Automatic classification of eating conditions from speech using acoustic feature selection and a set of hierarchical support vector machine classifiers

Abhay Prasad, Prasanta Kumar Ghosh

Electrical Engineering, Indian Institute of Science (IISc), Bangalore-560012, India
abhayprasad.337@gmail.com, prasantg@ee.iisc.ernet.in

Abstract

The problem of automatic classification of seven types of eating conditions from speech is considered. Based on the confusion among different eating conditions from a seven class support vector machine (SVM) classifier, a hierarchical SVM classifier is designed. Experiments on the iHEARu-EAT database show that the hierarchical classifier results in a better classification accuracy compared to a seven class classifier. We also perform a feature selection for each of the classifiers in the hierarchical approach. This further improves the unweighted average recall (UAR) to 73.7% compared to an UAR of 60.9% obtained from the baseline scheme of a direct seven-way classification.

Index Terms: Eating conditions classification, hierarchical classifier, feature selection

1. Introduction

Speech signal carries the linguistic information or the message from the speaker to the listener. Speech also carries paralinguistic information including affect, emotion, and personality of a speaker [1]. Recognition of eating conditions in speech is a challenging problem in computational paralinguistics [2]. In this work, we have addressed the task of classification of eating conditions from the information embedded in the speech of a speaker while eating. We recognize whether the subject is eating or not and if the subject is eating, the type of food the subject eats while speaking. Six types of foods are considered for the present work – Apple, Banana, Haribo, Biscuit, Crisp, and Nectarine.

Speech signal of a speaker undergoes several changes when simultaneously eating and speaking. Typically the speaker takes arbitrary pauses in order to chew food in the mouth while speaking, thus resulting in a number non-speech regions in the speech signal. This in turn could reduce the speaking rate and hinder the natural way of speaking of the subject. At the same time, when the speaker pauses to chew the food in the mouth, the non-stationary chewing sounds get embedded into the speech signal, thus, making the signal noisy. The chewing sound varies according to the nature of food being eaten. For example, the speech signal from a speaker while eating biscuit or crisp contains more crunchy sounds compared to when the speaker is eating other foods such as banana or haribo. Other factors affecting the speech signal while eating could be due to the changes in the shape of vocal tract oral cavity as it is filled with food resulting in a different articulation compared to in a no-food condition. The changes in articulation could also vary depending on the type of food being taken. This could in fact vary across different speakers due to different eating styles, which in turn may embed noise due to subject specific eating style in the speech signal. The nature of the food also plays a major role in determining the eating condition of a speaker from the speech.

In the literature, there are a few attempts for automatic classification of eating conditions from speech [3, 2]. Hantke et al [2] have proposed an automatic classification by using low-level acoustic features as well as high-level features related to intelligibility from an automatic speech recognizer. An audio-visual dataset is used for this classification. On the other hand, Schuller et al [3] have proposed a set of acoustic-only features with multi-class support vector machine (SVM) classifier for the eating conditions classification task, which is used as the baseline scheme for the work in this paper.

From the baseline scheme [3], we observe that the confusions between several food pairs are high including (Apple, Nectarine), (Banana, Nectarine), (Banana, Haribo) and (Biscuit, Crisp). Exploiting these confusion between the food types, we propose a hierarchical classification approach using SVM classifiers. The hierarchical classification scheme has been used in the past [4] for several problems including video genre categorization [5] and speech segmentation [6]. In this work, the hierarchical classification is found to perform better than a direct multi-class eating conditions classification. We also perform a forward block feature selection for each classifier in the hierarchy in order to select representative features for each classification task.

Our experiments with the iHEARu-EAT corpus [2] show that the unweighted average recall (UAR) obtained from the proposed hierarchical classification with acoustic feature selection is higher than the baseline scheme by 12.8% (absolute) on the training set in a cross-validation setup. Hierarchical approach also achieves a 1.4% (absolute) UAR improvement on the test set.

