Project Report :CSE666 (2013)

Study of Word-Level Accent Classification and Gender Factors

Xing Wang, Peihong Guo, Tian Lan, Guoyu Fu,
{wangxing.pku, peihongguo, welkinlan, fgy108}@gmail.com

Department of Computer Science and Engineering, Texas A&M University

Abstract

In this work, we conduct word-level accent classification. Different features, words, and learning methods are explored for accent classification, and results show that HMM-MFCC models show promising performance. Besides, we also explore the effect of gender on accent classification. Results show that models trained on male data do not generalized well on female data; Models trained on both male and female data is not always better than models trained on female data only. At last, we propose to use stacked ensemble classifier to classify gender firstly and then classify accent to improve accuracy.

Keywords: Speech Processing; Accent; HMM; GMM; MFCC; Formants

1 Introduction

Speaker and speech recognition is faced with bottleneck caused by speaker variability, in which accent is one of the most important factors [1] [2]. Accent is the way that particular person or group of people sound [1]. Identifying accent prior to automatic speech/speaker recognition produces a reduced search space and lower language modeling perplexity [3].

In our work, we extracted MFCC and formant features from words. Models are built using time-series HMM method and non time-series GMM method. Results show that HMM with MFCC features is the optimal combination. The most discriminant words are in accordance with human perception. Besides, we also tested how gender affects the accent recognition. Models built on one gender do not generalize well on samples of the other gender. We verified that a two-layer classifier by classifying the gender firstly can help improve the accuracy of the whole system.

The remainder of this paper is structured as follows. Section 2 provides an overview of previous works on accent classification. Section 3 describes our experiment design. Section 4 presents results and analysis. The article concludes with a discussion of results and future directions.

[1] http://dialectblog.com/2011/01/28/dialectvsaccent/

2 Related Work

Accent classification has been studied on serval levels: phoneme level [4] , word-level [2] [5] , and whole audio level [6]. Features extracted from the whole audio level contain too much noise. And the difference between classes is not strong enough to build accent classifier. Based on the description of India accented English [2], we find that the essence of accent is miss-pronouncing one phoneme as another. For example, Indians prefer to pronounce /au/ as /a:/, and voiceless plosives /p/ are always unaspirated. There are few studies working on word-level accent recognition. Tang et. al [5] extracted prosody features. Levent et. al [2] extracted formants features on word level. We extended Levent's work to try MFCC features and test gender's effect on the system.

After we decided the granularity of the audio excerpt, the next issue concerned is the feature used to describe every audio frame. Prior researches constructed accent recognition system with spectrum information as their features. One of the most frequently used features is the Mel-frequency Cepstral Coefficients (MFCCs) [2][7][8][9]. Another kind of spectral features is formant structure [2][6], especially the first and second formant frequencies (as Arslan and Hansen suggested in [10]). Liu et al. also stated that if ignoring the fundamental frequency F0, the accent classification accuracy receives a slight degradation [11]. In our work, comparison of these three kinds of features is performed.

In addition to feature extraction, another popular topic is choice of the learning models. Hidden Markov Model (HMM) is often utilized to model and classify temporal patterns [2][5]. HMM recognizer has an advantage in capturing the variation of spectral features. When the state number of HMM is reduced to one, the Gaussian Mixture Model (GMM) will serve as the actual classifier. The strength of GMM over HMM lies in that GMM is less time-consuming and that GMM does not needs transcripts [7]. We will make a comparison of these two classifiers in this work.

3 Experiment Design

3.1 Framework

The GMU Speech Accent Archive [6] is selected as data set. The speech was sampled and digitized. Then the audio is align with word transcript. We then used mel-frequency cepstrum and formant information to serve as feature vectors of each word. On the classifier level, we utilized GMM an HMM for comparing their performances. The framework of our system is shown in figure 1.

The optimal feature and classifier were noted from experiments on American and Indian speakers. Then the optimal combination of feature and classifier was implemented for experiments of the gender factor.

