Comparative study of voiced-unvoiced-silence classification of speech

Shridevi A. Jirli[(1)] Ramachandra G Turkani [(2)]
(1), (2) Assistant Professor, K L E College of Engineering & Technology, Chikodi

shridevi.jirli19@gmail.com rgturkani@gmail.com

Abstract **- In speech analysis, the voiced-unvoiced decision is usually performed in conjunction with pitch analysis. The linking of voiced-unvoiced (v-uv) decision to pitch analysis not only results in unnecessary complexity, but makes it difficult to classify short speech segments which are less than a few pitch periods in duration. Here, we describe a pattern recognition approach for deciding whether a given segment of a speech signal should be classified as voiced speech, unvoiced speech, or silence, based on measurements made on a signal. In this method, five different measurements are made on the speech segment to be classified. The measured parameters are the zero-crossing rate, the speech energy the correlation between adjacent speech samples, the first predictor coefficient from a 12-pole linear predictive coding (LPC) analysis, and the energy in the prediction error. The speech segment is assigned to a particular class based on a minimum distance rule obtained under the assumption that the measured parameter are distributed according to the multidimensional Gaussian probability density function. The means and covariances for the Gaussian distribution are determined from manually classified speech data included in a training set. The method has been found to provide reliable classification with speech segments as short as 10ms and has been used for both speech analysis-synthesis and recognition applications. A simple nonlinear smoothing algorithm is described to provide a smooth 3-level contour of an utterance for use in speech recognition applications.**

Entire Design is carried out using MATLAB.

Keywords **– Speech Measurement.**

I. Introduction

The need for deciding whether a given segment of a speech waveform should be classified as voiced speech, unvoiced speech, or silence (absence of speech) arises in many speech analysis systems. A variety of approaches have been described in the speech literature for making this decision - methods for voiced-unvoiced (V--UV) decision usually work in conjunction with pitch analysis. For example, in the well known ceptral pitch detector, the V-UV decision is made on the basis of the largest peak in the ceptrum. A classification of speech into voiced or unvoiced sounds provides a useful basis for subsequent processing, for example fundamental frequency estimation, formant extraction or syllable marking. A three-way classification into silence/unvoiced/voiced (hence the title, SUV) extends the possible range of further processing to tasks such as stop consonant identification and endpoint detection for isolated utterances. Strictly, speech sounds such as voiced fricatives (e.g. "z") can have characteristics of both voiced and unvoiced sources simultaneously which makes classification more difficult, so for the purposes of this study we will assume that a 3-way classification is sufficient for the needs of any further processing.

There are five types of measurements :

1) Energy of the signal.
2) Zero-crossing rate of the signal.
3) Autocorrelation coefficient at unit sample delay.
4) First predictor coefficient.
5) Energy of the prediction error.

We approach this problem of SUV from two different directions: zero-crossings and short-term energy. These two methods complement each other well, and prevent us from having to rely heavily on one single method to label the different parts of speech.

The classification of the speech signal into voiced, unvoiced, and silence provides a preliminary acoustic segmentation of speech, which is important for speech analysis. The nature of the classification is to determine whether **a** speech signal is present and, if so, whether the production of speech involves the vibration of the vocal folds. The vibration of vocal folds produces periodic or quasi-periodic excitations to the vocal tract for voiced speech whereas pure transient and/or turbulent noises are aperiodic excitations to the vocal tract for unvoiced speech. When both quasi-periodic and noisy excitations are

present simultaneously (mixed excitations), the speech is classified here **as** voiced because the vibration of vocal folds is part of the speech act. The mixed excitation could also be treated as an independent category.

The V/UV classification can be performed using a single feature, whose behavior could be significantly affected by the presence or absence of voicing activity. The accuracy of such an approach would not go beyond a certain limit, because the range of values of any single parameter generally overlaps between different categories. The confusion caused by overlapping between different regions is further intensified if speech has not been recorded in a high-fidelity environment. Although V/UV classification has been traditionally tied to the problem of pitch frequency determination, the vibration of the vocal cords does not necessarily result in periodicity in the speech signal. Therefore, a failure in the detection of periodicity in some voiced regions would result in V/UV classification. Voiced-unvoiced classification in noise, however, is a far more challenging task since the noise can potentially mask low-energy speech segments such as fricatives (e.g.,If/, lthl) and stop-consonants (e.g., *lbl,* Id/).

The note is organized as follows: Review of previous techniques is proposed in section II.
Study of speech measurement methods and block diagram in section III. Study of algorithms in section IV. Categorization result of speech, advantages and disadvantages are listed in section V. Conclusion is proposed in section VI.

