International Journal for Modern Trends in Science and Technology
Volume: 03, Issue No: 05, May 2017
ISSN: 2455-3778
http://www.ijmtst.com

A Survey on Vocalize Fusion and its Taxonomy

P.Arun Kumar[1] | S.P.Shantharajah[2]

[1]Assistant Professor, Department of Master of Computer Applications, Sona College of Technology, Tamilnadu, India.
[2]Professor, School of Information Technology and Engineering, VIT University, Vellore, Tamilnadu, India.

To Cite this Article
P.Arun Kumar and S.P.Shantharajah, "A Survey on Vocalize Fusion and its Taxonomy", *International Journal for Modern Trends in Science and Technology*, Vol. 03, Issue 05, May 2017, pp. 250-256.

ABSTRACT

Along with the rapid development of information technology, the amount of information generated at a given time far exceeds human's ability to organize, search and manipulate without the help of automatic systems. Even, as we have a drastic improvement and innovation in computing technologies, still physically challenged people have a big question that, How we can have the effective process of interaction and innovation in an area of computing. The benefits of this proposed research work are Users can move around their homes or offices freely and still interact with computers without having to sit down and use a mouse or keyboard; independently a person can write and hear the documents through speech. So, this research provides solution for the physically challenged people for their effective Process of using and learning through computers.

Index Terms *– Speech Recognition, ASR (Automatic Speech Recognition), STT, Speech Processing.*

I. INTRODUCTION

Automatic speech recognition or computer speech recognition means understanding voice of the computer and performing any required task or the ability to match a voice or acquired vocabulary. The task is to getting a computer to understand spoken language. By "understand" we mean to react appropriately and convert the input speech into another medium e.g. text. Speech recognition is therefore sometimes referred to as speech-to-text (STT). A speech recognition system consists of a microphone, for the person to speak into; speech recognition software; a computer to take and interpret the speech; a good quality soundcard for input and/or output; a proper and good pronunciation.

The speech is primary mode of communication among human being and also the most natural and efficient form of exchanging information among human in speech. Speech Recognition can be defined as the process of converting speech signal to a sequence of words by means Algorithm implemented as a computer program. Since the 1960s computer scientists have been researching ways and means to make computers able to record interpret and understand human speech. Throughout the decades this has been a frightening task. Even the most rudimentary problem such as digitalizing voice was a huge challenge in the early years. It took until the 1980s before the first systems arrived which could actually decipher speech. Off course these early systems were very limited in scope and power. Communication among the human being is dominated by spoken language, therefore it is natural for people to expect speech interfaces with computer.

Generally, in human computer interaction speech recognition is an important topic in recent research to develop a computer to understand and hear the spoken information [1].In fact, speech has the potential to be a better interface than other computing devices used such as keyboard or

mouse[2].In generally, speech interfacing involves speech synthesis and speech recognition. Speech Recognition is a technology that allows the computer to identify and understand words spoken by a person using a microphone and Speech recognizer converts the spoken word into text [10].It is otherwise called as speech-to-text (STT)due to its input speech into another medium e.g. text conversion [3]. The different speech recognition techniques used are Dynamic Time Warping (DTW) and Artificial Neural Network (ANN) [5]. Among the various models, Hidden Markov Model (HMM) is so far the most widely used technique due to its efficient algorithm for training and recognition [3].

The World Health Organization estimates there are about 314 million vision impaired people in the world, of which about 45 million are blind and various technologies were developed to assist the visually impaired people [9]. In addition, visual speech is of particular importance modality to the hearing impaired person as mouth movement is well known to play an important role in both sign language and simultaneous communication between the deaf [2].Generally, the machine recognition of speech involves generating a sequence of words best matches the given speech signal. Some of known applications include virtual reality, Multimedia searches, auto-attendants, travel Information and reservation, translators, natural language understanding and many more Applications [4].

