A Preliminary Study of Cross-lingual Emotion Recognition from Speech: Automatic Classification versus Human Perception

Je Hun Jeon, Duc Le, Rui Xia, Yang Liu

Department of Computer Science
The University of Texas at Dallas, Richardson, TX, USA
{jhjeon,duc,rx,yangl}@hlt.utdallas.edu

Abstract

The aim of this study is to investigate the effect of cross-lingual data on human perception and automatic classification of emotion from speech. We use four different databases from three languages (English, Chinese, and German) and two types (acted and improvised). For automatic classification, there is a significant degradation using cross-corpus than within-corpus setup. For human perception, we observe differences between native and non-native speakers when judging emotions for a language, and there is less performance loss in cross-language setup compared to automatic classification. In addition, we find that the automatic approaches work well in classifying the emotional activation category: positive and negative activated emotions, but are not good at classifying instances within the same activation category, which is different from the confusion patterns of the human perception experiment. This study provides insights to better understanding of cross-lingual human emotion perception and development of robust automatic emotion recognition systems.

Index Terms: emotion recognition, human perception, cross lingual

1. Introduction

Speech contains rich information beyond what is said, such as the speaker's emotion. Correctly recognizing emotions helps effective human-human communication and is also important when developing friendly human-computer interaction systems. For automatic emotion recognition, reasonable performance has been demonstrated for different data sets, though it is still questionable how robust the systems are, since they are typically evaluated using matched training/test conditions on specifically designed data sets. There have been some previous studies conducted across different languages/cultures. However, they are mostly on either human perception or automatic recognition of emotion. Therefore the language effect is especially not clear on how human perception and automatic recognition system performance differ.

Many theories about emotion perception assume that there are both universal and culture-specific cues to emotion. Elfenbein and Ambady [1] conducted a meta-analysis of 97 human perception experiments exploring cross-cultural recognition of emotion in both visual and auditory modes. Emotional stimuli (speech and visual expression) from one culture were presented to members and non-members of that culture. It is shown that cross-cultural emotion recognition accuracy using speech was better than chance performance, but lower than using facial expression and body language. For automatic emotion recognition from speech, there are a few studies conducted on cross-lingual corpora [2, 3, 4, 5, 6]. The experiment in [3] showed poor cross-data generalization performance when training was performed on one database and testing was on another language. They suggested this is because the same emotion occupies different regions of the feature space for different databases. Polzehl et al. [5] used feature selection in order to reduce the performance loss in cross-lingual experiments, but the improvement was not significant. Both studies showed that when merging the databases from different languages, classification results were comparable to within-corpus experiments. To better understand the high variance of cross-lingual experiments, Schuller et al. [6] evaluated the effect of different types of feature normalization and various emotion categories using six different corpora, and observed that as long as conditions remain similar, cross-corpus training and testing can work to a certain degree.

There is very little previous research that examines and compares the language effect on human perception and automatic recognition of emotion. In this paper, we perform a cross-lingual study with the aim to investigate the effect of language/culture and the difference between human perception and automatic classification of emotion from speech. We used data from three languages: Chinese, English, and German. For human perception experiments, we use two subject groups: native English and native Chinese speakers. For automatic recognition, we performed within- and cross-corpus experiments with feature normalization and selection. Our results show that, in general, automatic classification performs significantly better on within-corpus than cross-corpus conditions. For human perception, although there is some performance drop in cross-corpus/language experiments, the difference is much less than that in automatic classification. Additionally, in automatic cross-corpus/language experiments, we notice many errors within the same emotional activation categories, which is different from human perception. These results suggest that the current feature set focuses more on classifying the positive and negative activated emotions.

2. Selected Data

For the cross-lingual experiments, we used four different corpora in three languages: English, Chinese, and German. There are two corpora in English: one is acted speech, and the other one is improvised speech. The Chinese and German data are acted. For all these corpora, we use four common emotion categories ('*Angry*', '*Happy*', '*Neutral*', and '*Sad*'), and 40 utterances from each emotion category (20 from male, 20 from female speakers). The number of different speakers varies in these corpora. The following briefly describes these corpora.

