Enhancing Conversational ASR Training with Modified Read Speech Corpora

Team

1. Faculty: Dr. M Uttara Kumari: / uttarakumari@rvce.edu.in and Dr.Saba Farheen N S: / sabafarheenns@rvce.edu.in
2. Students:
   a. Aryan Porwal / aryanporwal.ec21@rvce.edu.in
   b. Kumari Anjali / kumarianjali.ec21@rvce.edu.in
   c. Akansha Tanu / akanshatanu.ec21@rvce.edu.in
   d. Ahan Tejaswi / ahantejaswi.ec21@rvce.edu.in
3. Department: Electronics and communication

16 Sept 2023

Conversational ASR

Problem Statement

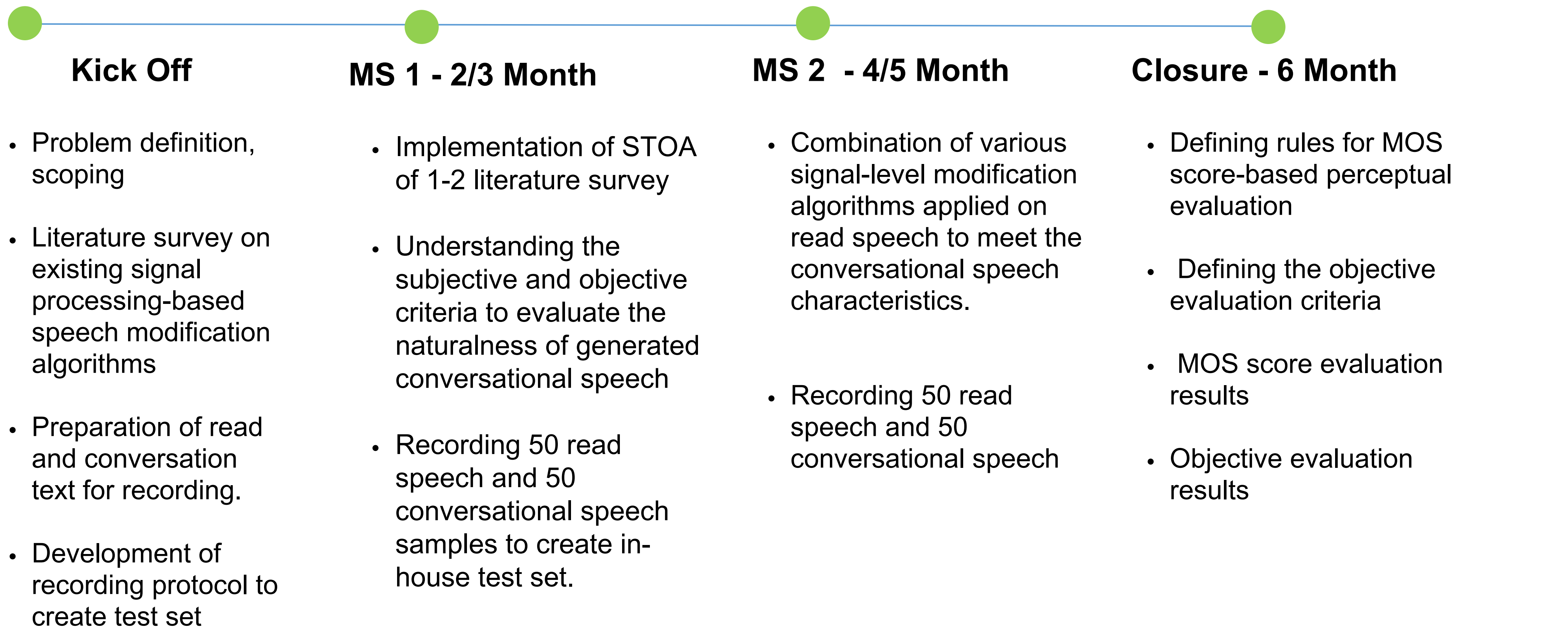
- Conversational ASR is important in speech understanding, call transcription, handling complex audios with code-switched, multi-lingual and multi-speaker cases.
- Most of the publicly available speech corpus are in reading mode but there is a limited availability of conversational corpus for ASR development.
- Modification of read speech into conversational speech may help in conversational ASR training.
- Generation of multi-speaker conversational speech by simply concatenating read speech recordings from multiple speakers may be natural as real-time conversations.
- The addition of filled pauses, dynamic speed and amplitude variation, repetitions of syllables and words and pitch modification on read speech will add conversational naturalness.
- The naturalness of conversational speech obtained by the modification of read speech is evaluated using subjective mean opinion scores and objective scores..

