An Efficient Framework of Human Voice Verification for Robotic Applications

Anuj Kumar Goel*, T Satyanarayana, and Mohammad Jabirullah
Department of Electronics and Communication Engineering, CMR Engineering College, Hyderabad, India
Email: *anuj40b@gmail.com

Abstract. Waves are considered to be used to decode the speech signal more efficiently. This study is an accessible and robust approach for obtaining voice recognition features. Here, we suggested a new text-related method for the identification of human voices (TDHVR) system, which utilizes the discrete wavelet transform (DWT) for low level feature extraction, Relative Spectral Algorithm (RSA) for denoising the voice signal and finally Additive Prognostication (AP) for estimating the formants. First, the proposed methods are used for voice signals, and then we construct a vector train function that includes the derived low level function and estimated formant parameters. The same technique is then applied for calculating speech signals and constructing a test feature vector. The Euclidean distance between the vectors will now be used to balance all vectors in order to distinguish the voice and voice. The simulated human voice would equal the educated person's speech if the difference between two vectors is almost null. Computation results were compared with the LPC Scheme and revealed, that by using fifty preconfigured six voice signals, verification trials were carried out, and a best accuracy of approximately 90 percent was reached, the suggested methodology surpassed the current methodology.

Keywords: Speech Processing, Human Voice Recognition (HVR), Wavelet analysis, Additive Prognostication (AP) and RASTA

1. Introduction

The audio signal is a real component in our everyday lives, especially voice indications, because it can be used as a true tool in our communication. In any case, the voice signal has improved or better regulated by its creative progression through the use of such technological technologies and can be found in many technologies such as security systems [1]. These discussion planning programmes take an imperative role in various uses, for example, the identification of discourses, voice interaction. Disk recognition is a method where eymological information is eventually deleted and measured, transmitted by the speech signal via PCs or electronic circuits. Technologies studied for a long time for the identification of translation and human PC association structures were primarily programmed for discourse acceptance [2]. Programmed discourse acknowledgment (PDA) features a portion of the following favorable circumstances:

- Language information is hard to execute because it does not require a specialized ability as is written or pushbuttons.
- The fact of whether customers go or do stretches on their elbows, feet, eyes and ears can not be data. Details cannot be data
• The inputs should be moderated, remotely suited for expertise on existing telecommunications networks and the Internet, as a recipient or telecommunications should be used to provide awareness [3].

In any event, the PDA errand is problematic because:
  a) There is a lot of redundancy in the speech flag, which creates difficulty in distinguishing the classes.
  b) Existence of transient variance and amplitude variation such as intra-speaker variance in word and phonemic expression and even bury speaker variability such as the influence of local vernaculars.
  c) Subordinate phonemic meaning pronunciation (co-articulation)
  d) Deterioration of the signals so no more material and convolution is visual or channel clamor
  e) Contrivance of signal attributable to the mark of the non-ideal channel [4].

2. Related Work
The following classifications will describe most discourse recognition systems:

2.1. Speaker Dependent Versus Speaker Independent
One that is ready to interpret the voice of the standing speaker is a voice-dependent entry system. These devices are only custom-made and are not financially reasonable for a single person. On the other hand, it is hard to create an autonomy-independent framework because speech recognition systems prefer to get tailored to the speakers that they are equipped to deliver error rates that are greater than subordinate systems for speakers [5].

2.2. Isolated versus Constant
With a single voice, the speaker pauses each word slightly, as the speaker talks in a straight, presumably long sequence, with no breaks in the middle. Lonely speech interfaces acknowledgement are all but impossible to operate, so it is difficult to point out when single words close and another begin and each word is all the more plainly spoken [6-7]. Words which have been spoken in a constant conversation then again are subject to the co-explaining influence, where the words accompanying a word are modified for pronunciation. This makes it troubling to establish a dialogue structure when the same word may be pronounced in several contrasting words. The data from the sensors was gathered on the basis of multi-layer Neural Network technology, then analyzed and passed to the server by means of an alarm, SMS, Email or voice message. [17].