- **Chinese: ACC Database (CA)** - This database from Chinese Corpus Consortium[1] covers eleven different emotion categories. For each emotion category, there are four different short stories that are designed as suitable scenarios for the emotion. We picked four sentences for each category, which are read by 5 male and 5 female speakers.
- **English: EMA Database (EA)** - Electromagnetic Articulography database (EMA) [7] contains a total of 680 utterances spoken in four emotions (anger, happiness, sadness, and neutrality). There are three native speakers of American English: two females and one male. The two female talkers produced 10 sentences, and the male produced 14 sentences (10 sentences overlap with those of the female speakers). Each sentence was repeated 5 times for each emotion. We randomly selected 20 instances for each emotion, by one female and male speaker.
- **German: EMO-DB Database (GA)** - The Berlin Emotional Speech Database (EMO-DB) [8] covers seven different emotion categories. It has 10 predefined emotionally neutral sentences. Five female and five male actors were asked to express each sentence in all seven emotional states. We randomly chose 20 instances for each of the four emotion categories (angry, happy, neutral, sad) for each gender. They are randomly selected from 5 male and 5 female speakers.
- **English: IEMOCAP Database (ES)** - The Interactive Emotional Dyadic Motion Capture English database (IEMOCAP) [9] contains approximately 12 hours of audiovisual data from five mixed gender pairs of actors. In this database, two techniques are used in order to elicit emotional displays: scripts and improvisation of hypothetical scenarios. In this study, we use randomly selected 20 instances for each emotion class from one improvisation session spoken by one female and male speaker. Again, this data set contains improvised speech, whereas the other three data sets above are all acted speech.

3. Methods

3.1. Automatic Classification

For the automatic recognition experiments, we used the acoustic feature set extracted using openSMILE [10]. There are 1,582 features in total, representing information such as loudness, MFCC, F0, jitter, shimmer, and various functions performed on them, as listed in Table 1. Details of these acoustic features are provided in [11].

Different classifiers have been used in previous research on emotion recognition. We choose the most frequently used classification algorithm: Support Vector Machines (SVM), implemented in WEKA [12]. A three-order polynomial kernel and multiclass (4-way classification) discrimination is used. To improve recognition performance, we normalize features to unit variance based on the sample mean and variance of the data set.[2] In cross-corpus experiments, we normalize each data set separately. Our experiments showed this is better than putting all the data together and then performing normalization.

Table 1: *Acoustic feature sets: 38 low-level descriptors and 21 functionals.*

Descriptors	Functionals
PCM loudness	position max./min.
MFCC [0-14]	arithmetic mean, std. deviation
log Mel Freq. Band [0-7]	skewness, kurtosis
line spectral pairs Frequency [0-7]	linear regression coeff. 1/2
F0	linear regression error Q/A
F0 Envelope	quartile 1/2/3
Voicing Prob.	quartile range 2-1/3-2/3-1
Jitter local	percentile 1/99
Jitter consec. frame pairs	percentile range 99-1
Shimmer local	up-level time 75/90

3.2. Human Perception

For the human listening experiments, we recruited 16 human subjects that are native English and Chinese speakers, 8 (4 male and 4 female) for each language. All the subjects are students from our university.

For each data set, we have 20 instances for each of the four emotions for each gender, i.e., 160 sound files in total. We designed a semi-random data distribution algorithm to distribute subsets of the 640 sound files to 16 annotators. Our design ensured that every sentence would be annotated by 6 different native and 6 different non-native speakers. For each annotator, we then duplicated 10% of the sentences from each language, in order to measure the annotators' internal consistencies. Our algorithm distributed the data roughly evenly among annotators speaking the same native languages. Because English sentences comprised 50% of our database, while Chinese and German sentences each comprised only 25% (there are two English data sets, and only one for Chinese and German), English native speakers would end up judging more than the others. On average, an English annotator would judge 396 sentences, and a non-English annotator would judge 330 sentences.

