INVESTIGATING GLOTTAL PARAMETERS FOR DIFFERENTIATING EMOTIONAL CATEGORIES WITH SIMILAR PROSODICS

Rui Sun, Elliot Moore II, Juan F. Torres

Georgia Institute of Technology
School of Electrical and Computer Engineering
210 Technology Circle, Savannah, GA, 31407

ABSTRACT

Speech prosodics (i.e., pitch, energy, etc.) play an important role in the interpretation of emotional expression. However, certain pairs of emotions can be difficult to discriminate due to similar displayed tendencies in prosodic statistics. The purpose of this paper is to target speaker dependent expressions of emotional pairs that share statistically similar prosodic information and investigate a set of glottal features for their ability to find measurable differences in these expressions. Evaluation is based on acted emotional utterances from the Emotional Prosody and Speech Transcript (EPST) database. While it is in no way assumed that acted speech provides a complete picture of authentic emotion, the value of this information is that the actors adjusted their voice quality to fit their perception of different emotions. Results show statistically significant differences ($p < 0.05$) in at least one glottal feature for all 30 emotion pairs where prosodic features did not show a significant difference. In addition, the use of single glottal features reduced classification error for 24 emotion pairs in comparison to pitch or energy.

Index Terms— Speech, Affect, Emotion, Glottal, Prosodics, Pitch

1. INTRODUCTION

Fundamentally, automated emotion detection is the attempt to quantify an abstract interpretation into objectively measured components of recorded human interaction. A review of the study of emotion for human computer interaction in [1] shows that prosodics (e.g., pitch, energy, speaking rate, etc.) are the most common form of speech analysis in literature. Additionally, [1] shows support that the general prosodic tendencies in distinguishing between different emotion categories can be extremely qualitative, subtle, and likely speaker dependent. For example, a person who is happy may tend to raise their prosody (e.g., increased pitch, energy, speaking rate, etc.) from their neutral state but may also show similar tendencies when expressing anger or panic. Work on the use of glottal features (i.e., features extracted from the estimated signal representing the air-flow through the vocal folds) in classifying emotion [2, 3, 4] has shown that these features can provide valuable insight into distinguishing different types of emotional expression. The purpose of this paper is to target speaker dependent expressions of emotional pairs that share statistically similar prosodic information and investigate a set of glottal features for their ability to find measurable differences in these expressions.

This work was supported in part by the National Science Foundation (Grant No. 0545772).

2. DATABASE

The speech used for this study was provided by the Emotional Prosody Speech and Transcripts (EPST)[5] database. The EPST database contains recordings of emotional expression on semantically neutral speech from 7 professional actors (4 females and 3 males) who are native speakers of standard American English. Each actor reads short (4-syllables) dates and numbers in 15 different emotional categories [6] (*"neutral", "disgust", "panic", "anxiety", "hot anger", "cold anger", "despair", "sadness", "elation", "happy", "interest","boredom", "shame", "pride", "contempt"*). The speech was recorded at a sampling frequency of 22.05 KHz with 2-channel interleaved 16-bit PCM format. The duration of each utterance varied from 1sec to 2sec. While it is in no way assumed that acted speech provides a complete picture of authentic emotion, the value of this information is that the actors adjusted their speech patterns to fit their *perception* of different emotions. These voice changes are objectively evaluated at this time without the need to explicitly determine the degree to which each utterance represents the intended emotion to an observer.

