A Review of Challenges in Automatic Speech Recognition

Harshalata Petkar
Assistant Professor
MGAHV,Wardha
Maharashtra,India

ABSTRACT

Speech is the nature's gift to the human being which contributes towards the intelligence and discrimination from rest of the animal kingdom. Taking into consideration technological aspects, speech recognition is the buzzword today, as communication and hands free computing evolving day by day. Speech is a very important mode of the communication and interaction with the digital computer. Speech recognition along with the wide range of applicability in domain of computer science, medical science, psychology, sports, neurology has many challenges while developing. Developing real time speech recognizer may hurdle from adverse environment to anatomy of the human body. It also involves linguistic aspects too. This paper explores various challenges in developing a robust ASR system.

General Terms

Signal Processing and linguistics

Keywords

Speech, Speech recognition, communication, linguistics

1. INTRODUCTION

ASR robustness are a captivating research area today. Challenges in automatic speech recognition need to address to make the ASR system robust. Developing a robust ASR system starts with the study of critical issues in human speech. This paper will try to explore the challenges and issues which make the ASR performed poorly. The potential challenges that are identified carefully can help in the further course of development process. To enable the machine to imitate the human speech and understand it, involves a series of critical steps. The researchers are still working from last 5 decades and still there are several unfold challenges that yet not addressed properly. In this paper, we discuss different challenges involved which make the automatic speech recognition difficult at the technical dimension.

2. AUTOMATIC SPEECH RECOGNITION

Automatic Speech Recognition can be viewed as a mapping from continuous time signal, the speech signal, to a sequence of discrete entities-for example, phonemes, words and sentences [1] Since 1930 researchers are taking effort to develop a machine which can produce and recognize and understand the human speech as that of human being [2]. Automatic speech recognition is a field of computer science that collaborate interfaces, which understand spoken words and identify them that the person speaks into a microphone or phone. The problem posed by the general speech recognition has not been solved yet for any language. This means there is a scope to develop a versatile speech recognition system. The domain where ASR can play vital role in facilitating daily activities and where the current level of accuracy has its own importance.

It is well known that speech signal can be transmitted virtually anywhere in the world. This fact reveals demand of the research in speech recognition, natural languages processing and speech synthesis. To meet with the down line and launch the product at par we need to find out the difficulties and challenges in developing the Speech Recognition system. In specific application where the eye and hands mean to be busy, the speech can be the best medium to instruct the computer. Speech spectrogram has its own significance in medical sciences. Cry of a new born is a speech, the original languages spoken by the male and female is also a speech which meant for communication certain information. For people suffering from lack of motor control, crippled hands, vision disabilities, speech recognition is a boon. Anatomical Structure of the vocal tract is unique for every person and hence speech can be a biometric identity where it can be used for remote person authentication [4].

The list of applications of automatic speech recognition is so long and is growing; some of known applications include virtual reality, Multimedia searches, auto-attendants, travel information and reservation, translators, natural language understanding and many more applications [5,6].

Technology spectrum in the underdeveloped countries may be widened if the solution is available in regional spoken languages. It will open the gateway to a human being to talk to the computer with his/her own mother tongue without botheration of internationally recognized languages.

3. SPEECH PERCEPTION AND PRODUCTION

After speech production speech recognition system begins with the processing of giving a speech signal for the purpose of obtaining the discriminatory acoustic information about utterance. This acoustic information computed in numerical form called as features.

The speech recognition system resembles with the human speech perception and production. The source - channel model is illustrated in fig1. [7]

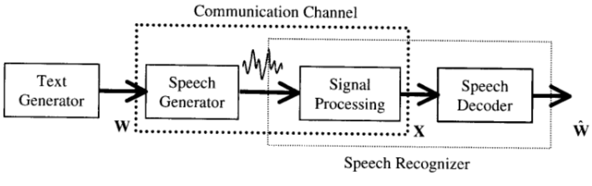

Fig 1: Source Channel Model for Speech recognition System

4. CHALLENGES WITH ASR

4.1 Psycho-Intellectual Aspect

4.1.1 Human Understanding of Speech:

Human have their own knowledge base which is resulted from the reading, experiments, experiences, examination, situation, interaction and communication. They may hear more than the speaker speaks to them. While speaking speakers have its own language model of native language. Human may understand and interpret the words or sequence of words, they never heard before.

