An acoustic study of emotions expressed in speech

Serdar Yildirim, Murtaza Bulut, Chul Min Lee, Abe Kazemzadeh, Carlos Busso, Zhigang Deng Sungbok Lee, Shrikanth Narayanan*

Emotion Research Group, Speech Analysis and Interpretation Lab
Integrated Media Systems Center, Department of Electrical Engineering, *Department of Computer Science
Viterbi School of Engineering, University of Southern California, Los Angeles
http://sail.usc.edu

Abstract

In this study, we investigate acoustic properties of speech associated with four different emotions (sadness, anger, happiness, and neutral) intentionally expressed in speech by an actress. The aim is to obtain detailed acoustic knowledge on how speech is modulated when speaker's emotion changes from neutral to a certain emotional state. It is based on measurements of acoustic parameters related to speech prosody, vowel articulation and spectral energy distribution. Acoustic similarities and differences among the emotions are then explored with mutual information computation, multidimensional scaling, and comparison of acoustic likelihoods relative to the neutral emotion. In addition, acoustic separability of the emotions is tested using the discriminant analysis at the utterance level and the result is compared with human evaluation. Results show that happiness/anger and neutral/sadness share similar acoustic properties in this speaker. Speech associated with anger and happiness are characterized by longer utterance duration, shorter inter-word silence, higher pitch and energy values with wider ranges, showing the characteristics of exaggerated or hyperarticulated speech. The discriminant analysis indicates that within-group acoustic separability is relatively poor, suggesting that conventional acoustic parameters examined in this study are not effective in describing the emotions along the valence (or pleasure) dimension. It is noted that RMS energy, inter-word silence and speaking rate are useful in distinguishing sadness from others. Interestingly, the between-group difference in formant patterns seems better reflected in back vowels such as /a/ (/father/) than in the front vowels. Larger lip opening and/or more tongue constriction at the mid or rear part of the vocal tract could be underlying reasons.

1. Introduction

Recently acoustic investigation of emotions expressed in speech has attracted increasing attention partly due to the potential value of emotion recognition for spoken dialogue management [1][2][3]. For instance, displeasure or anger due to frequent system errors in understanding user's requests could be dealt with smoothly by transferring the user to a human operator before premature man-machine dialogue disruption. However, in order to reach such a level of performance we need to identify a reliable acoustic feature set that is largely immune to inter- and intra-speaker variability in emotion expression. A prerequisite for this is to accumulate knowledge on how acoustic parameters of speech are modulated when emotion changes from normal to a certain emotional state. Such knowledge is also valuable for emotional speech synthesis through speech modification [4]. The speech database introduced and analyzed in this study has been designed and is currently being expanded with such purposes in mind. Some preliminary results of the acoustic analysis of the emotional speech database is presented.

In general, emotion has been described in a three dimensional space where arousal (activation), valence (pleasure) and control (power) represent each dimension [5]. Commonly analyzed acoustic parameters for such a description of emotion in speech have been pitch, duration at phoneme or syllable level, inter-word silence duration and voiced/unvoiced duration ratio in utterance level, energy related to the waveform envelop, the first three formant frequencies and spectral moment or balance. These are parameters related to speech prosody, vowel articulation and spectral energy distribution. Detailed reviews can be found in [6] [7][8][9].

Specifically, previous studies have shown that anger and happiness/joy are generally characterized by high mean pitch, wider pitch range, high speech rate, increases in high frequency energy, and usually increases in rate of articulation [6][10]. Sadness is characterized by decrease in mean pitch, slightly narrow pitch range, and slower speaking rate [6]. Recently, Kienast et al. [11] analyzed spectral and segmental changes due to emotion in speech. Their study on segmental reduction and vowel formants showed that anger has the highest accuracy of articulation compared to other emotions that they analyzed. They also analyzed the spectral balance of fricative sounds. Their analysis revealed that two different groups can be observed, one containing fear, anger and happiness (increased spectral balance compare to neutral), and the other containing boredom and sadness (decreased spectral balance compare to neutral).

