Senior Honors Project

Machine Recognition of Vocal Emotion

Rebecca Roseman

There are many aspects to communication. Specifically, paralinguistic information, such as emotion, can enrich the communicative process greatly by providing valuable context for spoken language comprehension. Additionally, emotion is a multimodal experience, which makes its perception an interesting phenomenon with many facets to study. Previous research has indicated that humans can accurately judge emotion from facial expressions cross culturally (Ekman et al., 1987). This research suggests that emotion expression and comprehension may be universal. In further investigations, Sauter et al. conducted research on whether detection of emotion in non-linguistic vocal utterances was universal and found encouraging evidence. Across drastically different Western and isolated Namibian populations, humans accurately identified emotions expressed in non-linguistic utterances such as laughs, sighs, and screams (Sauter et el., 2010). Sauter et al.'s findings lend support to the possibility that the vocalization of linguistic utterances of emotion, in addition to non-linguistic utterances, might be universal. However, research investigating whether humans can universally identify linguistic utterances of emotion throws a shadow of doubt on this hypothesis. Work by Scherer et al. looked at whether German, French, English, Dutch, Italian, Spanish, and Bahasa Indonesian speakers could identify the emotions of joy/happiness, sadness, fear, anger, and disgust regardless of language. They used stimuli collected

from four German actors (two male and two female), who were recorded speaking sentences that were created using the "standard sentence" approach, such that the semantic content of the sentences was meaningless and could not be used to judge the emotion intended. Their subjects were acquired through collaborations with Germany, Switzerland, Great Britain, Netherlands, United States, Italy, France, Spain, and Indonesia. Their results showed that people were good at judging emotion from speech from their native language (which was observed in the German speakers), but that people did not do as well across languages. The average recognition percentage for German speakers across emotions was 73.8%, while the average recognition percentage of emotions for the speakers of other languages was only 65.4% when they heard German speech. People were also shown to be significantly less accurate at identifying joy, which had an average recognition rate of 42% across all of the speakers, compared to an average rate of 71.5% for the other emotions across speakers. Joy was seen to be often confused with neutral, and the reasons for this remain unclear (Scherer et al., 2001).

We are now left wondering why people do so poorly at identifying emotion in linguistic expressions cross-culturally. Do people simply need to learn the relations between emotion and linguistic content by experience? Is there no relation across languages for how emotions are expressed in verbal language? It is possible that populations that speak different languages use differing proportions of the same linguistic features to convey emotions. Or, maybe, these different communities of speakers rely on different linguistic features entirely to convey emotion in speech.

Computational methods of investigation provide a nice opportunity for answering these questions. While we cannot directly investigate how humans solve the problem of the relation between emotional content and linguistic content, we can train a computational model to solve the same problem and observe how it performs. Work has been done with supervised learning algorithms to investigate whether they can learn to identify emotion in linguistic utterances within and across languages — and if they can, how well they can learn and perform. A paper by Ververidis and Kotropoulos reviews the quality of many different corpuses of emotionally affected speech samples and considers how many different classification models perform in this emotion detection task (Ververidis & Kotropoulos, 2006). They show that the emotion detection problem can be solved using computational methods within a language. However, it seems that computational research thus far has focused more on solving the problem within a single language than investigating the intricacies of the across language problem. Thus, the question still remains of why it is much more difficult to discriminate emotion in non-native speech.

The present study conducts investigations to begin answering this question. We conducted the following experiments. First, we investigated whether or not we could discriminate the emotions expressed in speech samples within a language in a computational way, using a support vector machine, or SVM. SVM is a supervised learning algorithm that can process a set of labeled data to learn the relations between them, becoming "trained" to label new data. The model learns the relations between a given set of inputs and outputs so that it can later predict the output of

new inputs. We trained four different support vector machine (SVM) models, one with English, one with Dutch, one with Mandarin, and one with Vietnamese samples, and tested each SVM model's accuracy within the language each one was trained on. The labeled data in our experiment were the speech segments that had been pre-labeled with the emotions they demonstrated. We made sure that the model could solve the task (i.e. identify and label new emotion inputs) to a satisfactory degree, given what similar studies have been able to find. After testing the SVM models within the languages they were trained on, we tested the SVM models on stimuli from outside the language they were trained on and then quantified their accuracy.

