Automatic Foreign Accent Recognition via Native Speech Models

By Alden Coots, Emil Lundberg and Joel Forsberg

DT2118 Speech and Speaker Recognition
Project Report

Abstract

Pronunciation variations in speech are one of the main difficulties in speech and speaker recognition tasks. Examples of these are accents and speech in a speaker's non-native language, which is treated in this report. Foreign accents tend to reduce the effectiveness of speech recognition systems. Speech classification has a number of application areas; for example identifying a foreign accent before speech recognition is applied allows a system to be more flexible by giving it the opportunity to make use of accent specific models. In this investigation, we attempt to classify non-native accented English speech using Gaussian mixture models trained only on speech corpora of native speech in other languages.

1 INTRODUCTION

If everybody spoke the same way every time they uttered something, speech recognition would be much more effective. Variations are of great importance for the characteristics of speech; a word never sounds the same twice. According to Sumner [1], accented speech makes the speech recognition task even harder.

Speaker variability and pronunciation variations have been a topic for researchers for some time and Benzeghiba et al. [2] write that the small variations are due to factors such as speaker, gender, age, regional accent, speaking style and speaking rate among others. Accented speech is associated with a shift within the feature space [2].

This report covers a project in the course *DT2118 Speech and Speaker Recognition*, carried out at KTH. The aim of the project was to investigate the possibilities of accent and non-native speaker recognition using training data consisting only of native speech of different languages. Tools used include the

Hidden Markov Model Toolkit (HTK) [3] and a CUDA implementation of Expectation Maximization for training Gaussian mixture models (GMMs) [4]. A smaller-size literature study on the topic was done, dealing with the possibilities of accent identification and the problems with pronunciation variations. These are presented in section 2. A description of the experiments and the speech corpora that were used is given in section 3, followed by a results, discussion and conclusions in sections 4, 5 and 6, respectively.

2 BACKGROUND

Chen et al. [5] claim that the main difficulties in speech recognition are due to speaker variability, where accent and gender are the two most important factors. They propose a successful method that makes use of GMMs to identify four different Mandarin accents, with better performance for female speakers than for male speakers. When using GMMs there is no need for transcriptions of the speech corpus, which differs from the use of hidden Markov models (HMMs). Chen et al. mention that accent identification is important because accent-independent systems generally perform 30 % worse than accent-dependent systems. The number of Gaussian components greatly affects the performance of a GMM; Chen et al. use 32 components in their experiments, but report that 64 components gave a better performance at the cost of being more time-consuming system. An accent identifier can be used as a model selector for the adaptation to a single model in a set of multiple models with smaller accent variations.

Techniques regarding pronunciation variations can also be used to help people learn a new language, according to Alsulaiman et al. [6]. They write about pronunciation issues for learners of Arabic as a second language, claiming that the difficulty of learning depends on the similarities between the learners' first language and the Arabic language. They use a method with GMMs for identifying the origin of three groups of speakers of Arabic as second language. If the speakers' accents are known, different acoustic and lexical models can be used, allowing the automatic speech recognition (ASR) system to perform better [7]. Hanani et al. [7] compare the human ability to recognize accents within the British English language with their language identification system using primarily GMMs. The recognition error rate was about four times greater for humans.

A model that is trained only on native speech, including accented native speech, cannot handle non-native accented speech properly [2]. Non-native speakers will often replace sounds that are not present in their native language with the native language's closest one. These sorts of errors cannot be handled by the usual triphone-based modeling, according to Benzeghiba et al [2]. If pronunciation variations are not taken into account, the performance of the ASR system will suffer. This is best done using several Gaussian mixtures instead of HMMs [8]. A native acoustic model can be adjusted using a non-native speech corpus, which in turn will make the ASR system's performance better [9].

Nguyen et al. [10] present a method for classifying Australian speakers into groups based on accent, gender and age using Mel-frequency cepstral coefficient (MFCC) features to train Gaussian speaker models. They conclude that it was easier to recognize a certain cultivated Australian accent for females than for males.