3.2 Data Preparation

In the corpus, speeches were collected in a quiet setting. It records speeches of one same paragraph that contains most of consonants and vowels. Basic information of every speaker, like gender, age,

[2] http://en.wikipedia.org/wiki/Indian_English

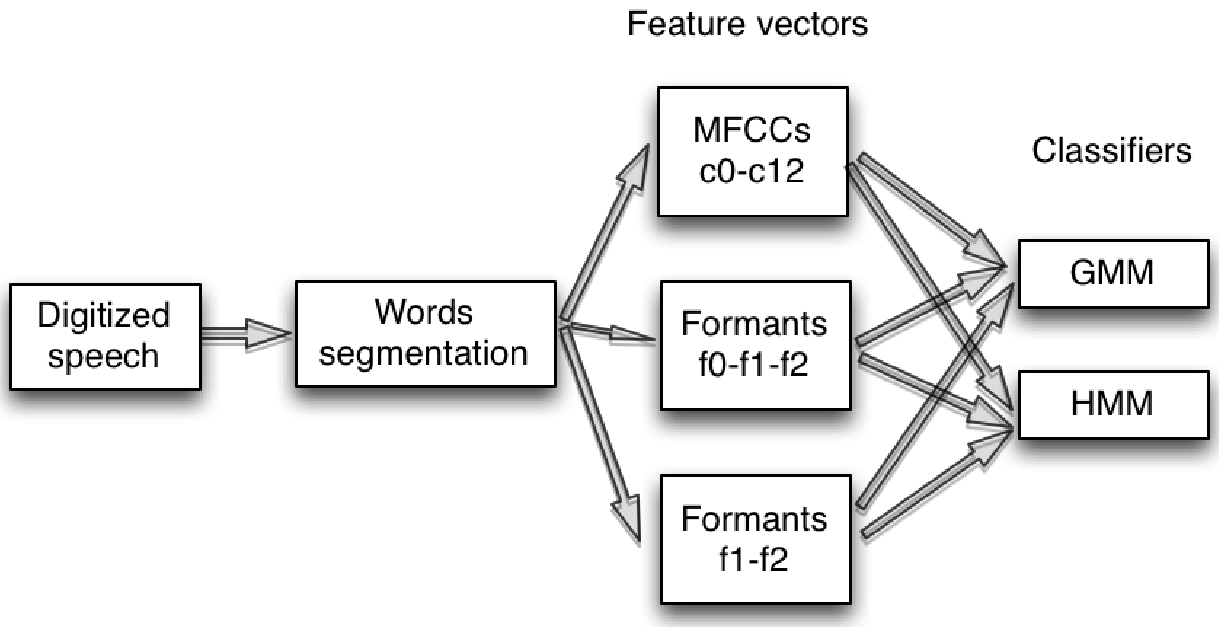


Fig. 1: System overview

	Male	Female
American	141	169
India	38	25

Table 1: Accent and gender distribution

etc., is provided and crawled from the website. The American-Indian data set we used is generated by restricting the nationality and the native language. The accent and gender distribution of the data set is shown in Table 1, and this dataset is unbalanced.

3.3 Word Alignment

Since the transcript of each speakerś speech is available, we utilized the tool [12] to align it with the speech and segment the speech word by word, after digitizing the speech. The output format of the toolkit is in Textgrid format, which can be processed by python script and can also be loaded in Praat as shown in Fig 2.

3.4 Feature Extraction

HTK MFCC package [3] and colea package [4] are used to extract MFCC and formant features. For each word, we have three types of features: MFCCs (trough c0 to c12), formant f0-f1-f2 and formant f1-f2

[3] http://www.mathworks.com/matlabcentral/fileexchange/32849-htk-mfcc-matlab

[4] http://www.mathworks.com/matlabcentral/fileexchange/108-colea

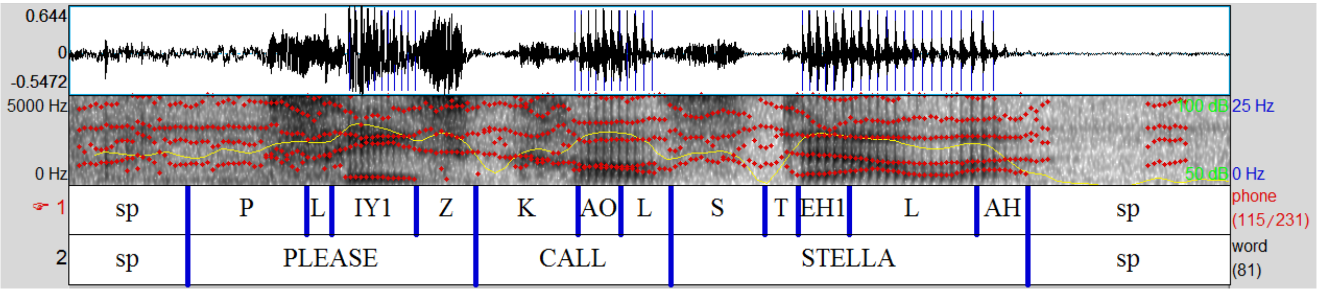


Fig. 2: Alignment Results

3.5 Machine Learning Method

GMM and HMM are adopted to train our classifiers. We conducted experiments 5 times and the average accuracy is reported.

