# Cascaded Acoustic Group and Individual Feature Selection for Recognition of Food Likability

Dara Pir
*Information Technology Program, Guttman Community College, City University of New York, New York, U.S.A.*

Keywords: Food Likability, Acoustic Features, Group Feature Selection, Large Acoustic Feature Sets, Computational Paralinguistics.

Abstract: This paper presents the novel Cascaded acoustic Group and Individual Feature Selection (CGI-FS) method for automatic recognition of food likability rating addressed in the ICMI 2018 Eating Analysis and Tracking Challenge's Likability Sub-Challenge. Employing the speech and video recordings of the iHEARu-EAT database, the Likability Sub-Challenge attempts to recognize self-reported binary labels, 'Neutral' and 'Like', assigned by subjects to food they consumed while speaking. CGI-FS uses an audio approach and performs a sequence of two feature selection operations by considering the acoustic feature space first in groups and then individually. In CGI-FS, an acoustic group feature is defined as a collection of features generated by the application of a single statistical functional to a specified set of audio low-level descriptors. We investigate the performance of CGI-FS using four different classifiers and evaluate the relevance of group features to the task. All four CGI-FS system results outperform the Likability Sub-Challenge baseline on iHEARu-EAT development data with the best performance achieving a 9.8% relative Unweighted Average Recall improvement over it.



## 1 INTRODUCTION

Computational Paralinguistics (CP) tasks attempt to recognize the states and traits of speakers. Whereas Automatic Speech Recognition's goal is to predict which words are spoken, CP is concerned with the manner in which those words are spoken (Schuller and Batliner, 2014). The ICMI 2018 Eating Analysis and Tracking Challenge's Likability Sub-Challenge, a CP task, aims at recognizing self-reported binary rating labels, 'Neutral' and 'Like', assigned by subjects to the food type they consumed while speaking. Food likability is a new research domain with potential applications in many fields such as emotion recognition, product evaluation, and smart assistance.

The Sub-Challenge baseline results using the audio mode alone outperform those using the video and the audio-plus-video modes on development data. We therefore choose an audio approach that employs the acoustic Sub-Challenge baseline feature set extracted by the openSMILE toolkit (Eyben et al., 2013) from the audio-visual tracks of the iHEARu-EAT database (Hantke et al., 2018; Hantke et al., 2016; Schuller et al., 2015).

The openSMILE generated baseline acoustic feature sets have been used in the Interspeech CP tasks since their inception in 2009 (Schuller et al., 2009). The baseline acoustic feature set is generated by applying statistical functionals like the mean to low-level descriptors (LLDs) like the spectral energy (Schuller et al., 2009; Weninger et al., 2013; Eyben, 2016). The number of features in the baseline feature set has increased from 384 in 2009 to 6373 in 2013 (Schuller et al., 2013) and using larger feature sets has resulted in improved accuracy performances. Large feature sets, however, may degrade accuracy performances by inducing the *curse of dimensionality* problem. Performing dimensionality reduction may therefore prove helpful in addressing this problem.

In this paper, we present the Cascaded acoustic Group and Individual Feature Selection (CGI-FS) method for automatic recognition of food likability. CGI-FS performs a group feature selection operation, which is followed by an individual feature selection one. First, we consider the acoustic feature space in groups and select an optimum subset of group features. We define an acoustic group feature as a collection of features generated by the application of a single statistical functional to a specified set of audio LLDs (Schuller et al., 2007). Next, the selected features of the previous step are used in the individual feature selection operation. Group features parti-



Pir, D.
Cascaded Acoustic Group and Individual Feature Selection for Recognition of Food Likability.
DOI: 10.5220/0007683708810886
In *Proceedings of the 8th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2019)*, pages 881-886
ISBN: 978-989-758-351-3

tioned by either statistical functionals or LLDs have previously been used together with feature selection methods (Schuller et al., 2007; Pir and Brown, 2015; Pir, 2018). To further improve classification performance, CGI-FS applies individual feature selection to the already selected group feature(s). We investigate the performance of CGI-FS using four different classifiers. In addition, we evaluate the relevance of group features to the task of food likability recognition.

This paper is organized as follows. Section 2 provides details about feature selection as a dimensionality reduction method. Section 3 describes the two steps of the CGI-FS method. Section 4 provides information about the corpus and the classifiers used in the implementations of the CGI-FS method. Experimental results are shown and explained in Section 5. We conclude and mention future work in the last section.

