Classifying Subject Ratings of Emotional Speech Using Acoustic Features.

Jackson Liscombe, Jennifer Venditti, Julia Hirschberg

Department of Computer Science
Columbia University, New York City
{jaxin,jjv,julia}@cs.columbia.edu

Abstract

This paper presents results from a study examining emotional speech using acoustic features and their use in automatic machine learning classification. In addition, we propose a classification scheme for the labeling of emotions on continuous scales. Our findings support those of previous research as well as indicate possible future directions utilizing spectral tilt and pitch contour to distinguish emotions in the valence dimension.

1. Introduction

Speech is a rich source of information, not only about what a speaker says, but also about what the speaker's attitude is toward the listener and toward the topic under discussion — as well as the speaker's own current state of mind. Until recently, most research on spoken language systems has focused on propositional content: what words is the speaker producing? Currently there is considerable interest in going beyond mere words to discover the semantic content of utterances. However, we believe it is important to go beyond semantic content as well, in order to fully interpret what human listeners infer from listening to other humans.

In this paper we present results from some recent and ongoing experiments in the study of emotional speech, designed to elicit subjective judgments of tokens of emotional speech and to identify acoustic and prosodic correlates of such speech based on these classifications. We discuss previous research as well as show results from correlation and machine learning experiments, and conclude with the implications of this study[1].

2. Previous Research

In recent years there has been considerable interest among speech researchers and technologists in the perception and production of emotional speech. Text-to-speech systems would be improved if they could model the emotional as well as the lexical content of messages to be conveyed [1] and speech recognition systems would profit from being able to identify emotions such as anger and frustration in users of spoken dialogue systems [2, 3, 4]. Promising work has been done recently in emotion detection in meetings, voicemail, and in spoken dialogue systems, especially for English and German [2, 3, 4]. These corpus-based studies have addressed the problem of emotion detection in natural or elicited speech, attempting to detect emotions such as anger and frustration with system problems in system-user interactions or urgency in voicemail by hand classifying or rating instances, extracting acoustic and prosodic features, such as duration, pitch, and energy as well as lexical cues, and employing machine learning techniques to develop predictive models. Success rates have ranged from 60-80%, depending upon the distinction attempted, with frustration and anger detectable in German with about 60% success [4, 5] and in English with 60-80% accuracy [2, 6], on different corpora and with differences in definition of target emotion.

However, work on perception and production suffers from the difficulty of obtaining reliable human judgments of emotional categories in speech, whether during perception experiments or during corpus labeling. Without reliable exemplars of particular emotional categories, it is difficult to identify those features which contribute to the effect of different emotional categories. In part, we believe that this problem is methodological. Perception studies and corpus labeling generally assume that a single emotion category can be assigned to any given speech token. We propose that a multiple ranking approach which allows listeners to rank tokens on multiple scales can provide more accurate training data, while also providing useful information on the relationship between features which signal multiple emotional categories.

Above and beyond the issue of categorizing emotions, however, is the issue of defining emotion. It is well-recognized that this is no trivial matter. In this paper, we adopt two approaches to emotion definition based on work of previous researchers. First, we maintain that emotions do not represent discrete phenomena and that a single utterance can simultaneously convey multiple emotions. Second, researchers have often found it useful to define emotions in some multidimensional space [7]. In this study we are interested in looking at two such dimensions: **valence** and **activation**. Valence is used to describe emotion in terms of positive and negative assessments (e.g. *happy* and *encouraging* have positive-valence whereas *angry* and *sad* have negative-valence). Activation is used to define emotion in terms of arousal or excitation (e.g. *happy* and *angry* have positive-activation while *sad* and *bored* have negative-activation). Past research has been relatively successful at discovering acoustic correlates distinguishing emotions on the grounds of activation [1, 8], but less successful for valence. These studies have found that positive-activation emotions have high mean F0 and energy as well as a faster speaking rate than negative-activation emotions.

3. Rating Emotion on Multiple Scales

To explore the methodological issue of eliciting reliable ratings, we have recently conducted a web-based study to discover whether it is possible to obtain multiple-emotion ratings of emotional speech tokens. We have preliminary results not only validating our hypothesis but also pointing us to further investigations of how perception of particular emotions is correlated and some hypotheses about some acoustic cues which may explain

[1]Thanks to Dan Jurafsky, Brian Pellom, Liz Shriberg, and Andreas Stolcke for useful discussions.

