Empirical Study of Dimensional and Categorical Emotion Descriptors in Emotional Speech Perception

Rui Sun, Elliot Moore II
Georgia Institute of Technology
210 Technology Circle
Savannah, Georgia 31407

Abstract

The dynamic between speaker intent and listener perception is played out in the variation of acoustical cues by the speaker that must be interpreted by the listener to determine in an appropriate way. Emotion speech research must rely on either acted intent (i.e., an actor attempting to express an emotion) or listener perception (i.e., listening tests to assign emotional categories to non-acted data) to define ground truth labels for analysis. The emotion labels are described either using emotion dimension or emotion category. This study examines the two emotion characterization strategies dimension and category in communication of emotion embedded in speech as expressed through acted intent and the perception of emotion determined by a group of listeners. The results reveal that, without context information, intended emotion categories could be perceived by listeners with the averaged accuracy rate five times of chance in category. Also, the trend of listener ratings between emotion dimensions (valence/arousal) and emotional word categories was shown to be well correlated. Furthermore, while listeners confused the specific identity of certain emotional expressions, they were generally very accurate at identifying the intended affective space of the actor as determined by intended valence and arousal.

Introduction

Emotion is an essential part of everyday communication between the speaker and listener. Listeners use their perception of the intended emotion of a speaker to gauge what response (if any) is warranted given the specific information that has been communicated. This is a critical component of social interaction as a misinterpretation of emotion on the part of the listener could lead to an inappropriate or unnecessary response. A significant challenge for emotional research is in finding appropriate emotional labels to represent ground truth. One labeling method is to use the intended emotion as determined by the speaker as the basis for ground truth (Busso, Sungbok, and Narayanan 2009; Barra-Chicote et al. 2010; Sun, Moore, and Torres 2009) . In one sense, this method of labeling represents the ground truth for the speaker's intent, but may not represent a ground truth for the perception by a group of listeners. Alternatively, other researchers evaluate emotional speech through the use of listening tests where subjects are asked to evaluate a group of utterances to determine the emotional ground truth (Grimm et al. 2007; Sun and Moore 2011a; 2011b). However, in this case, the ground truth provided is representative of the listener's ground truth and not necessarily the speaker's intent.

Another challenge of emotion research exists in describing and modeling emotion labels. The two most prominent means of emotion characterization have relied on either a discrete lexicon of emotional words or a dimensional scale for estimating levels of affect generated by a speech. The discrete lexicon of emotional words convey specific meanings and intent with the most commonly studied being "happy", "sad", "angry", "disgust", "surprise" and "fear". The advantage of this framework for analysis is that it helps to establish a controlled vocabulary for creating objective assessments. However, research literature does not always agree on a set lexicon of emotional words with there being hundreds of potential emotional labels that can be assigned to a particular human experience. This lack of agreement has produced conflicting research results and difficulties in comparing one set of research results to another given differences in emotion labeling strategies (Cowie et al. 2001). An alternative method of modeling emotion is based on using a dimensional scale. The dimensional approach assumes that humans are in a perpetual affective state that can be represented mainly along two dimensions: (1) **Valence**. A dimension that represents the degree to which a particular stimulus is viewed as being pleasant or unpleasant (i.e., positive or negative). Valence may also be referred to as Evaluation. (2) **Arousal**. A dimension that represents the degree of activity that is generated by a particular stimulus. It relates to the intensity of an emotional experience and may range from energized or alert to calm or drowsy (i.e., active or passive). Arousal may also be referred to as Activation. In this study, being aware that more dimensions available (e.g., power, expectation, etc.), we utilize a square representation of the 2-dimensional valence and arousal (VA) space where the values are allowed to vary from [-1, +1] in both dimensions.

A topic of interest for emotional speech research is determining how many emotions a system should be able to recognize. Most research limits to this number to the most popular emotion types (“happy”, “angry”, “sad”, etc) as a means to reduce the number of emotion stimuli that need be processed (Dromey, Silveira, and Sandor 2005). Other research regulates emotional expression to be “negative” or “positive” as this is likely the most important determination a listener must make in determining what type of response is warranted (Busso, Sungbok, and Narayanan 2009). Either way, it is clear that a robust emotion recognition system must be able to handle various emotional strengths whether it attempts to identify the specific emotion being expressed or simply makes a determination of whether the emotion is expressed “positively” or “negatively”. Therefore, it’s of great interest to compare the performance of emotion recognition using the labeling strategies of dimension and category.

