International Journal of Research, Science, Technology & Management
ISSN Online: 2455-2240, Volume 20 Issue III, September 2020

Feature Extraction Using Empirical Mode Decomposition of Speech Signal

Naresh Upadhyay

***Abstract*- Speech signal carries information related to not only the message to be conveyed, but also about speaker, language, emotional status of speaker, environment and so on. Speech is produced by exciting the time varying vocal tract system with a time varying excitation. Each sound is produced by a specific combination of excitation and vocal tract dynamics. This paper presents a speaker identification system using empirical mode decomposition (EMD) feature extraction method. The EMD is an adaptive multiresolution decomposition technique that appears to be suitable for non-linear, non-stationary data analysis. The EMD sifts the complex signal of time series without losing its original properties and then obtains some useful intrinsic mode function (IMF) components. The FFT is the most useful method for frequency domain feature extraction. Wavelet transform(WT) is yet another method for feature extraction.**

***Keywords*— Speaker identification, Empirical mode decomposition, Intrinsic Mode Function**

I. INTRODUCTION

Speech is composite signal that contains information about the message to be conveyed, the characteristics of the speaker and the language of communication. The unique characteristics of the voice of a speaker are due to anatomical and physiological factors. Anatomical factors relate to the physical aspects of speech production namely, the vocal tract system and the vocal folds. Physiological factors reflect the speaking habits of a person, such as speaking rate, accent and mannerisms. The features are embedded in the speech signal, and hence are useful in recognizing a speaker.

Speech signal is produced as a result of excitation of the time-varying vocal tract system. A schematic diagram of the speech production mechanism is given in Figure 1.1. From the acoustic point of view, speech mechanism can be viewed as a sound source coupled to a resonant system. The system consists of the region from trachea through vocal cord(glottis), vocal tract up to the front end of the mouth, and from trachea through soft palate up to the nostrils, if velum is open. Speech wave radiated is the result of characteristics of the system imposed on the glottal wave. For each sound there is a positioning for each of the vocal tract articulators: vocal folds (cords), tongue, lips, teeth, velum, and jaw [4]. After a preparatory inhalation of air into lungs, speech is produced as air is exhaled. Changes in articulatory positions influence their sound produced. The pulses of air produced by the abduction and abduction of the folds generate a periodic excitation of the vocal tract. Speech signal carries linguistic, speaker, and environmental information.

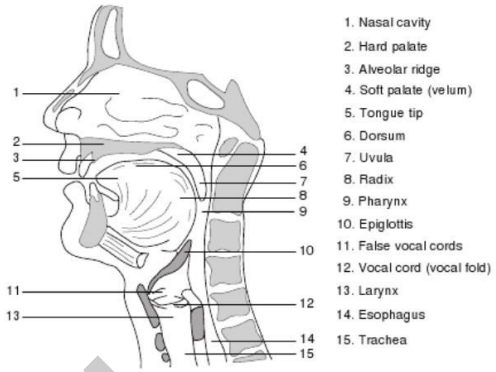

Figure 1. Speech Production System

Variability in the speech signal due to intentional changes in the vocal tract shape by the speaker to produce different sounds correspond to the linguistic part of speech signal. 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.

Feature vectors are in general some parameter vectors extracted from frames of speech signal to capture the characteristics of the speakers. The feature vectors span a feature space of dimensionality equal to the dimension of the vector. A large variety of acoustic parameters (feature vectors) can be extracted from speech signal either directly from the wave form or from the frequency domain representations. Depending on the analysis, the features may be majorly categorized into:

- Low-level features
- High-level features

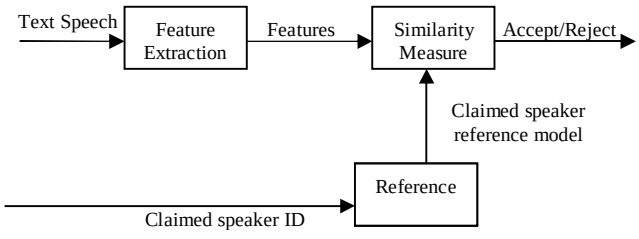

Figure 2. Block diagram of Speaker Verification

A. Low-Level features

Low-level features are mainly features representing the acoustic characteristics of the vocal tract shape and they are further categorized into

- Subsegmental features
- Segmental features

Subsegmental features are features extracted from 1-3ms (less than one pitch period) frames of speech. Generally, frames near the instants of major excitation are of great use here, because the signal-to-noise ratio is generally high in those regions. Segmental features are extracted from 10-30ms (one to three pitch periods) frames of speech.

B. High-Level features

Transitional features, representing how the vocal tract shape changes with time, are derived from more than one frame of speech. Features like pitch, intonation and duration which do not represent the vocal tract shape but considered to be specific to a speaker, come under the category of suprasegmental features which are representative of high-level features. Suprasegmental features generally represent speaker-specific features across frames. Higher-level features can provide useful metadata about a speaker, such as what topic is being discussed, how a speaker is interacting with another talker, whether the speaker is emotional or disfluent.

