WISE: Web-based Interactive Speech Emotion Classification

Sefik Emre Eskimez[⋆], **Melissa Sturge-Apple**[†], **Zhiyao Duan**[⋆] and **Wendi Heinzelman**[⋆]
[⋆]Dept. of Electrical and Computer Engineering
[†]Dept. of Clinical and Social Sciences in Psychology
University of Rochester, Rochester, NY

Abstract

The ability to classify emotions from speech is beneficial in a number of domains, including the study of human relationships. However, manual classification of emotions from speech is time consuming. Current technology supports the automatic classification of emotions from speech, but these systems have some limitations. In particular, existing systems are trained with a given data set and cannot adapt to new data nor can they adapt to different users' notions of emotions. In this study, we introduce WISE, a web-based interactive speech emotion classification system. WISE has a web-based interface that allows users to upload speech data and automatically classify the emotions within this speech using pre-trained models. The user can then adjust the emotion label if the system classification of the emotion does not agree with the user's perception, and this updated label is then fed back into the system to retrain the models. In this way, WISE enables the emotion classification models to be adapted over time. We evaluate WISE by simulating the user interactions with the system using the LDC dataset, which has known, ground-truth labels. We evaluate the benefit of the user feedback enabled by WISE in situations where manually classifying emotions in a large dataset is costly, yet trained models alone will not be able to accurately classify the data.

1 Introduction

Accurately estimating emotions of conversational partners plays a vital role in successful human communication. A social-functional approach to human emotion emphasizes the interpersonal function of emotion for the establishment and maintenance of social relationships [Campos *et al.*, 1989], [Ekman, 1992], [Keltner and Kring, 1998]. According to [Campos *et al.*, 1989] "Emotions are not mere feelings, but rather are processes of establishing, maintaining, or disrupting relations between the person and the internal or external environment, when such relations are significant to the individual." Thus, the expression and recognition of emotions allows the facilitation of social bonds through the conveyance of information about one's internal state, disposition, intentions, and needs.

In many situations, audio is the only recorded data for a social interaction, and estimating emotions from speech becomes a critical task for psychological analysis. Today's technology allows for gathering vast amounts of emotional speech data from the web, yet analyzing this content is impractical. This fact prevents many interesting large-scale investigations.

Given the amount of speech data that proliferates, there have been many attempts to create automatic emotion classification systems. However, the performance of these systems is not as high as necessary in many situations. Many potential applications would benefit from automated emotion classification systems, such as call-center monitoring [Petrushin, 1999; Gupta, 2007], service robot interactions [Park *et al.*, 2009; Liu *et al.*, 2013] and driver assistance systems [Jones and Jonsson, 2005; Tawari and Trivedi, 2010]. Indeed, there are many automated systems today that focus on speech [Sethu *et al.*, 2008; Busso *et al.*, 2009; Rachuri *et al.*, 2010; Bitouk *et al.*, 2010; Stuhlsatz *et al.*, 2011; Yang, 2015]. However, emotion classification accuracy of fully automated systems is still not satisfactory in many practical situations.

In this study, we propose WISE, a web-based interactive speech emotion classification system. This system uses a web-based interface that allows users to easily upload a speech file to the server for emotion analysis, without the need for installing any additional software. Once the speech files are uploaded, the system classifies the emotions using a model trained on previously labeled training samples. Each classification is also associated with a confidence value. The user can either accept or correct the classification, to "teach" the system the user's specific concept of emotions. Over time, the system adapts its emotion classification models to the user's concept, and can increase its classification accuracy with respect to the user's concept of emotions.

The key contribution of our work is that we provide an interactive speech-based emotion analysis framework. This framework combines the machine's computational power with human users' high emotion classification accuracy. Compared to purely manual labeling, it is much more efficient. Compared to fully automated systems, it is much more accurate. This opens up possibilities for large-scale speech emotion analysis with high accuracy.