Additionally, Scherer (Scherer 2000) mentioned, it was of vital importance to establish whether the vocal changes produced by emotional and attitudinal factors were universal or varied over cultures and/or languages. Some research has suggested that culture can play a role in emotion recognition. For example, Dromey (Dromey, Silveira, and Sandor 2005) determined that native English speakers perform better in recognizing English emotional speech in six emotion categories than non-native speakers. A similar conclusion was delivered by Pell (Pell, Monetta, and Paulmann 2009) for native Spanish speakers recognizing five emotions with higher accuracy in Spanish than in English, German and Arabic. Therefore, this study takes the culture/language factor into consideration when investigating the emotion perception using two descriptors by using bilingual listener groups (English and Mandarin) for emotion perception.

This article examines the concept of speaker intent and listener perception in emotionally acted speech using two emotion description strategies: dimension and category. A set of acted speech data is subjected to listening tests and the analysis including dimensional and categorical emotional descriptors of speaker intent and listener perception. The preconceived notion of a set of emotion words based on a subject’s culture/language and gender is examined along with their emotional characterization of several speech utterances from the acted database. Two main research issues are addressed in this study:

1. How accurately can a set of listeners correctly identify 15 distinct emotional categories from an acted speech database without other forms of context (e.g., linguistic, visual, etc.)?
2. What consistencies (if any) do listeners show in their labeling of emotional speech using two separate labeling strategies (VA and 15 emotion categories)?

Data Preparation

Emotional speech database

The speech material used for this study is provided by the Emotional Prosody Speech and Transcripts (EPST) database (Liberman et al. 2002). The EPST database contains recordings of emotional and semantically neutral speech spoken by seven native speakers (4 females and 3 males) of standard American English. All the speakers are professional actors. Each actor read short (4-syllables) dates and numbers (e.g. “five hundred one” or “August thirteenth”) with the intent to express 15 different emotional categories (“neutral”, “disgust”, “panic”, “anxiety”, “hot anger”, “cold anger”, “despair”, “sadness”, “elation”, “happy”, “interest”, “boredom”, “shame”, “pride” and “contempt”) selected according to the Banse and Scherer’s study (Banse and Scherer 1996). The speech was recorded at a sampling frequency of 22.05 kHz with 2-channel interleaved 16-bit PCM format (down sampled to 16 kHz for this study). The duration of each utterance varied approximately from 1 to 2 seconds and each actor produced 153 speech utterances. The use of short and emotionally neutral words allows the acoustics to carry all of the emotional expression and interpretation. While it is in no way assumed that acted speech provides a complete picture of authentic emotion, the value of the EPST database is to provide researchers with control over large number of discrete categories of emotional expression (i.e., 15 emoitons) that can be grouped and studied for analysis. Speech data from four (2 males and 2 females) of the seven actors in the EPST database were used in this study.

Listening tests subjects

While the EPST database contained specific labels indicating the actor’s intent for emotional expression, it was necessary for this study to analyze the emotional perception of this data by a group of listeners. Twenty listeners (5 American females, 5 American males, 5 Chinese females and 5 Chinese males) were recruited (largely from a population of graduate students) to participate in the listening experiments for this study. All 10 of the American subjects spoke American English as a first language and possessed no significant bilingual capability. All 10 Chinese subjects spoke Mandarin as a primary language. The Chinese subjects were drawn from a population of graduate students who had been in the U.S. less than 4 years upon arriving from China. While capable of using American English for communication, they used Mandarin as their primary method of communication the majority of the time. All of the subjects were in the age range of 22-30 years old.

Experiments and Results

All of the testing involving the listening subjects was conducted in a closed room where ambient noise was kept at minimum. The testing software was developed as a graphical user interface (GUI) in MATLAB and used from a computer with high quality headphones. The listening subjects were directly involved in three experiments as follows:

Experiment I: Introduced the subjects to the concept of the Valence-Arousal (VA) model and established a rating profile for the listeners in relating a set of emotional words to the VA scale. The averaged duration time for this experiment was about 15 minutes per listener.