II. TYPES OF SPEECH

Speech recognition systems can be separated in several different classes by describing what types of utterances they have the ability to recognize. These classes are classified as the following:

A. Isolated Words: Isolated word recognizers usually require each utterance to have quiet (lack of an audio signal) on both sides of the sample window. It accepts single words or single utterance (word) at a time. These systems have "Listen/Not-Listen" states, where they require the speaker to wait between utterances (usually doing processing during the pauses). Isolated Utterance might be a better name for this class.

B. Connected Words: Connected word systems (or more correctly 'connected utterances') are similar to isolated words, but allows separate utterances to be 'run-together' with a minimal pause between them.

C. Continuous Speech: Continuous speech recognizers allow users to speak almost naturally, while the computer determines the content. (Basically, it's computer dictation). Recognizers with continuous speech capabilities are some of the most difficult to create because they need to utilize special methods to determine utterance boundaries.

D. Spontaneous Speech: At a basic level, it can be thought of as speech that is natural sounding and not rehearsed. An ASR system with spontaneous speech ability should be able to handle a variety of natural speech features such as words being run together, "ums" and "ahs", and even slight stutters.

III AUTOMATIC SPEECH RECOGNITION SYSTEM CLASSIFICATIONS

The following tree structure emphasizes the speech processing applications. Automatic Speech Recognition systems can be classified as shown in figure 1.

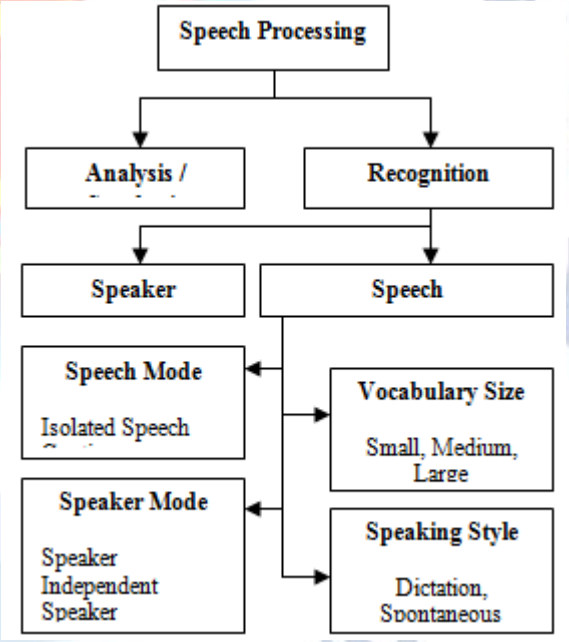

Figure 1: Speech Processing Classification

IV. LITERATURE SURVEY

A. Valentin Ion and Reinhold Haeb-Umbach, [12] have derived an uncertainty decoding rule for automatic speech recognition (ASR), which accounts for both corrupted observations and inter-frame correlation. The conditional independence assumption, prevalent in hidden Markov model-based ASR, was relaxed to obtain a clean speech posterior that was conditioned on the complete observed feature vector sequence. Here, the novel decoding was used to obtain a transmission-error robust remote ASR system,

where the speech capturing unit was connected to the decoder via an error-prone communication network. Also, they showed how the clean speech posterior can computed for communication links being characterized by either bit errors or packet loss. Recognition results were presented for both distributed and network speech recognition, where in the latter case common voice-over-IP codes were employed.

B. To improve the performance of speech recognition in the presence of background noise, Yu Shao and Chip-Hong Chang [13] have designed a framework of wavelet-based techniques; it was realized by implementing speech enhancement preprocessing, feature extraction, and a hybrid speech recognizer in the time–frequency space. In order to capture the most discriminative information in the time–frequency plane they developed a perceptual wavelet filter bank using a fixed base to imitate the human perceptual modus of speech. Then, a Bayesian scheme was applied in a wavelet domain to separate the speech and noise components in the designed iterative speech enhancement algorithm. It minimized the mismatch between the training and testing conditions of the classifier. The de-noised wavelet features were applied to the hybrid classifier founded on a hidden Markov model (HMM). It improved the recognition performance by combining the advantages of different methods into an integral system. The continuous digit speech recognition experiments conducted with the designed framework showed promising results. Also, it significantly improved the recognition performance at a low signal-to-noise ratio (SNR) without causing a poorer performance at a high SNR.

