Cross-cultural Emotion Classification: the Effect of Emotional Intensity and Acoustic Features

Anonymous ACL submission

Abstract

Cross-cultural emotion recognition is attracting increasing research attention; robustness to such differences in emotional expression is important for speech modality emotion recognition. In this work we quantify the accuracy loss when classifying cross-culturally for multiple emotional intensities, and investigate the effect of feature sets, including feature importance. We find that different emotional intensities yield a similar decrease in cross-culture accuracy relative to within-culture, and different acoustic feature sets also yield similar relative cross-culture accuracy. The top 10 important eGeMAPS features for within-cultural and cross-cultural classification share only one common feature, which partially explains differences in accuracy.

1 Introduction

Emotion recognition aims to predict a person's emotional state from external observations. Speech is one modality, where emotion is expressed in *what* a person says and *how* they say it, and is also important for multi-modal emotion recognition (Poria et al., 2017). Of increasing interest is *cross-cultural* emotion recognition (Schuller, 2018). Culture influences a person's language, accent (Prasad and Jyothi, 2020) and emotional expression (Laukka et al., 2016), among other effects. Emotional expression is encoded in acoustic features alongside confounding information such as pronunciation and speakers' vocal characteristics. Since there is some universality to emotional expression (Laukka et al., 2016) it is thus important for cross-cultural recognition to be predictive of emotion but robust to changes in acoustic information related to culture. Robust cross-cultural emotion recognition is thus important for affective computing in an international setting.

This study complements previous work on emotion classification across cultures. We use one of the largest balanced datasets for cross-cultural emotional speech research, which includes three emotional intensities, to investigate the effect of intensity on within and cross-culture emotion classification. We also investigate the effect of feature set on accuracy differences, and determine which features are most important for classification.

Our contributions are as follows: 1) We modify previous cross-culture testing methodology to include additional test set conditions. 2) We investigate the difference between within and cross-culture accuracy for 3 emotional intensities. 3) We compare the effect of cross-intensity prediction to cross-culture prediction. 4) We determine whether the choice of feature representation affects the difference between within and cross-culture recognition accuracy. 5) We measure feature importance for within and cross-culture classification, relating important features to differences in accuracy.

2 Related Work

The effect of culture on human emotion recognition has been investigated by Cowen et al. (2019). Comparing US and Indian ratings, they find an in-group advantage in recognising emotions within-culture, but moderate to high correlation between US and India affective ratings. They also find similar acoustic correlates to affective ratings between both countries' ratings.

Laukka et al. (2014) investigated the effect of culture on automatic emotion recognition, finding that within-culture/within-intensity testing yields higher accuracy than both within-culture/cross-intensity and within-intensity/cross-culture testing. In addition Laukka et al. (2016) quantify differences in a subset of GeMAPS features across the same cultures. Significant differences were found in 15 features across emotional expressions, with significant interaction effects for 9 features, and significant cross-cultural differences in some features for each emotion category.

Some papers have explored feature relevance for emotion prediction (Schuller et al., 2007; Ververidis et al., 2004; Busso et al., 2009), but not for cross-culture classification.

Our results complement previous work by comparing additional test conditions, determining the effect of emotional intensity and feature set, and analysing feature importance specific to cross-cultural emotion prediction.

3 Methodology

We use a subset of the VENEC data (Laukka et al., 2010) containing 10 emotions. Twenty speakers from each of Australia, India, Singapore, Kenya, and USA say the same English sentence in all emotions with high, medium and low intensity, yielding 3000 instances in total. We use a support vector machine (SVM) with linear kernel for multiclass prediction since it performs similarly to other classifiers when using utterance-level features (Keesing et al., 2021). The SVM cost parameter C is optimised over $\{2^{-6}, 2^{-4}, \ldots, 2^{6}\}$ using inner speaker-independent cross-validation.

3.1 Classification accuracy

We use 8 different train/test conditions outlined in Table 1 where #C is the number of combinations, and #E is the number of experiments using a consistent train set size.

Condition	Combinations	Experiments
wc_wi	15	15
wc_oi	5	15
wi_logo_cc	15	60
wi_pair_cc	60	60
oi_logo_cc	5	60
oi_pair_cc	20	60
wc_pair_ci	60	60
pair_cc_ci	120	120

Table 1: Different train/test conditions used in experiments. ‘wc’ = within-culture, ‘wi’ = within-intensity, ‘cc’ = cross-culture, ‘ci’ = cross-intensity, ‘oi’ = omni-intensity, ‘LOGO’ = leave-one-group-out, ‘pair’ = pairwise.