II. Speech Measurements

2.1 Short-Term Energy

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 1.

$$\boldsymbol{E_n} = \sum_{m=n-N+1}^{n} (\boldsymbol{x(m)w(n-m)})^2 \quad (1)$$

Equation 1: **Short-Term Energy (w is the window, n is the sample that the window is centered on,and N is the window size [1]).**

Unfortunately, unlike zero-crossings there are no standard values of short-term energy for specific window sizes. The choice of the window determines the nature of the short-time energy representation. In our model, we used Hamming window. The hamming window gives much greater attenuation outside the band pass than the comparable rectangular window.

h(n) = 0.54-0.46cos(2πn/(N-1)) , 0 ≤ n N-1

h(n) = 0 , otherwise

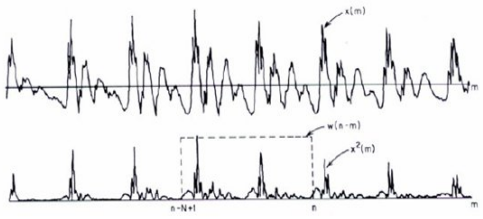

Fig. 1: Computation of Short-Time Energy

The attenuation of this window is independent of the window duration. Increasing the length, N, decreases the bandwidth, Fig 1. If N is too small, n E will fluctuate very rapidly depending on the exact details of the waveform. If N is too large, n E will change very slowly and thus will not adequately reflect the changing properties of the speech signal.

Short-term energy is purely dependent upon the energy in the signal, which changes depending on how the sound was recorded. For example, if a person is recorded saying the same phrase twice, one while whispering and once while shouting, 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 at what level to make the distinction between voiced and unvoiced speech. There is one thing that is standard though, and that is that short-term energy is higher for voiced than un-voiced speech, and should also be zero for silent regions in clean recording of clean speech. In a similar way to zero-crossings we calculate the short-term energy using a 10ms non-overlapping rectangular window. This, again, is not as accurate as using an overlapping hamming window but it is adequate for the SUV labeling of speech.

2.2 Zero-Crossing Rate

The notion of zero-crossing is defined to be:
"the number of times in a sound sample that the amplitude of the sound wave changes sign"
For a 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 also be useful for detecting regions of silence, as the zero-crossing rate should be zero. Unfortunately, very few sound samples are recordings of perfectly clean speech. This means that often there is some level of background noise that interferes with the speech, meaning that silent regions actually have quite a high zero-crossings 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-crossings 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-crossings for that. This enables us to filter out most of the zero-crossings that occur during silent regions of the sample due only to background noise. In this study to calculate zero-crossings we used a 10ms non-overlapping rectangular window. This does not produce such good zero-crossing results as an overlapping hamming window would, but since we are not interested in the fine details, this method works well when used to SUV a speech sample. In the context of discrete-time signals, a zero crossing is said to occur if successive samples have different algebraic signs. The rate at which zero crossings occur is a simple measure of the frequency content of a signal. Zero-crossing rate is a measure of number of times in a given time interval/frame that the amplitude of the speech signals passes through a value of zero.

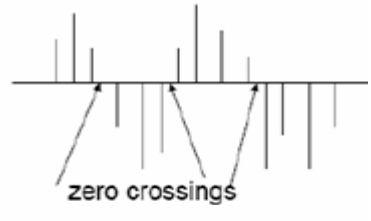

Fig. 2: Definition of zero-crossings rate

Fig 2 and Fig 3 .Speech signals are broadband signals and interpretation of average zero-crossing rate is therefore much less precise. However, rough estimates of spectral properties can be obtained using a representation based on the short-time average zero-crossing rate. Real world labeling of short term energy and zero crossing is shown in table 1.

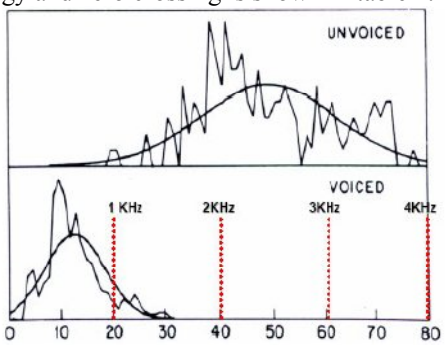

Fig 3: Distribution of zero-crossings for unvoiced and voiced speech

A definition for zero crossings rate is:

$$Zn=\sum_{m=-\infty}^{\infty}|sgn[x(m)]-sgn[x(m-1)]|w(n-m) \quad (2)$$

where

$$sgn[x(n)] = \begin{cases} 1, & x(n) \geq 0 \\ -1, & x(n) < 0 \end{cases}$$

and

$$w(n) = \begin{cases} \frac{1}{2N} & \text{for } 0 \leq n \leq N-1 \\ 0 & \text{for otherwise} \end{cases}$$

Zero crossings	Short term energy	Label
approx 0	approx 0	Silence
High	Low	Un-voiced
Low	High	Voiced
approx 0	High	Voiced
High	High	Voiced
Low	Low	Voiced
approx 0	Low	Un-voiced
Low	approx 0	Silence
High	approx 0	?