3. OBJECTIVE MEASURES

Pitch represents a high-level view of the motion of the vocal folds as it provides information on the rate at which air from the lungs is allowed into the vocal tract. The glottal waveform, on the other hand, provides a representation of the vol-

ume velocity of airflow through the vocal folds during voiced speech. While pitch information provides the rate, glottal features ideally provide a more detailed look at the phonatory process. This paper used prosodic features of speech based on the mean pitch and energy. Pitch was obtained using the RAPT pitch estimation algorithm in VOICEBOX[7] using a 10 ms frame rate. Energy was calculated as the squared sum of the values within each frame across the voiced sections in each utterance as indicated by the pitch information. The glottal waveform provides a representation of the shaping of the volume velocity of airflow *through* the vocal folds during voiced speech. The extraction of the glottal features for each speech utterance was processed in four steps: (1) each utterance was divided into frames 4 pitch periods long for feature extraction purposes (2) glottal closure instants (GCI's) were obtained using the DYPSA algorithm [8] on each frame (3) glottal waveform estimates were obtained for each frame using the Rank-Based Glottal Quality Assessment (RBGQA) algorithm [9], which iterates around approximate locations of GCI's to find the optimal analysis window position for deconvolution via the covariance method of linear predictive analysis (LPA) (for simplicity, an LPA order of 16 was used for all speakers) (4) for each frame, the 7 glottal features listed in Table 1 were computed using version 0.3.1 of the APARAT toolbox [10]. All features were quantified using only 1^{st} order statistics (i.e, the mean) across all frames of an utterance. The use of higher order statistics was excluded from the study at this time as the goal was to study the basic discriminatory power of the features themselves and not to build a complex model for general classification.

4. METHODOLOGY

Because of the high number of discrete emotion categories, most research on emotion has focused on smaller subsets of emotion (such as happy, anger, fear, etc.). However, in this paper a pairwise comparison is conducted on 14 distinct emotional categories in an effort to identify which emotional pairs statistically share the same prosody information. Four actors (2 females (F1, F2) and 2 males (M1, M2)) were chosen from the EPST database based on the speakers with the highest total number of observations (i.e., utterances). Pitch, energy, and glottal features were extracted on a speaker-dependent basis as described earlier. The feature extraction procedure resulted in 13644 frames with a 9 dimensional feature vector (i.e., mean pitch, energy, and glottal features). The average number of frames per speaker was 3411 with an average of 227 frames per emotion. There were approximately 25 utterances per emotion for each speaker on average with no emotion allowed to have less than 20 utterances for inclusion in the study (this resulted in the exclusion of *neutral* utterances).

Table 1. Glottal Waveform Parameters

clq	Closing quotient
dh12	Difference between 1^{st} and 2^{nd} glottal formants, in dB
hrf	Harmonic richness factor
naq	Normalized amplitude quotient
oq	Open quotient
oqa	Open quotient, derived from the LF model
sq	Speed quotient

The purpose of this study was to evaluate the discrimination power of glottal features on emotional categories that share statistically similar prosodics. Therefore, the mean pitch values of each of the pairwise groups of emotions (91 pairs total) was subjected to a Kruskal-Wallis (KW) significance test. Pairwise groups that showed no statistical difference in their pitch distributions at a significance level of $p < 0.05$ were targeted for further analysis. The discrimination of these emotional pairs was then evaluated by finding the error rates from using each of the 9 single features as classifiers and the error rates from using a Sequential Feature Selection (SFS) algorithm for selecting any combination of features for classification. SFS starts with an empty feature set and sequentially adds features that have not yet been selected. Every feature combination set is evaluated 10-fold cross validation until there is no improvement in the criterion function. For this study, the criterion was set to the error rate from a quadratic discriminant computed as the number of incorrect classifications divided by the total number of observations. The SFS algorithm added features in an effort to reduce the error rate as much as possible.

