Available online at www.elixirpublishers.com (Elixir International Journal)

Advanced Engineering Informatics

Elixir Adv. Engg. Info. 38 (2011) 4516-4519

ISSN: 2229-712X

Detection of similar pattern of chhattisgarhi dialects through speech recognition using wavelet transformation

Madhuri Gupta[1], Akhilesh Tiwari[1] and A S Zadgaonkar[2]

[1]SSCET, Junwani Bhilai, (C.G.)

[2]C. V. Raman University, Bilaspur, (C.G).

ARTICLE INFO

Article history:

Received: 22 August 2011;

Received in revised form:

26 August 2011;

Accepted: 31 August 2011;

Keywords

Chhattisgarhi language,

Speaker Modelling,

Speech Recognition,

Wavelet Transformation.

ABSTRACT

This paper will presents a method of finding a relations between similar speeches signals which are speaker independent taken from common set of samples of Chhattisgarhi language and dialects commonly spoken at different regions of Chhattisgarh. The objective of this paper is to detect similar and dissimilar patterns of different Chhattisgarhi dialects through wavelet analysis methods which are useful for complex speech signal analysis. Through wavelet transforms parameters are extracted from speech corpus and then relations are established.

Introduction

Chhattisgarhi is a dialect of Hindi Language or language of its own right and it is spoken and understood by the majority of people in Chhattisgarh. Chhattisgarhi was also known as Khaltahi to surrounding people and as Laria to Oriya speakers. Chhattisgarhi has several identified dialects of its own, in addition to Chhattisgarhi proper. Overall Chhattisgarhi can be divided into 26 different types of dialects. Wavelets are highly suitable for analysing transient signals (such as speech) because they are well localised in both the time and frequency domain, whilst exhibiting orthogonality and stability in a number of useful function spaces. A wavelet basis can be broken into subsets of transforms that still retain the spatially desirable features of the parent basis, yet extend the range of descriptive subsets available to characterise a signal. For the creation of this model the system used the voice samples database of the people from different region of the state. This database has a set of 15-20 statements which are used in day to day life by people. These statements have specific words through which we can differentiate the dialects. Using this database we will be able to differentiate a voice on the basis of: ◦ Native Language ◦ Geographic Origin

Analyzing Chhattisgarhi Speech Signal

Speech communication

The purpose of speech is communication. According to information theory, speech can be represented in terms of its Message Contents, Information, and Signal (Acoustic Waveform)

Signal Processing

The general problem of information manipulation and processing is depicted in Figure (2). In the case of speech signal the human speaker is the information source. The measurement is generally the acoustic waveform. Signal processing involves first obtaining a representation of the signal based on a given model and then the application of the some higher level transformation in order to put the signal into a more convenient form. The last step in the process is the extraction and utilization of the message information. This step may be performed by human listeners or automatically by machine.

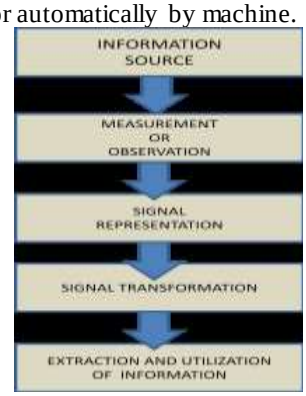

Figure 1: General view of information manipulation and processing

Analysis for this Model

For the analysis of this model the system load the signal from wavelet menu and then get the sample analyzed waveform for baheliya dialect for sentence —Main Jao Chul as depicted in figure 2.

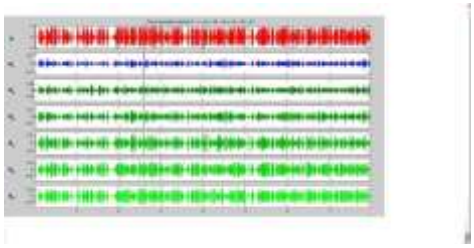

Figure 2: Analyzed acoustic wave signal of Sentence —Main Jao Chul

Wavelet based sd calculation

Statistical Distribution(SD) of Wavelet Coefficient

To calculate the statistical distribution of wavelet coefficients, the standard deviation is considered in terms of

Tele: 09039795781

E-mail addresses: mg.cit.rjn@gmail.com, akhilesh_tiwari@rediffmail.com

second momentum coefficient. It is noted that the mean of speech signal is zero. If we assume that the wavelet coefficients distribution is modeled as Laplace density function, then distribution of these coefficients is related directly to its standard deviation. In this way, assume that the output signal, y, is achieved as the summation of input, w, and noise, n. Then the variance of output is as follows (Eq. 1).

