Automatic Silence/Unvoiced/Voiced Classification of Speech Using a Modified Teager Energy Feature

ALEXANDRU CARUNTU GAVRIL TODEREAN ALINA NICA
Department of Telecommunications
Technical University of Cluj-Napoca
26-28 Baritiu str., Cluj-Napoca
ROMANIA

Abstract: - Labeling of speech signals is a very important task, which can not miss in any of the early stages of developing a system based on a speech technology. Because of the large amount of time consumed when it is done by hand, there is a major need for algorithms that perform it automatically. This paper investigates a few methods of automatic classification of speech in silence/voiced/unvoiced (SUV) regions, using both time and frequency domain parameters. The features include zero-crossings, root mean square energy, but also a modified version of Teager energy, which proves to give the best results.

Key-Words: - Zero-crossing rate, root mean square energy, Teager energy, frame-based Teager energy.

1 Introduction

The most common speech classification into voiced or unvoiced sounds can be found at the speech vs. sound discrimination, an issue of great importance in many areas of speech processing. For example, in automatic recognition of isolated words, is used a scheme for locating the beginning and end of an utterance, based on energy and zero-crossing rate [1]. Fundamental frequency estimation or formant extraction can benefit from this kind of classification also. A three-way classification into silence/unvoiced/voiced extends the possible range of further processing to tasks like syllable marking or stops consonant identification [2].

An important issue when performing a SUV classification is related to the features that must be used. Parameters like LPC derived cepstrum coefficients (LPCC), mel-frequency cepstrum coefficients (MFCC) and so on, have been found to work well, but are time and computationally intensive [3]. Compared to these, the calculus of the energy of the speech signal is a computationally simple operation, most of the algorithms being based on simple parameters such as energy contours and zero crossings.

The Teager energy is a feature that sums the effects of both energy and frequency. The benefits of using it have showed in noisy environments, where it outcomes the classical root mean square energy [4], [5]. In this paper we used it to SUV classification as an alternative to the traditional energy feature.

Rest of the paper is organized as follows: in section II we briefly review the algorithms for extracting the features for classification. Section III presents the classification criteria which we used to identify the nature of a speech portion. Section IV evaluates the performance of the methods of classification. Conclusions are given in Section V.

2 Features Used for SUV Classification

Most common features used for SUV classification are zero-crossing rate and energy. As an alternative to the last one, Teager energy can be used.

2.1 Zero-crossing rate

Zero-crossing rate (ZCR) is defined as the number of times in a sound sample that the amplitude of the sound wave changes sign:

$$ZCR = \sum_{n=0}^{N-2} \frac{1 - \mathrm{sgn}[s(n)]\mathrm{sgn}[s(n+1)]}{2}, \tag{1}$$

where *s* is the signal and *sgn* is the *signum* function.

An estimation of ZCR for 10 ms of clean speech would give an approximate value of 12 for voiced portions and 50 for unvoiced ones [2]. Theoretically a null value for ZCR would correspond to silence.

Usually, background noise interferes with the speech, which means that in silent regions we have a high number of zero crossings. To avoid the use of a noise detection algorithm a threshold is imposed: zero crossings that do not start and end above an absolute value of 0.001 are not taken into consideration since they are caused by the background noise [2].

2.2 Energy

Short – time energy of the speech wave is defined as [1]:

$$E(n) = \frac{1}{N} \sum_{m=0}^{N-1} [w(m)s(n-m)]^2 , \tag{2}$$

where N is the number of samples and w is a window used for analysis. In our experiments we used *root mean square energy (RMSE)* [4]:

$$E(n) = \frac{1}{N} \sum_{m=0}^{N-1} s(m)^2 , \tag{3}$$

which is the square root of the average of the sum of the squares of the amplitude of the signal samples.

A frame with high energy corresponds to a voiced portion of speech signal, while one with low energy to an unvoiced region.