Expectations

1. Understanding of existing signal processing algorithms for conversational speech generation
2. Working on public datasets like Librispeech to generate conversational speech.
3. Recording in-house 100 read speech and 100 conversational speech samples.
4. Perform subjective and objective tests to evaluate the naturalness of generated conversational speech.

Training/ Pre-requisites

1. Having done with Signals &systems and Signal processing with EC/TC/EE branch
2. Knowledge of speech signal processing techniques like short time processing, pitch, amplitude, anduration modification.
3. Hands on python scripting language
4. Speech recording and analysis tools like PRAAT and audacity.
5. Training listeners for conducting mean opinion score (MOS) score test



Kick Off

- Problem definition, scoping
- Literature survey on existing signal processing-based speech modification algorithms
- Preparation of read and conversation text for recording.
- Development of recording protocol to create test set

MS 1 - 2/3 Month

- Implementation of STOA of 1-2 literature survey
- Understanding the subjective and objective criteria to evaluate the naturalness of generated conversational speech
- Recording 50 read speech and 50 conversational speech samples to create in-house test set.

MS 2 - 4/5 Month

- Combination of various signal-level modification algorithms applied on read speech to meet the conversational speech characteristics.
- Recording 50 read speech and 50 conversational speech

Closure - 6 Month

- Defining rules for MOS score-based perceptual evaluation
- Defining the objective evaluation criteria
- MOS score evaluation results
- Objective evaluation results

Literature survey and study

Name of the Paper	Authors	Year Of Publication	Key Findings
Speech emotion recognition using machine learning — A systematic review	Samaneh Madanian, Talen Chen , Olayinka Adeleye a, John Michael Templeton, Christian Poellabauer , Dave Parry d, Sandra L. Schneider	2023	• This research intended to collect the reported challenges in SER development. This article summarizes the main methods for SER since 2010. • It can demonstrate the ML trends for SER and forms a comprehensive guideline for developing robust SER models
Emotional speech Recognition using CNN and Deep learning techniques	C. Hema , Fausto Pedro Garcia Marquez	2023	• To create SER system capable of identifying human emotions with the help of the CNN algorithm and MFCC feature extraction efficient enough and able to produce accurate results and reduce the overall false rates • The average accuracy noted in the SER system created is about 78%
A review on speech processing using machine learning paradigm	Kishor Barasu Bhangale, K. Mohanaprasad	2021	• PCA, ICA, LPC, PLP, RASTA, MFCC, ZCR, WT are the Feature extraction techniques examined. • GMM, HMM,DTW, VQ, KNN, SVM, ANN, NB, LDA are the classifiers examined.

Literature survey and study

Name of the Paper	Authors	Year Of Publication	Key Findings
Speech Emotion Recognition Using Machine Learning	Kumari S, Perinban D, Balaji M, Gopinath D, Hariharan S J	2021	• The paper discusses the challenges of detecting emotions in speech and how machine learning can be used to overcome them . - There are three classes of features during speech: lexical features , visual features and acoustic features. • The proposed model in this paper uses dual recurrent neural networks to encode knowledge from audio and text sequences and predict emotion class. The model outperforms previous state-of-the-art methods in assigning data to at least one of four emotion categories (angry, happy, sad, and neutral).

Literature survey and study

Name of the Paper	Authors	Year Of Publication	Key Findings
Context-aware transformer transducer for speech recognition	Feng-Ju Chang , Jing Liu , Martin Radfar, Athanasios Mouchtaris, Maurizio Omologo, Ariya Rastrow, Siegfried Kunzman	2021	**1.Contextual Information-** Describes the different types of contextual information that can be used to improve speech recognition, including personalized and contextual phrases, named entities, device settings, and device locations. 2.**CATT Architecture** - Describes the architecture of the CATT network, which consists of an encoder, a context encoder, and a decoder. - Explains how the network uses attention context vectors to incorporate contextual information during inference. 3.**Context Encoding Techniques**- Discusses different techniques for encoding contextual information, including BLSTM and pretrained BERT models. - Presents experimental results comparing the performance of different context encoding techniques.
Deep Learning Techniques for Speech Emotion Recognition, fromDatabases to Models	Babak Joze Abbaschian , Daniel Sierra-Sosa and Adel Elmaghraby	2021	• Deep learning approaches have shown promising results in SER compared to conventional machine learning techniques. • Attention mechanisms and LSTM networks have been found to be effective in improving the accuracy of SER. • The availability of large-scale emotional speech databases has facilitated the development of deep learning models for SER.

