Implementation Challenges in Sindhi Speech Recognition System

Author's Details:
[1]Tuba Qureshi, [1]Dil Nawaz Hakro, [2]Intzar Lashari, [1]Rajota Kharwal, [1]Maryam Hameed
[1]Institute of Information and Communication Technology, University of Sindh, Jamshoro,Pakistan
[2]Institute of Business Administration, University of Sindh, Jamshoro,Pakistan
Corresponding emails: dill.nawaz@gmail.com,
Contact Number: +923333774422

Abstract:
Significant efforts have been made on the recognition of speech and then it has become the important type of input for the computers as well machines. Latin scripts are easy to recognize and significant advancements regarding the speech recognition have been made. Arabic and its related languages like Sindhi are still lacking the speech recognition systems. An integrated system for Sindhi language is the need of time. This paper presents the challenges posed by various scripts and speech recognition system of Sindhi as a general. The challenges faced by researchers around the world specially for developing a system for Sindhi speech recognition system. The paper is organized as introduction followed by various challenges pose in developing speech recognition systems.
Keywords: *speech recognition, Sindhi, speaker, noise, controlling machines.*

1. Introduction

Speech recognition is an art of work that select the words from a trained vocabulary of words. Every spectral vector is identified by a label (phoneme), identification of word based the matching the string of label, phone machines based on label and transition probabilities and markov chains (CM). The acoustic models were used to match the phonetic elements or phonemes. Phonemes are generated by the label of each word by an acoustic processor in response of spoken word or input (Bahl et al.,1998). Automatic speech recognition is an advanced technology, and play a dynamic role in past decades. Application of automatics speech recognition can be benefits from a confidence measure (CM) to get a reliable output. The logistic regression (LR) classifier represent the simple input function to approximate naïve Bayes (NB) behavior (Cortina et al.,2016). Current speech recognition systems are capable in handling and understanding the speech (voice) efficiently resulting high accuracy, the present invention relates to recognize the efficiently when the speaker and moderately changing the objects and their places. An ASR system consist of one is sound source localization module and second is a sound source separation module, a sound source localization is used to localize a direction of the speaker sound noticed by different microphones, the second is a sound source separation module is which is based on the signal separation of speakers from the acoustic signals, a sound direction is grounded on a memory available in acoustic model which adds these models that adjust the plurality of direction at the time of interval, A module formed by acoustic model contains localization module along with the local sound, a sound direction (Nakadai .,et al 204). A natural speech is translated for the development of a device in automatic speech recognition. The main broad areas of the speech recognition are: 1) single letter or character recognition where pauses are the sign for the separation of words. 2) continuous speech recognition, where the process of recognition sentence is created continually such as natural order, 3) Understanding speech in which aim is transcription such as a data base query system or a robot systems. These systems are reply correctly to a spoken word, request and instruction (Bahl et al., 1983).

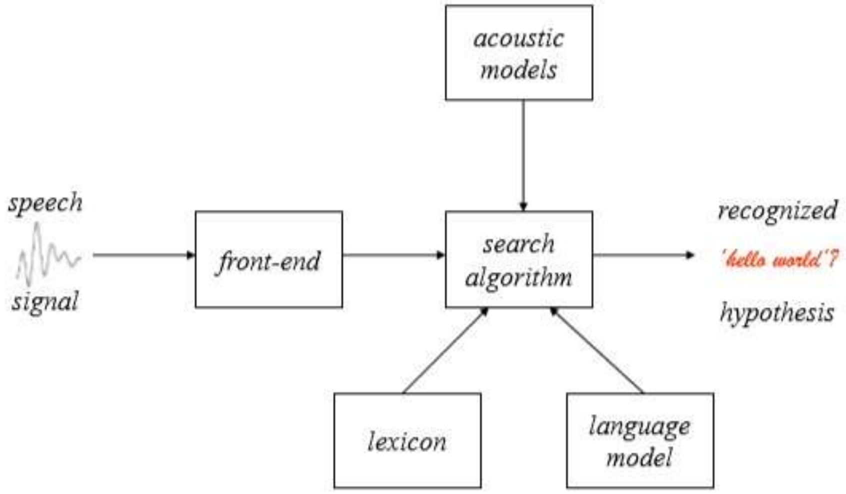

Fig:1 simple speech recognition system

2. World of languages and scripts

It is very difficult to count the spoken and written languages of the world. For attempting to count the languages we can visit the website named ethnologue.

