Speaker Information using Subsegmental and Segmental Analysis of LP Residual

Debadatta Pati and S. R. Mahadeva Prasanna
Department of Electronics and Communication Engineering
Indian Institute of Technology Guwahati, Guwahati-781039, India
Email: {debadatta, prasanna}@iitg.ernet.in

Abstract—Linear Prediction (LP) residual mostly contains the excitation source information. This work analyzes the LP residual once using frame size of 5 ms (subsegmental) and another time using frame size of 20 ms (segmental), each with a shift of 2.5 ms. The residual frames are then subjected to non-parametric Vector Quantization (VQ) to store the unique excitation sequences for each speaker. The testing of such codebooks seem to contain significant speaker information. Further the speaker information at the two levels are found to be different in nature. The performance of the speaker recognition system using subsegmental, segmental and combined subsegmental-segmental speaker information is found to be 71.67%, 55% and 83.33%, respectively, for a population of 30 speakers taken from TIMIT database using a codebook of size 128. Further combination of the proposed subsegmental-segmental LP residual based system with the vocal tract system feature based system providing stand alone performance of 95% is found to be 100%. This aspect reinforces the different speaker information available in the excitation component of speech.

Index Terms: speaker specific information, excitation source, LP residual, subsegmental, segmental

I. INTRODUCTION

Speaker recognition is the task of recognizing speakers from their speech signal [1]. The speaker information in the speech signal may be attributed to the physiological and behavioral aspects of the speaker. The physiological part is due to the vocal tract and the excitation source involved in the production of speech. The behavioral part is due to the habitual mode of the speaker for speech production. Among all these the mostly used one is the speaker information due to the vocal tract system [4]. The reasons for the same may be the availability of well developed spectral processing tools to extract it and also it provides reliable and good performance due to less intra-speaker variability. The reasons for the relatively less usage of speaker information from the excitation source and behavioral aspects may be due to the non-availability of suitable signal processing tools and also the large intra-speaker variability involved in them [2]. In case of human speaker recognition we may agree that the evidences from the excitation source and behavioral aspects are extensively used for recognizing speakers. Further, the speaker information from these aspects seem to be robust for external degradations like channel and environmental noises. Finally, since the speaker information from these represent independent aspect of speech production, they may combine well with the speaker information from the vocal tract component. Hence the scope for exploring methods for extracting speaker information from the excitation source and behavioral aspects of the speaker. This work focusses on developing methods for extracting speaker information from the excitation source component of speech.

There are several attempts that have been made for extracting speaker information from the excitation source [2], [3], [5], [6], [8]. These may be broadly grouped into two categories namely, methods which use speaker information from the excitation source as joint evidence along with the vocal tract features [3], [5], [6] and methods which perform independent modeling of speaker information from the excitation source [2], [8]. The joint modeling infers that speaker information from the excitation source added with the vocal tract provides improved performance. However, it provides no clue about the potential of the speaker information from excitation alone. Alternatively, the independent evaluation reveals that excitation source also contains significant speaker information [8]. Even though its performance is relatively less than those achieved using the vocal tract, they combine well to provide significantly improved combined performance [4]. These studies only demonstrated the potential of speaker information from the excitation source. Further explorations are needed in this direction for improving the performance of speaker recognition system using excitation source information. This may ultimately result in the development of a speaker recognition system using excitation source information whose performance is in par with that of the vocal tract feature based system.

The first step in extracting speaker information from the excitation source is to separate the excitation source component from the speech. This can be conveniently achieved using the Linear Prediction (LP) analysis [7]. For the proper choice of LP order, the LP Coefficients (LPCs) contains vocal tract and LP residual contains excitation information [2]. Analysis of signal using frame size less than 5 ms is termed as subsegmental analysis and frame size in the range 10-30 ms is termed as segmental analysis [2]. In the earlier attempts, the LP residual was blocked into frames of 5 ms with every sample shift and Auto-Associative Neural Network (AANN) models were trained [8]. It was experimentally demonstrated that the AANN models indeed capture speaker information from the excitation source which was attributed to the unique excitation sequences that may be present in each speaker [2]. Similar analysis may be carried out by looking for the speaker information at the segmental level. It has been shown that, humans can recognize people by listening to LP...
residual [9]. This may be attributed to the speaker specific excitation source information present at different levels. This work views the speaker specific excitation information at three levels namely, subsegmental (1-5ms), segmental (10-30ms) and suprasegmental (100-300ms) levels. The presence of speaker information at the subsegmental level has already been demonstrated [2]. This information may be attributed to glottal pulse characteristics in the samples between two pitch epochs. The presence of speaker information at the segmental and suprasegmental levels can be established by generating signals that retains specific features at this levels. For instance, the speaker specific information at the segmental and suprasegmental levels can be perceived in the signals which has amplitude of appropriate strength at each pitch epoch in the voiced region and at random instances in the unvoiced region [10]. In such a signal, the segmental level may be attributed to the pitch and energy values. Further the change in pitch and energy values may be attributed to the suprasegmental level.

