CHAPTER-3
EXPLORATION OF COMPLETE SOURCE FEATURES
FOR SPEAKER RECOGNITION

In the previous chapter, spectral and temporal based methods for capturing the speaker variability in terms of distribution of the speaker’s feature vectors have been reviewed. In this chapter, a new method for capturing speaker variability in terms of speaker-specific information, using source characteristics of speakers is introduced.

Section 3.1 is an introduction to understand the concept of source and system. To understand the concept of source and system, Section 3.2 describes the speech production mechanism. Speech production mechanism described in terms of source and system characteristics is explained in Section 3.3. Section 3.4 addresses the speaker recognition task using source features. In Section 3.5, the concept of residual is discussed. Section 3.6 reviews the work done on utilization of the LP residual and processing of LP residual in time-domain at subsegmental, segmental and suprasegmental levels for speaker recognition task. The proposed approach for text independent speaker recognition is discussed towards the end of the chapter in Section 3.6.1.

3.1 INTRODUCTION

A signal is defined as any physical quantity that varies with time, space or any other independent variables [81 and 82]. Speech sound is also a kind of signal for which the functional relationship between the signal and the independent variable is highly
complicated [81]. To a high degree of accuracy, a segment of speech may be represented as a sum of several sinusoids of different amplitudes and frequencies.

Whenever any kind of physical work has to be done, some supply of energy is required and the work actually consists in converting this energy from one form to another. The generation of sounds is also a physical work and hence is governed very much by the same laws as any other kind of phenomenon to be found in the universe [83]. They exemplify the effects of forces acting upon physical bodies to produce movements of various kinds and hence depend on a suitable source of energy. We can think of the energy supply as being the driving force and the system to which it is applied as the driven system. This driving force sets the systems into vibrations, which in turn produces sound. The amplitudes of these vibrations depend on how tightly the two elements are coupled together.

To have a better understanding of how this principle applies to speech signals, it is necessary to have knowledge about the speech production mechanism, which is discussed in the next section.

3.2 SPEECH PRODUCTION MECHANISM

Speech signal is produced as a result of time varying excitation of the time varying vocal tract system [30]. A schematic diagram of speech production mechanism is shown in the Fig. 3.1
Speech production mechanism consists essentially of a vibrating source of sound fundamentally coupled to a resonating system.

For a great deal of the time in speech, the larynx is the source and the air column from larynx to the lips, which is the vocal tract, is the system. But to produce some special sounds called nasal sounds, the vocal tract is replaced by nasal tract as the system. The nasal tract begins at the velum and ends at nostrils. When the velum is lowered, the nasal tract is acoustically coupled to the vocal tract to produce the nasal sounds of speech. But it is a known fact that no sound can be produced without a supply of force or energy. It is the breathing mechanism, constituting of the lungs and muscles of the chest and abdomen that constitute the

**Fig. 3.1: Speech Production Mechanism.**
energy supply. A simplified representation of the complete physiological mechanism for creating speech is as shown in Fig.3.2.

By the use of laryngeal muscles the vocal chords can be brought together so as to form as it were a shelf across the air way which leads from the lungs through trachea to pharynx and the mouth. There is a steady flow of air from the lungs into the trachea. While the edges of chords are held together, pressure on the underside of the shelf rises. When it reaches a certain level, it is sufficient to overcome the resistance offered by the obstruction and so the vocal chords open approximately, as shown in Fig. 3.3. The ligaments and muscle fibers that make up these structures have a degree of elasticity and have been forced out of position. They tend to return as rapidly as possible to their disposition. The pressure rises again and the cycle of opening and closing is repeated.

Fig.3.2: Representation of Speech Production Mechanism.
Fig. 3.3: Diagram of Vocal fold Motion at Different Time Instants.

