
By C.Moreno, A. Antolin and F.Diaz-de-Maria.

Summary
By

Maheshwar Jayaraman
1. Introduction

Voice Over IP is the technology, which is used to transmit Voice over the network using IP packets. This technology is similar to transmitting data over the web, the only difference being that the data happens to be audio data. Speech Recognition is a technology which is used to recognize human speech and reproduce that on the computer screen. This paper proposes a improvised model to recognize speech from the IP packets which carry the voice data. The paper explains the drawbacks of existing speech recognition system and proposes a solution for those problems. The proposed modal is compared with existing technologies and the experimental results are documented. This paper summarizes the proposal and gives the reader a brief overview of the approach. The flow of this overview is as follows; section 2 discusses about VoIP and section 3 talks about Speech Recognition. Section 4 discusses the author’s proposal followed by the conclusion in section 5.

2. VoIP

2.1. Introduction

VoIP stands for Voice Over Internet Protocol (IP) and this is the technology of transmitting voice or audio data over the Internet by embedding them in IP packets. Audio voice is analog in nature but IP packets carry digital data, so the analog data is converted to digital format before it is embedded in IP packets. The general operation of the Internet is that data is embedded in IP packets at the sending end. The source then sends the IP packets over the Internet, which is received by the receiver at the other end. The receiver extracts data from the IP packets and utilizes that data.

Data=>IP Packets=>Internet => retrieve Data from IP packets

For VoIP the input would be in analog format and so the analog data is first converted to Digital by using a Analog to Digital Converter (ADC). The digital data is then compressed, embedded in IP packets and then sent over the Internet. The receiver at the other end, gets the packet, extracts the Digital data and then converts it to analog by using a Digital to Analog Converter (DAC) and then feeds the analog voice data to a sound card to reproduce the sound.

Voice(Source) =>ADC=> Internet =>DAC=>Voice(receiver)

2.2. Problems associated with VoIP.

VoIP deals with huge amount of data and these huge data posses some constraints. The problems associated with transmitting Voice are

a. **Real-Time Constraints.**
   VoIP is used for real time applications, and then the data has to reach the other end within a stipulated amount of time. If the data is delayed then the receiver has to wait for a long time to receive the data and this might degrade the performance, as real time applications are very sensitive to data and delays. The other problem associated with VoIP is real-time applications are very sensitive to data loss. If a packet is lost, the amount of information lost is huge as data is sent in compressed form. Data has to be reconstructed when packets are lost on the network.

b. **Bandwidth Constraints.**
   VoIP involves huge amount of data and large bandwidth is required for transmitting such huge data. Real time streaming of audio data requires lot of bandwidth and so compression algorithms are used to reduce the size and thereby reduce the bandwidth requirements. The compression algorithms compression ratio are huge and because of this, the probability of errors increases. The errors caused due to compression, causes the reproduction/decompression of data at the receiving side to be less than perfect and quality of the voice degrades.

   VoIP uses ADC to convert Analog data to Digital format and uses G.723.1/G729 compression algorithms to compress the data. VoIP then uses Real Time Protocols (RTP) to encapsulate voice data in IP packets and then uses H323 protocol for signaling purposes.

3.1 Introduction.

Speech recognition can also be called as Voice Recognition and this technology is used to recognize voice and reproduce that on the computer. Speech recognition finds use in applications like Online Dictation, Remote Typing etc. Speech Recognition is classified in to two types.

a. Speaker Independent Type and
b. Speaker Dependent Type.

Speaker Independent Speech Recognition recognizes voice of any speaker and it is generalized for all kinds of speaker whereas Speaker Dependent Systems can recognize voice of one particular person only. Each System is further classified as Isolated Speech Recognition (ISR) and Continuous Speech Recognition (CSR). ISR systems can recognize single letters and characters only whereas CSR systems can recognize spoken words.

3.2 Problems associated with VoIP.

The problems associated with Speech Recognition are explained as follows.

a. Recognizing words requires the computer to be very powerful. The computer should process high processing speed to perform the complex tasks in small amount of time.
b. Cost of computers increase with processing power and complexity of processing.
c. The systems should be highly trained to recognize the set of words or characters and ISR systems require more complex training to recognize all peoples accent and speed.
d. Most Speech Recognition systems require the speaker to speak slowly.