Primary purposes of this study is to obtain detailed acoustic information on four emotions (anger, sadness, happiness and neutral) expressed in speech by an actress speaker. We analyze our proposed set of temporal and spectral parameters related to speech prosody, vowel articulation and spectral energy distribution as a function of emotion. Multidimensional scaling, mutual information computation, and acoustic likelihood estimation are used in order to investigate acoustic similarities and differences among the four emotions. Discriminant analysis is performed in order to investigate effectiveness of these parameters in emotion categorization and the results are compared with human performance. We describe differences in speech production strategy used for encoding different emotions. We also compare findings in this study with those of the previous studies mentioned above.

2. Speech material

The data analyzed in this study were collected from a semi-professional actress and consist of 112 unique sentences that are suitable, by design, to be uttered with any of the four emotions, i.e., angry, happy, sad, and neutral. Some example sentences are *She told me what you did*, *I know you were being serious*, *I am going shopping*. The recordings were made in a quite room at 48kHz sampling rate using a close-talk SHURE microphone. All files were segmented manually at the sentence level and downsam-

Table 1: *Confusion matrix of the subjective human evaluation. Columns represent the emotion selected for utterances from the emotion of each row.*

	Neutral	Sad	Angry	Happy	Other
Neutral	74	14	8	1	3
Sad	20	61	5	1	13
Angry	3	1	82	2	12
Happy	7	6	12	56	19

pled to 16kHz before analysis.

To obtain phoneme level segmentation of each utterance, first we trained a set of monophone HMMs using the TIMIT database because our database did not have enough speech material to obtain reliable HMMs. The initial HMMs were then adapted using the maximum likelihood linear regression method using our data [12]. Finally, segmentation was performed using the adapted HMMs.

Accuracy of the automatic segmentation procedure was examined based on the hand segmentation of 16 randomly selected utterances (4 from each emotion class) by a native English speaker. Absolute mean difference was 7.76 ms with a standard deviation of 18.8 ms.

3. Human evaluation

To determine how well the data represents each emotional state, we conducted human evaluation tests with 4 naive, native English speakers. 25 randomly selected utterances from each emotion category were played to the listeners and they were asked them to identify the emotional content in utterances. In addition to 4 emotion categories, listeners had the choice of assigning "none of the listed". The results of the evaluation test were moderate: 68.3% of the utterances were correctly identified. It can be observed from Table 1 that the most errors occurred between sad and neutral, and happy was generally confused with angry or *Other*, suggesting human evaluators experience difficulty to distinguish happy from the rest.

4. Acoustic measurements

4.1. Duration

Utterance durations, vowel durations, inter-word silence durations, voiced region durations, and unvoiced region durations, were measured from the corresponding label files produced by the automatic segmentation procedure. Speaking rate was also computed as the number of phonemes per second.

4.2. Fundamental and formant frequencies

We calculated the pitch and formant contours of each utterance using Praat speech processing software [13]. Resulting raw pitch and formant tracks were smoothed using a 3-point median filter. Global level statistics related to F0 such as minimum, maximum, mean, median, range (maximum-minimum), and standard deviation were calculated from smoothed F0 contours.

In order to minimize the effects of the automatic pitch and formant measurement errors, we grouped the raw pitch and formant tracks according to vowel identity and emotion category and then we calculated the standard deviation of each pitch and formant track. Next we removed the data set whenever one of the F0, F1,F2, or F3 values are outside the 2 standard deviation. After the cleaning procedure, fundamental frequency and the first three formant frequencies of vowels were estimated using the start and end times of each vowel segment in the label files and averaged values across the segments are computed.

4.3. Root mean square (RMS) energy

The energy measurement were obtained by using ESPS program *get_f0* function. We computed RMS energy of only the voiced segments in utterances.

