Utilization of Spectral and Temporal Acoustic Features for Vehicle-Centric Emotion Recognition

Tejal Udhan
Florida State University
tu13b@my.fsu.edu

Shonda Bernadin
Florida A&M University, Florida State University
bernadin@eng.fsu.edu

Abstract

It is an established fact that the emotional state of the human drivers influences their driving performance. Negative emotions while driving may have serious consequences, such as road-rage incidents and fatal crashes. If, however, a vehicle is "smart" enough to respond to a driver's undesirable emotional state, it may be able to thwart negative outcomes of these accidents. Positive and negative emotions are expressed differently by humans through their speech and facial expressions. Since speech-based systems are less distracting than visual interactive systems for in-car applications, this paper presents acoustic system to analyze four different human emotions anger, happiness, sadness, and neutrality (no emotion). However, speech emotion recognition is an emerging field and presents many challenges. The set of most powerful features that can differentiate between emotions is not defined; hence, the selection of features is a critical task. Since spectral features are primary indicators of human emotions and temporal features better model transitions in emotions, this paper analyzes these two different types of acoustic features for emotion recognition. Frequency formants are used as spectral and zero crossing rate as temporal features. A new algorithm based on a decision tree is designed to utilize these features for speaker-dependent emotion recognition.

Introduction

Speech is a method of communication or expression of thoughts in spoken words. It is the most common and fastest means of communication between humans. This fact compelled researchers to study acoustic signals as a fast and efficient means of interaction between humans and machines. For authentic human-machine interaction, the machines should exhibit sufficient intelligence to distinguish different human voices and their emotional states. It is well known that the emotional state of human drivers highly influences his/her driving performance. For example, many reports describe road-rage incidents where drivers become emotionally enraged due to the actions of another driver. This anger may lead to a high-speed chase, tailgating, and sometimes even death due to a traffic crash or physical contact. If a car is "intelligent" enough to respond to a driver's emotional state, it may be able to thwart negative outcomes of road-rage incidents.

Speech emotion recognition, extracting the emotional state of speakers from acoustic data, plays an important role in enabling machines to be “intelligent.” Most current research in this field focuses on using facial recognition techniques to characterize emotion (Tarnowski, Kołodziej, Majkowski, & Rak, 2017). However, for vehicle-centric applications, audio and speech processing may provide better noninvasive and less distracting solutions than other interactive in-vehicle infotainment systems (Lo & Green, 2013). Hence, utilization of acoustic features is preferred for emotion recognition in human drivers. This paper presents a new algorithm based on low-level acoustic features for emotion recognition of four common emotions: anger, happiness, sadness, and neutrality.

Speech is a complex signal containing information about messages contained in speech, speaker, language, and emotions. It contains linguistic (encoded in speech) and paralinguistic (related to speaker) information. The primary objective of speech is to convey information, encoded as linguistic content. Paralinguistic information includes a speaker’s age, sex, emotional state, and cognitive capacities. Due to their cognitive abilities, humans are capable of both conveying and understanding linguistic and paralinguistic parts of speech with minimal effort.

Speech processing can be defined as the field to determine speech features, understand how the features characterize the information contained in speech, and implement this knowledge to design machines capable of understanding human speech. Although speech processing deals with only the physiological nature of the speech signal, the speaker’s emotional state also imparts some of the features to human speech. Additionally, different human emotions affect speech features distinctively, and hence to have optimized speech recognition systems, the human emotions in speech also need consideration. Acoustic emotion recognition finds many applications in the modern world, ranging from interactive entertainment systems, medical therapies, and monitoring to various human safety devices (Cavazza, Charles, & Mead, 2002; Yang & Chen, 2012; Kessous, Castellano, & Caridakis, 2009; Ververidis & Kotropoulos, 2006).

Background

Emotions are specific and consistent collections of physiological responses triggered by internal or external stimuli as a representation of certain objects or situations. The internal stimuli consist of change in the person’s body that produces pain, or an external stimulus such as the sight of another person; or the representation, from memory, of a person, or object, or situation in the human thought process. The research also suggests that the basics of most emotional responses are preset in the genome (Damasio, 2000). In a general sense, emotions are a part of the bio-regulatory mechanism that humans have evolved to maintain life and survive. Emotions form an intermediary layer between stimulus and behavioral reaction, which replaces rigid reflex-like response patterns, allowing for greater flexibility in behavior (Scherer, 1982; Tomkins & Karon, 2008).

