Feature Extraction and Processing Analysis: A Review

Paras Nath Mishra, Assistant Professor

Department of Computer Science, Arka Jain University, Jamshedpur, Jharkhand, India

Email Id-paras.m@arkajainuniversity.ac.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, Pitch, Sinusoidal Model, Speech Recognition, Stressed Speech.

1. 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. 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. 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 [1]. 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[2].

Various challenges in the identification of practical speech can be described as follows:

- 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. Stressed speech analysis can give fresh and significant information that helps to better recognize speech, synthesize and verify the speaker automatically[3].

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. This indicates that speaking signal qualities alter under different moods or stressful circumstances[4].

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[5].

1.1. 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. 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. 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[6].

1.2. 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.

1.2.1. 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. This chamber has tools X1 x2, X3, each with a known ball mix, each with an 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. The choice of urn does not depend directly on the urns selected before this preceding urn; this is thus known as a Markov process[7].

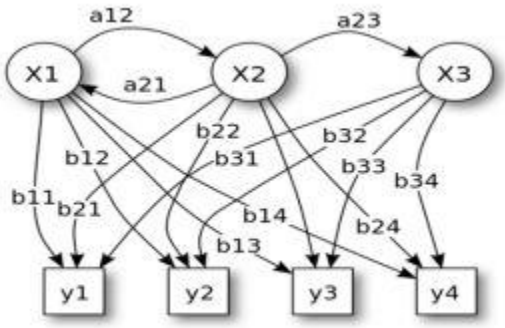

Figure 1: Hidden Markov Process as an Extension of the Substitute in its Discrete Version

1.2.2. 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. 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[8].

1.2.3. 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[9]. 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:

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

1.2.4. Teager Energy Operator(TEO):

The energy of the speech signal was prompted by Teager's speech and hearing tests, which gives a measure of the energy of a speech signal[10]. Teager established in these tests that the airflow in the vocal tract is divided and sticks to the vocal tract walls. Based on these data, vocal tract geometry and results of certain whistle cavity tests, Teager suggested the language productions hypothesis. In this concept air is released as a jet from the glottis and fastened to the closest vocal tract wall. As air travels through the cavity, the vortex of air is formed between genuine vocal folds and fake vocal folds. While the walls of the vocal tract, the bulk of the air continues to spill to the lips. Like the conventional power operator, the TEO is also used to determine signal energy. The TEO for a permanent time signal is:

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

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

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

Where $x(n)$ is the voice signal sampled, φ 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.

1.2.5. 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 voice signal database must be created to evaluate the signal under different conditions. Sound recordings are an electric or mechanical recording of sound waves, such speech, voice, singing, music or sound. Analog recording and digital recording are the two primary types of sound recording technology. A tiny microphone

diaphragm is used to detect changes in the air pressure (such as sound waves) and to graph them as the imagery of sound waves on a media like a phonograph. Acoustic analogical recording (in which a stylus senses grooves on a record). The sound waves vibrate the imitation in the magnetically tape recording and are converted into a different electrical current, converted into a different magnetic field with a magnetic coating, which represents the sound as magnetized areas on a plastic tape. The magnetic wave is then converted into a different magnetic field.

A broad range of speech databases for the development of language synthesis / acknowledgement and linguistic research are accessible. These statistics are generated over a database of 10 men and 10 women between the ages of 20 and 25 who have undergone stress examination. You recorded your speech shortly before the test and one hour after the exam. As is known, to make the analysis accurate, the sentence is taken into consideration that the weather is too hot today, the pattern of speech changes with the substance of the utterances of the speech. The entire database has been monitored and the pattern change throughout regular speech has been analyzed. The generated database is subject to several methods in which the different signal characteristics in various domains are estimated, namely time and frequency domains.

1.3. Sample:

The speech waveforms of each speech style are demonstrated to be substantially and identical to all other styles, therefore, demonstrating the relevance of waveform to the transfer of information about language styles and the causes of changes in speech waveform. When trained on one speaker and compared to another, the degree of variance in speech waveform has shown constant. The SUSAS database uses four stress types (neutral, furious, inquiry and soft) to evaluate the detection performance of an individual word system independent of a speaker.

1.4. Data Collection:

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. However, 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.

This loss of signal amplitude due to the window function can be mitigated by using the notion of window overlap as 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.

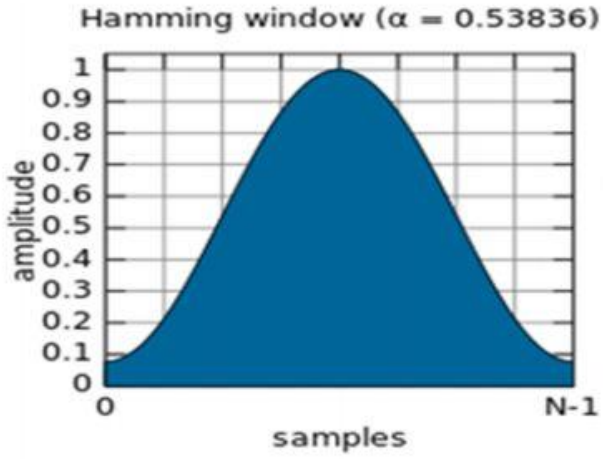

Figure 2: Overlapping Window Showing the Sinusoidal Behavior Including the Samples Obtained and their Respective Amplitudes

2. DISCUSSION

Evaluation of both VQ and VQ-based HMM classifications classification performance shows that the recognition increases with the VQ codebook size being increased. The recognition rate is dependent not just on the type of feature chosen, but on the stress class and speech signal. In English language, compassion emotion is well known. With the frequency functions for the speech in English, maximum average success is attained (87.1 percent).

