Scholar's Digest- Journal of Multidisciplinary Studies
Volume 02, Issue 03, March, 2023
ISSN (E): 2949-8856
Scholarsdigest.org

HUMAN-COMPUTER INTERACTIONS BY USING RECOGNIZER OUTPUT VOTING ERROR REDUCTION SYSTEM

Ali Najdet Nasret,
Zuhair Shakor Mahmood,
Abbas B Noori
Electronic Department, Kirkuk Technical Institute,
Northern Technical University

Abstract:
Interaction between humans and machines is central to the field of computer science. many researchers whose focus is human-computer interaction are actually located in unrelated fields. In recent years, research into human-computer interfaces has shown a keen interest in the incorporation of emotions into conversation design. Hidden Markov Models (HMMs) have been used to distinguish emotions from speech signals.in this study has been explaining the optimizations and improvements of an emotion recognizer that works in conjunction with automated speech recognition. This study presents findings from experiments conducted on recorded and spontaneous emotional speech to show that a post-processing algorithm that incorporates various speech emotion recognizers have been successfully implemented.

Keywords: Hidden Markov Models, emotions, speech signals.

INTRODUCTION

In the nowadays the challenges speech recognition become a very common. Many types of speech recognition were developed by researchers. in this study have been illustrate one most important type which is spoken language dialogue system Figure 1 shows the main components of adaptive conversation system. Consists of automatic speech recognition, natural language understanding, a conversation manager for launching applications and starting system activities, text creation and voice synthesis for creating system prompts and converting them into audio signals.

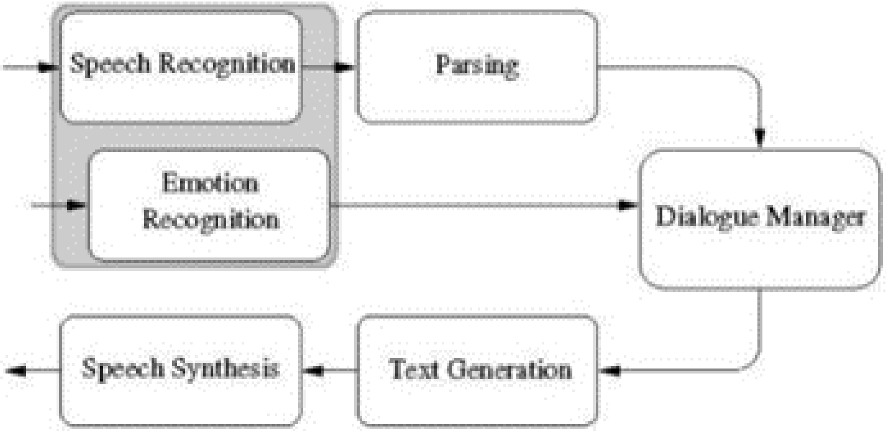

Figure 1. This is the design of an adaptive conversation system.

The adaptive conversation system conversation manager must react to the current condition of the user in order to improve user friend lines. Thus, the system has an emotional recognition component.

1. Different speech-based emotion identification techniques have been provided a comprehensive summary in Our approach is to use Hidden Markov Models (HMMs) to extract key characteristics from speech signals and then use that information to identify emotions.
2. As seen in Figure 1’s grey box on the left, we suggest a method that uses HMMs for speech recognition while also recognizing both speech and emotion.
3. After adapting the” emotional language model,” we use the ROVER notion to merge the output of numerous recognize into a single output and to lower the error probability of the speech and emotional recognition. In order to assess our suggested systems, we use both performed and spontaneous speech. Despite the fact that emotions can serve a variety of roles, they can also share traits, resulting in higher-order groups of emotions known as” families.” Researchers have proposed that good feelings can be classified into four types of emotion families based on the clustering of body language expressions of positive feelings (visual and physical gestures, speech phrasing, and body language vocalizations). These emotion families are epistemological, savoring, pro-social, and agency-approach positive emotions. Epistemological positive emotions are defined as those that are triggered by changes in an individual’s understanding of the world. Examples of such feelings include amusement, curiosity, relief, and amazement. Experiencing or thinking about different types of sensory delight can elicit good feelings such as contentment and sensory enjoyment. Loving-kindness-compassion-gratitude-admiration are examples of prosocial good emotions that are associated with caring for others. Positive emotions defined by approach inclinations are referred regarded as agency approach positive emotions, and they include elation and pride.