3 EXPERIMENT SETUP

3.1 Speech Corpora

As the goal of the project was to investigate the possibility of classifying foreign accents, the speech corpus used was a combination of corpora of different languages:

1. Native French speakers
2. Natve German speakers
3. Native Polish speakers
4. Native Spanish speakers
5. Native Swedish speakers
6. Native Turkish speakers
7. French speakers of English
8. German speakers of English
9. Polish speakers of English
10. Spanish speakers of English
11. Turkish speakers of English

Except for Swedish, these were all downloaded from the Backbone project and converted to WAV files using FFmpeg [11]. The Swedish corpus was downloaded from SweDia [12]. There was little time to assess the quality of

the recordings, but they were assumed to be good enough for the investigation in question. No distinction was made between male and female speakers, or speakers of different ages.

The native speech corpora were used only for training the GMMs, and the non-native speech corpora were used only for testing. A corpus of Swedish speakers of English was not found within the time frame of the project, but the native Swedish model was kept as an additional competing model in the classifier.

3.1 Feature Extraction

The first step in classifying speech is to convert the raw waveform data into vectors of features capable of reflecting the variances that occur in pronunciation. MFCCs attempt to do this on a scale that is indicative of the logarithmic perception of pitch in human hearing, and as such have long been a standard in the field of speech recognition. We begin by extracting 39-dimensional MFCC feature vectors from input speech using predefined methods in HTK. These vectors are composed of 13 Mel-frequency cepstral coefficients, 13 delta coefficients and 13 acceleration coefficients. The delta and acceleration coefficients are the finite first and second order derivatives, respectively, of the cepstral coefficients and aim to model the temporal nature of speech. Feature vectors are obtained from 25 ms long Hamming windows of speech taken every 10 ms.

3.3 Visualization of Features

Before training models and computing likelihoods, the 39-dimensional MFCC feature vectors of a selection of languages in the corpus were projected onto two dimensions for visualization, in the hope that some differences in the MFCC distributions might be discernible by manual inspection. This visualization was achieved using a self-organizing feature map (SOM) implementation of an artificial neural network (ANN). This is a vector quantization technique in the form of an unsupervised learning algorithm that attempts to find a topology preserving map from the feature space to a 2-dimensional space [13]. Details on the training parameters are given below. The training data consisted of 10,000 randomly selected MFCC feature points from each of the English, French, German, French English and German English corpora[1], for a total of 50,000 data points. The resulting trained network was then used to map the 50,000 training points and an additional

[1] Because of logistic problems, the MFCC values of the other corpora were not available at the time, which is why only these corpora were included for visualization.

20,000 randomly selected data points from each of the languages, for a total of 150,000 data points, to the nodes in the 2D grid. The density of MFCC feature vectors of the corpus could thus be visualized as seen in Fig. 1. Note that these results were not used in the later classification, this is only a way to visualize the MFCC feature vectors.

