Speech Task № 7

Vladyslav Bondarenko, University of Sheffield May 31, 2020

Instructions for setting up and running the solution are provided in the README.md file

Evaluation

Aside from accuracy other metrics extracted from confusion matrix are used. **Precision**, **recall** and **F1-score** are key in evaluating the success of classification system, particularly when there are class imbalances present. Other metrics that could be considered are: commonly used for optimization **categorical cross entropy** which summarizes how far away predicted distribution is from true one; receiver operating characteristic (ROC) curve and accompanying area under that curve (AUC) could also be used but is usually more appropriate for binary classification. Model training quality can be assessed from the speed of convergence represented by number of iterations.

Vowel Classification

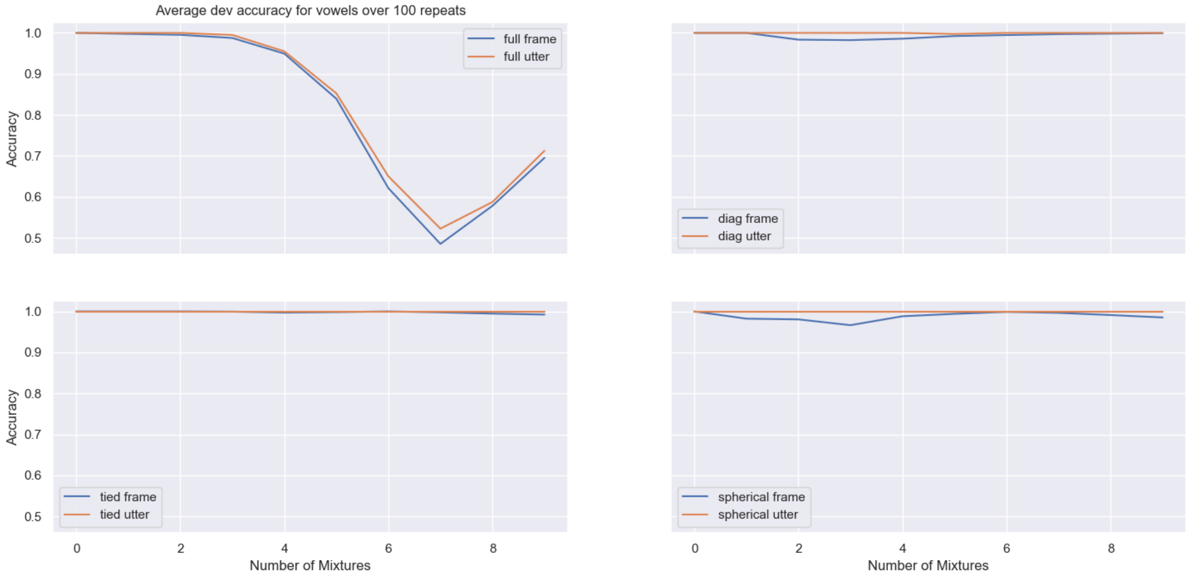

Figure 1: Optimization results on development set for vowels

GMM optimal hyper-parameter selection has been performed with results demonstrated above in Fig. 1. This has been performed by splitting the training set into two subsets with training and development data. The models are trained on train set and then evaluated on development set for each parameter combination. Two key parameters are optimized - number of mixture models and co-variance type. Although development data is drawn directly form training utterances and prone to over-fitting, it is still more robust than evaluating on training set (or test set for speakers).

This problem is clearly demonstrated in top left of Fig. 1 where as we increase the number of mixture models, the number of model parameters increases and the model starts to over-fit to the training set. As number of mixtures increases further development accuracy starts to increase again suggesting that the model is now over-fitting to the development set as well. This issue is only so apparent for **full** co-variance type as number of trainable parameters is much larger than for other set-ups. Convergence analysis demonstrated in left part of Fig. 2 confirm this with small number of iterations required by **full** covariance to train the model.

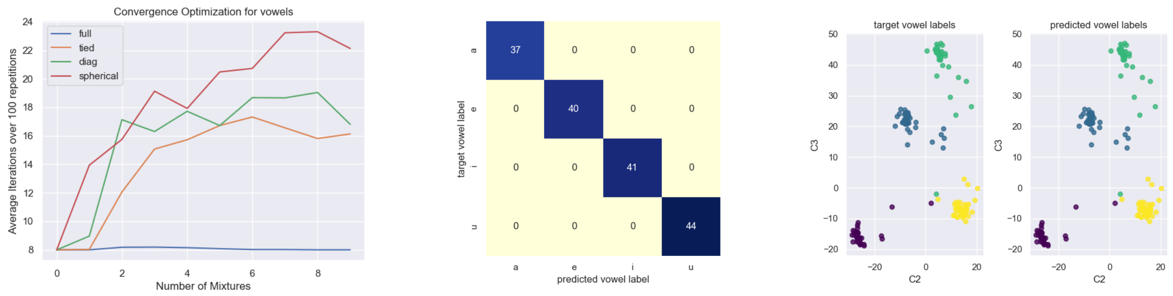

Figure 2: Vowel iteration optimization (left), final confusion matrix (mid) and comparison of true and predicted classes (right)

