Hindi Speaking Person Identification using Zero Crossing rate and Short-Term Energy

L. P. Bhaiya, Arif Ullah Khan

Abstract—language is man's most important means of communication and speech its primary medium. Speech recognition is the ability of a computer to recognize general, naturally flowing utterances from a wide variety of users. Differences of physiological properties of the glottis and vocal tracts are partly due to age, gender and/or person differences. Since these differences are related in the speech signal, acoustic measures related to those properties can be helpful for speaker identification. Acoustic measure of voice sources were extracted from 5 utterances spoken by10 peoples including 5 male and 5 female talkers (aged 19 to 25 years old). The differences of speech long term features including zero crossing rate and short term energy for different person is studied.

I. INTRODUCTION

A classification of speech into person's voice sound provides a useful basis for subsequent processing, for example fundamental frequency estimation, formant extraction or syllable marking. A classification into person's voice extends the possible range of further processing to tasks consonant identification and endpoint detection for isolated utterances.

We approach the problem of person identification from two different directions: zero crossing rate and short-term energy. These two methods compliment each other well, and prevent us from having to rely heavily on the single method to label the different part of the speech.

We do not provide an interface for viewing the results of this study instead the result of person identification are stored in a format that allows them to be loaded in tool.

II. LITERATURE REVIEW

Zero-crossing rate is proposed for sex identification and result of about 97% for gender classification is obtained [6]. Many attempts in speaker recognition have taken place in last thirty years. Major efforts have been made to develop methods for extracting information from audio-visual media, in order that they may be stored and retrieved in database automatically. Recent work has been done on segmentation of the audio signal and then classification into one of two main categories: speech or music [2, 3]. Zero-crossing rate is proposed for musical instrument identification and result reflects more effectively the difference of tones in musical instrument. [4]. An approach for separating the voiced/unvoiced part of the speech in a simple and efficient way, the algorithm shows good result in classifying the speech as the segmented speech into many frames [5].

Manuscript received September 02, 2012.

Lalit P Bhaiya, Department of Electronics and Telecommunications, Chhattisgarh Swami Vivekananda Technical University/Rungta College of Engineering and Technology/Bhilai, India.

Arif Ullah Khan, Department of Electronics and Telecommunication, Chhattisgarh Swami Vivekananda Technical University /RSR Rungta College Of Engineering and Technology/ Bhilai, India,

However the issue is yet far from being solved. The work on recognition from still remains crucial.

The performance achieved by listening, visual examination of spectrograms, and automatic computer techniques, attempt to provide a perspective with which to evaluate the potential of speaker recognition and productive direction for research into and application of speaker recognition technology [1].

An approach for automatic detection of mental illness from the speech signal has taken place using classifier configuration employed in emotion recognition from speech, evaluated on a speaker depressed sentence speech database [8]. The emotional speech recognition having three goals, having resources, features and methods, the feature used for emotion recognition is short-term energy which gives best result as compared with others [7].

III. PERSON IDENTIFICATION USING ZERO CROSSINGS

The notion of zero crossing is defned to be -

"*The number of times in a sound sample that the amplitude of the sign wave changes sign* "

For 10ms sample of clean speech, the zero-crossing rate is approximately 12 for voiced speech and 50 for unvoiced speech. For clean speech the zero-crossing rate should be useful for detecting region of silence, as the zero-crossing rate should be zero.

Unfortunately, very few sound samples are recorded in perfectly clean speech. This means there is some level of background noise, that interferes with the speech, meaning that the silent region actually have quite a high zero-crossing rate as the signal changes from just one side of zero amplitude to the other and back again. For this reason a tolerance threshold is included in the function that calculates zero-crossing to try and alleviate this problem. The thresholds work by removing any zero-crossings, which do not both start and end a certain amount from the zero value.

In this study we have used a threshold of 0.001 this means that any zero-crossings that start and end in the range of 'x', where $-0.001 < x < 0.001$, are not included in the total number of zero-crossing for that window. This enables us to filter out most of the zero-crossings that occur during silent region of the sample due to background noise.

IV. PERSON IDENTIFICATION USING SHORT-TERM ENERGY

Unfortunately, unlike zero-crossings there are no standard values of short-term energy for specific window sizes. Short term energy is purely dependent upon the energy in the signal, which changes depending upon whose sound was recorded, for example if two persons are speaking the same phrases, then the short-term energy values will be vastly different, although the zero-crossing values should be

roughly the same. This means that you have to inspect the recorded speech files to determine what level to make the distinction between the different persons.

There is one thing that is standard enough, and that is that short-term energy is different for different person and should be zero for silent region in clean recording of clean speech.

In a similar way to zero-crossing we calculate the short-term energy using a 10ms non-overlapping rectangular window.