Method

Participants

There are no active participants in this experiment per se, as it is a computational study. The languages investigated in this study are English, Dutch, Vietnamese, and Mandarin, so this data is applicable to those communities of speakers. A behavioral study conducted with human subjects from these language backgrounds had been conducted previously. Importantly, the results from this study looked at the accuracy of trials in which the participants heard low-pass filtered sounds only. This technique ensures that the subjects are using only paralinguistic cues and cannot rely on the semantic content of the utterances for emotion detection (Frye, 2013).

Materials

The linguistic utterance stimuli in each language had already been collected and used in a study by Frye (Frye, 2013). The stimuli consisted of sentences spoken by male and female speakers in the four languages for each of six emotions. The six emotions the experiment studied were fear, surprise, sadness, happiness, disgust, and anger. For each of the four languages studied, two male and two female speakers spoke sentences exhibiting each of the six emotions. This gave us 16 sentences for each of the six emotions across the four languages. We had access to the human results of this experiment using undergraduates from the University of California, San Diego (Frye, 2013). We used a free SVM library that was available online (LIBSVM), ran it, and analyzed the results in Matlab (Chang & Lin, 2011). We referred to Schuller et al. to choose which features to use in our analysis. Schuller et al. reports a list of 33 features, from which we pulled five to use in our experiments, all pitch-based features that we would use to conduct our analysis (Schuller et al., 2004). Specifically, we used maximum, minimum, and average pitch, standard deviation of pitch, and pitch range values normalized to each speaker.

Procedure

Experiment 1

The first question at hand is whether or not our computational model can discriminate the emotional content of a speech sound within one language. To evaluate this question, we trained an SVM classifier to identify emotions in sound clips in a single language. We trained four separate SVMs, one in each language, and

tested them on a set of unlearned stimuli to see if they had learned to discriminate the emotions in each speech sound to a comparable degree to what human subjects have demonstrated. We used the results from Frye's study as a baseline to determine how well we should expect our model to perform (Frye, 2013). Each model trained on a random 80% of the available data, and we used the RBF kernel and 5-fold cross validation settings while training each model. We searched for the best C and gamma parameters and saved them to use for the testing stage. In testing, each model tested on the remaining 20% of unseen data, used the RBF kernel setting, and the best C and gamma parameters from the training stage.

Results

To attain the within-language accuracy results, each model was trained and tested 10 times, using a randomly selected 80% of data to train and the remaining 20% to test, and those accuracies were averaged for the final results reported here. The SVMs perform slightly better than the human subjects, but are within the same general range, as we can tell from the results from Frye depicted in Figure 1 below. As depicted in Figure 2, which shows our SVM models' performance, the English model performed with an average accuracy of 43.55%, the Mandarin model performed with an average accuracy of 34.03%, the Vietnamese model performed with an average accuracy of 34.81%, and the Dutch model performed with an average accuracy of 42.47%.

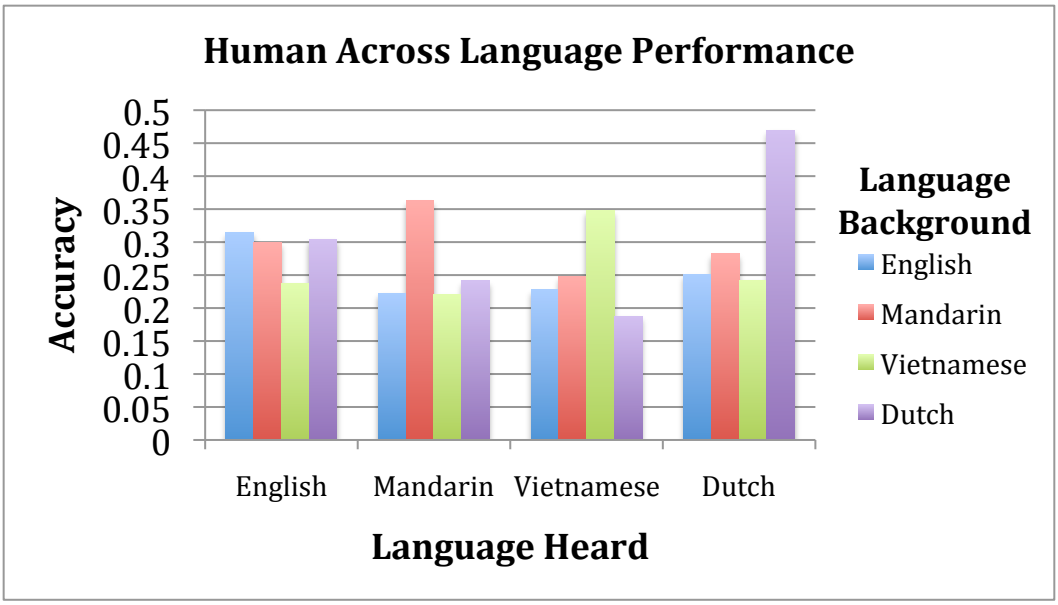

Figure 1. Human Across Language Performance Accuracies (Frye, 2013)