The proposed framework only considers offline labeling

Proceedings of the 4th Workshop on Sentiment Analysis where AI meets Psychology (SAAIP 2016), IJCAI 2016, pages 2-7, New York City, USA, July 10, 2016.

and returns labels in three categories: emotion, arousal and valance with time codes. To evaluate our system, we have simulated the user-interface interactions in several settings, by providing ground truth labels on behalf of the user. One of the scenarios is designed to be a baseline, with which we can compare the remaining scenarios. In another scenario, we test if the system can adapt to the samples whose speaker is unknown to the system. The next scenario tests how the system's classification confidence of a sample effects the system's accuracy. The full system is available for researchers to use. [1]

The rest of the paper is organized as follows. Section 2 contains a review of the related work. Section 3 describes the WISE web user-interface, while Section 4 explains the automated speech-based emotion recognition system used in this work. We evaluate the WISE system in Section 5, and conclude our work in Section 6.

2 Related Work

All-in-one frameworks for automatic emotion classification from speech, such as EmoVoice [Vogt *et al.*, 2008] and OpenEar [Eyben *et al.*, 2009], are standalone software packages with various capabilities, including audio recording, audio file reading, feature extraction, and emotion classification.

EmoVoice allows the user to create a personal speech-based emotion recognizer, and it can track the emotional state of the user in real-time. Each user records their own speech emotion corpus to train the system, and the system can then be used for real-time emotion classification for the same user. The system outputs the x- and y-coordinates of an arousal-valance coordinate system with time codes. It is reported in [Vogt *et al.*, 2008] that EmoVoice has been used in several systems including humanoid robot-human and virtual agent-human interactions. EmoVoice does not consider user feedback once the classifier is trained, whereas in our system, the user can continually train and improve the system.

OpenEar is an emotion classification multi-platform software package that includes libraries for feature extraction written in C++ and pre-trained models as well as scripts to support model building. One of its main modules is named SMILE (Speech and Music Interpretation by Large-Space Extraction), and it can extract more than 500K features in real-time. The other main module allows external classifiers and libraries such as LibSVM [Chang and Lin, 2011] to be integrated and used in classification. OpenEar also supports popular machine learning frameworks' data formats, such as the Hidden Markov Model Toolkit (HTK) [Young *et al.*, 2006], WEKA [Hall *et al.*, 2009], and scikit-learn for Python [Pedregosa *et al.*, 2011], and therefore allows easy transition between frameworks. OpenEar's capability of batch processing, combined with its advantage in transitioning to other learning frameworks, makes it appealing for large databases.

ANNEMO (ANNotating EMOtions) [Ringeval *et al.*, 2013] is a web-based annotation tool that allows labeling arousal, valence and social dimensions in audio-visual data. The states are represented between -1 and 1, where the user changes the values using a slider. The social dimension is represented by categories rather than numerical values, and those are agreement, dominance, engagement, performance and rapport. No automatic classification/labeling modules are included in ANNEMO.

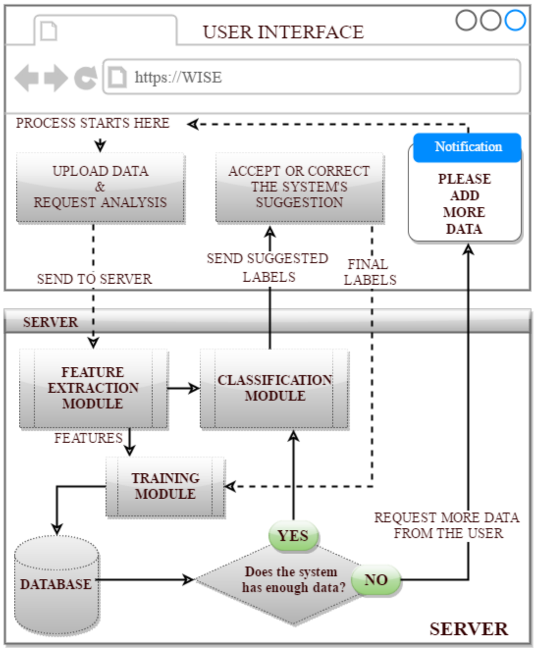


Figure 1: Flow chart showing the operation of WISE.

