DOI : https://doi.org/10.56452/7-791

ISSN: 0974-5823 Vol. 7 No. 1 January, 2022

International Journal of Mechanical Engineering

Classification of Voice Signals for Direct Speech-To-Image Translation with Enhance Feature Selection

Sumeshwar Singh

Asst. Professor, Department of Comp. Sc. & Info. Tech., Graphic Era Hill University, Dehradun, Uttarakhand India 248002

Abstract

Voice recognition and classification involves the analysis of various speech signals with the goal of enhancing the accuracy of either human recognition or machine decoding. Utilizing characteristics and feature matching, speech recognition algorithms aim to enhance the performance of communication systems. The matching name of the item will thus be detected in this procedure based on the speech characteristics added to the voice signal used as the input question. Also, a number of web-based picture annotation tools have been developed to enhance the quality of image retrieval in response to the rise of social web apps and the semantic web. The primary goals of this procedure are to categorize the voice signal, get the picture from the dataset, and enhance feature classification.

1. INTRODUCTION

Systems engineering, electrical engineering, and applied mathematics all include areas of study known as signal processing that deal with the manipulation or analysis of analogue and digital signals that reflect time-changing or spatially variable physical characteristics. Sound, electromagnetic radiation, pictures, sensor readings, including such electrocardiograms for biological measures, control system signals, telecommunication transmission signals, as well as several other types of signals can all be considered signals of interest.

Speech processing is the field of research that focuses on voice signals and the mechanisms that transform them. As speech signals are frequently processed in a digital form, speech processing may be seen as a special application of digital signal processing. Speech processing includes all steps involved in capturing, modifying, storing, transporting, and producing digital voice signals.

Although NLP applications might get both its output and its input from this system, it is likewise intimately related to NLP. For instance, text-to-speech synthesis may employ a syntactic parser on the text it receives as input, while information extraction techniques may use the output of voice recognition. Speech processing is primarily used to recognise, synthesise, and compress human speech.

When it comes to imaging science, any form of signal processing when an image, such as a photograph or video frame, is the input is referred to as image processing. The outcome of image processing may be another image or a group of variables or traits related to the original image. Most image-processing techniques start by treating the image as a two-dimensional signal and then proceed to traditional processing. Although though digital image processing is increasingly popular, analogue and optical image processing are also good alternatives.

The input and output signals for video processing, which typically employs video filters, are either video files or video streams, is a special form of signal processing in electrical engineering and computer science. Televisions, VCRs, DVD players, video scalers, video players, also many other gadgets need video processing techniques. For instance, most TV sets from various manufacturers merely differ in terms of design and visual processing.

When two or more locations are not physically linked by an electrical conduit, they can nonetheless communicate wirelessly. The most widely used wireless technologies make use of radio and other electromagnetic wireless telecommunications. It includes a range of stationary, transitory, plus portable applications, together with those for two-way radios, mobile phones, personal digital assistants, as well as wireless networking. Light, sound, magnetic, and electric fields may all be used to achieve wireless communications, however these are less often used techniques. A wireless sensor network (WSN) is a collection of geographically dispersed autonomous sensors that track various environmental or physical parameters, such as pressure, sound, and temperature.

Military uses, such battlefield monitoring, served as a driving force behind the creation of wireless sensor networks. Applications include environmental/Earth monitoring, air quality monitoring, monitoring of water quality, detection of forest fires, landslides, machinery health, data logging, industrial sensing & control applications, plus monitoring of water and waste water.

2. LITERATURE SURVEY

In this study, four distinct methods depend upon binarization and thinning are compared against a binarization-based technique. Ten Voices from NIST, an FBI sample, and an opto-electronic gadget make up the sample set. The findings are shown as type traded minutiae, non-existent minutiae, and undetected (dropped) minutiae. Normalization, local orientation estimation, local frequency estimation, along with filtering are the algorithm's primary phases. Three techniques are suggested for improving voice image quality: local histogram equalisation, Wiener filtering, and image binarization. The outcomes are contrasted with those attained using a few additional techniques. The efficiency or time required for each approach shows some progress in the minutiae detection process. Direct grey scale enhancement methods outperform methods that call for binarization and thinning as intermediary stages. Particularly due to inferiority in pictures besides damaged ridges plus blocks with single points, the modified Gabor filter H outperforms the original technique. The suggested improvement plan is also quicker and more effective. The second technique enhances direct grayscale by use of a special anisotropic filter. A ridge structure enhancement method is required to make the ridge structures of Speech pictures more distinct because they are not always clearly defined [1].