The objective is to model the dominating speaker information at the subsegmental and segmental levels separately and analyze their potential in preserving the speaker information. At the subsegmental level the sequence information present in less than or equal to a pitch period will be modeled. Alternatively, at the segmental level the sequence information present at two or more pitch periods will be modeled. Hence the motivation for the present work.

Rest of the paper is organized as follows: Modeling speaker information from the subsegmental and segmental analysis of LP residual are described in Section II and Section III, respectively. The combined modeling of subsegmental and segmental levels of speaker information are described in Section IV. The comparative study with respect to the speaker information from the vocal tract is made in Section V. Summary of the present work and the scope for future work are listed in Section VI.

II. SPEAKER INFORMATION BY SUBSEGMENTAL ANALYSIS OF LP RESIDUAL

A. Database

Thirty speakers are randomly chosen from the set of 630 speakers present in the TIMIT database. The speech signals of all these speakers are resampled from the original 16 kHz to 8 kHz sampling rate. Each speaker has 8 sentences and among them first six are used for training and the last two are used for testing. The same database is used in all the studies.

B. Subsegmental Analysis

10th order LP analysis is performed using a frame size of 20 ms with shift of 10 ms [5]. For speech signals sampled at 8 kHz, an LP order in the range 8-16 is found to be most suitable for extracting the LP residual [2]. For every frame of 20 ms, 10 LPCs are computed. These LPCs are used in the inverse filter formulation framework to extract the LP residual. The LP residual is blocked into frames of size 5 ms considered in shifts of 2.5 ms for extracting the speaker information. The subsegmental frames of 4 different speakers (2 male and 2 female) are shown in Fig. 1. The waveform nature and hence excitation sequences are different in each case inferring that they may have some unique information about the speaker. For each speaker on an average there are about 5500 subsegmental frames.

![Fig. 1: Subsegmental frames of 4 different speakers (2 male and 2 female).](image-url)

C. Non-parametric VQ of Subsegmental Frames of LP Residual

VQ is a data compression technique used for quantizing the signals expressed in terms of vectors [11], [12]. It can be either parametric or non-parametric depending on whether the vector is the parameter vector extracted from the signal using feature extraction method or normalized signal values itself. In the present case since the LP residual is more noise-like sequence, it is difficult to obtain parametric vectors and even in obtaining so using conventional spectral processing tools may result in the loss of speaker information. Hence the normalized LP residual sequences themselves are used for VQ.

The subsegmental frames of the LP residual from the training speech data are subjected to VQ using LBG algorithm [11], [12]. Codebooks of size 16, 32, 64, 128, 256 and 512 are built. The maximum codebook size of 512 is chosen on the thumb rule of every codebook vector should represent about 10 unique excitation sequences. These codebooks are then tested using the subsegmental frames of the LP residual from the testing speech data. The performance of the speaker recognition system for different codebook sizes is given in Table 1. The performance is poor for smaller codebook sizes indicating that the codebook does not store enough unique excitation sequences for each speaker. Alternatively, the performance improves significantly for larger codebook sizes. The performance of 75% for 512 codebook size indicates the significant amount of speaker information present at the subsegmental level of the LP residual.
TABLE I: Speaker recognition performance (%) for a set of 30 speakers using subsegmental frames of LP residual.

