International Journal of Innovative Research in Computer Science & Technology (IJIRCST)
ISSN: 2347-5552, Volume-9, Issue-6, November 2021
https://doi.org/10.55524/ijircst.2021.9.6.12
Article ID IRPS12022, Pages 52-55
www.ijircst.org

Feature Extraction and Processing Analysis in Speech Recognition

Mrinal Paliwal[1], and Pankaj Saraswat[2]

[1,2] SOEIT, Sanskriti University, Mathura, Uttar Pradesh, India

Correspondence should be addressed to Mrinal Paliwal; mrinalpaliwal.cse@sanskriti.edu.in

ABSTRACT- The difficulties with automated identification and synthesis of various speech patterns have become significant research issues in recent years. Stress-induced speech characteristics were compared to normal speech in a feature analysis. Due to stress, the performance of Stressed speech recognition decreases substantially. In the speech communication system, the voice signal is transmitted, stored, and processed in a variety of ways. The speech signal must be delivered in such a way that the information content may be easily extracted from human listeners or machine automation. To enhance speech recognition performance, a stressed compensation method is employed to compensate for stress distortion. To identify different moods in speech signals, these features are collected and assessed in English. The variations in glottal excitement of common speaking patterns are examined in depth in this article. The sinusoidal model effectively describes the different stress classes in a speech signal, according to the results. When it comes to detecting emotions in a pressured speaker, sinusoidal features outperform linear prediction features.

KEYWORDS-Feature Extraction, Processing, Sinusoidal Model, Speech recognition, Stressed.

I. INTRODUCTION

Two key processes are undertaken by the speech recognition system: signal modelling and pattern matching. Signal modelling is the process by which a signal is transformed into a collection of parameters[1]. The objective of matching the pattern is to locate the memory parameter set that closely matches the parameter set of the input speaker signal. Research in speech recognition began six decades ago and more academics and scientists still need to participate in this field, as voice-controlled applications should cover many parts of the future of everyday life[2]. Many business services and firms, like banks and payphone service providers, opened up touch-tone telephone services years ago, offering clients much convenience and better service and saving businesses time and labour. Stressed speech recognition technology in mental diagnostics, toys, and lying detector was widely utilized. Continuing study into stressed recognition of speech will certainly help people address many difficulties. Various challenges in the identification of practical speech can be described as follows[3]:

- variant of the speaker
- ambiguity and acoustic variables on phonemic variables are not precisely mapped on one
- different speech variants under stress
- interference and noise.

Stressed speech is defined as a language generated under any situation when the speaker differs from neutral speech output. Emotional and environmental variables like noise are the sources of perceptible stress. Classification of stress is an automated stress detection of the voice signal. Evaluation of speech stress has applications such as the sorting of emergency telephones, telephone banking and hospitals(1). Stressed speech analysis can give fresh and significant information that helps to better recognize speech, synthesize and verify the speaker automatically.

The speech that is spoken in rage is less long than that which is spoken in neutral emotion. When the voice signal is expressed with angry emotion, the average signal amplitude is greater. The spectrogram shows that the frequencies have moved upward or have greater values in the angered signal in comparison to the neutral emotionally expressed speech[4].This indicates that speaking signal qualities alter under different moods or stressful circumstances.

The patterns revealed by the characteristics of speech not only rely on emotions, but also on the language. The performance of the recognition depends on the characteristics of the voice signals. For analyzing and classifying stressed language signals, linear prediction (LP) model features and cepstral features utilized in different speech applications have been checked.

A. Analysis of Pitch in The Speech Samples

Pitch is the most frequently studied stress analysis parameter. The study covers pitch outlines, average pitch statistics, variable pitch and pitch distribution.

Connaissance of some pitch features and patterns may be gained following consideration of vast amount of pitch contours and observation of some remarks[5]. These remarks are supported by the example contours. For soft speech, the average pitch was usually reduced. Smoother than Neutral was soft speech too.Angry language exhibited very uneven contours of the pitch and also had the greatest mean and variability of all stress situations. Towards the end of every speech the questions pitch contours were increasing. The starting pitch and variability were equal to Neutral (in the time domain). Toward the conclusion of most utterances, variability was consistent[6]. Unlike the large variations seen on Angry, the rise in the pitch from Question style near the conclusion of the pronunciation was related to the lexical stress.