We consider each country and each intensity as a group, so there are five countries, three intensities, and 15 (country, intensity) pairs. For the within-culture conditions train and test data are from the same country. For the LOGO-cross-culture conditions the train set contains data from four countries and the test set has the remaining country. For the pairwise-cross-culture conditions we train on data from one country and test on data from another country for each pair of countries. Similar descriptions are for the within-intensity and cross-intensity conditions. The omni-intensity conditions are where the train and test sets have data from all intensities. For the pairwise-cross-culture/cross-intensity condition, a pair of countries and a pair of intensities are used such that both country and intensity differ between train and test sets.

To account for train set size as a factor, we maintain similar train set sizes for all experiments (approx. 200 instances), by selecting a random subset of speakers using stratified sampling. For example in the within-intensity/LOGO-cross-culture condition a quarter of the speakers from each of the four training countries is used as training set, and we have four such train sets. This maintains speaker-independent testing consistently for all train/test conditions. We also provide open-source code.

These experiments are performed for each of three feature sets: the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) (Eyben et al., 2016), the INTERSPEECH 2013 Computational Paralinguistics Challenge features (IS13) (Schuller et al., 2013), and mean-pooled wav2vec (Schneider et al., 2019) context embeddings. These features cover both high and low dimensionality, include hand crafted features, and have reasonable performance for emotion classification (Keesing et al., 2021). In each experiment features are individually standardised to zero mean and unit variance using parameters from the training set.

3.2 Feature importance

We use forward selection on the eGeMAPS set, since this is relatively small (88 features) and contains hand-crafted features meant to encode various aspects of speech, which are more interpretable than wav2vec or IS13 features. Forward selection greedily selects the feature which maximises accuracy each iteration. The resulting feature ordering indicates the most predictive features for classification.

Forward selection is a generic feature selection method which can be applied to any training task, so we use it to measure feature importance for both wc_wi and wi_pair_cc conditions. We again use SVM with linear kernel but fix the cost parameter $C = 1$ since this yielded accuracy within 1% of hyperparameter tuned accuracy in all cases.

4 Results

The results for the classification experiments is shown in Table 2. For all conditions, the Shapiro-Wilks test for normality shows no significant deviation from normality, so we use parametric statistical tests in our analysis.

We compare differences in accuracy between the conditions (wc_wi, wc_oi, wi_logo_cc, wi_pair_cc, oi_logo_cc, and oi_pair_cc) using repeated measures ANOVA for each feature set independently. This shows statistically significant results for all three features sets ($p < 0.001$ in all cases). Bonferroni corrected pairwise t-tests indicate statistically significant differences between the following pairs when using wav2vec features: (wi_logo_cc, oi_logo_cc), (wi_logo_cc, oi_pair_cc), (wi_pair_cc, oi_pair_cc). A boxplot for the within and cross-culture conditions for wav2vec features is shown in Figure 1.

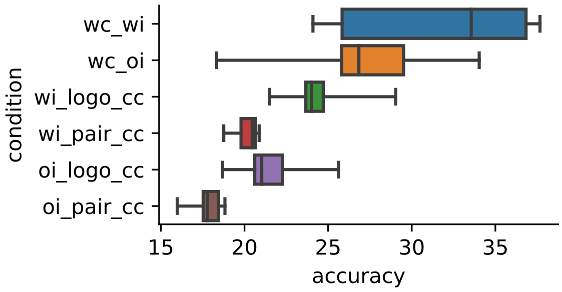

Figure 1: Boxplot of mean accuracy for within and cross-culture conditions using wav2vec embeddings.

We perform a 2x2 comparison of accuracies from the wc_wi, wc_pair_ci, wi_pair_cc and pair_cc_ci conditions, using two-factor repeated measures ANOVA, where the factors are intensity and culture, and each has ‘within’ and ‘cross’ levels. Each (country, intensity) pair is a subject, so we have 15 data points for each condition. Significant main effects were found for intensity for all feature sets ($p < 0.028$ in all cases), and country for all feature sets ($p < 10^{-5}$ in all cases). Classifying instances from a different emotional intensity than training data decreases accuracy by about 2.2% for wav2vec embeddings, while the decrease for cross-culture accuracy is about 11%, and for both together about 12%. Significant interaction effects were only found for eGeMAPS features ($p = 0.003$).

For each feature set and (country, intensity) pair, we calculate the values $d_{pair} = wc\_wi - wi\_logo\_cc$, and $d_{logo} = wc\_wi - wi\_pair\_cc$. Repeated measures ANOVA yielded no statistically significant differences in either d_{logo} or d_{pair} across feature sets ($p > 0.3$ in both cases). There were also no significant differences in either d_{logo} or d_{pair} across intensities ($p > 0.06$ in all cases). For each feature set d_{logo} is approx. 7% accuracy, while d_{pair} is between 9% and 11%. The within-culture accuracy is consistently 1.5 times the pairwise cross-culture accuracy, and 1.3-1.4 times the LOGO cross-culture accuracy.