C. Michelle Cutajar *et al* [14] have reviewed about the different methods which are widely used nowadays for the task of ASR. It has three components: feature extraction stage, classification stage and a language model. Various feature extraction methods were proposed, all of which achieved good performance. HMM was one of the widely used methods in classification stage. Although, considerable accuracies were obtained from ASR systems based on HMMs, these were still far from achieving an optimal ASR system by themselves. Hence, numerous hybrid models, based on the concept of merging HMMs with another approach were proposed. Initially, ANNs were being employed with HMMs. Most importantly the researchers showed that SVMs achieved, either comparable or even better results than the HMMs. The last component of an ASR system was the language model. In order to produce meaningful representation of the input speech signal the knowledge of the language being spoken was necessary. Advances in language processing were of fundamental importance for the development of ASR systems, mostly when it comes to large vocabulary speech recognition.

D. Qun Feng Tan[15] have presented a technique which improves upon previously proposed sparse imputation techniques relying on the least absolute shrinkage and selection operator (LASSO). LASSO was widely employed in compressive sensing problems. However, the problem with LASSO was that it does not satisfy oracle properties in the event of a highly collinear dictionary, which happens with features extracted from most speech corpora. When they say that a variable selection procedure satisfies the oracle properties, they mean that it enjoys the same performance as though the underlying true model was known. Through experiments on the Aurora 2.0 noisy spoken digits database, they demonstrate that the Least Angle Regression implementation of the Elastic Net (LARS-EN) algorithm was able to better exploit the properties of a collinear dictionary, and thus was significantly more robust in terms of basis selection when compared to LASSO on the continuous digit recognition task with estimated mask. In addition, they investigate the effects and benefits of a good measure of sparse on speech recognition rates. In particular, they demonstrate that a good measure of sparse greatly improves speech recognition rates, and that the LARS modification of LASSO and LARS-EN can be terminated early to achieve improved recognition results, even though the estimation error was increased.

E. In-Chul Yoo and Dongsuk Yook, [16] have presented a wearable sound recognition system to assist the hearing impaired. Here, they introduced a sound recognition algorithm which was optimized for mechanical sounds such as doorbells. The designed algorithm uses a distance measure called the normalized peak domination ratio (NPDR) that was based on the characteristic spectral peaks of these sounds.

F. For robust speech recognition, Jeih-weih Hung and Hao-Teng Fan,[17] designed a technique

that applied feature statistics normalization. In this designed method, the processed temporal-domain feature sequence was first decomposed into non-uniform sub bands using the discrete wavelet transform (DWT), and then each sub band stream was individually processed by well-known normalization methods, such as mean and variance normalization (MVN) and histogram equalization (HEQ). Finally, they reconstructed the feature stream with all of the modified sub band streams using the inverse DWT. With this process, the components that correspond to more important modulation spectral bands in the feature sequence can processed separately. They applied other types of wavelet functions in the DWT and IDWT processes of their method to investigate if a different analysis/synthesis operation will influence the recognition accuracy.

G. Qun Feng Tan and Shrikanth S. Narayanan, [18] have designed a method for variations of group sparse regularization. They expanded upon the Sparse Group LASSO formulation to incorporate different learning techniques for better sparse enforcement within a group and demonstrate the effectiveness of the algorithms for spectral denoising with applications to robust Automatic Speech Recognition (ASR). In particular, they showed that with a strategic selection of groupings greater robustness to noisy speech recognition can achieved when compared to state-of-the-art techniques like the Fast Iterative Shrinkage Thresholding Algorithm (FISTA) implementation of the Sparse Group LASSO. Moreover, they demonstrate that group sparse regularization techniques offered significant gains over efficient techniques like the Elastic Net. They also showed that the proposed algorithms were effective in exploiting collinear dictionaries to deal with the inherent highly coherent nature of speech spectral segments.