Table 1: Correlations of Emotion Judgments. Significant correlations in bold ($p < 0.001$).

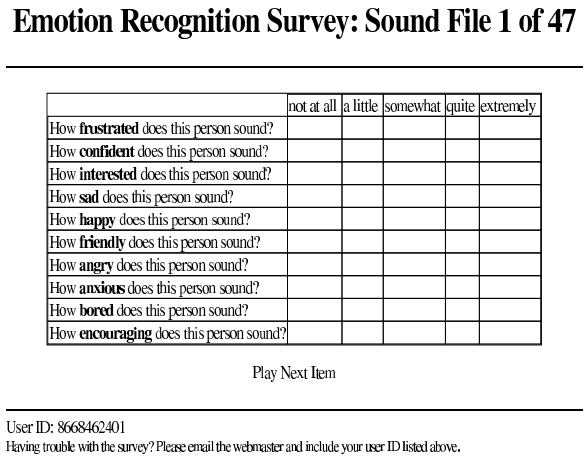
Emotion	sad	angry	bored	frust	anxs	friend	conf	happy	inter	encour
sad		0.06	**0.44**	**0.26**	**0.22**	**-0.27**	**-0.32**	**-0.42**	**-0.32**	**-0.33**
angry			0.05	**0.70**	**0.21**	**-0.41**	0.02	**-0.37**	**-0.09**	**-0.32**
bored				**0.14**	**-0.14**	**-0.28**	**-0.17**	**-0.32**	**-0.42**	**-0.27**
frustrated					**0.32**	**-0.43**	**-0.09**	**-0.47**	**-0.16**	**-0.39**
anxious						**-0.14**	**-0.25**	**-0.17**	0.07	**-0.14**
friendly							**0.44**	**0.77**	**0.59**	**0.75**
confident								**0.45**	**0.51**	**0.53**
happy									**0.58**	**0.73**
interested										**0.62**
encouraging										

some of these relationships.

For this study, we selected tokens from the LDC Emotional Prosody Speech and Transcripts corpus,[2] and prepared a web-based experiment, in which subjects were asked to rank each utterance on multiple scales. The Emotional Prosody corpus contains recordings of 8 professional actors (5 female, 3 male) reading short (4-syllables each) dates and numbers (e.g., "two thousand four") in 15 distinct emotional categories: *disgust*, *panic*, *anxiety*, *hot anger*, *cold anger*, *despair*, *sadness*, *elation*, *happy*, *interest*, *boredom*, *shame*, *pride*, *contempt*, and *neutral*. For this experiment, however, we modified the set of categories to be rated to represent emotions we felt were particularly important to the corpora we will ultimately examine and which were evenly distributed for valence. Our categories included the 'positive' emotion categories, *confident*, *encouraging*, *friendly*, *happy*, *interested*; and the 'negative' emotion categories *angry*, *anxious*, *bored*, *frustrated*, *sad*. One token representing each category plus *neutral* was selected from each of 4 actors from the corpus (2 male, 2 female), resulting in a total of 44 utterances. Selection was determined by listening to all of the LDC tokens and finding convincing exemplars matching each of our emotion categories. In addition, 3 more tokens were chosen from 3 other actors to use in practice trials.

Subjects participated in the survey over the internet. After answering introductory questions about their language background and hearing abilities, subjects were given written instructions describing the procedure. Subjects were asked to rate each token (which played out loud over headphones or speakers) on each of 10 emotional scales (see above, a 'neutral' scale was not included). For each emotion, subjects were asked *How X does this person sound?*. Subject responses could include: *not at all*, *a little*, *somewhat*, *quite*, or *extremely*. At the start of the experiment, subjects were presented with the 3 practice stimuli in fixed order. Then the remaining 44 test stimuli were presented one by one in random order. For each stimulus trial, a grid of blank radio-buttons appeared, as depicted in Figure 1. The sound file for that trial played repeatedly every two seconds until the subject selected one response for each emotional scale. Subjects were not allowed to skip any scales. The order in which the emotional scales were presented was rotated among subjects. Two randomized orders and their reverse orders were used. Each listener was presented with one of these fixed orders, shifted by one at each new token in a cyclic fashion. Forty native speakers of standard American English with no reported hearing impairment completed the survey, 17 female and 23 male. All were 18 or older, with a fairly even distribution among age groups.

