Speech Emotion Recognition Combining Acoustic Features and Linguistic Information using Network Architecture

Saumyadeep Singh
Department of CSE
Amity University Uttar Pradesh, India

Syed Wajahat Abbas Rizvi
Department of CSE
Amity University Uttar Pradesh, India

ABSTRACT
For more accurate speaker identification in emotion-driven human-robot interaction, we suggest a unique method for combining linguistic information and acoustic to improve automated speech recognition (ASR) performance. This study creates a model with two primary components and divides emotional states into seven distinct categories. Contour, pitch, and energy spectrum characteristics are important criteria for analysis in the first component, which focuses on emotion identification from audio information. Using emotional phrases, the second component uses linguistic information to identify emotions in conversational material. We investigate a number of classification techniques, such as neural networks, auxiliary vector machines, linear classifiers, and Gaussian mixture models, in order to assess the efficacy of our methodology. The accuracy with which these methods can categorize emotional states is the basis for their evaluation. Ultimately, a neural network is used to combine soft judgments from language and auditory models, guaranteeing a more thorough and reliable emotion identification system.

Two corpora of emotional speech are used for training and validation in order to evaluate performance. When compared to models that just use individual variables, the results show that combining language and auditory information greatly improves the accuracy of emotion identification. Enhancing ASR reliability and maximizing human-robot interaction depend on improvements in speaker emotion recognition, which this development helps to achieve. We also go over how our strategy stacks up against other approaches, emphasizing quantifiable benefits from our integration approach. The results show how well our model can identify emotions in a variety of speech situations, opening the door for more sophisticated and sensitive speech recognition systems. This work advances the creation of more responsive and intuitive human-robot communication by improving emotion identification algorithms, which is important for applications in assistive technology, customer service, and healthcare.

Keywords
Automatic Speech Recognition, Emotion Recognition, Human-Robot Interaction, Acoustic Features, Linguistic Information, Neural Network.

1. INTRODUCTION
At present, interest is growing in the detection, recognition and interpretation of user emotions during interactions between humans and machines. Individually, many applications exist in information retrieval and medical analytics [2]. In our work, we have focused on applying ER inthe vehicular environment i.e., to provide in-vehicle systems

with awareness about drivers mood be it for initiating safety strategies, enable proactive help or error forgiveness based on driver emotions. For human-computer interaction targeted research, non-invasive developments seem to have benefited from more attention due to the emotion user comfort level control and specific convenience it brings. While speech analysis and imitation seem to be the most promising, voice will be the input channel of this investigation. The sound characteristics of emotional speech is largely responsible regarding the advancements made within speech emotion identification. The integration of linguistic and auditory information appears to be the most reasonable result, but newer techniques have focused more on the verbal content itself [3–4]. Therefore, we aim to integrate these two knowledge sources as robustly as possible in the work that is presented. First, using only acoustic information, our goal is to demonstrate the best classification method and feature set in comparison that honors speaker autonomy and good performance. We then concentrate on linguistic data. We offer a belief network-based approach for emotional sentence recognition, while other works use the conditional probabilities of individual words in a statement to predict the probability of an emotion. This method's concept is to frame the entire speech as a denial of a feeling, letting the speaker specify the extent of the denial. Take this statement into consideration: "I don't feel good at all." Furthermore, "too" indicates the true degree whereas the phrase "good" isdisregarded. This talk on language- and sound-based emotion identification is followed by a new method for combining the two. We suggest soft-decision fusion, which retains current knowledge for the ultimate decision-making process, in contrast to the majority of fusion operations that have been carried out thus far, which use late semantic fusion. We will take into consideration given that there is currently no consensus on a general scheme for classifying emotions in technical applications and that researchers in the field of automatic emotion recognition frequently employ the user's discrete emotional states, the MPEG4 standard names the following emotional states: anger, joy, disgust, fear, sadness, and surprise. Immobility state is frequently divided into a neutral condition to complete this set. Within our investigation, we have selected these seven emotions for international comparison [5] [6]. Emotional assessment must adhere to the discussion statement in its entirety.