# 2 FEATURE SELECTION

Feature selection methods achieve dimensionality reduction by selecting a subset of features from the original feature set deemed more relevant to the associated task (Dougherty, 2013). An aim of feature selection may be to reduce the dimensionality of the feature space to render the training phase of complex learning algorithms more tractable. Another aim may be to improve the classification performance by removing irrelevant and/or redundant features. Since feature selection does not generate new features, e.g., as combinations of existing features, it offers an advantage when interpretability of the relevance of the selected features to the associated task is of importance.

Filters and wrappers are the two main types of feature selection methods (Kohavi and John, 1997). To evaluate feature subsets, filters rely on the properties of the data alone whereas wrappers use a classifier's accuracy scores. The wrapper, therefore, tends to provide superior performances as it uses a classifier's biases in subset evaluation (Ng, 1998).

# 3 METHOD

This section describes the two steps of the CGI-FS method. First, we describe the group feature selection method, which operates on the entire Sub-Challenge baseline acoustic feature set generated by the openSMILE toolkit. Next, we provide details on the individual feature selection method that operates on the resultant feature subset obtained in the previous step. Four implementation of the CGI-FS method are made each using a different classifier.

## 3.1 Group Feature Selection

In the group feature selection method we consider the acoustic feature space in groups where each group consists of all the features generated by the application of a single functional to a specified set of set of LLDs. The number of classifier evaluation cycles is reduced from 6373 (number of baseline features) to 56 (number of baseline functional groups). After evaluating each functional group, they are ranked, from high to low, according to their evaluation scores. Since the highest performing group, for each of the four implementations, already outperforms the baseline, and in the interest of achieving substantial dimensionality reduction, we select only the top performing group and use it as input to the individual feature selection step that follows.

## 3.2 Individual Feature Selection

The individual feature selection is performed only on the top ranking group feature. We use a wrapper-based Rank Search (RS) algorithm (Gutlein et al., 2009), which is a two-phase process. In the first phase, we rank feature subsets according to their subset evaluation scores, from high to low. Next, we use the RS algorithm to obtain the feature subset that achieves the highest evaluation score.

# 4 CORPUS AND CLASSIFIERS

## 4.1 Corpus

The audio-visual tracks of the iHEARu-EAT database were made by asking subjects to consume one of six food types (Apple, Nectarine, Banana, Gummi bear, Biscuit, and Crisps) or no food while speaking. Recordings were made from subjects' readings of phonetically balanced text as well as from their comments to various prompts.

At the end of the recordings, the subjects rated how much they liked each food type they had consumed by setting a continuous slider's position to a value between 0 and 1, associated with the cases of extreme dislike and extreme like, respectively. The chosen values were then mapped to binary labels of 'Neutral' and 'Like' based on the distribution of the ratings. Since the subjects did not consume the food type they disliked, a 'Dislike' label was not necessary. In addition, the 'Neutral' label was set for cases where

Table 1: Top 5 and bottom 5 group features ranked by group feature selection for the RF classifier. The ranking is based on UAR scores, from high to low. R: Rank of the group in the list of 56 ranked group features. Group: Name of the group feature. UAR: UAR evaluation score in %. S: Size represented by the number of features in the group. The name and UAR entries for the top group, which is chosen for the individual feature selection, are shown in bold.

| R | Group | UAR | S |
|---|---|---|---|
| 1 | ***amean*** | **70.5** | 71 |
| 2 | *meanFallingSlope* | 70.3 | 118 |
| 3 | *rqmean* | 70.1 | 130 |
| 4 | *meanRisingSlope* | 69.8 | 118 |
| 5 | *percentile1.0* | 69.8 | 130 |
| ⋮ | ⋮ | ⋮ | ⋮ |
| 52 | *meanSegLen* | 56.9 | 119 |
| 53 | *maxPos* | 56.1 | 130 |
| 54 | *upleveltime90* | 55 | 130 |
| 55 | *minSegLen* | 54 | 119 |
| 56 | *nnz* | 48 | 1 |

Table 2: Results for the SGD classifier.