The studies on musical wind instruments showed that the dimensions of the air column involved were all important in determining the frequency at which resonance would occur. The same principle applies equally in the case of the vocal tract also. Speech is produced as a sequence of sounds. Hence the state of the vocal chords and shape and size of various articulators change over time to reflect the sound being produced. To produce a particular sound, the articulators have to be positioned in a particular way. But when different speakers try to produce same sound, though their vocal tracts are positioned in a similar manner the actual shapes will be different due to differences in the anatomical structure of the vocal tract. The main objective of all the above studies is to effectively capture the speaker variability due to the anatomical structure of the vocal tract which will help in automatic speaker recognition.
The speech production mechanism and the concept of source and system to the speech signals is discussed in the next section.

3.3 CONCEPT OF SOURCE AND SYSTEM FOR SPEECH PRODUCTION MECHANISM

Speech signals, as any other real world signals, are produced by exciting a system with a source [29]. A simple block diagram representation of the speech production mechanism is shown in the Fig.3.4. A vibration of the vocal folds, powered by air coming from the lungs during exhalation, is the sound source for speech. It sets up a pulse wave in which the pulses are roughly triangular and of which the amplitude, fundamental frequency, and the waveform can be modified by the action of the laryngeal muscles. The sound generated in the larynx does not transmit linguistic information. It acts as the source for the information which is imposed upon it by modifications introduced by the vocal tract. Hence, as can be seen from the Fig.3.4, the glottal excitation forms the source, and the vocal tract forms the system.

![Source and System Representation of Speech Production Mechanism](image_url)

**Fig.3.4:** Source and System Representation of Speech Production Mechanism.
Speech is produced by exciting the vocal tract by the glottal excitation. From signal processing point of view, this block diagram can be replaced with excitation and filter representation as shown in the Fig.3.5. Here, the vocal tract is replaced with filter, and the filter coefficients depend on the physical dimensions of the vocal tract. Glottal excitation is replaced with two types of signal generators, impulse train generator for voiced sounds and random number generator for unvoiced and fricative sounds.

If we denote the Fourier Transform of the source as $U(f)$ and if we consider the vocal tract as a time-invariant linear system, represented as $H(f)$, however, is usually characterized by several peaks corresponding to resonances of the acoustic cavities that form the vocal tract. The spectral envelopes of the source and system for a sound unit /aa/ are as shown in Figs.3.6.
Fig. 3.6: Spectral Envelopes of Source and System for a Sound unit /aa/ a) Speech Signal, b) Spectral Envelope of System and c) Spectral Envelope of Source.

In brief, it is convenient to consider human speech production to be the result of the generation of one or more sources of sound and filtering of these sources by the vocal tract [76]. To a large extent, the mechanism of source generation is independent of the filtering process, i.e., the properties of source tend not to be strongly influenced by the acoustic properties of the filters. Thus, it is appropriate to discuss source mechanism separately from the behavior of the vocal tract in response to the sources. To an approximation, source properties can be considered to be independent of the shape of the vocal tract mechanism. In other
words, speech signal consists of two kinds of information, source information and system information. Having known this information about the speech signal, in the next section we shall discuss the usefulness of either of the kinds of information for speaker recognition task.

3.4 SOURCE CHARACTERISTICS FOR SPEAKER RECOGNITION

Speaker recognition systems have been developed mostly using spectral features for capturing speaker-specific information. Some of the spectral features used for speaker recognition task are short-time spectrum [84 and 85] predictor coefficients [29, 30, 86 and 87], cepstral coefficients [32, 88 and 89], formant frequencies and bandwidths [90 and 91], etc. In spectral analysis, the signal is considered as a band of sinusoidal signals. The sum of the responses of the system to these sinusoidal signals is computed at the output. Hence, spectral analysis yields information about the system.

Most of the present day speaker recognition systems capture the speaker variability in terms of system characteristics. But we have seen in the previous section that excitation characteristics are speaker specific and hence the source information can also be used to capture the speaker-specific information present in the given speech signal.