It has also been suggested that one of the major functions of emotion is the constant evaluation of stimuli in terms of relevance and the preparation of behavioral responses that may be required by these stimuli (Scherer, 1982; Arnold, 1963). Emotional reactions are essential in acquiring new behavior patterns and are a prerequisite for learning (Bower, 1981; Mowrer, 1973). The precise composition and dynamics of the emotions are specific to an individual and are based on environment and individual development. However, basic traits are consistent across all humans.

Speech is an informative source for the perception of emotions; for example, talking in a loud voice when feeling very happy, speaking in an uncharacteristically high-pitched voice when greeting a desirable person, or the presence of vocal tremor when something fearful or sad have been experienced. This mental recognition of emotions indicates that listeners are able to infer the speaker's emotional state reasonably accurately, even when the visual information about a speaker, such as the speaker's photo or video, is unavailable. This theory of cognitive emotion inference forms the basis for speech emotion recognition (Udhan & Bernadin, 2018).

Based on the definition of emotions as including a physiological component, both voluntary and involuntary effects on the human speech production apparatus can be expected, and the characteristics of vocal expression are the net result of these effects (Sethu, Epps, & Ambikairajah, 2014). Researchers have noted that characteristics affecting human movement also affect the voice production mechanism and consequently the voice. This theory is substantiated by the fact that vocal expressions of all basic emotions are similar in different languages (Udhan & Bernadin, 2018; Sethu, Epps, & Ambikairajah, 2014). Another research finding suggests that various aspects of a speaker's physical and emotional state, including age, sex, and personality, can be identified by voice alone (Kramer, 1963). This presence of low-level information even in short utterances can influence the interpretation of the words being uttered; moreover, the emotions can be recognized from segments of speech as short as 60ms (Pollack, Rubenstein, & Horowitz, 1960). Consequently, Scherer, Banse, & Wallbott have demonstrated that emotion can still be recognized even if the linguistic content of the message contained in speech is not interpreted; this serves as an evidence for the existence of vocal (acoustic) characteristics specific to emotions (2001).

Experimental listening studies with human subjects demonstrate a strong relation between qualitative acoustic features and perceived emotions (Gobl, 2003); many researchers studying the auditory aspects of emotions have been trying to define this relation (Cowie, Douglas-Cowie, Tsapatsoulis, Votsis, Kollias, Fellenz, & Taylor, 2001; Murray & Arnott, 1993). Different speech features, such as pitch, energy, frequency band ratios, jitter, shimmer, and frequency formants, are researched for the purpose of acoustic emotion recognition. However, feature selection for acoustic emotion recognition is in the early stages of research, since no set of ideal features is available to be readily used for optimal emotion recognition techniques.

Challenges of Acoustic Emotion Recognition

The development of machines capable of demonstrating human conversational skills is one of the long-sought goals of speech recognition. However, understanding linguistic and paralinguistic parts of the speech using a machine has not yet been achieved. Specifically, extraction of paralinguistic parts of speech involving emotions is a challenging task, since the machines do not have cognitive capacities as do humans. The importance of emotion recognition systems has increased with the need to improve naturalness and efficiency of speech based human-machine interfaces (Cowie et al., 2001).

The aim of an emotion recognition system is to extract features that are representative of the speech patterns characterizing only the emotional state of the speaker, while simultaneously masking the patterns that are characteristic of all other information (Udhan & Bernadin, 2018). Such features can then be utilized for automatically determining the emotional state of the speaker. However, no ideal features are identified, and the search for the best features that maximize emotion-specific information, while minimizing dependence on other aspects, is one of the central challenges in emotion recognition.

Since ideal features do not exist, pattern recognition techniques are used to make a decision about the emotional state based on chosen features. Depending which aspect of the speech signal they describe, features are broadly categorized into low-level or high-level descriptors. Low-level features describe the acoustic, prosodic, or spectral properties of the speech signal, without considering the linguistic content of the speaker's message (Udhan & Bernadin, 2018). High-level features, on the other hand, are based explicitly on linguistic content without taking into account any variations in the acoustic features of the speech signal. Even though evidence suggests that both contain emotion-specific information (Chul & Narayanan, 2005), to limit the complexity of the emotion recognition system, most acoustic emotion recognition systems rely on low-level acoustic, prosodic, and spectral features (Ververidis & Kotropoulos, 2006; Kwon, Chan, Hao, & Lee, 2003).