It has been found that even using just one mixture model with **full** covariance (where each component has its own covariance matrix) results in 100% accuracy on frame vowel classification. Confusion matrix and scatter plot comparison for true labels are shown in middle and right of Fig. 2 accordingly. Summing up log probabilities over the frames for each utterance, obviously also results in 100% accuracy. All other metrics such as F1 also report perfect scores as expected.

Although great results are demonstrated, they should be take with a grain of salt. Evaluation has been done on the same set as training which is prone to over-fitting. There are essentially no limits on model performance for the training set since we could set number of mixtures equal to number of data points and always achieve a perfect training score. Bringing an unseen data point to such a model would result in an unpredictable behaviour since it would almost always fall outside any trained distribution.

Speaker Classification

Similar optimization procedure has been undertaken for speaker classification for each feature type. Results for formant feature optimization are shown in Fig. 3 where **full** covariance perform the best again achieving almost 100% accuracy on full utterances with only 2 mixture models. Other features produced similar results with **spherical** consistently under-performing and other two reaching high accuracy scores but only with more mixture components. Visual are included as a reference at the end of the documents (Fig. 5, 6)

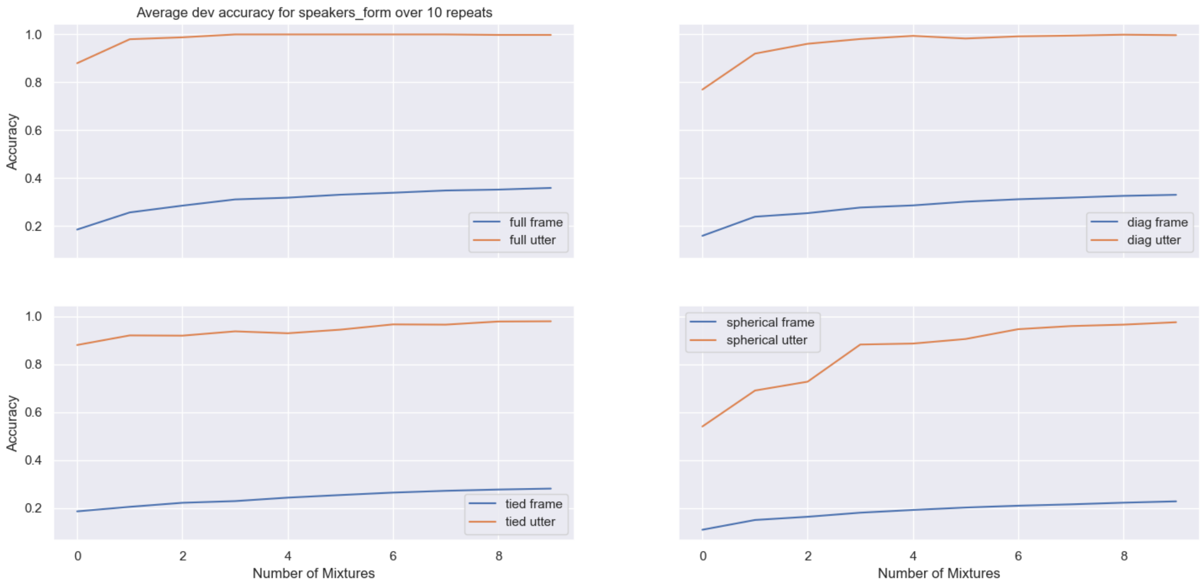

Figure 3: Optimization results on development set for speaker classification with formant features

It can be clearly observer that utterance level prediction is consistently more accurate than at the frame level. This can be explained by some noise in features at local level which is eliminated with aggregation of probabilities. It is still important to optimize the frame level accuracy as length of utterances that need to be recognised could be arbitrarily short.