4. Front-End for Recognizing VoIP.

4.1 Goal of the Modal.

The authors C. Moreno and et.all. proposed a robust front end to recognize the VoIP packets. The modal proposed a modified approach to recognize voice and also had a different way to deal with packet loss and data assumption. The modal proposed to recognize voice directly from the encoded data as opposed to existing Speech Recognition systems, which recognize voice from the decoded version of the data. Such an approach resulted in better performance and exhibited better performance in case of packet loss. The modal compared its approach with existing Automated Speech Recognition (ASR) systems and the results are tabulated.

4.2 Problems with Existing Systems.

Voice is encoded, compressed before transmitting and the existing ASR systems receive the packet, decompress the data, decode it and then use it. The encoding and compression algorithms are lossy and some error is introduced in the data during encoding and compression process. The data is decoded and decompressed at the receiver, and this decoding and decompressing process is also lossy. Errors are introduced both at transmitting and receiving end and this propagation of error affects the performance of the ASR systems. In addition to the lossy nature of the algorithms itself, the error induced in the initial stages of the encoding would mean that reproduction of data would be erroneous. This implies that the decoded data would not be same as the encoded data even if the decoding algorithm was perfect. Thus, errors are introduced at all the stages and all these errors add up to degrade the performance of the ASR systems. One other problem in existing systems is of packet loss. If the receiver losses a packet, it tries to estimate the contents of the lost packet from the data it received from the previous packet. This means that if the systems has a burst of lost packets the quality of estimated data reduces and finally the system reaches a point from which it cannot recover as the quality of the estimated data decreases.

The types of errors that are introduced in the ASR systems are explained below.

Coding Distortion:

Voice is encoded before transmission on the net and VoIP uses G.723.1 and G.729 codec for encoding the data. G.723.1 and G.729 codec are standards used for encoding. H323 protocol is used for signaling purposed between the source and the receiver. G.723.1 and G.729 codec provides the option of dual coding rate of 5.3 and 6.3 Kbps. Both codec achieve low bit rates by using Source-filter modal. These codec introduces two kinds of distortions namely,

a. Distortion due to quantization of parameters to be transmitted
b. Inadequacy of the source-filter modal.
Packet Loss:
Packet loss deteriorates the overall performance of the voice recognizing front end. The systems use compression algorithms, which have a high compression ratio, and so the amount of data sent in each packet is significant and large. Loss of one packet means loss of significant amount of data and the lost data has to retrieved. Some suggested methods for recovering from packet loss are Forward Error Correction (FEC) algorithms and Interleaving. The codec used in VoIP support low frame rate and so loss of one packet implies loss of large amount of data. When a packet is lost, the pattern of the missing packet is assumed or estimated from previously correctly received packets. When a stream of data packets is lost, the system degrades a lot and recovery is practically impossible after that.

The proposed modal tries to achieve better performance in case of burst of packet loss and also tries to achieve better performance during normal reception of VoIP packets.

4.3 Existing Approach and Proposed Approach.
Existing ASR sytems uses Speech Codec to code the speech. G.723.1 is used as Speech codec in ASR systems. The encoded data is used for recognizing speech at the receiver. The diagram depicting existing recognition modal and the approach specified in this paper is provided in the original paper in Fig 1. Details of speech recognition by the two modals in comparison can be refered from the paper. The modal proposed in the paper recognized using a different approach. The modal recognizes speech from the digital encoded form of data. The modal selectively accesses the needed and important parts of the encoded speech and develops the spectral envelope. This envelope is same as the spectral envelop derived from original speech except for that fact that it is quantized. This envelope will be affected by only the encoding process whereas the envelop derived from decoded speech would be affected by errors introduced during both the encoding and decoding stages.

