2007 IEEE International Conference on Signal Processing and Communications (ICSPC 2007), 24-27 November 2007, Dubai, United Arab Emirates

A NON-LINEAR OPERATOR BASED METHOD FOR HARMONIC FEATURE EXTRACTION FROM SPEECH SIGNALS*

Darren F. Kavanagh, Frank Boland

Department of Electronic and Electrical Engineering,
Trinity College Dublin, University of Dublin,
Dublin 2, Ireland.
Email: kavanadf@tcd.ie

ABSTRACT

An important pre-processing stage in speech recognition systems is that of extracting phonetically pertinent acoustic features from the speech signal. These features form the basis for discriminative classification and serve as cues for the identification of phonetic events in speech. The paper addresses this by presenting a novel method for the classification of harmonic (short-term periodic) and non-harmonic segments in speech signals. Classification is accomplished by proposing two new features derived from the non-linear Teager energy operator (TEO). The features proposed are the TEO-Weighted Harmonic Product (TEO-WHP*) and the TEO-Weighted Harmonic Sum (TEO-WHS*). Experiments are reported and discussed that demonstrate the effectiveness and the importance of these features as a valuable pre-processor for many speech systems.

Index Terms— Harmonic, feature, extraction, classification, Teager energy operator (TEO).

1. INTRODUCTION

Feature extraction and signal classification are a common pre-processor in many speech processing applications, such as: Speech enhancement, Speaker identification and Automatic speech recognition. The research literature on the role of features such as linear prediction coefficients, energy (Logarithmic or Euclidean), the spectrum, the cepstrum, zero crossing rate, duration and formants has been reviewed by Kotropolous [1]. More recently Gu and Rose [2] have introduced new features, based on harmonic cepstral coefficients. The existing pre-processing stages are considered to be suboptimal, thus advances in this area would have the scope for achieving major gains in the performance of such systems.

In this paper two new features are introduced, based on the Teager Energy Operator [3], that enable the classification of speech as harmonic or non-harmonic. The features are termed the TEO-Weighted Harmonic Sum (TEO-WHS) and the TEO-Weighted Harmonic Product (TEO-WHP) and they provide a simultaneous characterization of both pitch and formant content of a speech signal. This produces features which effectively extract the level of periodicity and aperiodicity content present in a particular frame using non-linear energy estimates. For comparison, two equivalent features are obtained using the Short-time Fourier Transform (STFT). These are obtained using essentially the same method; however, the magnitude of the short-time spectrum is used instead of the non-linear spectrum. Subsequently, to aid clarity these features have been named in a similar fashion. They are the Short-time Spectrum-WHS (StS-WHS) and the Short-time Spectrum-WHP (StS-WHP). These counterpart features are compared by means of experimental studies. The effectiveness and importance of the features as valuable pre-processors for speech systems is discussed and demonstrated using experiment results.

The paper is organized as follows. Section 2 details the background of Teager's energy Operator (TEO). An overview of the proposed system is presented in Section 3. The system description is presented in Section 4. Section 5 describes the experiments and discusses the results. Finally, Section 6 presents a conclusion and highlights future work.

2. BACKGROUND: NON-LINEAR SOURCES OF SPEECH SOUND EXCITATION

Speech processing systems have in general been based on a linear plane wave model for the airflow propagation along the vocal tract [4]. A classic example of this was presented by Schafer and Rabiner [5], using a linear source filter in the model. However, studies by Teager and Teager [6],[7] strongly suggest that non-linear processes are the primary source of sound excitation in the vocal tract during phonation. This contribution of non-linear excitation sources is something neglected by source filter theory [4]. The energy operators shown below in (1) and (2) were developed by Teager during his work on modeling speech production and were first introduced by Kaiser [3],[8]. Hence the operator is known as Teager's Energy Operator (TEO).

$$\Psi_c[x(t)] \triangleq (\frac{dx}{dt}(t))^2 - x(t)\frac{d^2x}{dt^2}(t) = [\dot{x}(t)]^2 - x(t)\ddot{x}(t), \quad (1)$$

where $\dot{x} = \frac{dx}{dt}$, and its discrete counterpart:

$$\Psi_d[x(n)] \triangleq x^2(n) - x(n-1)x(n+1), \quad (2)$$

for discrete-time signals $x(n)$, $\{n = 0, \pm1, \pm2, \pm3, ...\}$.