In automatic speech recognition we need to develop the annotated corpus and language model which provide the system with the limited grammatical structures. To increase the chance of pattern matching one can use the statistical models like Hidden Markov model. At the level of human understanding the knowledge can be represented by the exhaustive content. In ASR the limited knowledge needs to build and can be trained accordingly. Vocabulary size plays a crucial role in speed of the system. Search time and resources get increase due large vocabulary. Limited vocabulary has the restricted application domain.

4.1.2 Spoken Language and Written Language:

Spoken language that has been spoken by the people is not similar to that of written language. Spoken languages are less complicated than written languages. In spoken language people use shortened unit, repetition of words to emphasize important ideas and information, use of more familiar words to increase audience understanding, In ASR this problem needs to identify and address. Speech is a two way communication and is dialogue oriented. The dialogue is delivered and is interpreted in the form of feature vector and to understand the meaning of words, we adopt and train the receiver speech data. In written communication the dialogue is one way and is bounded by the complex structure of language. The grammatically spoken language is quite different from written language at many different levels.

Some difference we pointed are as follows:

- In spoken language there is often a radical reduction of morphemes and words in pronunciation.
- The frequencies of words collection and grammatical construction are highly different between spoken and written language.
- The grammar and semantic of spoken language is also significantly different from that of written language: 30-40% of all utterance consists of short utterances of 1-2-3 words with no predicative verb.
- Use of more colloquial words and shortened form make the conversation lively
- Much less use of terms and phrases that work in writing but can lose their meaning or become confusing in speaking. Examples: "as mentioned above," "the former… the latter," and "respectively."[9]

The speech recognition system should address these issues at psycho-intellectual level, as spoken language and written language is different from each other.

4.1.3 Multimodality in human-human communication:

The indispensable need for speech system design is to understand the multimodal communication between human. Human contributes more with the body language while he communicates. It includes eye movement, postures, symbolic gestures, hand movement, human-human communication which is mediated by computer [8]. When knowledge from one modality is weak, it may be complimented by another modality. Auditory and -visual information can be used to its optimum for speech recognition. Speech recognizer can use this cue to ensure improvement in speech recognition accuracy.

4.1.4 Background Noise:

In a natural environment, human speaks the words and sentences which incorporate the background noise. The noise is unwanted information in the clearest speech signal. It may be an echo effect, another speaker, speaking in the background, playing instruments in the background and so on. The speech signals mixed with noise is difficult to recognize accurately. Noise contaminated speech signal gives rise to speech variability and recognition become difficult in real time events. Reverberation and noise elimination are most challenging tasks. A robust noise reduction algorithm can be deployed dynamically for this type of challenge.

4.2 Continuous Speech

The real time ASR system works with the continuous speech. Phonetic, syllabic and word level recognition is not complex as recognition performed in isolation, but when the sequence of words/sentence is spoken out recognition become more critical. We need to deal with the word boundary ambiguity which is a difficult problem for ASR. One way to address this difficulty is to give the pauses while speaking so that adjacent words can be identified clearly, but it tends to loss naturalness in the speaking style and also it needs the training of the speaker while speaking for increased length of the sentences. Current continuous speech recognition (CSR) systems with large vocabulary are strictly based on the principles of statistical pattern recognition [18], [19].

4.3 Channel Variability and Noise

The acoustic of the speech signal is affected by the noise as well as a microphone that is used for acquiring the speech signal. This is called as channel variability. This phenomenon can be addressed by applying the filters as well as using the high end microphones.

4.4 Speaker Variability

All speakers have their unique voice as fingerprint due to the unique anatomy of the vocal tract and personality The speaker variability can be,

4.4.1 Phone-Realization:

The way of speaking the same word may result in different pronunciation every time. Hence the acoustic wave of the speech may vary over time for the same utterance. This is the phone realization of speech.

4.4.2 Accent:

While investigating the variability between speakers through statistical analysis methods the first two principal components of variation correspond to the gender (and related to physiological properties) and accent respectively [11]. Indeed, compared to native speech recognition, performance degrades when recognizing accented speech and non-native speech

[12,3] In fact accented speech is associated with a shift within the feature space [13]. The accents of speaker differentiate with respect to their personality. Each one is having the unique way of pronunciation and emphasize. The speaking accent may vary according to the social and personal situations. Speaker uniqueness results from the complex combination from pschysiological and cultural aspects. [10, 11] We may vary in accent while speaking to parents, friends. It also affects with emotional level, which may affect the loudness, pressure while speaking.

