Toward Building Automatic Affect Recognition Machine Using Acoustics Features

Andreas Marpaung, Avelino Gonzalez

Intelligent System Lab
School of Electrical Engineering and Computer Science
University of Central Florida Orlando, FL USA
amarpaung@knights.ucf.edu, gonzalez@ucf.edu

Abstract

Research in the field of Affective Computing on affect recognition through speech has used a "fishing expedition" approach. Although some frameworks could achieve certain success rates, many of these approaches missed the theory behind the underlying voice and speech production mechanism. In this work, we found some correlation among the acoustic parameters (paralinguistic/non-verbal speech content) in the physiological mechanism of voice production. Furthermore, we also found some correlation when analyzing their relationships statistically. Aligned with this finding, we implemented our framework using the K-Nearest Neighbors (KNN) algorithm. Although our work is still in its infancy, we believe this context-free approach will bring us forward toward creating an intelligent agent with affect recognition ability. This paper describes the problem, our approach and our results.

Introduction

Recent advances in Computer Science have allowed us to envision the development of smart machines that can be integrated into our daily activities. In *2001: A Space Odyssey*, the main character HAL (Heuristically programmed ALgorithmic computer) controls the spacecraft system with its artificial intelligence. Its original cognitive circuits are designed to help the astronaut crews to achieve their missions in Jupiter. However, HAL ends up killing two crew members and falsifies the causes of death as accidental. These murders happen because HAL is unable to resolve an internal conflict between its general mission and its motives derived from its heuristic knowledge and logic. If HAL were to exist, its logical rules could continue to be a threat to many lives and its mission could be steered away from its original one. Will we allow it to happen or can we introduce intelligence into these future machines to avoid such threats?

Affective Computing [1] has inspired many scientists to work not only in creating machines with linguistic and mathematical-logical reasoning abilities but also in developing agents that demonstrate their own emotions, as well as recognize affect in other agents, primarily humans. Here we focus on affect recognition in a human by a computer (e.g., robot, virtual human, avatar, etc.).

Even though there are no known signals that can be read directly from the human's body to tell a computer how he/she is feeling, several modalities exist that can be used to detect human's emotion. Some of them have been integrated into some machines. These modalities include: (a) body posture [2], (b) facial expressions [3], (c) combination of facial expressions and voice [4], (d) combination of facial expressions, body postures, and touch [5], (e) physiological signals [6, 7], (f) skin conductance and galvanic skin response [8], (g) touch [9], (h) combination of conversational cues, body posture, and facial expressions [10, 11], and (i) speech [12, 13, 14]. Zeng et al. [15] gives an in-depth overview of the current state-of-the-art research and the challenges faced by many scientists in this affect recognition domain.

Related Works

Many researchers have proposed several models of affect recognition through speech in the past several decades. Fernandez & Picard [12, 13, 14] designed a system that automatically recognizes affect through speech using machine learning techniques. It does this by fusing three different features: (1) loudness, (2) intonation, and (3) voice quality. Four emotions were investigated in this study: fear, anger, joy, and sadness. The overall recognition rate achieved was 49.4% (compared to 60.8% rate by human listeners) [12].

Another work by Iliev et al. [16, 17] proposed an approach that combines (1) the glottal symmetry feature, (2) Tonal and Break Indices (ToBI) of American English intonation, and (3) Mel Frequency Cepstral Coefficients (MFCC) of the glottal signal. Six emotions were investigated through this study: joy, anger, sadness, neutral, fear, and surprise. Combining classical features and ToBI domains led to 64.49% performance accuracy [16].

We believe that affect recognition through context-free speech alone can achieve a certain degree of accuracy. By integrating several modalities in our human body, we will be able to have intelligent agents with more accurate and robust affect recognition capabilities. In this initial work, the acoustics features were used to classify emotion. In our future work, the knowledge of the context and the context shifts during interaction with the agent will be integrated.