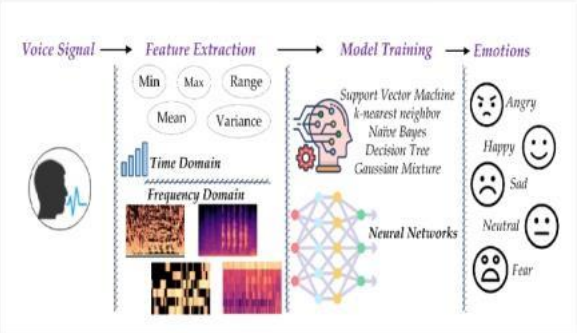

Figure 1: Speech Emotion Recognition

2. EMOTIONAL SPEECH CORPUS

The FERMUS III project, which focuses on emotion recognition in automotive environments, collected the emotional speech corpus. The AKG-1000S MK-II dynamic microphone was used to record the audio in an acoustically isolated space. The corpus consists of statements made by 13 participants—one of whom was female— in both German and English. In order to reduce the impact of actors' expectations, 2,829 staged emotional samples were gathered over the course of a year and make up the first portion of the corpus. These examples provide as a baseline dataset for prosodic and linguistic analysis training and assessment. Although these produced emotions are a good place to start, real emotions give a more accurate portrayal, especially when taking the conversational context into account. Seven hundred utterances taken from automobile infotainment voice interface conversations make up the second component of the corpus, which is meant to be evaluated for fusion. Other than contempt and melancholy, the project's main focus was on emotions, thus more usability tests were carried out to make sure that every emotion in the dataset was distributed fairly. At the conclusion of the test series, speakers were asked to reorder their own samples using a random arrangement to verify accuracy of emotions stated. The average results demonstrated that, at just 2.11%, the overall standard variation between human classifiers was very small. The figures utilize the following abbreviations: dis for disgust, fea for fear, ang for anger, neu stands for neutral, sad for melancholy, sur for surprise, and joy for delight.

Emotion	Ang	Dis	Fea	Joy	Ntl	Sad	Sur
Error, %	7.9	19.8	18.6	14.6	16.6	23.8	12.4

Figure 2: Human reclassification error rates, mean 16.31%.

3 ACOUSTIC FEATURE SET

In this work, we focus on the learned static features due to their better classification performance. Since a wide variety of noises might affect the raw pitch and energy contours, we first compute them. Spectral qualities, on the other hand, seem unduly dependent regarding phonemes and, thus, the phonetic content of the speech. Assuming phonetic content independence throughout acoustic analysis has the disadvantage of this dependence. Thus, we only take into account spectrum energy in the 251–651 Hz range while including spectral data. Using a Hamming window technique, we examine a speech frame every 10 ms lasting 20 ms. The energy figure is comparable to the frame's log-average energy. The Mean Amplitude The distinction Pitch is determined using function contour (AMDF). Because of its summing restriction, Compared to the autocorrelation function, the first-order AMDF offers a quicker solution. This method, like all height estimate techniques, depends on variations from the initial height, which can only be determined by evaluating the larynx. AMDF exhibits sensitivity to dominating formants but resilience against noise. We apply low-pass filtering using a symmetric moving average filter with a width of three to smooth down sharp edges before statistical analysis.

Subsequently, we use the contour to extract higher-level characteristics by taking its mean and adjusting it to its standard deviation. We approximate the temporal characteristics of voiced sounds relative to the pitch contour zero level because unvoiced sounds are inharmonic. The calculation of silence time relies on a threshold energy value.

We first took into consideration a thorough set of more than 200 features, in light of the continuing discussion over the ideal set of global static properties. The components of our finished feature vector in 33 dimensions are listed the following table. These characteristics are classified using linear discriminant analysis. A straight comparison shows that using all terrain-related features results in an accurate recognition percentage of 69.80%, whereas using only energy-related features results in a rate of 36.58%.