4.4.3 Gender of the Speaker and Anatomy of Vocal Tract:

Men and women have different fundamental frequencies while speaking. It is because women have shorter vocal tract than men have. The fundamental frequency of a woman's voice is roughly two times higher than a man’s voice. The shape and length of the vocal cords, formation of cavities and size of the lungs may change over time and it may cause the speaker variability. Children have shorter vocal tract and vocal folds compared to adults. This results in higher positions of formants and fundamental frequency. The high fundamental frequency is reflected in a large distance between the harmonics, resulting in poor spectral resolution of voiced sounds. The difference in vocal tract size results in a non-linear increase of the formant frequencies. In order to reduce these effects, previous studies have focused on the acoustic analysis of children's speech [14] This variability issue has been addressed by vocal tract length normalization [15] as well as spectral normalization [16]

4.4.4 Speed of the speech:

The speed while speaking may vary and depend upon the situations, physical stress. Rate-of-speech (ROS) is considered as an important factor which makes the mapping process between the acoustic signal and the phonetic categories more complex. Timing and acoustic realization of syllables are affected due in part to the limitations of the articulatory machinery, which may affect pronunciation through phoneme reductions, time compression/expansion, changes in the temporal patterns, as well as smaller-scale acoustic–phonetic phenomena. Due to rate of speech significant degradation in performance has been reported [17]

4.5 Regional and Social Dialects

Dialects are group related variant within language. Regional dialects are the features of pronunciation, vocabulary and grammar which differ according to the geographical area of the speaker. [20]

In many cases we may be forced to consider dialects as another language in ASR due to large differences between two dialects. Dialects of the specific language differ from each other, but they are still understandable by the speaker of another dialect of the same language. Differences among dialects are mainly due to the social and regional factor. During the past few years there have been significant attempt to automatically recognize the dialect or accent of the a speaker given his her speech utterances [21,22,23,24], Recognition of dialects or accents of speakers prior to automatic speech recognition (ASR) helps in improving performance

The ASR systems by adapting the ASR acoustic and/or language models appropriately [25].

4.6 Amount of Data and Search Space

Day to day communication produces the largest amount of speech data. This has to match with the group of phones, sounds, the words and the sentences.

The quality of input can be regulated by the number of samples of the input signal, but the quality of the speech signal will decrease with the lower sampling rate.

4.7 Ambiguity

Spoken language have the ambiguity i.e. the set of words have different meaning. There is ambiguity that needs to deal with speech recognition are homophones and word boundary ambiguity.

4.7.1 Homophones:

Homophones are the words that sound the same but have different orthography. The words or sounds same, but the meaning is different. How ASR will distinguish between homophones? It is impossible at the world level is ASR, we need a larger context to decide which is intended.

4.7.2 Word boundary ambiguity:

When the sequences of groups of phones are put into a sequence of words, we sometimes encounter word boundary ambiguity. Word boundary ambiguity occurs when there are multiple ways of grouping phones into words. There are some examples which naturally occur during the spoken communication which ends with the rhyming sounds. This can be viewed as a specific case of handling the continuous speech where even a human can have problems with finding the word boundaries.

5. CONCLUSION

As technology continues to cultivate and shape up new application which are integral part of human life. The challenges of designing a speech system that mimics the human being is still going forward Indeed, improving the accuracy in speech recognition system, regardless of the challenges pertaining to speaker variability, channel variability and environment variability is a matter of counteracting the effects. Characterization of effects can be studied first to accomplish the robustness in ASR. This paper reviewed the various challenges, if addressed properly can result in real time robust ASR.

State of the art speech recognition systems is concentrating on modeling the speech variance parameters such as psycho-intellectual aspects, dialects, accents, rate of speech, speaker gender, age, health and emotional state as well as some physiological variables such as differences in vocal tract length. These variations lead to the difficulties in modeling large-scale, robust speaker-independent systems. Human want to speak to computer without losing naturalness then all the challenges and difficulties must be addressed.

6. ACKNOWLEDGMENT

I am thankful to my supervisor Dr. Vijaykumar Kaul and Dr. Lobiyal and member of center for technological studies, Mahatma Gandhi International Hindi Univesity, Wardha, India

7. REFERENCES

[1] John Makhoul and Richard Schwartz, “State of art in continuous speech recognition” proceeding National Academy of Science,USA Vol 92 pp9956-9963 october1995

[2] H. Dudley, The Vocoder, Bell Labs Record, Vol. 17, 122-126, 1939.