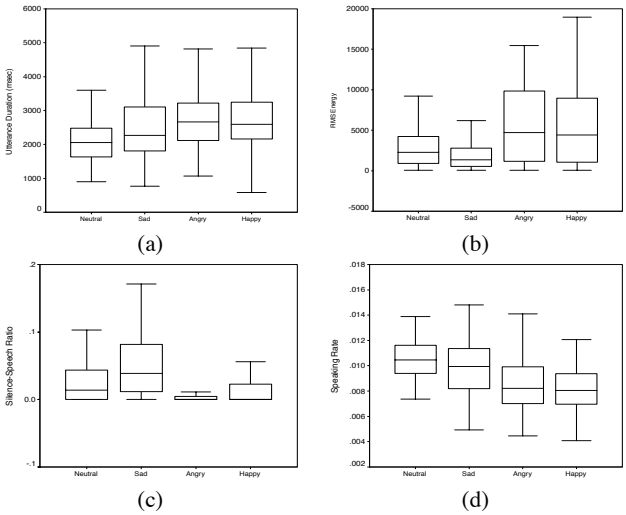


Figure 1: (a) Box plot of utterance duration for each emotion category. (b) Box plot of RMS energy. (c) Box plot of inter word silence/speech ratio.(d)Box plot of speaking rate.

4.4. Spectral balance

To investigate effects of emotion on spectral envelop, we calculated spectral balance of voiced segments of each utterance using Praat speech processing software [13]. The spectral balance is defined in Eq.1.

$$f = \frac{\sum_i |S(f_i)|\, f_i}{\sum_i |S(f_i)|} \tag{1}$$

where $|S(f)|$ is the amplitude of the spectrum and f is the frequency. Kienast et al. [11] calculated the spectral balances for voiceless fricatives. In this study, we computed the spectral balance of vowel sounds.

4.5. Acoustic likelihood comparison

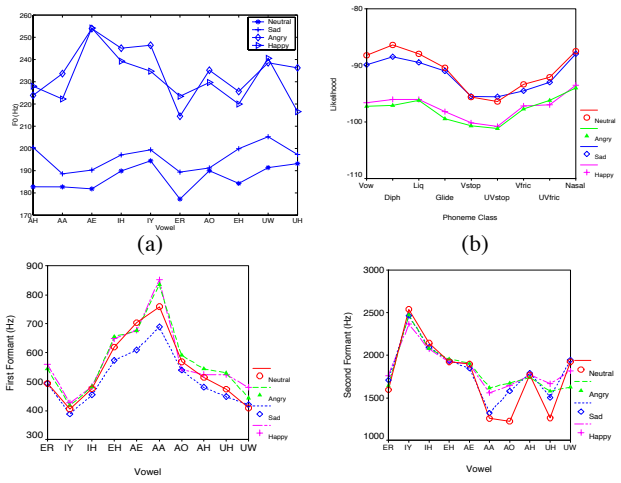
Another way to examine acoustic similarity of speech utterances associated with the different emotional states is to compute likelihood using hidden Markov models (HMMs) trained by normal or neutral speech. Forced alignment procedure was applied to all the speech utterances analyzed in this study and averaged likelihoods in phoneme level were estimated with pre-trained speaker-independent triphone HMMs using normal microphone speech of about 60 hours. The acoustic model set is believed to reflect fairly well the general acoustic properties of emotionally neutral speech.

5. Results and discussion

5.1. Duration

A box plot of utterance durations for each emotion is shown in Fig.1a. The middle line of the quantile boxes is the median, and the 25th and 75th quantiles are the ends. It is clear that sad, angry, and happy have higher median values and greater spread in the utterance duration than neutral. A simple factorial analysis (ANOVA) indicates that the effect of emotion on utterance duration is somewhat significant [F(3,444)=3.317, p=0.02]. Also, two-tailed t test showed that mean difference between neutral and happy is significant (t=2.932, df=222, p=0.004).

The box plot of inter-word silence/speech ratio within utterance is given in Fig.1c. According to the box plots, the speaker tends to use more pauses between words with the sad emotion. Also, ANOVA shows that effect of emotion on this durational parameter is significant [F(3,444)=26.390,p<0.001]. Multiple comparisons test indicates that mean differences among emotions are significant except between angry and happy.