The list of languages 6909 is shown the existing languages available in the world, the calculating of the existing scripts is a very complicated task (Besacier et al., 2014) languages is based on the gravity on that language and its speaker's crystal (2000), spoken languages is the primary source of human communication, some languages are mostly used for verbal communication not for the written. According to the (krauwer,2003) they defined the position of the languages and the idea of BLARK (basic language resource kit) were well-structured and a combined effort among the network of languages in Europe. ELSNET and ELRA are the projects based on the negligible set of resources to make the availability for many languages as possible.

3. language Recognition

Spoken language recognition is the emerging branch of the speech signal processing, usually spoken language recognition is also called the (language recognition) the main purpose is to recognize the language of spoken words with the help of application such as multilingual speech recognition, speech translation (Muthusamy et al.,1994; zissman and berkling,2001). Language recognition consist on the two comprehensive types such as: acoustic model methods and phono tactic methods.

3.1 Acoustic model methods

Acoustic spectra feature vector model advert to as a spectrum methods, the standard model methods contain the (Torres-carrasquillo,2002) support vector machine (SVM) and Gaussian mixture model(GMM), (Zhang et al.,2006), SVM and support with the GMM super vector(GSV) (Torres-carrasquillo,2008) they mostly work on the I-vector method (Dehak et al.,2011)

3.2 Phono tactic Methods

Phono tactic methods is used to apply on the phone recognizer, this method also bring up token method, core purpose is tracked by the language model zissman and singer (1994), this model is used by the phone recognizer to get the string token and n-gram language models as the backed to the co-existence of the tokens. Phototactic methods initially decode the noise into a token string or lattice and then model the token string using n-gram lexicon model (Hazen and Zue, 1993; Zissman and Singer, 1994). Phone recognizer is tracked by the vector space model (PR-VSM) is the part of the phototactic method for spoken language recognition

4. Speech synthesis

Rebai and Ayed (2015) proposed that synthesis system based on the text to speech (TTS) is significantly available for many of the world languages. TTS system is a modern technology due to growing field of application such as aids for handicap, multimedia and telecommunication. There are 442 million speakers of Arabic language and approximately there are five million Arabic speakers are blind around the world (zaki et al.,2010). Speech synthesis is a process in which text output is generated from speech. Hunt and black (1996) proposed the two way of speech synthesis 1) is the concatenative speech synthesis which is also known as (corpus based method) and 2) is the also known as knowledge based approach statistical parametric speech synthesis for generating of output as speech from the input which is given in form of text (black et al.,2007).

4.1 Speaker Modeling

Today's Visual and audio technology is used widely. There are many definition of speech recognition systems such as defined by (Jurafsky,2000) in which he defines speech recognition is the structure of system for recording signals to a string of words. Speech recognition is the Voice or it is the capability of a machine or program to recognize and carry out spoken commands. Speech is consisted on the text-dependent or the text independent, the application of text dependent the system has existing knowledge of the spoken words Reynolds (2002). The main attributes of the speaker model are that1) theoretical underpinning in which speaker understand the performance, mathematical approach improvement, 2) generalizable this is used to simplify the new data therefore suitable data matched the new data,3) parsimonious representation is based on the size and the computation of words, there are many type off technique that have been used in the speaker verification system

4.2 Noise Reduction Technique

Hirsch and Enricher (1995) proposed the two-new techniques for estimating the area of noise and the characteristics which make noise noisier beyond any explicit in detection of speech pause, mostly reduction of unwanted sound is consisting on single channel. It is hard to eliminate the unwanted noise between the pure noise. (Campernolle 1989; Martin 1993; Hirsch 193) presented that a noise which is available in the ground is not considered as signal to noise ratio is not considered low or it may not be the part of the stationary. Few methods are used to cope by avoiding noise problem from the noisy speech or its any of the segment.