V. PERSON IDENTIFICATION USING BOTH METHODS

From the descriptions of the methods that are used to identify the person through speech signal, in this study, it should be clear that the two methods compliment each other well. For every person short-term energy is different and zero-crossing is same for some persons, whether that person is male or female. This can be seen clearly in table 6-1(a).

In a perfect world, all speech samples would be clean and then table 6-1 could be used to classify the speech as different persons voice.

VI. RESULTS

Twenty people independent of each other, is manually labeled as the five sound files by each person. As the speech signal is text dependent, all the speakers were requested to say the hindi alphabets ”cha” ”tta” “*ka*” for speech based speaker identification.

A. Result Using Zero-Crossing Rate

It indicates the frequency of signal amplitude sign changes. To some extent, it indicates the average signal frequency as:

$$ZCR = \frac{\sum_{n=1}^{N} |\operatorname{sgn} x(n) - \operatorname{sgn} x(n-1)|}{2N}$$

Where *sgn[]* is a signum function and *x(m)* is discrete audio signal.

In mathematical terms, a “zero-crossing” is a point where the sign changes(e.g from positive to negative), represented by a crossing of the axis(zero value) in the graph of the function. The zero-crossing is important for systems which send digital data over AC circuits, such as modem, X10 home automation control systems for Lionel and other AC model trains. Counting zero-crossing is also a method used in speech processing to estimate the fundamental frequency of the speech.

Zero-crossing rate is important because they abstract valuable information about the speech and they are simple to compute.

Cha:-

S.No	Person	Frames					
		0	5	10	15	20	25
1	F1	36	50	42	37	37	45
2	F2	48	48	44	36	46	42
3	F3	51	52	52	51	52	49
4	F4	49	56	57	53	62	55
5	F5	52	41	46	43	40	48

TTa:-

S.No	Person	Frames					
		0	5	10	15	20	25
1	F1	39	56	39	39	40	38
2	F2	64	66	43	48	44	41
3	F3	46	4	55	47	53	49
4	F4	38	48	46	48	56	43
5	F5	28	44	40	39	48	55

Ka:-

S.No	Person	Frames					
		0	5	10	15	20	25
1	F1	39	56	39	39	40	38
2	F2	64	66	43	48	44	41
3	F3	46	4	55	47	53	49
4	F4	38	48	46	48	56	43
5	F5	28	44	40	39	48	55

Cha:-

S.No	Person	Frames					
		0	5	10	15	20	25
1	M1	40	154	55	38	38	38
2	M2	51	194	46	32	51	34
3	M3	39	143	42	112	34	45
4	M4	40	158	31	35	40	41
5	M5	48	157	60	60	46	44

Tta:-

S.No	Person	Frames					
		0	5	10	15	20	25
1	M1	49	52	41	41	46	42
2	M2	57	59	59	51	51	37
3	M3	43	48	36	45	44	45
4	M4	39	41	63	38	45	45
5	M5	43	51	44	58	59	27

Ka:-

S.No	Person	Frames					
		0	5	10	15	20	25
1	M1	41	45	37	37	52	52
2	M2	46	75	44	37	40	43
3	M3	47	49	49	35	55	46
4	M4	47	75	36	36	35	30
5	M5	48	44	57	59	45	45

B. Result Using Short-Term Energy

The short-term energy measurement of a speech signal can be used to determine voiced vs. unvoiced speech. Short-term energy can also be used to detect the transition from unvoiced to voiced speech or vice versa. The energy of voice speech is much greater than the energy of unvoiced speech.

We record i\p signal at fs=8KHz. Now using hamming window with the following specifications: window size = 256 samples, window step = 100 samples, window overlap = 156 samples and number of frames = (length of i\p – window size)/ (window step), we calculate the STE for each frame using the following formula.

Short-term energy allows us to calculate the amount of energy in a sound at a specific instance in time, and is defined in equation 3-1.

$$E_n = \left(\sum_{m=-\infty}^{\infty} x^2(m) h(n-m) \right) \quad \text{--- (1)}$$

Eq. (1) defines the short time energy for a sampled signal where *h(n-m)* is a windowing function. For simplicity a rectangular windowing function is used as defined in eq. (2).

$$H(n) = \begin{matrix} 1 & 0 \le n \le N-1 \\ 0 & otherwise \end{matrix} \quad \text{--- (2)}$$

N in eq. (2) is the length of the window in samples.