Details of the Voice Recording

In this study have been utilize the English Emotional Speech Database of the University of Arizona state for our performed emotion identification studies noise, poor, honor, frank, deputy, and narrow are all included in the database, as are neutral recordings that serve as references. Every day-speak utterance from 10 distinct actors, five female and five male speakers, were performed ten times to achieve excellent comparability across moods and speakers. Ratings were made by 20 people, allocating the statements to various emotions and defining their level of conviction. The Arizona corpus had 705 utterances for training and 104 for testing in the studies that follow.

In addition, a database of emotional and spontaneous speech has been utilized. noise, frank, poor, deputy and narrow are all included in this collection of 586 statements. Three female and 10 male speakers were recorded while they took a personality test and a quiz. We recorded in a quiet room with no outside noise, and we manually classified the utterances according to our own interpretations of what we heard. The utterances have been relabeled alliteratively using an emotion recognizer and manually corrected in order to increase the credibility of the data. Words are expanded into word-emotion pairs in our emotion recognition and speech method. Instead of (evening), the recognizer might choose" evening-NOISE," evening-POOR," etc. Consequently, the dictionary's HMM count increases by seven since each phoneme now has seven possible pronunciations (one for each of the human emotions). In order to get the feature vectors needed by the recognizer, is used to extract them from the speech stream. Training and recognition are carried out utilizing the Hidden Markov Model.

Extraction of Features

Thirty-eight MFCCs, as well as their first and second order derivatives derived using HTK, make up the feature set. Praat has been used to extract an additional 38 characteristics from the voice data at 10 MS intervals. Prosodic characteristics and associated statistical calculations, such as pitch, formants, intensity, jitter, and harmonicity, are regarded to be the most significant aspects for emotion identification. Mean, variance, and range are all included in their statistical calculations. These parameters, together with MFCCs, have been used to create a wide variety of feature vectors, such as those with parameters 30, 31, 32, and so on.

Evaluation

Three models of language have been employed in our investigations. Only the sentences in the corpus will be included in the sentence grammar, which will include all possible word combinations, and the likelihood of nearby words will be included in the word loop model (bi-grams). Word-emotion pairs are the output of the recognizer in combined emotion recognition and speech. As a result, in addition to an overall review, we also assess the accuracy of words and the accuracy of emotions.

Evaluation of a Word

The three distinct grammars have a considerable impact on word correctness, as predicted. While the word loop is used for the lowest word accuracy, the sentence grammar provides the highest word accuracy (the upper limit) (lower bound). Sentences have a maximum word accuracy of 100%, a word loop of 56.5 percent, and 89.1 percent (bi-gram).

Emotional Assessment

There is just one feeling in each phrase, and we presume that in our approach. If a sentence contains many emotions, which is the most usual scenario, the most dominating emotion is chosen as the emotion of this sentence.

TABLE 1. The matrix of emotional misclassification

	Noi.	Poo.	Pon.	Dep.	Fra.	Nar.
Noise.	2.08					
Poor.	0.71	1.06				
Honor.	0.00	2.56	6.04			
Deputy.	0.34	1.37	1.60	0.87		
Frank	3.22	5.03	0.17	0.83	0.50	
Narroow	0.00	0.13	6.66	1.60	1.12	0.32

Emotional accuracies are significantly different from word accuracy's, which is to be anticipated. We can attain great emotion accuracy with any language model since the choice of language models has no significant impact on them at all. After adding three formants, intensity, pitch, and statistical computations of pitch to the mel frequency cepstrum coefficients, we have a 70.7 percent accuracy rate. Using frequency measure instead of a time measure may provide more accurate findings. the frequency of mistakes in emotion pairings as shown in Table 1. noise, poor and deputy are the three most common misinterpretations. There are three emotion pairings that account for 48% of the overall error rate: noise and poor, frank and honor. deputy and narrow don't play a significant part in conventional conversation applications.