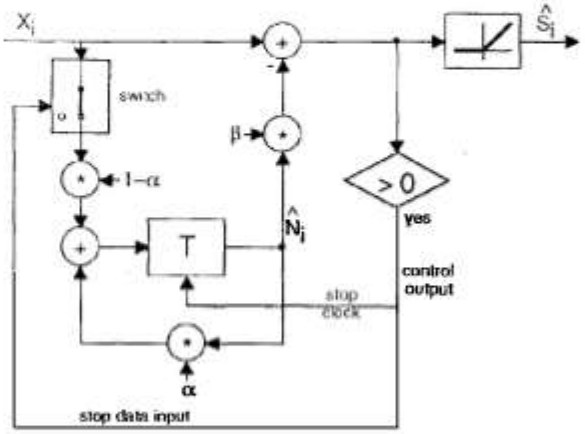

Fig:2 Simple noise reduction scheme is one sub band

5. Technique of speech recognition

Speech recognition is the part of pattern recognition, haykin (1999) presented the three solution to handle that problem such as1) Artificial Neural Networks (ANNs) 2) hidden Markov modeling and 3) dynamic time warping, (Bourobua et al., 2006). Many authors presented the many ways but the main part of each authors research is that they all are approved that, neural network made up of neurons, each processing units are pre-maintained by using input-output data set which are deliver to the network to evaluate the good outcomes, recognizer help to test each data set to produce the best outcomes (haykin 1999; kohonen 1987; widrow, darpa 1988). Dede and sazli (2009) presented artificial neural network is used to fulfil and scattered the speech recognition, which is consist on two technique in which the first step is pre-processing for the digital signal processing (DSP) and the other is the post-processing of artificial neural networks(ANN) techniques.

5.1 pre-processing

Every speech signal is typically well defined and structured in any of the type of the speech recognition, each word recognized properly which is fit to that word then, feature vector is used to get the applicable information, therefore pre-processing is the combination of the speech signals from recording step to feature extraction step.

5.2 post-processing

According to the model post-processing is based on the action on which feature vector are demonstrated on that word which is to be recognized or classified as similar to that model. (Bourobua et al., 2006; Ahad et al.,2002; Azam et al.,2007; Alotaibi 2005), These authors presented the artificial neural network model there are three different type of neural network are such as, Multilayer Perceptron (MLP), probabilistic neural networks and Elman where the relative accuracy and performance is calculated.

6. Prior Effort Using Neural Network Acoustic Models

Dahl et al., (2012) proposed the large vocabulary speech recognition (LVSR) which is using in phone recognition, this research presents the mixture of artificial neural networks (ANNs) and hidden markov models (HMMs), in the last of 1980s and the start of 1990s, perceive the all-inclusive survey of different architectures and a training algorithms which is presented by the different authors (Trentin and M.

Gori,2001), most relevant works of the ANNs are to estimate whereas the HMM state the following probabilities (Boulard and Morgan.,1993; Robinson et al.,2002).

7. Conclusion and Future work

The use of evolutionary approaches the problems of the Sindhi speech recognition system will be coped and an integrated speech recognition system can be developed to the acceptable maturity level. A high performance with enhanced accuracy Sindhi speech recognition system may be achieved if the available approaches may be enhanced and refined. This paper presented the problems, issues and challenges in the way of developing of Sindhi speech recognition system. A Simple speech recognition will be the first step followed by a comprehensive speech recognition system which can help to communicate with machines and control various types of machines by voice in different languages.

References

1. A. J. Robinson, G. D. Cook, D. P. W. Ellis, E. Fosler-Lussier, S. J. Renals, and D. A. G. Williams, “Connectionist speech recognition of broadcast news,” Speech Commun., vol. 37, pp. 27–45, May 2002.
2. Ahad, A. Fayyaz, T. Mehmood, Speech recognition using multilayer perceptron, in: Proc. of the IEEE Conference ISCON’02, vol. 1, 2002, pp. 103–109.
3. B. Widrow (Ed.), DARPA: Neural Network Study, AFCEA International Press, 1988.
4. Bahl, L. R.; DeGennaro, S. V.; Mercer, R. L. & others (1988), 'Speech recognition system', Google Patents, US Patent 4,718,094.
5. Bahl, Lalit R., Frederick Jelinek, and Robert L. Mercer. "A maximum likelihood approach to continuous speech recognition." IEEE transactions on pattern analysis and machine intelligence 2 (1983): 179-190
6. Besacier, L.; Barnard, E.; Karpov, A. & Schultz, T. (2014), 'Automatic speech recognition for under-resourced languages: A survey', Speech Communication 56, 85 - 100.
7. Black, A., Zen, H., Tokuda, K., 2007. Statistical parametric speech synthesis. In: International Conference on Acoustics, Speech, and Signal Processing, pp. 1229–1232
8. Crystal, D.,2000 language death. Cambridge CPU
9. D. Van Campernolle, "Noise Adaptation in a Hidden Markov Model Speech Recognition System·, Computer Speech and Language, Vol. 3, pp. 151-167, 1989
10. Dahl, G.E., Yu, D., Deng, L. and Acero, A., 2012. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, *20*(1), pp.30-42.
11. Dede, G. and Sazlı, M.H., 2010. Speech recognition with artificial neural networks. *Digital Signal Processing*, *20*(3), pp.763-768.
12. Dehak, N., Kenny, P., Dehak, R., 2011. Front-end factor analysis for speaker verification. IEEE Trans. Audio Speech Lang. Process. 19 (4), 788–798.
13. E. Trentin and M. Gori, “A survey of hybrid ANN/HMM models for automatic speech recognition,” Neurocomputing, vol. 37, pp. 91–126, 2001.
14. H. Boulard and N. Morgan, “Continuous speech recognition by connectionist statistical methods,” IEEE Trans. Neural Netw., vol. 4, no. 6, pp. 893–909, Nov. 1993.
15. H. Bourobua, M. Bedda, R. Djemili, Isolated words recognition system based on hybrid approach, Informatica 30 (2006) 373–384.
16. H.G. Hirsch, "Estimation of Noise Spectrum and its Application to SNR Estimation and Speech Enhancement", Technical Report TR-93-012, International Computer Science Institute, Berkeley, USA, 1993
17. Hazen, T.J., Zue, V.W., 1993. Automatic language identification using a segment-based approach. In: Proc. Eurospeech, vol. 2, Berlin, pp. 1303–1306