The parameters for the model are obtained through 5-fold cross validation on the training data set. In each validation experiment, every combination of different parameters was implemented, and the combination of parameters that produced the maximal accuracy was selected as the one to train the classifier.

During the EM training process, the number of GMM components for each class is viewed as a prior. However, this parameter M may vary from circumstance to circumstance. In [6], 12 and 13 were used to model American accented speech and Indian accented speech, respectively. In [8], M=32 was set. Considering our classification is based on word-level, it is necessary to lower the range of this number. In our experiment, this range is set as $M \in \{1, 2, 3\}$ for both classes.

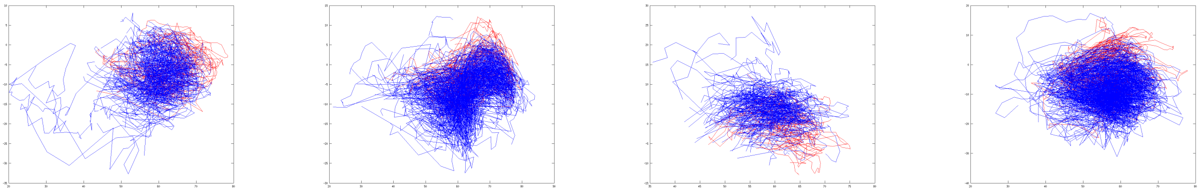
The HMM classifier we used is based on a continuous HMM with single Gaussian component for each hidden state. The observation sequences are the set of feature vectors extracted from each test word, where each feature vector is a single observation. The models are trained using EM algorithm with random restart. The number of hidden states used in the model are chosen from [5,9,13] which is determined by cross validation. Generally, we expect more accurate results with more hidden states, as well as longer training time.

4 Results and analysis

In this section, we will report our results on features comparison, machine learning methods comparison, words comparison and check the effect of gender on accent classification.

4.1 Features Comparison

The comparison of features is shown in Fig 4. The HMM was fixed as the classifier. MFCCs, F0F1F2 and F1F2 are extracted from each frame of every word in these three experiments, respectively. For systems of each pair, the training-test process was repeated 5 times. The average of these 5 accuracies was collected and drawn in Fig 4 in Appendix. As mentioned above, before training process, 5-fold cross validation was performed to select the optimal parameters for HMM.



(a) CAN (b) ASK (c) WILL (d) SCOOP

Fig. 3: Trajectory of discriminant words in C1-C2 of MFCC, blue lines represent American while red lines represent Indian

In this figure, we can find that: the MFCCs-HMM recognition system receives a highest accuracy, which states that MFCCs are more adequate features than formants; accuracy with fundamental frequency F0 is slightly better than that without F0; for classification consistency across words, accuracy of formants varies in a range of 15-20 percents from word to word, while this range of MFCCs based model is less than 10 percents. The feature of MFCCs contributes to such a high accuracy that we will consider it as the optimal feature extraction approach in the following experiments.

4.2 Machine Learning Methods Comparison

The comparison of temporal and non-temporal machine learning methods is shown in Fig 5 in Appendix. MFCC is fixed as the approach to extract features from frames of all words. After the same process with the previous experiment, the accuracies of GMM and HMM of 5 experiments were obtained. In this figure, both the accuracy and its stability across words of GMM are lower than those of HMM, for most of words; performance of GMM are worse than HMM, which proves that even on the word level, better results can be achieved if capturing the temporal variation of frames.

4.3 Words Comparison

The accuracy of each pair of recognition system varies from word to word. Experiments were carried out to discover that relation between performance of classification system and the words' characteristics. The result is shown in Fig 6 in Appendix. For word 'OF', most Indians do not differentiate between /v/ and /w/. For word 'TO', the /t/ sounds like /d/ by Indians. For word 'BRING' and 'BROTHER', the /r/ is a rolling r. For other word will good discriminant ability, we can find that C1-C2 for these words occupy different space as shown in Fig 3 .