4. A flight booking system, for example, should appeal to both noise and honor consumers equally. This raises the accuracy of emotion detection to 75.8% by excluding deputy utterances and classifying honor as noise instead of the more common emotions like rage and terror.

Recognition in Two Steps

It is demonstrated in Figure 2 that an improved language model is constructed for the combined speech-emotion recognizer utilizing the output of a speech recognizer, which extends the previous combined speech-emotion recognizer.

The voice recognition component uses a bi-gram model, resulting in a word accuracy rate of 93.6 percent. As a result of using this speech recognizer, the whole system's emotional accuracy rises from 71.7 percent to 72.6 percent.

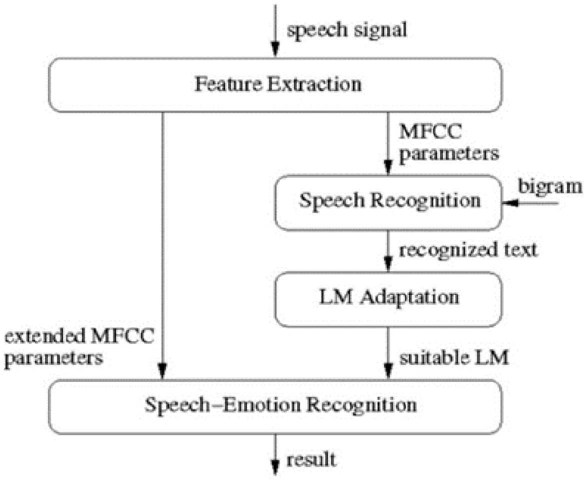

Figure 2. It's a two-step process.

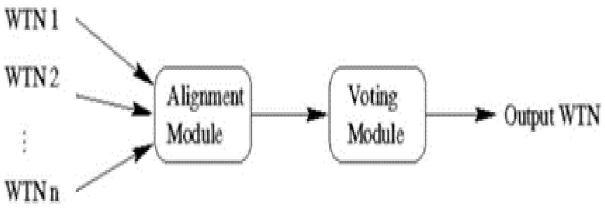

Figure 3. The ROVER system's internal structure.

There is no significant difference in performance between the two-step recognizer and the ordinary speech recognizer assuming flawless voice recognition (100 percent word accuracy).

Multiple Recognizers Put Together

The Recognizer Output Voting Error Reduction system (ROVER) takes use of the variances in how mistakes arise in various recognition systems since it was initially created for automated voice recognition. The mistakes of two or more speech recognizers might vary significantly despite the fact that their recognition capabilities are quite comparable. This also holds true when speech-emotion recognizers are employed in conjunction with each other.

An overview of the system

The Recognizer Output Voting Error Reduction system is made up of two parts: a voting module and an alignment module, as shown in Figure 3. For alignment, word transition networks (WTNs), which are recognized outputs, are aligned according to the distance between the Levenshtein distance between them. A new WTN will be generated by this module, which will then be used by a voting module to estimate each branch's score and choose its highest-scoring phrases according to Equation (1).

$$S(e)=\left[\sum_{i,j} C_e(i,j)\right]\Big/\sum_{e,i,j} C_e(i,j)\cdot\alpha\cdot N_e\big/N_s+(1+\alpha) \qquad (2)$$

For example, N_e is the total duration of emotion e, N_s the total duration of all emotions in all input ETNs and the confidence score of emotion e at time instant i in ETN $j\alpha\in[0,1]$ is again a weighting factor determining to what extent the duration or the confidence measures of an emotion are considered in the score calculation. The input WETNs will be provided by five different recognizers that we've tested in the past. Accuracy rates for words vary from 82.5 to 89.1 percent, while accuracy rates for emotions range from 66.0 percent to 68 percent. ROVER systems are ranked according to their WETNs. All five of the stand-alone recognizers are included in the most sophisticated system (ROVER-D) which incorporates the output WETNs of all five of the recognizers.

Analyzing the Meaning of the Word

The word accuracy of several ROVER systems is shown in Table 2. The results vary from 87.9% to 89.7%, which represents a minimal variance of −1.2% to +0.6% when compared to the top single recognizer. The correctness of the term is unaffected by the use of different forms of the letter a.