Before proceeding there is a clear requisite to distinguish what is meant by the term: *the energy of a signal*. Traditionally in signal processing literature [3], [9] the energy of a discrete time signal is defined as (3).

$$E_x \stackrel{def}{=} \sum_{n=-\infty}^{\infty} |x(n)|^2 \quad (3)$$

and $x(n)$ is called an energy signal if $[0 < E_x < \infty]$.

The non-linear energy operator $\Psi_{c|d}$ provides an estimate of the energy over time required to generate, in a certain sense, a signal [3]. A derivation exists in [3] that uses basic physics of motion for a simple spring and mass oscillator, to show that the

Manuscript received April 5, 2007. *This work was supported by the Irish Research Council for Science Engineering and Technology (IRCSET).

total energy in the system is proportional to both the amplitude and the frequency. This is very different to the energy measure in (3); where, from Parseval's relation all frequencies are treated uniformly. The operator Ψ_d is considered to be a high resolution energy estimator [4]. This property has provided additional motivation for investigating its use for obtaining features with high temporal resolution.

3. OVERVIEW OF THE PROPOSED SYSTEM

The non-stationary nature of speech can be modeled in terms of voiced and unvoiced excitations. Voiced excitation is periodic or quasi-periodic, it is produced by the vibration of the vocal folds whereas unvoiced sounds (noise-like) are aperiodic and is produced by forcing air past some constriction in the vocal tract [9]. Certain sounds such as voiced fricatives are a combination of both voiced and unvoiced speech. Feature extraction is used to capture these various acoustic characteristics of the signal. The main characteristics considered in this approach are the pitch and the formants. These are combined to effectively extract the four main features via the (STFT). This obtains features that extract the level of harmonic content present. These are outlined in Figure 1. Short frames of speech that are periodic in nature (i.e.

Figure 1. Overview of the features.

voiced) will tend to have the majority of their energy contained within the harmonic spectral components. As opposed to aperiodic speech which will tend to have its energy spread across the spectrum. The method in essence exploits this fact to have different energy ranges for the features to work on, thus forming the basis for the classification by comparing the features against predetermined threshold values. High temporal resolution is gained by using a sliding window technique which averages the overlapping speech frames to obtain a mean value.

4. SYSTEM DESCRIPTION

4.1. Feature Extraction

4.1.1. Pre-Filtering

The system has been tested using speech samples from the TIMIT data corpus. The sample frequency F_s is equal to 16 kHz for the speech samples, corresponding to a sample period T_s of 62.5-μs. The speech signals are pre-filtered using a low-pass filter with a cut-off frequency of 4 kHz.

4.1.2. Windowing and Frames (Short-terms)

The speech signal $x(n)$ is analyzed using symmetric windows w_n. This selects a frame (or short-term) of the speech signal labeled f_x, which spans N samples, N of 640 (40-ms duration) and 320 (20-ms duration) have been used during experimentation; assume an N of 320 herein. The first frame $f_1(n) = x(1:320)$ moving along by one sample so that the second frame $f_2(n) = x(2:321)$, this iterative process continues taken successive frames, hence we refer to it as a sliding. The frames $f_x(n)$ are obtained by simple multiplication process as in (4). The different window types that have been used are the basic Rectangular, the Hamming and the Kaiser windows.

$$f_x(n;m) = x(n)w(m-n) \quad (4)$$

4.1.3. Short-time Discrete Fourier analysis.

The Short-time Fourier transform (STFT) $F_x(k)$ of the frame $f_x(n)$ is obtained using the Fast Fourier Transform FFT. During experimentation different FFT point sizes were used, e.g. (512, 1024 and 2048). The absolute value is then obtained $|F_x(k)|$.

4.1.4. Nonlinear Energy Spectrum

The non-linear energy spectrum is obtained by computing the TEO for each of spectral components in $|F_x(k)|$. This may be achieved using each of the following techniques: (a) Band-pass filtering; (b) Obtaining the IFFT for each conjugate pair respectively and (c) Calculate the TEO directly from the spectrum. The latter two techniques have been used. The preferred method is (c) due to its reduced computational complexity. This is obtained by computing a energy estimate for each spectral component in $|F_x(k)|$ using the right-most term in (5).

$$\Psi_d[x(n)] \triangleq x^2(n) - x(n-1)x(n+1) = A^2\sin^2(\Omega) \quad (5)$$

Note: Ω must be less than $(\pi/2)$ which is equivalent to $(F_s/4)$, [3] . This gives us a non-linear energy spectrum $E(k)$ for the frame $f_x(n)$.

