International Journal of Scientific Research in Science, Engineering and Technology
Print ISSN: 2395-1990 | Online ISSN : 2394-4099 (www.ijsrset.com)
doi :https://doi.org/10.32628/IJSRSET

A Study on Automatic Speech Recognition

Viren Nivangune, Argade Siddhi Anil, Mrudula Milind Patankar
Department of Computer Engineering, Zeal College of Engineering and Research, Pune, Maharashtra, India

Article Info
Volume 9, Issue 1
Page Number : 318-326

Publication Issue :
January-February-2022

Article History
Accepted : 05 Feb 2022
Published: 20 Feb 2022

ABSTRACT

Speech is an easy and usable technique of communication between humans, but nowadays humans are not limited to connecting to each other but even to the different machines in our lives. The most important is the computer. So, thiscommunication technique can be used between computers and humans. This interaction is done through interfaces, this area called Human Computer Interaction (HCI). This paper gives an overview of the main definitions of Automatic Speech Recognition (ASR) which is an important domain of artificial intelligence and which should be taken into account duringany related research (Type of speech, vocabulary size... etc.). It also gives a summary of important research relevant to speechprocessing in the few last years, with a general idea of our proposal that could be considered as a contribution in this areaof research and by giving a conclusion referring to certain enhancements that could be in the future works.

I. INTRODUCTION

Humans communicate with each other in many ways such as speech, hand gestures, facial expressions… etc. But speech is considered the most important means that a human uses, as it facilitates communication and it is the most widely used between speakers.

Speech is a useful expression and has a particular meaning and it is composed of several words, which in turn contain several letters accompanied by voices. This voice can spread in objects of air and empty and inanimate in the form of waves; a wave that overlaps between them or begins in the form of small circles of the sound source. This situation is characterized by force a then widens these circles little by little until they disappear completely when they spread over long distances. Logical speech is done in speakers who talk the same language, meaning that the sender and recipient have the same keys that help them decipher the meaning.

The researchers applied this phenomenon and developed it to become a key branch in human communication with the machine where the sound has helped to facilitate the use of the machine with the user and to make a natural communication between them. Automatic speech recognition has greatly contributed to the development of artificial intelligence, which seeks to create very flexible methods of handling the machine, this allows the user to communicate and exchange information without using known input/output modules such as the keyboard. Voice-based input/output techniques are very useful in several areas, such as the care of

disabled people, the use of cars, in particular when driving, distress calls, etc.

In this paper, we present a review of the latest works that focused on the ASR mechanism where we show their most important characteristics, advantages, and disadvantages, as we compare these works to explain our vision as another possible approach.

In Section 2 and Section 3, an overview of the automatic speech recognition system and the characteristics for the speech recognition system, we present the definition of automatic speech recognition, and in Section 4, description about the architecture of the automatic speech recognition systems, with an explanation of each part's work. In Sections 5 and 6, we present some recent work on automatic speech processing, discuss and present our vision of the work presented in the previous section. We explain to what extent hybrid models are used and neural network models.

Finally, we finish the article by a conclusion and display the list of references.

II. AUTOMATIC SPEECH RECOGNITION

Automatic speech recognition is one of the most automatic speech processing areas, allowing the machine to understand the user's speech and convert it into a series of words through a computer program, thus creating a kind of natural communication between man and machine. [12] [13] [14]

Automatic speech recognition is also called speech recognition; it can be defined as graphical representations of frequencies emitted as a function of time. All speech processing techniques (speech synthesis and processing, speaker identification, Speaker verification make it possible to create voice interfaces (Human Machine Interface) or perform voice interaction. Voice recognition can be applied in several applications such as:

- Voice services: speaking clock, weather, race results, etc.,
- Quality control, data entry.
- Avionics, Training
- Disabled assistance, Vocal dictation

We can also mention embedded voice recognition modules, such as in mobile phones or in cars: car radio, air conditioning, on- board navigation on the Internet with voice commands and others.