The lack of agreement about a theory of emotions complicates this process of data collection. Human languages exhibit many "emotion denoting" adjectives. A "Semantic Atlas of Emotion Concepts" lists 558 words with "emotional connotations" (Sethu, Epps, & Ambikairajah, 2014; Averill, 1975). It is very challenging to represent these high numbers in both collecting emotion data and constructing automatic recognizers that are capable of distinguishing such a large number of classes (Cowie & Cornelius, 2003). However, it may be that not all of these terms are equally important and, given the specific research aims, it could be possible to select a subset of these terms fulfilling certain requirements.

While the aim of these approaches is to reduce the number of emotion-related terms, it has also been argued that emotions are a continuum and these terms, even a very large number of them, do not capture every shade of emotion a person can distinguish. The dimensional approach to emotion categorization is also related to this line of argument; *i.e.*, it describes shades of emotions as points in a continuous two- or three-dimensional space. Emotional states are described in terms of a two-dimensional circular space, with its axes labelled "activation" or "arousal" (going from passive to active) and "evaluation" or "valence" (going from negative to positive) (Cowie et al., 2001). Figure 1 shows a two-dimensional emotion states model depicting different emotional states.

An important question with the dimensional approach is then if these emotion dimensions capture all relevant properties of the emotion concepts or if they are simplified and reduced descriptions. For the analysis and recognition point of view of acoustic emotion recognition, a continuum of emotions is an intractable problem, and a finite and relatively small number of emotional categories are a necessity.

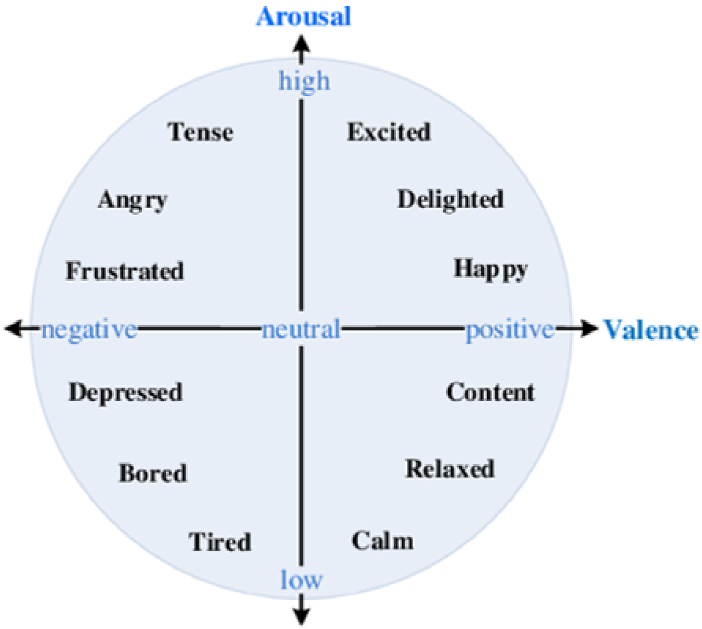

Figure 1. Two-dimensional emotion state model.

Another challenge in the emotion recognition is the lack of a common database to compare the recognition rates. Scherer (2003) stated that a review of about 30 studies yielded an average recognition rate of about 60%. However, direct comparisons of the recognition rates are futile since different datasets analyze different emotions. The datasets collected from acted and elicited emotional speech are also one of the challenging factors. There is no clear consensus since the acted speech data may not reflect what emotions people would produce spontaneously. However, research stated that even elicited emotions are “acted,” although for different reasons (Cowie & Cornelius, 2003). Using speech based on acted emotions has numerous advantages, namely, control over the verbal and phonetic content (different emotional states can be produced using the same emotionally neutral utterance) and ease of producing full-blown emotions. The high level of control over the linguistic content could also potentially allow direct comparisons of prosodic and voice quality parameters for different emotional states.

Data Description

The emotional speech database used in these experiments is the LDC Emotional Speech and Transcripts Corpus. This database was mainly chosen on the basis of language and variety of emotions. The dataset is in English and contains 14 different emotions along with the neutral state. It contains data from three male and four female speakers, including audio recordings and the corresponding transcripts. The audio is recorded at a sampling rate of 22050 Hz. Professional actors were used as subjects for recording the data. The emotion categories are neutral, hot anger, cold anger, panic, anxiety, despair, sadness, elation, happiness, interest, boredom, shame, pride, contempt, and disgust (Liberman, Davis, Grossman, Martey, & Bell, 2002).