C. Desirable Characteristics for Features

A set of desirable characteristics for speaker recognition feature vectors are as follows.

1. Efficient in representing the speaker-specific information.
2. Easy to measure.
3. Stable over time.
4. Occurs naturally and frequently in speech.
5. Change little from one speaking environment to another, and
6. Not susceptible to mimicry.

II. FEATURE EXTRACTION

Certain features from .wav file is extracted after finding IMF(Intrinsic Mode Function of each file. Log energy is the feature which is extracted. There are several methods to do the feature extraction.

A. Fast Fourier Transform

In traditional techniques, the sound features are usually obtained by fast Fourier transform (FFT) or short time Fourier transforms (STFT), the FFT is the most useful method for frequency domain feature extraction. Nevertheless, this method loses some information in the time domain while the signals are converted into the frequency domain, therefore, the STFT was developed to overcome the above drawback and evaluate the sinusoidal frequency. Fast Fourier Transform algorithm yields an implementation with a level of parallelism proportional to the radix r of factorization of the discrete Fourier transform, allows 100 percent utilization of the arithmetic unit, and yields properly ordered Fourier coefficients without the need for pre- or postordering of data. The speed of moving data, elimination of addressing, and simplicity of machine organization constitute, in this context, some of their advantages over core-type memories. The algorithms are also well suited for an implementation using integrated circuit random access memories (RAM)[3].

B. Wavelet Packet Adaptive Network based Fuzzy Inference System(WPANFIS)

In contrast to the Fourier transform, the characteristic of wavelet transform (WT) gives short time intervals for the high-frequency bands and long time intervals for the low-frequency band. WPANFIS consists of two layers: wavelet packet and adaptive network based fuzzy inference system. The main advantage of WT is it has an adjustable window size. Unfortunately, Continuous wavelet transform usually generates an immense amount of wavelet coefficients, and therefore will be highly redundant. Thus, the discrete wavelet transform (DWT) was developed to improve on CWT, and it can avoid generating redundant information. The DWT permits the systematic decomposition of a signal into its sub-band levels. It can be performed with minimum distortion of the signal, even for stationary signal analysis. Speech recognition using a wavelet packet adaptive network based fuzzy inference system deals with the combination of feature extraction and classification for real speech signals[2].

C. Empirical Mode Decomposition(EMD)

Empirical Mode Decomposition is one of the method of feature extraction. It is advantageous compared to the other methods. EMD method is able to decompose a complex signal into a series of intrinsic mode functions(IMF) and a residue in accordance with different frequency bands. EMD is self-adaptive because the IMF works as the basis functions determined by the signal itself rather than what is pre-determined. It is highly efficient in non-stationary data analysis[1].

III. PROPOSED METHODOLOGY

Feature extraction is done using empirical mode decomposition. For this there are two procedures.
a) To find Intrinsic Mode Functions(IMF)

b) Sifting Process-Since sifting is a recursive process, a sifting stopping rule is required.

A. Intrinsic Mode Function

A multi-resolution decomposition technique is presented, empirical mode decomposition (EMD), which is adaptive and appears to be suitable for non-linear and non-stationary signal

analysis. It was carried in the time domain to form the basis functions adaptively. The major advantage of EMD basis functions can be directly derived from the signal itself. IMFs can be both amplitude and frequency modulated. The EMD decomposes the original signal into a definable set of adaptive basis of functions called the intrinsic mode functions. Each IMF must satisfy two basic conditions : in the whole data set, the number of extrema and the number of zero crossings must either equal or differ at most by one. Note, either local minima or local maxima are extrema. Moreover, a sample S_i in a time-series is a local maximum if $S_i > S_{i-1}$ and $S_i > S_{i+1}$, and a sample S_i is a local minimum if $S_i < S_{i-1}$ and $S_i < S_{i+1}$, where i is a discrete time; and (2) at any point, the mean value of the envelope, one defined by the local maxima (upper envelope) and the other by the local minima (lower envelope) is zero.

B. Sifting process

The purpose of sifting is to subtract the large-scale features of the signal repeatedly until only the fine-scale features remain. First, the original speech signal, x(t), should be enclosed by the upper and lower envelope in the time domain. Using a cubic spline, the local maxima is connected forming the upper envelope $u_+(t)$ and the local minima is connected forming the lower envelope $l_-(t)$. The two envelopes cover all the data points. The envelope mean m(t) is determined as follows, $m(t) = (u_+(t) + l_-(t))/2$. The first component is described as

$$h_1(t) = x(t) - m(t) \quad (1)$$

The component $h_1(t)$ is now examined to see if it satisfies the conditions to be an IMF. If $h_1(t)$ doesn't satisfy the conditions, $h_1(t)$ is regarded as the original data, the sifting process would repeat, obtaining the mean of the upper and lower envelopes, which is designated as m_{11}; therefore,

$$h_{11}(t) = h_1(t) - m_{11}(t) \quad (2)$$

Then, repeat the procedure until k_{1k} is an IMF, i.e.,

$$h_{1k}(t) = h_{1(k-1)}(t) - m_{1k}(t) \quad (3)$$

After k siftings, we explained the first intrinsic mode to be

$$C_1 = h_{1k} \quad (4)$$

Finally, c_1 revealed the higher frequency component of IMF. To obtain enough physical definitions of IMF, the sifting stop criteria, namely, the stop conditions, are of great importance and are found by the following equation:

$$SD = \sum \frac{|h_{1(k-1)}(t) - h_{1(k)}(t)|^2}{h_{1(k-1)}(t) * h_{1(k-1)}(t)} \quad (5)$$