4.4 Gender Effects

The comparison of machine learning methods is shown in Fig 7 in Appendix. Firstly, we can find that models trained using male are not as good as models trained using female data, and less stable. Secondly,if we use both male and female training data to train the model, we do not get consistent better result than models training using female data only. Finally, if we use a two layer classifier to classify the gender first and classify accent using model trained using data of the same gender, we can get better results most of the time.

5 Conclusion

In this work, we conduct word-level accent classification. Results show that HMM-MFCC models show promising performance. Model trained on one gender do not generalized well on the other gender. Stacked ensemble classifier can help alleviate this problem by classifying gender firstly and then accent.

6 Acknowledgement

The authors would like to appreciate Dr. Gutierrez-Osuna, Mr. SandeshAryal and Mr. Joseph Sung Lee for their valuable suggestions and continuous helps.

References

[1] Chao Huang, Tao Chen, Stan Z. Li, Eric Chang, and Jian-Lai Zhou, "Analysis of speaker variability.," in *INTERSPEECH*, Paul Dalsgaard, Brge Lindberg, Henrik Benner, and Zheng-Hua Tan, Eds. 2001, pp. 1377–1380, ISCA.

[2] Levent M. Arslan, John H. L. Hansen, Prof John, and H. L. Hansen, "Language accent classification in american english," 1996.

[3] Fadi Biadsy, Hagen Soltau, Lidia Mangu, and Jiri Navratil Julia, "Discriminative phonotactics for dialect recognition using context-dependent phone classifiers," 2010.

[4] F. William, A. Sangwan, and J.H.L. Hansen, "Automatic accent assessment using phonetic mismatch and human perception," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 21, no. 9, pp. 1818–1829, 2013.

[5] Hong Tang and Ali A. Ghorbani, "Accent classification using support vector machine and hidden markov model," in *Proceedings of the 16th Canadian Society for Computational Studies of Intelligence Conference on Advances in Artificial Intelligence*, Berlin, Heidelberg, 2003, AI'03, pp. 629–631, Springer-Verlag.

[6] S. Deshpande, S. Chikkerur, and V. Govindaraju, "Accent classification in speech," in *Automatic Identification Advanced Technologies, 2005. Fourth IEEE Workshop on*, 2005, pp. 139–143.

[7] Tao Chen, Chao Huang, E. Chang, and Jingchun Wang, "Automatic accent identification using gaussian mixture models," in *Automatic Speech Recognition and Understanding, 2001. ASRU '01. IEEE Workshop on*, 2001, pp. 343–346.

[8] Phuoc Nguyen, D. Tran, Xu Huang, and D. Sharma, "Automatic classification of speaker characteristics," in *Communications and Electronics (ICCE), 2010 Third International Conference on*, 2010, pp. 147–152.

[9] M.A. Zissman, "Comparison of four approaches to automatic language identification of telephone speech," *Speech and Audio Processing, IEEE Transactions on*, vol. 4, no. 1, pp. 31–, 1996.

[10] L.M. Arslan and J.H.L. Hansen, "Frequency characteristics of foreign accented speech," in *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on*, 1997, vol. 2, pp. 1123–1126 vol.2.

[11] Liu Wai Kat and P. Fung, "Fast accent identification and accented speech recognition," in *Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on*, 1999, vol. 1, pp. 221–224 vol.1.

[12] Jiahong Yuan and Mark Liberman, "Speaker identification on the scotus corpus," in *In Proceedings of Acoustics 2008*, 2008.