Approach

To evaluate the relationships among several acoustics parameters and their significances for each emotion and to establish the foundation of our work, a multivariate statistical analysis using these parameters was performed.

Speech Material

Audio files as part of the Geneva Multimodal Emotional Portrayal (GEMEP) database [18] were used for this study. In this corpus, 10 (five male and five female) professional French-speaking theater actors (mean age of 37.1 years; age range: 25-57 years) portrayed 18 affective states. Each speaker in the corpus enacted expression by saying one of these two pseudo speech sentences (in French): *"nekal ibam soud molen!"* (equivalent to a declarative statement or an exclamation) or *"koun se mina lod belam"* (equivalent to a question) for each affective state. Due to space limitation, interested readers can refer to [18] for further details on the emotion elicitation technique, recording procedures and setups, and the statistical analysis of the perceptual accuracy for the believability, authenticity, and plausibility.

Voice Source Analysis

Our work focuses on five emotions: anger, joy, fear, relief, and sadness. From 50 speech samples (five emotions x 10 speakers), in this initial work, five acoustic parameters are extracted using Praat voice analysis software [19]. These include jitter/RapJitter (varying pitch in the voice, which causes a rough sound), shimmer (a frequent back and forth change in amplitude from soft to louder), mean harmonics-to-noise ratio (HNR, the ratio between multiples of fundamental frequency and the noise), mean noise-to-harmonics ratio (NHR, the ratio between multiples of the noise and fundamental frequency), and mean frequency (mf0, the average of the sound's pitch, which shows the highness or lowness of the human voice, measured in Hertz).

Non-Parametric Statistical Analysis

To assess the relationships among five acoustic parameters, the Spearman Rank-Order Correlation test was run across all speakers and emotions using SPSS. The Spearman Rank-Order Correlation was used when the data did not pass the monotonicity test or the normality test.

Human visual inspection of the entire scatterplot graphs shows that all data parameters pass the monotonicity test. Next, we ran the normality test with our data. Based on the p-values (0.000, 0.003, and 0.008), all parameters are not normally distributed as assessed by Shapiro-Wilk's test ($p < 0.05$). Thus, we accepted the alternative hypothesis and concluded that the data did not come from a normal distribution.

The Spearman Correlation Test results have suggested strong positive correlations for mean frequency-HNR ($r_s(48) = 0.486$), jitter-shimmer ($r_s(48) = 0.667$), jitter-NHR ($r_s(48) = 0.735$), shimmer-NHR ($r_s(48) = 0.781$) and strong negative correlations for mean frequency-jitter ($r_s(48) = -0.715$), mean frequency-shimmer ($r_s(48) = -0.423$), mean frequency-NHR ($r_s(48) = -0.486$), jitter-HNR ($r_s(48) = -0.680$), shimmer-HNR ($r_s(48) = -0.768$), and NHR-HNR ($r_s(48) = -0.953$).

The Spearman Correlation Test result does not only provide us with the significant association between each pair of parameters, but it also gives us the knowledge on the magnitude and directional changes as their pairs increasing or decreasing.

Discussion

Our directional and significance results above are also parallel to the findings in [20, 21] done on the same dataset but using different approaches. These authors suggested that "Mf0, Jitter, Shimmer, and HNR are all related to vocal fold vibration, which would be influenced by vocal fold length and tension, glottal adduction, and sub-glottal pressure" [20]. The frequency of vocal fold vibration represents the number of times the vocal folds open and close per second, and directly determines the fundamental frequency (lowest frequency) of the produced sound. Typically, men's average fundamental frequency is approximately 125 Hz, over 200 Hz for women, and over 300 Hz for children [22]. The size of vocal folds also can affect the fundamental frequency; men have vocal lengths between 17 and 24 mm and women have the lengths between 13 mm and 17 mm.