| R | Group | UAR | S |
|---|---|---|---|
| 1 | ***amean*** | **68.1** | 71 |
| 2 | *posamean* | 67.5 | 71 |
| 3 | *stddev* | 67.2 | 130 |
| 4 | *rqmean* | 67 | 130 |
| 5 | *percentile99.0* | 66.6 | 130 |
| ⋮ | ⋮ | ⋮ | ⋮ |
| 52 | *qregc1* | 53.5 | 71 |
| 53 | *minRangeRel* | 53 | 118 |
| 54 | *maxSegLen* | 52.4 | 119 |
| 55 | *minSegLen* | 51.4 | 119 |
| 56 | *nnz* | 50 | 1 |

the subjects were not eating while speaking. Further detail about the corpus can be found in (Hantke et al., 2018) and (Hantke et al., 2016).

## 4.2 Classifiers

Each of the four CGI-FS systems presented in this paper uses one of WEKA toolkit's classifier implementations: RandomForest (RF) (Breiman, 2001), Stochastic Gradient Descent (SGD), VotedPerceptron (VP) (Freund and Schapire, 1999), and SimpleLogistic (SL) (Sumner et al., 2005). Our preprocessing step standardizes all features to zero mean and unit variance prior to classification. The training is performed with 10-fold cross-validation in all cases.

Table 3: Results for the VP classifier.

| R | Group | UAR | S |
|---|---|---|---|
| 1 | ***amean*** | **68** | 71 |
| 2 | *quartile3* | 66.9 | 130 |
| 3 | *posamean* | 66.7 | 71 |
| 4 | *percentile99.0* | 65.7 | 130 |
| 5 | *stddev* | 65.4 | 130 |
| ⋮ | ⋮ | ⋮ | ⋮ |
| 52 | *upleveltime90* | 53.1 | 130 |
| 53 | *peakRangeRel* | 52.8 | 118 |
| 54 | *minRangeRel* | 52.7 | 118 |
| 55 | *minSegLen* | 51.8 | 119 |
| 56 | *minPos* | 51.6 | 130 |

Table 4: Results for the SL classifier.

| R | Group | UAR | S |
|---|---|---|---|
| 1 | ***flatness*** | **68.3** | 130 |
| 2 | *amean* | 67.7 | 71 |
| 3 | *rqmean* | 67.7 | 130 |
| 4 | *iqr1-3* | 67.4 | 130 |
| 5 | *posamean* | 67 | 71 |
| ⋮ | ⋮ | ⋮ | ⋮ |
| 52 | *upleveltime90* | 54.5 | 130 |
| 53 | *minRangeRel* | 54.1 | 118 |
| 54 | *peakRangeRel* | 53.7 | 118 |
| 55 | *nnz* | 53.4 | 1 |
| 56 | *minSegLen* | 53.1 | 119 |

# 5 EXPERIMENTAL RESULTS

The experimental results obtained by all four systems, i.e., implementations using the RF, SGD, VG, and SL classifiers, are shown for each of the two steps of the CGI-FS method in this section. We then briefly describe previous work in the Likability Sub-Challenge.

## 5.1 Group Feature Selection Results

Table 1 shows the top five and bottom five group features ranked, from high to low, according to their Unweighted Average Recall (UAR) evaluation scores for the CGI-FS implementation using the RF classifier. The *amean* group feature, which includes 71 features, achieves the highest performance with an evaluation score of 70.5% UAR. The *amean* functional is defined as the arithmetic mean of the underlying contour (Eyben, 2016; Eyben et al., 2013). Similarly, Tables 2, 3, and 4 display results obtained by the CGI-FS systems implemented using the SGD, VP, and SL classifiers, respectively.

We note that for three out of four systems the *amean* group achieves the highest performance while attaining the second best score in the system that uses

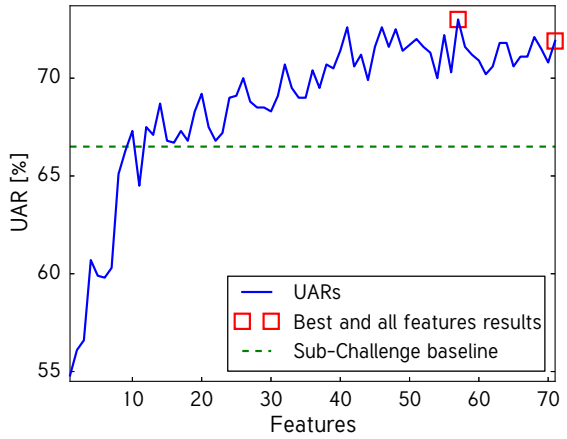


Figure 1: UAR evaluation scores of the individual feature selection step of the CGI-FS system using the RF classifier. The best UAR score of 73.0% is achieved with a feature subset size of 57. The x-axis represents the selected feature subset size. Both the best and the baseline (which uses all features) results are shown using square shapes. The Sub-Challenge baseline is displayed with the dashed line.