4.1.5. Spectrum Down-sampling

Using this non-linear energy spectrum $E(k)$ down-sampled versions $E_\alpha(k)$ are obtained. These are compressed replicas of the original; this is illustrated in Figure 2. The spectrum $E(k)$ is down-sampled by a factor α, for $\alpha = 1, 2, ...\beta$, where β is the number of down-sampled spectra. Consequently, β is also the number of harmonics (fundamental integer multiples) we wish to include. Note: $E_1(k) = E(k)$.

4.1.6. Spectral Peak Detection

The spectral peaks in each of down-sampled energy spectra $E_\alpha(k)$ are identified in the range $[2 \leq k \leq FFT\_size/4]$. Note: $(k = 1)$ has been omitted as this is dc. A one dimensional numerical gradient (directional derivative) $G_r(k)$ is computed for each $E_\alpha(k)$. Then using condition (6), we can determine the spectral peaks in $E_\alpha(k)$ through inspection of changes in sign. The peaks are arranged in the order of their magnitude and the dominant peaks are selected.

$$if[(Gr_x(k) < 0)\&(Gr_x(k-1) >= 0)] \quad (6)$$

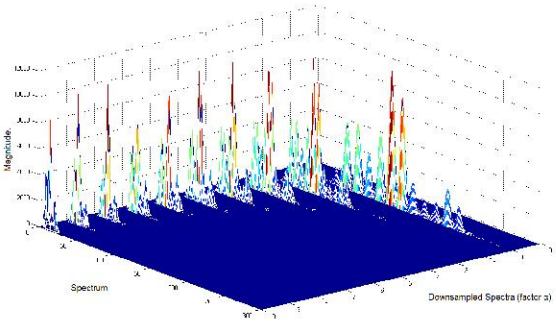

Figure 2. Down-sampled non-linear energy spectra for a 20-ms frame of voiced speech.

4.1.7. *Weighting Criterion*

A weighting scheme is used to emphasize the harmonic peaks and conversely de-emphasize the non-harmonic components. Assume a frame contains speech of a periodic nature. Since the original energy spectrum has been down-sampled, therefore the harmonic peaks, the formants in the down-sampled spectra will therefore align at corresponding spectral locations. Further, if we take the ideal case where there are harmonics at all the multiples of the fundamental, the first harmonic h_1 in $E_1(k)$ will line up with the second harmonic h_2 in $E_2(k)$ and so on. Accordingly, each spectra is assigned an associated weighted mask $W_\alpha(k)$. The down-sampled spectra $E_\alpha(k)$, are compared for aligning peaks. The locations (k) in $W_\alpha(k)$, are given an exponential weight based on the number of peaks (t) that aligned at that particular location as shown in (7), the weighting coefficients are contained in Table 1.

$$W_\alpha(k) = \begin{matrix} w_t, \text{ [according to the no. } (t) \text{ aligned]} \\ 1, \qquad \text{[otherwise for } t \leq 1\text{]} \end{matrix} \tag{7}$$

4.1.8. *TEO-WHS and TEO-WHP features*

The two new features are introduced in (8) and (9). They are the TEO-WHS* and the TEO-WHP*. The features are labeled $\Upsilon^{\dagger}_{whs}$ and $\Upsilon^{\dagger}_{whp}$ and are obtained by getting the weighted sum and weighted product for the down-sampled non-linear energy spectra $E_\alpha(k)$.

$$\Upsilon_{whs}(k) = \left[\sum_{\alpha=1}^{\beta} W_\alpha(k)[E_\alpha(k)]\right] \tag{8}$$

$$\Upsilon^{\dagger}_{whs} = \max\left\{\Upsilon_{whs}(k_p)\right\}$$

$$\Upsilon_{whp}(k) = \exp\left[\sum_{\alpha=1}^{\beta} \ln([E_\alpha(k)]^{W_\alpha(k)})\right] = \left[\prod_{\alpha=1}^{\beta} ([E_\alpha(k)]^{W_\alpha(k)})\right] \tag{9}$$

$$\Upsilon^{\dagger}_{whp} = \max\left\{\Upsilon_{whp}(k_p)\right\}$$

Where β is the number of down-sampled spectra and $[k_l \leq k_p \leq k_h]$ is the maximum range (bandwidth) in which the pitch resides. The maximum values in both $\Upsilon_{whs}(k)$ and $\Upsilon_{whp}(k)$ within the pitch range are identified to give $\Upsilon^{\dagger}_{whs}$ and $\Upsilon^{\dagger}_{whp}$.