The annotation interface sequentially presented the assigned sentences to the corresponding user (sentences from their native language first, and then non-native ones). For each sentence, the annotator was asked to listen to the sound file, select a label which most accurately described the emotional state of the speaker, and choose a label denoting the certainty level of his/her selection. The available emotion labels were '*Happy*', '*Angry*', '*Sad*', '*Neutral*', and '*I don't know*'. The first 4 labels corresponded to the emotion categories in our database. The annotator was instructed to select the '*I don't know*' option if he/she could not possibly decide what the speaker's emotional state was. The available certainty labels were 'Very Certain', 'Certain', and 'Somewhat Certain', corresponding to decreasing certainty levels. The annotator was instructed to ignore the sentence's semantic meaning and to identify the underlying emotion based only on how the sentence was being uttered, i.e., acoustic cues.

[1] http://www.d-ear.com/CCC/corpora.htm

[2] We also evaluated other normalization methods, such as linear normalization and rank normalization, but found they did not outperform unit variance normalization.

4. Results

4.1. Automatic Classification

We first evaluate the cross-corpus effect on automatic classification. Table 2 shows the classification results. For cross-corpus setup, a separate training and test set were used. For within-corpus conditions using one data set (i.e., the diagonals in the table), we performed 10-fold cross-validation evaluation and averaged the results.

Table 2: *Automatic classification results (accuracy in %) of the within-corpus and cross-corpus setups.*

		test data			
		CA	**EA**	**ES**	**GA**
train data	**CA**	85.6	61.3	51.9	51.9
	EA	61.9	92.5	60.0	47.5
	ES	49.4	56.3	71.9	57.5
	GA	50.6	53.8	60.6	89.4

From Table 2 we can see that overall the within-corpus performance is much better than the cross-corpus conditions, which is as expected. Among the four databases, the results are generally worse for the improvised data (ES), and much better for the acted speech (the other three data sets).

For an analysis, we focus on the results on CA and EA data since we have human subjects for these two languages and they are both acted speech. Table 3 shows the confusion matrix for the within-corpus and cross-corpus setups for these two data sets. Compared to within-corpus results, the performance loss on cross-corpus experiments happens for all the emotion classes, especially within the same activation group (i.e., '*Angry*' and '*Happy*'; '*Neutral*' and '*Sad*'). If we consider the binary classification performance – whether it is activated emotion or not, rather than 4-way classification, the performance difference between within- and cross-corpus experiments is much smaller: 99% vs. 97% when tested on EA, and 94% vs. 91% on CA. Based on these results, we can infer that the currently used feature set is very good at capturing the emotional activation, but weak at other aspects such as emotional valence.

To analyze the feature effectiveness for the two languages, we performed feature selection for each corpus. We notice that the contributing features in the two data sets are quite different. For example, for 'Angry' class, 'logMelFreqBand' (MFB) features are more effective on EA data, while 'MFCC' and 'F0' features contribute more on CA. We also combined different corpora as training data and tested on individual data sets, and found that the increase of training data by adding other corpora does not yield performance gain, which is similar to findings in previous work [3, 5, 6]. This is mainly because of the inherent difference among corpora.

4.2. Human Perception

Table 4 summarizes the accuracies of human perception experiments for the two language populations. As we expected, each language group performs well on their own language data. Note that the native Chinese speakers are all fluent in English (they are university students); however, the native English speakers do not know Chinese. Compared with automatic classification (Table 2), the accuracies of human perception experiments for

Table 3: *Confusion matrices for automatic classification. Reference emotion classes are in the rows, and system outputs are in the columns.*