Feature	LDA,%
Pitch maximum gradient	31.5
Pitch relative position of maximum	28.4
Pitch standard deviation	27.6
Pitch mean value gradient	26.1
Pitch mean value	25.6
Pitch relative maximum	25.2
Pitch range	24.8
Pitch relative position of minimum	24.4
Pitch relative absolute area	23.8
Pitch relative minimum	23.7
Pitch mean distance between reversal points	23.0
Pitch standard dev. of dist. between reversal points	23.0
Energy mean distance between reversal points	19.0
Energy standard dev. of dist. between reversal points	18.6
Duration mean value of voiced sounds	18.5
Spectral energy below 250 Hz	18.5
Energy standard deviation	18.1
Energy mean of fall-time	17.8
Energy median of fall-time	17.8
Energy mean value	17.7
Energy mean of rise-time	17.6
Duration of silences mean value	17.5
Rate of voiced sounds	17.0
Signal number of zero-crossings	16.9
Signal median of sample values	16.8
Energy median of rise-time	16.7
Signal mean value	16.7
Energy relative maximum	16.6
Spectral energy below 650 Hz	16.3
Energy relative position of maximum	15.9
Energy maximum gradient	15.7
Duration of silences median	15.7
Duration of voiced sounds standard deviation	15.1

Figure 3: A linear discriminant analysis is used to rank the auditory characteristics.

3.1 Classification of Acoustic Sets

Techniques for acoustic layer classification are investigated: For each method, the best parameter settings and results are discussed.

3.2 Linear Classifiers

As a benchmark for performance, a basic classifier based on a Euclidean distance metric that finds the closest mean class vector (kMeans) was employed. In a later phase, k-nearest neighbor (kNN) classifiers were also assessed. The k closest references to the input vector cast a majority vote, which determines the outcome. When was set to 1, the best outcome was chosen directly, resulting in the highest performance. The outcomes produced by these classifiers amply demonstrate the nonlinear character of the issue and the need for a more sophisticated strategy.

3.3 Gaussian Mixture Models

By combining a variety of weighted Gaussians, GMMs offer a reliable estimate of the probability distribution function of the first observed feature. An method was optimized was used to determine the mixing coefficients. GMM is used to represent each emotion, and the maximum likelihood model is used to guide decision-making. With 16 combinations, the greatest recognition result was obtained.

3.4 Neural networks

Using neural networks is conventional practice for classifying templates. They are renowned for their selective learning, independent weight capabilities, and nonlinear transfer functions. In light of the limited amount of accessible information for emotion training, its better performance on a short training set than GMM appears to be beneficial.The number of input characteristics was represented by 33 input neurons in the multi-layer perceptron, which had seven output neurons for every emotion and a sigmoid transfer function in the buried layer. The concealed layer's optimal performance was seen when 100 neurons were used. The softmax function was employed in order to standardize the output after the fact. Cross-entry as a mistake in learning, 1000 repetition distribution function as well as several cross-checks were employed.

3.5 Vector machine support

Support vector machines (SVM) have garnered much attention lately for classification challenges because of their great generalization abilities resulting from structural risk minimization-based training. Using a mapping function that allows for linear separability, SVMs convert feature vectors that are entered into a feature space that is usually high-dimensional in order to solve nonlinear issues. The separation hyperplane between two class borders is properly positioned by the approach to ensure optimum classification performance. Support vectors define this hyperplane, reducing the number of required references. There are several approache handling issues. In assessment, we are offered 3 distinct methods: first, comparing each class's SVM to all others and choosing the one with the biggest separation from the others; The second method involves putting the distances into a multilayer perceptron (MLP), which has seven inputs, 400 hidden neurons, and matching outputs. The third method involves using a Multi-Layer SVM (ML-SVM), the idea of which is depicted in the diagram that follows. The MLP is described in sections.

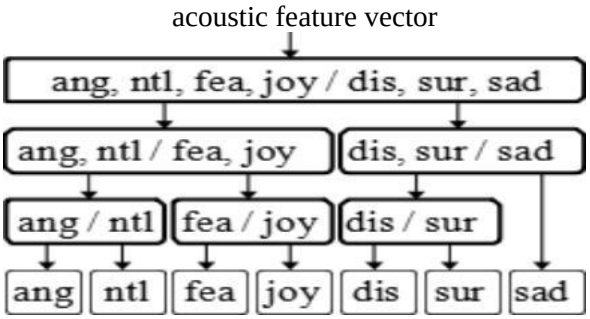

Figure 4: Best synchronization of emotions utilizing ML-SVMs.