The experimental result on the Aurora 2.0 continuous digit database and the Aurora 3.0 realistic noisy database demonstrated that the performance improvement with the designed methods, including showing that their execution time was comparable to FISTA, which made their algorithms practical for application to a wide range of regularization problems.

H. Zheng-Hua Tanand Borge Lindberg, [19] have presented a low-complexity and effective frame selection approach based on a posteriori signal-to-noise ratio (SNR) weighted energy distance. It makes the method computationally efficient and enables fine granularity search, and the use of a posteriori SNR weighting emphasizes the reliable regions in noisy speech signals. It was experimentally found that the method able to assign a higher frame rate to fast changing events such as consonants, a lower frame rate to steady regions like vowels and no frames to silence, even for very low SNR signals. The resulting variable frame rate analysis method was applied to three speech processing tasks that were essential to natural interaction with intelligent environments. First, it was used for improving speech recognition performance in noisy environments. Second, the method was used for scalable source coding schemes in distributed speech recognition where the target bit rate was met by adjusting the frame rate. Third, it was applied to voice activity detection. Very encouraging results were obtained for all three speech processing tasks.

I. In order to improve the performance of the commercial speech recognizers, Kit Yan Chan *et al*, [20] have presented a multichannel signal enhancement method. The designed method aims to optimize speech recognition accuracy of a commercial speech recognizer in a noisy environment based on a beam former, which was developed by an intelligent particle swarm optimization. It overcomes the limitation of the existing signal enhancement approaches whereby the parameters inside commercial speech recognizers were required to be tuned, which is impossible in a real-world situation.

Also, it overcomes the limitation of the existing optimization algorithm including gradient descent methods, genetic algorithms and classical particle swarm optimization that were unlikely to develop optimal beam formers for maximizing speech recognition accuracy. The performance of the proposed methodology was evaluated by developing beam formers for a commercial speech recognizer, which was implemented on warehouse automation. The experimental result indicated a significant improvement regarding speech recognition accuracy.

V. PROPOSED TECHNIQUE

The problem of speech recognition has received a significant amount of research attention over the past several decades. It has been applied to wide spread areas including helping disabled persons to perform major tasks. We aim to help blind persons

to perform major tasks such as writing exams with the aid of speech recognition technique. The proposed technique consists of three phases, namely pre-processing phase, feature extraction phase and recognition phase. In the pre-processing phase, the input signal will be broken down into time-frequency units.

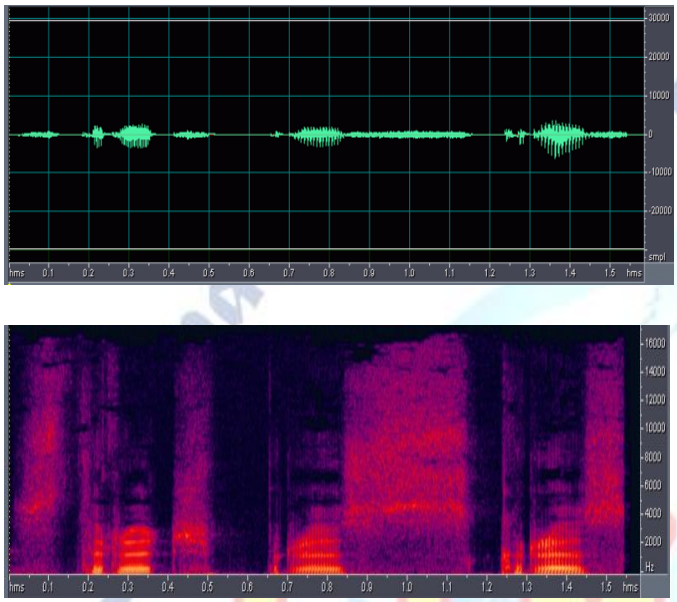

Figure 2: Input signal with their frequency

Subsequently in the feature extraction phase, the features will be extracted using efficient feature extraction technique such as MFCC. The speech signal is also enhanced by the use of masked weights generated with the employment of learning techniques such as Bayesian Networks. In the recognition phase, the signals are matched to the database and recognition is made with the use of K-NN technique. This experimentation is conducted with the various datasets and various evaluation matrices such as recognition accuracy would be employed to validate the proposed technique.

