A Comparative Study of the Various Emotional Speech Databases

Vishal B. Waghmare, Ratnadeep R. Deshmukh, Pukhraj P. Shrishrimal
Department of Computer Science and IT,
Dr. Babasaheb Ambedkar Marathwada University,
Aurangabad-431004 (MS) India

***Abstract*— Speech emotional database and recognition is the challenging part of human computer interaction. The current research focuses towards the detection of emotion in various situations, while the database demands more to fetch out the work of recognition. The study investigates the various existing speech databases containing various basic emotions, enhancing the appropriate database development as well as transformation system for the emotion in speech. It will differentiate various existing techniques to distinguish the databases.**

Keywords- Speech Databases, Emotion, Speech analysis, TV, Hidden Markov Model.

I. INTRODUCTION

There are the different ways to express the emotions by humans. Humans express their emotions by speech and actions like crying, yelling, dancing, laughing, stamping, and many other things. But when it comes to speech human emotions affects the tone and the speaking style of the person. The emotional speech affects the Speech recognition accuracy. Now the researchers are taking interest in emotion detection by Speech. In human computer interaction, many researchers are finding the depth of the area for emotion detection from speech. The designing of emotional speech database is essential for the speech and emotion analysis to fulfill the need of researchers. The emotional speech database is needed for the further study of automatic speech recognition (ASR) and for robotics. The purpose of this work is to find out the better techniques to build up a proper database for the appropriate work to design and develop an emotional speech database which will allow emotional corpus based synthesis and the definition of the prosodic models of emotions for appropriate detection. The database might be useful to challenge the robustness of a variety of speech applications in automatic speech recognition systems. [1]

This paper gives comparative study of the various emotional speech databases developed by the various researchers, as well as the procedures used to build up database for the research requirements.

Mostly, the emotional speech databases are used for automatic speech recognition and speech synthesis. This paper contains the collection of various emotional speech databases containing the various basic emotions such as Happy, Sad, Angry, Surprise, Afraid, Neutral etc. in different spoken languages. For the comparative study we have taken some databases like Russian language speech database, a database with Emotional Utterances also German, Danish, Croatian and English. The major speech processing areas in a world are developing efficient algorithms for emotional speech recognition as well as emotional speech synthesis. To find out which technique is better to develop a proper emotional speech database, the collection of emotional speech databases is a prerequisite. In this paper, we have compared some emotional speech databases and also discussed a brief description and the features of collected databases.

The rest of paper contains one by one the description of various emotional speech databases, the comparison on the basis of comparative study and the conclusion to develop a proper database for emotional speech recognition.

II. EMOTIONAL SPEECH DATABASES

A. Russian Emotional Speech Database

The Russian emotional speech database contains the affective emotional utterances for the Russian language. Database contains the recorded data of 61 speakers (12 Male and 49 Female) with the age ranges from 16 to 28 yrs, which all are come from all over the country. The database contains ten pronounced sentences, which include all the phonemes and consonants in Russian, in five states of emotional expressions surprise, happiness, anger, sadness and fear. The database has been created to use it for the emotion recognition as well as for the production systems on speaker-and-gender-independent, gender-dependent and speaker-dependent levels. The database was created along with larger Map-task corpus of modern conversations in Russian; the database contains 3,660 utterances. All collected data was recorded on a portable Digital Audio Tape recorder Sony

TCD-D8 at 48 kHz sampling rate via Sennheiser headphone set in a soundproof recording studio of Department of Phonetics, St. Petersburg State University, St. Petersburg, Russia. The obtained recordings were converted into monophonic Windows PCM format at 32 kHz sampling frequency and 16 bits resolution. After creation of a database the prosodic features has been calculated for each audio file, fundamental frequency (F0) is calculated for all voice parts of speech signal with 40ms window and 10ms step between consecutive measurement, Energy or intensity is calculated with 10 ms window and 10 ms step. First three formants (F1, F2, and F3) and their bandwidths (BW1, BW2, and BW3) are calculated with 40 ms window and 10 ms step, also it varies length of utterance, percentage of pause and voiced part in utterance, speaking part, average length as well as mean, median and standard deviation. The database also provides a browser to explore the database without doing a programming.
The general issues occurs in this way is the types of emotions expressed in speech, and the relationship between genuine and simulated emotion. It has been suggested that emotions may be represented in speech at different levels. [2]