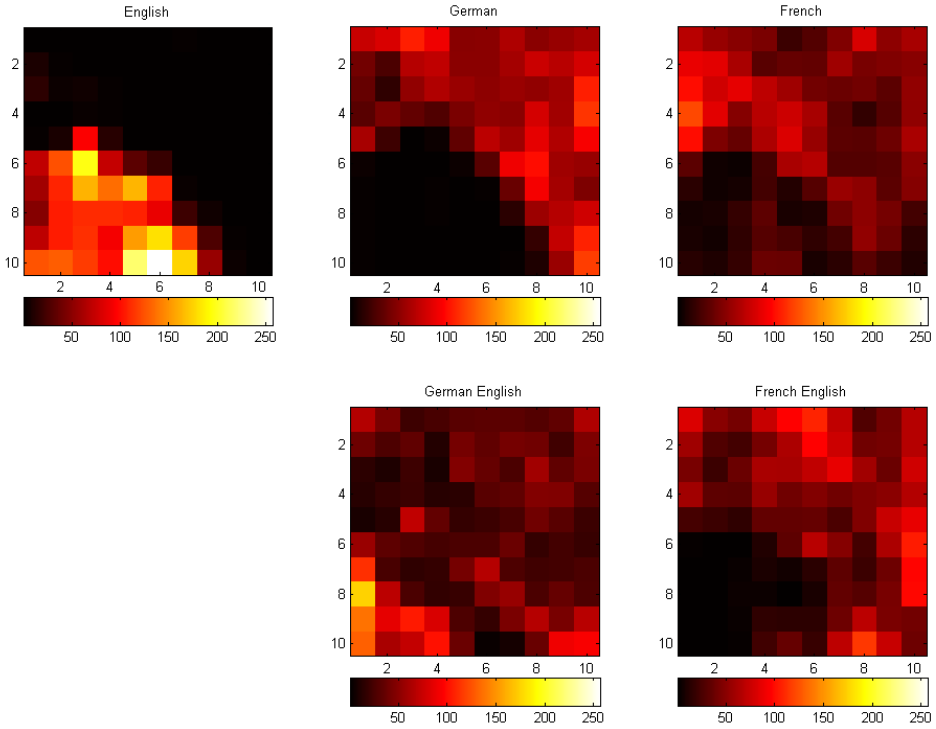

Fig. 1. Heatmaps of projections of MFCC vectors from a selection of the languages of the corpus onto a 10-by-10 grid using a SOM ANN. From top left: English, German, French, English with German accent, English with French accent. High values indicate a high density of feature vectors. Relative values and coordinates are meaningful, absolute values and coordinates are meaningless.

The images indicate that there might be some systematic differences in which MFCC vectors appear in the different languages.

3.3.1 SOM Training Parameters

The network had 100 nodes arranged in a 10-by-10 grid whose edges did not wrap around. The network was trained for up to 100 epochs, until more than 95% of the features were mapped to the same node as in the last epoch. The learning rate was initially set to 0.2 and exponentially decreased towards 0.001 in epoch 100. The multiple-winners neighbourhood was all nodes at a Manhattan distance of n or less from the winner, where n was initially set to 4 and linearly decreased towards 0 in epoch 100.

3.4 Model Training

Speech models are trained for a given language using Bodzár et al.'s [4] CUDA implementation of the Expectation-Maximization (EM) algorithm for GMMs. CUDA is a parallel computing platform developed by NVIDIA that allows for a dramatic reduction in computing time for parallelizable operations by taking advantage of concurrency of computation on a GPU. A well-known problem with the EM algorithm for mixture models occurs when one component converges to a single point, resulting in zero variance and infinite likelihood. This ultimately causes the algorithm to fail if not handled appropriately. Bodzár et al.'s implementation of the EM algorithm handles this issue by deleting components when their variance becomes zero. As a result, the algorithm itself has a tendency to determine the number of components a model trained on a particular dataset contains. Initially, models were trained on native speech data from each language corpus with a large number of components, allowing the algorithm's elimination of faulty components to determine a good value. All language models were then trained with the minimum number of components that resulted from initial training. This resulted in 8 Gaussian components being used when training the models for Swedish, German, French, Spanish, Polish and Turkish. The algorithm was run for 20 iterations on 1-2.5 million MFCC vectors randomly selected from the corpus for each language.

3.4.1 Visualization

The same SOM as used in section 3.3 was used to visualize 30 000 random points drawn from each of the GMM distributions for German and French, for comparison with the visualizations of the corpora. The result is shown in figure 2.