2. Dataset and experimental setup

For the experiments in this work, we have used the iHEARu-EAT corpus [2]. The corpus consists of 30 speakers speaking partially while eating six types of foods (Apple (Ap), Banana (Ba), Haribo (Ha), Biscuit (Bi), Crisp (Cr), Nectarine (Ne)) as well as in no-food (NF) condition (overall seven classes). The corpus has both training and test sets. Classification experiments are performed using a cross validation within the training set (train-cv) as well as on the test set as described in [3]. A set of 6373 features are computed using openSMILE [7, 8] using the steps outlined in [3]. SVM with sequential minimal optimization (SMO) training algorithm is used for classification using Weka 3 data mining toolkit [9]. Following the work in [3], the UAR is used as the performance metric for the classification.

3. Hierarchical SVM classifiers

The baseline scheme described in [3] involves running a seven-class SVM classifier. Table 1[1] shows the UAR obtained using the baseline scheme with different complexity parameter (C) values on the train-cv. The baseline scheme achieves the highest UAR of 60.9% for C=0.01 with the confusion matrix across seven classes as shown in Table 2. It is clear that some of the test categories are misclassified although majority of them are correctly classified (diagonal entries). Higher percentage of misclassification between two classes indicates that the seven-way classifier fails to discriminate the respective class pairs. This could be due to the fact that the decision boundaries between the confusing class pairs could not distinguish the classes well when a multi-class SVM is trained. This limitation of the seven-way classification could be overcome by developing a hierarchical classifier where in each level of the hierarchy the confusing class pairs could be grouped together to first perform a broad class classification followed by a finer classification within each broad level. Only 8.6% of the NF class were misclassified suggesting that the Food and NF categories are well-separated in the feature space. Hence, in the first level of the hierarchical classification, a Food vs NF binary classification is performed. Further, depending on the confusion between different food categories, the classifiers in the following level of the hierarchy are designed. In order to select the confusing class pairs, we take an average of two percentages from confusion matrix (Table 2) corresponding to the respective class pairs and consider all pairs for which the average percentage is more than 15% – Ap-Ne (23.76%), Ba-Ne (22.35%), Ba-Ha(16.62%), Bi-Cr(16.13%).

C	1	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-5}	10^{-6}
UAR	59.3	59.3	**60.9**	53.8	51.6	51.3	49.2

Table 1: UAR on train-cv using the baseline scheme with different C values.

	Ap	Ba	Bi	Cr	Ha	Ne	NF
Ap	**45.7**	3.6	9.3	6.4	5.7	25.7	3.6
Ba	7.1	**48.6**	2.1	0.7	16.4	22.1	2.9
Bi	11.3	0.8	**67.7**	16.5	0.7	2.3	0.8
Cr	4.3	0.7	15.7	**76.4**	0.0	2.9	0.0
Ha	6.7	16.8	0.8	1.6	**58.8**	12.6	2.5
Ne	21.8	22.6	5.3	7.5	5.3	**37.6**	0.0
NF	0.0	2.9	0.7	0.7	1.4	2.9	**91.4**

Table 2: Confusion matrix of the baseline scheme for $C = 10^{-2}$. Entries are in percentage. Each row sums up to 100% indicating the percentage of a test class getting classified to all seven classes.

We group the confusing class pairs to create broad food categories and perform classification among them. Considering the above mentioned top four most confusing class pairs, we design a three class classification in three different ways – 1) ApNe vs BaHa vs BiCr, 2) ApBa vs NeHa vs BiCr, 3) ApHa vs NeBa vs BiCr. In the following level, three binary classifiers are used for classification within respective broad food classes. For all the classifiers in the hierarchy, we use SVM. We also consider a hierarchy using a six-way food classification following Food vs NF binary classification. Thus, we consider four following hierarchical schemes:

1. HS1: Food vs NF — ApNe vs BaHa vs BiCr — Ap vs Ne and Ba vs Ha and Bi vs Cr
2. HS2: Food vs NF — ApBa vs NeHa vs BiCr — Ap vs Ba and Ne vs Ha and Bi vs Cr
3. HS3: Food vs NF — ApHa vs NeBa vs BiCr — Ap vs Ha and Ne vs Ba and Bi vs Cr
4. HS4: Food vs NF — Ap vs Ba vs Ha vs Bi vs Cr vs Ne

The UAR obtained on the train-cv set using these hierarchical schemes for different choices of C are shown in Table 3.