[2] http://www.ldc.upenn.edu/Catalog/LDC2002S28.html



Emotion Recognition Survey: Sound File 1 of 47

	not at all	a little	somewhat	quite	extremely
How **frustrated** does this person sound?					
How **confident** does this person sound?					
How **interested** does this person sound?					
How **sad** does this person sound?					
How **happy** does this person sound?					
How **friendly** does this person sound?					
How **angry** does this person sound?					
How **anxious** does this person sound?					
How **bored** does this person sound?					
How **encouraging** does this person sound?					

Play Next Item

User ID: 8668462401
Having trouble with the survey? Please email the webmaster and include your user ID listed above.

Figure 1: Sample page from web-based perception experiment.

A correlation matrix for subject ratings of each token on each emotional 'scale' is presented in Table 1, where correlations were calculated for each pair of emotions from each subject's rating of each utterance on those scales.

From this table, we see that *frustration* patterns as we might expect, with strong positive correlation only with *angry*, and strong negative correlations with *encouraging*, *happy*, and *friendly*. There are also intuitively plausible correlations between *friendly* and *encouraging*, *happy*, *interested*, and *confident*. Note also that *bored* is positively correlated with *sad* and negatively with *happy*.[3] *Sad* is negatively correlated with *confident*, *friendly*, *encouraging*, *interested*, and, of course, *happy*. It is interesting that the speaker's own personal state, *sad* or *happy*, seems to carry over into more other-directed states, such as *encouraging* and *interested*.

When we look at the actual rankings given by subjects for each token and each emotion category, we see that roughly half of all rankings for each emotion are ranked "*not at all*", while the other half is split more evenly between the other four rankings. However, *interested* and *confident* differ markedly from this pattern, with more even distribution over all rank categories.

[3] Since TTS systems are routinely criticized as sounding *bored*, do they also sound unhappy?

Table 2: Correlation between emotion rankings and continuous acoustic features ($p < 0.001$).

Feature	sad	angry	bored	frust	anxs	friend	conf	happy	inter	encour
F0_MIN	-0.36		-0.36	-0.11		0.32	0.20	0.39	0.35	0.30
F0_MAX	-0.38	0.08	-0.51			0.31	0.24	0.39	0.42	0.29
F0_MEAN	-0.35		-0.53		0.10	0.32	0.23	0.39	0.43	0.29
F0_RANGE	-0.35	0.09	-0.47			0.28	0.23	0.34	0.38	0.25
F0_STDV					0.09					
F0_ABOVE		0.12	-0.09	0.12	0.14					
RMS_MIN	-0.16				-0.08		0.13	0.10		
RMS_MAX	-0.27	0.14	-0.37	0.10	0.08	0.11	0.22	0.21	0.26	0.14
RMS_MEAN	-0.28	0.12	-0.36		0.12	0.13	0.23	0.22	0.28	0.16
RMS_RANGE	-0.27	0.14	-0.37	0.10	0.08	0.11	0.22	0.20	0.27	0.14
RMS_STDV	-0.27	0.15	-0.35	0.10	0.08	0.10	0.23	0.20	0.26	0.13
VCD	-0.19		-0.10	-0.14	-0.17	0.16	0.23	0.23	0.14	0.20
SYL_LNGTH	0.23		0.23			-0.15	-0.09	-0.19	-0.19	-0.17
TILT_STRESS	-0.12	0.17		0.10	-0.11		0.18			
TILT_RMS		0.25	0.09	0.22		-0.17		-0.11		-0.13

We hypothesize that these tokens were more difficult to assign rankings to.

4. Identifying Emotion Categories

To discover underlying causes for differences in subject rankings of tokens in the survey, we examined acoustic and prosodic features of each token. For this analysis, we defined the corpus to include 1760 tokens, each representing one ranking of one token by one subject on one scale. The features we analyzed were divided into two sets, one automatically extracted from the speech tokens and the other labeled by hand. The hand-labeled features included continuous as well as discrete valued features. Features were chosen based upon earlier findings in the literature as well as our own intuitions.