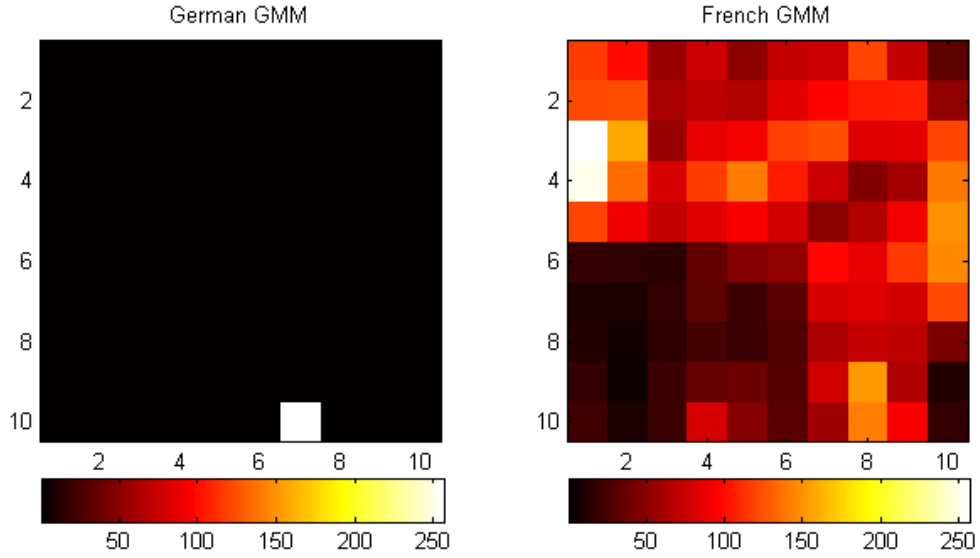

Fig. 2. Heatmaps of 30,000 random points drawn from each of the German (left) and French (right) GMMs, mapped to a 2D grid using the same SOM as in figure 1. Refer to figure 1 for interpretation. Unlike the plots in figure 1, these two plots use different scaling factors for the pixel values. In each of these plots, the pixel values have been scaled so that the highest value is 256, while in figure 1 the pixels of all plots were scaled by the same factor. This change in scaling was done because the pixel values for the German GMM would otherwise dominate the plots, making the French GMM plot render as entirely black.

3.5 Accent Classification

Given the GMM models trained using only native speech, MFCC data from the non-native speech corpora was classified as follows. For each recording of non-native speech, the log-likelihood of each MFCC point extracted given each model was computed. These points were then binned based on the model that gave the highest log-likelihood.

4 RESULTS

In Fig. 3, each group on the x-axis represents a model. The colors and positions within the groups of the bars represent different recordings of non-native speech. Each bar in each group represents how many MFCC points in that file received the highest likelihood of being produced by that language's native speech model.

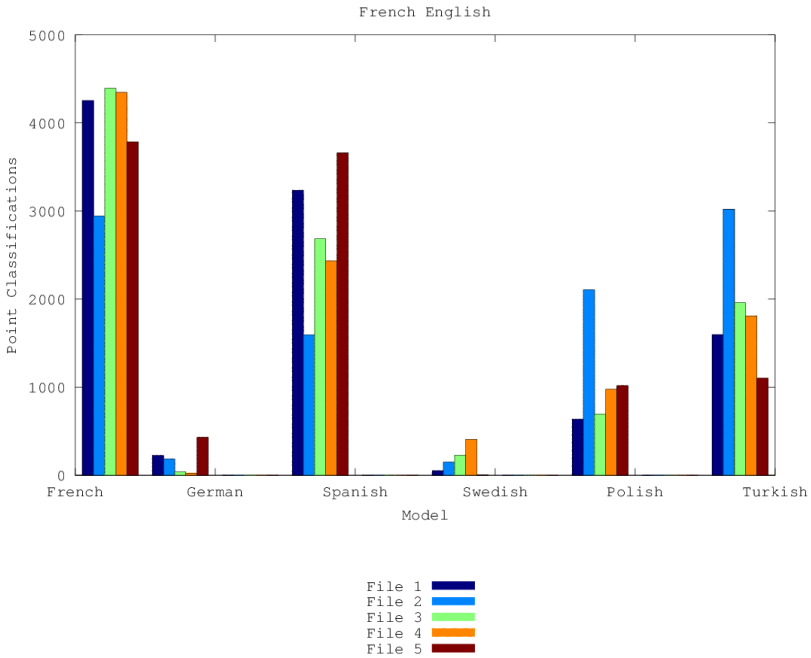

Fig. 3. Histogram for MFCC point classification of French English.

Fig. 4 shows pointwise likelihood classifications for 5 native Spanish speakers speaking English.