(a) (b) (c) (d)

Figure 2: (a) Mean Vowel fundamental frequencies. (b) Average Likelihood. (c) Average first formant frequencies for vowels.(d)Average second formant frequencies for vowels.

The box plots of speaking rate for each emotion are shown in Fig.1d. It is clear that sad, angry, and happy have greater variability in speaking rate than that of neutral speech. Also, ANOVA showed that effect of emotion on speaking rate is significant [(F(3,444)=38.058, p<0.001)]. Mean differences in speaking rate are significant among all emotions except between angry and happy emotions.

Another durational parameter that has been used for recognizing emotion from speech is voiced-unvoiced duration ratio in utterance level. ANOVA indicates that the emotional changes do not significantly affect this ratio.

With regard to vowel durations, ANOVA showed that effect of emotions on vowel durations are significant [F(3,3737)=25.914, p<0.001]. Multiple comparisons test indicates that two separate groups can be observed when vowel durations are taken into account. One group comprising neutral and sad, and the other including happy and angry emotions.

5.2. Fundamental frequency

ANOVA shows that the effect of emotion on fundamental frequency (F0) is significant (p<0.001). The mean (standard deviation) of F0 for neutral are 188 (49) Hz, for sad 195 (66) Hz, for angry 233 (84) Hz, and for happy 237 (83) Hz. Earlier studies report that the mean F0 is lower in sad speech compared to that of neutral speech [6]. This tendency is not observed for this particular subject. However, it is confirmed that angry and happy speech have higher F0 values and greater variations compared to that of neutral speech. Individual mean vowel F0 values for each emotion category is shown in Fig.2a. It is observed that mean vowel F0 values for neutral speech are less than that of other emotion categories. It is also observed that anger/happy and sad/neutral show similar F0 values on average, suggesting that F0 modulation between the two within-group emotions.

In order to visualize pattern of proximities (i.e. similarities or distances) among emotion categories based on the distribution of F0 usage, we employed multidimensional scaling technique which is used to project distance relations among three or more variables into a two or three dimensional space. Results based on the Kullback-Leibler distance measure using F0 histograms are shown in Fig.3. Similar patterns of F0 in terms of averaged value and distribution is evident between angry/happy and between sad/neutral. However, separation along dimension 2, possibly the valence dimension, is not very clear.

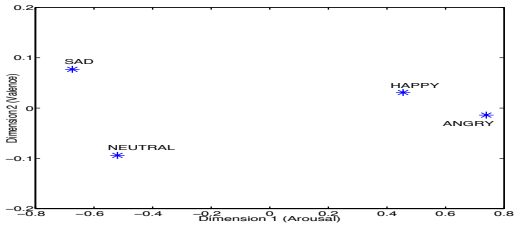


Figure 3: Multidimensional Scaling corresponding to F0 for emotion categories.

Table 2: *Mutual Information between pitch and RMS energy for each emotion.*

Emotion	Mutual Info (bits)
Neutral	0.4810
Sad	0.5202
Angry	0.8189
Happy	0.7988

5.3. Formant frequencies

The first two formant frequencies for each emotion are shown in Fig. 2c and in Fig.2d as a function of vowel identity. Two-factor ANOVA indicates that both the effect of emotion and the interaction between emotion and vowel identity are significant [F=115.3, $p < 0.001$ and F=37.4, $p < 0.001$ for the first formant and F=64.3, $p < 0.001$ and F=78.3, $p < 0.001$ for the second formant, respectively]. This suggests that the tongue positioning for a given vowel production can significantly vary depending on emotion to be expressed. Interestingly, difference in the formant patterns between the two groups of emotion (i.e., anger/happiness and sadness/neutral) are better reflected in back vowels such as /a/ than in the front vowels in this speaker. Difference in the manipulation of the lip opening and/or the tongue positioning at the rear part of the vocal tract could be underlying factors. It may be noted that we can't draw any conclusion on the variability of formant frequencies as a function of emotion as they vary depending on which formant is considered. For instance, the sad speech shows the smallest variability for F1, but it is the happy one for F2.