7 Appendix

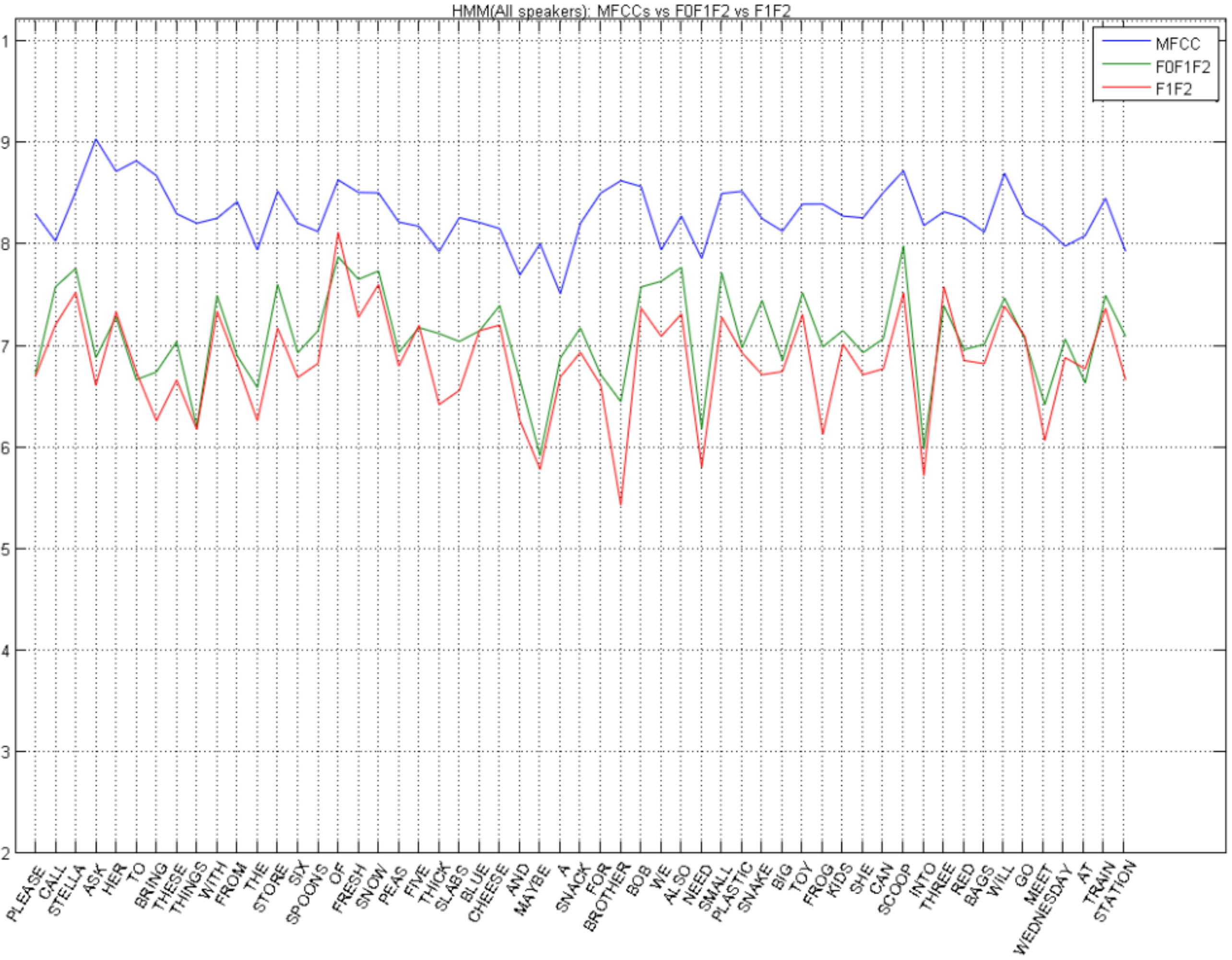


Fig. 4: Comparison of features using HMM training method

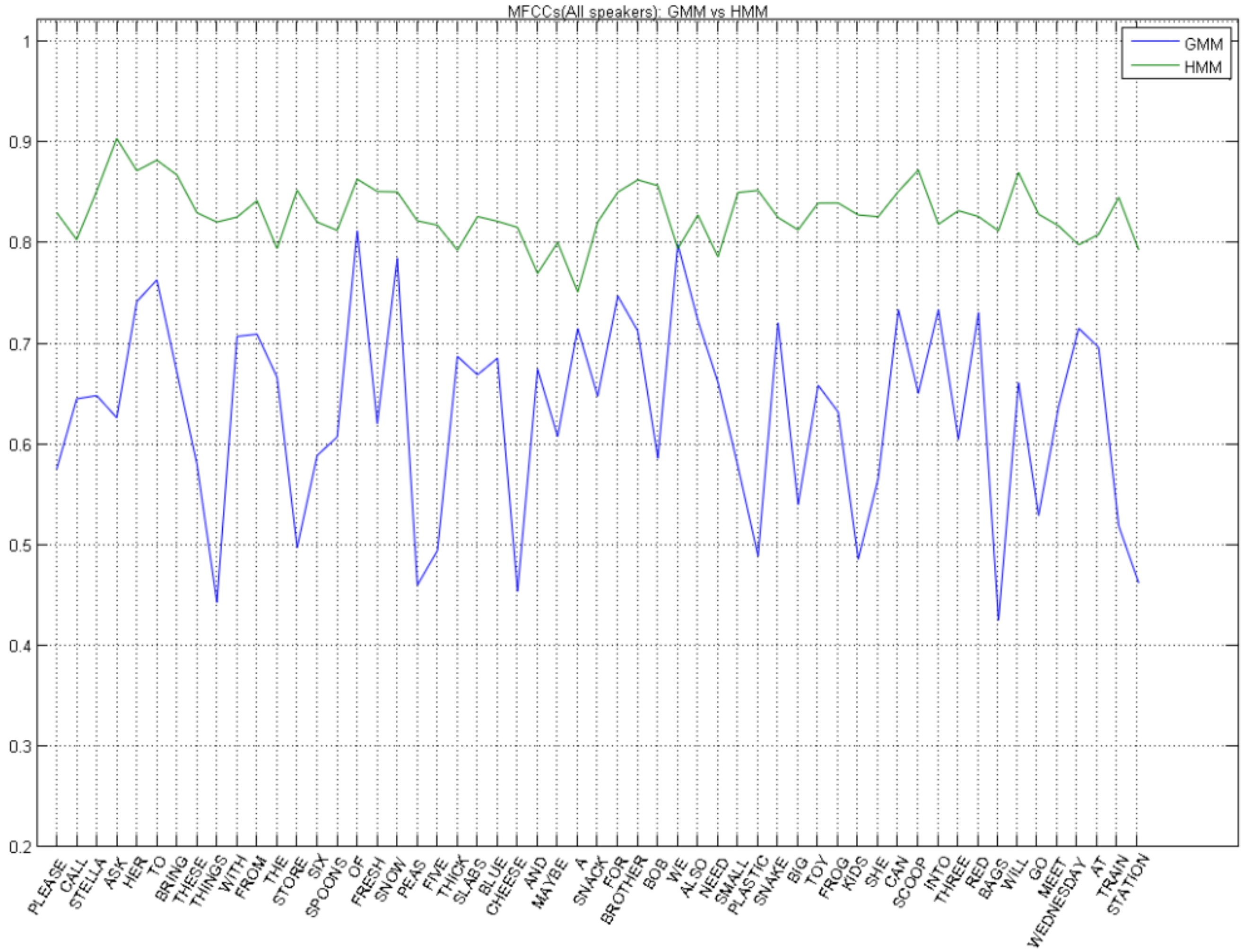