<table>
<thead>
<tr>
<th>Frame size/shift</th>
<th>Codebook size</th>
</tr>
</thead>
<tbody>
<tr>
<td>40/20</td>
<td>16 32 64 128 256 512</td>
</tr>
<tr>
<td>40/20</td>
<td>33.33 41.67 51.67 71.67 73.33 75</td>
</tr>
</tbody>
</table>

III. SPEAKER INFORMATION BY SEGMENTAL ANALYSIS OF LP RESIDUAL

A. Segmental Analysis

The LP residual is extracted using the 10th order LP analysis, frame size of 20 ms with a shift of 10 ms. The LP residual is then decimated by a factor of 4 so that the sampling rate of the LP residual becomes 2 kHz. In the decimated LP residual we have excitation source information up to 1 kHz. The reason for the decimation operation is to make the length of the vector considered in the segmental analysis same as that of subsegmental analysis. That is, in case of segmental analysis the frame size is considered to be 20 ms. For 20 ms at 8 kHz we will have 160 samples per vector, which will be a very large dimension vector for VQ. Hence decimation by 4 makes the vector length equal to 40 which is same as that of the subsegmental vector length. Another reason for the decimation is also that, since the information present as fine variations represented by high frequency components have already been modeled by subsegmental analysis. Hence only other information which can be observed at the segmental level needs to be emphasized. The main speaker information at the segmental level are the pitch and energy, which can be preserved even after decimation by 4. Thus in segmental analysis the decimated by four LP residual is blocked in frames of 20 ms (40 samples at 2 kHz) and 2.5 ms (5 samples) shift. Since the shift is 2.5 ms, in this case also on an average we have about same number of frames that is, 5500 frames for training. The segmental frames for 4 different speakers (2 male and 2 female) are shown in Fig. 2. The waveform nature and hence excitation sequences at this level are also different in each case which may be exploited for extracting the speaker information.

B. Non-parametric VQ of Segmental Analysis of LP Residual

The segmental frames of the decimated LP residual are subjected to non-parametric VQ using LBG algorithm. Codebooks of size 16, 32, 64, 128, 256 and 512 are built. The segmental frames of decimated LP residual from the testing speech data are compared with the codebooks and the performance is evaluated. The performance of the speaker recognition system using segmental speaker information from the LP residual is given in Table 2. The significantly better performance for lower codebook sizes as compared to subsegmental case infers that there are not much variations among the unique excitation sequences in each speaker. This is also true because by decimating in effect we have smoothed the excitation sequences and accordingly they match well with many excitation sequences. However, the performance does not improve significantly as the codebook size increases as in the case of subsegmental analysis. This is because there may not be significantly different number of unique excitation sequences. This is due to the decimation process employed. However, since the nature of excitation sequences are different at subsegmental and segmental levels, it may be possible to combine the two systems.

TABLE II: Speaker recognition performance (%) for a set of 30 speakers using Segmental frames of LP residual.

<table>
<thead>
<tr>
<th>Frame size/shift</th>
<th>Codebook size</th>
</tr>
</thead>
<tbody>
<tr>
<td>40/5</td>
<td>16 32 64 128 256 512</td>
</tr>
<tr>
<td>40/5</td>
<td>46.67 61.67 68.33 55 66.67 63.33</td>
</tr>
</tbody>
</table>

IV. COMBINED SUBSEGMENTAL-SEGMENTAL SPEAKER INFORMATION FROM LP RESIDUAL

Comparing Figs. 1 and 2 it can be observed that the excitation sequences represented in the two cases are different. Therefore the non-parametric VQ of these sequences may also capture speaker information which is different in each case. Hence the evidences from them can be combined to see whether it provides improved performance or not.

For each test signal the average minimum distance from each speaker codebook are normalized with respect to their maximum value. The evidence from each system are then combined using one the following combination schemes:

1) COMB1: For given test speech signal the normalized distance values are added and speaker of the codebook with combined minimum average distance is identified as the speaker. Mathematically

\[ E_c = E_{ss} + E_{sg} \] (1)

2) COMB2: For given test speech signal suppose \( E_{ss} \) and \( P_{ss} \) are the normalized distance and performance
of the subsegmental system and \(E_{ss}\) and \(E_{sg}\) are the normalized distance and performance of the segmental system, then the combined distance is given by

\[
E_c = \left(\frac{P_{ss}}{P_{ss} + P_{sg}}\right)E_{ss} + \left(\frac{P_{sg}}{P_{ss} + P_{sg}}\right)E_{sg}
\]  

(2)

3) COMB3: For given test speech signal, the COMB2 scheme is further modified as

\[
E_c = \left(\frac{P_{ss}}{P_{ss} + P_{sg}}\right)E_{ss} + \left(\frac{P_{sg}}{P_{ss} + P_{sg}}\right)E_{sg}
\]  

(3)

The performance of these combined systems are tabulated in Table 3.