Many speech-language pathologists also have confirmed experimentally the relationship between these parameters and the vocal fold vibratory patterns [23, 24, 25]. So what is the relationship between vocal fold vibratory patterns and emotion elicitation?

Vocal fold vibration patterns, which are not only involved in voice production and filtering, also relate directly to the motor effects of the emotion-related physiological processes [20]. The vibration patterns for joy is characterized by high frequency and high subglottal

pressure while sad is characterized by low frequency, weak glottal adduction, and low subglottal pressure. On the other hand, high subglottal pressure, high adduction, and high frequency are related to fear while anger corresponds to high subglottal pressure and low shimmer value. These physiological processes are controlled by the limbic system of human's brain, which is activated by autonomic nervous system and somatic nervous system. The causes and effects of affect arousal on the vocalization process are cumbersome. In a highly simplified manner, the autonomic nervous system (ANS), part of the peripheral nervous system that controls heart rate, digestion, respiratory rate, salivation, perspiration, pupil dilation, urination, and sexual arousal, activates both sympathetic and parasympathetic systems when they receive commands from the human's brain. In parallel, the limbic system also activates the somatic nervous system (SNS) that controls the muscle tone and motor commands. The activation of both ANS and SNS influences other systems such as respiration, which controls vocal intensity and frequency, phonation, which controls vocal intensity, frequency, and quality, and articulation, which controls vocal quality and resonances (formants). Due to space limitation, readers interested in more in-depth information on how vocal and emotional arousal is intertwined can refer to [26].

The fact that the relationships between the mean pitch (mf0) has negative correlation with both jitter and shimmer (through the human visual inspection) has also been confirmed by several investigations for other datasets [27 – 33]. In addition, Orlikoff et al. [33] discovered that the relationship between jitter and mf0 over the entire range of phonatory frequency was nonlinear. Our finding on the relationship's strength between each pair of emotion parameters has also been confirmed by [20, 21] that the mf0 has positive relationship with NHR and negative relationship with HNR.

Implementation

After establishing the fundamental foundation of our work through a multivariate statistical analysis, we describe the implementation of our proposed architecture next.

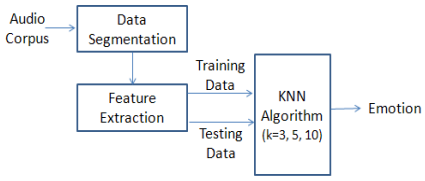

Figure 1: Proposed Architecture Block Diagram

Architecture

Fig. 1 shows our initial architecture. In this implementation, the entire 50 speech samples of the GEMEP corpus, used in our statistical analysis, are utilized in this implementation.

Data Segmentation & Feature Extraction

EasyAlign [34] was used to segment the audio corpus and generate the training and testing datasets. EasyAlign, a freely available system with a plug-in to the Praat, a well-known speech analysis software, includes two external components: a grapheme-to-phoneme[1] conversion system and a segmentation tool for alignment at the phone level.

The segmentation process produces 188 word segments whose acoustic parameters (jitter, shimmer, pitch, NHR, and HNR) are measured using the Praat software. The dataset distribution is as follows: anger (39 segments), joy (39 segments), fear (32 segments), relief (38 segments), and sadness (40 segments). These measured data serve as the training and testing datasets for the KNN algorithm; each datum belongs to one dataset only.

KNN Algorithm & Results

A non-parametric method, the KNN algorithm, classifies the testing data to the most common class amongst its k nearest neighbors (k is a positive integer, k = 3, k = 5, and k = 10). From our measured data, N datasets are chosen randomly; 10, 20, and 30 are the selected values of N for the testing datasets while using the remaining data, which are different from the testing datasets, as the training datasets.

Table 1 shows the average value of the performance results for given k and N values. Although our results do not perform to our expectation, this work has given us good direction to our research effort. Combining this method with more advanced classification algorithms, such as Gaussian Mixture Model (GMM), ARTMAP, Support Vector Machine (SVM), we believe that better performance accuracy can be achieved.

