Lacking Vision: Insights into the Automatic Classification of Emotion in AMCs The Walking Dead

Joanne Quinn

Montclair State University
Montclair, NJ
quinnj11@montclair.edu

Abstract

Speech Emotion Recognition (SER) is in huge demand in our high-tech world, but can SER detect emotion with near-human accuracy? To explore this question, we must first explore others: What is near-human accuracy in SER? And: How much is that accuracy influenced by visual prosody? This study consists of two parts: The first contrasts the difference in emotional perception in near-natural speech when audio is presented alone, then in conjunction with visual stimuli. The results create a baseline for human auditory SER, which is used to judge a basic automatic SER classification model using prosodic, semantic, and temporal features.

Index Terms: speech emotion recognition, prosody, emotion detection

1. Introduction

Emotions in speech have been studied over decades and across disciplines. This research has provided a basic framework for selecting features that differentiate among strong emotions [1]. However, emotion is present in different levels of language structure, so a one-size-fits-all guideline for feature selection may be an overly simplistic model. Since emotion is implied, rather than entailed, and shaped by both culture and cognition, emotional encoding and decoding are complex processes that are also heavily speaker and listener dependent [2]. While verbal prosodic cues are believed to have patterns that can communicate emotion within and across languages and cultures, visual prosodic cues are also believed to convey information about the speaker's emotional state and temper [3],[4].

With advances in computational ability, researchers in many fields have turned their attention to automatic speech recognition (SER). Many of the features previously identified have proven useful in automatic SER, however, an optimal feature set for automatic recognition has not yet been established [5]. Additionally, the corpora used for SER are not necessarily representative of realistic emotions. Past research has included work with the Switchboard and Fisher corpora, prescribed, acted corpora, and fixed utterances in stories designed to elicit specific emotions [6],[3],[1]. Schuller et al. notes that "the types of emotions that normally are prompted are definitely not the same as one would encounter in realistic scenarios", and Rakov and Rosenberg criticize that the language used among strangers is much more formal than language used among friends [7],[6].

In order to address both the role of visual prosody in the interpretation of emotion and the limitations of past corpora, the first half of this study focuses on assembling and evaluating a corpus of acted, yet authentic speech. The second half of the study explores the effectiveness of common prosodic and semantic features in the automatic classification of SER with Weka's RandomForest algorithm.

2. Motivation

Graf et al. studied visual prosody in the facial movements of speakers and concluded that it is identifiable in the speech of most people, and, while varying from person to person, it is strongly correlated with the prosodic structure of the text. Although Graf et al. focused mainly on visual prosody of the face and head (the scope of their research was confined to modeling authentic talking heads), they do note that body language is used to facilitate turn taking and emphasize point of view [4]. For this research, the definition of visual prosody is expanded to include any gesture or movement that the speaker performs whilst he/she is speaking.

Rakov and Rosenberg explored the use of sarcasm with clips from MTV's animated series *Daria*. *Daria* was chosen for this task because it is not a traditionally acted corpus and it is rich in sarcasm. Furthermore, the authors felt that the animated nature of the show would lend to a more exaggerated expression of sarcasm [6]. In keeping with the idea that scripted television more closely mirrors human emotion, the corpus for this experiment was collected in the same manner.

The main prosodic feature selection was inspired by prior work in SER. Many studies have emphasized the importance of fundamental frequency in the detection of emotion. Additional features such as formant values and formant bandwidth values have been judged useful in automatic SER [8]. Additionally features related to energy and speech rate are calculated [9],[5].

Finally, the semantic features selected in this study were inspired by the norms of valance, arousal, and dominance collected by Warriner et al. [10].

3. Materials

3.1. Corpus creation

The corpus for this project is comprised entirely of clips from AMC's *The Walking Dead*. Evaluating realistic emotion is the goal of this study, and *The Walking Dead* was chosen for this task because, despite its fantastical premise, it has been nominated for numerous awards. In 2015, episodes of *The Walking Dead* occupied all 10 chart spots for the top 10 most watched scripted cable television shows. These facts imply that both critics and the general public alike believe the acting to be quite realistic, even if the situation is not as believable.

In total, 152 clips were selected from seasons 1, 2, 4, and 5 and recorded as avi files in November 2015. The clips were selected for the target emotions: "Happy/playful", "neutral", "sad/upset", and "angry". Context was considered in the judgments, and subtitles were recorded on screen. The final breakdown of stimuli per emotion for the researcher-selected clips was (N: 152):

- Happy/Playful: 45
- Neutral: 13
- Sad/Upset: 47
- Angry: 47

3.2. Corpus evaluation

In January and February of 2016, 4 male and 4 female audio/visual raters were recruited to view and rate all 152 clips. The raters ranged in age from 20 to 37.