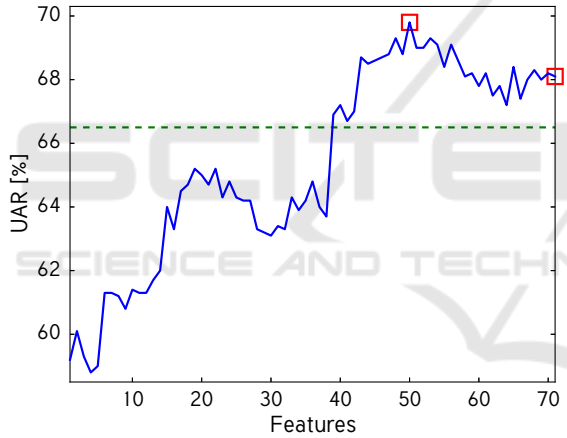


Figure 2: Scores obtained using the SGD classifier.

the SL classifier, where the *flatness* group ranks first. The *flatness* functional is defined as the ratio of the geometric mean to the arithmetic mean, both in absolute values (Eyben, 2016). In addition, the *rqmean* and *posamean* groups are ranked in the top five groups list for three of the classifiers.

On the low performing end, the *minSegLen* group is shared by all four systems while the *upleveltime90*, *minRangeRel*, and *nnz* groups are shared by three systems in their respective lists of the five lowest performing groups.

The degree of similarity among the tables suggests that the relevance of the functional-based feature groups to the task, indicated by the group rankings, is potentially valid in general regardless of the specific classifier used. (Eyben, 2016) provides further

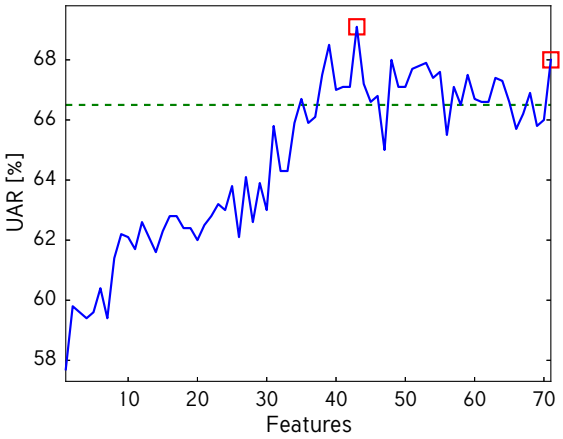


Figure 3: Scores obtained using the VP classifier.

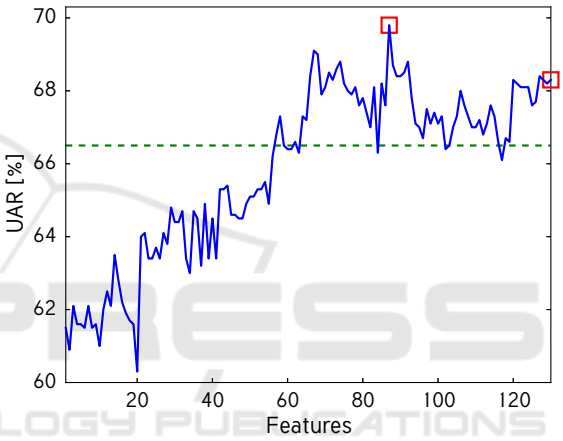


Figure 4: Scores obtained using the SL classifier.

detail about the statistical functionals used for generating the group features employed in this paper.

## 5.2 Individual Feature Selection Results

Figure 1 displays the result of the wrapper-based individual feature selection method using the RS algorithm for the system implemented with the RF classifier. The highest UAR evaluation score of 73.0% is attained using a feature subset of size 57 out of the original 71 features of the *amean* group feature. The baseline for the *amean* group shown in Figure 1 is greater than the value of 70.5% reported in Table 1 due to the reordering of the positions of the features in the ranked feature subset. Both the highest and the baseline values, indicated by the squares, are above the Sub-Challenge baseline of 66.5% indicated by the dashed line.

