Distributed Speaker Recognition

Veena Desai and Hema A. Murthy

Department of Computer Science and Engineering
Indian Institute of Technology, Madras, Chennai. 600 036.

hema@tenet.res.in

Abstract

Speech recognition systems are gaining increasing importance with the wide-spread use of mobile and portable devices and other interactive voice response systems. Because of the resource constraints on such devices and the requirements of specific applications, the need to perform speech recognition over a data network becomes inevitable. The requirements of such a system with a human at one end and a machine at the other end are clearly asymmetric. The major focus of this work is to enable speaker recognition for information access over the network. Assuming that at the client end the device is either a Personal Digital Assistant (PDA) or a cellphone, an attempt is made to perform part of computation at the client end, thus conserve bandwidth. Experiments have been performed on both TIMIT data and TIMIT data passed through a speech codec. The results indicate that by performing feature extraction at the client end, the bitrate can be reduced significantly to 13.6 kbps with 96% recognition performance.

1. Introduction

With the convergence of data, voice and other multi-media traffic over the Internet, the need for higher demands for network resources seems to be rising. When we look at the traditional VoIP (Voice over Internet Protocol) based applications which ensure strict quality and delays, there is an unavoidable need to reserve bandwidth and resources all the way from source to destination [1]. With the advent of ubiquitous computing, information access by voice has become imminent. Today most of the PDAs and cellular phones can be used for information access and hands-free dialing especially in car like environments. The transmission of speech over telephone seemed to be the primary solution over the period. Because of the attendant problem of channel noise, the recognition performance still remains extremely poor which has disguised the future of speaker recognition over a network. Today there is an integration of voice and data over a single network. Sending voice over the data network has a significant advantage over sending voice over a telephone network in that data integrity is guaranteed. In this paper we address the problem of performing speaker recognition over a network. We compare the performance of this system with that of data sent over a voice network. If voice is sent over the data network, the constraints on bandwidth becomes an issue. To overcome this problem, voice may have to be compressed by using a speech codec (coder-decoder) [2]. But this results in distortion. Alternatively, part of the computation can be performed at the client end and transmitted over the network. This may prove to be a better approach as the feature vectors are extracted from the clean speech and the data integrity is ascertained. Thus we make the client device perform limited computing or processing such as feature extraction and then transmit these features to a remote server for recognition [3] [4] [5]. Hence this work addresses the QoS requirements that are specific to Distributed speaker recognition (DSR) applications.

Unlike the streaming audio/video and other real-time applications where there are strict constraints on the delays and quality, the QoS requirements for such a system seem to be less stringent [6].

The important observations regarding the system are:

- At one end the client is a human.
- At the other end the server is a machine.
- Delays can be tolerated by the server.
- For the client, intelligibility of the message is important.
- For the server, data integrity is important.
- The server can tolerate a fairly significant packet loss.
- On the other hand, the client can tolerate only a small amount of packet loss which does not result in the loss of intelligibility of the message.

Different modes for the transmission of speech from the client device to the server such as using the voice network (telephone) and data network (VoIP, feature vectors) are being exploited in the following sections. The issues associated with the transmission of speech over a voice network and the speaker recognition performance are presented in Section 3. Section 4 deals with the performance and the related issues of transmitting speech over a data network. Section 5 includes a discussion on the results obtained and characterizes the protocol for transmission over a data network.
2. Baseline System

A GMM based speaker recognition system has been used for the experiments. The results stated below are obtained on a given set of speakers using 17-dimensional feature vectors of MFCCs (Mel Frequency Cepstral Coefficients) [7]. The database used is TIMIT [8]. The results are compared against the results obtained with the NTIMIT which stands for noisy telephone speech database. The TIMIT data is passed through G.729A(8kbps) [9] speech codec to simulate VoIP data.