5. RESULTS

Table 2 shows the emotional pairs that showed no significant difference ($p < 0.05$) in their pitch distributions after the Kruskal-Wallis test on a speaker-dependent basis. Intuitively, many of the emotional pairs reflect an expected similarity in prosodic tendencies. For example, many of the pairs reflect a confusion between two high (such as *elation* and *hot anger* for speaker F1) or low arousal states (such as *pride* and *sadness* for speaker M1). Additionally, there is very little overlap in the confused emotional states across actors, which reflects the highly speaker dependent nature of emotional interpretation and expression. For all of the emotion pairs listed in Table 2, at least one glottal feature showed a statistically significant difference and 19 out of 30 pairs had 4 or more of the 7 glottal features show statistical significance. Further evaluation was conducted by finding the error rate (ER) for each of the individual features in discriminating the emotional pairs using 10-Fold cross validation. The error rate was computed as the number of incorrect classifications divided by the total number of observations (i.e., utterances). The number of observations was approximately equal for each of the emotional pairs, making the chance error rate roughly equal to

50%. Due to the relatively small number of observations, the 10-Fold cross validation was repeated 50 times, where each iteration randomized the data in a way ensure that enough variations on the combinations of data observations for training and testing were used. Table 2 shows the mean of the error rate computed across all 50 runs of the 10-fold cross-validation. Only the best performing glottal feature is shown in the table. A lower error rate is achieved by a glottal feature in 24 out of the 30 pairs. Of the the 6 pairs where a glottal feature is not the best feature, energy has the lowest error rate in 4 pairs and pitch has the lowest error rate in 2 pairs. That pitch could have the lowest error rate (though slight) even though there was no statistically significant difference highlights the reasons for evaluating the classification performance of each feature.

Table 3 shows the resulting mean error rates from the SFS procedure along with the percentage change from the lowest error rate achieved for a single feature. The SFS was run on each subset of the 9 features 50 times using 10-fold cross validation to ensure enough randomization of training and testing combinations in the data. The '%Change' column indicates the percentage change in the error rate that resulted from using multiple features over the single feature with the lowest error rate shown in Table 2. Table 3 shows that pitch and energy continue to play an important role in emotional classification even when the emotional pairs are selected based on non-significant differences in pitch distributions. In only one instance (M1, $contempt, sadness$) was neither pitch nor energy selected for the classifier. For the females, 'hrf' feature was among the most prominent glottal features selected while for the males the 'oqa', 'clq', and 'naq' were the most prominent glottal features. The 'dh12' feature was a prominent feature across all speakers while the 'sq' feature showed little impact on discrimination and was rarely chosen. While the discrimination for most emotional pairs was greatly improved through the multiple feature classifier, the discrimination for speaker F2 with the emotional pairs $(sadness, disgust)$, $(happy, panic)$ and $(happy, elation)$ could not be improved over the performance of the single features of energy and harmonic richness factor, respectively.

Table 2. Minimum error rate (ER) for emotional pairs using single features.(A=Actor);(Pch=Pitch);(Eng=energy)

A	Emotional Pairs	Pch	Eng	(Glottal, ER)
F1	$pride, anxiety$	0.46	0.24	(hrf,0.16)
	$elation, hotanger$	0.48	0.49	(hrf,0.21)
	$boredom, coldanger$	0.52	0.33	(oq,0.21)
	$contempt, coldanger$	0.43	0.47	(dh12,0.29)
	$happy, sadness$	0.36	0.38	(oqa,0.40)
	$interest, sadness$	0.20	0.29	(hrf,0.23)
	$pride, interest$	0.51	0.36	(hrf,0.13)
F2	$coldanger, disgust$	0.32	0.15	(hrf,0.30)
	$sadness, disgust$	0.32	0.23	(hrf,0.32)
	$despair, panic$	0.38	0.20	(oq,0.05)
	$happy, panic$	0.41	0.08	(hrf,0.06)
	$despair, hotanger$	0.40	0.34	(oq,0.21)
	$elation, hotanger$	0.32	0.57	(dh12,0.16)
	$happy, hotanger$	0.35	0.32	(oq,0.23)
	$sadness, coldanger$	0.36	0.39	(hrf,0.30)
	$elation, despair$	0.55	0.31	(oq,0.08)
	$contempt, sadness$	0.40	0.41	(oq,0.30)
	$happy, elation$	0.45	0.26	(hrf,0.08)
	$contempt, boredom$	0.40	0.36	(hrf,0.18)
M1	$shame, anxiety$	0.38	0.51	(oqa,0.32)
	$elation, coldanger$	0.46	0.38	(clq,0.30)
	$interest, coldanger$	0.46	0.40	(naq,0.31)
	$pride, sadness$	0.70	0.36	(dh12,0.30)
	$contempt, sadness$	0.39	0.60	(naq,0.25)
M2	$shame, disgust$	0.62	0.16	(oqa,0.34)
	$happy, panic$	0.68	0.47	(naq,0.28)
	$despair, anxiety$	0.40	0.43	(oqa,0.22)
	$contempt, anxiety$	0.51	0.36	(clq,0.37)
	$interest, coldanger$	0.45	0.46	(oqa,0.07)
	$contempt, despair$	0.40	0.42	(oqa,0.17)