In contrast, WISE is a web-based system and can be used easily without installing any software, unlike EmoVoice and OpenEar. WISE is similar to ANNEMO in terms of the web-based labeling aspect, however WISE only considers audio data and provides automatic classification as well.

3 Web-based Interaction

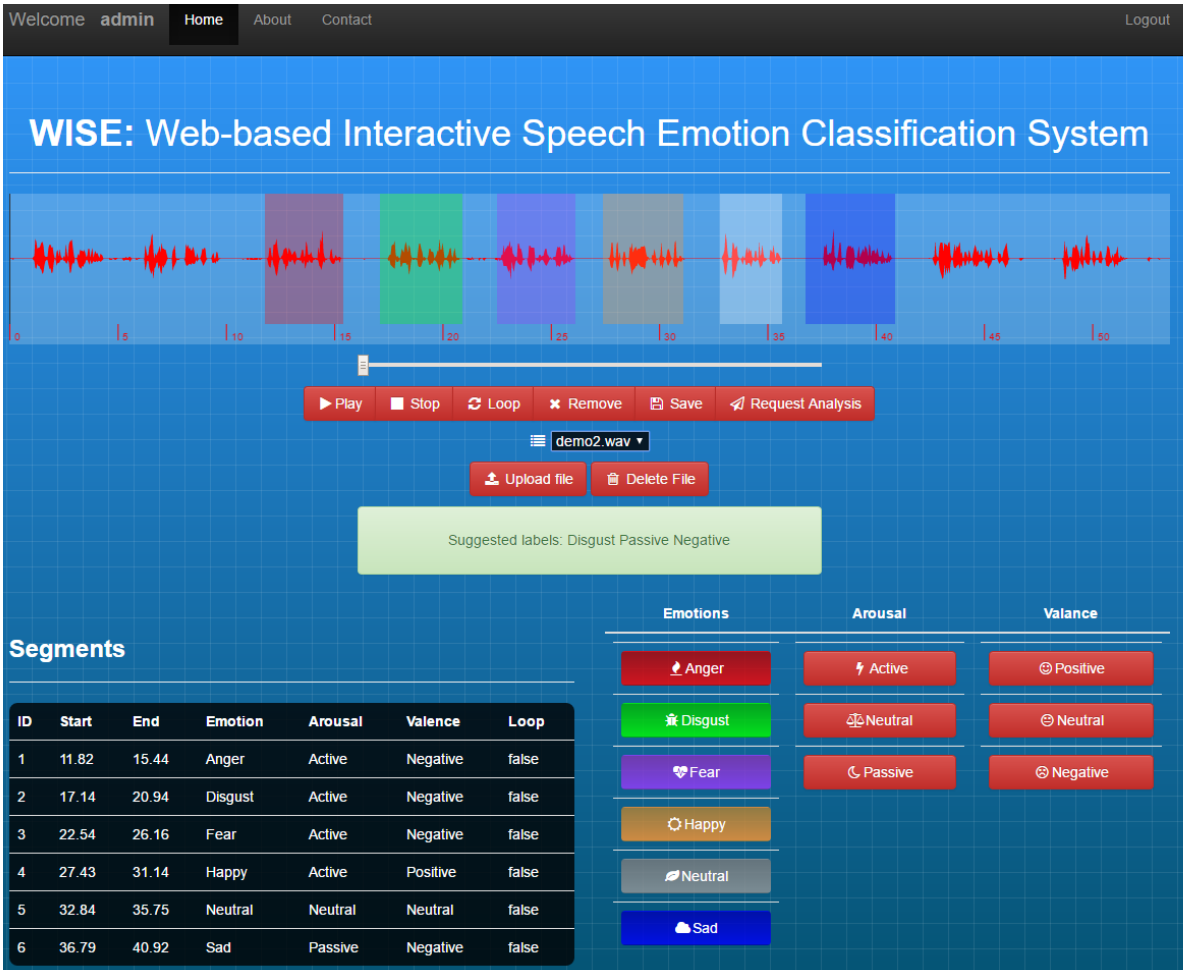
Our system's interface, shown in Figure 2, is web-based, allowing easy, secure access and use without installing any other software except a modern browser.