3. Speaker Recognition over telephone channel

The major functional units of a speaker recognizer in general are feature extraction and recognition. The input to the recognizer is the speech signal. The subject of interest is to deal with a situation where the recognizer is remotely located and a communication channel separates the device from the recognizer. We assume that the results of the recognizer are communicated by voice as for example an e-mail reader. Our requirements are clearly laid out in the sense that we need a bidirectional VoIP kind of arrangement as against the VoIP applications such as music or video on demand and the interactive voice systems wherein there is voice data only on one side of the bidirectional connection, i.e., a system wherein the speech data is being sent from the client and the response from the server perceived as speech. The applications under consideration might include identity verification for banking transactions over the phone, voice identification for smart voice mail systems, etc.

The traditional ways to transmit the speech signal to a remote server is either through a telephone channel or as VoIP. In the former case, we need to be aware of several issues such as transmission channel distortion which have direct influence on the recognition output. Since we know that the application at the back end is speaker recognition and not play-back, the data that is being transmitted appears to be very large. The recognition performance is shown in Table 1. The effect of distortion or noise on the recognition output is apparent from Table 1. This result is obtained on the NTIMIT database for a set of 400 speakers which represents data sent over a telephone channel [10]. The requirements in terms of bandwidth also seem to be expensive which works out to be 64kbps. Hence the problem with telephone data is that the channel distortion is significant and the data rates are very high [11]. The alternative to overcome the problems associated with telephone channel transmission is speech being sent over the data network. The transmission through data network not only eliminates the need for channel compensation techniques but also contributes to a significant reduction in data rates. In this context, the data consists of either the coded speech as in the case of VoIP or the processed speech data for recognition. Whereas VoIP provides very low data rates by using speech codec and succeeds to reconstruct the signal for play-back, it introduces distortions due to coding and decoding. In the latter approach, partial computation of the recognizer such as feature extraction can be ported to the client device thus allowing the feature vectors to be transmitted through the network (Fig 1). The issues involved with the transmission of speech over a data network are dealt in Section 4.

4. Speaker Recognition over data network

The same baseline system mentioned in Section 2 is used for the experiments to be listed.

4.1. Experimental Results

Case 1: The most popular approach towards getting over the transmission issues is using a speech codec to transmit the coded waveform. The present day codecs have come up with extremely low data rates and are well suited for VoIP based applications wherein reservations of the resources such as bandwidth are made. As an attempt to verify the recognition performance the recognizer is trained and tested with the G.729A codec data. Though the bitrates seem to be very low, the results in Table 1 clearly reflect the effect of distortion due to coding and decoding.

Table 1: Recognition Performance for different modes of transmission

<table>
<thead>
<tr>
<th>TrainType</th>
<th>TestType</th>
<th>NumSpeakers</th>
<th>Recognition%</th>
</tr>
</thead>
<tbody>
<tr>
<td>NTIMIT</td>
<td>NTIMIT</td>
<td>400</td>
<td>40</td>
</tr>
<tr>
<td>CODEC</td>
<td>CODEC</td>
<td>561</td>
<td>91.29</td>
</tr>
<tr>
<td>TIMIT</td>
<td>TIMIT</td>
<td>561</td>
<td>98.75</td>
</tr>
</tbody>
</table>

Case 2: An interesting approach suggested in literature for DSR kind of tasks [3, 4, 5] is to make the client device extract the feature vectors from the speech signal and then transmit these features to the server for recognition as shown in Fig 2. By adopting this approach we are taking advantage of the requirements of the speaker recognizer by simply transmitting the feature vectors as against the entire speech waveform. Further we have several ways to optimize the data which corresponds to the feature vectors and to address the issues of transmission. Assuming that the feature vectors

---

Note: The image containing the diagram is not transcribed here as it is not essential for understanding the text.
for transmission are acquired from clean speech, the results from column 3 of Table 1 for TIMIT database show very high recognition performance. The current requirements of bandwidth for transmitting the feature vectors as floating point values turns out to be 54.4kbps.