This research presents data analysis from one male and one female speaker. Data samples for only four emotions are considered for emotion recognition, since these four emotions are the most frequently occurring emotions in everyday life: anger, happiness, sadness, and neutrality (no emotion), and the impact of negative emotions like anger and sadness are more harmful in driving applications than the other emotions (James & Nahl, 2000).

Method

This paper evaluates two different acoustic features, zero crossing rate and frequency formants, temporal and spectral features, respectively, to recognize the four different emotions for an acted speech dataset. The choice of features is based on readily available tools for calculation of these features. Since these are one-dimensional features, they can be easily analyzed for vehicle-centric applications. Values for total zero crossings and frequency formants are acquired using PRAAT software. Mean zero crossing rate and first four frequency formants are used for emotion recognition. The brief description of these selected features is as follows:

Zero Crossing Rate (ZCR)

It is a feature that characterizes only a part of the spectrum. It provides a rough estimate of the dominant frequency in the speech signal encapsulated in a single dimensional frame-based feature. ZCR has been used as a feature for emotion recognition (Huang & Ma, 2006; Lugger, Janoir, & Yang, 2009). For discrete time, it can be calculated as

$$ZCR = \frac{1}{N}\sum_{i=0}^{N-1}\left|sign(x[i]) - sign\left(x_{-1}(i-1)\right)\right| \qquad (1)$$

Where $x[i]$ is the speech signal and $x_{-1}(N)$ is a temporary array created to store previous frame values and N is the total number of samples in a frame (Lugger, Janoir, & Yang, 2009). A zero crossing occurs if the successive samples have different algebraic signs where the values of sign are

$$\begin{aligned}&\text{If } x(i) > 0,\ \text{then } sign\ (x[i]) = 1\\ &\text{If } x(i) = 0, \text{then } sign\ (x[i]) = 0\\ &\text{If } x(i) < 0, \text{then } sign\ (x[i]) = -1\end{aligned} \qquad (2)$$

For this research, mean zero crossing rates each are evaluated for each utterance of a single emotion and used as a feature for emotion recognition.

Frequency Formants

Formants can be defined as resonances of vocal tract and estimation of their location and frequencies at that location which is significant for emotion recognition (Khulage & Pathak, 2012). In this research, the first four frequency formants corresponding to the maximum pitch are used for emotion recognition.

Figure 2 shows the block diagram of emotion recognition. The features acquired from PRAAT are evaluated using a decision-tree based algorithm designed in MATLAB. While analyzing emotion data, 80% is used for training and 20% is used for testing the accuracy of emotion recognition algorithm. A total of 20 test signals are used for each speaker.

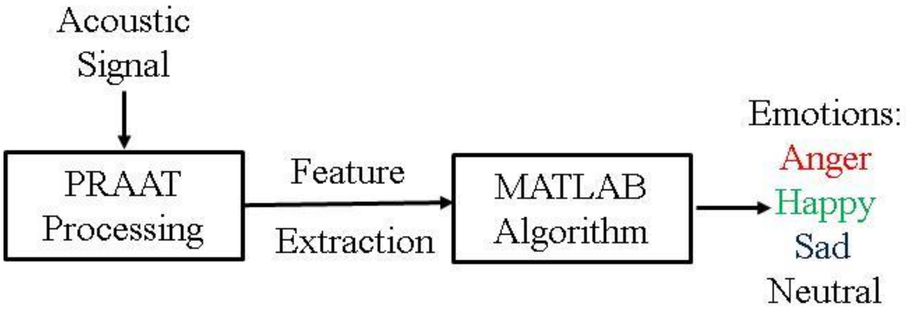

Figure 2. Block diagram of acoustic emotion recognition.

Results and Discussion

The frequency formants for male speakers do not have many variations across the same emotion category and hence provide a reliable feature for emotion recognition. Although mean ZCR is a good criterion for speech recognition, it does not establish any concrete pattern for emotion recognition and gives similar values for all the emotions. As a result, emotion recognition in male speakers has an accuracy of 85% for acoustic test signals. The first two frequency formants have distinct values for each emotion, which result in higher emotion recognition accuracy.

For female speakers, the frequency formants have a very wide range for three different emotions: happiness, sadness, and anger. Hence, the mean ZCR becomes a critical criterion for emotion recognition. The mean ZCR, however, is almost similar for happiness and anger, since both emotions are high arousal. The formants for emotions sadness and neutrality have quite similar values, which resulted in lowest accuracy for sad emotion in the female speaker of about 60%. The overall accuracy of emotion recognition for the female speaker using this method is 71%.