Table 1: Real world labeling scheme

III Block Diagram

In our design, we combined zero crossings rate and energy calculation. Zero-crossing rate is an important parameter for voiced/unvoiced classification. It is also often used as a part of the front-end processing in automatic speech recognition system. The zero crossing count is an indicator of the frequency at which the energy is concentrated in the signal spectrum. Voiced speech is produced because of excitation of vocal tract by the periodic flow of air at the glottis and usually shows a low zero-crossing count, whereas the unvoiced speech is produced by the constriction of the vocal tract narrow enough to cause turbulent airflow which results in noise and shows high zero-crossing count.

Energy of a speech is another parameter for classifying the voiced/unvoiced parts. The voiced part of the speech has high energy because of its periodicity and the unvoiced part of speech has low energy. The analysis for classifying the voiced/unvoiced parts of speech has been illustrated in the block diagram in Fig 4. And also the speech segments for voiced and unvoiced speech are shown in Fig 5.

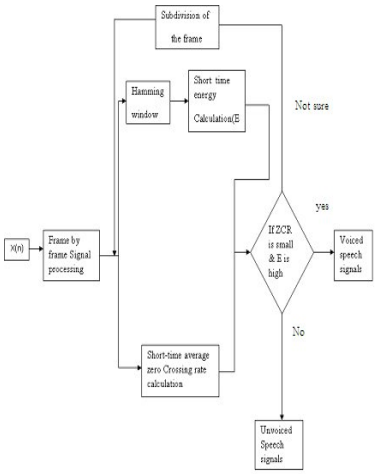

Fig 4: Block diagram

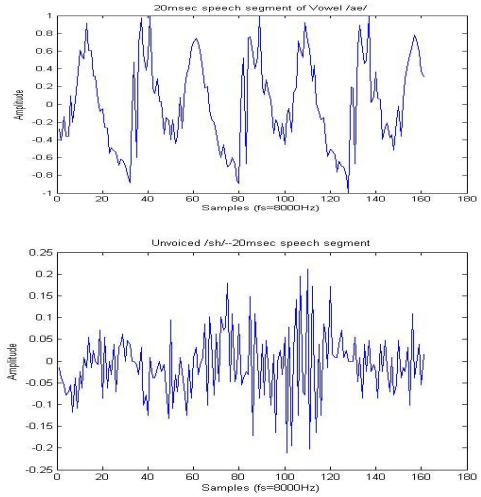

fig 5: Speech Segments

IV. Algorithms

Algorithm for Speech Measurement

Step 1: Read the input speech signal.
Step 2: Divide the signal into frames.
Step3: Find the zero-crossing rate and the energy of the signal in each frame using Matlab.
Step 4: Display the figure along with values.

Algorithm for Speech Classification

Step 1: Divide the sound wave into frames of 10ms.
Step 2: Assign a scaled energy value to each sample.

Step 3: Using the resulting scaled energy levels, obtain two thresholds by which classification is conducted: silence and un-voiced thresholds.
Step 4: Cycle through each frame and assign them speech classifications in a marker array

a. If the scaled energy of the frame is less than the silence threshold, mark as 0.
b. If the scaled energy of the frame is less than the unvoiced threshold, mark as 0.5
c. All other frames, mark as 1.

Step 5: Using the markers the voiced, un-voiced, and silence portions of the sound can be identified.

V Result and Analysis

MATLAB 7.0 is used for our calculations. We choose MATLAB as our programming environment as it offers many advantages. It contains a variety of signal processing and statistical tools, which help users in generating a variety of signals and plotting them. MATLAB excels at numerical computations, especially when dealing with vectors or matrices of data.

In the frame-by-frame processing stage, the speech signal is segmented into a non-overlapping frame of samples.

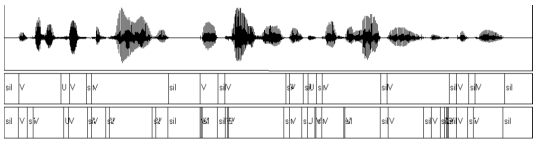

Figure 6: The waveform, manual transcription, and automatic transcription for the file eg1.au, where the percentage correctness is 88%. Produced using the *slt* tool.