Literature survey and study

Name of the Paper	Authors	Year Of Publication	Key Findings
Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis.	Ye Jia,Yu Zhang, Ron J. Weiss, Quan Wang , Jonathan Shen, Fei Ren , Zhifeng Chen , Patrick Nguyen , Ruoming Pang, Ignacio Lopez Moreno, Yonghui W	2019	**1.Neural network-based system-**The proposed model can generate natural-sounding speech audio in the voice of different speakers, even those unseen during training. Three independently trained components work together to produce natural-sounding speech. **2.Speaker Encoder Network**- The speaker encoder network is trained on a large and diverse speaker set, and embeds speaker identity information into a fixed-length vector that is used to condition the decoder network during synthesis. Embeds speaker identity information into a fixed-length vector . It is used to condition the decoder network during synthesis. **3.Decoder Network-** The decoder network generates a mel-spectrogram from text input and speaker embedding, and is trained on a multispeaker dataset. It can synthesize speech in the voice of any speaker seen during training. It converts mel-spectrogram to time-domain waveform . It is trained on a single-speaker dataset . It can also be fine-tuned for each speaker during inference.
Emotion Perception and Recognition from Speech	Chung-Hsien Wu, Jui-Feng Yeh, and Ze-Jing Chuang	2019	• Various recognition methods, including support vector machines, K-nearest neighbors, and neural networks, can be used for emotion recognition. • A practical approach to emotion recognition involves extracting intonation groups from speech signals and using Gaussian Mixture Models for each emotional state.

Literature survey and study

Name of the paper	Authors	Year	Key findings
End-to-End ASR: from Supervised to Semi-Supervised Learning Modern Architectures	Gabriel Synnaeve * 1 Qiantong Xu * 1 Jacob Kahn * 1 Tatiana Likhomanenko * 1 Edouard Grave * 1 Vineel Pratap 1 Anuroop Sriram 1 Vitaliy Liptchinsky 1 Ronan Collobert * 1 -.	2019	1. **OBJECTIVE**: The paper aims to enhance speech recognition models by incorporating semi-supervised learning techniques and modern architectures. 2. **DATA UTILIZATION**: The authors explore the impact of incorporating unlabeled audio data alongside labeled data and its effect on acoustic modeling. 3**PERFORMANCE IMPROVISATION:** Semi-supervised learning improves speech recognition model performance by combining labeled and unlabeled data. 4. **DATA QUALITY AND SPEAKER CHARACTERISTICS:** The quality of unlabeled audio data and the characteristics of speakers within it significantly influence model performance.
THE MICROSOFT 2017 CONVERSATIONAL SPEECH RECOGNITION SYSTEM	W. Xiong, L. Wu, F. Alleva, J. Droppo, X. Huang, A. Stolcke	2019	1.**Achievement of Low Word Error Rate:** The paper highlights the success of a neural network-based approach in achieving a 5.1% word error rate on the Switchboard evaluation set, demonstrating significant progress in conversational speech recognition. 2.**Neural Network Architecture:** The system utilizes a combination of convolutional and recurrent neural networks (CNNs and RNNs) to model acoustic features. This neural network architecture plays a crucial role in improving the acoustic modeling of speech. 3.**Advanced Language Modeling:** The paper discusses the incorporation of character-based and dialog session-aware LSTM language models. These models contribute to more accurate language modeling, enhancing the overall recognition system. 4.**Exploitation of Conversation-Level Consistency**: The authors highlight the importance of exploiting global conversation-level consistency in language modeling. This approach helps capture the context and coherence of conversations, leading to improved recognition accuracy.

Name of the Paper	Authors	Year Of Publication	Key Findings
Listen, Attend and Spell	William Chan	2019	1.High Accuracy Transcription: The LAS model is introduced as a neural network designed for transcribing speech into characters with high accuracy, surpassing the performance of traditional DNN-HMM models and other state-of-the-art models across various benchmark datasets. 2.Two-Component Architecture: The LAS model consists of two primary components: a listener and an attention-based spell. The listener encodes input speech into a sequence of feature vectors, while the attention-based spell decodes these vectors into character sequences. 3.Attention Mechanism: The paper emphasizes the significance of the attention mechanism, which enables the spell component to dynamically focus on different parts of the input sequence at different times. This attention mechanism enhances transcription accuracy by capturing context effectively. 4.Sensitivity and Mitigation: The LAS model's sensitivity to factors such as beam width, utterance length, and word frequency is acknowledged. However, the paper suggests that these effects can be mitigated through appropriate normalization and tuning techniques. 5.Versatile Applications: The LAS model is described as versatile, with potential applications in various domains, including speech recognition, speech synthesis, and natural language processing. Additionally, it can be trained on large speech datasets and adapted to new speakers with minimal additional data, making it adaptable for various scenarios.