Our automatically-extracted features include: **F0_MIN**, minimum non-zero F0 value; **F0_MAX**, maximum F0 value; **F0_MEAN**, mean of all non-zero F0 values; **F0_RANGE**, difference between highest and lowest F0 values; **F0_STDV**, standard deviation from F0 mean; **F0_ABOVE**, of all voiced samples, the ratio of those above the center of F0 range to those below the center; **RMS_MIN**, minimum amplitude; **RMS_MAX**, maximum amplitude; **RMS_MEAN**, mean amplitude; **RMS_RANGE**, difference between highest and lowest amplitudes; **RMS_STDV**, standard deviation from amplitude mean; and **VCD**, ratio of voiced samples to total segments.

Our hand labeled features include: **SYL_LNGTH**, mean length of syllables; **TILT_STRESS**, spectral tilt[4] of vowel with nuclear stress; **TILT_RMS**, spectral tilt of vowel with highest amplitude; **NUC_ACNT**, type of nuclear accent (!H* collapsed with H*); **CONTOUR**, type of intonational contour; and **PHR_END**, type of phase accent and boundary tone.

We calculated cross correlation statistics between all of the continuous features and all of the emotions based on subject ratings. Table 2 shows a number of strong correlations between individual emotion categories and these features.[5]

Most of the significant correlations shown in this table support earlier findings: F0, RMS, and speaking rate are good at distinguishing emotions on the grounds of activation [7]. Positive-activation emotions correlate with higher F0 and RMS measurements than negative-activation emotions and they have a faster speaking rate, generally speaking. Also consistent with previous studies, Table 2 indicates that these same features are **not** useful for distinguishing between emotions on the valence dimension.

The features we examined involving spectral tilt, however, seem to identify new correlations we have not observed in previous studies.[6] In particular, TILT_RMS appears to group positive-activation emotions with different valences into separate categories, with *friendly*, *happy*, and *encouraging* falling into one and *angry* and *frustrated* into another. This suggests a possible method for distinguishing emotions with regard to their valency.

We also calculated correlations for ToBI labellings [9] of each token, identifying type of nuclear pitch accent, contour type, and phrase accent/boundary tone as features. The most striking correlations occur for the final feature. The negative emotions are positively correlated with the plateau (H-L%) contour, while the positive emotions are negatively correlated. The latter are, however, positively correlated with the standard declarative (L-L%) contour, while the negative emotions are either negatively correlated or not significantly correlated with this contour.

5. Automatic Emotion Classification

We next examined how well our feature sets could predict rankings for the various emotion categories. For these experiments, we divided our 1760 token corpus into training (90%) and test (10%) sets. The test set was randomly selected, but evenly distributed among speakers and tokens. We adopted a binary classification scheme based on the observed ranking distributions ("*not at all*" as the absence of emotion *x*, all other ranks as the presence of emotion *x*).

We used a machine learning program, RIPPER [10] to automatically induce prediction models. RIPPER takes as input training data specifying the class and feature values for each training example and outputs a classification model for use in classifying future examples. This model is constructed using

[4] First harmonic subtracted from second harmonic, measured in dB, over a 30 ms window centered over the middle of the vowel

[5] ToBI-labeled features not shown.

[6] Note only that [1] mentions that positive-activation emotions have a flat spectral slope.

Table 3: Performance for Combined Feature Set by Emotion Category. Baseline computed using the most frequent ranking for that emotion.

Emotion	Baseline	Accuracy	Improvement
angry	69.32%	77.27%	11.47%
confident	75.00%	75.00%	0.00%
happy	57.39%	80.11%	39.59%
interested	69.89%	74.43%	6.50%
encouraging	52.27%	72.73%	39.14%
sad	61.93%	80.11%	29.36%
anxious	55.68%	71.59%	28.57%
bored	66.48%	78.98%	18.80%
friendly	59.09%	73.86%	25.00%
frustrated	59.09%	73.86%	25.00%

Table 4: The best performing feature(s) for each emotion.

Emotion	Feature	Accuracy
angry	F0_*, RMS_*, TILT_*, VCD	77.27%
confident	F0_RANGE, F0_MEAN	76.14%
happy	F0_MIN	81.25%
interested	F0_STDV	75.57%
encouraging	VCD	73.86%
sad	F0_MAX	81.25%
anxious	TILT_RMS	78.41%
bored	TILT_RMS	80.11%
friendly	TILT_STRESS	75.00%
frustrated	F0_MAX	75.00%

greedy search guided by an information gain metric, and is expressed as an ordered set of if-then rules.