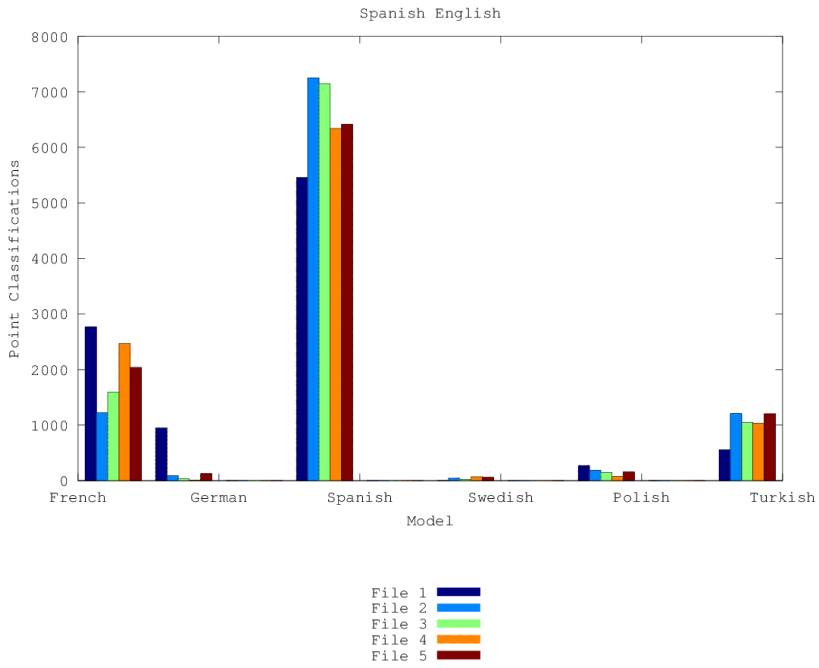

Fig. 4. Histogram for MFCC point classification of Spanish English.

Fig. 5 shows pointwise likelihood classifications for 5 native Polish speakers speaking English.

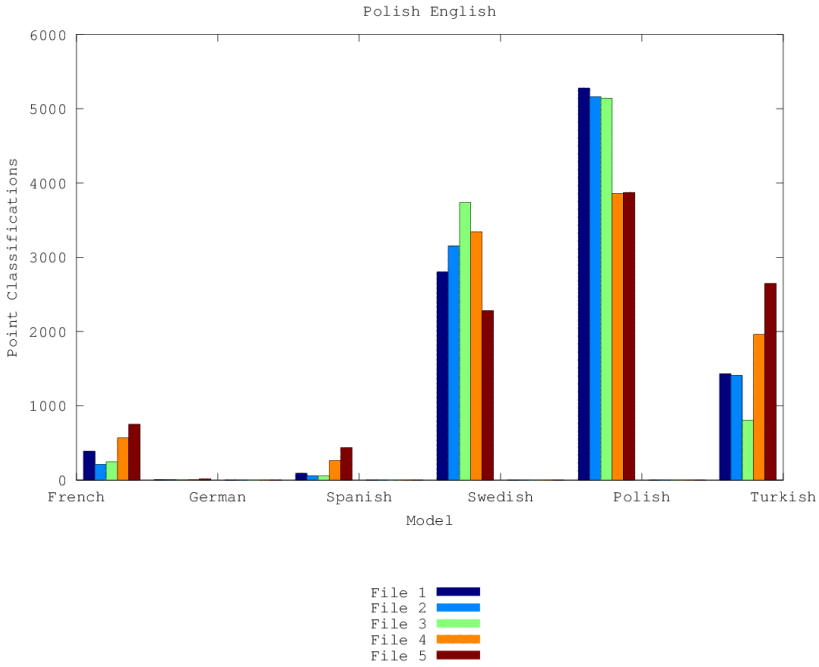

Fig. 5. Histogram for MFCC point classification of Polish English.

Fig. 6 shows pointwise likelihood classifications for 5 native Turkish speakers speaking English.