Experiment II: Listening test on emotion words choice. The listening subjects were asked to assign one of the fifteen emotional categories to the speech utterances. The purpose of this experiment was to re-label EPST data by assigning the perceived emotion category and evaluate emotion category recognition. The test duration for this experiment was about 35 minutes per listener. By repeating for four speakers, the total required time for experiment III was about 140 minutes.

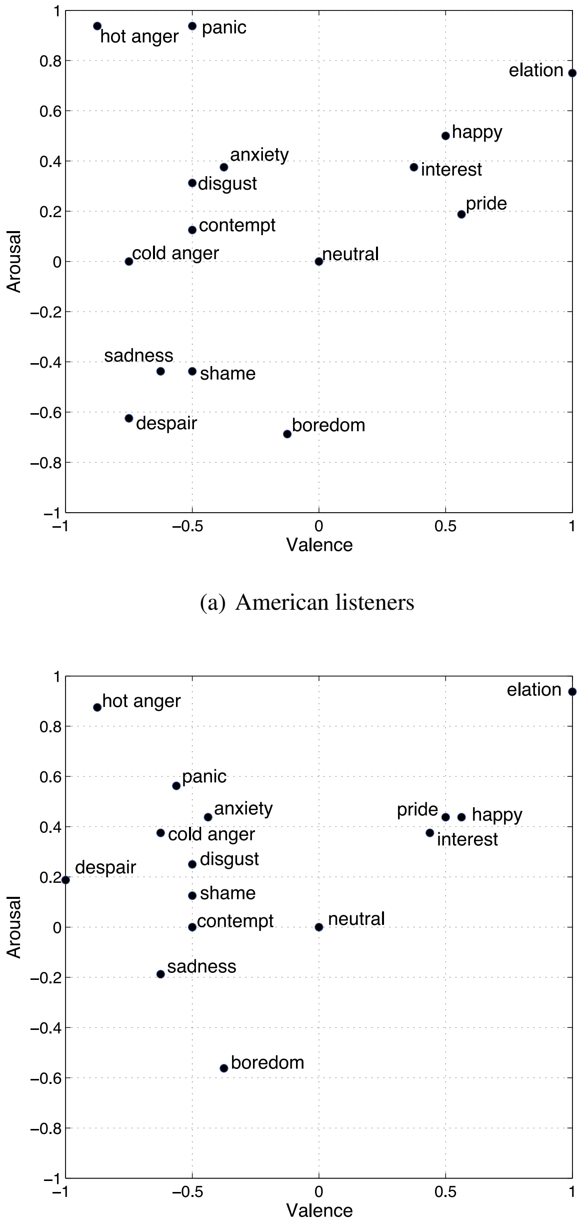
Experiment III: Listening test on VA degrees. The listening subjects were asked to assign a VA metric to each spoken utterance. The purpose of this experiment was to assign a set of valence and arousal metrics to the EPST data based on the listener perception and determine the degree to which the listeners could agree on the ratings. The test duration for this experiment was about 35 minutes per listener. By repeating for four speakers, the total required time for experiment II was about 140 minutes. The total duration of the whole listening test was about 5 hours, which was finished within 6-8 weeks for all listeners. To guarantee results, no more than two experiments were allowed in one day and a at least three hours break was required between two experiments.

Experiment I: Profile data on VA ratings

Before asking the volunteers to listen to speech utterances and provide emotional labels, it was necessary to: (1) ensure they understood the concepts of arousal and valence in the context of emotional expression and (2) establish a profile of the rating tendencies of each listener group (e.g., one group may consider the same emotional word as having a stronger/weaker implication than another rater). This experiment also provided an initial glimpse of differences that may occur due to culture. The test involved a graphical user interface designed in MATLAB that explained and gave examples of circumstances fitting certain types of high/low arousal and valence in order to explain the concepts. After this, the volunteers were asked to examine each of the 15 emotional labels (e.g., "happy", "angry", etc.) from the EPST database and assign VA values from -1 (Negative/ Passive) to 1 (Positive/ Active) (in increments of 0.125) based on their own perception of these words. The volunteers were given no time limit and were able to see their ratings for all of the words simultaneously so they could create relative associations on the VA scale among the emotional word choices (e.g., considering "elation" to be a higher arousal than "happy" even though both would be positive).