As stated in [2], unlike ZCR case, there are no standard estimates for energy values for voiced and unvoiced regions. Even if the same person utters the same phrase twice, there is a significant chance for energy levels to be different. As a consequence, we decided to normalize the energy values and to impose a lower threshold of 0.15, and a higher one of 0.85.

2.3 Teager energy

In modeling speech production, Teager proposed a new method for computing the energy of a signal [6]. If a signal sample is given as $x_i = A\cos(\Omega i + \varphi)$, where A is the amplitude of the oscillation, Ω is the digital frequency, and φ is the initial phase, the instantaneous energy E_i of the sample x_i is:

$$E_i = x_i^2 - x_{i+1}x_{i-1} \tag{4}$$

$$= A^2 \sin^2(\Omega)$$

$$= A^2\Omega^2 \tag{5}$$

From the above equation it can be seen that the energy of the signal depends not only on the amplitude but also on the corresponding frequency component, which means that this feature has the ability to track rapid changes in both amplitude and frequency, unlike RMSE, which takes into account only amplitude changes.

2.4 Frame-based Teager energy

Since Teager energy depends on amplitude and energy, instead of calculating the instantaneous energy for each signal sample using equation (4), we will use the algorithm described in [3] and [4].

First of all, speech signal is divided into frames. For each frame spectrum is calculated by applying an FFT:

$$X(w) = \sum_{i=-\infty}^{\infty} s_i e^{-jwi} . \tag{6}$$

The magnitudes obtained in this manner are weighted by the square of the corresponding frequency component:

$$f_i = w_i^2 X(w_i) . \tag{7}$$

Finally, the modified Teager energy is the square root of the sum of the weighted power spectrum:

$$T_i = (\sum_{k=1}^{K} f_k)^{1/2} . \tag{8}$$

What we obtained is called the *Frame-based Teager Energy Measure*.

3 Classification Criteria

Given ZCR and energy we need some thresholds to help us to take a decision regarding the character of a region of speech. The classical criterion is depicted in Table 1.

ZCR	Energy	Label
Low	High	Voiced
High	Low	Unvoiced
0	0	Silence

Table 1. Classification of speech signals after energy and zero crossing rate

Since in the real world background noise affects speech signal, a different classification scheme is proposed in [2] (Table 2).

ZCR	Energy	Label
approx. 0	approx. 0	Silence
High	Low	Unvoiced
Low	High	Voiced
approx. 0	High	Voiced
High	High	Voiced
Low	Low	Voiced
approx. 0	Low	Unvoiced
Low	approx. 0	Silence
High	approx. 0	?

Table 2. The classification scheme that we used in our experiments

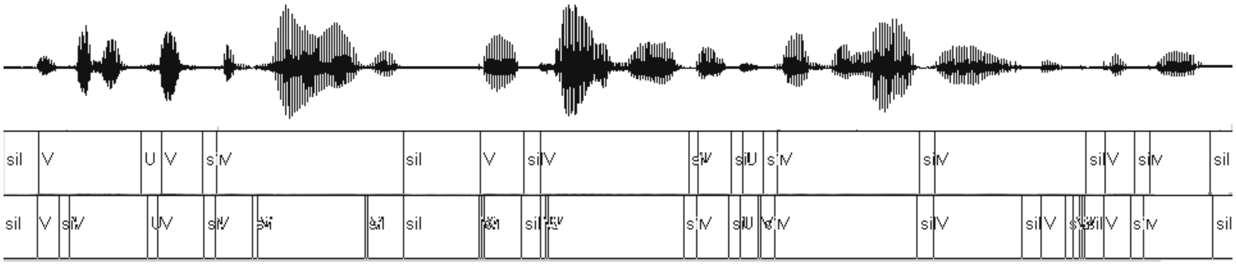

Figure 1. The waveform, the manual and the automatic transcriptions, for a file
where the percentage of classification correctness is 92.66 %.
The features used are ZCR and FTE

Apart from background noise there are other factors which make the classification difficult [1]. It is hard to distinguish between silence and sound when we are dealing with weak fricatives (/f/, /th/, /h/) or weak plosives (/p/, /t/, /k/). Other problems may be caused by the voiced fricatives that sometimes become devoiced, and by the nasals sound at the end of a word.