III. CHARACTERISTICS OF SPEECH RECOGNITION SYSTEM

There are many variables in the systems of speech recognition and it is necessary to know these variables so that we can determine the algorithm appropriate to the system and the most important of these variables:

Types of Speech in most studies, speech is resumed into four types:

- **Isolated Words:** This type usually requires a quiet (silence state) between utterances.
- **Connected Words:** Word systems are similar to isolated words, the only difference between themes to allow separate words to "run-together" with a minimum of pausing between them.
- **Continuous Speech:** The users of this type talk almost normally, while the computer selects the content. It is one of the most difficult systems.
- **Spontaneous Speech:** At a basic level, it can be thought of as speech that is natural-sounding and not rehearsed. [15]

Size of the Vocabulary the volume of vocabulary used in the speech recognition system is important because it affects the complexity and processing requirements and determines the accuracy of the system. We note that there are applications that use only a few words, while others require the use of a very large number. There are no specific definitions, but we can define them as follows: [17]

- Small vocabulary - tens of words,
- Medium vocabulary - hundreds of words,

- Large vocabulary - thousands of words,
- Very-large vocabulary - tens of thousands of words.

Speaker Dependence

- Speaker dependent system: Systems that require the user to train the system according to the user's voice.
- Speaker independent system: Systems developed for any speaker.
- Speaker adaptable system: Developed to adapt to the characteristics of new speakers.

IV. THE ARCHITECTURE OF AUTOMATIC SPEECH RECOGNITION SYSTEMS

The basic goal of a speech recognition system is that the device is capable of listening and understanding of spoken or acoustic information to make the right decision, but how can this be possible?

The first phase of the system is the analysis of speech signal to be the last result, a series of spoken words. Between these two phases, the system shows several stages that are often based on the statistical approach. Generally, the speech recognition system consists of five units (see Figure. 1) which are:

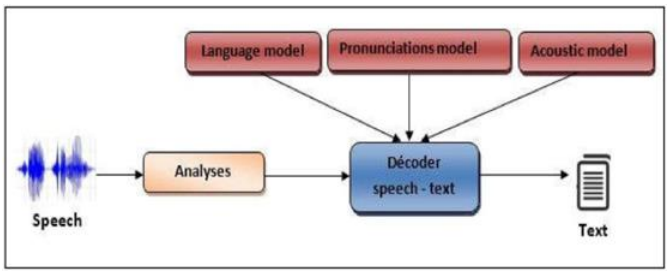

Figure 1. General Architecture of Automatic Speech Recognition Systems

Analysis Speech analysis is the beginning of speech processing, it allows to select the frame size to segment the input signal is intended to make another analysis on speech segment; speech analysis can be performed in three methods:

- **Segmentation Analyses:** The purpose of this step is to extract the speaker's information using frame usage of a size of 10 to 30 ms. [25]
- **Sub Segmental Analysis:** This technique is used to analyze and extract the characteristics of the excitation state [24], and for that use a frame of size 3 to 5 ms.
- **Supra Segmental Analysis:** In this step analyze and characteristic of the speaker's behavioral character. [24]

Language Model The language model is divided into twogroups:

- **Deterministic (or grammatical):** Is designed by language experts.
- **Stochastic (or Statistical):** Statistical language models are the result of an unsupervised language model estimate on a learning corpus. Most cases start with a set of empty parameters that are estimated during the observation of linguistic data. [26]
- **Pronunciations Model:** Is building a language model: How to write a word.
- **Acoustic Model:** This model makes it possible to predict the most likely phonemes in the audio that has been input. [27]
- **Decoder Speech – text:** This is a combination of previous models to provide the most likely text transcription for a given speech statement. [27]

V. RELATED WORK

Deep neural networks are the latest methods that have contributed significantly to the development of speech recognition. Zied Elloumi & al [30], is proposed a multitasking system for performance prediction. This system is based on the convolutional neural network. This came after a comparison between this approach, which is based on learned features, and an approach based on predefined traits (engineered features).