When a user uploads an audio file, the waveform appears on the main screen, allowing the user to select different parts of the waveform. Selected parts can be played and labeled independently. These selected parts will also be added to a list, as shown in the bottom-left side of Figure 2. The user can download this list by clicking on the "save" button in the interface.

The labeling scheme is restricted to three categories: emotion, arousal and valence. Emotion category elements are anger, disgust, fear, happy, neutral, sadness. Arousal category elements are active, passive and neutral, and valance category elements are positive, negative and neutral. Our future work includes adding user defined emotion labels into the system.

The user can request labels from the automated emotion classifier by clicking on the "request label" button. The system then shows suggested labels to the user.

[1] http://www.ece.rochester.edu/projects/wcng



WISE: Web-based Interactive Speech Emotion Classification System

Suggested labels: Disgust Passive Negative

Segments

ID	Start	End	Emotion	Arousal	Valence	Loop
1	11.82	15.44	Anger	Active	Negative	false
2	17.14	20.94	Disgust	Active	Negative	false
3	22.54	26.16	Fear	Active	Negative	false
4	27.43	31.14	Happy	Active	Positive	false
5	32.84	35.75	Neutral	Neutral	Neutral	false
6	36.79	40.92	Sad	Passive	Negative	false

Figure 2: WISE user interface screenshot.

The next section describes the automated speech-based emotion classification system used in WISE.

4 Automated Emotion Classification System

There are various automated speech-based emotion classification systems [Sethu *et al.*, 2008; Busso *et al.*, 2009; Rachuri *et al.*, 2010; Bitouk *et al.*, 2010; Stuhlsatz *et al.*, 2011] that consider different features, feature selection methods, classifiers and decision mechanisms. Our system is based on [Yang, 2015], which provides a confidence value along with the classification label.

4.1 Features

Speech samples are divided into overlapping frames for feature extraction. The window and hop sizes are set to 60 ms and 10 ms, respectively. For every frame that contains speech, the following features are calculated: fundamental frequency (F_0), 12 mel-frequency cepstral coefficients (MFCCs), energy, frequency and bandwidth of first four formants, zero-crossing rate, spectral roll-off, brightness, centroid, spread, skewness, kurtosis, flatness, entropy, roughness, and irregularity, in addition to the derivatives of these features. Statistical values such as minimum, maximum, mean, standard deviation and range (i.e., max-min) are calculated from all frames within the sample. Additionally, speaking rate is calculated over the entire sample. Hence, the final feature vector length is 331.

4.2 Feature Selection

The system employs the support vector machine (SVM) recursive feature elimination method [Guyon *et al.*, 2002]. This approach takes advantage of SVM weights to detect which features are better than others. After the SVM is trained, the features are ranked according to the order of their weights. The last ranked feature is eliminated from the list and the pro-

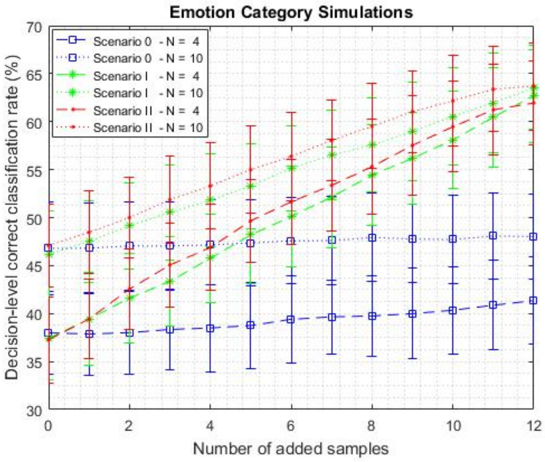


Figure 3: The results of emotion category for Scenarios I-III.