It is interesting to note that human beings recognize people mostly from the source characteristics such as glottal vibrations, and prosodic features such as intonation and duration, which are unique to a speaker. Moreover, human listeners have proved themselves as robust speaker recognizers when presented with degraded speech and session variability [92]. This shows that
source feature might be more robust to degradations and session variability. It is this fact which has motivated us to address the speaker recognition task using complete source characteristics at sub-segmental, segmental and supra-segmental levels.

The source features are very significant to find fact the speaker variability. The significance of the source feature is illustrated in the Fig. 3.7. The speech utterances sampled at 8 kHz were collected from two male speakers over a microphone. All the speakers uttered the sound unit /aa/. The significant instants of the glottal excitation are computed for the two speakers. The instants corresponding to the steady section of the utterances were displayed in the Fig. 3.7. It is clearly seen from the figure that the periodicity of the instants of glottal excitation for each of the two speakers is different from that of the others.

Thought it is a known fact this high level source information is certainly valuable in terms of speaker-specific information, not much effort has gone in developing speaker recognition system using this information due to practical difficulties involved in using this information. In the present work we have derived a speaker-specific model using predominantly the source characteristics of the speech signal of LP residual at multi levels, and used this model for speaker recognition. GMM and HMM models are proposed to capture the speaker specific characteristics, and LP residual signal is used as a representation of the source characteristics. The procedure for extraction of LP residual for speaker recognition is explained in the next section.
Fig. 3.7: Instants of Significant Excitations for Two Male Speakers.

3.5 LINEAR PREDICTION RESIDUAL

One of the most powerful speech analysis techniques is the method of linear predictive analysis [87]. The philosophy of linear prediction is intimately related to the basic speech production model. LPC analysis approach performs spectral analysis on short segments of speech with an all-pole modeling constraint [87]. Since speech can be modeled as the output of a linear, time-varying system excited by a source, LPC analysis captures the vocal tract system information in terms of coefficients of the filter representing the vocal tract mechanism. Hence, analysis of speech signal by LP results in two components, namely the synthesis filter on one hand
and the residual on the other hand. In brief, the LP residual signal is generated as a by-product of the LPC analysis, and the computation of the residual signal is given below.

In the previous sections we have seen the source and system representation of speech production mechanism. The discrete speech production representation of the same is as shown in the Fig.3.8.

![Filter and Inverse Filter Representation of the Speech Production Mechanism.](image)

If the input signal is represented by \( u(n) \) and the output signal by \( s(n) \), then the transfer function of the system can be expressed as,

\[
H(z) = \frac{S(z)}{U(z)}
\]  (3.1)

Where \( S(z) \) and \( U(z) \) are z-transforms of \( s(n) \) and \( u(n) \) respectively.

Consider the case where we have the output signal and the system and have to compute the input signal. The above equation can be expressed as
Where $A(z) = 1/H(z)$ is the inverse filter representation of the vocal tract system.

LP models the output $s(n)$ as the linear function of past outputs and present and past inputs. Assuming an all-pole model for the vocal tract, the signal $s(n)$ can be expressed as a linear combination of past values and some input $w(n)$ as shown below.

\[ s(n) = -\sum_{k=1}^{p} a_k s(n-k) + Gw(n) \]  \hspace{1cm} (3.6)

Where $G$ is a gain factor.

Now assuming that the input $w(n)$ is unknown, the signal $s(n)$ can be predicted only approximately from a linear weighted sum of past samples. Let this approximation of $s(n)$ be $s^\wedge(n)$, where

\[ s^\wedge(n) = -\sum_{k=1}^{p} a_k s(n-k) \]  \hspace{1cm} (3.7)

Then the error between the actual value $s(n)$ and the predicted value $s^\wedge(n)$ is given by

\[ e(n) = s(n) - s^\wedge(n) = Gw(n) \]  \hspace{1cm} (3.8)

This error $e_n$ is nothing but the LP residual of the signal. For a detailed description of LP analysis, see Appendix A.
This LP residual, which is generated by LP analysis, is usually ignored in all the major applications of speech analysis like speaker recognition. Only LPC coefficients are used to compute the feature vectors. But the residual signal is rich with source characteristics, which are also speaker-specific. Hence, the information present in the residual signal can be used for speaker recognition task. In the next section we shall review the work done in the direction of using the LP residual signal for speaker recognition task.