$VAR[y] = VAR[w] + VAR[n]$ (1)

The mean of speech and noise signals is zero, so we have:

$VAR[y] = MEAN\ [y2]$ (2)

The standard deviation is calculated as follows:

$\hat{\sigma} == \sqrt{mean\,[y2]} - \sigma_n^2$ (3)

Standard Deviation for wavelet transform

To calculate standard deviation based on the wavelet transform, the speech signal is applied to two filters. One of these filters is low pass and another is high pass. The original signal is shown in figure 3, on the basis of original signal the system will generate a histogram and cumulative histogram. Histogram and Cumulative histogram of calculated standard deviations is shown in figure 3 and 4.

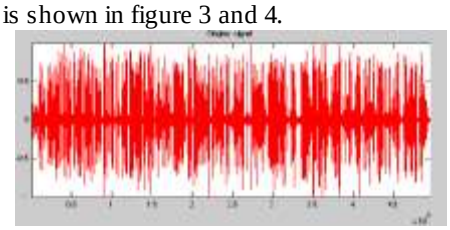

Figure 3: Original Wave signal of Bhunjwari Dialect

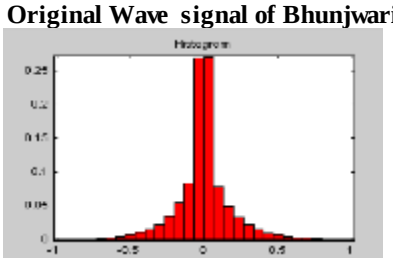

Figure 4: Sample Calculated Histogram of Bhunjwari Dialect

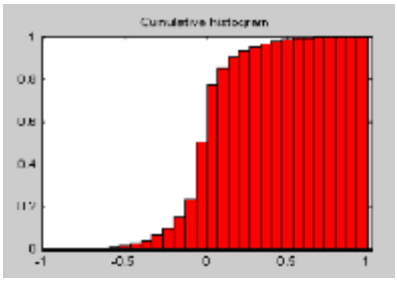

Figure 5: Sample Calculated Cumulative Histogram of Bhunjwari Dialect

Speech parameterization

Speech parameterization consists in transforming the speech signal to a set of features vectors. The aim of this transformation is to obtain a new representation which is more suitable for statistical modelling and the calculation of a distance or any other kind of score. Most of the speech paramerization used in speaker verification systems relies on a cepstral representation of speech.

Filter bank –based cepstral parameters

The speech signal is first pre-emphasized, that is, a filter is applied to it. The goal of this filter is to enhance the high frequencies of the spectrum, which are generally reduced by the speech production process. The pre-emphasized signal is obtained by applying the following filter:

$xp(t) = x(t) - a \cdot x(t-1)$. (4)

Values of a are generally taken in the interval [0.95, 0.98]. This filter is not always applied, and some people prefer not to pre-emphasize the signal before processing it. There is no definitive answer to this topic but empirical experimentation. The analysis of the speech signal is done locally by the application of a window whose duration in time is shorter than the whole signal. This window is first applied to the beginning of the signal, and then moved further and so on until the end of the signal is reached. Each application of the window to a portion of the speech signal provides a spectral vector. For the length of the window, two values are most often used: 20 milliseconds and 30 milliseconds. These values correspond to the average duration which allows the stationary assumption to be true. For the delay, the value is chosen in order to have an overlap between two consecutive windows; 10 milliseconds is very often used. Once these two quantities have been chosen, one can decide which window to use.