Cha:-

S.No	Person	Frames					
		0	5	10	15	20	25
1	M1	2	2	7	10	9	8
2	M2	2	2	8	9	7	7
3	M3	1	2	4	8	5	4
4	M4	3	4	6	7	5	3
5	M5	2	3	6	6	5	5

Tta:-

S.No	Person	Frames					
		0	5	10	15	20	25
1	M1	2	3	4	7	6	5
2	M2	2	3	5	9	8	6
3	M3	4	4	7	8	4	3
4	M4	2	4	6	8	5	2
5	M5	3	3	7	7	7	3

Ka:-

S.No	Person	Frames					
		0	5	10	15	20	25
1	M1	3	4	7	9	9	4
2	M2	4	6	9	11	9	7
3	M3	3	3	8	10	7	3
4	M4	4	5	6	9	9	2
5	M5	3	5	5	9	5	2

Cha:-

S.No	Person	Frames					
		0	5	10	15	20	25
1	F1	6	6	8	7	4	3
2	F2	9	7	6	5	3	2
3	F3	5	7	7	6	4	3
4	F4	6	7	6	5	4	2
5	F5	9	7	7	6	4	3

Tta:

S.No	Person	Frames					
		0	5	10	15	20	25
1	F1	7	7	5	4	3	2
2	F2	8	9	7	6	5	3
3	F3	9	9	8	7	6	4
4	F4	8	9	7	6	5	3
5	F5	6	8	7	6	5	2

Ka:-

S.No	Person	Frames					
		0	5	10	15	20	25
1	F1	2	6	5	3	2	1
2	F2	3	7	5	2	1	1
3	F3	2	8	6	4	3	1
4	F4	4	7	8	5	1	1
5	F5	1	6	5	3	2	1

VII. DISCUSSION

To label the sample a zero-crossing function and short-term energy function were applied to the sample. These functions are complimentary, for example zero-crossing are high when the speech is unvoiced, but the short-term energy is low at this point, the vice versa is true for voiced speech, and both are approximately zero for silence. The function uses the frequency of a sample and a window size of 10ms to split the sample into sections and produce the results.

There are cases where the function produce different values, these problem is partially caused by background noise. This causes the cutoff for silence to be raised, as it may not be quite zero due to noise being interpreted as speech by the functions.

The way different people talk, such as volume and speed also causes problem identifying endpoints of words. As the samples are of different volumes the cutoff value would need to be changed for each sample making accuracy hard from sample to sample. It would also be a very time consuming activity having to tweak the cutoff values for each sample, which would also defeat the object of this study.

We decided that if the result of the functions didn't match then if the short-term energy implies person 1 voiced speech and zero-crossing implies person 2 voices speech, then the result should be person 1 voiced speech. This is because zero-crossing sometimes have same values for different persons therefore there is more chance of an error between these values, but the short-term energy is only different for different persons voice speech occurs.

In retrospect these assumption seems to have been valid assumption to make, as the person identification technique produced under these assumptions seems on the whole to be correct.

VIII. CONCLUSION

The part of the speech labeling produced using the algorithm, obtained in this study are reasonably accurate and well recorded.

The accuracy, of the algorithms outlined in this study, could be improved in two ways.

Firstly more time could be spent on tweaking the cut-off values used by algorithms to label the different part of the speech. The problem with this, however, is that if the values are fine tuned for one speech sample it is unlikely that they will be as accurate on other speech samples.

The other possible way of increasing the accuracy of the algorithms would be to use an over-lapping hamming window, when calculating the zero-crossing and short-term energy.

In this paper, we examined the role of voice source measure in person identification and compare the results to perceptual experiment performed on the same database, voice source measures were extracted from a large database of utterances spoken by 10 peoples with equal number of males and females.

We used different parameters in the analysis and from the experiment we could observe evident results for zero-crossing rate and short-term energy. Zero-crossings rate sometimes give the same value for different persons and short-term energy shows different values for different persons present in the database.

These coefficients contain useful information about different person identification and employing them in such a process lead to decrease in person identification error rate.

REFERENCES

1. Yiu - Kei Lau and Chok- Ki Chan,"Speech recognition based on zero-crossing rate", IEEE Transactions on acoustics, speech and signal processing, Vol.ASSP-33, No.1.
2. Costas panagiotakis and George tziritas, "A speech/music discriminator based on RMS and zero-crossings", IEEE transactions on multimedia.vol.7, no.1, February 2005.
3. Sumit Kumar Banchhor, Om Prakash Sahu, Prabhakar, "A Speech/Music Discriminator based on Frequency energy, Spectrogram and Autocorrelation", IJSCE, Volume-1, Issue-6, January 2012
4. Sumit kumar Banchhor and Arif Khan, "Musical Instrument Recognition using Zero Crossing Rate and Short-time Energy", Volume 1– No.3, February 2012.
5. Bachu R.G, Kopparthi S, Adapa B, Barkana B.D, "Separation of voiced and unvoiced using zero crossing rate and energy of the signal".
6. Sumit kumar Banchhor and S. K. Dekate, "Text-dependent Method for Gender Identification through synthesis of voiced segments",IJEST, Volume- 3, No. 6, June 2011
7. Dimitrios Ververidis and Constantine Kotropoulos, "Emotional Speech Recognition: Resources, Features, and methods".
8. Nicolas Cummins, Julien Epps, Miachael Breakspear, and Roland Goecke, " An Investigation on Depressed Speech Detection: Features and Normalization".