### Table III: Speaker recognition performance (%) for 30 speakers using combined subsegmental-segmental information from LP residual.

<table>
<thead>
<tr>
<th>Combination scheme</th>
<th>Codebook size</th>
</tr>
</thead>
<tbody>
<tr>
<td>COMB1</td>
<td>16 32 64 128 256 512</td>
</tr>
<tr>
<td>COMB2</td>
<td>60 70 71.67 81.67 76.67 76.67</td>
</tr>
<tr>
<td>COMB3</td>
<td>63.33 71.67 76.67 83.33 78.33 78.33</td>
</tr>
</tbody>
</table>

In all the combination schemes, the combined system provides improved performance indicating different speaker information present in each case. The improvement is maximum in the COMB3 scheme which emphasizes the need for using the knowledge about the individual system performance.

### V. SPEAKER INFORMATION FROM VOCAL TRACT SYSTEM

For comparison purpose, the conventional speaker recognition system using speaker information from the vocal tract is also developed. Speech signal is processed in blocks of 20 ms with a shift of 10 ms. For every frame of 20 ms, 13 dimensional Mel Frequency Cepstral Coefficients (MFCCs) are computed [13]. The parametric VQ using LBG algorithm is built for different codebook sizes 16, 32, 64, 128 and 256. Since there about 2000 feature vectors, the maximum codebook size is chosen to be 256. The performance of this speaker recognition system is given in Table 4.

### Table IV: Speaker recognition performance (%) for a set of first 30 speakers for combination and Conventional System

<table>
<thead>
<tr>
<th>Frame size</th>
<th>Codebook size</th>
</tr>
</thead>
<tbody>
<tr>
<td>shift</td>
<td>16 32 64 128 256</td>
</tr>
<tr>
<td>160/80</td>
<td>86.67 90 95 95 95</td>
</tr>
</tbody>
</table>

The performance is significantly better compared to those achieved using the excitation source information. This is expected since MFCC provides a compact representation of speaker information present in the vocal tract. Further the intra-speaker variability is less in the vocal tract based features. However, what will be more interesting is to check whether the proposed subsegmental-segmental LP residual based speaker recognition system combines well with the vocal tract based system to provide combined improved performance or not. The performance of the combination system using subsegmental-segmental LP residual based system and MFCC based system using COMB3 scheme is given in Table 5. The proposed subsegmental-segmental LP residual speaker recognition system further improves the performance of the vocal tract based speaker recognition system. This aspect reinforces the different speaker evidence available in the excitation source.

### Table V: Speaker recognition performance (%) for a set of first 30 speakers for combination and Conventional System

<table>
<thead>
<tr>
<th>Frame size/shift</th>
<th>Codebook size</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>16 32 64 128 256</td>
</tr>
<tr>
<td>normalized</td>
<td>86.67 95 98.33 100 100</td>
</tr>
</tbody>
</table>

### VI. SUMMARY AND CONCLUSIONS

The objective of this work is to develop a method for extracting speaker information present at the subsegmental-segmental levels of the LP residual. This is demonstrated using the non-parametric VQ scheme. Further it is demonstrated that speaker information present at these two levels are different. Finally it is demonstrated that the combined subsegmental-segmental based system contains speaker information which is different to that of the vocal tract based system.

The present work is an improvement compared to earlier attempts for extracting the excitation source based speaker information in the aspect of exploring non-parametric VQ for modeling and also capturing speaker information at the segmental level and demonstrating its different nature compared to the subsegmental level. Further studies are required to improve the performance at each of these levels and also methods may be explored to extract the speaker information present at the suprasegmental level of the LP residual. This work is currently underway. In the earlier work using AANN models for speaker recognition using LP residual, the speaker information is learnt by the model. Alternatively, the present work quantizes the temporal sequences to store typical sequences for comparison. Thus the working principle of the two modelling techniques are different and hence it may be possible to combine the evidences from them.

### VII. ACKNOWLEDGEMENTS

This work is supported by the ongoing UK-India Education and Research Initiative (UKIERI) project titled “Study of Source Features for Speech Synthesis and Speaker Recognition” between CSTR, University of Edinburgh, UK, Indian Institute of Technology Guwahati (IITG), India and International Institute of Information Technology Hyderabad (IIITH), India.

### REFERENCES