Hamming Window

The Hamming and the Hamming windows are the most used in speaker recognition. One usually uses a Hamming window or a Henning window rather than a rectangular window to taper the original signal on the sides and thus reduce the side effects.

Detection of similar pattern via likelihood detection

Single and multi speaker Detection

Given a segment of speech Y and a hypothesized speaker S, the task of speaker verification, also referred to as detection, is to determine if Y was spoken by S. An implicit assumption often used is that Y contains speech from only one speaker. Thus, the task is better termed single speaker verification. If there is no prior information that Y contains speech from a single speaker, the task becomes multi speaker detection.

Likelihood ratio test

The system essentially implements a likelihood ratio test to distinguish between two hypotheses: the test speech comes from the claimed speaker or from an imposter. Features extracted from the speech signal in front-end processing are compared to a model representing the claimed speaker, obtained from a previous enrolment, and to some model(s) representing potential imposter speakers (i.e., those *not* the claimed speaker). The ratio (or difference in the log domain) of speaker and imposter match scores is the likelihood ratio statistic (□), which is then compared to a threshold (□) to decide whether to accept or reject the speaker. The general techniques used for the three main components, front-end processing, speaker models, and imposter models, are briefly described next.

The optimum test to decide between these two hypotheses is a likelihood ratio (LR) test1 given by

$$\frac{p(Y|H0)}{p(Y|H1)} \begin{cases} > \theta, accept\ H0, \\ < \theta, accept\ H1, \end{cases} \quad (5)$$

Where p (Y|H0) is the probability density function for the hypothesis H0 evaluated for the observed speech segment Y, also referred to as the "likelihood" of the hypothesisH0 given the speech segment.2 the likelihood function for H1 is likewise p (Y|H1). The decision threshold for accepting or rejecting H0 is θ. One main goal in designing a speaker detection system is to determine techniques to compute values for the two likelihoods p (Y|H0) and p (Y|H1).

Speaker modeling

Desirable attributes of a speaker model are:

(1) A theoretical underpinning so one can understand model behaviour and mathematically approach extensions and improvements;
(2) Generalizable to new data so that the model does not over fit the enrolment data and can match new data;
(3) Parsimonious representation in both size and computation.
There are many modeling techniques that have some or all of these attributes and have been used in speaker verification systems. The selection of modeling is largely dependent on the type of speech to be used, the expected performance, the ease of training and updating, andstorage and computation considerations. A brief description of some of the more prevalent modelling techniques is given next.

Template Matching

In this technique, the model consists of a template that is a sequence of feature vectors from a fixed phrase. During verification a match scored is produced by using dynamic time wrapping (DTW) to align measure the similarity between the test phrase and the speaker template. This approach is used almost exclusively for text-dependent applications.

Nearest Neighbor

In this technique, no explicit model is used; instead all features vectors from the enrolment speech are retained to represent the speaker. During verification, the match score is computed as the cumulated distance of each test feature vector to its k nearest neighbors in the speaker's training vectors. To limit being modeled and some alternative speakers. Training can be computationally expensive and models are sometimes not generalizable.

Neural networks

The particular model used in this technique can have many forms, such as multi-layer perceptions or radial basis functions. The main difference with the other approaches described is that these models are explicitly trained to discriminate between the speaker Models (GMMs), are used. From published results, HMM based systems generally produce the best performance.

Hidden Markov Model

This technique uses HMMs, which encode the temporal evolution of the features and efficiently model statistical variation of the features, to provide a statistical representation of how a speaker produces sounds. During enrollment HMM parameters are estimated from the speech using established automatic algorithms. During verification, the likelihood of the test feature sequence is computed against the speaker's HMMs. For text-dependent applications, whole phrases or phonemes may be modeled using multi-state left to right HMMs. For text-independent applications, single state HMMs, also known as Gaussian Mixture Models (GMMs), are used. From published results, HMM based systems generally produce the best performance.