Fig. 5: Comparison of machine learning methods using MFCC features

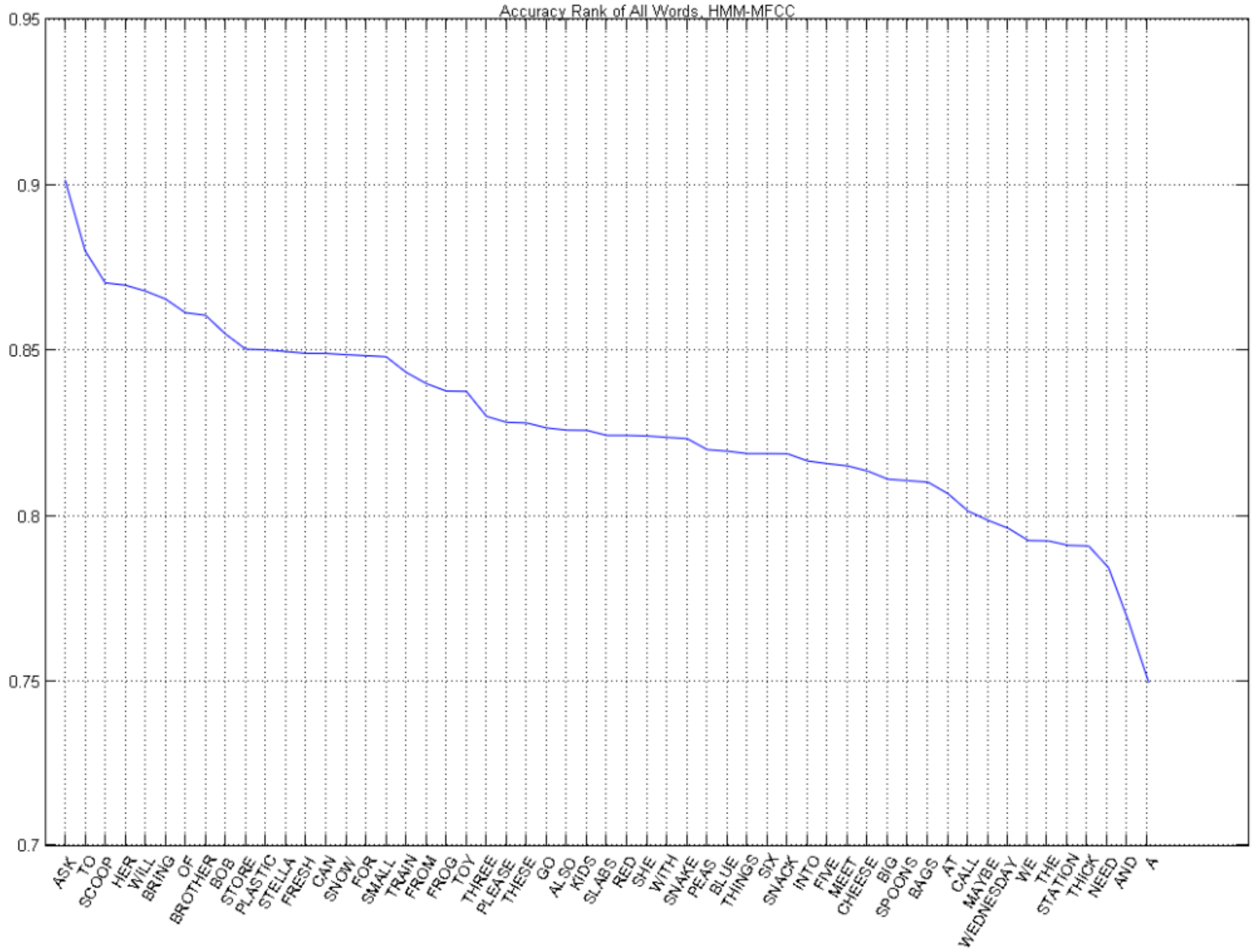