C	HS1	HS2	HS3	HS4
1	61.2	59.0	56.5	58.6
10^{-1}	61.2	59.0	56.5	58.6
10^{-2}	**62.5**	60.1	58.5	61.1
10^{-3}	55.7	52.8	53.5	53.3
10^{-4}	49.9	47.9	48.9	50.7
10^{-5}	50.1	48.6	49.0	50.5
10^{-6}	48.8	47.5	46.8	48.4

Table 3: Comparison of different hierarchical classifiers on the train-cv set in terms of UAR for different C

It is clear from Table 3 that HS1 (as shown in Fig. 1) performs the best (bold entry in Table 3) among the four hierarchical schemes considered. This is also better than the seven-way classification (baseline scheme) as described in [3]. The confusion matrix corresponding to this best hierarchical scheme is shown in Table 4. Comparing Table 2 and 4, it can be seen that for four classes the UAR improves because of the hierarchical approach. We observe an absolute UAR improvement for Ap (5.7%), Cr (5.0%), Ne (9.0%), and NF(3.6%). However, for the remaining classes the UAR decreases although, on average, the overall UAR improves by 1.6%.

	Ap	Ba	Bi	Cr	Ha	Ne	NF
Ap	**51.4**	3.6	6.4	5.0	3.6	26.4	3.6
Ba	7.1	**47.1**	1.4	0.7	15.7	18.6	9.3
Bi	5.04	0.7	**59.4**	19.5	0.00	3.8	1.5
Cr	5.0	0.0	10.7	**81.4**	0.0	2.1	0.7
Ha	10.1	20.2	0.0	0.8	**56.3**	9.2	3.4
Ne	25.6	15.0	0.7	5.3	5.3	**46.6**	1.5
No_Food	0.7	0.0	0.7	0.7	0.7	2.1	**95.0**

Table 4: Confusion matrix of the best hierarchical scheme (HS1).

It is important to note that there are five different classifiers in the proposed hierarchical approach. We hypothesize that an identical set of features is not equally discriminative for all classification tasks. Thus selection of a subset of features that best discriminates the classes in a classification task would be more appropriate. The acoustic feature selection is described in the next section.

[1] These UAR values are obtained by running the scripts provided with the Interspeech 2015 Computational Paralinguistics Challenge (ComParE) - Eating condition sub-challenge

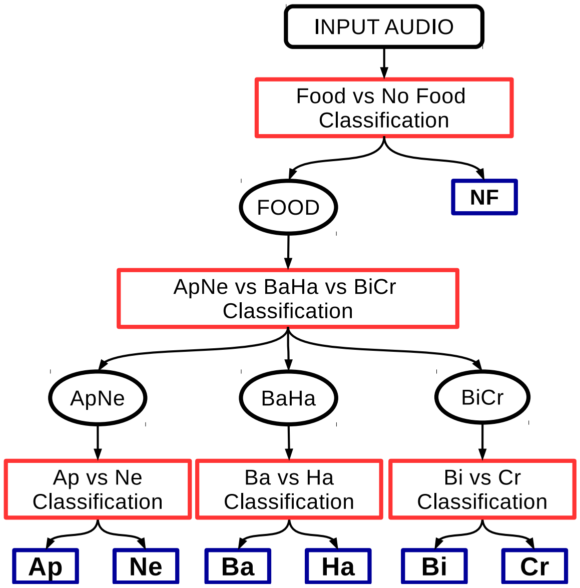

Figure 1: The best hierarchical scheme for automatic classification of eating conditions.