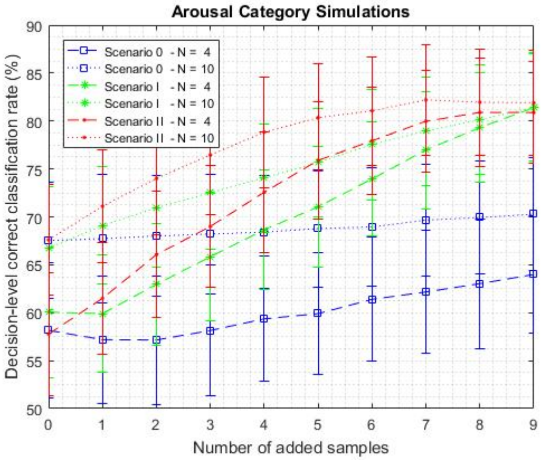


Figure 4: The results of arousal category for Scenarios I-III.

cess starts again, until there are no features left. Features are ranked in reverse order of elimination order. The top 80 best features are chosen to be used in the classification system. Note that in Section 5.2, the features are selected beforehand and not updated when a new sample is added to the system.

4.3 Classifier

Our system uses a one-against-all (OAA) binary SVM with radial basis function (RBF) for each emotion, arousal and valance category element, for a total of 12 SVMs. The trained SVMs calculate confidence scores for any sample that is being classified. The system labels the sample with the class of the binary classifier with maximum classification confidence on the considered sample.

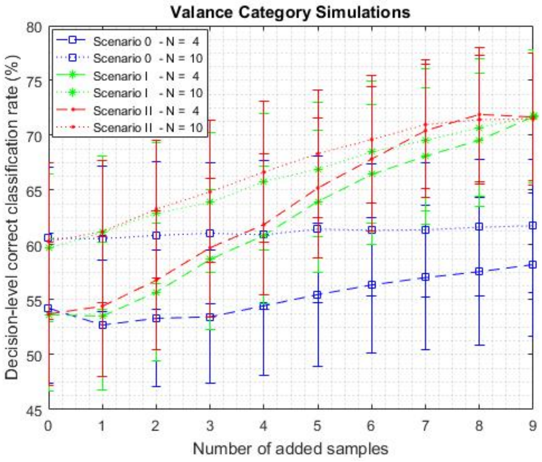


Figure 5: The results of valence category for Scenarios I-III.

5 Evaluation

To evaluate WISE and the benefit of user-assisted labeling of the data, we have simulated user-interface interactions using the LDC database as the source of data for training, validation and testing.

5.1 Dataset

We use the Linguistic Data Consortium (LDC) Emotional Prosody Speech and Transcripts [Liberman *et al.*, 2002] database in our simulations. The LDC database contains samples from 15 emotion categories; however, in our evaluation, we only use 6 of the emotions as listed in Section 3. The LDC database contains acted speech, voiced by 7 professionals, 4 female and 3 male. The transcripts are in English and contain semantically neutral utterances, such as dates and times.

5.2 Simulations

We have simulated user-interface interactions in different scenarios for which WISE can be used to enable user feedback to improve classification accuracy. In these simulations, there are three data groups: training, test and validation. We assume that validation data represents the samples where the user provides the "correct" label. In each iteration, the system evaluates the test data using the current models, and at the end of each iteration, a sample from the validation data is added to the training data to update the models. Next, we describe the different scenarios in detail.