3.6 WORK DONE ON LP PRESIDUAL FOR SPEAKER RECOGNITION

Though it is a known fact that residual contains information regarding the complete source, which is speaker-specific, not much work has been done in utilizing this information for speaker recognition task [2].

A few attempts that have been made for extracting speaker information from the excitation source [76, 93, 94 and 95]. 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 [93 and 94] and the method which perform independent modeling of speaker information from the excitation source [76 and 95]. The joint modeling infers that speaker information from the excitation source added with the vocal tract features provides improved performances. However, it provides no clue about the potential of the speaker information from the excitation alone. Alternatively, the independent evaluation reveals that excitation source also contains significant speaker information [95]. Even though its performance is relatively less than those achieved using the vocal tract, they combine well to provide significantly improved combined
performance [96]. These studies only demonstrated the potential of speaker information from excitation source. Further explorations are needed in this direction for improving the performances of speaker recognition system using complete excitation source information. Whose performance is better than that of the vocal tract features based system. Thus to improve the performance of source features, methods need to developed that tries to capture the complete source information. Source information contained in the LP residual of the speech signal [76]. The LP residual can be processed in time, frequency, cepstral or time–frequency domains to extract and model information [76]. Processing of LP residual in time-domain has the advantage that of the artifacts of digital signal processing like block processing or windowing effect that creeps in other domains of processing will be negligible. Thus processing LP residual in time-domain is expected to model the speaker information in the best possible manner. Much work is not done on processing of LP residual at different levels. A unified frame work may be evolved where a given LP residual is processed at subsegmental, segmental and suprasegmental levels. Extract features at each level and model them at each level of speaker-specific information using GMM and ergodic HMM.

It is quoted in the literature that information extracted from the residual signal using the above mentioned techniques yielded improvement in the performance when combined with the existing information [94]. From signal processing concepts, it is a known fact that the spectrum based features (like MFCCs), capture the gross level information. But the residual signal has much flatter spectrum (like white noise) representing the source characteristics rather than those of the vocal tract. The spectrum of the signal and
the residual for a speech segment are as shown in the Fig. 3.9. Spectral representation of the residual signal might not yield complete source information present in the residual since the spectrum of the LP residual has crept. The challenge is to extract complete source information effectively. For this, we process LP residual in time-domain directly at subsegmental, segmental and suprasegmental levels. The procedure for extraction of subsegmental, segmental and suprasegmental features from LP residual for speaker recognition are explained in the next subsections.

Fig. 3.9: (a) Short-Time Spectrum of a Speech Signal, (b) Short-Time Spectrum of the Corresponding Residual Signal.
3.6.1. Speaker Information from Subsegmental Processing of LP Residual.

At the subsegmental level, speaker information present mostly within one glottal cycle is modeled. This information may be attributed to the activity like opening and closing glottal characteristics. To model this information, the LP residual is blocked into frames of 5 msec with a shift of 2.5 msec. For 5 msec at 8 kHz, they have 40 samples one such frame is shown in fig 3.10 (b) and its spectrum is shown in fig 3.10(c). The largest amplitude of the samples of the vector indicates the strength of excitation. The samples in the vector represent the sequence information of glottal. Since these frames are obtained from the LP residual sampled at 8 kHz, they will have excitation source present as the fine variations represented by frequency components up to 4 kHz. These frames of LP residual samples in the time-domain are used as the feature vectors to represent speaker information at the sub segmental level and used for developing speaker recognition experiments using GMM will be discussed in chapter 4.

The nature of the LP residual that will be processed at the subsegmental level is one shown in fig 3.10(a). This is nothing but the original LP residual and its subsegmental level is shown in fig 3.10 (b).
Fig. 3.10: Subsegmental of a) LP Residual (b) No. of Samples and c) its Spectrum.