4. Acoustic feature selection

6373 features computed using the openSMILE contain a set of 131 feature categories and their various statistics including range, standard deviation, quartiles, skewness. We perform a forward feature selection in a batch mode in which a group of features of the same category is selected in each iteration of the feature selection. Feature selection is separately performed for each of the five classifications in the hierarchical approach. For four binary classifications (Food vs NF, Ap vs Ne, Ba vs Ha, and Bi vs Cr), F-score is used as the metric for the feature selection. This is because F-score combines precision and recall into a single measure. For the ApNe vs BaHa vs BiCr classification, the UAR is maximized in each iteration of the feature selection. Suppose $\mathbf{u}$ denotes the 6373-dimensional feature vector consisting of K feature categories $\mathbf{u}_k$, $1 \leq k \leq K$. Consider a classification task for which the class labels are given by the variable Ω. Then the steps of the acoustic feature selection are outlined in Algorithm 1, which returns the indices of the ranked feature categories (η) and the corresponding classification performance ($\mathcal{P}$).

Algorithm 1 Forward acoustic feature selection - inputs: $\mathbf{u} = [\mathbf{u}_1, \mathbf{u}_2, \cdots \mathbf{u}_K]$ and Ω

```
$\mathbf{u}^s = \emptyset$. $\mathcal{P}$, $\eta$ are initialized as empty vectors. $\mathcal{I} = \{1, 2, \cdots, K\}$.
for $l$=1 to $K$ do
    for $i \in \mathcal{I}$ do
        $\zeta_i \leftarrow$ classification performance (F-score or UAR) using $[\mathbf{u}^s \;\; \mathbf{u}_i]$ and $\Omega$.
    end for
    $\mathcal{P}_l \leftarrow \max_i \zeta_i$
    $\eta_l \leftarrow \arg\max_i \zeta_i$
    $\mathbf{u}^s \leftarrow [\mathbf{u}^s \;\; \mathbf{u}_{\eta_l}]$
    $\mathcal{I} \leftarrow \mathcal{I} \setminus \{\eta_l\}$
end for
Return $\mathcal{P}$ and $\eta$
```

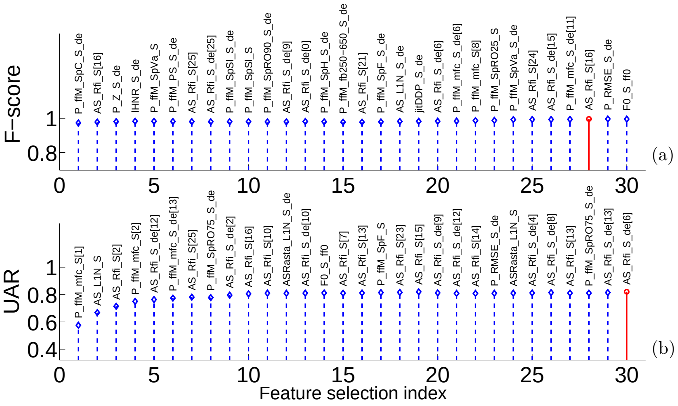

Figure 2: Feature groups from forward feature selection for (a) Food vs NF classification (b) ApNe vs BaHa vs BiCr classification. The graphs of the respective performance measure vs selected features are also shown along with.

Top 30 selected feature categories are shown in Fig. 2 (a) and (b) for the Food vs NF classification and ApNe vs BaHa vs BiCr classification respectively. The highest classification performances (Fscore=0.9963 and UAR=82.1%) are indicated by solid red lines and they occur when top 28 and 30 feature categories are considered for these two classifications respectively. Similarly the selected features for Ap vs Ne, Ba vs Ha, and Bi vs Cr classifications are shown in Fig. 3 (a), (b), and (c). The highest classification performances (Fscore=0.8027, 0.8530, and 0.9366) are attained when 30, 24, and 23 feature categories are used for these three binary classifications. The name of the selected feature categories are also indicated on top of the stem plot of the classification performance. For the lack of space, the feature names are abbreviated using the initial letter in each word of the original names [2]. For example, the original name 'pcm_fftMag_spectralCentroid_sma_de' is abbreviated as 'P_ffM_SpC_S_de'.