3. Wavelet Analysis
Discrete Transform Wavelet (DWT) is a modified Continuous Transform Wavelet (CWT) rendition. The DWT collects the tremendous amount of CWT knowledge. DWT is like CWT in its main operating concepts, but the wavelet's scales and sites are two. In both components [8-9] the term dysadic refers to the connexion and position [11-12].As in many approved systems, the key characteristics of a sign lie in the low-frequency region. The low frequencies material for voice flags is the section or component of the symbol, which makes the symbol distinctive, while the higher frequencies material is regarded as component of the sign, which gives the sign cleverness. It's like giving the sign a flavor. In a speech signal, the speech appears recognizable as the high frequency material is evacuated, but right today, a speech may be received or transmitted. That is valid that, except for arbitrary disruptions, what is said cannot be heard, the low frequency contents of the sign are removed. The key trait of DWT is that the sign was moved to a high-pass and low-pass channel to get the sign with high heat and low frequency levels. Approximations [10] are the low-frequency substance of the symbol. Which means that approaches are reached by the use of high waves compared with low frequencies.
The hotel elements called delicate elements are obtained by using the low-scale wavelets related to the higher frequencies. Figure 1 explains the Decomposition and Reconstruction of multi-level wavelet.

\[
P(x) = -\sum_{l=1}^{P} a_l x(n - l)
\]

The meaning of the expected and \(x(n-l)\) value is the value expected or projected. By broadening the equation

\[
\hat{x}(n) = -a_1 x(n) - a_2 x(n-1) - a_3 x(n-2) ...
\]

By measuring or forecasting the formants, the LPC will analyse the symbol. The influence of formants is expelled from the talk signal at this stage. The remainder of the buzz is measured in strength and frequency. Thus, by eliminating the formants from the voices sign, we can destroy the effect of resonance. This is the reverse filtering method. Once the formant is removed, the keeping signal is regarded as the aggregation. The coefficients of the LPC must be calculated with a clear intention to calculate the formants. The correlations are assessed with the square error between some of the sign and the first flag. In reducing the error, higher precision of the coefficients is recognised and voice sign formants are gained.

5. Proposed Td-Hvr Model

5.1. RASTA (Relative Spectral Algorithm)

This [13] approach uses a band pass channel for vitality in each frequency subband to cover short-term clamour varieties and expel any continuous offset. This strategy has a clear end purpose. Figure 2 shows the Block diagram of RASTA process.
Stationary disturbances are also differentiated in speech signals. Stationary concussions are clamours that have no diminishing feature [16] that are made available for the entire duration of a certain flag. For some time, the assets will not shift. The suspicion is that the upheaval is increasingly shifting in rhetoric. This allows the RASTA an immaculate instrument for the main steps of the speech sign philtres to evacuate steady clamors[18]. Established stationary clamours are stopping at the frequency range of 1Hz-100Hz.

5.2. Formant Estimation
Formant is one of the most important speech elements. The movements of the entire pinnacles are known as training frequencies or primarily training frequencies [19]. The formant of the sign can be obtained by separating the vocal tract frequency reaction. The frequency response of the vocal tract is seen in Figure 8. The X-stroke is about the frequency level, while the y-stroke is about the size of the symbol. As can be shown, the symbol shapers are F1, F3, F2 and F4. There are three to five formants regularly on a speech symbol. Up to four formants can be differentiated in most speech signals, in any case. The AP strategy is used to achieve the formant of the speech flags. This is taken from the direct term estimation. A kind of scientific activity is a clear expectation, as the word implies. This scientific function is used to determine the potential output on the basis of a direct function of past examples[8] as part of a covert time signal. Figure 3 explains the Formant estimation.

With a specific end goal to actualize the framework, a specific procedure is executed by disintegrating the voice sign to its estimation and point of interest. From the estimation and subtle element coefficients that are extricated, the methodology is revised with the final purpose of the identification protocol in mind.

The observable figure is the methodology recommended for the identification level. The coefficients are determined by four unique kinds of factual calculations. Means, confidence interval, difference and average absolute deviation are observable figures. The wavelet used for the frame is the 7 wavelet and this transform is related to the voice signal. Different experiments and failures overcome this. As the two local contain the bulk of related voice knowledge, the coefficients are the second-level parameters that are segregated from the time domain disintegration process. The information in higher quantities includes absolutely no information which it is unusable for the
identification process. Therefore, the level two coefficients are used for initial framework implementation. The correlations are an additional boundary for the evacuation of the low values and are observable using this. As result of the voice flag analysis, precise association calculation is used along with approximation of the formant and the vitality of the wavelet. All data deleted is used by the speech signals as a "special mark." By comparing the current signal characteristics with the voice signal characteristics, the control rate is determined. Figure 4 illustrates the Proposed TD-HVR block diagram.