The last labeling decision from Table 2 refers to a situation which is very improbable to appear in the real world. The authors decided to label it with "?" instead of trying to fit it to one of those three other labels available.

4 Experiments and Results

For our experiments we used the same database [7] that was used by Greenwood and Kinghorn [2]. Our implementation is slightly different but the results obtained are within the same range of values (for the test concerning the energy and zero crossings, the only one that is common between our paper and theirs). The database consists of 10 phrases and their manually labeled transcriptions.

Speech was divided into 10 ms frames and no windowing was applied. Three sets of experiments were performed using as features zero-crossing rate and one of the energy measures (RMSE, Teager, and frame-based Teager). The results can be seen in Table 3.

For the low threshold of zero crossings we used a value of 12, while for the high threshold we used 50. All the energy values were normalized to 1 and thresholds of 0.15 and 0.85 were imposed. For comparing the similarity between two transcriptions we used the ratio between the number of correct matched frames and the total number of frames.

Features	Min	Max	Avg
ZCR + RMSE	53,90%	76,92%	**66,35%**
ZCR + TE	53,90%	76,92%	**66,17%**
ZCR + FTE	80,72%	94,08%	**87,38%**

Table 3. Results of the SUV classification

As it can be seen from the table, there is practically no difference between RMSE and instantaneous Teager energy. The major improvement is given by frame-based Teager energy: no classification has a less then 80% accuracy.

A representation of a waveform and the corresponding manual and automatic transcriptions can be seen in Figure 1. They were generated using *Matlab Auditory Demonstrations*.

5 Conclusion

A few methods for silence/unvoiced/voiced classification of speech signals were investigated in this paper. The features which we used were zero crossing rate and an energy measure (RMSE, Teager energy, and frame-based Teager energy). Some constrains were imposed to them in order to discriminate between the three categories of speech regions. The experiments were conducted on a manually labeled database.

Results proved that RMSE and Teager energy in conjunction with zero crossings have an accuracy of around 66%. Spectacular results are obtained for frame-based Teager energy (no classification has less than 80 percent accuracy) and this seems to be the solution that needs to be improved further.

References:

[1] Rabiner, L. R., Schafer, R. W., *Digital Processing of Speech Signals*, Prentice-Hall, Englewood Cliffs, 1978.

[2] Greenwood, M, Kinghorn, A., *SUVing: Automatic Silence/Unvoiced/Voiced Classification of Speech*, Undergraduate Coursework, Department of Computer Science, The University of Sheffield, UK, 1999.

[3] Gu, L., Zahorian, S. A., A New Robust Algorithm for Isolated Word Endpoint Detection, *IV-4161 International Conference on Acoustics, Speech, and Signal Processing*, Orlando, Florida, May 13-17, 2002.

[4] Ying, G.S., Mitchell, C. D., Jamieson, L. H., Endpoint Detection of Isolated Utterances Based on a Modified

Teager Energy Measurement, *Proceedings IEEE ICASSP-92*, 1992, pp. 732-735.

[5] Jabloun, F., Cetin, A., E., Erzin, E., *Teager energy Based Feature Parameters, for Speech Recognition in Car Noise, IEEE Signal Processing Letters*, Vol. 6, No. 10, 1999, pp. 259-261.

[6] Teager, H. M., Some observations on oral air flow during phonation, *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-28, 1980, pp. 599-601.

[7] *** http://www.ee.columbia.edu/~dpwe/classes/e6820-2001-01/matlab/MAD/data/swstracks/

[8] *** http://www.dcs.shef.ac.uk/~martin/MAD/docs/mad.htm