B. German Emotional Speech Database

German Emotional Speech database contains data collected from a German TV show recording, the 12 broadcasts of the talk show has been taken, which includes the dialogues and utterances, the utterances varies from the entire show were segmented in serious emotions of the speakers. The speaker's age ranges from 16 to 69 yrs. The TV show was chosen to build up a database because; of the spontaneous discussion between the guests is affective and such type of interpersonal communication leads to a variety of emotional states depending upon the topics. By using the three dimension emotional concept, an emotion is discussed in three basic entities (primitives): valence (negative – positive), activation (calm – excited) and dominance (weak – strong), due to this the expression pattern establishes emotion categories like Happy, Sad, Angry, Surprised etc. This method provides a 5-point scale for each emotion primitive in the range of [-1, +1]. The database is a collection of the spontaneous speech collected overall from the show, the communication between the guests on the various topics discussed; the topics are personal issues like friendship, fatherhood questions and romantic affairs. The collected audio was recorded at a sampling rate of 44.1 kHz stereo, and to extract as a wave file the signal was down sampled to 16 kHz (16 bit). Finally, the database has been carried out with 1018 emotional utterances by 47 speakers (11m/36f) with an average of 21.7 sentences per speaker. The average sentence duration was 3.0 s. The disadvantage of such a database from a TV show is that it is not possible to control the affective states that would occur in the cause of the dialogues. [3]

C. Danish Emotional Speech Database

The aim to develop a database for Danish emotional speech is to study the inter attitude variations from the recorded dataset, the Danish emotional speech database contains 4 speakers out of that 2 are male and 2 female, expressing the 5 emotions that are Neutral, Surprise, Happiness, Sadness and Anger. Recorded files are of 30 sec, total is of 10 min. The recording was done in an acoustically damped sound studio. The whole session was recorded on DAT tape at 48K sampling frequency in stereo mode. Later it was transferred as encoded data with 16 bits/sample and stored as 20 kHz binary. In order to get good approximation for the database the recordings were taken from the actors performing during the dramatic event. Mainly, the confidential part here in this database is that, the utterances has been taken is very short means the short sentences like "Yes", "No", this is because the short sentences or the short utterances are appropriate to analyze the emotional features, also to study the pause and specific emotional sounds like laughter or sighs. For the database, the recordings has been taken in such a manner as 2 single words, 9 sentences and 2 passages of fluent speech. The emotions were correctly identified in 67% of the cases, ranging from 55% to 80%. [4]

D. English Emotional Speech Database

This Emotional Speech Database was developed to explore how the acoustic and prosodic information can be used to detect the emotional state of speaker. For developing the corpus they have generated 50 sentences. These sentences comprised of questions, statements and orders. The mean sentence length was 5.8 words. The sentences were recorded from 5 drama students. They were asked to pronounce all 50 sentences in 5 different emotional moods i.e. happy, sad, angry, afraid and normal. They collected 250 sentences from each student. The data was recorded using Sennheiser HMD 410 or Sennheiser HMD 414 microphones with a sampling frequency of 16 kHz. The experiment on the database shows how acoustic and prosodic information can be combined and integrated into a HMM-based speech recognition system using suprasegmental states. It was shown that prosodic information is essential for a reliable detection of the underlying emotional state of a Speaker. [5]

E. Modern Greek Emotional Speech Database

The purpose of this database is to support the advance of emotion recognition technology and to assist for adaptation of this technology for real-world environments. The Modern Greek Real World Emotional Speech

Database Contains utterances recorded from 43 Speakers out of which 20 are females and 23 Females. The experimental setup used for developing this database was similar to smart home dialogue system. Each participant was asked to interact with the dialogue system in one session for approximately 25 minutes. The speech recognizer was fed from the output of the microphone array. The speech signal, as it was segmented by the speech recognizer, was stored to waveform files with sampling rate of 8 kHz, single channel, resolution 16 bits. The audio signal captured from the close-talking microphone was sampled at 16 kHz and stored into the same format. The participants used 10 cards consisting different tasks. The equal error rate (EER) observed for the speaker dependent system was 24.6%, while for the speaker independent system it was 33.4%. The experimental results got from the Real World Modern Greek emotional speech database can be considered as baseline performance which can be improved in future. [6]