5 Pros and Cons of Proposed System
The advantages of the existing system are
• Performance of system is affected only by quantization distortion of speech signal. Since the voice is recognized from the encoded speech, the estimated voice would not be affected by the errors of the decoder. Estimating the voice from a previous stage prevents the output from being contaminated by the extra stage.
• Distortions involved during decoding of encoded speech do not affect the performance.
• Proposed front-end is more effective during packet loss. The reason for this is that better estimation can be performed as the previously estimated packet is much more precise than the estimation from the decoded version and this implies better estimation in case of packet loss. If burst of packets are lost then even this modal cannot recover as the system cannot recover from that loss.
• Does not introduce any extra computational cost.

The shortcomings of the modal are
• The proposed modal is adapted to specific codec though the author suggests a probable solution to overcome this problem.
• Spectral envelope is available at the frame rate of the codec, which is very slow for voice recognition.

5.1 Experimental Results:
Experiments were performed for two Speech Recognition systems namely Speaker Independent (ISR) and Speaker independent (CSR). Performance of existing ASR systems was compared with the proposed modal. Two different databases were used for the ISR and CSR systems. Information about the databases can be got from Table I from the paper. The database was recorded at 16KHz in clean conditions. The modal also simulated the packet loss scenario to test their modal in case of packet loss. Packet loss was simulated by using Gilberts Model represented in Fig.2, in the paper, and it used a Markov chain with 2 states, good and bad. Good state represented a state, which has very less packet loss rate, and the bad state represented a state with high packet loss rate. The probabilities were set such that the system was highly unlikely to move from good to bad state, but once the system moved to the bad state, it will not return to good state that quickly. Bases on various loss rates the paper devised 6 kinds of channels each with variegated Packet Loss Rates (PLR) and variegated Mean Burst Rates (MBR). The channel was designed such that 90% of the bursts consisted of 3 packets or less. Details of the channel can be got from Table II from the paper.
Existing ASR system was used for speech recognition over the 6 different channels both for ISR and CSR systems. The ASR system performed well for the ISR systems over the different channels and the general trend was that the confidence level of the recognized speech decreases with increase in PLR. The system had a confidence level of 99% for ISR task. The systems performance degraded very much for channel F as it had large PLR and MBR. The ASR system had a low confidence rate of 87% or less for channels A through E but had a very low confidence level of 81% for channel F. This proves that the performance of existing ASR systems degrades faster during packet loss. The proposed modal had a confidence level of 99% for the ISR task and it was found out that the proposed modal performed slightly better than the existing ASR modal. The proposed modal had a noticeable improved performance for the CSR task. The proposed approach had a confidence level of 88% and less as opposed to 87% for the ASR system. The system had a confidence level of 83% for channel F whereas the ASR system had a confidence level of only 81% for the same channel. Thus the proposed system performed very well during packet loss. Detail figures can be referred from Table III through Table V from the paper.

5. Conclusions.

It was evident from the experimental results that the proposed modal performed much better than the ASR system for CSR task and had little improvement for the ISR task. The modal performed better in networks with high packet loss rates and the performance improved as the channel degraded further. As mentioned above the chances of error being introduced are at the stages of Compression, Coding, Decoding and Decompression. Errors introduced at one stage will definitely affect the performance of the other stage as the output of one stage happens to be the input for another stage. The proposed modal estimates the output from the encoded data and so the last two stages are bypassed. This implies that the error which would have hampered the recognized speech by the decoding part is avoided. This in turn improves the performance of the end recognized speech. The speech is recognized only from select components of the encoded speech and so not all errors of the encoded data would affect the output. To summarize, the modal tries to estimate speech by using minimum amount of data and avoids error by bypassing the decoding stage and thereby should produce better results than the traditional ASR system. The above conclusion is strengthened by the experimental results which proves that the proposed modal does perform much better than traditional ASR system. The modal also provided solutions for existing problems associated with VoIP and Speech Recognition. Comparison results with existing ASR (Automated Speech Recognition systems) prove that the proposed modal is effective. This approach could be always extended to other coding standards and it is not restricted to G.723.1 standard. The system performs well for ISR systems when PLR is greater than 1.13% but outperforms the ASR system in all situations for a CSR system.