**SPEECH DIARAIZATION RESEARCH AND EMOTION DETECTION**

CODING ASSIGNMENT-1

**SUBMITTED BY :**

R.GOPIKA SRI

STUDENT AT NXTWAVE

**INDEX**

|  |  |  |
| --- | --- | --- |
| S.NO | CONTENT | PAGE NO |
| 1 | Abstract | 1 |
| 2 | **Introduction** | 1 |
| 3 | **Selection of emotion categories** | 1,2 |
| 4 | **Type of emotion** | 2 |
| 5 | **Context of the speech diaraization** | 2 |
| 6 | **Results** | 3 |
| 7 | **Normality test** | 4,5 |
| 8 | **Conclusion** | 6 |

**Abstract:**

Speech diarization is defined as the process of partition an audio stream containing human speech into homogenous segments accordind to the identifynof each speaker. The total duration of the speech diraization is in excess of 7 hours containing 7000 utterances, and it is the largest emotional speech diarization available for this language. Twenty native speakers participated in the gender-balanced set, each recording of 10 sentences simulating seven targeted emotions. participated in the evaluation of Each audio clip of this corpus, except those of Disgust emotion, was validated four times by male and female raters perception tests. Kappa statistics and intra-class correlation coefficient scores indicated highlevel of inter-rater.

**Introduction:**

While communicating, people try to understand each other’s content of speech as well as the active emotion of the speaker. This is depicted by their body language and speech delivery. The study of emotion recognition is important to perceive human behaviors and their relationships both for social studies and human-computer interaction (HCI). It is also important to understand some physiological changes in humans. Research on Speech Emotion Recognition (SER) has been drawing increasing attention of researchers since the last two decades. The first requirement of a functional SER system is to develop a corpus containing useful emotional contents. Studies show that emotional expressions vary from culture may be different for other cultures. A specific language does not only represent some alphabetic symbols and rules, it also points to a specific culture. For cross-language experiments of emotions, the researchers need different emotional speech. Development of such an emotional corpus is considered to be relatively expensive as professional speakers are needed to express natural-like emotional speeches. Also, a high-quality audio lab is required for recordings to preserve the clear spectral details in the speeches for scientific analysis. The ultimate goal of this study was to build a validated emotional speech corpus as a linguistic resource that will be useful for prosodicanalysis .

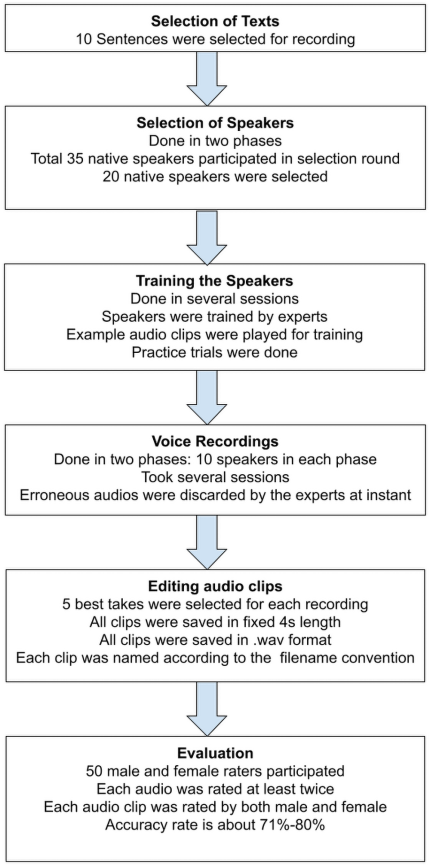
**Selection of emotion categories :**

The most challenging task of emotion recognition is to find the distinguishing features of target emotions. Researchers are still searching for a consensual definition of emotion. Since only voice data is being considered here for the data base design, for correct emotion recognition it is necessary to consider the emotional states which have intense impacts on voice data. For the speech dairaization development, the set of six fundamental emotions like anger, disgust, fear, happiness, sadness, surprise were considered along with neutral emotional state. These are probably the most frequently occurring emotions in all cultures around the world . Neutral can be compared to the Peaceful or Calm state. For speech dairaization this dataset, all the speakers had to simulate these seven emotions for the sentences. This s consists of an equal number of recorded audio clips for all emotions, indicating that the speech dairaization has balanced data in terms of desired emotional states.