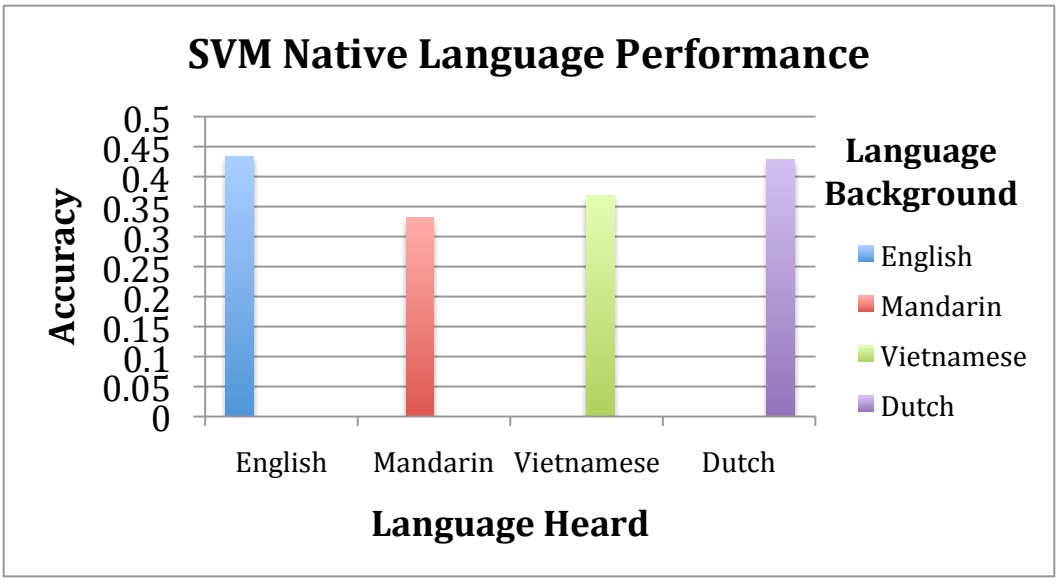

Figure 2. SVM Native Language Performance Accuracies

Experiment 2

Next, we investigated whether we could discriminate the emotional content of a speech sound from another language in a computational way. This time, we trained each SVM on all of the data from one language and tested it on all of the data from each of the other languages. Again, we used the RBF kernel and 5-fold cross validation settings while training each model, and we searched for the best C and gamma parameters and saved them to use during the testing stage. In the testing stage, each model was tested separately on the data from each foreign language and then compared to the average accuracy attained for its "native" language in the previous experiment.

Results

These results are interesting for several reasons. First, they are similar to the human behavioral results in that each model performed best when tested on the sounds of the language it had been trained on, mimicking that of human speakers performing best at identifying emotions in the sounds from the language they speak. In addition, all of the SVM accuracies were slightly higher than the human results, as seen in the first experiment, and each model also seemed to perform quite well at detecting emotion in the English and Dutch language samples compared to the other languages. Each model performed as well as or nearly as well in its "native" language as it did on the English and Dutch tests, as can be seen in Figure 3 below.