All of the raters considered "neutral" as an emotion with 0 valence and 0 arousal. It was therefore assumed that volunteers extended their concept of emotion from the base concept of "neutral" being a state of no emotional activity. The standard deviation of the listener ratings was analyzed to determine the amount of conceptual variation in the listeners' concept of particular emotional words in valence/arousal dimensions. An emotional word with less variation suggested a more consistent view among the listeners in regard to that particular emotion. For valence, "elation" showed the lowest variation (0.12) in listener ratings (all positive) while "hot anger" had the highest variation (0.46) in listener ratings (all negative). For arousal, the lowest variation occurs in "hot anger" (0.21) while the highest in "despair" (0.6).

Figure 1 shows the VA coordinates for each emotion category based on the median values by culture. The results of Figure 1 are relatively consistent with the expected locations of these emotions based on work by Cowie (Cowie et al. 2001), which implies that the listeners had a good foun-



(a) American listeners

(b) Chinese listeners

Figure 1: Distribution of emotion categories using median valence and arousal values for (a) American and (b) Chinese listeners collected as profile data (without acoustic stimuli).

Table 1: The recognition rate of emotional word choice in Experiment II based on groups (American: American listeners, Chinese: Chinese listeners).

	mean	American	Chinese
hot anger	0.70	0.66	0.74
interest	0.58	0.68	0.49
boredom	0.56	0.66	0.47
sadness	0.52	0.49	0.55
panic	0.51	0.56	0.45
neutral	0.48	0.55	0.42
elation	0.46	0.50	0.42
happy	0.44	0.52	0.36
anxiety	0.37	0.39	0.35
despair	0.36	0.37	0.34
pride	0.34	0.41	0.28
contempt	0.34	0.33	0.35
shame	0.32	0.30	0.35
disgust	0.29	0.27	0.32
cold anger	0.27	0.23	0.31
overall ave	0.44	0.46	0.41
baseline	0.07		

dation for participating in the listening tests of Experiments II and III. While Figure 1 gives the sense the listeners as a whole understood the concepts of arousal and valence as intended for relating to emotional words, it is also clear that some emotional words carried apparent differences across cultures (e.g., "panic", "shame", and "despair").

Experiment II: Listening test for emotion words

The second experiment involved all 20 volunteers in a listening test that required the listeners to assign each of the 153 emotional utterances per actor with one of the 15 emotional categories from the EPST database. The listeners were not given any information on the intended emotion of the utterances they were rating and each utterance was presented in random order. Listeners were encouraged to listen to the utterance as many times as they needed in order to make as accurate a rating as possible. Additionally, listeners were allowed to change their rating on an utterance at any time during the test. Table 1 shows the recognition results of listeners in selecting the intended emotions of speakers.

The recognition rate of emotion recognition shown in Table 1 represents the percentage of utterances whose intended emotion labels were correctly recognized. The numbers of utterances of fifteen emotions are approximately same (11 or 10 except that "neutral" has 7 per speaker per listener), so the recognition rate by chance is 0.07. From Table 1, "hot anger", the most selected emotion by all subgroups for all speakers, gains the highest overall accuracy rate 0.70 and "cold anger", all subgroups recognized poorly, has the lowest overall rate 0.27. The average recognition rate over all emotions for all listeners is 0.44, which is five times of chance. A further investigation by comparing the recognition accuracy among culture subgroups provides more details about the recognition pattern. The higher average recognition rate of American than Chinese listeners is not significant. However, for certain emotions, e.g., "boredom", "panic", and "neutral", the difference between listener groups could be obviously observed, which indicates the difference in emotion recognition between culture subgroups.

Experiment III: Listening tests for VA ratings

The third experiment involved all 20 subjects in a listening test similar to the second experiment except that that the listeners were required to assign ratings for valence and arousal based on their perception of 153 speech utterances from each of the four actors selected from the EPST database (totally 612 utterances for four actors). The purpose of this experiment was to examine the agreement among the listeners in assigning ratings for valence and arousal based on what they hear.