The typical values of SD are 0.2 and 0.3. To obtain the second and subsequent intrinsic mode functions, the residual signal can be calculated as

$$x(t) - c_1(t) = r_1(t) \quad (6)$$

r_1 considers the original data, and by repeating the above procedures, x(t) could be obtained by the second IMF component c_2. The procedure as described above is repeated for n times, then the n-IMFs of signal x(t) could be obtained

$$\begin{aligned} &r_1(t) - c_2(t) = r_2(t) \\ &. \\ &. \\ &. \\ &. \\ &r_{n-1}(t) - c_n(t) = r_n(t) \end{aligned} \quad (7)$$

The decomposition procedure can be stopped when the residue, r_n, becomes a constant, a monotonic function, or a function containing only a single extrema, from which no more IMF can be extracted. By summing Eqns. and the original signal can be reconstructed as follows:

$$x(t) = \sum_{j=1}^{n} \left(a_j(t) + r_n(t)\right) \quad (8)$$

Energy of each IMF is calculated as follows:

$$energy = \sum_{i=1}^{n} \log\left(S_i^2\right)$$

IV. RESULTS AND DISCUSSIONS

Speech samples are extracted from CHAINS corpus[5]. It has recordings of 36 speakers obtained in two different sessions with a time separation of about two months. The first recording session was in a sound proof booth while second one is in a quite office environment. The Duration, IMF's and Log Energy of various speech signals has been calculated and is represented in a table(Table I).

V. CONCLUSION AND FUTURE WORK

The proposed system is an effective recognition method for speaker identification. The EMD can effectively sift the riding wave from every complex signal of the time series. The sifted IMF components represented important information in the entire signal set. The energy correlates closely with every IMF component. The proposed procedure of feature extraction reduced the dimension of feature more effectively.

Empirical Mode Decomposition method of feature extraction reduced the dimension of feature more effectively. Further improvement can be done using these features to give

as inputs to BPNN, GRNN and AANN and the performance of the three neural networks can be compared.

File	Duration	No. of IMF's	Log energy of each IMF
irf01_s10_solo	4s	12	
		IMF1	-4.36E+05
		IMF2	-4.40E+05
		IMF3	-4.48E+05
		IMF4	-4.51E+05
		IMF5	-4.61E+05
		IMF6	-4.72E+05
		IMF7	-4.60E+05
		IMF8	-4.31E+05
		IMF9	-4.36E+05
		IMF10	-4.55E+05
		IMF11	-3.71E+05
		IMF12	-3.99E+05
irf01_s11_solo	3s	12	
		IMF1	-4.68E+05
		IMF2	-4.37E+05
		IMF3	-3.88E+05
		IMF4	-3.59E+05
		IMF5	-3.58E+05
		IMF6	-3.80E+05
		IMF7	-4.19E+05
		IMF8	-3.95E+05
		IMF9	-4.01E+05
		IMF10	-3.89E+05
		IMF11	-1.53E+05
		IMF12	-2.08E+05
irm01_s10_solo	2s	12	
		IMF1	-4.43E+05
		IMF2	-3.76E+05
		IMF3	-3.38E+05
		IMF4	-3.06E+05
		IMF5	-3.01E+05
		IMF6	-3.23E+05
		IMF7	-3.80E+05
		IMF8	-3.95E+05
		IMF9	-4.17E+05
		IMF10	-4.01E+05
		IMF11	-4.20E+05
		IMF12	-3.90E+05

Table I: Duration, IMF and Log Energy

REFERENCES

[1] I Jian-Da Wu, Yi-Jang Tsai, "Speaker identification system using empirical mode decomposition and an artificial neural network," *Expert Systems with Applications*, 38 ,6112–6117.

[2] Avci, E., & Akpolat, Z. H. (2006). "Speech recognition using a wavelet packer adaptive network based fuzzy inference system," *Expert Systems with Applications*, 31, 495–503.

[3] Corinthios, M. J. (1971). A fast Fourier transform for high-speed signal processing. *IEEE Transaction on Computer*, C-20, 843–846.

[4] Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). *Learning representations by back-propagating errors*. Nature, 323, 533–536.

[5] F.Cummins, M.Grimaldi, T.Leonard, and J.Simko, "The chains corpus: Characterizing individual speakers," in *Proc.SPECOM'06, St. Petersburg, Russia*, 2006, pp.431-435.