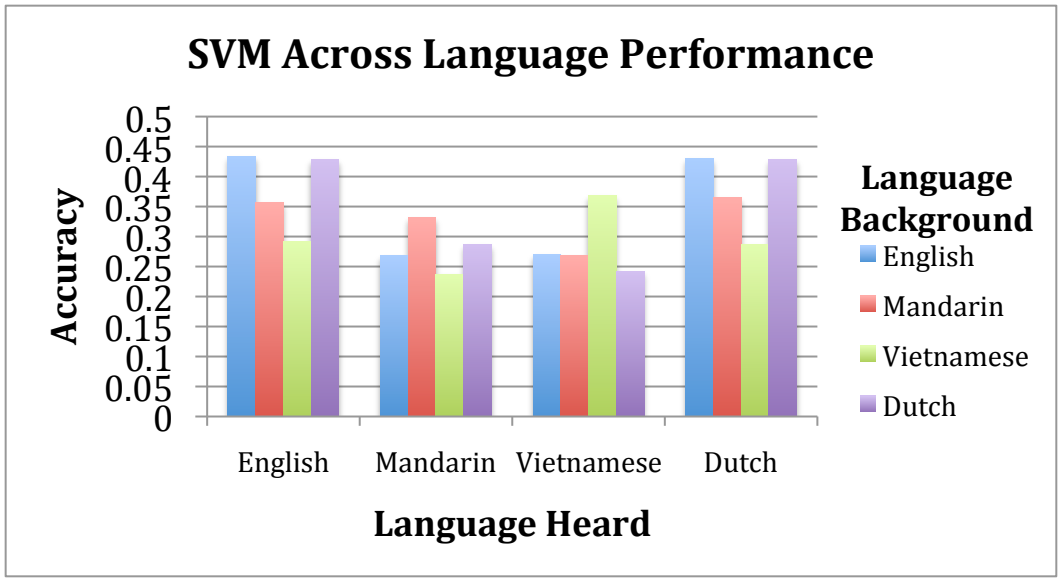

Figure 3. SVM Across Language Performance Accuracies

Discussion

The result that the SVMs do poorly at identifying and labeling emotions in speech from other languages leaves room for more questions. Why can we discriminate emotions in our own language, but not in others? It is possible that every language encodes emotion in a different way, using different combinations of different features. Or maybe languages use the same features, but in such vastly differing combinations that it is difficult to see similarities across languages. So we wonder, can we determine which features the classifier uses to distinguish between emotions within each language, and can we see whether they track similar or different features to identify emotions across languages? This would be an excellent direction for further research. Additionally, the SVM had slightly higher accuracies than human subjects across languages. So why does the computer beat the people? It seems plausible that the computer might be aware, more attentive to, and have

better memory for technical aspects of sounds than humans. With this in mind, it would be interesting to see if we could use this algorithm to see what combination of features most closely approximates human performance to provide some insight into how humans attempt to classify emotions in speech. And lastly, we came across the interesting result that we seemed to see better performance on English and Dutch stimuli and worse performance of Mandarin and Vietnamese stimuli. Why might this be? Possibly, the features we used are less informative for emotion in the Mandarin and Vietnamese languages. Or maybe, it is because Mandarin and Vietnamese are both tonal languages, and extra tonal information contained in those utterances distracted from the relevant cues to emotion. A deeper investigation into this would be interesting as well.

Conclusion

To conclude, this study investigated whether we can computationally categorize emotional utterances in unfamiliar languages. We used an SVM classifier to obtain classification accuracies within and across languages for English, Mandarin, Vietnamese, and Dutch. The model performed comparably to human subjects both within language and across languages. Each of the models performed best within its own language, though with a slightly higher accuracy than humans, and worse across languages, though not as poorly as humans did outside their native language. Additionally, there appears to be an advantage for detecting emotion in English and Dutch utterances.

References

Chang, C., & Lin, C., LIBSVM : a library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2:27:1--27:27, 2011. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm

Ekman, P., Friesen, W. V., O'Sullivan, M., Chan, A., Diacoyanni-Tarlatzis, I., Heider, K., . . . Tzavaras, A. (1987). Universals and cultural differences in the judgments of facial expressions of emotion. *Journal of Personality and Social Psychology, 53*(4), 712-717. doi:http://dx.doi.org/10.1037/0022-3514.53.4.712

Frye, C. (2013) Emotional Speech Processing and Language Knowledge. Second Year Project, University of California, San Diego.

Sauter, D. A., Eisner, F., Ekman, P., & Scott, S. K. (2010). Cross-cultural recognition of basic emotions through nonverbal emotional vocalizations. *PNAS Proceedings of the National Academy of Sciences of the United States of America, 107*(6), 2408-2412. doi:http://dx.doi.org/10.1073/pnas.0908239106

Scherer, K. R., Banse, R., & Wallbott, H. G. (2001). Emotion inferences from vocal expression correlate across languages and cultures. *Journal of Cross-Cultural Psychology, 32*(1), 76-92. doi:http://dx.doi.org/10.1177/0022022101032001009

Schuller, B.; Rigoll, G.; Lang, M., "Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture," *Acoustics, Speech, and Signal Processing, 2004. Proceedings. (ICASSP '04). IEEE International Conference on* , vol.1, no., pp.I,577-80 vol.1, 17-21 May 2004
doi: 10.1109/ICASSP.2004.1326051

Ververidis, D., & Kotropoulos, C. (2006). Emotional speech recognition: Resources, features, and methods. *Speech Communication, 48*(9), 1162-1181. doi:http://dx.doi.org/10.1016/j.specom.2006.04.003