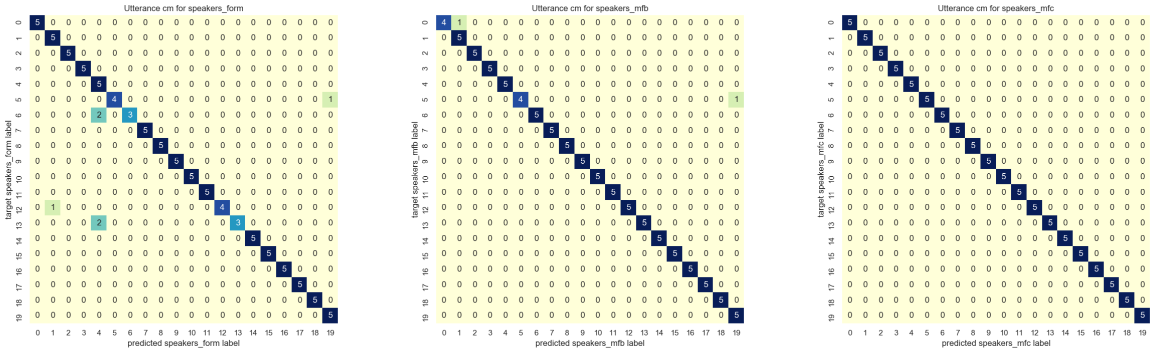

Figure 4: Utterance level test results for different features for speaker classification. Left to right - formants, mel filterbanks, mfcc's

Final evaluation has been performed for each feature on test set. Test set consists of the same speakers producing different utterances making evaluation on it less prone to over-fitting. Utterance level classification results are reported in a form of confusion matrices in Fig. 4. Top accuracy for formant features is 24% per frame and up to 91% on full utterances. For mel filter-banks some instabilities in training have been observed particularly for higher number of mixture models. This has been solved by scaling the features to be between 0 and 1. Scaling has resulted in the increase in utterance level performance from 90% to 96%. Finally, MFCCs performed the best achieving 40% accuracy for individual frames and often perfect 100% accuracy for utterances.

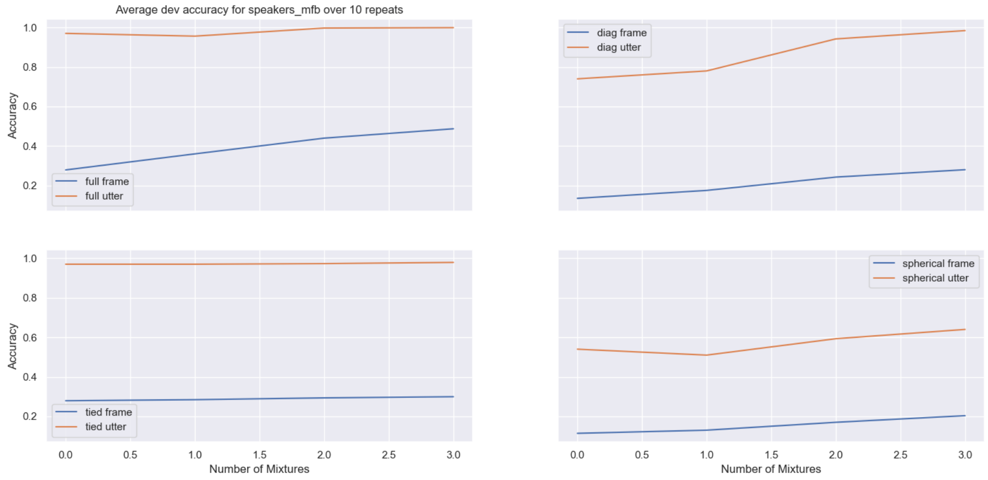

Figure 5: Optimization results on development set for speaker classification with mel filterbank features

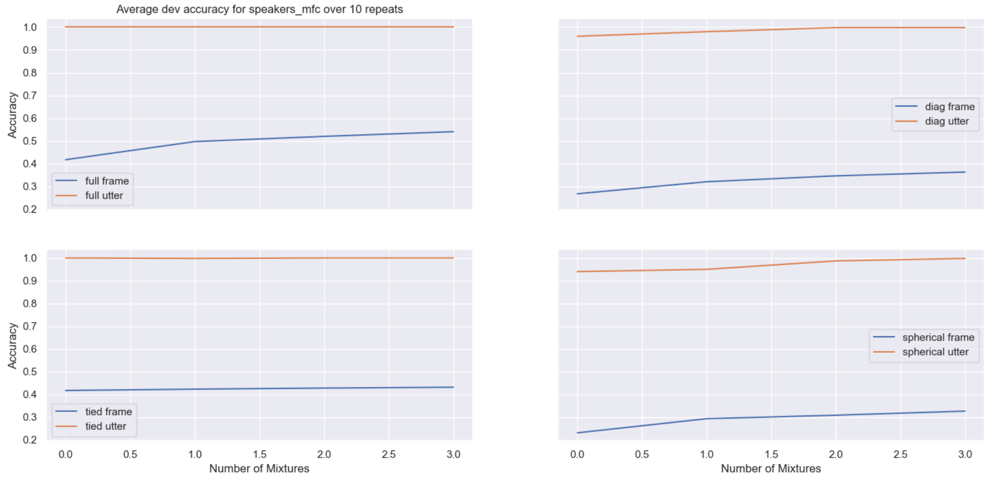

Figure 6: Optimization results on development set for speaker classification with MFCC features