We repeated all experiments and found the general trends to be stable, with the exception of interaction effects between intensity and country, and absolute differences in d_{pair} between feature sets, which were sometimes quite similar and sometimes quite different.

We report the top 10 features selected using forward selection in Table 3 for within and cross-culture classification. Accuracy levelled off at around 15 features, before decreasing when using more than 50 features. Accuracy when using features from the opposite column, and baselines using all 88 features and 10 random features, are given for comparison.

5 Discussion

We obtain highest accuracy when predicting on data from the same country and intensity as the train set, which is expected because optimal classification is achieved on test data from the same distribution. Hence on data with a different distribution, as in the cross-culture conditions, the model performs more poorly. The omni-intensity conditions have somewhat lower accuracy than their corresponding within-intensity conditions because when intensities are pooled, the model must predict accurately for all three intensities together instead of only one. This yields lower accuracy for each intensity, and thus lower accuracy than the within-intensity accuracies.

Predicting with four countries in the train set (LOGO) yields 4-5% higher accuracy than when using only one country (pairwise), because the train-

	wc_wi	wc_oi	wi_logo_cc	wi_pair_cc	oi_logo_cc	oi_pair_cc	wc_pair_ci	pair_cc_ci
AUS	33.6 ± 1.9	26.8 ± 2.1	24.0 ± 1.4	20.7 ± 1.6	22.3 ± 0.6	17.8 ± 1.4	28.1 ± 1.3	19.6 ± 0.8
IND	36.8 ± 2.0	34.0 ± 2.0	23.7 ± 1.4	19.8 ± 1.4	20.6 ± 0.9	17.5 ± 1.4	33.5 ± 1.3	18.1 ± 0.9
KEN	24.1 ± 1.6	18.3 ± 1.6	21.5 ± 0.9	18.8 ± 1.0	18.7 ± 0.9	16.0 ± 1.3	26.5 ± 0.9	17.4 ± 0.8
SIN	25.8 ± 1.6	25.8 ± 2.3	24.7 ± 1.0	20.9 ± 0.6	21.0 ± 0.7	18.5 ± 1.1	26.4 ± 1.2	20.6 ± 0.7
USA	37.7 ± 1.9	29.5 ± 2.0	29.0 ± 1.1	20.5 ± 1.5	25.6 ± 0.8	18.8 ± 1.4	32.2 ± 1.3	19.7 ± 0.9

Table 2: Within and cross-culture accuracy % (mean ± std. err.) for 10-class classification using wav2vec embeddings, mean over all results per country.

# Features	wc_wi	Accuracy	wi_pair_cc	Accuracy
1	loudness mean falling slope	17.2 ± 0.6	**F0 percentile range**	13.6 ± 0.3
2	MFCC 1 mean	20.0 ± 0.6	logRelF0-H1-A3 mean	16.2 ± 0.4
3	**F0 percentile range**	22.1 ± 0.7	HNRdBACF mean	17.8 ± 0.4
4	F1 logRel F0 std. dev.	23.2 ± 0.8	MFCC 2 mean	18.5 ± 0.5
5	loudness std. dev.	24.6 ± 0.8	alphaRatio std. dev.	18.8 ± 0.5
6	voiced segment rate	24.7 ± 0.8	shimmerLocaldB mean	19.2 ± 0.4
7	logRel F0-H1-H2 mean	24.9 ± 0.8	MFCC 4 mean	19.5 ± 0.4
8	loudness percentile range	25.5 ± 0.8	F0 median	19.7 ± 0.4
9	loudness mean rising slope	25.8 ± 0.8	voiced segment mean	19.7 ± 0.5
10	F2 logRel F0 std. dev.	$\mathbf{25.8 \pm 0.8}$	Hammarberg index std. dev.	$\mathbf{19.8 \pm 0.5}$
88		$\mathbf{25.6 \pm 0.8}$		$\mathbf{16.3 \pm 0.5}$
Random 10		$\mathbf{19.3 \pm 0.4}$		$\mathbf{14.2 \pm 0.2}$
10 from opposite column		$\mathbf{21.7 \pm 0.7}$		$\mathbf{15.7 \pm 0.5}$

Table 3: Top 10 features cumulatively selected with forward selection for within and cross-culture classification, accuracy % (mean ± std. err.) for each feature added from top to bottom.

ing distribution is more broad and hence the trained model is more general. As the number of training countries increases, the accuracy should approach that of the within-culture conditions.

The two-factor ANOVA results suggest there is no interaction between cross-intensity conditions and cross-culture conditions, except perhaps for eGeMAPS features. Additionally, the difference between within and cross-culture accuracy is similar for each intensity separately. This suggests that low intensity emotions are not relatively more difficult to recognise in a cross-culture setting than high intensity emotions, and emotional intensity does not significantly influence cross-cultural emotion recognition.