The data that are used in this experiment was a collection of French-language programs: a subset of the Quaero1 corpus, the data from the ETAPE project [28], the data from the ESTER 1 & ESTER 2 [31]

evaluation campaigns, the data from the REPERE evaluation campaign [29]. The results obtained in this experiment, the prediction by CNN is better than the comparative approach in terms of MAE (Mean Absolute Error) and Kendall scores. also the joint inputs of texts and signals give positive results and better performance, also the CNN (convolutional neural networks) correctly predict the distribution of word error rates on a collection of records.

Laszlo Toth [1] propose a simple approach, which combines two approaches are the deep neural networks standard ReLU and the linear augmentation approach of Ghahremani et al [20], he proposed to skip the calculation of the activation function on the subsets of neurons in each network layer which therefore act as linear units and the database was used in this experiment is TIMIT. The positive thing about this experiment is that this change is simple and decreases the cost of computing and in all experimental setups the linearly increased ReLU network outperformed the standard ReLU network, it is effective or slightly better than a maxout network when it is driven on a larger network, even it outperformed the maxout network for a larger lot size. But the downside of this experiment is it could not beat the maxout network performance on the TIMIT2 database [1], and achieve the same error rate for smaller batch size.

Yuki Saito & al [2] proposed an approach to solving the discourse quality problem, this approach allows to create a training algorithm for high-quality parametric vocal synthesis based on the deep neural network (DNN) and using the basic ATR3 data which consists of two neural networks:

- Discriminator (discriminator) to distinguish natural samples and generates and can be interpreted as anti-spoofing
- Generator to deceive the discriminator.

In this algorithm, the acoustic models are the forms make the distribution of the parameters of the generic words close to that of the natural speech. The proposed method was carried out both in the DNN-based TTS (Text-to-Speech) and VC (voice conversion) systems and at the same time applied to statistical parametric approaches and also to glottal wave synthesis.

Also, the algorithm to offset the overall Gregory Gelly & Jean-Luc Gauvain [8], proposed an optimization method based on a neuron network; the process optimizes all SAD system parameters:

- Characteristics extraction parameters
- The parameters of the weights NN (NN weights)
- Variance and correlation between speech parameters generated.

The result demonstrates that the algorithm produces significant improvements in speech quality in both TTS and VC. Incorporating Wessertien's GAN4 improved synthetic speech quality over various GANs. They also compared three types of RNNs: RNN5 basic, LSTM6, and CG-LSTM7 that were proposed in [9]. In addition, they compared three types of methods: The standard MLP, method based on functionality (feature-based method) and the LongTerm Signal Variability (LTSV) method use long-term signal variability. This process based on a QPSO8 algorithm that allows us to treat a non-differential part of the SAD model and also to adopt the difficult optimization tasks and more effective than the genetic algorithms. This experiment was applied to four different corpora: the REPERE9 evaluation data, the AMI10 meeting corpus, the NIST11 OpenSAD'1 evaluation data, and the NIST OpenKWS'13 multilingual conversation corpus. The obtained result is the superiority of the optimization method proposed for gradient descent training as well as the CG-LSTM network surpasses the original network LSTM and a basic RNN on MLP and two other basic SAD systems.

Michael Price & al [5], seek to decode a string of audio to text to achieve good energy efficiency and scalability. For this purpose, it represents an ASR12 approach to implement VDA (Voice Activity Detection) digital ICs13 such that this

implementation accepts audio samples from a digital microphone, label regions of the waveform as speech, non-speech. And as an output, the ASR do text transcription with models stored in external memory. With this approach, the use of a VAD powered portal ASR, where the latter has performed a variety of tasks in real-time with vocabularies ranging from 11 words to 145,000 words, and designed interfaces allowing subsystems to operate together on a single chip and full-chip power consumption ranging from 172 μ W to 7.78 mW.

In 2017 Lilia LAZLI & al [3], make the hybrid type of stochastic/connectionist model with the use of FCM14 / GA15 clustering for French and Arabic. They proposed a hybrid learning algorithm stochastic model/connectionist based on FCM clustering with the optimization of this result using GA, FCM / GA clustering with hybrid HMM / ANN modeling has improved ASR accuracy such that the recognition rate for BDSON [3] Corpuses is 100 and for AD [3] Corpora is exceeded 80, which shows that the hybrid model is more efficient than the HMM model and hybrid system using the hybrid Fuzzy Genetic approach providing superior results to the use of the K-means algorithm for classification.