A method of user authentication known as "biometric authentication" makes use of physical and behavioural traits such a person's face, voice, hand geometry, iris, keystroke, signature, and speech. Biometric features cannot be lost or forgotten, are challenging to replicate, reproduce, or disseminate, and need the presence of the individual being validated at the moment and place of authentication, making them more trustworthy than password-based verification. Also, it is challenging to fake biometrics, therefore, it is uncommon for a user to prohibit using biometrics to access digital data. By using biometrics in conjunction with passwords (or tokens) and other authentication techniques, the security offered by the authentication system may in certain situations be increased. There are a number of advantages when contrasting biometric systems to traditional authentication techniques. Biometric characteristics are inherently more reliable than password-based verification since they cannot be misplaced or forgotten. The crypto-biometric template securely stores both secret and biometric templates, making the second approach—known as the biometrics-based key generation technique—more secure. Nevertheless, the quality of the audio signal via the phone is often diminished by the communication channel, making voice-based recognition less suitable for phone-based applications [2].

Due to the numerous applications, face recognition has drawn a lot of interest as a difficult topic in image analysis and computer vision. Depending on how face data is acquired, it may be separated into three categories: those that use intensity photographs, people who work with video clips and those who require more sensory information like 3D information or infrared photography. In addition to a discussion of the motivation for utilising face recognition technology, its applications, and some of the challenges facing existing systems with regard to this job, this article gives a summary of some of the popular approaches used in all of these areas, both the advantages as well as disadvantages of the schemes discussed there, also an analysis of the plans referred to therein. The two primary sources of diversity in face appearance are inherent and extrinsic variables. The same person's facial appearance might vary depending on intrapersonal elements like age, facial expression, and facial accessories, while interpersonal ones like gender and race can produce distinct distinctions. Illumination, posture, scale, and imaging parameters are examples of extrinsic variables. Since it has so many uses across so

many different fields, face recognition has drawn a lot of interest within fields of image analysis and computer vision [3].

In order to enhance the performance of DHMM, this work provides a text-dependent speaker verification method employing a mix of VQ plus DHMM. This work has demonstrated that a VQ plus HMM combo technique can enhance the HMM's effectiveness in a silent setting. The system was tested and validated using the Malay spoken digit database, thus a total success rate (TSR) of 99.97% was attained compared to HMM's 89.87%. For FRR, FAR, and EER, the TSRs performance increases dramatically, and more study will be required to assess the system's resilience. The technique of mapping vectors from one vector space to a limited number of its regions is known as vector quantization (VQ). In speaker identification, the speaker codebook is produced when VQ is applied to the collection of feature vectors taken from the audio sample. The most successful design likelihood is chosen. Practically speaking, decision-making is not always straightforward; for instance, in the so-called open-set identification problem, the solution can lie in the fact that the incoming voice signal doesn't match any of the speakers' registered patterns [4].

In order to efficiently extract features from speech signals for a text-dependent speaker identification system, this research suggests a robust feature extraction approach that combines the wavelet transform with the MFCCs. The voice signal is split up into two distinct frequency channels by the wavelet transform, which is then iterated with subsequent approximations being deconstructed. The multi-resolution characteristics of the speech signal are extracted employing the MFCCs of the approximations and detail channels. HMMs and DTW template-based models are contrasted during the identification step. Speech is a complex signal that is created by a number of changes taking place at various levels, thus it's crucial to choose an effective representation for eliminating the informational content of voice signals. The results show that utilising a single reference template has the drawback of being less resistant to the unpredictability of the speech input. The identification of a representation that is best suited for extracting the information content of voice signals is a significant challenge for speech recognition systems [5].