**Type of emotion:**

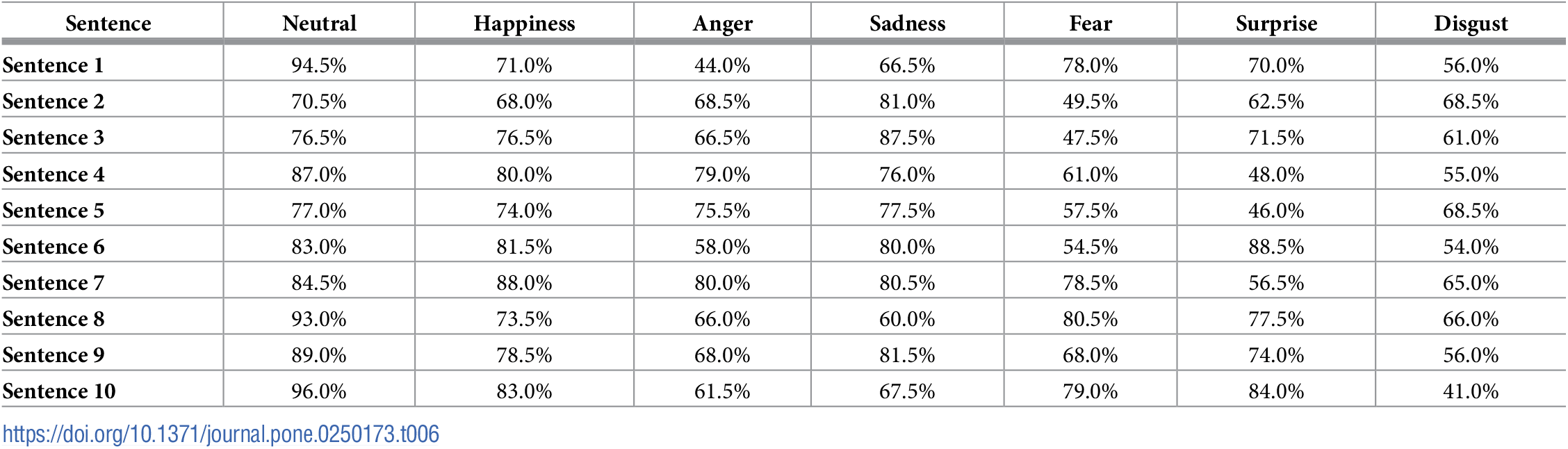
Databases of emotional speech corpus may be classified into three types based on the nature of the speech collected: speeches collected in-the-wild, simulated or acted emotional speech database, and elicited emotional speech database. In type 1, real-life emotional speeches are collected from a free and uncontrolled environment for analysis. For example, speeches collected from customer services, call centers, etc. But, the problem is that in-the-wild datasets have no ‘ground truth’, that is, there is no way to know the intended emotion of the speaker at the time of capture which is important for ML models. This kind of dataset also involves some copyright and legal issues. For this reason, they are often unavailable for public use. Type 2 database is developed by collecting acted or simulated speeches. Trained actors are asked to deliver speeches in different emotional states for predefined texts. Most of the available emotional speech databases are acted or simulated. The problem of these types of databases is that sometimes emotions are exaggerated and fail to represent naturally experienced emotions authentically [36]. For controlled scientific experiments, balanced data recorded in a laboratory environment are needed [37] which can be achieved by type 2 database. Type 3 is an elicited emotional speech database in which emotions are induced in speakers using some context. This is not acted in the sense that speakers are provoked to show the real emotions