3.6.2 Speaker Information from Segmental Processing of LP Residual

At the segmental level, speaker information present in two to three glottal cycles is modeled. This information may be attributed mostly to pitch and energy. Speaker information represented by variations within a glottal cycle has already been modeled by subsegmental analysis. In segmental level processing of LP residual, other information that can be observed at the segmental level needs to be emphasized. For this we propose to decimate the LP residual by a factor 4 so that the sampling rate becomes 2 kHz and we may have source information up to 1 kHz. The decimated LP residual is shown in Fig 3.11(a). Even after decimation, the dominant speaker information at the segmental level, that is, pitch and energy information still can be preserved. Moreover, in segmental level processing, LP residual frames of 20 msec duration are used as the feature vectors. For 20 msec at 8 kHz, these feature vectors with 160 samples are of very large dimension for building the models. By decimating the LP residual by a factor 4, the dimension of the feature vectors is reduced to 40 samples per vector which is equal to the subsegmental feature vectors length.
Since the LP residual is decimated by a factor 4, we prefer to compute the feature vectors for every 2.5 msec frame shift so that the number of feature vectors will remain same as the subsegmental features. One such feature vector derived from the decimated LP residual is shown in Fig. 3.11(b) and its spectrum is shown in fig 3.11(c). It contains mainly the pitch and energy information. The fine variations within the glottal cycle are suppressed by smoothing. The periodicity and the amplitude of the segmental level clearly represent the pitch and energy information. This observation indicates that segmental feature vectors reflect different aspect of source information compared to subsegmental feature vectors. This will also be confirmed from the comparison study in Sect 3.6.4. The effectiveness of these features is evaluated from the identification experiment and the results have been discussed in chapter 4.

![Segmental of (a) LP Residual (b) No. of Samples and (c) its Spectrum](image)

**Fig. 3.11**: Segmental of (a) LP Residual (b) No. of Samples and (c) its Spectrum
3.6.3 Speaker Information from Suprasegmental Processing of LP Residual

Subsegmental processing models speaker information up to 4 kHz. Segmental processing models speaker information up to 1 kHz. Beyond that LP residual also contains some speaker information at very low frequency range, that is, may be less than 100 Hz. For example, the variation in pitch and energy across several glottal cycles [97 and 112]. In capturing such information, we need to process the LP residual at the suprasegmental level, for example, with frames of 100–300 msec range. For the LP residual sampled at 8 kHz, the feature vectors from such frames will be of very large dimension for building models.

We prefer to decimate the LP residual by a factor 50 so that the sampling rate becomes 160 Hz and we may have the source information up to 80 Hz. The dimension of the feature vector is also reduced by 50 factors. Further, the high frequency information that is already modeled by subsegmental and segmental level processing will be smoothed out. Therefore in suprasegmental level processing of LP residual, we decimate the LP residual by a factor of 50 and process in frames of 250 msec with shift of 6.25 msec. The frame size is decided so that the dimension of the feature vectors will remain same as in subsegmental and segmental processing. However, the minimum possible frame shift in this case is 6.25 msec which corresponds to one sample shift. Fig 3.12(b) shows a suprasegmental feature vector derived from the decimated residual shown in Fig. 3.12(a). The fast varying components of the original LP residual is eliminated and it mostly represents the long term variations. This can also be observed from the spectrum of the shown feature vector from Fig. 3.12(c). Information present in the
smoothed spectrum is up to 80 Hz. The periodicity and other high frequency related information are absent. The speaker information present in these features is verified from the recognition experiments as performed earlier. The results of the identification and are given in next chapter.

Results show that suprasegmental level features contain some speaker information. Further, the recognition performance is significantly poor compared to sub segmental, segmental and vocal tract information. The poor result indicates that the suprasegmental features may have large intra-speaker variability. The other major factor is text independent mode of operation. However, it may contain different aspect of speaker information and hence may combine well with other features.