6. CONCLUSION

The results highlight a few critical points about emotion in speech. The first confirms that there are emotional pairs that carry subtle differences that can be difficult to express and interpret based on prosody alone. Additionally, while the types of emotional ambiguities are largely speaker dependent, there are subtleties that can exploited from features of the glottal flow to help resolve some of them. The presented work examined the ambiguities present in an actors' *intended* emotional expressions. Future work for this continuing study will involve subjective tests for assessing the emotional categories into which the actors utterances are *interpreted*. Additionally, more complex classifiers will be designed to consider higher order statistics as well as additional prosodic and glottal features.

7. REFERENCES

[1] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor, "Emotion recognition in human-computer interaction," *IEEE Signal Processing Mag.*, pp. 33–80, 2001.

[2] K. Cummings and M. Clements, "Analysis of the glottal excitation of emotionally stressed speech," *J. Acoust. Soc. Am.*, vol. 98, no. 1, pp. 88–98, 1995.

[3] R. Fernandez and R. Picard, "Classical and novel discriminant features for affect recognition from speech," in *INTERSPEECH*, 2005, pp. 473–476.

Table 3. Minimum error rate (ER) for emotional pairs using SFS and the top five selected features

Actor	Emotional Pair	ER	%Change	(Feature, Selection Percentage(%))				
F1	$pride, anxiety$	0.10	-40%	(pitch,94),	(hrf,92),	(sq,6),	(dh12,4)	
	$elation, hotanger$	0.19	-7%	(hrf,100),	(sq,14),	(eng,10),	(naq,8),	(oq,4)
	$boredom, coldanger$	0.20	-3%	(oq,100),	(eng,20),	(dh12,6),	(hrf,6)	
	$contempt, coldanger$	0.17	-36%	(eng,96),	(dh12,92),	(hrf,34),	(oq,8),	(sq,8)
	$happy, sadness$	0.16	-54%	(pitch,98),	(naq,60),	(oqa,60),	(clq,42),	(hrf,38)
	$interest, sadness$	0.08	-59%	(pitch,96),	(oq,74),	(hrf,72),	(dh12,28),	(clq,16)
	$pride, interest$	0.09	-31%	(hrf,98),	(pitch,76),	(clq,40),	(oqa,10),	(dh12,4)
F2	$coldanger, disgust$	0.03	-80%	(pitch,100),	(eng,100),	(oq,44),	(naq,30),	(hrf,14)
	$sadness, disgust$	0.23	0%	(eng,100),	(pitch,6),	(clq,4),	(oq,4),	(oqa,4)
	$despair, panic$	0.04	-16%	(oq,70),	(eng,26),	(hrf,26),	(dh12,22),	(pitch,16)
	$happy, panic$	0.06	0%	(hrf,76),	(dh12,18),	(eng,4),	(oqa,2),	(sq,2)
	$despair, hotanger$	0.17	-20%	(hrf,74),	(oqa,36),	(oq,32),	(eng,24),	(pitch,22)
	$elation, hotanger$	0.09	-41%	(pitch,100),	(dh12,100),	(hrf,2),	(sq,2)	
	$happy, hotanger$	0.18	-23%	(oq,92),	(naq,60),	(oqa,48),	(eng,32),	(pitch,26)