**Context of the speech diaraization:**



**Results:**

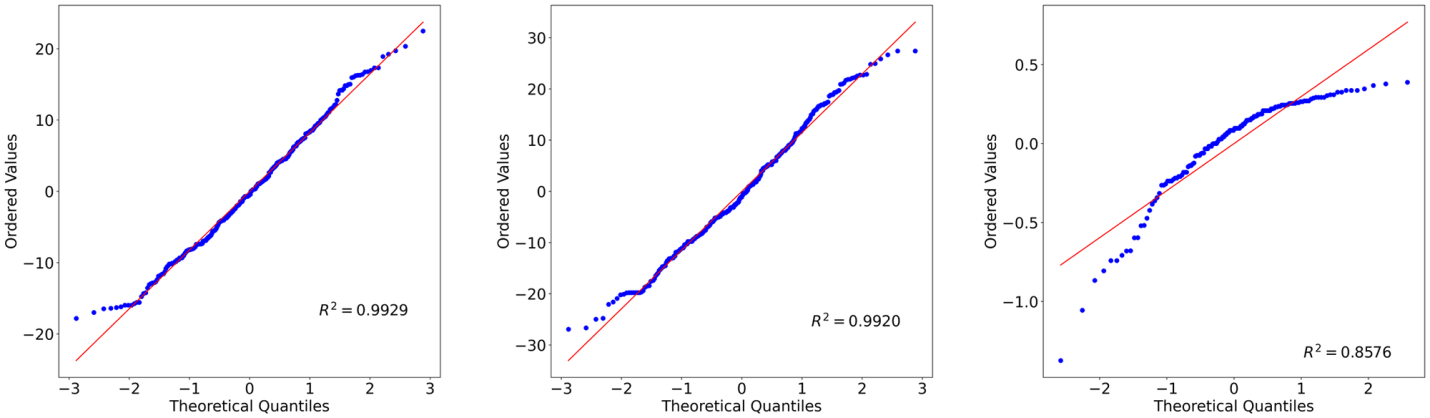
Phase of evaluation There were 7000 audio clips in this phase. Each audio was rated twice so that the total number of ratings was 14000 in this phase. The overall raw hit rate for this step is 71% with mean SD = 8.24. Table 3 represents the raw hit rate and unbiased hit rate for each emotion. It can be seen that the highest recognition rate was achieved by Neutral (raw = 85%, SD = 7.6), whereas Happiness achieved the second-highest rate (raw = 77.4%, SD = 9.8). Disgust has the lowest recognition rate (raw = 59.1%, SD = 9.6). From Table 4 it is obvious that the largest confusion occurred between Anger and Disgust (18.4%), which is more than 5% of the total ratings for this phase. More than one fourth Anger emotion audios were incorrectly recognized as





**Normality test:**

Before applying any statistical test on a dataset, it is worth investigating its probability distribution type. The Jarque–Bera and Shapiro-Wilk tests were applied on one-way ANOVA residuals to examine the probability distributions of data based on Rater Gender, Speaker Gender and Emotion. The null hypotheses of these tests assume that the target population is normally distributed. According to the test statistics, they were rejected for both tests except Rater Gender data for Phase 1 for the Jarque-Bera statistics .That means the normality assumption was not satisfied for those factors. Figs 5 and 6 present the probability plot distributions of the factors. It can be seen that the distributions are not normal, data points are skewed and removed from the fitted lines. Therefore, log transformation was applied to the data to normalize it. Transformation could not normalize all data except for Phase 1 Rater Gender and Emotion. Still, the Two-way ANOVA was performed as it was assumed that it is robust against normality assumption violation when the sample size is large and homogeneity assumption is satisfied.



**Conclusion:**

it is an audio-only emotional speech database containing 7000 audio files which was evaluated by 50 validators. Several statistical methods were applied to analyze reliability of the corpus. Good perception rates were obtained for human perception tests (up to 80%). Reliability indices also showed quite satisfactory results. Two-way ANOVA was executed to analyse the effects of Gender and Emotion. The normality and homogeneity of data for these factors was also investigated using Jarque-Bera, Shapiro-Wilk, Levene, and Bartlette tests. A high rate of reliability and consistency of evaluation task shows that this corpus should be considered as a valuable resource for the research on emotion analysis .