Strengths and Weaknesses

Strengths

It is clear that speaker verification technology is indeed ready for use. But, as stated before, it is not the universal solution. The main strength of speaker verification technology is that it relies on a signal that is natural and unobtrusive to produce and can be obtained easily from almost anywhere using the familiar telephone network (or internet) with no special user equipment or training. This technology has prime utility for applications with remote users and applications already employing a speech interface. Additionally, speaker verification is easy to use, has low computation requirements (can be ported to cards and handhelds) and, given appropriate constraints, has high accuracy.

Weaknesses

Some of the flexibility of speech actually lends to its weaknesses. First, speech is a behavioral signal that may not be consistently reproduced by a speaker and can be affected by a speaker's health (cold or laryngitis). Second, the varied microphones and channels that people use can cause difficulties since most speaker verification systems rely on low-level spectrum features susceptible to transducer/channel effects. Also, the mobility of telephones means that people are using verification systems from more uncontrolled and harsh acoustic environments (cars, crowded airports), which can stress accuracy. Robustness to channel variability is the biggest challenge to current systems. Spoofing of systems is often cited as a weakness, but there have been may approaches developed to thwart such attempts (prompted phrases, knowledge verification).

There is current effort underway to address these known weaknesses. Some of these weaknesses may be overcome by combination with a complementary biometric, like face recognition.

Conclusion

Speech recognition is the process by which a computer (or other type of machine) identifies spoken words. Wavelet analysis is capable of revealing aspects of data that other signal analysis techniques miss aspects like trends, breakdown points, discontinuities in higher derivatives, and self-similarity. This model will helpful to find out patterns and parameters for speeches and come out with general model to answer the questions like Locality. 22 set of sentences have been used whose conversion in different dialects are recorded for the creation of the database. Basically three types of dialects among 26 dialects are used in this model. These dialects are: Baigani, Baheliya,and Bhujwari. These standard set of sentences contains almost all major differences in the dialects through which analysis and recognition of these dialects is easily possible.

Acknowledgment

The real spirit of achieving a goal is through the way of excellence and asteroids discipline. I would have never succeeded in completing my task without the cooperation, encouragement and help provided to me by various personalities. I want to thank SSCET, Bhilai for providing me the necessary software, tools and other resources to deliver my research work. With deep sense of gratitude I express my sincere thanks to my esteemed and worthy Guide Prof. Akhilesh Tiwari ,SSCEC , Bhilai for their valuable guidance in carrying out this work under their effective supervision. Encouragement, enlighten and cooperation.

References

[1] Madhuri Gupta & Akhilesh Tiwari .―Speech Analysis of Chhattisgarhi (dialect) Speech signal of different regions of Chhattisgarh. International Conference on emerging trends in soft computing (SCICT), March 2011,Bilaspur,India.
[2]Leh Luoh,Yu-Zhe Su and Chih-Fan Hsu. "Speech signal processing based emotion recognition".2010 International Conferences on System Science and Engineering. 978-1-4244-6474-61101$26.00 © 2010 IEEE.
[3] Mansour Sheikhan1, Mohammad Khadem Safdarkhani2, Davood Gharavian3. ―Presenting and Classification Based on Three Basic Speech Properties, Using Haar Wavelet

Analyzing‖. 2010 2nd International Conference on Signal Processing Systems (ICSPS).
[4] Fr ´ed´ eric Bimbot,1 Jean-Franc¸ois Bonastre, A Tutorial on Text-Independent Speaker Verification‖. EURASIP Journal on Applied Signal Processing 2004:4, 430–451.
[5] Douglas A. Reynolds "Automatic Speaker Recognition: Current Approaches and Future Trends" ICASSP 2001.
[6] Ilyas Potamitis and George Kokkinakis, ―Speech Separation of Multiple Moving Speakers Using Multisensory Multistage Techniques‖, IEEE TRANSACTION ON SYSTEM, MAN, AND CYBERNETICS—PART A: SYSTEMS AND HUMANS, VOL. 37, NO. 1, JANUARY 2007.
[7] G. R. Doddington, ―Speaker Recognition based on Idiolectal Differences between Speakers,‖ Eurospeech 2001.
[8] A. D. Andrews, M. A. Kohler and J. P. Campbell, ―Phonetic Speaker Recognition,‖ Eurospeech 2001.