Scenario 0 - Baseline

In this scenario, the data from 1 of the 7 speakers is used for testing, while the remaining 6 speakers' data are used for training and validation. Since only a limited amount of data is available from each speaker in the next scenarios, we also limit the amount of the validation data in this scenario. In this way, the baseline becomes more comparable to the other scenarios.

The training data starts with N samples from each class for each category. For the emotion classification, there are only 2 samples available in each class (emotion) for the validation data. However, the arousal and valance categories have half the number of classes that the emotion category has, therefore, there are 3 samples available in each class that can be used in validation data for these categories. After the data are chosen randomly, the system simulates the interaction process. This process is repeated for all speakers, and the results are averaged over all 7 speakers and 200 trials.

Scenario I

This scenario has the same settings as Scenario 0, except this time, the testing data, as well as the validation data are chosen from a speaker, and the training data is chosen among the remaining 6 speakers' data.

Scenario II

This scenario has the same settings as Scenario I with a single difference: in each round, the validation data has been ordered in ascending order of the classifier's confidence level in classifying them. Therefore in each iteration, the sample, on which the system has the least confidence, is added to the training data from the validation data.

Discussion

Figures 3-5 show the classification accuracy versus the number of added samples for each scenario for the emotion, arousal and valence, respectively. Note that the error bars represent the standard deviation of the results over the 7 speakers and 200 trials.

Scenario I shows the ability of WISE to enable adaptation of the models. In many situations, trained models of automatic systems have no information on the speaker to be classified. The comparison of classification accuracy between Scenario 0 and Scenario I shows that adaptation to unknown data is vital for accurate emotion estimation, as the accuracy increases greatly when data from the new user are added.

For example, in Scenarios I and II, when N is 4 for the emotion category, the system's initial accuracy starts around 37% and increases to approximately 63%, as can be seen in Figure 3, where on the other hand in Scenario 0, accuracy can only increase to approximately 41%. In Scenarios I and II , when N is 10, the classification accuracy starts higher then the previous case, yet with the same number of added samples, they converge to the same percentage. This enables the possibility of using pre-trained models in our system that are trained on available databases.

The results of Scenario II suggest that adding the samples with low classification confidence are slightly more beneficial than adding a sample for which the system already has more confidence. Figures 3-5 show that the classifier in Scenario II converges to a slightly higher classification accuracy than the one in Scenario I. This can be seen especially in the arousal category results.

6 Conclusion

This study introduced and evaluated the WISE system, which is an interactive web-based emotion analysis framework to assist in the classification of human emotion from voice data. The full system is available for the community to use. The evaluation results show that the system can adapt to the user's choices and can increase the future classification accuracy when the speaker of the sample is unknown. Hence, WISE will enable adaptive, large scale emotion classification.

References

[Bitouk *et al.*, 2010] Dmitri Bitouk, Ragini Verma, and Ani Nenkova. Class-level spectral features for emotion recognition. *Speech Commun.*, 52(7):613–625, 2010.

[Busso *et al.*, 2009] C. Busso, S. Lee, and S. Narayanan. Analysis of emotionally salient aspects of fundamental frequency for emotion detection. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(4):582–596, May 2009.

[Campos *et al.*, 1989] Joseph J Campos, Rosemary G Campos, and Karen C Barrett. Emergent themes in the study of emotional development and emotion regulation. *Dev Psychol.*, 25(3):394, 1989.

[Chang and Lin, 2011] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011.

[Ekman, 1992] Paul Ekman. An argument for basic emotions. *Cognition Emotion*, 6(3-4):169–200, 1992.

[Eyben *et al.*, 2009] Florian Eyben, Martin Wllmer, and Bjrn Schuller. openear - introducing the munich open-source emotion and affect recognition toolkit. In *In ACII*, pages 576–581, 2009.

[Gupta, 2007] Purnima Gupta. Two-Stream Emotion Recognition For Call Center Monitoring. In *Interspeech 2007*, 2007.