3. PROPOSED SYSTEM

One of the earliest biometric methods, voice identification has been utilised in law enforcement for more than a century. Although normally trustworthy, it can be impacted by finger wetness and dust. A basic image of a Voice cannot deceive optical voice readers, but any 3D voice model creates a serious issue. Some readers come with extra finger liveness detectors. Silicon Voice readers do not work well for those with extremely damp or dry fingertips. As voices are neither compared nor saved as bitmaps, automated voice comparisons have typically relied on correlation- and minutiae-based methods.

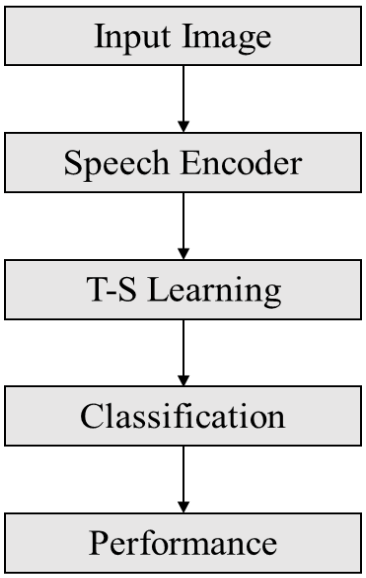

Fig 1: Block Diagram

DOI : https://doi.org/10.56452/7-791

People are identified using Automatic Voice Identification Systems (AFIS), which are managed by seasoned professionals who personally enact change of minutiae extraction with matching. Access control systems are entirely automated, although their precision is a little bit inferior, and the quality of the voice image produced by an automatic voice reader from a novice (non-professional) user is often worse. As opposed to today, when many manufacturers provide solely Voice hardware or software firms provide device-independent Voice processing software, producers of Voice readers used to give the Voice processing software together with the hardware.

The issue with minutiae is that poor-quality voice makes it challenging to effectively extract the detail elements. Moreover, the global pattern of ridges and furrows is not taken into consideration by this technique. Single biometric systems include drawbacks such as uniqueness, a high rate of spoofing, a high rate of mistake, non-universality, and noise. Since that face identification systems are not very accurate, especially in settings with crowded backgrounds and varied lighting conditions, the frequency of misses and false alarms would be significantly higher if faces are employed to identify people rather than voices.

Without using transcription, we try to transform voice inputs into visual signals in this approach. In order to improve generalisation on new classes, a voice encoder is specifically created to represent the input speech signals as an embedding feature. It is then trained utilising teacher-student learning with a pretrained picture encoder. Next, based on the embedding feature, high-quality pictures are created using a stacked generative adversarial network. Experimental findings on both simulated and actual data demonstrate the effectiveness of our suggested strategy for converting raw voice signals into graphics devoid of an intermediate text representation. The following benefits of the suggested strategy are listed:

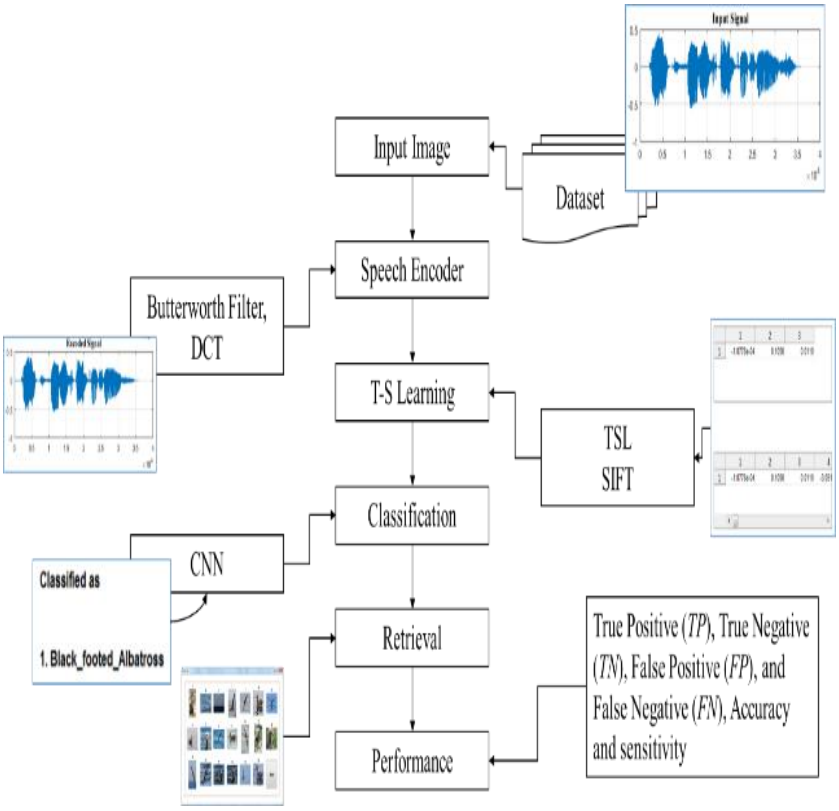

Fig 2: Flow Diagram

- Because this approach recognises both the speaker and the item, it is more effective than the current method.
- The retrieval is quite reliable.

DOI : https://doi.org/10.56452/7-791

4. RESULTS

Voice recognition and classification involves the analysis of various speech signals with the goal of enhancing the accuracy of either human recognition or machine decoding. Based on features and feature matching, speech recognition algorithms try to make communication systems work better. So, in this method, a speech signal was used as the input inquiry, and depending upon this voice characteristics, the appropriate item name would subsequently be identified. To be able to enhance The level of of image retrieval, a range of web-based picture annotation tools have been developed in response to the growth of social web applications and the semantic web.

Experimental findings on both simulated and actual data demonstrate the effectiveness of our suggested strategy for converting raw voice signals into graphics devoid of an intermediate text representation.

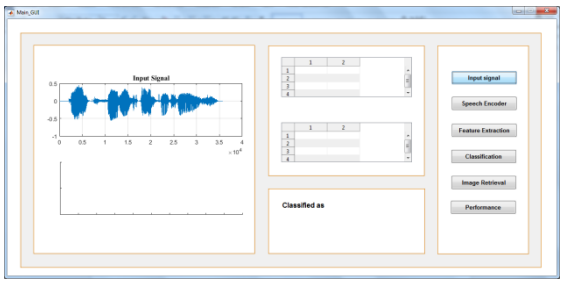

Fig 3: Input Signal

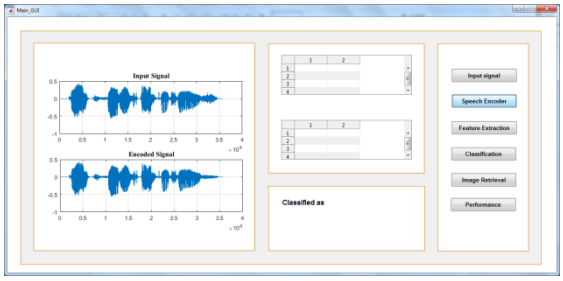

Fig 4: Speech Encoder

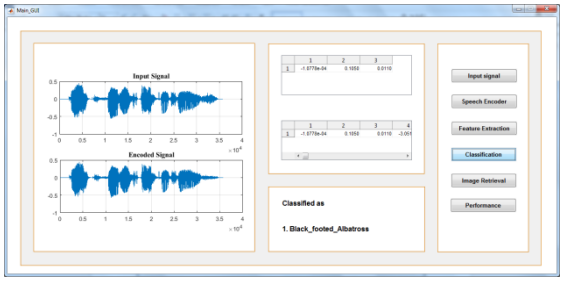

Fig 5: Classification

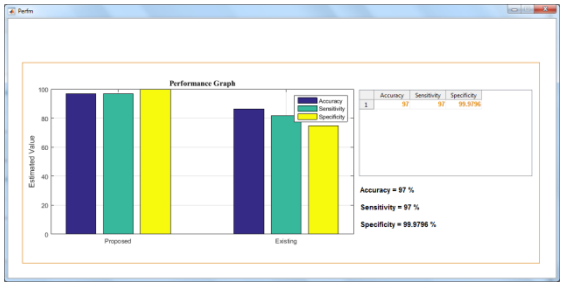

Fig 6: Performance Analysis

5. CONCLUSION

Here, we provide a novel method for converting voice signals into visuals without the need of an intermediate text representation. In order to solve this issue, we extracted a low-dimensional embedding feature from the voice descriptions and used a stacked GAN to create pictures from this feature. For both simulated and actual data, we have shown that our suggested model is capable of creating visuals that are semantically coherent with the input spoken description. On the synthesised datasets, our model outperformed the "twostage" technique, the classifier-based method, and even reached performance comparable to the text-to-image models. We believe that synthesising pictures from voice signals without text is fresh viewpoint to interpret the semantic content in the speech signals and can open up new research paths.