The clips were randomized and a Java program was written that allowed a rater to replay a clip until he/she decided on an emotional category. Participants were not provided with a definition or an example of the emotions.

After all evaluations were completed, the inter-rater reliability for each stimulus was calculated using Fleiss' Kappa. Any item that did not receive a kappa score of 0.31 or above was removed from consideration. Additionally, any item that was evenly judged between 2 or more categories was removed. The final breakdown of stimuli per emotion after the audio/visual rating phase was (N:134):

- Happy/Playful: 35
- Neutral: 34
- Sad/Upset: 38
- Angry: 27

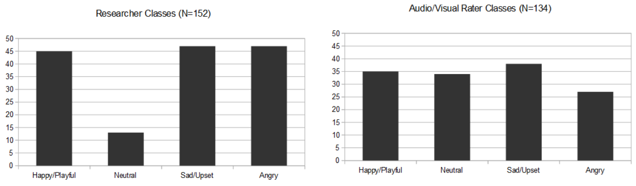

Figure 1: Researcher Selected Classes vs AV Rater Labeled Classes

4. Methods

4.1. Experiment 1

Between January and April, 2016, Amazon's Mechanical Turk, an online marketplace for human intelligence tasks, was used to survey the 28 audio-only participants. MP3 and ogg (audio) files were created for each of the 152 stimuli and text files were created with the transcribed audio of each clip. Participants were asked to listen to the audio while reading the text and then assign each clip to an emotional category. Again, participants were not provided with a definition or an example of the emotions. The survey took approximately 50 minutes to complete.

4.1.1. Evaluation

Although the audio-only raters were surveyed on all 152 stimuli from the original set, their results were pre-processed to remove the ambiguous clips and those that did not receive a kappa score of .031 or above in section 3.2.

Responses were judged on a binary scale, receiving a score of 1 if the correct category was identified and a score of 0 if an incorrect category was identified.

4.1.2. Results

Overall, the 28 audio-only respondents correctly identified the stimuli 59.68% of the time. Previous research by Banse and Scherer has reported accuracy rates of around 55% on similar tasks; additionally the authors propose a recognition rate of around 50% as a stable estimate of acoustic emotional recognition rate [11]. Below is the combined confusion matrix for all 28 audio-only respondents. Note that, while the majority class of the stimuli was "sad/upset", the majority of respondents selected "neutral" for every class, excepting that of "happy/playful".

a	b	c	d		← Classified As
5	10	16	4	a	= Happy/Playful
3	13	11	7	b	= Neutral
7	14	6	11	c	= Sad/Upset
7	9	6	5	d	= Angry

Table 1: Confusion Matrix for Audio Participants

The percent of correctly identified stimuli in this experiment was used as a benchmark to evaluate an automatic SER classifier.

4.2. Experiment 2

4.2.1. Feature selection

All stimuli in the corpus were manually annotated in Praat [12]. Values for the word and sentence level features described below were then automatically extracted using a series of Praat scripts. In the pre-processing step, audio was converted to mono, and each file was denoised with Praat's denoising feature. In total, each stimulus was represented by a 50 item feature vector, though some features were used for normalization, not classification. At this point, 5 additional stimuli were removed from the corpus because automatic extraction of features was unsuccessful. The final corpus contained 129 items labeled as such:

Category	Count	Percent
Happy/Playful	32	24.8
Neutral	33	25.6
Sad/Upset	37	28.7
Angry	27	20.9
Total	129	100

Table 2: Final Class Totals

The majority class was "sad/upset", comprising 28.7% of the data. No effort was made to ensure that speaker genders were evenly distributed. At this point, the data was split by speaker gender (two clips spoken by young children were categorized as "female"):

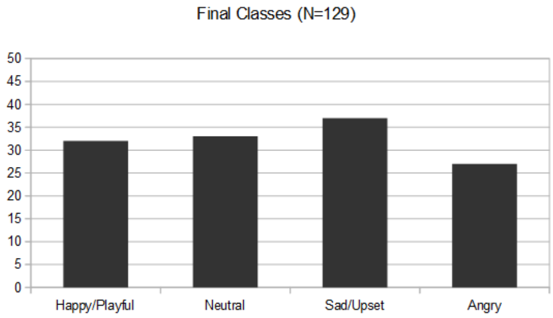

Figure 2: Final Classes for Automatic Evaluation

Gender	Count	Percent
Female	60	46.5
Male	69	53.5

Table 3: Speaker Gender

4.2.2. *Word-level and temporal features*

A Praat script was created to iterate through each word in each file, extracting the average measurement for that word. Word level features include:

- Average Word Length
- Speech Rate
- F0: Average, Minimum, Maximum
- Intensity: Average, Minimum, Maximum
- Changes in pitch and intensity were calculated from the ranges

4.2.3. *Sentence-level features*

Another Praat script was written to iterate through all files, extracting average measurements for each utterance. For all averages, log(10) was also calculated. Utterance features included:

- Average formant values and formant bandwidths for the first, second, and third formants
- Word Count
- F0: Average, Minimum, Maximum
- Intensity: Average, Minimum, Maximum
- Total Change in Pitch and Intensity
- Absolute value of change in Pitch and Intensity
- Average F0 and Intensity were separately normalized over the average for all utterances and the gender-average for all utterances

4.2.4. *Semantic features*

Bash scripting was used to assign overall affect scores to each utterance using the theory of arousal, valance, and dominance. Each word of each utterance was first lemmatized using the Stanford Core NLP Toolkit [13]. The Lemmas were then compared to a database of word norms collected by Warriner et al. [10]. Total raw scores for valence (the pleasantness of a stimulus), arousal (the intensity of emotion provoked by a stimulus), and dominance (the degree of control exerted by a stimulus) were calculated. Each of these values was then calculated as a percent over the total affect value of the utterance.

- Arousal/TotalAffect
- Valance/TotalAffect
- Dominance/TotalAffect

5. Results and discussion

For the classification task, in order to test the features and select the best subset of them, Weka's CorrelationAttributeEval was run on the full set of features [14]. Using the output of the algorithm, a total of 18 features were chosen for classification.

Selected Features
Speech Rate
log(10) F1, F2, and F3 Bandwidth
log(10) F2 and F3
Average Pitch Utterance over Average for all samples
Average Pitch Utterance over Average for gender
Log(10) Average Pitch
Max pitch over average
Change in Pitch and Intensity
Min and Max intensities over average intensity
Each affect value as a percent over all

Table 4: Final Selected Features

Weka's RandomForest classifier was most successful in classifying emotions into their labeled categories [15]. This result was surprising as none of the prior research consulted in this study made use of decision trees for classification. This classifier was run via the Weka GUI using 10-fold cross validation. The algorithm creates a collection of decision trees, which then vote for the most popular class. Overall, the RandomForest classifier attained an accuracy of 40.31%. Precision, recall, and F-measures for each class are reported in table 5 below.

Precision	Recall	F-Measure	Class
0.267	0.250	0.258	Happy/Playful
0.414	0.364	0.387	Neutral
0.417	0.541	0.471	Sad/Upset
0.545	0.444	0.490	Angry

Table 5: Precision, Recall, and F-Measures by Class

The confusion matrix in table 6 shows the algorithm's correctly and incorrectly classified instances for each emotion. A comparison of this matrix to the one in Table 1 above, reveals that the algorithm correctly classified every class except "neutral" more accurately than did the human, audio-only respondents. (The audio-only respondents correctly identified "neutral" utterances 38.2% of the time, whereas the algorithm achieved a correct classification 36.4% of the time; however, the audio-only respondents also incorrectly classified most "sad/upset" and "angry" utterances as "neutral".) One possible explanation for the audio-only respondents majority "neutral" classification is that they may have defaulted to this class when they were unsure of which other class to select.

A problem for both the automatic SER and the audio-only respondents was correctly classifying the "happy/playful" emotions. Both experiments have a large number of "happy/playful"

a	b	c	d		← Classified As
8	6	13	5	a	= Happy/Playful
10	12	9	2	b	= Neutral
7	7	20	3	c	= Sad/Upset
5	4	6	12	d	= Angry

Table 6: Confusion Matrix for Classification

items classified as “sad/upset”. There are three possible reasons for this:

The first reason is the actual construction of the classes. While minimization of classes was intentional in this research, it may have benefited the outcome to have more fine-grained class definitions. For example, the class of “sad/upset” comprises both the low arousal, sad, and the high arousal, upset. Splitting this class into two separate classes and providing a high arousal and low arousal option for the “happy/playful” class (e.g. “contentment” and “excitement”) may solve some of the classification errors. It is possible that people heard excitement or laughter and confused the emotion with being upset or crying.

Another possible reason for these classification errors is that, in the hopes of experimenting on near-natural speech, exaggerated speech emotion was avoided, making the classification task harder than it would be on a traditionally acted corpus.