RIPPER models using our binary classification and training either on the automatically-derived, hand-labeled, and combined feature sets, each predict about 75% of the data correctly, representing a 22% improvement over the baseline.

When we look at individual emotion categories, we again see that each feature set provides equivalent predictive power. We present results for the combined features set in Table 3. Note in particular that while a number of emotion categories are predicted with considerable improvement over the baseline (e.g. *happy*, *encouraging*, *sad*, and *anxious* in particular), emotions such as *confident* and *interested*, which, as noted above, show very different ranking distributions from other emotions, are not well predicted.

Finally, we tested the performance accuracy of each feature individually in order to estimate the importance of each in predicting particular emotion categories. Table 4 lists the highest performing feature(s) for each emotion category. Note that some single features performed as well or better than the entire feature set. While a variety of features show superior predictive power for the various emotion categories, it is interesting that F0 features are among the best predictors of *angry*, *confident*, *happy*, *interested*, *sad*, and *frustrated*, while spectral tilt serves as a good predictor of *angry*, *anxious*, *bored* and *friendly*.

6. Discussion

This study has shown that emotions can be distinguished in terms of activation using any one of a multitude of easily-obtainable features: pitch, energy, and speaking rate. Our results thus provide further evidence for a growing body of work on the description of emotional speech based on acoustic features.

Beyond previous studies, however, this study suggests that spectral tilt and type of phrase accent and boundary tone may be useful in discriminating between the valency of emotions, that is, whether an emotion is 'positive' or 'negative'. The ability to make such a distinction is widely recognized as critical in applications such as spoken dialogue systems, where it is crucial to determine whether users are satisfied with the interaction or not. These findings will be important for future study.

Finally, we believe that our methodology of eliciting multiple emotion rankings for speech tokens has led to a better representation of emotion. We have seen that utterances systematically convey several emotions simultaneously; an observation that can be exploited in various ways. For example, if a system was designed to detect user satisfaction and a user was in fact frustrated, the types of errors our detection system might make are non-trivial. For example, it is important to know that guessing *angry* is substantially better than guessing *friendly* in this instance. The labeling method explored here allows us to do this in a systematic and reliable way. In addition, with our classifiers, we can label an utterance on all 10 emotion scales, which could provide us with a much richer and more informative representation of the emotional state of the user.

7. References

[1] Schröder, M., Cowie, R., Douglas-Cowie, E., Westerdijk, M., Gielen, S., "Acoustic Correlates of Emotion Dimensions in View of Speech Synthesis", Eurospeech-2001, Aalborg, 87–90.

[2] Ang, J., Dhillon, R., Krupski, A., Shriberg, E., Stolcke, A., "Prosody-based Automatic Detection of Annoyance and Frustration in Human-Computer Dialog", ICSLP-2002, Denver, 2037–2039.

[3] Litman, D., Hirschberg, J., Swerts, M., "Predicting User Reactions to System Error", ACL-2001, Toulouse.

[4] Batliner, A., Fischer, K., Huber, R., Spilker, J., Nöth, E., "Desperately Seeking Emotions or: Actors, Wizards, and Human Beings", ISCA Workshop on Speech and Emotion-2000, Belfast.

[5] Huber, R., Batliner, A., Buckow, J., Nöth, E., Warnke, V., Niemann, H., "Recognition of Emotion in a Realistic Dialog Scenario", ICSLP–2000, Beijing, 665–668.

[6] Lee, C. M., Narayanan, S. S., "Combining Acoustic and Language Information for Emotion Recognition", ICSLP-2002, Denver, 873–876.

[7] Cowie, R., "Describing the Emotional States Expressed in Speech", Speech Comm., vol. 40, April 2003, 5–32.

[8] Tato, R., Santos, R., Kompe, R., Pardo, J. M., "Emotional Space Improves Emotion Recognition", ICSLP-2002, Denver, 2029–2032.

[9] Silverman, K., Beckman, M., Pierrehumbert, J., Ostendorf, M., Wightman, C., Price, P., Hirschberg, J., "ToBI: A Standard Scheme for Labeling Prosody", ICSLP-1992, Banff, 867–879.

[10] Cohen, W., "Learning Trees and Rules with Set-valued Features", AAAI-1996, Portland.