Since valence/arousal information is not inherently part of the EPST database, we used the valence/arousal trends (+ or -) as determined by the American subgroup of listeners in Figure 1(a) (since all of the actors were American speakers) to assign a positive (+) or (-) valence/arousal trend for each intended emotional expression for the actors. Table 2 show the percentage of listeners by culture that agreed with the broader trend (+/-) of the actors' intended valence and arousal, respectively, based on the emotional word. Since the choice of the listeners has now been reduced to their assignment of a positive (+) or negative (-) value, the baseline for chance had been increased from 0.07 (1 out of 15) in Table 1 to 0.5 (1 out of 2). The baseline increase in chance provided a natural increase in overall accuracy which made it more difficult to make direct comparisons of Table 2 to Table 1. However, the information in Table 2 is meant simply to quantify the types of emotions for which the listeners would be able to accurately assess an emotional space for the actor. The recognition rate (*RR*) for Table 2 was calculated in the following way as shown in Eq. 1

$$RR = \frac{m}{n * l}, \tag{1}$$

where *n* represents the number of utterances intended as a certain emotion category *i*, and *l* shows the number of listeners per subgroup. Therefore, *(n*l)* valence/arousal ratings per subgroup per actor would be obtained. Among the *(n*l)* ratings, *m* means the number of ratings having the same sign (+ or -) as that of emotion *i* in the profile data collected in Experiment I for each subgroup (Figure 1). Then, *m/(n*l)* represents the percentage of entries for which a correct assessment of the intended valence/arousal of the actor was made.

From Table 2, the intended valence and arousal trends across all actors are identified by the listeners at above chance (0.5). However, most of the valence/arousal trends of emotions are recognized at a rate greater than 70%, with several achieving greater than 90% recognition in the arousal domain. It's shown that the emotions of "pride" and "contempt" for both American and Chinese listener groups and "cold anger", "elation", "happy", and "interest" for Chinese listeners possessed the lowest recognition rates overall in identifying the actors' trends in valence. The lowest recog-

Table 2: The percentage of subjective ratings that agreed with the profile data assignments (negative (-) or non-negative (+)) in valence and arousal based on groups (AM: American listeners, CH: Chinese listeners).

	valence			arousal		
	Int	AM	CH	Int	AM	CH
disgust	-	0.83	0.76	+	0.6	0.56
panic	-	0.78	0.75	+	0.97	0.97
anxiety	-	0.76	0.73	+	0.58	0.66
hot anger	-	0.88	0.9	+	0.94	0.97
cold anger	-	0.71	0.66	+	0.8	0.85
despair	-	0.87	0.76	-	0.57	0.51
sadness	-	0.93	0.92	-	0.8	0.72
elation	+	0.78	0.66	+	0.97	0.98
happy	+	0.8	0.62	+	0.9	0.88
interest	+	0.72	0.59	+	0.81	0.79
boredom	-	0.81	0.83	-	0.79	0.77
shame	-	0.73	0.73	-	0.68	0.68
pride	+	0.68	0.53	+	0.78	0.79
contempt	-	0.69	0.53	+	0.69	0.65
average		0.78	0.71		0.78	0.77

nition of valence overall was for the Chinese subgroup ratings of "pride" and "contempt". Table 2 also shows that the emotions of "disgust", "anxiety", "despair", "shame", and "contempt" exhibited the lowest recognition rate (¡70%) for both American and Chinese listeners, with several actors having arousal recognized at or below chance. However, while arousal had fewer emotions that exhibited recognition rates greater than 70% when compared to valence, the arousal ratings for "panic", "hot angler", and "elation" were nearly perfect on almost all of the actors. As these emotions were considered to have high arousal ratings (from Figure 1) it was clear that the listeners were able to perceive the intended arousal strength as conveyed by the actors very clearly. Overall, both the valence and arousal trend for most emotions could be heard correctly while the American group seemed to be better at the recognition of valence. However, the difference is for valence, American listeners performed better than Chinese but for arousal, both groups exhibited similar performance.

