Speech Emotion Recognition Experiments on a Speech2Text Dataset

Aayusha Siriano *

Department of Information Technology, Maharaja Agrasen Institute of Technology, New Delhi, India

Short Communication

Received: 10-Aug-2022

Reviewed: 24-Aug -2022

Published: 31-Aug-2022

***For Correspondence:**

Aayusha Siriano, Department of Information Technology, Maharaja Agrasen Institute of Technology, New Delhi, India

E-mail: **aayushasiriano@gmail.com**

Keywords: ASR; Wavelet; HMM; MFCC; GMM classifier; Random forest classifier

ABSTRACT

The main objective for this project was to study the performance of non-linear speech analysis methods in automatic speech recognition. Specifically, we selected wavelet transform as a promising non-linear tool for signal analysis that has been already successfully applied in many tasks, such as image recognition and compression leading to standards such as JPEG2000. The plan was to perform a comparative analysis between the standard mel–cepstral and wavelet-based set of features and to evaluate the baseline speech recognition rates of the two aforementioned parameterization methods

INTRODUCTION

Speech is the most natural means of communication between humans. With the development of information technology and the massive use of computers. Man-Machine Dialogue (MMD) using the word as a means of communication has been an increased interest from both the scientific and the industrial community. Automatic Speech Recognition (ASR), the main component of the MMD system, is a central topic in the broader one Natural Language Processing (NLP) domains. The general structure of an HMM-based speech recognition system [1] consists of two phases: a learning phase whose goal is the construction of acoustic models (HMM models) and recognition phases in which the most likely word being imposed. Generally, ASR systems use cepstral parameters called standard parameters as an acoustic representation of the speech signal. Cepstral parameters currently the most successful are the MFCCs coefficients (Mel Frequency Cepstral Coefficients). The procedure of calculation of

5	0.9893287	0.930556	0.7466161
6	0.786531	0.8695604	0.888889
7	0.642375	0.833333	0.9684608
8	0.4356444	0.833333	0.9867778

Table 2: We also used the same skip rate of speech window with the value of 10 ms.

Emotion	MFCC 128ms window width, 50-PC's	Continuous WaveletTransform: 30-PC's	Discrete Wavelet Transform: 25-PC's
1	0.656575	0.888889	0.6754172
2	0.65403	0.9345857	0.888889
3	0.33332	0.7534839	0.861111
4	0.656575	0.6754172	0.888889
5	0.9893287	0.7466161	0.930556
6	0.4356444	0.833333	0.9867778
7	0.642375	0.9684608	0.833333
8	0.4356444	0.9867778	0.833333

DISCUSSION

Tests that give the most relevant information are the tests on city names and the phonetically rich words. These tests are representative due to the diverse phonetic content and can serve as a baseline for judging the overall success of the parameterization methods involved [6]. Slovenian SD2 experiments exhibit a small improvement of the recognition results with the wavelet features. We could hypothesise that the variable frequency resolution in the wavelet transform enhances the overall recognition rate. We also tested the recognition performance using a 32 ms speech window in the MFCC calculation. Unequal MFCC and WPP window durations were therefore not considered to be problematic since the MFCC recognition scores were found to be consistently worse for longer window durations. The English SD2 experiments yielded consistently better results obtained by the wavelet features. This could imply that the wavelet features are more robust in the variable noise conditions. During the experiments we observed the appearance of side lobes in band pass filters that cut out frequency content of the signal [7]. This is due to the non-optimal separability of conjugate mirror filter that implements the Daubechies 2 mother wavelet. Another observation was a different phone level alignment between MFCC and WPP features.

CONCLUSION

Our study is to apply the Wavelet Transform (WT) in the acoustic analysis of speech signals to an ASR task .To accomplish this task and to evaluate the acoustic analysis based on the WT, we built a reference ASR system based on HMM models. Acoustic analysis of this reference system is based on the extraction of MFCC parameters. This system is built under the platform HTK and evaluated on the basis of OLLO database. The construction of the reference system calls for the type of acoustic parameters and the number of Gaussian components for each active state HMM. To this end, we conducted various experiments to determine these two parameters. The results showed us that the most relevant acoustic parameters are MFCC_E_D_A coefficients. The results also showed that the choice of 25 Gaussian components provides a good compromise between recognition accuracy and computation time. In the present work, we chose the DWT based on dyadic decomposition to extract the WCC parameters.