	A	**H**	**N**	**S**
A	0.90	0.10	0.00	0.00
H	0.05	0.85	0.08	0.03
N	0.00	0.13	0.80	0.08
S	0.00	0.00	0.05	0.95

(a) Within: train: CA, test: CA

	A	**H**	**N**	**S**
A	0.88	0.08	0.00	0.05
H	0.08	0.93	0.00	0.00
N	0.00	0.00	0.95	0.05
S	0.00	0.00	0.03	0.98

(b) Within: train: EA, test: EA

	A	**H**	**N**	**S**
A	0.50	0.50	0.00	0.00
H	0.18	0.65	0.00	0.18
N	0.00	0.18	0.63	0.20
S	0.00	0.03	0.23	0.75

(c) Cross: train: EA, test: CA

	A	**H**	**N**	**S**
A	0.68	0.25	0.03	0.05
H	0.35	0.65	0.00	0.00
N	0.00	0.00	0.75	0.25
S	0.00	0.05	0.08	0.88

(d) Cross: train: CA, test: EA

spontaneous speech (ES) are much better. This is because the annotators are more affected by the content in the utterances. For Chinese native speakers, their performance on ES is much better than EA. They perform similarly as English native speakers on ES, but much worse on EA. These suggest that for native Chinese speakers, understanding of the content helps determine emotion categories for ES data; however, as non-native speakers, they are not as good as native speakers in judging emotions just based on acoustic/prosodic cues in speech. Even though no participants knows German, both groups do quite well in the German data (GA), with slightly better performance by the English group, which is consistent with results in [3], but the difference between the two groups is much smaller in our study ([3] compared United States and Indonesia). The good performance in GA can be explained by the database itself – the instances in GA were filtered with human perception [8]; only the most confident utterances are kept in the database and are thus relatively easy, even for listeners who do not speak the language.

Table 4: *Human perception results. The values in parenthesis are the standard deviation of the annotators.*

		listener	
		Chinese	**English**
test data	CA	80.59 ($\pm$6.4)	67.78 ($\pm$6.5)
	EA	64.62 ($\pm$5.9)	77.04 ($\pm$7.2)
	ES	78.46 ($\pm$5.2)	77.98 ($\pm$4.2)
	GA	85.01 ($\pm$5.6)	88.20 ($\pm$8.2)

Table 5 shows the human annotator agreement using Fleiss' Kappa statistics. The English group showed relatively lower inter-annotator agreement rate. On the CA and EA data, as we expected, the interannotator agreement is higher for within-language than cross-language conditions. For the other two data, the difference of inter-annotator agreement between each group is relatively lower. These are consistent with human perception accuracies on different data.

We also computed confusion matrix results for human perception, shown in Table 6. Compared to automatic classification (Table 3), the error patterns are a bit different. The major differences on within-corpus/language experiments are that, in the human perception, (a) more '*Happy*' instances are labeled as '*Neutral*', (b) '*Neutral*' are labeled as '*Sad*'. These patterns are shown in both language groups. The first case (a) is more

Table 5: *Fleiss' Kappa statistics for inter-annotator agreement rate.*

	CA	EA	ES	GA
Chinese	0.714	0.514	0.752	0.734
English	0.473	0.581	0.696	0.725

Table 6: *Confusion matrices for human perception. Reference emotion classes are in the rows, and system outputs are in the columns.*

	A	H	N	S
A	0.92	0.02	0.06	0.01
H	0.05	0.53	0.40	0.01
N	0.01	0.00	0.86	0.12
S	0.00	0.00	0.07	0.93

(a) Chinese speakers testing on CA

	A	H	N	S
A	0.84	0.05	0.10	0.00
H	0.05	0.77	0.17	0.01
N	0.06	0.00	0.72	0.21
S	0.03	0.00	0.11	0.86

(b) English speakers testing on EA

	A	H	N	S
A	0.84	0.13	0.01	0.02
H	0.09	0.57	0.29	0.06
N	0.02	0.03	0.60	0.35
S	0.03	0.03	0.12	0.82

(C) English speakers testing on CA

	A	H	N	S
A	0.52	0.10	0.37	0.01
H	0.05	0.60	0.34	0.02
N	0.01	0.00	0.72	0.27
S	0.00	0.01	0.24	0.75

(d) Chinese speakers testing on EA

often in Chinese group, and the second case (b) is more in English group. The '*Angry*' and '*Sad*' categories are relatively less confusable classes on the within-corpus/language setup. For cross-language, we can still notice similar emotion confusions to within-language human perception experiments. In the English group, the mislabeled cases mentioned above are more often, and thus the overall performance is lower than within-language results; while the Chinese group made fewer errors in those emotion classes, but has another mislabeling, such as '*Angry*' to '*Neutral*'. The '*Sad*' class is relatively stable in both groups.