Discussion

While the EPST database contains intended emotional word labels, it does not contain annotations for valence and arousal. The results of Experiments II and III provided each utterance with listener assignments of perceived valence/arousal values as well as a perceived emotional word. As seen in Table 2, the recognition of the intended emotional word as expressed by the actor exhibited a great deal of variation. Part of this is a result of the number of emotions being considered (15) with many of them apparently sharing a similar emotional space as indicated by Figure 1. However, one of the research issues of this article was also to assess whether listeners were able to show consistency in perceiving emotions using two emotion characterization strategies (i.e., dimension and category). Leading up to this determination, we proceeded on this investigation by evaluating the agreement among the listener subgroups in their emotional word assignments and their valence/arousal.

This first phase of this study involved combining all the listeners into their cultural subgroups (American or Chinese) and evaluating their consistency within a cultural group for specific emotional words assigned to the utterances. In this case, each utterance was considered to have 10 (one for each listener in a particular culture) sets of annotations (emotional word, valence/arousal) per subgroup. We collected all the valence/arousal ratings from each listener and grouped them based on the emotional words to which they were assigned. For the purposes of this specific investigation, a particular utterance was allowed to belong to more than one emotional category. For example, if a specific utterance was rated as "happy" by American listener L1 and "panic" by American listener L2, we used the valence/arousal ratings from listener L1 to represent an example of "happy" and the valence/arousal ratings from L2 to represent an example of "panic" for the American subgroup. The key here was to examine the listener's perception of the emotional word and subsequent valence/arousal ratings without regard (at this time) to what the actor intended to express. This was used to determine which emotions the listeners were most consistent on in regard to their perception over several sessions. Valence arousal was reduced to represent on the trend related to positive (+) or negative (-) without consideration of the degree of valence or arousal as indicated by a specific value.

Table 3 reflect the level of consistency among the subgroups in assigning valence and arousal trends (+/-, e.g., the trend of -0.2 is -), respectively, to their choices of emotional words (Experiment Test II) individually according to listeners' profile data (Experiment Test I). Tabler̃eftbl:cons were used to better quantify the amount of random assignments being made by the listener subgroups on a whole for specific emotions. It's shown that the valence assignments for 12 of the 14 emotions ("neutral" was excluded) were consistent with their associated trend (+ or -) in at least 75% of their observations for valence. The exceptions to this were "boredom" for the American subgroup and "pride" and "contempt" for the Chinese subgroup. For arousal, 8 of the 14 emotions were consistent with their associated trend (+ or -) in at least 71% of their observations. "Cold anger", "shame", and "contempt" exhibited less than 70% consistency for both culture groups. Overall, however, every emotion exhibited a consistency of at least 50% with their associated valence/arousal trends.

A closer examination of Tables 1-3 was conducted to better interpret the results of the listening tests. As shown in Table 1, "shame" and "cold anger" is perceived with the lower rate but relatively higher rate in Table 2, which suggested that listeners (generally) perceived the intended valence and arousal successfully but did not identify the discrete emotion. Also, "despair" was poorly recognized as shown in Table 1, but the valence trend was recognized with a high rate (see Table 2). However, a potential reason why the emotion could not be identified correctly may be explained by the

Table 3: The percentage of subjective ratings that agreed with the perceived emotion word assignments by listener subgroups (American: American listeners, Chinese: Chinese listeners). The **bold** shows emotion with high agreement (> 0.70) and the *italic* shows emotions with lowest agreement.

	valence		arousal	
	American	Chinese	American	Chinese
sadness	**0.90**	**0.93**	0.68	**0.70**
boredom	*0.58*	**0.82**	**0.77**	**0.82**
interest	**0.91**	**0.78**	**0.80**	**0.94**
anxiety	**0.79**	**0.78**	**0.79**	**0.79**
happy	**0.88**	**0.82**	**0.85**	**0.76**
hot anger	**0.80**	**0.93**	**0.97**	**1**
pride	**0.82**	0.68	**0.79**	**0.85**
contempt	**0.75**	*0.67*	0.58	*0.51*
despair	**0.93**	**0.90**	**0.84**	0.63
panic	**0.82**	**0.79**	**0.92**	**0.96**
disgust	**0.89**	**0.82**	**0.71**	0.68
cold anger	**0.87**	**0.79**	*0.50*	0.61
shame	**0.87**	**0.78**	0.69	0.58
elation	**0.89**	**0.83**	**0.98**	**0.99**
average	0.84	0.81	0.78	0.77

fact that the arousal trend was only recognized at a rate near chance. Emotions recognized correctly by most subgroups ("hot anger", "boredom", "sadness", "interest" and "panic" in Table 1) also show high recognition rate for both valence and arousal trends in Table 2. A similar explanation is possible for all of the recognition rates reported in Table 1 which supports the notion that the listeners were generally better at identifying the general emotion space of the actor as opposed to the specific intended emotion.