Overall, among the 131 feature categories, 91 feature categories appear in the top selected features in one or more classification tasks. It is observed that many selected features are common across these classification tasks while few others are found to be unique for each classification. For example, pcm_fftMag_spectralSlope_sma_de and pcm_fftMag_spectralHarmonicity_sma_de appear in the top selected feature categories for Food vs NF, Ap vs Ne, and Bi vs Cr classifications. Similarly pcm_fftMag_mfcc_sma[1] and F0final_sma_ff0 appear for ApNe vs BaHa vs BiCr, Ap vs Ne, and Ba vs Ha classifications. On the other hand, jitterLocal_sma_de appears in top selected feature only for Ap vs Ne; jitterDDP_sma_de appears only for Food vs NF classification. It is interesting to note that while jitter related features are selected from the forward feature selection, no shimmer related features appear in the top selected feature list. Similarly, no features related to voicingFinalUnclipped were selected. This could be because the presence of food in the mouth does not affect the glottal source signal significantly.

For Food vs NF, Ap vs Ne, and Bi vs Cr, the F-score using the topmost feature is more than 80% of the highest F-score obtained for the respective classification tasks. For ex-

[2] The names of the features are available in the arff files given along with the dataset for the Interspeech 2015 Computational Paralinguistics Challenge (ComParE) - Eating condition sub-challenge

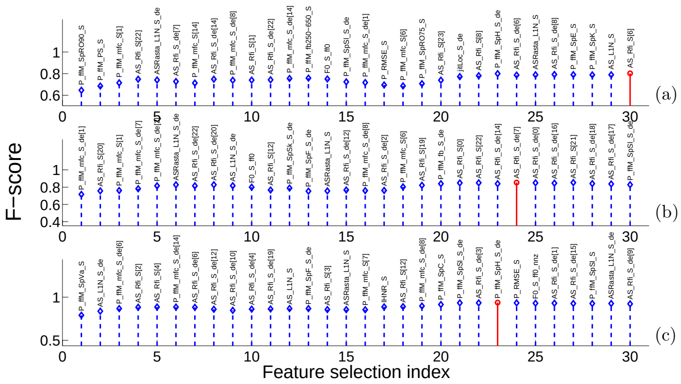

Figure 3: Feature groups from forward feature selection for (a) Ap vs Ne classification (b) Ba vs Ha classification and (c) Bi vs Cr classification. The graphs of the respective performance measure vs selected features are also shown along with.

ample, using only pcm_fftMag_spectralCentroid_sma_de, the Food vs NF classification task achieves an Fscore of 0.973. This appears to be an important feature for only Food vs NF classification. This could be because of the noise due to chewing food may cause changes in the spectral centroid in many frames compared to neighboring speech frames. pcm_fftMag_spectralRollOff90.0_sma turns out to be the first selected feature for Ap vs Ne classification. This could be due to different amount of spectral distortions caused by the chewing of apple and nectarine. pcm_fftMag_spectralVariance_sma is the topmost selected feature for the Bi vs Cr classification suggesting that chewing biscuit causes a different amount of spectral spread when compared to chewing crisp.

5. Results and discussions

We run the proposed hierarchical scheme based eating conditions classification with the best set of selected features on the train-cv. The UAR values obtained for different choices of C are shown in Table 5. The best UAR=73.7% is obtained for C=10 unlike the best UAR for C=0.01 when all features are used (Table 3). There is 11.2% absolute UAR improvement due to the feature selection when compared to the hierarchical classification with all features suggesting the benefit of feature selection. High value of C(=10) corresponding to the highest UAR could be due to the matched training and test acoustics in each fold of the train-cv. A higher value of C could exploit the acoustic conditions of the training set for evaluating on the test set. The confusion matrix corresponding to UAR=73.7% is shown in Table 6.

C	100	10	1	10^{-1}	10^{-2}	10^{-3}	10^{-4}
UAR	72.2	**73.7**	65.2	59.2	53.4	47.7	47.8

Table 5: UAR on train-cv using hierarchical with the best selected features with different C values.