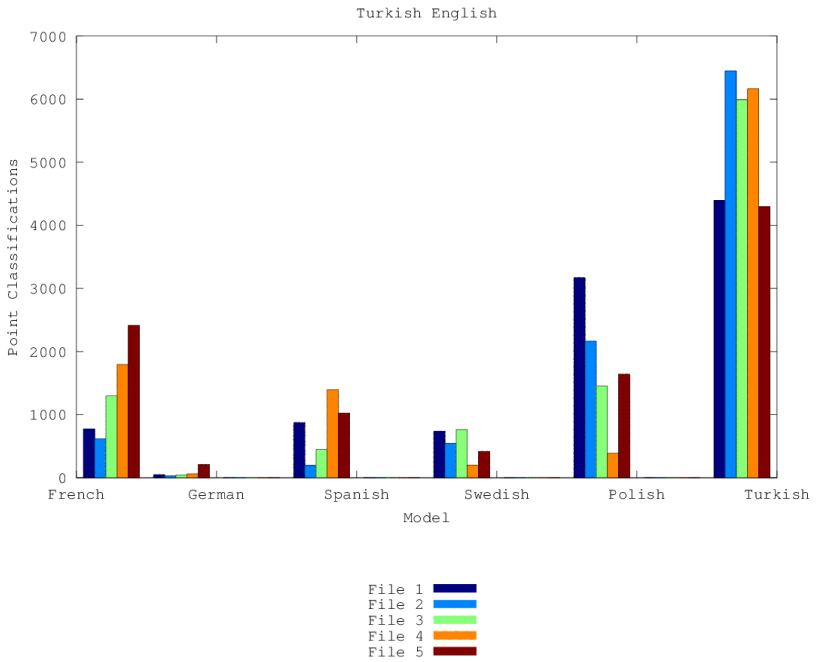

Fig. 6. Histogram for MFCC point classification of Turkish English.

Fig. 7 shows pointwise likelihood classifications for 5 native German speakers speaking English.

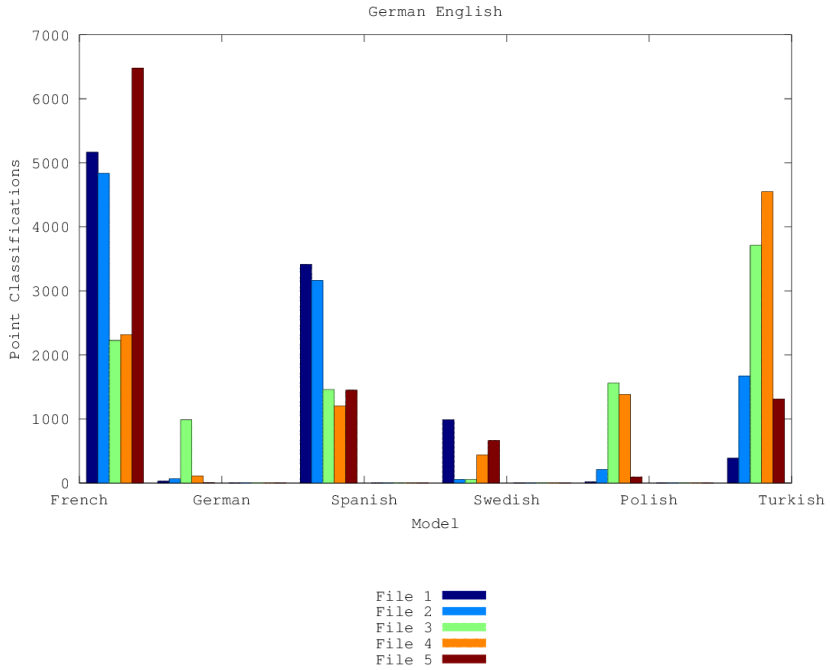

Fig. 7. Histogram for MFCC point classification of German English.

5 DISCUSSION

One should keep in mind that one of the assumptions in this project is that it is known in advance which language the foreign speakers are speaking, in this case English. This may not always be the case in applications, and the proposed method may perform worse if such a thing cannot be assumed. However, considering the way non-native speakers tend to replace sounds with ones from their native language, it might also achieve comparable performance even for accents of other languages.

Based on the results obtained, the potential of an accent identification system trained on native speech appears to be very promising. Excluding the German speakers, only 1 out of the 20 remaining non-native English speakers tested was misclassified using a pointwise classification method. One of the native French speakers was classified as a native Turkish speaker. This error could likely have been avoided if some degree of a confidence measure was applied to classification results, since, in this instance, the point total difference between the two closely-competing models was a meagre 1% of the total points in the recording.