B. Analytical Models Regarding Stressed Speech

There are several algorithms that may be used to monitor pattern differences in different fields. Some of these algorithms are HMM, LPC, TEO, etc. Some of them are HMM.

1) Hidden Markov Model (HMM)

A hidden Markov process may be seen as an extension of the issue of urn substitute in its discrete version (where each item from the urn is returned to the original urn before the next step). Please take this example: there is a genius in a room not apparent to an on-looker[7]. This chamber has tools X1 x2, X3, each with a known ball mix, each with a y1-y2, y3 marked ball in this area the genie picks an urn and draws a ball from it allegedly as shown in the Figure 1. Then it places the ball in a conveyor belt, where the spectator may see the sequence of the balls, but not the urn sequence.The genie has some method for selecting urns; the urn selection for the nth ball relies simply on the random numbers and the urn selection for the (n1)th ball[8]. The choice of urn does not depend directly on the urns selected before this preceding urn; this is thus known as a Markov process.

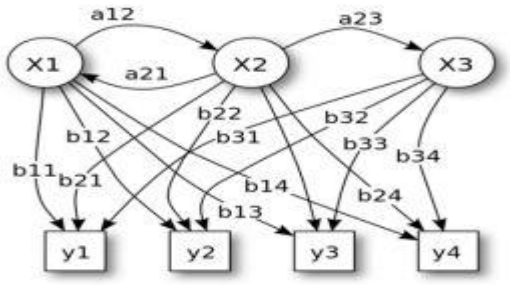

Figure 1: Hidden Markov Process as an Extension of the Substitute in its Discrete Version

2) Linear Predictive Coding (LPC)

Linear predictive coding (LPC) is an audio and speech processing technique used mainly to represent a compressed digital speech signal's spectral envelope utilizing linear predictor model information. This is one of the most efficient speech analysis techniques and one of the most effective ways for coding high speech quality at a low bit rate and gives very precise assessments of voice characteristics. LPC begins on the idea that a buzzer produces a voice signal on the end of a tube, with hissing and popping sounds occasionally added (sibilants and plosive sounds). Although this model appears to be rudimentary, it really is near to the reality of speaking production[9]. The glottis creates a buzz that is defined by strength (loudness) and frequency (distance between vocal folds) (pitch). The tube creates the vocal tract (throat and mouth), which is distinguished by its resonances that give birth to shapes and increased sound frequency ranges. Tongue, lips and throat movement during sibilants and plosives generates hisses and pops.LPC evaluates the spoken signal by calculating the formants, eliminating their effects from the spoken signal and evaluating residual buzz strength and frequency. Reverse filtering is termed the process of eliminating the formants, and the residual signal is called the residuum following the subtraction of the filtered model signal[3].

3) Speech Processing and Feature Extraction

Voice processing and extraction of characteristics is a very essential stage in obtaining representatives from a speech signal ready to analyses, compensate and recognize characteristics. Speech enhancement is another stage prior to speech processing. Noise has been cut at the front and back-end and the speaking signal length is fixed in order to simplify it and if required, is added to zero. Speaking signals are captured at a rate of 8 kHz in this article. The LPC analysis is performed to extract characteristics in the frequency domain from the spoken stream[10].A 32ms hamming weighing window is added to each frame, i.e. w(n), to reduce the pressure from the original continuous speech signal to cuts the finite sampling window (256 points). The weighing function of the hamming window is the following:

$$w(n+1) = 0.54 - 0.46 * \cos\left(\frac{2\pi n}{N-1}\right).$$

4) Teager Energy Operator(TEO)