Table 1: Experiment Results

	N = 10	N = 20	N = 30
K = 3	58.0%	51.0%	56.6%
K = 5	58.0%	58.0%	57.7%
K = 10	62.0%	56.6%	54.33%

Conclusion & Future Work

The relationship among several acoustics features using a multivariate statistical analysis method is presented. These features have become the inputs to our KNN algorithm. This simple implementation has shown promising results. In the future, besides segmenting the corpus to more refined tiers, phones and syllables, we also want to integrate this work with other acoustic features and more advanced classification algorithms to have a more robust and enhanced affect recognition intelligent agent.

[1] Grapheme is a minimal unit of a writing system consisting of sequences of written symbols to represent phoneme.

References

[1] Picard, R. 1997. Affective Computing. MIT Press.

[2] Mota, S.; and Picard, R. 2003. Automated Posture Analysis for Detecting Learner's Interest Level. Workshop on Computer Vision and Pattern Recognition for Human-Computer Interaction. CVPR HCI, Madison, WI, IEEE.

[3] El Kaliouby, R.; and Robinson, P. 2005. Real-time Inference of Complex Mental States from Facial Expressions and Head Gestures. Real-Time Vision for Human-Computer Interaction, Springer-Verlag. pp. 181-200.

[4] Huang, T.S.; Chen.L.S; and Tao, H. 1998. Bimodal emotion recognition by man and machine. ATR Workshop on Virtual Communication Environments.

[5] Kapoor, A.; Ahn, H.; and Picard, R. 2005. Mixture of Gaussian Processes for Combining Multiple Modalities," In Proceedings of the Multiple Classifier Systems. 6th International Workshop MCS 2005, Seaside, CA.

[6] Picard, R.; Vyzas,E.; and Healey, J. 2001. Toward Machine Emotional Intelligence: Analysis of Affective Physiological State. IEEE Transactions Pattern Analysis and Machine Intelligence 23(10).

[7] Healey, J.; and Picard, R. 2005. Detecting Stress During Real-World Driving Tasks Using Physiological Sensors. IEEE Trans. on Intelligent Transportation Systems Vol. 6. pp. 156-166.

[8] Nasoz, F.; Lisetti, C; and Vasilakos, A. 2010. Affectively intelligent and adaptive car interfaces. Inf. Sci. 180 (20), pp. 3817-3836.

[9] Ark, W.; Dryer, D.; and Lu., D. 1999. The Emotion Mouse. HCI International '99. Munich, Germany.

[10] D'Mello, S.; Jackson, T.; Craig, S; Morgan, B; Chipman, P; White, H; Person,N; Kort, B. El Kaliouby,R.; Picard, R; and Graesser, A. AutoTutor Detects and Responds to Learners Affective and Cognitive States. 2008. Workshop on Emotional and Cognitive Issues at the International Conference of Intelligent Tutoring Systems. June 23-27, Montreal, Canada.

[11] D'Mello, S.; Craig, S.; Fike, K.; Graesser, A. 2009. Responding to learners' cognitive-affective states with supportive and shakeup dialogues. HCI, Part III, HCII2009. LNCS 5612, J.A Jacko (Ed.). pp. 595-604.

[12] Fernandez, R.; and Picard, R. 2005. Classical and Novel Discriminant Features for Affect Recognition from Speech. Interspeech 2005 - Eurospeech - 9th European Conf on Speech Communication and Technology, Lisbon, Portugal.

[13] Fernandez, R. 2003. A computational model for the automatic recognition of affect in speech. Ph.D. dissertation. MIT. http://www.media.mit.edu/˜galt/phdthesis.pdf.

[14] Fernandez, R.; and Picard, R. 2011. Recognizing affect from speech prosody using hierarchical graphical models. Speech Communication. Vol. 53. Issues 9-10. pp. 1088-1103.