[Guyon *et al.*, 2002] Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. Gene selection for cancer classification using support vector machines. *Mach. Learn.*, 46(1-3):389–422, March 2002.

[Hall *et al.*, 2009] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The weka data mining software: An update. *SIGKDD Explor. Newsl.*, 11(1):10–18, November 2009.

[Jones and Jonsson, 2005] Christian Martyn Jones and Ing-Marie Jonsson. Automatic recognition of affective cues in the speech of car drivers to allow appropriate responses. In *Proceedings of the 17th Australia Conference on Computer-Human Interaction: Citizens Online: Considerations for Today and the Future*, OZCHI '05, pages 1–10, Narrabundah, Australia, Australia, 2005.

[Keltner and Kring, 1998] Dacher Keltner and Ann M Kring. Emotion, social function, and psychopathology. *Rev. Gen. Psychol.*, 2(3):320, 1998.

[Liberman *et al.*, 2002] Mark Liberman, Kelly Davis, M Grossman, N Martey, and J Bell. Emotional prosody speech and transcripts. In *Proc. LDC*, 2002.

[Liu *et al.*, 2013] Chih-Yin Liu, Tzu-Hsin Hung, Kai-Chung Cheng, and Tzuu-Hseng S Li. Hmm and bpnn based speech recognition system for home service robot. In *Advanced Robotics and Intelligent Systems (ARIS), 2013 International Conference on*, pages 38–43. IEEE, 2013.

[Park *et al.*, 2009] J. S. Park, J. H. Kim, and Y. H. Oh. Feature vector classification based speech emotion recognition for service robots. *IEEE Transactions on Consumer Electronics*, 55(3):1590–1596, August 2009.

[Pedregosa *et al.*, 2011] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[Petrushin, 1999] Valery A. Petrushin. Emotion in speech: Recognition and application to call centers. In *In Engr*, pages 7–10, 1999.

[Rachuri *et al.*, 2010] Kiran K Rachuri, Mirco Musolesi, Cecilia Mascolo, Peter J Rentfrow, Chris Longworth, and Andrius Aucinas. Emotionsense: a mobile phones based adaptive platform for experimental social psychology research. In *Proc. 12th ACM Int. Conf. on Ubiquitous Computing*, pages 281–290, 2010.

[Ringeval *et al.*, 2013] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne. Introducing the recola multimodal corpus of remote collaborative and affective interactions. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–8, April 2013.

[Sethu *et al.*, 2008] Vidhyasaharan Sethu, Eliathamby Ambikairajah, and Julien Epps. Empirical mode decomposition based weighted frequency feature for speech-based emotion classification. In *Proc. IEEE ICASSP*, pages 5017–5020, 2008.

[Stuhlsatz *et al.*, 2011] A. Stuhlsatz, C. Meyer, F. Eyben, T. Zielke, G. Meier, and B. Schuller. Deep neural networks for acoustic emotion recognition: Raising the benchmarks. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pages 5688–5691, May 2011.

[Tawari and Trivedi, 2010] A. Tawari and M. Trivedi. Speech based emotion classification framework for driver assistance system. In *Intelligent Vehicles Symposium (IV), 2010 IEEE*, pages 174–178, June 2010.

[Vogt *et al.*, 2008] Thurid Vogt, Elisabeth Andr, and Nikolaus Bee. Emovoice - a framework for online recognition of emotions from voice. In *In Proceedings of Workshop on Perception and Interactive Technologies for Speech-Based Systems, Springer, Kloster Irsee*, 2008.

[Yang, 2015] Na Yang. *Algorithms for affective and ubiquitous sensing systems and for protein structure prediction*. PhD thesis, University of Rochester, 2015. http://hdl.handle.net/1802/29666.

[Young *et al.*, 2006] S. J. Young, G. Evermann, M. J. F. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. C. Woodland. *The HTK Book, version 3.4*. Cambridge University Engineering Department, Cambridge, UK, 2006.