Teager's speech and hearing tests, which provide a measure of the energy of a speech signal, inspired the measurement of the speech signal's energy. In these experiments, Teager discovered that airflow in the vocal tract is split and adheres to the vocal tract walls. Teager proposed the language productions theory based on these facts, vocal tract geometry, and the results of specific whistle cavity experiments. The air is expelled as a jet from the glottis and attached to the nearest vocal tract wall in this idea. A vortex of air forms between real vocal folds and false vocal folds when air flows through the cavity. The majority of air continues to flow to the lips despite the walls of the vocal tract. The TEO, like the traditional power operator, is used

to calculate signal energy. The TEO for a permanent time signal is:

$$\varphi(x(n)) = \frac{d}{dt}x(t)^2 - x(t)(\frac{d^2}{dt^2}x(t))$$

Where this TEO is the discrete signal used, is defined as:

$$\varphi(x(n)) = x(n)^2 - x(n+1)x(n-1)$$

Where $x(n)$ is the voice signal sampled, $\varphi$is the energy operator Teager. Teager is a non-linear operator to compute energy instantaneously with an enhanced signal to noise ratio (SNR). The recording and processing equipment is always linked with some noise. The noisy portion of the signal is removed by TEO when its energy is calculated. Noise energy is therefore not taken into account. In contrast, the normal squared energy operator takes the input signal and calculates the energy together with the noise present.

5) Stressed Speech Using Sinusoidal Model Features

The voice signal may be seen via a time-differing linear filter which simulates the resonant features of the vocal tract, by means of a glottal excitement waveform. This is illustrated by the sinusoidal speech model. The voice signal is split into fixed signal segments or frames as shown:

$$s[n] = \sum_{(k)=1}^{M} \hat{s}[n-(k)N]$$

where k is the frame index and N is the length of the frame. $\hat{s}$is a sum of sinusoids given by:

$$\hat{s}[n] = \sum_{j=1}^{L} A_j^{(k)} \cos(2\pi f_j^{(k)} \frac{n}{F_s} + \emptyset_j^{(k)})$$

Where Fs is sample frequencies as operated in regards to $\hat{s}[n]$ and the value of L is always taken as 10 here.

A mathematical instrument to restrict the input signal is a window function. That is, only a specific input signal interval is allowed while the outer signal interval is restricted. This study can thus show that a window function is a time domain filter that only permits the signals to pass a specific interval while the signal is attenuated beyond the given period. There are several different sorts of window functions, such as rectangle, hamming, blackmann, etc. There is a rectangle window:

$$w(n) = \begin{cases} 1, & 0 < n \leq (N-1) \\ 0, & otherwise \end{cases}$$

where,

N is the overall number of signal samples. The window function in this work is the spectral efficiency hamming window, which is studied further. Window hamming is defined as:

$$w(n) = 0.54 + 0.46\cos\left(\frac{2\pi n}{N}\right)$$

Where, N is the total number of input signals samples. The idea of digital communication shows that a band signal is not time-bound, and that a time-bound signal is not restricted to a band. Therefore, if you do not use a window approach, you obviously use a rectangle window without knowing it This study limits the signal time which resulted in a significant spectrum flow of data into the frequency domain. But if a hamming window is used here but this analysis compromises the amplitude of the signal, the frequency field representation of the signal and a less frequency leakage will be increased [10].

This loss of signal amplitude due to the window function can be mitigated by using the notion of window overlapas shown in the Figure 2. Not alone will the signal be approximated better but spectral leakage will be reduced as well. A 50% overlapping window feature is utilized in this study.

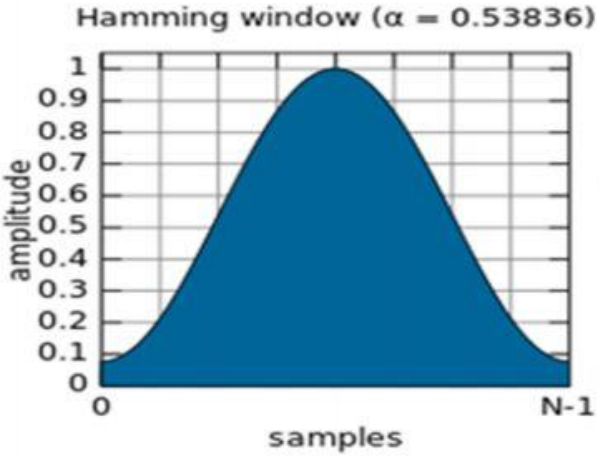

Figure 2: Overlapping Window Showing the Sinusoidal Behavior Including the Samples Obtained and their Respective Amplitudes

II. DISCUSSION

The classification performance of both VQ and VQ-based HMM classifications indicates that as the VQ codebook size grows, so does the recognition. The recognition rate is affected not only by the feature type used, but also by the stress class and voice signal. Compassion is a well-known emotion in the English language. Maximum average success is achieved using the frequency functions for English speech (87.1 percent).

