, applying other methodologies which have been listed in section 4.2 and comparing their outcome to our results could be interesting, for future attempts. In terms of deep network architectures, in some literature, it is indicated that there is a threshold for the number of neurons for a network to learn well enough. However, in the scope of the course, I could not boost the neurons to run a larger network on my device. To conclude, I tried to answer if gender generalization is feasible by train a deep neural model. Meanwhile, attacking the challenge of imbalanced data got the other objective. The result showed gender generalization in phone classification, and generally, speech recognition can be possible in some specific experimental setup. Finally, defining a cost-sensitive loss function was practiced that showed this approach could confront the gap among classes in the TIMIT dataset.

References

[1] Oh, D. et al. (2021) 'Hierarchical phoneme classification for improved speech recognition', Applied Sciences (Switzerland), 11(1), pp. 1–17.

[2] Toledano, D. T., Fernández-Gallego, M. P. and Lozano-Diez, A. (2018) 'Multi-resolution speech analysis for automatic speech recognition using deep neural networks: Experiments on TIMIT', PLoS ONE, 13(10), pp. 1–24.

[3] Zhang, Y. et al. (2016) 'Towards end-to-end speech recognition with deep convolutional neural networks', Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, 08-12-September-2016, pp. 410–414.

[4] Halberstadt, A. K. (1998). Heterogeneous Acoustic Measurements and Multiple Classifiers for Speech Recognition. Ph.D. thesis, Department of Electrical Engineering and

[5] Bai, L. et al. (2018) 'Exploring how phone classification neural networks learn phonetic information by visualising and interpreting bottleneck features', Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, 2018-September(September), pp. 1472–1476.

[6] Povey, D. et al. (1968) 'Sensory-perception testing box.', Canadian Journal of Occupational Therapy, 35(4), p. 140.

[7] Lopes, C. and Perdigao, F. (2011) 'Phoneme Recognition on the TIMIT Database', Speech Technologies, (June).

[8] Fan, W., Stolfo, S. J. and Chan, P. K. (no date) 'AdaCost : Misclassification Cost-sensitive Boosting'.

[9] Sun, Y., Wong, A. K. C. and Kamel, M. S. (2009) 'Classification of imbalanced data: A review', International Journal of Pattern Recognition and Artificial Intelligence, 23(4), pp. 687–719.

[10] Yao, Yanli Yu, Qiang Wang, Longbiao Dang, Jianwu. (2019). An integrated system for robust gender classification with convolutional restricted Boltzmann machine and spiking neural network. 2348-2353.

[11] Alsulaiman, M., Ali, Z. and Muhammad, G. (2011) 'Gender classification with voice intensity', Proceedings - UKSim 5th European Modelling Symposium on Computer Modelling and Simulation, EMS 2011, pp. 205–209. doi: 10.1109/EMS.2011.37.