2 TIMIT: Texas Instruments/Massachusetts Institute of Technology.

3 ATR: Advanced Telecommunications Research Institute International.

4 GAN: Generative Adversarial Network.

5 RNN: Recurrent Neural Network.

6 LSTM: Long Short-Term Memory.

7 CG-LSTM: Long Short-Term Memory with Coordinated Gates.

8 QPSO: Quantum-Behaved Particle Swarm Optimization.

9 REPERE: Campagne d'évaluation REconnaissance de PERsonnes dans des Emissions audiovisuelles

10 AMI: Augmented Multi-party Interaction.

11 NIST: National Institute of Standards and Technology.

12 ASR: Automatic Speech Recognition.

13 IC: Integrated Circuit.

14 FCA: fuzzy c-means.

15 GA: Genetic Algorithms

In the same year, Abhijit Mohanta and Vinay Kumar Mittaln [4], classified the four emotional states of low (happy, angry, frightened and nature) with the use of the parts of the vowels presented on the speech signal, they proposed an approach that allows to analysis changes in the speech production functionalities and in particular at the regions of the vocal signal vowel. The "wave surfer" tool for SVM detection and classifier was used for the binary classification and the multi-class ranking, with the calculation of the characteristic values F0, Formants. The emotional state of anger for male and female speakers, the vowel / a / e / gives the highest F0 value compared to other states. In the classification, the happy emotional state gives the highest classification accuracy of 76% and the emotional state of fear gives the lowest accuracy rate of 60%.

Edwin Simonnet & al [7], studied the problem of automatic speech recognition (ASR), and how to detect errors and use them to improve language understanding systems (SLUs). They proposed an approach aims to enrich the set of semantic tags with specific error tags. For this, they used an approach two SLU architectures respectively based on random field conditions (CRF "concept error rate") and a structured NN-EDA 16NN-structured neural coded-decoder network, such that each ASR error detection subsystem has been provided with dependency functions based on syntactic dependencies and relevance words on the semantic plane. The best combination of these architectures provides improvements with a relative reduction of the conceptual error rate (CER) of 18.9% and a relative error concept value (CVER) relative reduction of 10.3% compared to a written reference in [6].

Dominique F & al [10] introduced the neural networks as the most widely used product in the field

of artificial intelligence and they presented and studied acoustic and linguistic models of automatic speech recognition system relying on the corpus "radio broadcast news". These models are deep neural networks (DNN) and Hidden Markov Model (HMM) models where they compared with conventional GMM/HMM, and the result of this comparison is a significant relative improvement estimated at 24%.

Luiza Orosanu's [11] approach is an integral part of the RAPSODIE (Automatic Speech Recognition for the Deaf or Disabled) project, such as this approach, which involves the generation of automatic speech transcriptions, with a display adapted to deaf people. They proposed the use of the so-called hybrid model combining words and syllables in a single language model, the inclusion of syllables in the recognition model allows to approximate the pronunciations of words out of vocabulary.

During decoding so they have to make hybrid language models based on the transcription of speech pronunciations (at the word level, and at the phoneme level through a forced alignment). The hybrid language models considered for an embedded version of the recognition system have sizes ranging from 1K words and 6K syllables up to 31K words and 2K syllables with theuse of the ESTER2[21], ETAPE[23] and, EPAC[22] corpus. This work highlighted the interest of the model syllabics which, for a small memory footprint (8,100 syllabic units) and a low calculation cost, gave good phonetic decoding performance, with a phonetic error rate (2%). but the problem posed by this approach concerns the means of passing from text to phonetic representation of pronunciations and of combining data transcribed manually with text data, another point the evaluation of corpora that can evaluate the real performance of the system. and improve it.

