Tuesday, 21st Nov '23

International Institute Of
Information Technology

PROFESSOR

Dr. Anil Kumar Vuppala

SPEECH SIGNAL PROCESSING

Final evaluation
Presentation

Excitation Features for Emotion Recognition

Soma Satya Pradhith Vulichi 2021102015

Ajay Ray 2021102032

Table of Contents

I Introduction

II Extraction of Excitation Features

III Work done

IV Work done and Observations

V References

Introduction

- In generation of emotional speech, there are deviations in the speech production features when compared to neutral (or) non-emotional speech.

- In this project we have captured the deviations in these features that are related to the excitation component of speech.

- The excitation features used in the project are the instantaneous fundamental frequency (F0), the strength of excitation (SoE), the energy of excitation (EoE) and the ratio of the high-frequency to low-frequency band energy (β)

II Extraction of Excitation Features

- The excitation features which were discussed earlier are extracted using the zero frequency filtering (ZFF) method, linear prediction (LP) analysis and short-time Fourier transform (STFT).

- The glottal closure instants (GCIs) of speech are obtained using the ZFF method, where the speech signal is passed through a cascade of two ideal resonators located at 0 Hz, followed by trend removal. The resultant signal is called ZFF signal.

- The negative-to-positive zero crossings of the ZFF signal correspond to the GCIs.

- The interval between two successive GCIs gives the fundamental period T0. The instantaneous fundamental frequency is given by F0 = 1/T0.

II

Extraction of Excitation Features

- The slope of the ZFF signal at each GCI is called the strength of excitation (SoE), which is related to the amplitude of the impulse-like excitation in most cases.

- Linear prediction (LP) residual gives an approximation of the excitation component of the speech signal. The energy of excitation (EoE) parameter is computed from the samples of the Hilbert envelope of the LP residual over a 2-ms region around each GCI.

- A segmental feature, which is the ratio between the high-frequency and low frequency spectral energy (β) used for discriminating shouting and neutral speech.

- The β measure is computed as the ratio of the high-frequency band (800–4000 Hz) energy to the low-frequency band (0–550 Hz) energy from short-time Fourier magnitude spectrum of speech signal.

III Work done

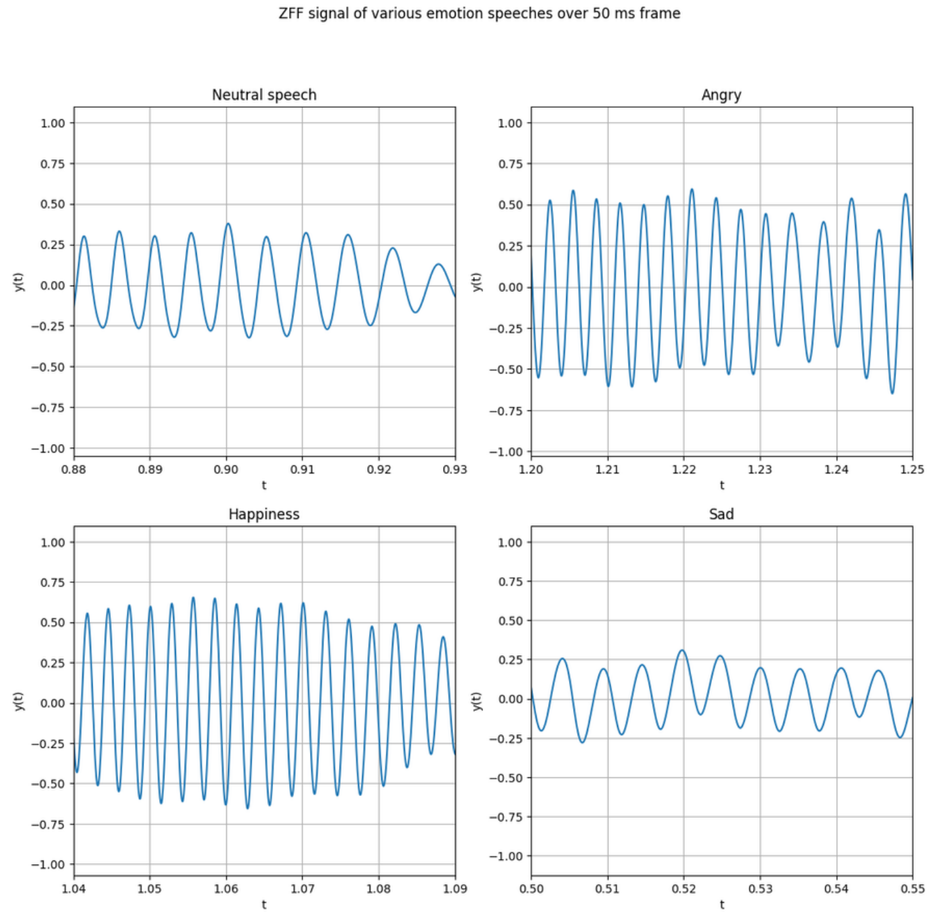
- We are using the Berlin Emotional Database (**EMO-DB**) for our project. We have annotated all the audio files by scripting to categorize them under different emotions and different genders as well.