Name of the Paper	Authors	Year Of Publication	Key Findings
Speech based Emotion Recognition using Machine Learning	Girija Deshmukh , Apurva Gaonkar , Gauri Golwalkar , Sukanya anil Kulkarni	2019	• Three feature vectors (pitch, Mel frequency cepstral coefficients and short term energy) are used to classify three emotions (anger, happiness, and sadness).The dataset used for training and testing consists of audio samples in male and female voice and is divided in a 4:1 ratio. • The paper compares the accuracy of two methods (mode and mean) for classifying emotions and finds that the mode method is not satisfactory.The authors use the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) to test their system. • The system is programmed using MATLAB R2014a and uses a multiclass support vector machine (SVM) classifier to form a model corresponding to every emotion. The test signal is tested with every model in order to classify and detect its emotion.

Literature survey and study

Name of the Paper	Authors	Year Of Publication	Key Findings
Speech Recognition Using Deep Neural Networks	Ali Bou Nassif , Ismail Shahin , Imtinan Attili, Mohammad Azzeh , And Khaled Shaalan	2019	• Deep learning models, such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), were commonly utilized. The papers highlighted the advantages of deep learning in improving speech recognition accuracy compared to traditional methods. • The machine learning algorithms used in the papers include deep neural networks, support vector machines (SVMs), hidden Markov models (HMMs), Gaussian mixture models (GMMs) and k-nearest neighbors (KNNs). • The techniques used for feature extraction from speech include the linear discriminate analysis (LDA) transform, the HLDA transform and the short time Fourier transform (STFT).

Literature survey and study

Name of the Paper	Authors	Year Of Publication	Key Findings
Emotional Speech Generation by Using Statistic Prosdy Conversion Method	Jianhua Tao and Aijun Li	2018	• The Deviation of Perceived Expressiveness (DPE) measure is created to evaluate the expressiveness of the output speech. • The GMM method is found to be more suitable for a small training set, while the CART model gives better emotional speech output if trained with a large, context-balanced corpus. • The models aim to map the subtle prosody distributions between neutral and emotional speech.
Developments in Corpus-Based Speech Synthesis: Approaching Natural Conversational Speech	Nick CAMPBELL	2018	• A framework for specifying affect through differences in speaking style and voice quality can be used to create more natural and conversational speech synthesis systems • The focus of unit selection in speech synthesis should shift from segmental or phonetic continuity to prosodic and discoursal appropriateness in order to synthesize conversational speech • Consideration of both linguistic and paralinguistic information, such as affective states, is crucial in speech synthesis to make it more conversational and natural

Literature survey and study

Name of the Paper	Authors	Year Of Publication	Key Findings
Human speech emotion recognition	Maheshwari Selvaraj, Dr.R.Bhuvana, S.Padmaja	2016	• Gender speech classifier is based on pitch analysis. MFCC approach for emotion recognition from speech • Radial Basis Function and Back Propagation Network is used to recognize the emotions based on the selected features, Radial basis function produces more accurate results for emotion recognition than the back propagation network
Expressive Speech Synthesis: Past, Present, and Possible Futures	Marc Schroder	2016	• Adequate annotation of speech units and careful manual selection are necessary to produce conversational speech of unprecedented naturalness • The inclusion of nonverbal vocalizations, such as laughs and speech "grunts," is crucial for capturing the subtleties of meaning in conversational speech • Converting read speech to conversational speech requires a database of everyday speech as the unit selection database

Literature survey and study

Name of the Paper	Authors	Year Of Publication	Key Findings
Machine Learning Approach for Emotion Recognition in Speech	Martin Gjoreski, Hristijan Gjoreski,Andrea Kulakov	2014	• This approach improves upon the existing methods by performing a thorough ML analysis, including methods for: feature extraction and standardization, feature selection analysis, algorithm selection analysis, and algorithm parameters optimization
Speech Recognition by Machine	M.A.Anusuya, S.K.Katti	2009	• Automatic Speech Recognition (ASR) is a challenging task due to the variability in speech signals, including differences in pronunciation, speaking rate, and background noise. ASR systems typically involve three main components: acoustic modeling, language modeling, and decoding. • Acoustic modeling involves mapping acoustic features of speech to phonetic units, while language modeling involves predicting the probability of word sequences.

Thank you