Comparing Table 4 and 6, it is clear that the feature selection improves the class specific UAR for all seven classes. The absolute improvements in UAR are 17.2%, 20.0%, 21.8%, 7.87%, 2.5%, 7.5%, 1.4% for Ap, Ba, Bi, Cr, Ha, Ne, and NF respectively. From Table 2 and 6, it can be seen that the UAR values obtained using the proposed hierarchical classification with selected features are either equal (for Ha) or better than the baseline scheme indicating the effectiveness of the proposed approach.

	Ap	Ba	Bi	Cr	Ha	Ne	NF
Ap	**68.6**	6.4	2.1	2.9	6.4	13.6	0.0
Ba	8.6	**67.1**	1.4	0.7	11.4	10.0	0.7
Bi	6.0	2.3	**81.2**	7.5	2.3	0.8	0.0
Cr	0.7	2.1	4.3	**89.3**	2.9	0.7	0.0
Ha	8.4	11.8	3.4	1.7	**58.8**	16.0	0.0
Ne	22.6	9.0	0.8	1.5	12.0	**54.1**	0.0
NF	1.4	1.4	0.0	0.7	0.0	0.0	**96.4**

Table 6: Confusion matrix for train-cv using hierarchical approach with the best selected features and C=10.

The proposed hierarchical classification approach with all features results in a UAR of 67.33% and 65.79% for C=0.01 and C=10 respectively. While the former is better than the UAR (=65.9%) obtained by the baseline scheme on the test set [3], the later is not. This could be because of the mismatch between the training and test acoustics resulting in a smaller value of C preventing the classifier from over-fitting to the training data. This also demonstrates the effectiveness of the hierarchical approach on the test set. However, no improvement on the test set is observed when the selected features are used in the hierarchical classification in place of all features.

6. Conclusions

We propose a hierarchical classification approach for automatic classification of seven types of eating conditions by exploiting the confusion between different food types from a direct seven-way classification scheme. Using the iHEARu-EAT database, we show the merit of the hierarchical approach over a direct seven-way classification. We also show that the UAR of the eating conditions classification improves further by performing feature selection for each of the classifiers in the proposed hierarchical scheme. We find that the proposed approach achieves an absolute improvement of 12.8% UAR on the train-cv set and 1.4% on the testset over a direct seven-way classification. The classification performance could further be improved by automatically selecting speech segments from the test sentence that best discriminates the food types. This is part of our future work.

7. Acknowledgements

We thank Department of Science and Technology (DST), Govt. of India for the support in this work.

8. References

[1] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. MüLler, and S. Narayanan, "Paralinguistics in speech and language state-of-the-art and the challenge," *Computer Speech & Language*, vol. 27, no. 1, pp. 4–39, 2013.

[2] S. Hantke, F. Weninger, R. Kurle, A. Batliner, and B. Schuller, "I hear you eat and speak: automatic recognition of eating condition and food type," *ms. to appear*.

[3] B. Schuller, S. Steidl, A. Batliner, S. Hantke, F. Hönig, J. R. Orozco-Arroyave, E. Nöth, Y. Zhang, and F. Weninger, "The interspeech 2015 computational paralinguistics challenge: Nativeness, parkinson's & eating condition," *Proceedings INTERSPEECH*, 2015.

[4] N. Cesa-Bianchi, C. Gentile, and L. Zaniboni, "Hierarchical classification: combining Bayes with SVM," *23rd international conference on Machine learning*, pp. 177–184, 2006.

[5] X. Yuan, W. Lai, T. Mei, X.-S. Hua, X.-Q. Wu, and S. Wu, "Automatic video genre categorization using hierarchical SVM," *International Conference on Image Processing*, pp. 2905–2908, 2006.

[6] A. Juneja and C. Espy-Wilson, "Speech segmentation using probabilistic phonetic feature hierarchy and support vector machines," *International Joint Conference on Neural Networks*, vol. 1, pp. 675–679, 2003.

[7] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," *Proceedings of the international conference on Multimedia*, pp. 1459–1462, 2010.

[8] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in openSMILE, the Munich open-source multimedia feature extractor," pp. 835–838, 2013.

[9] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: an update," *ACM SIGKDD explorations newsletter*, vol. 11, no. 1, pp. 10–18, 2009.