Until only one class is left, the two-class choice procedure is repeated at each tier. Accurate identification depends critically on how emotion groups are arranged and how the layer structure is put together. According to our analysis, classes that are challenging to divide up should be separated last. This can be accomplished automatically with confusion matrices from the original SVM technique, or based on expert knowledge. This method's drawback, though, is that it cannot determine confidence for individual classes, which renders it inappropriate for merging. As a mapping function, the radial basis function kernel worked the best.

3.6 Classification Results

The classifier was assessed on a big corpus of voice data. Over the course of three cycles, two thirds, two-thirds of the data were used for training, while the remaining one-third was used for testing. The average error rates are shown in the table below, with standard deviations varying between ±0.01% and ±0.03%.With an exclusive training that is dependent on the speaker (S DEP) and evaluation that is independent of the speaker (S IND) were both taken into consideration.

Classifiers	S IND, Error, %	S DEP, Error, %
kMeans	57.04	27.38
kNN	30.41	17.39
GMM	25.16	10.89
MLP	26.84	9.35
SVM	23.88	7.05
SVM-MLP	24.55	11.3
ML–SVM	18.71	9.05

Figure 5: Comparison of the categories of acoustic features.

4. LANGUAGE INFORMATION

Usually, emotional information is only partially conveyed by a user's speech. Even in cases where emotional content is present, it often only shows up in little doses throughout the whole speech. Therefore, to efficiently discover keywords and sentiment expressions in natural language, a tracking strategy is essential. We use a typical (ASR.) system using a model based on a zero-gram hidden Markov model that offers the top n theories, including one-word confidence scores, to assess spoken language. Because belief networks are capable of handling ambiguous and partial data, we chose them as the mathematical foundation for our investigation. Though this page provides only a cursory introduction to belief network theory, it is becoming more and more popular as a solution to problems with pattern recognition. A belief network is made

up of a number of nodes that each have a limited number of alternative states and represent state variables, X. The directed edges that link these nodes indicate the conditional probabilities between every node as well as the parent node. The combined probability distribution represents the network's whole structure and conditional probabilities. Calculating the distribution may be done as follows, where N is the overall quantity of random variables.

$$P(X1,..., XN) = \tilde{O}P(Xi \mid parents(Xi)) \ldots\ldots\ldots \quad (1)$$

The network may infer the status of certain inquiry variables according to evidence variable observations. The objective is to identify the sentiment hypothesis that increases the likelihood of a collection of words given the acoustic data, much like in conventional techniques to natural speech interpretation. Every emotion has a unique network that represents it. During the initialization phase, root probabilities are equally distributed, corresponding to the prior probability of each emotion. A maximum likelihood choice is reached if emotional language data is understood independently. If not, each emotion's root probability is sent to a higher-level fusion algorithm, which works similarly to the method for acoustic confidence. The following graphic shows how the hierarchical clustering reduces mistakes to 18.9%: at the bottom four levels, words cluster together with superwords, followed by sentences, supersentences, and feelings.

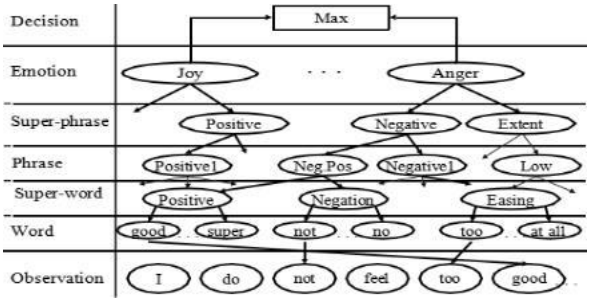

Figure 6: Using belief networks to identify phrases.

"...I don't feel well at all...."

Evidence supplied into the network at the word level according to the degree of confidence in the words that are actually seen. Tradition dictates that definite evidence be extended as uncertain evidence by integrating the ASR hypotheses' confidence level. Using a large corpus of manually labeled emotions, the training phase computes the numerical contribution P(ej|w) of each word w to the perception of emotion ej derived on the emotion's frequency of occurrence during observed speech.