Alternatively, the test feature vectors can be quantized using Vector Quantization (VQ) [7] with an appropriate codebook size and the codebook index is communicated to the server. A codebook of size 32768 is obtained from the training data of the 561 speakers. It is assumed that the client device is initialized with the codebook computed at the server. The recognition performance is still 90.6%. This shows a substantial reduction in bitrate to 1.6kbps compared to that of speech codec for a marginal degradation in performance. Further, since the data is unaffected by channel noise, it may be possible to increase the bitrate and improve performance.

As an effort to reduce the computation at the client end we have attempted to convert the Feature Extraction module to fixed-point arithmetic. The results with 4 decimal places is found to be 98.75% which seems to be the same as that for single-precision floats (Table 1). The conversion of various stages of feature extraction module to fixed-point arithmetic is illustrated in Fig 3. These feature vectors can be transmitted as integers(short) which reduces the bandwidth to 27.2kbps (Table 2). The performance is found to be 98.75% which shows no degradation. The feature vector values are multiplied by 10 to restore up to 1 decimal place so as to make these values fit into a short integer. This clearly shows a significant reduction in bandwidth against transmitting the entire sampled waveform as data which is 256kbps for 16kHz sampled data with 16bits/sample (Table 2). Further an attempt is made to explore the loss tolerance towards prioritizing the frames by dropping the frames based on an energy threshold. The recognition performance for various energy thresholds are listed in Table 3. The numbers indicate a fraction of the total number of frames transmitted. The frames are sorted based on the energy of the frame. Different percentage of frames are dropped from the top and the bottom (Table 3). It is observed that the increase in the error for a 50% drop rate is only 2% as seen from the last row of Table 3. If 50% of the frames are ignored the bitrate required is only 13.6kbps.

5. Discussion

5.1. Results

The different modes of transmission between the client device and the recognizer are speech sent over a voice network or through a data network as VoIP or as feature vectors (Fig 1). Table 1 gives the results of the system for each of these three scenarios. The results on NTIMIT database from Table 1 show that the speech signal is corrupted by channel noise and hence the results are extremely poor. The transmission of the coded signal shows considerable reduction in bandwidth but the recognition performance is found to be 91% which shows reduction in performance by 7% to that of clean speech which shows 98% performance. The degradation in performance is due to the distortion due to coding and decoding. The results from thresholds on energy from Table 3 show better recognition performance as compared to that of NTIMIT and codec results from Table 1 and claim considerable recognition even when 50%(0.5) of the transmitted frames are dropped at the server over the network. Also the issues arising from channel and coding distortions could be avoided when transmitting the feature vectors.

5.2. Issues in protocol

The transmission of feature vectors allows a relaxation on the constraints to the protocols unlike those that cater to the requirements of VoIP based applications. The transmission
of feature vectors enables them to be treated at par with other traffic with no special attention. The QoS requirements are clearly asymmetric in that the uplink (client → server) is not delay sensitive but is sensitive to accuracy. Further the link can afford some packet loss. The sequence information is not very important. The network can afford to lose most or all the low priority packets at times and the system can carry out the recognition process even under adverse conditions as shown in Table 3. Hence we may use the services of a Partially Order/Partially Reliable (PO/PR) protocol on the forward link [12].

On the other hand on the downlink (server → client) which involves synthesized speech playback, we need the behaviour of real time protocols such as RTP to provide timely and orderly delivery of data [13] suggesting an environment as shown in Fig 4.

6. Conclusion

It has been shown that the transmission of feature vectors for speaker recognition yields better recognition results over the traditional VoIP based techniques. Further the need for a strict and completely reliable protocol and the need to reserve the network resources is eliminated. Attempts such as conversion of floating point feature vectors to integers have been made to reduce the bitrate requirements comparable to that of the VoIP based applications but still ensure better recognition performance. We are currently addressing the issues of identifying a)frames that are of high priority for speaker recognition other than those based on energy alone b)interframe similarities to further reduce the bitrate. Further we are trying to tune the VQ codebook size [14] and to explore new features [10] towards improving the speaker recognition performance.

7. Acknowledgements

Our sincere thanks to Mr. T. Nagarajan and Mr. Rajesh M. Hegde for their constructive criticism which enabled us to improve the quality of the paper.

8. References