[15] Zeng, Z.; Pantic, M.; Roisman, G.; and Huang, T. 2009. A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions. IEEE Transactioncs on Pattern Analysis and Machine Intelligence. Vol. 31. No. 1. pp. 39-58.

[16] Iliev, A. 2009. Emotion Recognition Using Glottal and Prosodic Features. Open Access Dissertation. Paper 515. http://scholarlyrepository.miami.edu/oa_dissertations/515

[17] Iliev, A.; and Scordilis, M. 2011. Research Article: Spoken Emotion Recognition Using Glottal Symmetry. EURASIP Journal on Advances in Signal Processing – Special issue on emotion and mental state recognition from speech.

[18] Bänziger, T.; Mortillaro, M.; and Scherer, K.R. 2011. Introducing the Geneva Multimodal Expression corpus for experimental research on emotion perception. *Emotion*. Advance online publication. doi: 10.137/a002582, http://www.affective-sciences.org/gemep/coreset.

[19] Boersma, P. 2001. Praat, a system for doing phonetics by computer. Glot International. 5(9/10). pp. 341–345.

[20] Sundberg, J.; Patel, S.; Bjorkner, E.; and Scherer, K. 2011. Interdependencies among Voice Source Parameters in Emotional Speech. IEEE Transactions on Affective Computing. Vol. 2. No. 3. pp. 162-173.

[21] Patel, S.; Scherer, K.R.; Bjorkner, E.; Sundberg, J. 2011. Mapping emotions into acoustic space: The role of voice production. Biological Psychology. Vol. 87. pp. 93-98.

[22] Borden, G.; Harris, K.; and Raphael, L. 1994. Speech Science Primer – Physiology, Acoustics, and Perception of Speech. 3rd edition, Baltimore, Maryland: Williams & Wilkins.

[23] Horii, Y. 1979. Fundamental frequency perturbation observed in sustained phonation. Journal of Speech and Hearing Research. Vol. 22. pp. 5-19.

[24] Horii, Y. 1980. Vocal shimmer in sustained phonation. Journal of Speech and Hearing Research. Vol. 23, pp. 202-209.

[25] Gelfer, M. 1995. Fundamental Frequency, Intensity, and Vowel Selection: Effects on Measures of Phonatory Stability. Journal of Speech and Hearing Research. Vol. 38, pp. 1189-1198.

[26] Scherer, K. 1989. Vocal Correlates of Emotional Arousal and Affective Disturbance. In Handbook of Social Psychophysiology. New York: John Wiley & Sons. pp. 165-197.

[27] Beckett, R. 1969. Pitch perturbation as a function of subjective vocal constriction. Folia phonation. Vol. 21, pp. 416-425.

[28] Hollien, H.; Michel, J. and Doberty, E. 1973. A method for analyzing vocal jitter in sustained phonation. Journal of Phonetic. Vol. 1, pp. 85-91.

[29] Horii, Y. 1979. Fundamental frequency perturbation observed in sustained phonation. Journal Speech Hearing Res. Vol. 22. pp. 5-19.

[30] Horii, Y. 1980. Vocal shimmer in sustained phonation. Journal of Speech Hearing Res.. Vol 23. pp. 202-209.

[31] Koike, Y. 1973. Application of some acoustic measures for the evaluation of laryngeal dysfunction. Studia phonology. Vol. 17. pp. 17-23.

[32] Lieberman, P. 1963. Some acoustic measures of the fundamental periodicity of normal and pathologic larynges. Journal Acoustic Society of America. Vol. 35. pp. 344-353.

[33] Orlikoff, R.F.; and Baken, R.J. 1990. Consideration of the relationship between the fundamental frequency of phonation and vocal jitter. Folia phoniatrica. Vol. 42. Issue 1. pp. 31-40.

[34] Goldman, J. 2011. EasyAlign: an automatic phonetic alignment tool under Praat. Proceedings of InterSpeech. Firenze, Italy.