![Fig.3.12: Suprasegmental of (a) LP Residual (b) No. of Samples and (c) its Spectrum.](image)

3.6.4 Combining evidences from Subsegmental, Segmental and Suprasegmental Levels of LP Residual

By the way of deriving each feature, the information present at sub segmental, segmental and suprasegmental levels are different and hence may reflect different aspect of speaker specific
source information. By comparing their recognition performances it can be observed that the subsegmental features provide best performance. Thus the subsegmental features may have more speaker-specific evidence compared to other level features. The different performances in the recognition experiments indicate the different nature of speaker information present.

The speaker-specific excitation source information contains both the amplitude and the sequence information. In this the effect of amplitude dominates over the sequence information. Hence need to separate both the amplitude and phase information by using the method called analytical representation of LP residual. Since the amplitude and sequence information are two different aspects of speaker-specific information, their combined effect may provide improved performance. We describe an analytical representation of LP residual for extracting speaker-specific complete source characteristics in terms of amplitude and phase information are discussed in the next section.

3.7 SPEAKER INFORMATION USING ANALYTICAL SIGNAL REPRESENTATION OF LP RESIDUAL

In the previous section, Speaker information from the LP residual was derived by direct processing of the LP residual at the subsegmental, Segmental and suprasegmental levels. The dominant speaker information present in these three levels of processing mostly represents the amplitude and sequence information of the source. When the LP residual is processed directly, effects of amplitude values dominate over the sequence, especially around the instants of glottal closure [114]. It may therefore be better to separate the amplitude and sequence information and then process
them independently. One approach to achieve this is with the use of analytic signal representation of the LP residual [115]. In this representation, the magnitude of the analytic signal of LP residual and the cosine of the phase of the analytic signal represents the sequence information. Thus the analytic signal representation of the LP residual may help in exploiting the amplitude and sequence information separately. We propose to derive the subsegmental features from the analytic signal representation of the LP residual.

The analytic signal of the LP residual \( r_a(n) \) corresponding to the LP residual \( r(n) \) is given by [115]

\[
r_a(n) = r(n) + j\hat{r}_a(n)
\]

Where \( \hat{r}_a(n) = \text{IFT} [R_a(w)] \)

Where

\[
R_a(w) = \begin{cases} 
-jR(w) & 0 \leq w \leq \pi \\
-jR(w) & 0 > w \geq -\pi 
\end{cases}
\]  

(3.10)

\( R(w) \) is the Fourier transform of \( r(n) \) and IFT denotes the inverse Fourier transform. The magnitude of the analytic signal, called as the Hilbert envelope (HE) of the LP residual is given by [114]

\[
|a_r(n)| = \sqrt{\hat{r}_a^2(n) + \hat{r}_a^*(n)}
\]

(3.11)

and the cosine of the phase, called as the residual phase (RP) is given by [114]

\[
\cos(\theta(r(n))) = \frac{r(n)}{|a_r(n)|}
\]

(3.12)

The procedure to compute subsegmental, segmental and the suprasegmental features vectors from HE and RP of the LP residual
is same as described earlier except the input sequence. In one case the input will be HE and the other case it will be RP.

3.8 SUMMARY

In this chapter, we introduced the general idea of source and system, and extended this idea to the speech production mechanism. The concept of LP residual is introduced and a review of the work done on utilization of speaker-specific information present in LP residual for speaker recognition task is presented.

We described a method for the extraction of complete speaker-specific excitation source information by the time-domain analysis of the LP residual. Speaker specific information in the LP residual includes those within one glottal cycle. Pitch and energy across two to three cycles, and variation of the pitch and energy across several glottal cycles. In the proposed method, speaker information within one glottal cycle is extracted by the subsegmental processing of the LP residual. Pitch and energy contour information is extracted from the suprasegmental processing of the LP residual. A method is discussed to separate amplitude information and phase information using HT since by direct processing of LP residual dominates amplitude information over phase information. The different nature of both amplitude and phase information increases the performance of speaker recognition compared with the LP residual alone.