[3] Lawson, A.D., Harris, D.M., Grieco, J.J., 2003. Effect of foreign accent on speech recognition in the NATO N-4 corpus. In: Proceedings of Eurospeech, Geneva, Switzerland, pp. 1505–1508.;

[4] Vibha Tiwari, International Journal on Emerging Technologies 1(1): 19-22(2010) ISSN : 0975-8364 MFCC and its application in speaker recognition

[5] Scan soft (2004). Embeded speech soloutions retrieved January 25, 2005 from http://www.speechworks.com/

[6] Robertson, J., Wong, Y.T., Chung, C., and Kim, D.K., (1998) Automatic Speech Recognition for Generalised Time Based Media Retrieval and Indexing, Proceedings of the sixth ACM International Conference on Multimedia(pp 241-246) Bristol, England.

[7] Huang, X., Acero, A., Hon, H., 2001. Spoken Language Processing. Prentice-Hall, PTR, Upper Saddle River, NJ.

[8] *Multimodality in Language and Speech Systems Björn Granström, David House, and Inger Karlsson (Eds.).* Text, speech and Language Technology, Dordrecht,(2002)

[9] Article from url https://www.hamilton.edu/oralcommunication/spoken-language-vs-written-language

[10] Garvin, P.L., Ladefoged, P., 1963. Speaker identification and message identification in speech recognition. Phonetica 9, 193–199. (Garvin and Ladefoged, 1963; Nolan, 1983)

[11] *Nolan, F., 1983. The Phonetic Bases of Speaker Recognition. Cambridge University Press, Cambridge*

[12] Kubala, F., Anastasakos, A., Makhoul, J., Nguyen, L., Schwartz, R., Zavaliagkos, E., 1994. Comparative experiments on large vocabulary 782 M. Benzeghiba et al. / Speech Communication 49 (2007) 763–786speech recognition. In: Proceedings of ICASSP, Adelaide, Australia,pp. 561–564

[13] Van Compernolle, D., 2001. Recognizing speech of goats, wolves, sheep and ... non-natives. Speech Communication 35 (1–2), 71–79.

[14] Lee, S., Potamianos, A., Narayanan, S., 1999. Acoustics of children speech: developmental changes of temporal and spectral parameters. The Journal of the Acoustical Society of America 105, 1455–1468.

[15] Das, S., Nix, D., Picheny, M., 1998. Improvements in children speech recognition performance. In: Proceedings of ICASSP, vol. 1. Seattle, USA, pp. 433–436.

[16] Lee, L., Rose, R.C., 1996. Speaker normalization using effcient frequency warping procedures. In: Proceedings of ICASSP, vol. 1. Atlanta, Georgia, pp. 353–356.

[17] Martinez et al., 1997; Mirghafori et al., 1995; Siegler and Stern, 1995

[18] RABINER, L.R., JUANG, B., Fundamentals on Speech Recognition, New Jersey, Prentice Hall, 1996.

[19] HUANG, X., ACERO, A., HON, H.W., Spoken Language Processing: A Guide to Theory, Algorithm and System Development, New Jersey, Prentice Hall, chapter 11, 2001.

[20] [20]Linguistics: An introduction to language and communication

[21] Louis Boves and Johan de Vethd. Comparison of channel normalization techniques for automatic speech recognition over the phone. In Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on, volume 4, pages 2332 {2335 vol.4, oct 1996.

[22] Gang Liu and John L. Hansen. A systematic strategy for robust automatic dialect identi_ cation. In EUSIPCO2011, pages 2138{2141, 2011.Gang Liu, Yun Lei, and John H.L. Hansen. Dialect identi_ cation: Impact of di_erences between read versus spontenous speech. In EUSIPCO2010,pages 49{53, 2010.

[23] J ohn Nerbonne. Linguistic variation and computation. In Proceedingsof the tenth conference on European chapter of the Association for Computational Linguistics - Volume 1, EACL '03, pages 3{10, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.

[24] Pedro A. Torres-Carrasquillo, Douglas A. Reynolds, and P. Gleason.Dialect identi_cation using gaussian mixture models. In ISCA, pages757{760, 2004.

[25] Mingkuan Liu, Bo Xu, Taiyi Hunng, Yonggang Deng, and Chengrong Li. Mandarin accent adaptation based on contextindependent/context-dependent pronunciation modeling. In Proceedings of the Acoustics, Speech, and Signal Processing, ICASSP '00, pages II1025{II1028, Washington, DC, USA, 2000. IEEE Computer Society.