Table 1 shows the confusion matrix for overall accuracy of the emotion recognition system for both male and female speakers using this method. The low accuracy in emotions sadness and neutrality is specifically attributed to the female speaker. An increased number of sample errors in emotions anger and happiness are because both are high-arousal emotions resulting in high formant frequencies. Figure 3 shows the comparison of individual accuracies in male and female speakers for each emotion.

Table 1. Confusion matrix for overall accuracy of emotion recognition.

Emotions	**Happy**	**Anger**	**Sadness**	**Neutral**
Happy	32	4	1	3
Anger	3	33	0	4
Sadness	2	0	29	9
Neutral	3	0	6	31

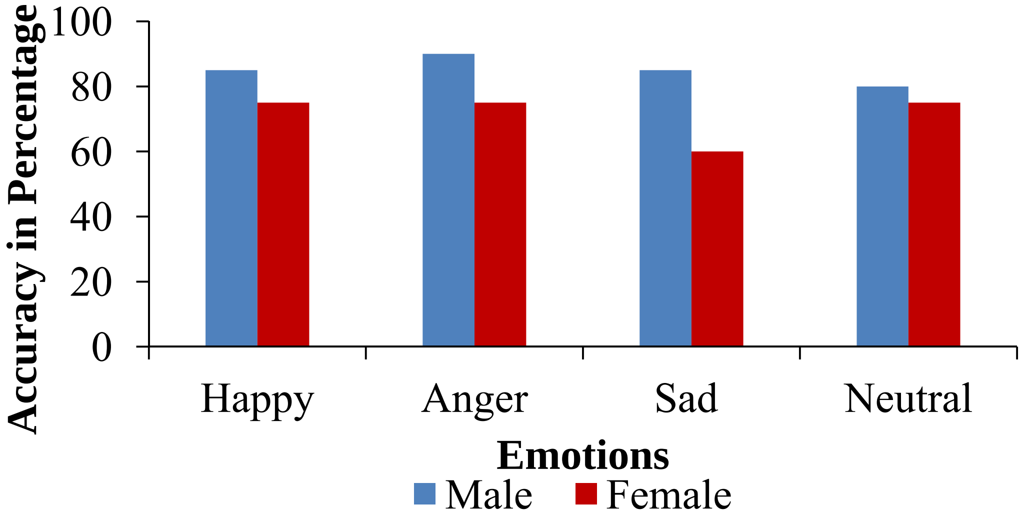

Figure 4. Emotion recognition accuracy comparison.

Conclusion

This algorithm successfully recognized emotions in the male speaker. Out of four frequency formants, the first two are distinctive for each emotion, which resulted in better accuracy of emotion recognition. However, for sad and neutral emotions in male speakers, the accuracy slightly drops due to their similarity in frequency formants. Qualitative voice features should be explored for these emotions. For female acoustic data, the selected features are insufficient to describe the selected emotions. Hence, other features that are dependent of voice quality such as pitch, intensity, and mean signal energy should be evaluated.

References

Arnold, M. (1963). *Emotion and personality*. New York: Columbia University Press.

Averill, J. R. (1975). A semantic atlas of emotional concepts. *JSAS Catalog of Selected Documents in Psychology*, *5*, 330.

Bower, G. H. (1981). Mood and memory. *American Psychologist*, *36*, 129-148.

Cavazza, M., Charles, F., & Mead, S. (2002). Character-based interactive storytelling. *IEEE Intelligent Systems*, *17*(4), 17-24.

Chul, M. L., & Narayanan, S. (2005). Toward detecting emotions in spoken dialogs. *IEEE Transactions on Speech and Audio Processing*, *13*(2), 293-303.

Cowie, R., & Cornelius, R. (2003). Describing the emotional states that are expressed in speech. *Speech Communication*, *40*(1-2), 5-32.

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

Damasio, A. R. (2000). A second chance for emotion. In R. D. Lane & L. Nadel (Eds.). *Cognitive neuroscience of emotion* (pp. 12-23). New York: Oxford University Press.

Gobl, C. (2003). The role of voice quality in communicating emotion, mood and attitude. *Speech Communication*, *40*, 189-212.

Huang, R., & Ma, C. (2006). Toward a speaker-independent real-time affect detection system. *Proceedings of the 18th International Conference on Pattern Recognition*. Los Alamitos, CA: IEEE.

Liberman, M., Davis, K., Grossman, M., Martey, N., & Bell, J. (2002). *Emotional prosody speech and transcripts*. Retrieved from https://catalog.ldc.upenn.edu/LDC2002S28

James, L., & Nahl, D. (2000). *Road rage and aggressive driving*. Amherst, N.Y.: Prometheus Books.