VI. DISCUSSION

Previous studies have tried to focus on the major gaps in the Automatic Speech Recognition System and identified solutions that reduce those gaps and the percentage of errors. This research examined all the phases of the automatic speech recognition systems (Acoustic pre-processing, pronunciation model, acoustic model, language model) but also agreed on a common objective: use of neural networks; with the exception of the work of Lilia LAZLI et al [3] and Dominique .F et al [10] when they used the HMM/DNN hybrid model. All these studies have validated the use of neural networks to get better results both with other techniques or without them and this is what is observed in all the results of this work, also the use of this type of networks brings several advantages to the field of automatic speech processing. The performance of Automatic Speech Recognition

16 NN-EDA : Encoder-Decoder Neural Network structure with a mechanism of Attention

Systems are determined in both accuracy and rapidity, where most searchers have used the word error rate (WER) to measure the precision in a system and speed has been measured with the real-time factor. Other precision measures include: Command Success Rate (CSR), Frame Error Rate (FER), NIST Detection Cost Function (DCF), Recognition rate, concept value error rate (CVER), concept error rate (CER), Phone Recognition Error Rates (PHER), and Frame Classification (FRER). The graph below shows the use of neural networks in automatic speech processing during the work we presented in the previous section 2017/ 2019.

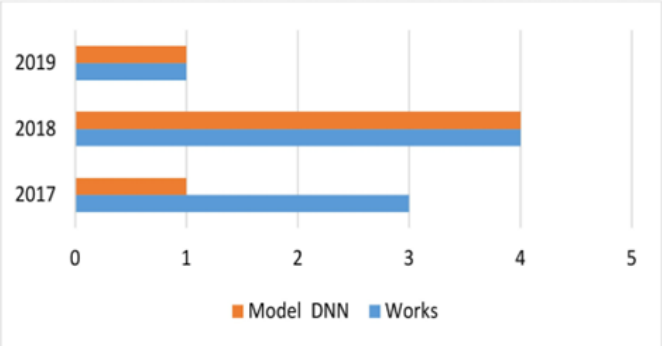

Figure 2. Use the model of DNN in Automatic Speech Processing

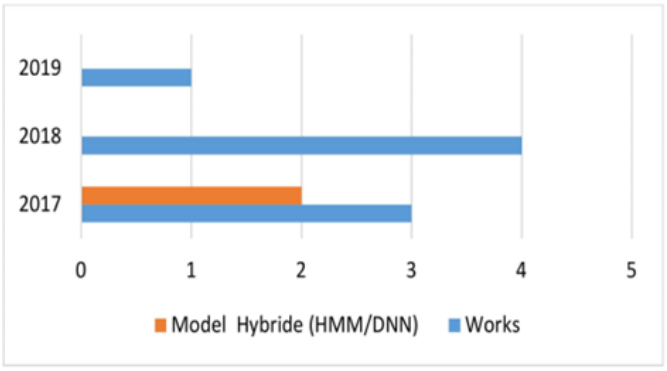

Figure 3. Use the model of Hybrid (HMM/DNN) in Automatic Speech Processing

Through analysis, we note that the most commonly used approach is neural networks because of its large capacity to process the vast amount of information, which made it one of the most important models used in artificial intelligence. The recent development of artificial intelligence is what made the use of neural networks take up a lot of use. Neural networks consist of three basic layers: the input layer, the hidden layer, and the output layer. Each layer plays an important role in the processing of information, and the difference in the work of the hidden layer is what led to the emergence of many types of neural networks. Our research aims to use sound as a means of use in a smart interface so that the latter does not depend only on speech, but on many sensations adopted in humans. We thought of using neural networks because of their great usefulness in processing information in a good way. The following sample shows an idea for our future work.