TABLE 2. The matrix of emotional misclassification

TABLE 2. The matrix of emotional misclassification

α	Rover-A	Rover-B	Rover-C	Rover-D
0.0	89	88.5	89.2	89.3
0.24	87.7	86.8	88.7	87.8
0.51	87.7	87.0	88.7	88.1
0.65	87.7	86.9	88.7	88.2
1.0	87.1	87.3	88.7	88.2

Emotional Intelligence

A comparison of each Recognizer Output Voting Error Reduction system accuracy in detecting emotions is shown in Table 3. A gain of 7.5% in absolute accuracy over the best-performing single recognizer is shown here, with the highest emotion accuracy reaching 76.4 percent. Even the value of has an impact on emotional correctness. Best results may be achieved if an emotion's confidence ratings and duration are taken into account equally in the computation of $S(e)\alpha = 0.5$. For each row in this table, there is an increase in emotion accuracy that corresponds to a rise in input systems (WETNs). The top theoretical limit of a Recognizer Output Voting Error Reduction system reaches 99% if the single recognizer's error distributions are diverse enough. The introduction of a fourth WETN (with a greater mistake rate than the previous five WETNs) reduces emotion identification accuracy in reality, however, as later tests have shown. In (66.0 percent, 73.7 percent, 68.9 percent, 69.8 percent and 70.8 percent).

TABLE 3. ROVER systems' ability to accurately convey emotions.

α	Rover-A	Rover-B	Rover-C	Rover-D
0.0	66.0	64.2	64.1	65.0
0.35	68.8	67.9	72.7	73.6
0.4	72.7	73.5	74.5	74.4
0.65	73.7	73.7	73.5	72.5
1.0	71.7	70.6	70.6	71.8

Recognition Of Unplanned Emotional States

Combining speech and emotion identification on acted emotional data was the goal of the investigations outlined here. Testing data is the Affective Speech and Small Database of English Spontaneous (in section 2), which contains examples of spontaneous and emotional English speech. Earlier attempts at labeling utterances based on subjective impressions of one individual have been abandoned in favor of a more scientific approach. In order to do this, additional labelers and emotion models trained on the acted Arizona database were used in conjunction with a five-fold Recognizer Output Voting Error Reduction system. on the other side training and testing, 484 utterances from the database were employed, whereas 96 were used for testing, and the highest emotion accuracy of 58% was obtain. It's important to remember that the human emotion recognition performance is lower for spontaneous speech.

Conclusion

In this study, have been examined several methods for recognizing the emotional content of speech. There are 38 MFCCs, as well as three formants (pitch and intensity), that contribute to a successful recognition of emotional states, according to the findings of the simulations Emotion identification accuracy of up to 75.6% was achieved for all emotions and all speakers in the English corpus using a two-step recognition strategy. Emotion detection accuracy rises to 74.8 percent when the number of emotions is reduced to only four and neutral.

The accuracy of emotion identification is greatly improved by combining numerous recognizer outputs and adapting a Recognizer Output Voting Error Reduction. Emotion accuracy for a ROVER system with five inputs is 76.4% and 78.1% (also with 10 speakers and all emotions. Comparatively, 86.2% of the 22 labelers in the corpus can identify human emotions. Up to 58.2% of emotional speech data may be used to identify emotions.

Semantic analysis, for example, will be used in order to boost identification rates and make emotion recognition more robust. Conversation systems are now including an emotion recognition component, and the output will be used in adaptive dialogue management. The conversation flow is formed by an evolving user state model, which takes into account the identified emotions as one of many dialogue-influencing parameters.