Feature importance results show that only a handful of features are necessary to attain high accuracy. Using only the 10 most important features yields similar or higher accuracy than using all 88 eGeMAPS features and the difference is more pronounced for cross-culture classification. This suggests using too many features hinders performance, although it may depend on the classifier used, and we only tested linear SVM for these results. The top 10 features for within-culture classification are mostly different than the top 10 features for cross-culture classification, even to the extent that using each set of features on the opposite task causes a significant drop in accuracy. The exception is *F0 percentile range*, present in both columns. The top features for within-culture classification should be those that are correlated with emotion class, while the top features for cross-culture classification should be those which correlate with emotion class but do not correlate with country and hence are robust to different countries. Different important features account for some difference in accuracy, since top features for one case yield lower accuracy when used for the other. However, top features for cross-culture recognition yield similar accuracy for both within and cross-culture recognition, supporting the idea they are predictive of emotion but robust to differences across countries.

6 Conclusion

In this study we quantify differences between within and cross-cultural accuracy for five English-speaking countries, for three emotional intensities. While emotional intensity has little effect, the important features for each case explain some differences in accuracy. In future we plan to investigate accuracy and feature importance for individual emotions, and also determine *unimportant* features for classification.

References

Carlos Busso, Sungbok Lee, and Shrikanth Narayanan. 2009. Analysis of Emotionally Salient Aspects of Fundamental Frequency for Emotion Detection. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(4):582–596.

Alan S. Cowen, Petri Laukka, Hillary Anger Elfenbein, Runjing Liu, and Dacher Keltner. 2019. The primacy of categories in the recognition of 12 emotions in speech prosody across two cultures. *Nature Human Behaviour*, 3(4):369–382.

Florian Eyben, Klaus R. Scherer, Björn W. Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y. Devillers, Julien Epps, Petri Laukka, Shrikanth S. Narayanan, and Khiet P. Truong. 2016. The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing. *IEEE Transactions on Affective Computing*, 7(2):190–202.

Aaron Keesing, Yun Sing Koh, and Michael Witbrock. 2021. Acoustic Features and Neural Representations for Categorical Emotion Recognition from Speech. In *Interspeech 2021*, pages 3415–3419. ISCA.

Petri Laukka, Hillary Anger Elfenbein, Wanda Chui, Nutankumar S. Thingujam, Frederick K. Iraki, Thomas Rockstuhl, and Jean Althoff. 2010. Presenting the VENEC corpus: Development of a cross-cultural corpus of vocal emotion expressions and a novel method of annotating emotion appraisals. In *Proceedings of the LREC 2010 Workshop on Corpora for Research on Emotion and Affect*, pages 53–57, Paris, France. European Language Resources Association.

Petri Laukka, Hillary Anger Elfenbein, Nutankumar S. Thingujam, Thomas Rockstuhl, Frederick K. Iraki, Wanda Chui, and Jean Althoff. 2016. The Expression and Recognition of Emotions in the Voice Across Five Nations: A Lens Model Analysis Based on Acoustic Features. *Journal of Personality*, 111(5):686–705.

Petri Laukka, Daniel Neiberg, and Hillary Anger Elfenbein. 2014. Evidence for cultural dialects in vocal emotion expression: Acoustic classification within and across five nations. *Emotion*, 14(3):445–449.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. 2017. Context-Dependent Sentiment Analysis in User-Generated Videos. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 873–883, Vancouver, Canada. Association for Computational Linguistics.

Archiki Prasad and Preethi Jyothi. 2020. How Accents Confound: Probing for Accent Information in End-to-End Speech Recognition Systems. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3739–3753, Online. Association for Computational Linguistics.

Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. 2019. Wav2vec: Unsupervised Pre-training for Speech Recognition. *arXiv preprint arXiv:1904.05862*.

Björn Schuller, Anton Batliner, Dino Seppi, Stefan Steidl, Thurid Vogt, Johannes Wagner, Laurence Devillers, Laurence Vidrascu, Noam Amir, and Loic Kessous. 2007. The relevance of feature type for the automatic classification of emotional user states: Low level descriptors and functionals. In *Eighth Annual Conference of the International Speech Communication Association*.

Björn Schuller, Stefan Steidl, Anton Batliner, Alessandro Vinciarelli, Klaus Scherer, Fabien Ringeval, Mohamed Chetouani, Felix Weninger, Florian Eyben, and Erik Marchi. 2013. The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism. In *14th Annual Conference of the International Speech Communication Association*, pages 148–152, Lyon, France.

Björn W. Schuller. 2018. Speech Emotion Recognition: Two Decades in a Nutshell, Benchmarks, and Ongoing Trends. *Commun. ACM*, 61(5):90–99.

Dimitrios Ververidis, Constantine Kotropoulos, and I. Pitas. 2004. Automatic emotional speech classification. In *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages I–593.