18. Hirsch, H.G. and Ehrlicher, C., 1995, May. Noise estimation techniques for robust speech recognition. In *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on* (Vol. 1, pp. 153-156). IEEE.
19. http://www.ethnologue.com/
20. Hunt, A., Black, A., 1996. Unit selection in a concatenative speech synthesis system using a large speech database. In: International Conference on Acoustics, Speech, and Signal Processing, IEEE, pp. 373–376.
21. J. H. M. Daniel Jurafsky. Speech and Language Processing, An introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Prentice Hall, Upper Saddle River, New Jersey 07458, 2000
22. Muthusamy, Y.K., Barnard, E., Cole, R.A., 1994. Reviewing automatic language identification. IEEE Signal Process. Mag. 11 (4), 33–41.
23. Nakadai, K.; Tsujino, H. & Okuno, H. (2004), 'Automatic Speech Recognition System', Google Patents, US Patent App. 10/579,235.
24. R. Martin, "An Efficient Algorithm to Estimate the Instantaneous SNR of Speech Signals", Eurospeech - 93, pp.1 093-1 096, 1993
25. Rebai, I. & BenAyed, Y. (2015), 'Text-to-speech synthesis system with Arabic diacritic recognition system', *Computer Speech & Language* **34**(1), 43 - 60.
26. Reynolds, D.A., 2002, May. An overview of automatic speaker recognition technology. In *Acoustics, speech, and signal processing (ICASSP), 2002 IEEE international conference on* (Vol. 4, pp. IV-4072). IEEE.
27. S. Haykin, Neural Networks: A Comprehensive Foundation, 2nd ed., Prentice–Hall, Inc., Englewood Cliffs, NJ, 1999.
28. S.M Azam, Z.A. Mansoor, M.S. Mughal, S. Mohsin, Urdu spoken digits recognition using classified MFCC and backpropagation neural network, in: Computer Graphics, Imaging and Visualization Conference, 2007.
29. Sanchez-Cortina, I.; Andrés-Ferrer, J.; Sanchis, A. & Juan, A. (2016), 'Speaker-adapted confidence measures for speech recognition of video lectures', Computer Speech & Language 37, 11 - 23.
30. T. Kohonen, State of the art in neural computing, in: IEEE First International Conference on Neural Networks, vol. 1, 1987, pp. 79–90.
31. Torres-Carrasquillo, P., Singer, E., Campbell, W.M., et al., 2008. The MITLL NIST LRE 2007 language recognition system. In: Proc. Interspeech, Brisbane, pp. 719–722.
32. Torres-Carrasquillo, P.A., 2002. Language Identification using Gaussian Mixture Models (Ph.D. thesis), Michigan State University.
33. Y.A. Alotaibi, investigating spoken Arabic digits in speech recognition setting, Inform. Sci. 173 (2005) 113–129
34. Zaki, M., Khalifa, O., Naji, A., 2010. Development of an Arabic text-to-speech system. In: International conference on computer and communication engineering, pp. 1–5.
35. Zhang, W., Li, B., Qu, D., Wang, B. 2006. Automatic language identification using support vector machines. In: Proc. ICSP, vol. 1, Guilin
36. Zissman, M.A., Berkling, K.M., 2001. Automatic language identification. Speech Commun. 35 (1-2), 115–124.
37. Zissman, M.A., Singer, E., 1994. Automatic language identification of telephone speech messages using phoneme recognition and N-gram modeling. In: Proc. ICASSP, vol. 1, Adelaide, pp. 305–308
38. Biswas, A.; Sahu, P. & Chandra, M. (2015), 'Multiple camera in car audio visual speech recognition using phonetic and visemic information ', Computers & Electrical Engineering 47, 35 - 50.