In this study the characteristics of stressed speech were examined in three areas: pitch, intensity and glottal spectrum. The statistical study shows that the speech can be recognized and the stress style can be detected. Stress styles: wrathful, questionable, gentle and neutral for the processing, analysis and acknowledgment of communication. The results indicate that pitch analysis is a powerful instrument for the detection of stress. HMM was used to acknowledge the spoken word as a pattern recognition technique.

Stress is a crucial component in the perception of speech. Stressed syllables in every word are usually the finest articulated syllable. Stressed syllables thereby demonstrated the dependability of the Sound Islands in the typical voice blur. The vowels in stressed syllables are generally longer and louder. They tend to maintain their full vowel value. More significantly. Vowels, on the other hand, all of them in unstressed syllables (decreased syllables at fast speaking speeds) tend to be neutral or centre vowel sound. So it is vital to understand how stress and emotion impact language output in the actual world as voice technology continues to improve.

This article has taken a thorough look at the variations in excitement via different styles to evaluate the value of the glottal excitement in emotional communication. In order to improve a robust automatic detection of style and stressed speech, a human perception of styled and stressed speech, and natural speech synthesis this work adds to the latest information regarding excitation. A deeper knowledge of the variations in the speech waveform when the speaker is under stress would allow scientists to compensate and replicate both for these effects in recognition and perception.

With a series of statistical tests, these results were quantitatively verified. Six waveform characteristics were parametrized: closing slope, opening slope, closing time, closed time, opening time and high duration. The 11 styles of glottal excitement then proved to be quite varied. Finally, numerous major trends outside from typical glottal excitation have been proven to be constant for a second speaker, especially as regards the form characteristics such as a closure path, opening slope and closed duration.

3. CONCLUSION

The study indicates that a detailed examination of the sinusoidal model, for example, offers information for identifying distinct stress classes in a speech signal, such as individual frequency, amplifications and phases. Higher values of lower frequencies average stressed language frequency indexes show lower frequencies more in a voice signal with different stress classes. Finally, many major trends away from typical glottal excitation were demonstrated to be consistent for the second speaker, notably with regard to the form characteristics of closure path, opening path and closed duration. These results demonstrate the usefulness of the sinusoidal model

for classifying stress classes in a voice stream. Proper selection of sinusoidal models can enhance stress classification performance. For the study and identification of speech signals, the sinusoidal model was not thoroughly examined. With this investigation, it became clear that the sinusoidal model characterizes emotions and the recognition of emotions in a speech signal extremely effectively.

REFERENCES

[1] H. Liang, X. Sun, Y. Sun, and Y. Gao, "Text feature extraction based on deep learning: a review," *Eurasip Journal on Wireless Communications and Networking*. 2017. doi: 10.1186/s13638-017-0993-1.

[2] D. Ververidis and C. Kotropoulos, "Emotional speech recognition: Resources, features, and methods," *Speech Commun.*, 2006, doi: 10.1016/j.specom.2006.04.003.

[3] M. A. Rahurkar, J. H. L. Hansen, J. Meyerhoff, G. Saviolakis, and M. Koenig, "Frequency band analysis for stress detection using a teager energy operator based feature," 2002.

[4] M. N. Mohanty and B. Jena, "Analysis of stressed human speech," *Int. J. Comput. Vis. Robot.*, 2011, doi: 10.1504/IJCVR.2011.042273.

[5] H. K. Palo, M. N. Mohanty, and M. Chandra, "Design of neural network model for emotional speech recognition," 2015. doi: 10.1007/978-81-322-2135-7_32.

[6] K. Waghmare, S. Kayte, and B. Gawali, "Analysis of Pitch and Duration in Speech Synthesis using PSOLA," *Commun. Appl. Electron.*, 2016, doi: 10.5120/cae2016652061.

[7] S. P. Panda and A. K. Nayak, "Automatic speech segmentation in syllable centric speech recognition system," *Int. J. Speech Technol.*, 2016, doi: 10.1007/s10772-015-9320-6.

[8] E. Keller, "The analysis of voice quality in speech processing," 2005. doi: 10.1007/11520153_4.

[9] S. P. Panda and A. K. Nayak, "An efficient model for text-to-speech synthesis in Indian languages," *Int. J. Speech Technol.*, 2015, doi: 10.1007/s10772-015-9271-y.

[10] H. K. Palo, M. N. Mohanty, and M. Chandra, "Efficient feature combination techniques for emotional speech classification," *Int. J. Speech Technol.*, 2016, doi: 10.1007/s10772-016-9333-9.