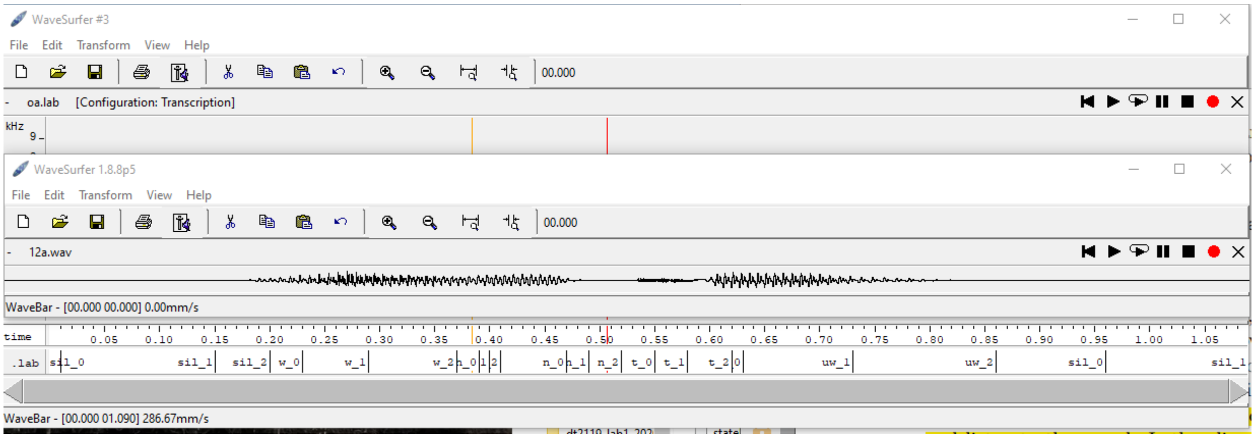
DT2119 Lab3: Phoneme Recognition with DNN

Zhanpeng Xie, zxie@kth.se

4.2 Forced alignment

I select the audio from '../data/tidigits/disc_4.1.1/tidigits/train/man/ae/12a.wav'. From the WaveSurfer result, the transcription is correctly aligned but not very accurate.



4.6 Feature Standardization

- *In the third case, what will happen with the very short utterances in the files containing isolated digits?*

Case 2: This normalization strategy could lead to poor performance when testing the model that is not in the training dataset, since we don't have enough data in general to estimate a mean and variance for each speaker. Case 3: We may increase the amplitude of the background noise which lead to a noisy sample. Case 1: I think it is the most balanced solution. However, it can also perform very poorly if the speaker's voice is very different from the majority.

5 Phoneme Recognition with DNN

- *why you chose the specific activation and what alternatives there are.*

I used Relu for the activation functions for the hidden layers and softmax layer for the output layer. Softmax layer allows us to have probabilities values for every class. Relu activation function is a common choice for NN, it is able to avoid problems like gradient vanishing or gradient explode for very deep neural network. Since my networks weren't very deep and only had three layers, I think it is fine to use activation function like Sigmoid (), Tanh().

- *What is the purpose of the validation data?*

To monitor the training process. Since validation data was not used for training, we can select the final model for testing based on the validation accuracy to prevent having an overfitted model.

- *What can you say comparing the results on the training and validation data?*

Training data had higher accuracy score and it was increasing throughout the whole training process. The validation accuracy is lower than the training one and stopped increasing after several epochs of training. This result shows the model was overfitting to the training data for later epochs.

5.1 Evaluation

- *Plot the posteriors for each class for an example utterance and compare them to the target values. What properties can you observe?*

The posteriors from target values and the predicted values have very similar pattern in general. Most probability mass are at the correct labels. However, models without dynamic features have more values at the wrong labels than ones with dynamic features.

5.2 Possible questions

- *what is the influence of feature kind and size of input context window?*

According to my experiment results, both feature kinds performed very similarly. Using dynamic feature (context window) has brought some improvements in performance. The validation accuracy without dynamic feature is about 0.65 whereas the validation accuracy with dynamic feature reached around 0.80.

- *what is the purpose of normalising (standardising) the input feature vectors depending on the activation functions in the network?*

Normalization is a common technique in deep learning. It can make my models more robust against factors that could change the range of input data values like varying volume. It also allows the models to weight different features more equally.

- *what is the influence of the activation function (when you try other activation functions than ReLU, you do not need to reach convergence in case you do not have enough time)*

Same answer as in section 5 "*why you chose the specific activation and what alternatives there are.*"

- *what is the influence of the learning rate/learning rate strategy?*

If the learning rate is too large, the model will not converge. If the learning rate is too small, the learning to become very slow and it would take much more epochs for the model to converge. Using different learning rate or learning rate strategy will influence the performance a little bit if they were in a reasonable range.

- *How stable are the posterior-grams from the network in time?*

According to four posterior-grams I obtained with trained models, I found there is less stability when the phoneme is changing or when the utterance is short.

- *how do the errors distribute depending on phonetic class?*

From the figure in the "confusion matrix.png", most errors distributed in the same phonetic class but different phoneme index. For example, our models sometimes predicted 'sil_1' for 'sil_0' leading to an error. Note that this confusion matrix is based on the model trained without dynamic features which implies the models didn't have very much information about the context.