Figures 2, 3, and 4 display the results obtained by systems using the SGD, VP, and SL classifiers, respectively. In all four cases, the highest attained

Table 5: Classification results of the CGI-FS systems. CLS: Classifier name. G-FS: UAR evaluation score of the group feature selection step in %. I-FS: UAR evaluation score of the individual feature selection step in %. N: The final feature subset size in number of features (baseline feature set size is 6373). ↑ BL: Performance improvement of the CGI-FS system over the Sub-Challenge baseline of 66.5% in %. The highest score and the highest improvement are shown in bold.

| CLS | G-FS | I-FS | N | ↑ BL |
|---|---|---|---|---|
| RF | 70.5 | **73.0** | 57 | **9.8** |
| SGD | 68.1 | 69.8 | 50 | 5.0 |
| VP | 68.0 | 69.1 | 43 | 3.9 |
| SL | 68.3 | 69.8 | 87 | 5.0 |

values are above the baseline, which uses all of the features in the subset. Comparison of graph patterns displayed in the figures reveal that although the RF-based system starts out with the lowest evaluation score the accumulative subset selection process helps it bypass the Sub-Challenge baseline in fewer steps as well as achieve the highest performance among all four systems.

Furthermore, both the highest and the baseline values are above the Sub-Challenge baseline. The displayed results show that each step of our two-step feature selection process has improved accuracy performances using each of the four systems. In addition, the dimensionality of the problem has been greatly reduced from the original feature set's 6373 features to subset sizes in double digits. Table 5 displays, for each of the four CGI-FS systems, the evaluation scores obtained in the two feature selection steps, the final feature subset size, and the relative UAR improvement achieved over the Sub-Challenge baseline.

### 5.3 Previous Work

Addressing the Likability Sub-Challenge, (Guo et al., 2018) and (Haider et al., 2018) use development data to report results that outperform the baseline. A fusion of systems using deep representation, bag-of-audio-words, and functional-based features obtains the best performance in (Guo et al., 2018). The best result for (Haider et al., 2018) is obtained with a fusion of systems including those that use active feature transformation and active feature selection. The lack of performance report and the unsurpassed baseline performance result on test data (achieved using the video mode), highlight the fact that development models do not always generalize to the test data.

## 6 CONCLUSIONS AND FUTURE WORK

This paper presents the Cascaded acoustic Group and Individual Feature Selection (CGI-FS) method for automatic recognition of food likability addressed in the ICMI 2018 Eating Analysis and Tracking Challenge's Likability Sub-Challenge. CGI-FS employs an audio approach and is performed in a sequence of two feature selection operations. First, group feature selection is used to select the best performing functional-based group feature. Second, individual feature selection is performed on the previous step's resultant subset using a wrapper-based Rank Search algorithm for feature subset evaluation.

Four classifier-specific CGI-FS systems are implemented. All four CGI-FS system results outperform the Sub-Challenge baseline on iHEARu-EAT data suggesting the effectiveness of the method in general. The system implemented using the RandomForest (RF) classifier attains the best UAR score of 73.0% achieving a 9.8% relative UAR improvement over the Sub-Challenge baseline. The RF-based system reduces the number of baseline features from 6373 to 57, achieving a greater than 99% reduction in dimensions.

Future work includes investigating the use of other dimensionality reduction methods for both the group and the individual feature selection steps of our cascaded approach. In addition, to further improve classification performance, various fusions of the predictions made by our four CGI-FS systems will be considered.

## REFERENCES

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Dougherty, G. (2013). Feature extraction and selection. In *Pattern Recognition and Classification: An Introduction*, pages 123–141. Springer.

Eyben, F. (2016). *Real-time Speech and Music Classification by Large Audio Feature Space Extraction*. Springer.

Eyben, F., Weninger, F., Groß, F., and Schuller, B. (2013). Recent developments in opensmile, the munich open-source multimedia feature extractor. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 835–838. ACM.

Freund, Y. and Schapire, R. E. (1999). Large Margin Classification Using the Perceptron Algorithm. *Machine Learning*, 37(3):277–296.