The rate of check is given by:  
\[
\text{Verification}\% = \frac{\text{Test value}}{\text{Registered value}} \times 100
\]

6. Simulation Results

RASTA the experiments show for many typical and suggested TD-HVR architectures were seen in this section. In real time, we checked human voice signals, i.e. captured voice was directly taken to the format and translated into format, as MATLAB reads it. The initial and de-noise voice signals for instruction, testing have been seen in Figure 5 (a) and 5 (b). We can see both were similar and therefore the database file has checked the voice. The unknown voice was found in figures 6 (a) and 6 (b) and eventually, the LPC received 66.66% accuracy while approximately 90% of the proposed technique was completed. Figure 7. (a) and 7 (b) illustrates the Initial speech signal and De-noise test signal. Figure 8 (a) and 8 (b) demonstrates the Initial speech and de-noised test signal.

![Proposed TD-HVR block diagram](image-url)
**Figure 6.** (a) Initial speech signal (b) De-noise test signal

**Figure 7.** (a) Initial speech signal (b) De-noise test signal
7. Conclusion

Introduced and analyzed in real time with various test signals a new text-dependent human speech recognition device. A proposed method has also been used to validate an individual's identity by means of the mathematical computing, the formant calculation and the energy of wavelet dependent on its own speech signals. Verification tests were conducted, and the proposed Algorithm obtained an accuracy rate of about 90 per cent while the LPC was 66.66 per cent. Through analysing the simulation outcome on different speaker signals, we can infer that in contrast with LPC the suggested algorithm accuracy was increased.

References

[1] SoontornOraintara, Ying-Jui Chen Et.al. IEEE Transactions on Signal Processing, IFFT, Vol. 50, No. 3, March 2002
[2] Reddy, G., and R. Paviarasi. "Design of fire fighting biped robot with human detection module." Indian Journal of Public Health Research & Development 8, no. 4 (2017): 1200-1202.
[3] Kumar, B. Prasanth, and Y. Bhaskarao. "A Distributed Framework for Surveillance Missions Robots to Detect Intruders." Indian Journal of Public Health Research & Development 8, no. 4 (2017): 1080-1083.
[4] Aravind, K. "Automation of space management in vehicle parking using PLC and SCADA." International Journal of MC Square Scientific Research 9, no. 2 (2017): 135-144.
[5] Amara Graps, An Introduction to Wavelets, Istituto di FisicadelloSpazioInterplanetario, CNR-ARTOV
[6] BraniVidakovic and Peter Mueller, Wavelets For Kids – A Tutorial Introduction, Duke University
[7] O. Farooq and S. Datta, A Novel Wavelet Based Pre Processing For Robust Features In ASR
[8] Giuliano Antonioli, Vincenzo Fabio Rollo, Gabriele Venture, IEEE Transactions on Software Engineering, LPC & Cepstrum coefficients for Mining Time Variant Information from Software Repositories, University Of Sannio, Italy
[9] Michael Unser, Thierry Blu, IEEE Transactions on Signal Processing, Wavelet Theory Demystified, Vol. 51, No. 2, Feb’13
[10] C. Valens, IEEE, A Really Friendly Guide to Wavelets, Vol.86, No. 11, Nov 2012.
[11] James M. Lewis, C. S Burrous, Approximate CWT with An Application To Noise Reduction, Rice University, Houston
[12] Ted Painter, Andreas Spanias, IE EE, Perceptual Coding of Digital Audio, ASU
[13] D P. W. Ellis, PLP, RASTA, MFCC & inversion Matlab, 2005
[14] Ram Singh, Proceedings of the NCC, Spectral Subtraction Speech Enhancement with RASTA
Filtering IIT-B 2012

[15] Nitin Sawhney, Situational Awareness from Environmental Sounds, SIG, MIT Media Lab, June 13, 2013

[16] Rami Al-Hmouz, Khaled and Ali, “Multimodal Biometrics Using Multiple Feature Representations to Speaker Identification System”, International Conference on Information and Communication Technology Research (ICICTR), 2015.

[17] Shahada, S.A.A., Hreiji, S.M., Atudu, S.I. and Shamsudheen, S., 2019. Multilayer Neural Network Based Fall Alert System Using IOT. International Journal of MC Square Scientific Research, 11(4), pp.1-15.

[18] B. Pattanaik and S. Chandrasekaran, "Recovery and reliability prediction in fault tolerant automotive embedded system," 2012 International Conference on Emerging Trends in Electrical Engineering and Energy Management (ICETEEEM), Chennai, 2012, pp. 257-262, doi: 10.1109/ICETEEEM.2012.6494477.

[19] Manjula P., Balachandra P. (2018) An Analysis on Pricing Strategies of Software ‘I-Med’ in Healthcare Industry. In: Mishra D., Nayak M., Joshi A. (eds) Information and Communication Technology for Sustainable Development. Lecture Notes in Networks and Systems, vol 9. Springer, Singapore. https://doi.org/10.1007/978-981-10-3932-4_25