6. FUTURE ENHANCEMENT

With the purpose of enhancing the speech descriptors' qualities. The goal is to increase the signal's feature stability. The identification and retrieval rate will be elevated because of an improvement in feature stability.

REFERENCE

[1] A. Canedo-Rodriguez, S. Kim, J. Kim and Y. Blanco-Fernandez, 'English to Spanish translation of signboard images from mobile phone camera', IEEE Southeastcon 2009, 2009.

[2] H. Nakajima, Y. Matsuo, M. Nagata and K. Saito, 'Portable Translator Capable of Recognizing Characters on Signboard and Menu Captured by Built-in Camera', in Proceedings of the ACL Interactive Poster and Demonstration Sessions, 2005.

[3] L. Zhifang, L. Bin and G. Xiaopeng, 'Test automation on mobile device', Proceedings of the 5th Workshop on Automation of Software Test - AST '10, 2010.

[4] Who.int, 'WHO | Visual impairment and blindness', 2015. [Online]. Available: http://www.who.int/mediacentre/factsheets/fs282/en/. [Accessed: 28- Sep- 2015].

[5] A. Shaik, G. Hossain and M. Yeasin, 'Design, development and performance evaluation of reconfigured mobile Android phone for people who are blind or visually impaired', Proceedings of the 28th ACM International Conference on Design of Communication - SIGDOC '10, 2010.

[6] S. Mori, C. Suen and K. Yamamoto, 'Historical review of OCR research and development',Proceedings of the IEEE, vol. 80, no. 7, pp. 1029- 1058, 1992.

[7] M. Laine and O. Nevalainen, 'A Standalone OCR System for Mobile Cameraphones', 2006 IEEE 17th International Symposium on Personal, Indoor and Mobile Radio Communications, 2006.

[8] P. C. Loizou, Speech Enhancement: Theory and Practice. Boca Raton, FL: CRC, 2007.

[9] D. S. Brungart, P. S. Chang, B. D. Simpson, and D. L. Wang, “Isolating the energetic component of speech-on-speech masking with ideal time–frequency segregation,” J. Acoust. Soc. Amer., vol. 120, no. 6, pp. 4007–4018, 2006.

[10] C. Ludvigsen, C. Elberling, and G. Keidser, “Evaluation of a noise reduction method—Comparison between observed scores and scores predicted from STI,” Scand. Audiol. Supplement., vol. 38, pp. 50–55, 1993.

[11] F. Dubbelboer and T. Houtgast, “The concept of signal-to-noise ratio in the modulation domain and speech intelligibility,” J. Acoust. Soc. Amer., vol. 124, no. 6, pp. 3937–3946, 2008.

[12] Y. Hu and P. C. Loizou, “A comparative intelligibility study of singlemicrophone noise reduction algorithms,” J. Acoust. Soc. Amer., vol. 122, no. 3, pp. 1777–1786, 2007.

[13] C. H. Taal, R. C. Hendriks, R. Heusdens, J. Jensen, and U. Kjems, “An evaluation of objective quality measures for speech intelligibility prediction,” in Proc. Interspeech, 2009, pp. 1947–1950.

[14] J. Ma, Y. Hu, and P. Loizou, “Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions,” J. Acoust. Soc. Amer., vol. 125, no. 5, pp. 3387–3405, 2009.

DOI : https://doi.org/10.56452/7-791

[15] C. Christiansen, M. S. Pedersen, and T. Dau, “Prediction of speech intelligibility based on an auditory preprocessing model,” Speech Commun., vol. 52, pp. 678–692, 2010.

[16] J. B. Boldt and D. P. W. Ellis, “A simple correlation-based model of intelligibility for nonlinear speech enhancement and separation,” in Proc. EUSIPCO, 2009, pp. 1849–1853.