Dataset	Classification
After Removal of Noise	⊗ Noise ○ Absorbed Values

Table 1: Noise and Absorbed Values

VI. APPLICATIONS OF SPEECH RECOGNITION

Various applications of speech recognition domain have been discussed in the following table 1.

SL.NO	PROBLEM DOMAIN	APPLICATION	INPUT PATTERN & PATTERN CLASSES
1	Speech / Telephone/ Communication Sector/Recognition	Telephone directory enquiry without operator assistance	Speech wave form & Spoken words
2	Education Sector	Teaching students of foreign languages to pronounce vocabulary correctly.	Speech wave form & Spoken words
3	Outside education sector	Computer and video games, Gambling, Precision surgery	Speech wave form & Spoken words
4	Domestic sector	Oven, refrigerators, dishwashers and washing machines	Speech wave form & Spoken words
5	Artificial Intelligence sector	Robotics	Speech wave form & Spoken words
6	Medical sector	Health care, Medical Transcriptions	Speech wave form & Spoken words
7	Military sector	High performance fighter aircraft, Helicopters, Battle management, Training air traffic controllers, Telephony and other domains, people with disabilities	Speech wave form & Spoken words
8	General	Automated transcription, Air traffic control, Multimodal interacting, court reporting, Grocery shops	Speech wave form & Spoken words
9	Translation	It is an advanced application which translates from one language to another language.	Speech wave form & Spoken words

Table 2: Applications of Speech Recognition

VII. CONCLUDING DISCUSSIONS

Speech is the primary, and the most convenient means of communication between people. Whether due to technological curiosity to build machines that mimic humans or desire to automate work with machines, research in speech and speaker recognition, as a first step toward natural human-machine communication, has attracted much enthusiasm over the past five decades. We have also encountered a number of practical limitations which hinder a widespread deployment of application and services.

In most speech recognition tasks, human subjects produce one to two orders of magnitude less errors than machines. There is now increasing interest in finding ways to bridge such a performance gap. What we know about human speech processing is very limited. Although these areas of investigations are important the significant advances will come from studies in acoustic phonetics, speech perception, linguistics, and psychoacoustics. Speech recognition is one of the most integrating areas of machine intelligence, since humans do a daily activity of speech recognition.

Speech recognition has attracted scientists as an important discipline and has created a technological impact on society and is expected to flourish further in this area of human machine interaction.

REFERENCES

[1] Mustafa Nazmi Kaynak, Qi Zhi, Adrian David Cheok, Kuntal Sengupta and KO Chi Chung,"Audio-Visual Modeling for Bimodal Speech Recognition", International Conference on system Man and Cybernetics",Vol.1, pp.181 – 186, 2001.

[2] C. Y. Fook, M. Hariharan, Sazali Yaacob and Adom AH,"C. Y. Fook, M. Hariharan, Sazali Yaacob, Adom AH",A Review: Malay Speech Recognition and Audio Visual Speech Recognition", International Conference on Biomedical Engineering, pp.479 – 484, 2012.

[3] Preeti Saini and Parneet Kaur,"Automatic Speech Recognition: A Review", International Journal of Engineering Trends and Technology, Vol.4, No.2, 2013.

[4] Santosh K.Gaikwad, Bharti W.Gawali and Pravin Yannawar, "A Review on Speech Recognition Technique", International Journal of Computer Application, International Journal of Computer Applications, pp.0975 – 8887, Vol.10, No.3, 2010.

[5] Shing-Tai Pan, Ching-Fa Chen and Yi-Heng Tsa, "Genetic Algorithm on Speech Recognition by Using DHMM", 7th IEEE Conference on Industrial Electronics and Applications, pp.1333-1338, 2012.