F. Croatian Emotional Speech Corpus

It is the second speech database in Croatian language; the database consists of the neutral speech utterances consisting reports from radio speakers on news and weather. It is the first of its kind of emotional speech database in Croatian language. This corpus consisted of Approximately 40 minutes of affective speech segments. The first part called “real-life emotions” was collected from Internet, mostly from Croatian reality shows and from different documentaries from internet. The second part called “acted emotions” was collected from Croatian movies, TV Shows and Books-Aloud programs. The collected utterances were normalized and stored in ‘WAV’ format with 11 KHz sampling frequency, 16 bits per sample, monaural. A total of 714 utterances were collected with durations of 56:22 minutes. This was the first attempt to develop a Croatian Emotional Speech Corpus which provides encouragement for further improvements of training procedures. [7]

G. Serbian Emotional Speech Database

It is the first emotional Speech Database in Serbian language. The database consists of the recordings of following emotions neutral, anger, happiness, sadness and fear. The data was collected from 6 actors 3 of each gender. The database was collected in an anechoic studio. The data was collected from the speakers in separate sessions to prevent the influence of each other speaking style. The actors were asked to use their own everyday way of expressing the emotional states and not that of stage acting. The database consists of 32 isolated words, 30 short semantically neutral sentences, 30 long semantically neutral sentences and one passage with 79 words in size. The speech was recorded with a high quality microphone at a 44.1 kHz sampling frequency. Lately the recordings were transferred from DAT to PC computer with reduced sampling frequency of 22.050 kHz and stored in .WAV format. The listening test of Serbian Emotional Speech Database showed correct identification of emotions is 95% and the confusion that occurred between anger and happiness and between neutral and fear. [8]

III. COMPARISON

Overall the paper demonstrates the importance of the emotional speech database, for emotional speech recognition and synthesis. We have described briefly, the collected databases and compared some of them, with that of the initial standard input taken; the data collection from various situations, audio recordings, the instruments used for recordings, speakers and the percentage of the robustness comes after recognition.

We have studied and compared the various techniques used to develop a speech emotional database. The table shows that the basis on which we have compared these different databases, the taken standard Audio Recordings, the instruments or gadgets used for recordings, the Speakers and the scenario used for data collection.

After Studying the Databases we observed that in the Russian Emotional Speech Database the researchers observed a relative independence of 'genuine' and 'fake' emotions is demonstrated while acting and watching acting. For German Emotional Speech Database they presented the collection, segmentation and emotional labeling of many samples of spontaneous speech extracted from unscripted, natural discussions in a TV talk show. The emotion labels were given on a continuous valued scale for three emotion primitives: valence, activation and dominance, using a large number of human evaluators. For the Danish Emotional Speech Database the emotions were correctly identified up to 67% of the cases, ranging from 55% to 80%. After each listening test, the subjects were asked to state whether they found the task very easy, easy, neither easy nor difficult, difficult or very difficult. 75% of the listeners found it difficult or neither easy nor difficult to identify the emotions. For the English Emotional Speech Database they observed that the prosodic information is essential for a reliable detection of the underlying emotional state of a Speaker (7% Absolute Improvement). The researchers observed the equal error rate of 24.6% for Speaker dependent Modern Greek Database and an equal error rate of 33.4% for speaker independent system. The Recognition accuracy achieved for Croatian Emotional Speech database was 40% for which they have used Hidden Markov Models (HMM) with 5 basic emotions. The listening test of Serbian Emotional Speech Database showed correct identification of emotions is 95% and the confusion that occurred for emotion recognition was between anger and happiness and between neutral and fear in all basic emotions (i.e. Happy, Sad, Anger, Fear, and Neutral or Normal).