Kessous, L., Castellano, G., & Caridakis, G. (2009). Multimodal emotion recognition in speech-based interaction using facial expression, body gesture and acoustic analysis. *Journal on Multimodal User Interfaces*, *3*(1-2), 33-48.

Khulage, A. A., & Pathak, B. V. (2012). Analysis of speech under stress using linear techniques and non-linear techniques for emotion recognition system. *Computer Science & Information Technology*, *6*, 285-294.

Kramer, E. (1963). Judgment of personal characteristics and emotions from nonverbal properties of speech. *Psychological Bulletin*, *60*(4), 408-420.

Kwon, O., Chan, K., Hao, J., & Lee, T. (2003). Emotion recognition by speech signals, *Proceedings of the 8th European Conference on Speech Communication and Technology*. Baixas, France: ISCA.

Liberman, M., Davis, K., Grossman, M., Martey, N., & Bell, J. (2002). *Emotional prosody speech and transcripts* Retrieved from https://catalog.ldc.upenn.edu/LDC2002S28

Lugger, M., Janoir, M.-E., & Yang, B. (2009, January). Combining classifiers with diverse feature sets for robust speaker independent emotion recognition. *Proceedings of the 17th European Signal Processing Conference*. Glasgow, Scotland: EUSIPCO.

Mowrer, O. (1973). *Learning theory and behavior*. Huntington, New York: Krieger.

Murray, I., & Arnott, J. (1993). Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. *The Journal of the Acoustical Society of America*, *93*(2), 1097-1108.

Pollack, I., Rubenstein, H., & Horowitz, A. (1960). Communication of verbal modes of expression. *Language and Speech*, *3*, 121-130.

Scherer, K. (1982). The nature and function of emotion. *Social Science Information*, *21*(4-5), 507-509.

Scherer, K. (2003). Vocal communication of emotion: A review of research paradigms. *Speech Communication*, *40*(1-2), 227-256.

Scherer, K., Banse, R., & Wallbott, H. (2001). Emotion inferences from vocal expression correlate across languages and cultures. *Journal of Cross-Cultural Psychology*, *32*, 76-92.

Sethu, V., Epps, J., & Ambikairajah, E. (2014). Speech based emotion recognition. In T. Ogunfunmi, R. Togneri, & M. S. Narasimha (Eds.). *Speech and audio processing for coding, enhancement and recognition* (pp. 197-228). New York: Springer.

Tomkins, S., & Karon, B. (2008). *Affect imagery consciousness*. New York: Springer.

Udhan, T., & Bernadin, S. (2018). Speaker-dependent low-level acoustic feature extraction for emotion recognition. *The Journal of the Acoustical Society of America*, *143*(3), 1747-1747.

Ververidis, D., & Kotropoulos, C. (2006). Emotional speech recognition: Resources, features, and methods. *Speech Communication*, *48*(9), 1162-1181.

Yang, Y., & Chen, H. (2012). Machine recognition of music emotion. *ACM Transactions on Intelligent Systems and Technology*, *3*(3), 1-30.

Biographies

TEJAL UDHAN is a PhD candidate at the Department of Electrical and Computer Engineering, Florida State University. She is a research assistant in the Speech Processing and Data Analysis Laboratory. Her research interests include speech processing systems, data mining, and pattern recognition techniques. She graduated from Marathawada Institute of Technology (Aurangabad, India) with a BE in Electronics and Communication. Currently, she is working with Dr. Shonda Bernadin in the Department of Electrical and Computer Engineering on low-level acoustic feature utilization, which will help in speech pattern recognition. Her expertise includes PRAAT, MATLAB, and fuzzy clustering analysis techniques. Ms. Udhan may be reached at tu13b@my.fsu.edu.

SHONDA BERNADIN is an associate professor in the Electrical and Computer Engineering Department the Florida A&M University-Florida State University College of Engineering. Dr. Bernadin received her BS degree in Electrical Engineering from Florida A&M University in 1997, her MS degree in Electrical and Computer Engineering from University of Florida in 1999, and her PhD degree in Electrical Engineering from Florida State University in 2003. She is currently the director of the Speech Processing and Data Analysis Laboratory. Her research interests include speech analysis and pattern recognition, feature extraction, data mining, instructional design, and engineering education. Dr. Bernadin may be reached at bernadin@eng.fsu.edu.