4.1 Gentle integration of the solution

In this section, we focus on integrating the verbal and auditory datawe've gathered. Some studies suggest combining these sources using a late semantic Boolean OR approach, but this method is limited when dealing with more than two classes. Instead, we propose a more sophisticated approach: first, we calculate a paired average rating for every feeling informed by both sound and linguistic scores. Then, we use an adjacent maximum likelihood solution to make the final decision. This method has the advantage of using soft scores from both aspects before reaching a conclusion. However, this simple fusion method doesn't account for the varying levels of confidence in acoustic and linguistic estimations for each emotion.To address this, we adopt a more discriminative approach that considers all available emotion confidences in one decision process. We recommend using a Multi-Layer Perceptron (MLP) for this fusion task, as introduced in section 3.1. The MLP takes a 14-dimensional input feature vector, which includes seven confidence measures from both sound and language evaluations. The final sentiment probability is generated by the seven output neurons using a softmax function. Our tests demonstrated that optimal results were attained with 100 hidden neurons. MLP was trained on a separate dataset from the original training set, and its effectiveness was evaluated using a third dataset. The results, as shown in the table, were obtained utilizing the FERMUS III dialogue corpus with the best setup. Importantly, 12% of the statements included solely auditory details regarding the concealed emotion. The performance gain achieved by using the MLP-based fusion over the means-based fusion is evident, highlighting the benefits of our proposed approach.

Model	Acoustic Information	Language Information	Fusion by means	Fusion by MLP
Error., %	25.7	40.3	16.8	7.9

Figure 7: Performance enhancement refers to means-based and MLP integration.

5. CONCLUSION

Combining linguistic and auditory information has greatly propelled the field of speech emotion recognition (ser). By integrating these different approaches, researchers have made significant advancements in accurately recognizing human emotions through speech. A significant study conducted by Schuller et al. proposed a hybrid model that combines acoustic features, including pitch, energy, and spectral contours, with linguistic cues identified through belief networks. This method resulted in a significant decrease in error rates, achieving an impressive 7.9% error rate, underscoring the effectiveness of multimodal analysis in speech recognition. Researchgate: Support vector machines (svms) have gained recognition as a dependable classification method in ser due to their ability to handle complex data and their capacity to generalize well. When used to evaluate well-ranked acoustic features, svms have shown excellent accuracy in identifying emotional states. Additionally, incorporating linguistic cues, such as emotional keywords identified through belief networks, improves the system's performance. The achievement of integrating acoustic and linguistic data emphasizes the significance of multimodal analysis in speech recognition systems. By combining the analysis of both the words spoken (linguistic content) and the way they are spoken (acoustic features), these models can achieve a higher level of accuracy in understanding the speaker's emotional state. This progress has the potential to be applied in various fields, such as human-computer interaction, mental health monitoring, and customer service, where comprehending human emotions is of utmost importance.

6. REFERENCES

[1] Emotion recognition in human-robot interaction. Inf. Sci. **2020**, 509, 150–163.

[2] Hansen, J.H.; Cairns, D.A. Icarus: Source generator based real-time recognition of speech in noisy stressful and lombard effect environments. Speech Commun. **1995**, 16, 391–422.

[3] Koduru, A.; Valiveti, H.B.; Budati, A.K. Feature extraction algorithms to improve the speech emotion recognition rate. Int. J. Speech Technol. **2020**, 23, 45–55.

[4] Zheng, W.; Zheng, W.; Zong, Y. Multi-scale discrepancy adversarial network for crosscorpus speech emotion recognition. Virtual Real. Intell. Hardw. 2021, 3, 65–75.

[5] Schuller, B.; Rigoll, G.; Lang, M. Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture. In Proceedings of the 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing, Montreal, QC, Canada, 17–21 May 2004; pp. 577–580.

[6] Spencer, C.; Koç, İ.A.; Suga, C.; Lee, A.; Dhareshwar, A.M.; Franzén, E.; Iozzo, M.; Morrison, G.; McKeown, G. A Comparison of Unimodal and Multimodal Measurements of Driver Stress in Real-World Driving Conditions; ACM: New York, NY, USA, 2020.