Fig. 6: Rank of accuracy across words for HMM-MFCC

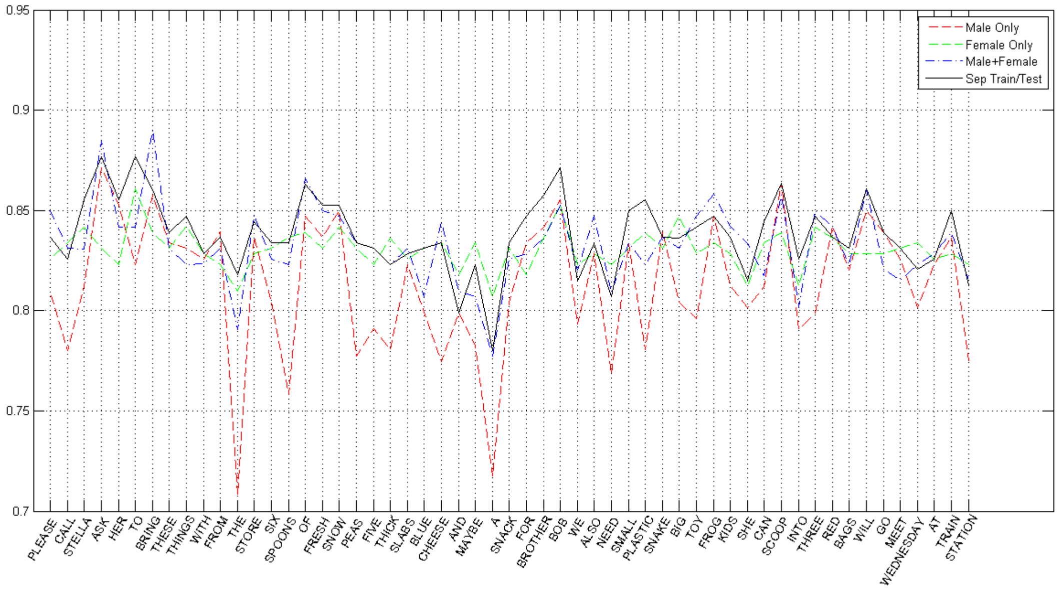


Fig. 7: Gender effects on classification.