The features of stressed speech were investigated in three areas: pitch, intensity, and glottal spectrum in this research. The statistical research indicates that the stress style can be identified and the speech may be recognized. For the processing, analysis, and acknowledgement of communication, stress types include wrathful, doubtful, mild, and neutral. The findings suggest that pitch analysis is an effective tool for detecting stress. As a pattern recognition method, HMM was utilized to recognize the spoken word.

The perception of speech is heavily influenced by stress. Every word's stressed syllables are typically the best articulated syllables. In the usual voice blur, stressed phrases showed the Sound Islands' reliability. In stressed syllables, vowels are usually longer and louder. They usually keep their vowel value intact. Significantly more. Vowels, on the other hand, tend to be neutral or center

vowel sound in unstressed syllables (decreased syllables at rapid speaking rates). As speech technology improves, it's critical to understand how stress and emotion affect language production in the real world.

The significance of glottal excitation in emotional communication has been evaluated after a comprehensive examination of the differences in excitement through various styles. This study contributes to the most recent knowledge on excitation in order to enhance a robust automated identification of style and stressed speech, a human perception of styled and stressed speech, and natural speech synthesis. Scientists would be able to adjust and reproduce both of these impacts in recognition and perception if they had a better understanding of the changes in the speech waveform when the speaker is stressed.

These findings were quantitatively confirmed via a number of statistical tests. Closing slope, opening slope, closing time, closed time, opening time, and high duration were the six waveform properties that were parametrized. The 11 different types of glottal excitation were subsequently discovered to be very diverse. Finally, with a second speaker, many significant trends outside of normal glottal excitation have been shown to remain consistent, particularly in terms of form features such as a closure route, opening slope, and closed duration.

III. CONCLUSION

A thorough analysis of the sinusoidal model, for example, provides information for distinguishing different stress classes in a speech signal, such as individual frequency, amplifications, and phases, according to the research. Lower frequencies are more present in a speech signal with various stress classes when lower frequencies have higher values in average stressed language frequency indices. Finally, several significant trends away from normal glottal excitation were shown to be constant for the second speaker, particularly in terms of the closure route, opening path, and closed duration form features. These findings show how the sinusoidal model may be used to identify stress classes in a speech stream. The use of the right sinusoidal models may help in stress categorization. The sinusoidal model was not extensively explored for the analysis and detection of speech signals. With this research, it was discovered that the sinusoidal model accurately describes emotions and the identification of emotions in speech signals.

REFERENCES

[1]. Rahurkar MA, Hansen JHL, Meyerhoff J, Saviolakis G, Koenig M. Frequency band analysis for stress detection using a teager energy operator based feature. In: 7th International Conference on Spoken Language Processing, ICSLP 2002. 2002.

[2]. Wang Y. Speech recognition under stress. ProQuest Dissertations and Theses. 2009.

[3]. Keller E. The analysis of voice quality in speech processing. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2005.

[4]. Panda SP, Nayak AK. An efficient model for text-to-speech synthesis in Indian languages. Int J Speech Technol. 2015;

[5]. Waghmare K, Kayte S, Gawali B. Analysis of Pitch and Duration in Speech Synthesis using PSOLA. Commun Appl Electron. 2016;

[6]. Panda SP, Nayak AK. Automatic speech segmentation in syllable centric speech recognition system. Int J Speech Technol. 2016;

[7]. Ghahramani Z. An introduction to hidden Markov models and Bayesian networks. Int J Pattern Recognit Artif Intell. 2001;

[8]. Mohanty MN, Jena B. Analysis of stressed human speech. Int J Comput Vis Robot. 2011;

[9]. Palo HK, Mohanty MN, Chandra M. Design of neural network model for emotional speech recognition. In: Advances in Intelligent Systems and Computing. 2015.

[10]. Palo HK, Mohanty MN, Chandra M. Efficient feature combination techniques for emotional speech classification. Int J Speech Technol. 2016;