REFERENCES

1. Hassan, M. D., Nasret, A. N., Baker, M. R., & Mahmood, Z. S. (2021). Enhancement automatic speech recognition by deep neural networks. Periodicals of Engineering and Natural Sciences, 9(4), 921-927.
2. Mahmood, Z. S., Coran, A. N. N., & Aewayd, A. Y. (2019, October). The impact of relay node deployment in vehicle ad hoc network: Reachability enhancement approach. In 2019 Global Conference for Advancement in Technology (GCAT) (pp. 1-3). IEEE.
3. Mahmood, Z., Nasret, A., & Awed, A. (2019, September). Design of new multiband slot antennas for wi-fi devices. In International Journal on Communications Antenna and Propagation (IRECAP) (Vol. 9, No. 5).
4. Mahmood, Z. S., Nasret, A. N., & Mahmood, O. T. (2021, October). Separately excited DC motor speed using ANN neural network. In AIP Conference Proceedings (Vol. 2404, No. 1, p. 080012). AIP Publishing LLC.
5. Mahmood, Z. S., Coran, A. N. N., esam Kamal, A., & Noori, A. B. (2021, August). Dynamic spectrum sharing is the best way to modify spectrum resources. In 2021 Asian Conference on Innovation in Technology (ASIANCON) (pp. 1-5).
6. Nasret, A., & Mahmood, Z. (2019). Optimization and integration of rfid navigation system by using different location algorithms. International Review of Electrical Engineering (IREE), 14(4).
7. 7. Mahmood, Z. S., Coran, A. N. N., & Kamal, A. E. (2018). Dynamic approach for spectrum sharing in cognitive radio. International Journal of Engineering & Technology, 7(4), 5408-5411.

8. Nasret, A. N., Noori, A. B., Mohammed, A. A., & Mahmood, Z. S. (2021). Design of automatic speech recognition in noisy environments enhancement and modification. Periodicals of Engineering and Natural Sciences, 10(1), 71-77.
9. NASRET, A. N., KAMAL, A. E., & MAHMOOD, Z. S. Radar Target Detection by Using Levenberg-Marquardt Algorithm.
10. Mahmood, Z. S., Kadhim, I. B., & Nasret, A. N. (2021). Design of rotary inverted pendulum swinging-up and stabilizing. Periodicals of Engineering and Natural Sciences, 9(4), 913-920.
11. Coran, A. N. N., Mahmood, Z. S., & Kamal, A. E. (2021, October). Classification of Acoustic Data Using the FF Neural Network and Random Forest Method. In 2021 International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON) (pp. 1-4). IEEE.
12. Coran, A. N. N., Sever, P. D. H., & Amin, D. M. A. M. (2019). Acoustic data classification using random forest algorithm and feed forward neural network. In IEEE global conference for advancement in technology.
13. Laylani, L. A. A. S. S., Coran, A. N. N., & Mahmood, Z. S. (2022, January). Foretelling Diabetic Disease Using a Machine Learning Algorithms. In 2022 International Conference for Advancement in Technology (ICONAT) (pp. 1-5). IEEE.
14. Kamal, A. E., Salih, A. B., & Coran, A. N. N. (2018). Spectrum sensing algorithm using ANN in cognitive radio. International Journal of Engineering & Technology, 7(4), 5151-5155.
15. Kadhim, I. B., Khaleel, M. F., Mahmood, Z. S., & Coran, A. N. N. (2022, August). Reinforcement Learning for Speech Recognition using Recurrent Neural Networks. In 2022 2nd Asian Conference on Innovation in Technology (ASIANCON) (pp. 1-5). IEEE.
16. Kadhim, I. B., Nasret, A. N., & Mahmood, Z. S. (2022). Enhancement and modification of automatic speaker verification by utilizing hidden Markov model. Indonesian Journal of Electrical Engineering and Computer Science, 27(3), 1397-1403.
17. Coran, A. N. N., Ali, A. H. M., Mahhmood, Z. S., & Mohammed, S. F. (2021, December). Design Speech Recognition Systems in the nosily Environment by Utilizing intelligent Devices. In 2021 Second International Conference on Smart Technologies in Computing, Electrical and Electronics (ICSTCEE) (pp. 1-6). IEEE.
18. Coran, A. N. N., Mahmood, Z. S., & Salih, A. B. (2018). Satellite network underlying LMS for coverage and performance enhancement. International Journal of Engineering & Technology, 7(4), 5404-5407.
19. Mezaal, Y. S., & Nasret, A. N. A New Microstrip Bandpass Filter Design Based on Slotted Patch Resonator.
20. Mahmood, Z. S. (2012). YÜKSEK LISANS TEZI.
21. Aaref, A., & Mahmood, Z. (2021). Optimization the accuracy of ffnn based speaker recognition system using pso algorithm. International Journal on Communications Antenna and Propagation (IRECAP), 11(4).