Guo, Y., Han, J., Zhang, Z., Schuller, B., and Ma, Y. (2018). Exploring a new method for food likability rating based on dt-cwt theory. In *Proceedings of the 20th*

*ACM International Conference on Multimodal Interaction*, ICMI '18, pages 569–573, New York, NY, USA. ACM.

Gutlein, M., Frank, E., Hall, M., and Karwath, A. (2009). Large-scale attribute selection using wrappers. In *Computational Intelligence and Data Mining, 2009. CIDM'09. IEEE Symposium on*, pages 332–339. IEEE.

Haider, F., Pollak, S., Zarogianni, E., and Luz, S. (2018). Saameat: Active feature transformation and selection methods for the recognition of user eating conditions. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, ICMI '18, pages 564–568, New York, NY, USA. ACM.

Hantke, S., Schmitt, M., Tzirakis, P., and Schuller, B. (2018). Eat – the icmi 2018 eating analysis and tracking challenge. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, ICMI '18, pages 559–563, New York, NY, USA. ACM.

Hantke, S., Weninger, F., Kurle, R., Ringeval, F., Batliner, A., Mousa, A. E.-D., and Schuller, B. (2016). I hear you eat and speak: Automatic recognition of eating condition and food type, use-cases, and impact on asr performance. *PloS ONE*, 11(5):e0154486.

Kohavi, R. and John, G. H. (1997). Wrappers for feature subset selection. *Artificial Intelligence*, 97(1):273–324.

Ng, A. Y. (1998). On feature selection: Learning with exponentially many irrelevant features as training examples. In *Proceedings of the Fifteenth International Conference on Machine Learning*, ICML '98, pages 404–412.

Pir, D. (2018). Functional-based acoustic group feature selection for automatic recognition of eating condition. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, ICMI '18, pages 579–583, New York, NY, USA. ACM.

Pir, D. and Brown, T. (2015). Acoustic group feature selection using wrapper method for automatic eating condition recognition. In *Interspeech 2015 – 16th Annual Conference of the International Speech Communication Association, September 6-10, 2015, Dresden, Germany, Proceedings*, pages 894–898.

Schuller, B. and Batliner, A. (2014). *Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing*. John Wiley & Sons.

Schuller, B., Batliner, A., Seppi, D., Steidl, S., Vogt, T., Wagner, J., Devillers, L., Vidrascu, L., Amir, N., Kessous, L., and Aharonson, V. (2007). The relevance of feature type for the automatic classification of emotional user states: Low level descriptors and functionals. In *Interspeech 2007 – 8th Annual Conference of the International Speech Communication Association, August 27-31, Antwerp, Belgium, Proceedings*, pages 2253–2256.

Schuller, B., Steidl, S., and Batliner, A. (2009). The interspeech 2009 emotion challenge. In *Interspeech 2009 – 10th Annual Conference of the International Speech Communication Association, September 6–10, 2009, Brighton, UK, Proceedings*, pages 312–315.

Schuller, B., Steidl, S., Batliner, A., Hantke, S., Hönig, F., Orozco-Arroyave, J. R., Nöth, E., Zhang, Y., and Weninger, F. (2015). The interspeech 2015 computational paralinguistics challenge: Nativeness, parkinson's & eating condition. In *Interspeech 2015 – 16th Annual Conference of the International Speech Communication Association, September 6–10, Dresden, Germany, Proceedings*, pages 478–482.

Schuller, B., Steidl, S., Batliner, A., Vinciarelli, A., Scherer, K., Ringeval, F., Chetouani, M., Weninger, F., Eyben, F., Marchi, E., Mortillaro, M., Salamin, H., Polychroniou, A., Valente, F., and Kim, S. (2013). The interspeech 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism. In *Interspeech 2013 – 14th Annual Conference of the International Speech Communication Association, August 25–29, Lyon, France, Proceedings*, pages 148–152.

Sumner, M., Frank, E., and Hall, M. (2005). *Speeding Up Logistic Model Tree Induction*, pages 675–683. Springer Berlin Heidelberg, Berlin, Heidelberg.

Weninger, F., Eyben, F., Schuller, B., Mortillaro, M., and Scherer, K. (2013). On the acoustics of emotion in audio: what speech, music, and sound have in common. *Frontiers in Psychology*, 4:292.