5. Conclusion and Future Work

In this paper, we investigate cross-lingual/corpus effect of emotion recognition from speech. We perform both human perception and automatic recognition on four different databases from three languages. Our results show similarities as well differences between human perception and automatic classification. In general, automatic classification performs much better on within-corpus than cross-corpus conditions, and its performance degrades more on cross-corpus setups compared to human perception. We also notice that automatic approaches work well in binary classification – classifying the positive and negative activated emotions regardless of within- or cross-corpus conditions; however, they are not good at classifying four emotions on cross-corpus experiments, suggesting the limitations of the current features used for emotion recognition or possible differences of cross cultural emotion cues.

In future work, for the human perception experiments, we plan to ask human subjects based on what factors they choose the emotion category, such as loudness, pitch, voice quality, speaking rate, etc. In addition, a larger scale study is needed to verify the findings using a larger size, more speakers, and a variety of data sets. In particular, for the automatic recognition experiments, we would like to conduct a speaker independent evaluation. Finally, using the analysis from the human perception results, we will develop more robust approaches to automatic emotion recognition in cross-language setups.

6. Acknowledgments

This work is supported by an award from the US Air Force Office of Scientific Research, FA9550-10-1-0388, and DARPA under Contract No. FA8750-13-2-0041. The views expressed in this paper are those of the authors and do not represent the funding agencies.

7. References

[1] H. A. Elfenbein and N. Ambady, "On the universality and cultural specificity of emotion recognition: a meta-analysis." *Psychological Bulletin*, vol. 128(2), pp. 203–235, 2002.

[2] F. Eyben, A. Batliner, B. Schuller, D. Seppi, and S. Steidl, "Cross-corpus classification of realistic emotions - some pilot experiments," in *Proc. of the Third International Workshop on EMOTION (satellite of LREC)*, 2010.

[3] M. Shami and W. Verhelst, "Automatic classification of expressiveness in speech: A multi-corpus study," *Speaker Classification II*, pp. 43–56, 2007.

[4] D. Neiberg, P. Laukka, and H. A. Elfenbein, "Intra-, inter-, and cross-cultural classification of vocal affect," in *Proc. of Speech Prosody*, 2011.

[5] T. Polzehl, A. Schmitt, and F. Metze, "Approaching multilingual emotion recognition from speech - on language dependency of acoustic/prosodic features for anger detection," *Proc. of the Fifth International Conference on Speech Prosody*, 2010.

[6] B. Schuller, B. Vlasenko, F. Eyben, M. Wllmer, A. Stuhlsatz, A. Wendemuth, and G. Rigoll, "Cross-corpus acoustic emotion recognition: Variances and strategies," *IEEE Transactions on Affective Computing*, vol. 1(2), pp. 119–130, 2010.

[7] S. Lee, S. Yildirim, A. Kazemzadeh, and S. Narayanan, "An articulatory study of emotional speech production," *Proc. of InterSpeech*, pp. 497–500, 2005.

[8] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A database of german emotional speech," *Proc. of Interspeech*, pp. 1517–1520, 2005.

[9] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Journal of Language Resources and Evaluation*, vol. 42 (4), pp. 335–359, 2008.

[10] F. Eyben, M. Wllmer, and B. Schuller, "openEAR - introducing the munich open-source emotion and affect recognition toolkit," *Proc. of ACII*, pp. 576–581, 2009.

[11] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Mller, and S. Narayanan, "The interspeech 2010 paralinguistic challenge," *Proc. of Interspeech*, pp. 2794–2797, 2010.

[12] I. Witten, E. Frank, L. Trigg, M. Hall, G. Holmes, and S. Cunningham, "Weka: Practical machine learning tools and techniques with java implementation," *ICONIP/ANZIIS/ANNES International Workshop*, pp. 192–196, 1999.