TABLE I. COMPARISON OF THE DATABASES

Name of Database	Audio Recordings	Instruments used for recording	Profession of Speakers	Situations used for Data Collection
Russian Emotional Speech Database	Sound Proof Recording Studio	Portable Digital Audio Tape recorder Sony TCD-D8 via Sennheiser headphone set	Common People	Sentences For Various Scenarios
German Emotional Speech Database	Studio	High Quality Microphones	Celebrities	Talk Shows
Danish Emotional Speech Database	Acoustically Damped Studio	AKG 414 ULS microphone, Amek Angela 36 ch. in -line Mixerdesk, PANASONIC DAT SV 3500	Actors	Various Scenarios
Detecting Emotion in Speech	General	Sennheiser HMD 410 or Sennheiser HMD 414 microphones	Drama Students	Questions, statements and orders.
Modern Greek Emotional Speech Database	Studio	SONY DCR-VX1000E, ACOUSTIC MAGIC Voice Tracker TM Array Microphone, AKG UHFPT40 (863.100 MHz)	Common People	Smart Home Scenario
Croatian Emotional Speech Database	Studio and Outdoor Location	High Quality Microphone, Cameras	Actors	Various Situations
Serbian Emotional Speech Database	Anechoic Sound Studio	High Quality Microphone	Actors	Various Situations

IV. CONCLUSION

Initially looking towards the comparative study, it is found that the total work for the development of emotional speech database requires more efforts and time, on the basis of speech recognition and speech synthesis. The comparative study illustrates the existing techniques to develop the emotional speech databases and also summarizes how to manipulate the emotion recognition needs. The quality of the database will be decided on the basis of intensity of emotions in the database. From the above Studied various seven databases we observed that for developing only two databases common people were been asked to come for the recordings. Whereas the other databases have been developed artificially using professional actors, celebrities and drama students who can imitate the emotions in somewhat real world affective states. Some of the databases contain utterances which are not spoken during the natural communication. These emotional speech databases are not classified with the intensity levels i.e. Very low, Low, Middle, High and Very high in natural emotions. That is why it does not measure level of intensity of emotions in the database.

As per our comparative study we found that the Serbian Emotional Speech Database which achieved an accuracy of 95% recognition of emotions in speech. The proposed methodology adopted for developing Serbian Emotional Speech Database is suitable for developing appropriate emotional speech database in various languages. The outcome of this comparative study will help other researchers in the field of emotional speech recognition to conduct and enhance their research by using the appropriate method for developing their own emotional speech database.

After studying the above emotional speech databases we have developed an emotional speech database from Marathi movies (i.e. artificial emotions). We will try to achieve a higher recognition rate for the basic emotions (i.e. Happy, Sad, Anger, Afraid and Neutral) for the developed database of emotions from Marathi movies.

ACKNOWLEDGMENT

The authors would like to thank the University Authorities for providing the infrastructure to carry out the research. This work is supported by University Grants Commission.

REFERENCES]

[1] Cowie, R. Douglas-Cowie, E. Tsapatsoulis, N. Votsis, G. Kollias, S. Fellenz, W. Taylor, J. G., "Emotion recognition in human-computer interaction", IEEE Signal Processing Magazine JAN 2001.

[2] Veronika Makarova1 and Valery A. Petrushin Meikai University, National Institute of Advanced Industrial Science and Technology, Japan, Accenture Technology Labs, Accenture, Chicago, USA "Ruslana: A Databse of "Russian Emotional Utterances" ICSLP-2002.

[3] Michael Grimm, Kristian Kroschel, Shrikanth Narayanan, "The Vera Am mittag German Audio-Visual Emotional Speech Database", IEEE, ICME 2008.

[4] Inger Samsø Engberg & Anya Varnich Hansen, "Documentation of the Danish Emotional Speech Database", Aalborg September 1996.

[5] Polzin T.S. and Waibel A. H., “Detecting Emotions in Speech”, Proc. CMC 1998.
[6] Theodoros Kostoulas, Todor Ganchev, Iosif Mporas, Nikos Fakotakis, “A Real-World Emotional Speech Corpus for Modern Greek”, LREC-2008.
[7] Branimir Dropuljić, Miłosz Thomasz Chmura, Antonio Kolak, Davor Petrinović,“Emotional Speech Corpus of Croatian Language”, ISPA-2011.
[8] Slobodan T. J., Zorka Kašić, Miodrag Đorđević, Mirjana Rajković, “Serbian emotional speech database: Design, Processing and Evaluation”, Specom’2004