The results of classifying English with German accent is peculiar. No language scores particularly well against the native German model, and German accented English is scored most highly with French. One, if not both, of the German corpora used - or at least the MFCC data that was extracted from them - appears to be defective in some way. This is also reflected in figures 1 and 2. The French corpus and GMM look similar under this map, but the German ones do not. The German GMM appears to have degenerated into a single mode with very small variance. As noted in section 3, the quality of the corpora was not assessed prior to the experiment.

Performance could perhaps be enhanced further by taking factors such as age and gender into account, since these are two of the most important factors in speaker variability [5]. They were not taken into account in this project since the goal was to determine whether the proposed method was feasible or not.

6 CONCLUSIONS

Non-native speech identification using GMMs trained only on native speech was fairly reliable for large speech samples of French, Spanish, Polish and Turkish. German speech of English, however, was more frequently classified as French than German. This may have been caused by one or both of the corpora being defective in some way. The proposed approach to identifying foreign accents does indeed seem feasible, but using a confidence-based classification strategy would probably be preferable to a hard strategy such as the raw argmax used here.

7 REFERENCES

[1] M. Sumner, “The role of variation in the perception of accented speech,” *Cognition*, vol. 119, no. 1, pp. 131–136, Apr. 2011.

[2] M. Benzeghiba, R. De Mori, O. Deroo, S. Dupont, T. Erbes, D. Jouvet, L. Fissore, P. Laface, A. Mertins, C. Ris, R. Rose, V. Tyagi, and C. Wellekens, “Automatic speech recognition and speech variability: A review,” *Speech Communication*, vol. 49, no. 10–11, pp. 763–786, Oct. 2007.

[3] “HTK Speech Recognition Toolkit,” 2013. [Online]. Available: http://htk.eng.cam.ac.uk/. [Accessed: 13-May-2013].

[4] E. Bodzár, B. Daróczy, I. Petrás, and A. A. Benczúr, “GMM Based Fisher Vector Calculation on GPGPU,” Budapest, Hungary, 2012.

[5] T. Chen, C. Huang, E. Chang, and J. Wang, “Automatic Accent Identification Using Gaussian Mixture Models,” in *IEEE Workshop on Automatic Speech Recognition and Understanding*, 2001, pp. 343–346.

[6] M. Alsulaiman, B. Mohamed, G. Muhammad, Z. Ali, M. Al-Gabri, G. H. Al Shatter, and S. A. Al-Kahtani, “Automatic Identification of Arabic L2 Learners Origin,” in *IS ADEPT*, 2012, pp. 107–112.

[7] A. Hanani, M. J. Russell, and M. J. Carey, “Human and computer recognition of regional accents and ethnic groups from British English speech,” *Computer Speech & Language*, vol. 27, no. 1, pp. 59–74, Jan. 2013.

[8] H. Strik and C. Cucchiarini, “Modeling pronunciation variation for ASR: A survey of the literature,” *Speech Communication*, vol. 29, no. 2–4, pp. 225–246, Nov. 1999.

[9] H. Wei, Y. Pu, and J. Yang, “Non-native Speech Recognition Based on Speaker Adaptation,” in *ICNC*, 2010, pp. 2024–2027.

[10] P. Nguyen, D. Tran, X. Huang, and D. Sharma, “Australian Accent-Based Speaker Classification,” *Knowledge Discovery and Data Mining, 2010. WKDD ’10. Third International Conference on*. pp. 416–419, 2010.

[11] “Backbone.” [Online]. Available: http://webapps.ael.uni-tuebingen.de/backbone-search/faces/initialize.jsp;jsessionid=3072B53B47649B46E747D62C1923F96D. [Accessed: 10-May-2013].

[12] “SweDia,” 2000. [Online]. Available: http://swedia.ling.gu.se/snabbmeny.html. [Accessed: 10-May-2013].

[13] S. Marsland, *Machine Learning: An Algorithmic Perspective*. CRC Press, 2009, p. 406.