5.4. RMS energy

Box plots of RMS energy for each emotion class are shown in Fig.1b. It is clear that sad speech has less median value and lower spread in RMS energy than that of other emotions. Angry and happy speech have higher median values and greater spread in RMS energy. Also, ANOVA indicates that effect of emotion is significant (p<0.001). Mean RMS energy differences are also significant among emotion classes. According to our statistical analysis, RMS Energy is the best single parameter to separate emotion classes. This result was also confirmed by our discriminant analysis. Detailed results on discriminant analysis are given in Section 6.

5.5. Mutual information analysis

Pitch and RMS energy are two important prosodic cues to describing emotions. We used mutual information to find out the relation between pitch and RMS energy in a given emotion category. We estimate mutual information based on a joint probability estimation using 2-D histogram method. The results are given in Table 2. Higher value (in bit) implies more synchrony between F0 and RMS energy in speech production. As expected from other observations mentioned previously, anger and happiness use similar F0-RMS energy relation. The same tendency also holds for neutral and sadness, although sadness shows a slightly larger value due to higher pitch and wider pitch distribution.

Table 3: *Results of Discriminant Analysis*.

Features	Accuracy
F0	50.9%
Energy	55.4%
Duration	44.4%
Spectral Balance	40.2%
F0+Energy	64.7%
F0+Energy+Duration	66.1%
All features	67%

Table 4: *Confusion matrix of discriminant analysis (all features)*

	Neutral	Sad	Angry	Happy
Neutral	90	19	1	2
Sad	23	82	2	5
Angry	6	4	60	42
Happy	12	1	31	68

5.6. Spectral Balance

With regard to spectral balance analysis, ANOVA showed that effect of emotion is significant [F(3,3031)=78.813, p<0.001]. However, according to multiple a priori comparisons test, mean difference between angry and happy is not statistically significant. The mean(standard deviation) of spectral balance for neutral is 1190 (357) Hz, for sad 1244 (426) Hz, for angry 1484 (522) Hz, and for happy 1449 (458) Hz. These observations indicates the spectral slope increases toward angry speech. More air flow through the vocal folds may emphasize energy in higher frequency region. However, the exact reason is unclear yet.

5.7. Comparison of acoustic likelihood

Acoustic likelihoods associated with each emotional state are shown in Fig. 2b. It is evident that anger and happiness share similar acoustic properties and the same tendency also holds for neutral and sadness. Two-factor ANOVA indicates that the effect of emotion on likelihood as well as the effect of phoneme class are significant [F=584.3, $p < 0.001$ and F=201.3, $p < 0.001$, respectively]. Furthermore, the interaction between emotion and phoneme class is significant [F=9.1, $p < 0.001$]. This suggests that the acoustic effects of emotional change are realized differently for different phonemes, probably depending on voiced or unvoiced distinction. A related observation is the larger separation of sad and neutral likelihoods for sonorants such as vowel and diphthongs. It seems that acoustic information associated an emotional change is conveyed more in sonornats than in obstruent sounds.

6. Discriminant Analysis

In order to see how effectively these acoustic cues could be used to discriminate emotions, Fisher's linear discriminant analysis is performed. We used global level acoustic features obtained from the acoustic parameters explained in section 3. These included the mean, median, standard deviation, maximum, minimum, range (maximum-minimum) values obtained from pitch (F0), RMS energy, and spectral balance. Average utterance duration, average voice durations, average unvoiced durations and average inter-word silence durations are included as duration parameters in the analysis. Using only RMS energy features gave better performance compare to other features. 55.4% accuracy rates were achieved using energy features. Best performance was achieved using all features, 67% of the emotional data was correctly classified. Results are summarized in Table 3. The confusion matrices are given in Table 4. Comparison with Table 1 indicates that human listeners also show the same tendency. Happy and angry emotions expressed by the current speaker seem harder to distinguish with the discriminant analysis as well as by human evaluators.