- Since in literature, most of the study on emotion recognition have been done using features representing vocal tract system characteristics, we have built a model using random forest method that takes **MFCC features** for each file in the dataset as input data and was giving accuracy of approximately **59%**.

- We then, extracted four excitation features and built model using random forest method, taking different combinations among the features and compared the resultant accuracy with that of MFCC features.

III Work done

- Since for extracting instantaneous fundamental frequency and SoE is done from ZFF signal, we extracted ZFF signal by differentiating the signal and then pass through the filter twice followed by mean subtraction. The ZFF signal for different emotions are plotted below:



III Work done

- From the plots, it can be said that the time interval between two successive GCIs is less for happiness and angry speech as compared to neutral. Hence, the fundamental frequency is observed to be high.

- Similarly, the time interval between two successive GCIs is observed to be slightly higher for sad speech as compared to that of neutral. Hence, pitch is found to be lower.

- Also, it can be seen that due to lesser T0 in case of angry and happy speech when compared with neutral, the strength of excitation appears to be lower than that in neutral speech.

- In case of sadness, SoE is observed to be slightly higher since T0 is greater for sad than that of neutral speech.

- Probably, the reason for such trend could be that in order to maintain high rate of vibration, the vocal folds may not close with suction which results in lower values of SoE for lower T0.

III Work done

- We have computed loudness parameter by computing STFT at window level first, with the window length and window shift of 512 and 256 respectively.

- We set up upper and lower limits for defining the high frequency and low frequency band and then computed the energy of both bands and found out the ratio of them.

- For extracting energy of excitation, we found out GCIs. Then, took the Hilbert envelope of LP residual of the part of signal which is 2ms around each GCI. And then found the energy of the envelope found.

IV Work done and Observations

- After extraction of all four features, we then built models training them with each feature individually and also by taking pairwise. On doing so, we observed the following accuracies:

Feature	Accuracy of the model with that feature (in %)	Feature combination	Accuracy of the model with the feature combination (in %)
F_0	54.62	F_0 and β	44.75
β	46.29	F_0 and SoE	45.37
SoE	42.59	β and SoE	49.07
EoE	33.33	β and EoE	41.66

IV Work done and Observations

- We then built model by training them by taking combination of three features at a time and at the end trained by taking all four features at a time:

Feature combination	Accuracy of the model with the feature combination (in %)
F_0, β and SoE	53.70
F_0, β and EoE	45.37
F_0, SoE and EoE	43.51
F_0, β, SoE and EoE	51.85

IV Work done and Observations

- It can be observed that among all possible combinations, MFCC was giving better accuracy because they capture information related to the spectral characteristics of the signal.

- On the other hand, excitation features may primarily capture information about the source characteristics, such as pitch or voicing, which might be less informative in this case.

- It can be observed that the model gives better accuracy when trained only with pitch as pitch also contains lesser information about frequency components(only F0) which might be the reason for giving better accuracy.

- It was observed that model gave lesser accuracy when trained with EoE only. Hence the net accuracy when the model was trained with 2-3 features at a time was lesser compared to that of pitch only.

- On training the model with all features, we could observe accuracy of **51.85%** which is approximately **7% lesser** than that of MFCC.

References

- Excitation Features of Speech for Emotion Recognition Using Neutral Speech as Reference by Sudarsana Reddy Kadiri, P. Gangamohan, Suryakanth V. Gangashetty, Paavo Alku and B. Yegnanarayana, for feature extraction.

- Epoch Extraction From Speech Signals by K. Sri Rama Murty and B. Yegnanarayana, for ZFF extraction.

Tuesday, 21st Nov '23

International Institute Of
Information Technology

PROFESSOR

Dr. Anil Kumar Vuppala

SPEECH SIGNAL PROCESSING

Final evaluation
Presentation

Thank you for listening!

Soma Satya Pradhith Vulichi 2021102015

Ajay Ray 2021102032