Finally, visual prosody also may have played a role in the incorrect classification of the “happy/playful” stimuli, especially in regard to the playful stimuli. Sometimes, when one person is joking with another person, that playfulness may be conveyed with a smile or a wink. While the audio-visual participants were able to evaluate the utterance in context with the visual prosody, these visual cues were unavailable to the audio-only respondents who may have mistaken the playfulness for another emotion.

While the overall automatic classification rate is still rather low and did not attain the benchmark goal of 59.68%, it does represent an increase of 11.6% over a majority-class baseline, while also correctly identifying 3 of the 4 classes more frequently than the audio-only respondents did. Additionally, it is important to consider that the rate was attained without the use of short-term spectral features such as linear prediction cepstrum coefficients (LPCC) and mel-frequency cepstrum coefficients (MFCC), which have both shown success in speech recognition algorithms.

6. Conclusion

This paper aims to advance the field of automatic SER by first providing a baseline of how well humans can identify emotion without visual prosody in near-human, acted speech. The creation and categorization of the new *The Walking Dead* Emotion Corpus was crucial in the task of evaluating human competency in SER. The automatic classification of that corpus based only on prosodic and semantic features not only provides a glimpse into which features are important for automatic SER, but it also highlights an important classification method, decision trees, which may deserve more consideration in future work. Improvements to this corpus could be made with the addition and evaluation of more data and the partitioning of classes into high arousal/low arousal variants of each emotion. Improvements on the automatic SER classification rates may be achieved with the addition of the short-term spectral features (such as MFCC), which were discussed earlier.

7. Acknowledgments

I would like to thank Dr. Jonathan Howell, Dr. Anna Feldman, and Dr. Jin Peng of Montclair State University for their helpful guidance and suggestions, all the participants who donated their time to make this study possible, and the reviewers who offered thoughtful, thought-provoking feedback.

8. References

[1] Sobin, C. and Alpert, M., “Emotion in speech: The acoustic attributes of fear, anger, sadness, and joy.”, Journal of Psycholinguistic Research, 28: 347-65, 1999.

[2] Majid, A., “Current Emotion Research in the Language Sciences.”, Emotion Review, 4: 432-443., 2012.

[3] Baum, K.M. and Nowicki, S.N., “ Perception of emotion: measuring decoding accuracy of adult prosodic cues varying in intensity.”, Journal of Nonverbal Behavior, 22: 89-106., 1998.

[4] Graf, H.P., Cosatto, E., Strom, V., and Huang, F.J., “Visual prosody: Facial movements accompanying speech.”, Paper presented at the 5th International Conference on Automatic Face and Gesture Recognition, Washington, DC., 2012.

[5] Vogt, T., Andre, E., and Bee, N., “Emovoice - a framework for online recognition of emotions from voice.”, OC. of an IEEE Tutorial and Research Workshop on Perception and Interactive Technologies for Speech-Based Systems., 5078, 188-99., 2008.

[6] Rakov, R. and Rosenberg, A., “Sure, I Did The Right Thing: A System for Sarcasm Detection in Speech.”, Proceedings of the 14th Annual Conference of the International Speech Communication Association, Lyon, France, 2529 August 2013.

[7] Schuller, B., Batliner, A., Steidl, S., and Seppi, D., “Recognising realistic emotions and affect in speech: state of the art and lessons learnt from the first challenge.”, Speech Communication, 53(9/10), 10621087., 2010.

[8] Petrushin, V.A., “Motional Recognition in Speech Signal: Experimental Study, Development, and Application.”, ICSLP-2000, 2: 222-5., 2000.

[9] Joshi, A., and Kaur, R., “A Study of speech emotion recognition methods.”, Int. J. Comput. Sci. Mob. Comput.(IJCSMC), 4: 28-31., 2013.

[10] Warriner, A.B., Kuperman, V. and Brysbaert, M., “Norms of valence, arousal, and dominance for 13,915 english lemmas.”, Behavior Research Methods, 44(4)., 2013.

[11] Banse R. and Scherer, K., “Acoustic profiles in vocal emotion expression.”, Personality Social Psych, 70(3): 614-636., 1996.

[12] Boersma, P., “Praat, a system for doing phonetics by computer.”, Glot International, 5(9/10):341-345. 2001.

[13] Manning, C.D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S.J., and McClosky D., “The Stanford CoreNLP Natural Language Processing Toolkit.”, Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, 55-60. 2014.

[14] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I.H., “The WEKA Data Mining Software: An Update.”, SIGKDD Explorations, 11(1)., 2009.

[15] Breiman, Leo., “Random Forests.”, Machine Learning, 45(1):5-32. 2001.