[7] France, D.J.; Shiavi, R.G.; Silverman, S.; Silverman, M.; Wilkes, M. Acoustical properties of speech as indicators of depression and suicidal risk. IEEE Trans. Biomed. Eng. 2000, 47, 829–837.

[8] Uddin, M.Z.; Nilsson, E.G. Emotion recognition using speech and neural structured learning to facilitate edge intelligence. Eng. Appl. Artif. Intell. 2020, 94, 103775.

[9] Jahangir, R.; Teh, Y.W.; Hanif, F.; Mujtaba, G. Deep learning approaches for speech emotion recognition: State of the art and research challenges. Multimed. Tools Appl. 2021, 80, 23745–23812.

[10] Fahad, M.S.; Ranjan, A.; Yadav, J.; Deepak, A. A survey of speech emotion recognition in natural environment. Digit. Signal Process. 2021, 110, 102951.

[11] Jahangir, R.; Teh, Y.W.; Mujtaba, G.; Alroobaea, R.; Shaikh, Z.H.; Ali, I. Convolutional neural network-based cross-corpus speech emotion recognition with data augmentation and features fusion. Mach. Vis. Appl. 2022, 33, 41.12. Ayadi, M.E.; Kamel, M.S.; Karray, F. Survey on speech emotion recognition: Features, classification schemes, and databases. Pattern Recognit. 2011, 44, 572–587.

[12] Abdel-Hamid, O.; Mohamed, A.-R.; Jiang, H.; Deng, L.; Penn, G.; Yu, D. Convolutional neural networks for speech recognition. IEEE/ACM Trans. Audio Speech Lang. Process. 2014, 22, 1533–1545.

[13] Trigeorgis, G.; Ringeval, F.; Brueckner, R.; Marchi, E.; Nicolaou, M.A.; Schuller, B.; Zafeiriou, S. Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network. In Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 20–25 March 2016; pp. 5200–5204.

[14] Anvarjon, T.; Kwon, S. Deep-net: A lightweight CNN-based speech emotion recognition system using deep frequency features. Sensors 2020, 20, 5212.

[15] Rybka, J.; Janicki, A. Comparison of speaker dependent and speaker independent emotion recognition. Int. J. Appl. Math. Comput. Sci. 2013, 23, 797–808.

[16] Akçay, M.B.; Oğuz, K. Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. Speech Commun. 2020, 116, 56–76.

[17] Zhang, S.; Tao, X.; Chuang, Y.; Zhao, X. Learning deep multimodal affective features for spontaneous speech emotion recognition. Speech Commun. 2021, 127, 73–81.

[18] Pawar, M.D.; Kokate, R.D. Convolution neural network based automatic speech emotion recognition using Melfrequency Cepstrum coefficients. Multimed. Tools Appl. 2021, 80, 15563–15587.

[19] Issa, D.; Demirci, M.F.; Yazici, A. Speech emotion recognition with deep convolutional neural networks. Biomed. Signal Process. Control. 2020, 59, 101894.

[20] Sajjad, M.; Kwon, S. Clustering-based speech emotion recognition by incorporating learned features and deep BiLSTM. IEEE Access 2020, 8, 79861–79875.

[21] Badshah, A.M.; Rahim, N.; Ullah, N.; Ahmad, J.; Muhammad, K.; Lee, M.Y.; Kwon, S.; Baik, S.W. Deep features-based speech emotion recognition for smart affective services. Multimed. Tools Appl. 2019, 78, 5571– 5589.

[22] Er, M.B. A Novel Approach for Classification of Speech Emotions Based on Deep and Acoustic Features. IEEE Access 2020, 8, 221640–221653.

[23] Nicholson, J.; Takahashi, K.; Nakatsu, R. Emotion recognition in speech using neural networks. Neural Comput. Appl. 2000, 9, 290–296.

[24] Noroozi, F.; Sapiński, T.; Kamińska, D.; Anbarjafari, G. Vocal-based emotion recognition using random forests and decision tree. Int. J. Speech Technol. 2017, 20, 239–246.