Two people independent of each other, manually labeled the ten sound files with silence, unvoiced or voiced. The results of the manual labeling were then compared and discussed so that the most accurate manual SUV results could be obtained. The technique that was used to label the samples was to inspect the spectrogram and the speech waveform to identify silence and speech. Once areas of silence and non-silence are established, the non-silence parts of speech are labeled as voiced or unvoiced. Voiced speech can be distinguished from unvoiced speech as it has a much greater amplitude displacement, when the speech is viewed as a waveform. Another way of telling voiced from unvoiced speech is through examining the spectrogram. In a spectrogram areas of voiced speech have obvious structure (actually the formants), whereas unvoiced speech lacks any real speech. As has already been stated, the automatic

translations are then produced using a comparison of zero-crossings and short-term energy for 10ms non-overlapping rectangular windows of the speech signal. These two sets of transcriptions are difficult to compare by eye, as can be seen in figure 6, so a method of comparison is needed. The method employed in this study is to calculate the percentage of windows in the two transcriptions that match.

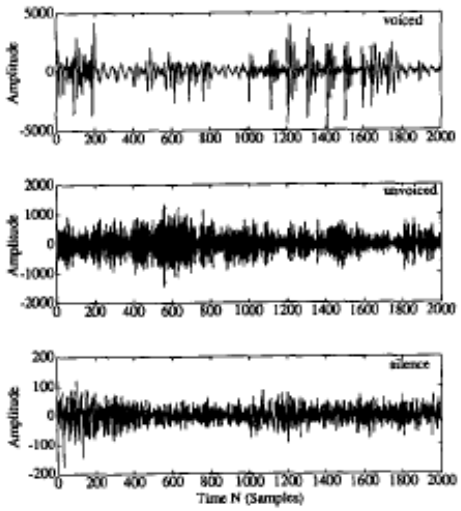

Fig 7: Training samples for the three sound categories.

Figure 7 shows training samples for three sound categories like voiced, un-voiced and silence classification of speech. We can see that voiced speech have high amplitude compared to un-voiced and silence speech. But un-voiced speeches have more frequency.

We can analyze pitch period of all samples and we can determine which type of speech signal it is. If signal having high pitch period then it is voiced speech and if signal is having low pitch period then we decide it as unvoiced speech signal.

As mentioned in section II zero crossing rate and energy of the signal are very easiest methods to determine voiced-unvoiced and silence speech signals. So here the algorithms considered are very advantageous because as we are very familiar with thresholds, here we will divide sound wave into frames and later by their scaled energies will get different thresholds so we can determine the type of signal.

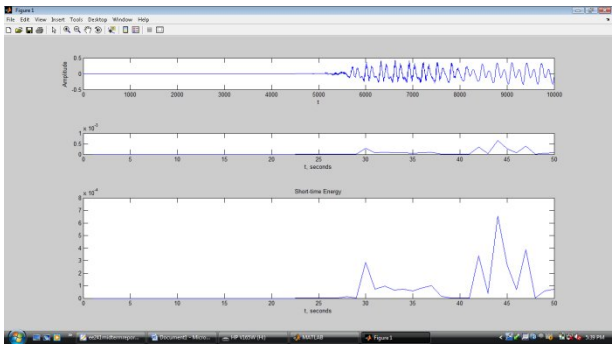

Figure 8: V/UV speech classification

Conclusion

The parts of speech labeling produced using the algorithms outlined in this study are reasonably accurate for well recorded, fairly clean speech but are not nearly as accurate for quiet recordings of speech. 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 the algorithms to label the different parts of 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-crossings and short-term energy. This would, however, mean that many more calculations were necessary for each speech file, which would drastically increase the time taken to label an entire speech file. If, however, the speed of SUVing is not an issue, then this method of improving the algorithms is preferred to fine tuning the cut-off values. The Matlab code of the algorithms outlined in this paper and the manual and automatic transcriptions of the speech samples, for use with the *slt* tool.

References

[1] Bishnu s. atal, member, ieee, and lawrence r. rabiner, fellow,IEEE A Pattern Recognition approached to voiced/unvoiced silence classification with application to speech .

[2] *Mark Greenwood, Andrew Kinghom* SUVING "Automatic silence/voiced/unvoiced classification of speech".

[3] B. Atal and **S.** Hanuer, "Speech analysis and synthesis by linear prediction of the speech wave," *J. Acoust. Soc. Amer.,* vol. *50,* Aug. 1971

[4] "Two-Feature Voiced/Unvoiced Classifier Using Wavelet Transform"
A.E. Mahdi* and E. Jafer