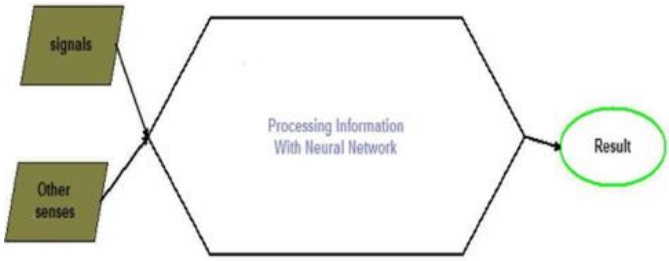

Figure 4. Model for smart interface

VII. CONCLUSION

In this paper, we have presented a review of Automatic Speech Recognition Systems. Where we have explained in the first sections the system and its major features, also the general architecture of the Automatic Speech Recognition System. The second section has discussed the results of some of the recent research works after the presentation of most of the encountered problems and the different solutions proposed.

The most used solution is the neural networks according to the work we have presented in this article. Through what we have presented in this article, several research perspectives can be considered. First of all, we want to use neural networks in our approach that is concerned with automatic speech processing to create an intelligent interface based on computer vision and works to receive the voice of users so that these interfaces allow intelligent interaction with users and establish a natural and easy communication between the machine and human. In addition, we want to use speech with other human senses to improve our interface. Finally, we want to create an interface that allows us to make a decision in real-time.

VIII. REFERENCES

[1]. Laszlo, T. (2018). Deep Neural Networks with Linearly Augmented Rectifier Layers for Speech Recognition, SAMI 2018 IEEE 16th World Symposium on Applied Machine Intelligence and Informatics February 7-10 Košice, Herl'any, Slovakia.

[2]. Yuki, S., Shinnosuke, T. (2018). Statistical Parametric Speech Synthesis Incorporating Generative Adversarial Networks, IEEE/ACM Transactions on Audio, Speech, and Language Processing, 26 (1).

[3]. Lilia, L., Mohamed, T. L., Rachid, B. (2017). Discriminant Learning for Hybrid HMM/MLP

Speech Recognition System using a Fuzzy Genetic Clustering, Intelligent Systems Conference 20177-8 | London, UK.

[4]. Abhijit, M., Vinay, K. M. (2017). Human Emotional States Classification Based upon Changes in Speech Production Features in Vowel Regions, 2017 2nd International Conference on Telecommunication and Networks (TEL-NET 2017).

[5]. Michael, P., James, G., Anantha, P. C. (2018). ALow-Power Speech Recognizer and Voice Activity Detector Using Deep Neural Networks, IEEE Journal of Solid-state Circuits, 53(1).

[6]. Stefan, H., Marco, D., Christian, R., Fabrice, L., Patrick, L., Renato, D., Alessandro, M., Hermann, N., Giuseppe, R. (2011). Comparing stochastic approaches to spoken language understanding in multiple languages, IEEE Transactions on Audio, Speech, and Language Processing, 19 (6) 1569–1583.

[7]. Edwin, S., Sahar, G., Nathalie, C., Yannick, E., Renato, D. (2017). ASR error management for improving spoken language understanding, arXiv: 1705.09515v1[cs.CL].

[8]. Gregory, G., Jean-Luc, G. (2018). Optimization of RNN-Based Speech Activity Detection, IEEE/ACM Transactions on Audio, Speech, and Language Processing, 26 (3).

[9]. Gregory, G., Jean-Luc, G. (2015). Minimum Word Error Training of RNN-based Voice Activity Detection, INTERSPEECH 2015,16th Annual Conference of the International Speech Communication Association, Dresden, Germany.

[10]. Dominique, F., Odile, M., Irina, I. (2017). New Paradigm in Speech Recognition: Deep Neural Networks, the ContNomina project supported, French National Research Agency (ANR).

[11]. Luiza, O. (2015). Reconnaissance de la parole pour l'aide à la communication pour les sourds et malentendants, Université de Lorraine, Laboratoire Lorrain de Recherche en Informatique et ses Applications - UMR 7503 .

[12]. https://www.voicebox.com/wp-content/uploads/2017/05/Automatic-Speech-Recognition-Overview-and-Core-Technology.pdf, © 2017 Voicebox Technologies Corporation, voicebox.com.

[13]. Xuedong, H., Li, D. (2009). An Overview of Modern Speech Recognition, Indurkhya/Handbook of Natural Language Processing C5921_C01, 339 -344, Microsoft Corporation.