7. Summary

In this study, we investigate acoustic properties of speech associated with four different emotions (sadness, anger, happiness, and neutral) intentionally expressed in speech by an actress. Results show that happiness/anger and neutral/sadness share similar acoustic properties in this speaker. Speech associated with anger and happiness are characterized by longer utterance duration, shorter inter-word silence, higher pitch and energy values with wider ranges, which agrees with [6][10]. However, we observe slightly higher pitch with wider range in sad speech, compared to neutral one. It is also found that RMS energy, inter-word silence and speaking rate are useful in distinguishing sadness from others. RMS energy is found to be the only single parameter that is significantly different among the all emotion classes. Acoustic separability between anger and happiness is poor, suggesting that conventional acoustic parameters examined in this study are not effective in describing the emotions along the valence (or pleasure) dimension. As the current speech database also includes facial expression data for this speaker, however, a joint analysis of audio and video data is expected to classify emotions more accurately. Acoustic likelihood estimation using HMMs indicates that acoustic information associated emotional changes is conveyed more in sonorants than in obstruents. Interestingly, the between-group difference in formant patterns seems better reflected in back vowels such as /a/ (/father/) than in the front vowels. Larger lip opening and/or more tongue constriction at the mid or rear part of the vocal tract could be underlying reasons. Finally, it is noted that the current results are based on speech data collected from a single and some observation such as formant frequency relations among emotions may not be generalized to other speakers. As we are collecting more data from a number of speakers, however, we will be able to address the inter-speaker variability topic soon.

8. References

[1] Lee, C. M., and Narayanan, S., "Towards detecting emotion in spoken dialogs," *IEEE Trans. on Speech & Audio Processing*, in press.

[2] Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, ans Taylor, J., "Emotion recognition in human-computer interactions", *IEEE Sig. Proc. Mag.*, vol. 18(1), pp. 32-80, Jan 2001.

[3] Litman, D., and Forbes, K. "Recognizing Emotions from Student Speech in Tutoring Dialogues". *In Proceedings of the ASRU'03*,2003.

[4] Cahn, J.E., "Generating Expressions in Synthesized Speech", Master's Thesis, MIT, 1989. (www.media.mit.edu/ cahn/masters-thesis.html)

[5] Schlosberg, H., "Three Dimensions of Emotion", *Psychol. Rev.*, vol. 61(2), pp. 81-88, 1954.

[6] Murray, I.R., Arnott, J.L., "Toward to simulation of emotion in synthetic speech: A review of the literature on human vocal emotion.", *JASA*, vol. 93(2), pp. 1097-1108, 1993.

[7] Banse, R., Scherer, K. R., "Acoustic Profiles in Vocal Emotion Expression, *J. Pers. Soc. Psy.*,vol. 70(3), p. 614-636, 1996.

[8] Nwe, T. L., Foo, S. W., De Silva, L. C., "Speech emotion recognition using hidden Markov models", *Speech communication*, vol. 41, pp. 603-623, 2003.

[9] Douglas-Cowie, E., Campbell, N., Cowie, R., and Roach, P., "Emotional speech: Towards a new generation of databases", *Speech Communication*, vol. 40, pp. 33-60, 2003.

[10] Davitz, Joel R., "Auditory correlates of vocal expression of emotional feeeling.", In *The communication of emotional meaning*, p. 101-112. New York, MacGrav-Hill, 1964.

[11] Kienast, M., Sendlmeier, W., "Acoustical analysis of spectral and temporal changes in emotional speech", *ISCA Workshop on Speech and Emotion, Northern Ireland*, 2000.

[12] Young, S., Evermann, G., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., Woodlan, P., *The Htk Book 3.2*, 2002.

[13] Boersma, P., Weenink, D., "Praat Speech Processing Software," Institute of Phonetics Sciences of the University of Amsterdam. http://www.praat.org