Conclusion

This article examined the correlation between the assignment of 15 distinct emotional categories and the labeling of a general affective space as intended by an actor using words with no inherent emotional implication. The results showed that, while listeners confused the specific identity of certain emotional expressions, they were generally very accurate at identifying the intended affective space of the actor as determined by intended valence and arousal. Since social interaction is largely based on a general "sense" of the speaker's intent, the results suggest that the listeners would understand the basic nature of the actor's intent (positive or negative) even if the specific intent was not identified accurately. Additionally, a comparison of American and Chinese listening groups suggested that the American group was better overall at recognizing valence when compared to the Chinese group of listeners while both performed equally well at the recognition of arousal. This suggested that, in the absence of linguistic context, the American group was able to utilize acoustic cues more effectively in determining the valence. As the acoustic cues are the only clue regarding an emotion annotation, a thorough acoustic analysis is underway to find features that will correlate better with the ratings provided by the listeners as the future work.

References

Banse, R., and Scherer, K. 1996. Acoustic profiles in vocal emotion expression. *Journal of ersonality and Social Psychology* 70:614–636.

Barra-Chicote, R.; Yamagishi, J.; King, S.; Montero, J. M.; and Macias-Guarasa, J. 2010. Analysis of statistical parametric and unit selection speech synthesis systems applied to emotional speech. *Speech Communication* 52(5):394–404.

Busso, C.; Sungbok, L.; and Narayanan, S. 2009. Analysis of emotionally salient aspects of fundamental frequency for emotion detection. *Audio, Speech, and Language Processing, IEEE Transactions on* 17(4):582–596.

Cowie, R.; Douglas-Cowie, E.; Tsapatsoulis, N.; Votsis, G.; Kollias, S.; Fellenz, W.; and Taylor, J. G. 2001. Emotion recognition in human-computer interaction. *Signal Processing Magazine, IEEE* 18(1):32–80.

Dromey, C.; Silveira, J.; and Sandor, P. 2005. Recognition of affective prosody by speakers of english as a first or foreign language. *Speech Communication* 47(3):351–359.

Grimm, M.; Kroschel, K.; Mower, E.; and Narayanan, S. 2007. Primitives-based evaluation and estimation of emotions in speech. *Speech Communication* 49(10-11):787–800. doi: DOI: 10.1016/j.specom.2007.01.010.

Liberman, M.; Davis, K.; Grossman, M.; Martey, N.; and Bell, J. 2002. Emotional prosody speech and transcripts. *http://www.ldc.upenn.edu/ Catalog/ CatalogEntry.jsp? catalogId= LDC2002S28*.

Pell, M. D.; Monetta, L.; and Paulmann, S. 2009. Recognizing emotions in a foreign language. *Journal of Nonverbal Behavior* 33(2):107–120.

Scherer, K. R. 2000. A cross-cultural investigation of emotion inferences from voice and speech: Implications for speech technology. In *Sixth International Conference on Spoken Language Processing*, 379–382.

Sun, R., and Moore, E. I. 2011a. Investigating acoustic cues in automatic detection of learners' emotion from auto tutor. In D'Mello, S.; Graesser, A.; Schuller, B.; and Martin, J.-C., eds., *the 4th Affective Computing and Intelligent Interaction*, volume 6975 of *Lecture Notes in Computer Science*, 91–100. Memphis, TN: Springer.

Sun, R., and Moore, E. I. 2011b. Investigating glottal parameters and teager energy operators in emotion recognition. In D'Mello, S.; Graesser, A.; Schuller, B.; and Martin, J.-C., eds., *the 4th Affective Computing and Intelligent Interaction*, volume 6975 of *Lecture Notes in Computer Science*, 425–434. Memphis, TN: Springer.

Sun, R.; Moore, E.; and Torres, J. 2009. Investigating glottal parameters for differentiating emotional categories with similar prosodics. In *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*.