[14]. Julien, A. (2003). Approche Dela Reconnaissance Automatiquede La Parole, Rapport cycle probatoire, CNAM.

[15]. Anusuya, M. A., Katti, S. K. (2009). Speech Recognition by Machine: AReview, (IJCSIS) International Journal of Computer Science and Information Security, 6(3).

[16]. Preeti, S., Parneet, K. (2013). Automatic Speech Recognition: A Review, International Journal of Engineering Trends and Technology, 4(2) 2013,http://www.internationaljournalssrg.org

[17]. Santosh, K. G., Bharti, W. G., Pravin, Y. (2010). A Review on Speech Recognition Technique, International Journal of Computer Applications (0975 – 8887) 10(3), (November).

[18]. Vrinda1, Shekhar, Chander. (2013). Speech Recognition System For English Language, International Journalof Advanced Research in Computer and Communication Engineering, 2(1), January 2013, ISSN (Print): 2319-5940 ISSN (Online): 2278- 1021, www.ijarcce.com, Copyright to IJARCCE.

[19]. http://www.speech.cs.cmu.edu/comp.speech/Section6/Q6.1.html, date of consultation: 29/08/2018.

[20]. Pegah, G., Jash, D., Michael, L. S. (2016). Linearly augmented deep neural network, in Proc. ICASSP, 5085–5089.

[21]. Sylvain,G., Guillaume, G., Laura, C. (2009). The ESTER2 Evaluation Campaign for the Rich Transcription of French Radio Broadcasts, Brighton UK, Proceedings of Interspeech.

[22]. Yannick, E.,Thierry, B., Jean-Yves, A., Frédéric, B. (2010). The EPAC corpus: manual and automatic annotations of conversational speech in French broadcast news, In: Proceedings of the International Conference on Language Resources and Evaluation (LREC).

[23]. Guillaume, G., Gilles, A., Niklas, P., Matthieu, C., Aude, G., Olivier. (2012). The ETAPE corpus for the evaluation of speech- based TV content processing in the French language, In: Proceedings of the International Conference on Language Resources, Evaluation and Corpora (LREC).

[24]. Nicolás, M., John, H., Hansen, L., Doorstep, T. T. (2005). MFCC Compensation for improved recognition filtered and band limited speech, Center for Spoken Language Research, University of Colorado at Boulder, Boulder (CO), USA.

[25]. Santosh, K. G., Bharti, W. G., Pravin, Y. (2010). A Review on Speech Recognition Technique, International Journal of Computer Applications (0975 – 8887), 10(3).

[26]. Manoj, K. S., Omendri, K. (2015). Speech Recognition: A Review, Special Conference Issue: National Conference on Cloud Computing & Big Data.

[27]. Benjamin., B. (2016). Reconnaissance Automatique de la Parolepour la transcription et le sous-titrage de contenus audio et vidéo, 52 av. P. Sémard -94200 Ivry-sur-Seine, Authôt.com - November 2016, 64.

[28]. Gravier, G., Adda, G., Paulson, N., Carré, M., Giraudel, A., Galibert, O. (2012). The ETAPE corpus for the evaluation of speech- based TV content processing in the French language, LREC Eighth International Conference on Language Resources and Evaluation, p. na.

[29]. Kahn, J., Galibert, O., Quintard, L., Carré, M., Giraudel, A., Joly, P. (2012). Apresentation of the REPERE challenge, Content- Based Multimedia Indexing (CBMI), 2012 10th International Workshop on, IEEE, 1-6, 2012.

[30]. Zied, E., Benjamin, L., Olivier, G., Laurent, B. (2019). Prédiction de performance des systèmes de reconnaissance automatique de la parole à l'aide de réseaux de neurones convolutifs, HAL Id: hal-01976284, TAL. Volume 59 - no 2/2018.

[31]. Galliano, S., Geoffrois, E., Mostefa, D., Choukri, K., Bonastre, J.-F., Gravier, G. (2005). The ESTER phase II evaluation campaign for the rich transcription of French broadcast news, Interspeech, 1149-1152.