[6] Catherine J Nereveettil, M.Kalamani and S.Valarmathy, "Feature Selection Algorithm for Automatic Speech Recognition Based On Fuzzy Logic" International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, Vol. 3, No. 1, 2014.

[7] M.A.Anusuya and S.K.Katti, "Speech Recognition by Machine: A Review",(IJCSIS) International Journal of Computer Science and Information Security, Vol. 6, No. 3, 2009

[8] Shing-Tai PanandXu-YuLi, "An FPGA-Based Embedded Robust Speech Recognition System Designed by Combining Empirical Mode Decomposition and a Genetic Algorithm", IEEE Transactions On Instrumentation and Measurement, Vol. 61, No. 9, 2012.

[9] Bruce Moulton, Gauri Pradhan and Zenon Chaczko, "Voice Operated Guidance Systems for Vision Impaired People: Investigating a User-Centered Open Source Model", International Journal of Digital Content Technology and its Applications, Vol.3, No. 4, 2009.

[10] S.Ganesh,Saravana Kumar, Shankar, Samuel D.Raj and R.Kartik, "A Novel Voice Recognition System for Dumb People" Journal of Theoretical and Applied Information Technology, Vol. 53 No.1, 2013.

[11] R.Vijayasarathi and C. Manoharan, "Speech Based Search Engine System Control and User Interaction", Journal of Computer Engineering (IOSRJCE), Vol.4, No.4, pp. 24-30, 2012.

[12] Valentin Ion and Reinhold Haeb-Umbach,"A Novel Uncertainty Decoding RuleWith Applications to Transmission Error Robust Speech Recognition", IEEE Transactions on Audio, Speech, and Language Processing, Vol. 16, No. 5, 2008.

[13] Yu Shao and Chip-Hong Chang, "Bayesian Separation with Sparsity Promotion in Perceptual Wavelet Domain for Speech Enhancement and Hybrid Speech Recognition", IEEE Transactions on Systems, Man, And Cybernetics—Part A: Systems And Humans, Vol. 41, No. 2, 2011.

[14] Michelle Cutajar, Edward Gatt, Ivan Grech, Owen Casha and Joseph Micallef, "Comparative study of automatic speech recognition techniques", IET Signal Process, Vol. 7, No. 1, pp. 25–46., 2013.

[15] Qun Feng Tan, Panayiotis G. Georgiouand Shrikanth Narayanan, "Enhanced Sparse Imputation Techniques for a Robust Speech Recognition Front-End", IEEE Transactions on Audio, Speech, and Language Processing, Vol. 19, No. 8, 2011.

[16] In-Chul Yoo and Dongsuk Yook, "Automatic Sound Recognition for the Hearing Impaired", IEEE Transactions on Consumer Electronics, Vol.54, No.4,pp.2029-2036, 2008.

[17] Jeih-weih Hungand Hao-Teng Fan, "Sub band Feature Statistics Normalization Techniques Based on a Discrete Wavelet Transform for Robust Speech Recognition", IEEE Signal Processing Letters, Vol. 16, No. 9, 2009.

[18] Qun Feng Tan and Shrikanth S. Narayanan,"Novel Variations of Group Sparse Regularization Techniques with Applications to Noise Robust Automatic Speech Recognition", IEEE Transactions on Audio, Speech, and Language Processing, Vol. 20, No. 4, 2012.

[19] Zheng-Hua Tanand Borge Lindberg, "Low-Complexity Variable Frame Rate Analysis for Speech Recognition and Voice Activity Detection", IEEE Journal of Selected Topics in Signal Processing, Vol. 4, No. 5, 2010.

[20] Kit Yan Chan, Cedric K. F. Yiu, Tharam S. Dillon, Sven Nordholm, and SaiHoLing," Enhancement of Speech Recognitions for Control Automation Using an Intelligent Particle Swarm Optimization", IEEE Transactions on Industrial Informatics, Vol. 8, No. 4, 2012.