	$sadness, coldanger$	0.13	-58%	(hrf,96),	(oq,90),	(oqa,70),	(naq,70),	(eng,46)
	$elation, despair$	0.04	-44%	(pitch,64),	(clq,60),	(dh12,30),	(eng,24),	(hrf,6)
	$contempt, sadness$	0.12	-60%	(oq,100),	(oq,76),	(eng,68),	(hrf,58),	(pitch,44)
	$happy, elation$	0.08	0%	(hrf,98),	(naq,98),	(pitch,6),	(dh12,2)	
	$contempt, boredom$	0.18	-3%	(hrf,84),	(dh12,18),	(sq,10),	(oqa,8),	(eng,4)
M1	$shame, anxiety$	0.28	-14%	(oqa,84),	(pitch,36),	(clq,32),	(naq24),	(hrf,18)
	$elation, coldanger$	0.23	-25%	(clq,100),	(pitch,90),	(eng,40),	(sq,26),	(hrf,14)
	$interest, coldanger$	0.24	-23%	(naq,96),	(eng,76),	(oq,36),	(dh12,24),	(oqa,20)
	$pride, sadness$	0.25	-16%	(dh12,90),	(eng,36),	(pitch,28),	(hrf,28),	(naq,24)
	$contempt, sadness$	0.11	-55%	(naq,98),	(clq,92),	(oqa,76),	(oq,62),	(dh12,34)
M2	$shame, disgust$	0.13	-18%	(eng,100),	(oq,42),	(clq,36),	(naq,8),	(oqa,4)
	$happy, panic$	0.22	-20%	(naq,100),	(clq,56),	(pitch,26),	(oqa,16),	(dh12,14)
	$despair, anxiety$	0.09	-57%	(oqa,100),	(pitch,90),	(dh12,78),	(clq,28),	(naq,18)
	$contempt, anxiety$	0.27	-25%	(eng,68),	(oqa,44),	(clq,32),	(dh12,22),	(pitch,14)
	$interest, coldanger$	0.06	-9%	(oqa,54),	(hrf,50),	(pitch,46)		
	$contempt, despair$	0.15	-13%	(oqa,98),	(oq,54),	(hrf,32),	(eng,4),	(dh12,4)

[4] E. Moore, M. Clements, J. Peifer, and L. Weisser, "Critical analysis of the impact of glottal features in the classification of clinical depression in speech," *IEEE Trans. on Biomedical Engineering*, vol. 55, no. 1, pp. 96–107, 2008.

[5] M. Liberman, Davis K., M. Grossman, N. Martey, and J. Bell, "Emotional prosody speech and transcripts," *Rand Corporation Report*, (online document: http://www.ldc.upenn.edu /Catalog /CatalogEntry.jsp?catalogId=LDC2002S28) 2002.

[6] R.Banse and K.R.Scherer, "Acoustic profiles in vocal emotion expression," *Journal of Personality and Social Psychology*, vol. 70, pp. 614–636, 1996.

[7] M. Brooks, "Voicebox:speech processing toolbox for matlab," (online document: http://www.ee.ic.ac.uk /hp /staff /dmb /voicebox /voicebox.html) 2007.

[8] J.Gudnason P.A.Naylor, A.Kounoudes and M.Brooks, "Estimation of glottal closure instants in voiced speech using the dypsa algorithm," *IEEE trans*, vol. 15, no. 1, pp. 34–43, 2007.

[9] E. Moore and J. Torres, "A performance assessment of objective measures for evaluating the quality of glottal waveform estimates," *Speech Communication*, vol. 50, no. 1, pp. 56–66, 2008.

[10] T.Backstrom M.Airas, H.Pulakka and P.Alku, "A toolkit for voice inverse filtering and parametrisation," *INTERSPEECH*, pp. 2145–2148, 2005.