4.1.9. *StS-WHS and StS-WHP features*

As aforementioned, these features are obtained using the same approach; the only exception is that the computation described in section 4.1.4 does not form part of the procedure.

4.2. Classification

4.2.1. *Sliding Window Technique*

A sliding window scheme is exploited to gain high temporal resolution. This is achieved by obtaining an average value for the overlapping frames; that is for a given percentage of overlap. As outlined in 4.1.2, the overlap is chosen to be the maximum to achieve the highest resolution possible. Other overlaps of 50%, 75%, etc. may be used. A trade-off exists here in the amount of processing versus temporal resolution required.

Table 1. Weighting Coefficients

$no.(t)$	1	2	3	...	...	β
w_t	$e^{(0)}$	$e^{(0.25)}$	$e^{(0.50)}$	...	...	$e^{[(\beta-1)\times 0.25]}$

4.2.2. *Threshold Stage*

$$T_{high} = [\bar{x} + c_1(s)] \,||\, [T_{pre\ \ high}]\,; \tag{10}$$

$$T_{low} = [\bar{x} + c_2(s)] \,||\, [T_{pre\ \ low}]\,; \tag{11}$$

To perform classification upper T_{high} and lower T_{low} energy thresholds are established initially. A section of frames at the start are assumed to be silence, that is a time duration in the region of 1000-ms. The sample mean $\bar{x}$ and standard deviation s for the features $\Upsilon^{\dagger}_{whs}$ and $\Upsilon^{\dagger}_{whp}$ of these frames is computed to estimate the energy of the background noise present. Alternatively, predetermined energy thresholds $T_{pre\ \ high}$ and $T_{pre\ \ low}$ for a given background noise level may be used. Classification is achieved by comparing the subsequent frames outside of the training period, against these threshold values, see equations (10) and (11), where $c_1 = 3$ and $c_2 = 2$. The upper and lower thresholds prevent erratic decisions and hence yield better classification.

5. EXPERIMENTS AND DISCUSSION

The system has been applied to an extensive set of speech utterances from the TIMIT corpus to confirm its consistency for various utterances of speech. An example is presented in Figure 3. In the first panel we have the speech waveform x(n) with the corresponding TIMIT phonetic symbols located above and below in the form of textures to aid clarity. The corresponding speech manner of articulation and phonetic symbols for these textures are tabulated in the appendix, see Table 2. The second panel shows the harmonic|non-harmonic classification for each of the features. Finally, the last two panels show the energy profiles for each of the features. From Figure 3. we can observe that the TEO-WHP and the TEO-WHS classification signals are similar in terms of overlap. This is also the case for the Short-time spectrum features but to a greater extent. Looking at the classification and feature profiles for each of the features along with the phonetic labels, we can see that the TEO based features are more selective in terms of their harmonic classification. That is they classify the speech containing strong voicing (i.e. vowel, semivowel and glide sounds) to be harmonic over the other phones of voiced speech. This is in contrast to the short-time spectrum features, which tends to classify all phones of voiced speech in general as harmonic.

Further, take the first two words in the utterance, "The emblem". Here we can see that the TEO-based features for the vowel sounds /iy/, /eh/, /ax/, and for the semivowel /l/ have classified these phones as harmonic (short-term periodic). We can see that this tendency holds true for the rest of the utterance. However, the TEO-WHS has classified the unvoiced fricative /s/ at the end of the word "acropolis" as being harmonic, whereas the TEO-WHP feature has classified it correctly. The TEO-WHP is more robust under the conditions of the high frequency content of the unvoiced fricative. Throughout the course of our experiments and analysis we have found these results to be consistent for an extensive assortment of speech utterances. This leads us to conclude that the TEO features, in particular the TEO-WHP, to be important features for the identification of speech containing strong harmonic content such as vowels, semivowels, etc. Hence, the study suggests these to be very valuable features for the improvement of existing pre-processing stages, for reliably identifying (or) distinguishing phones that are characterized by strong stable voicing.

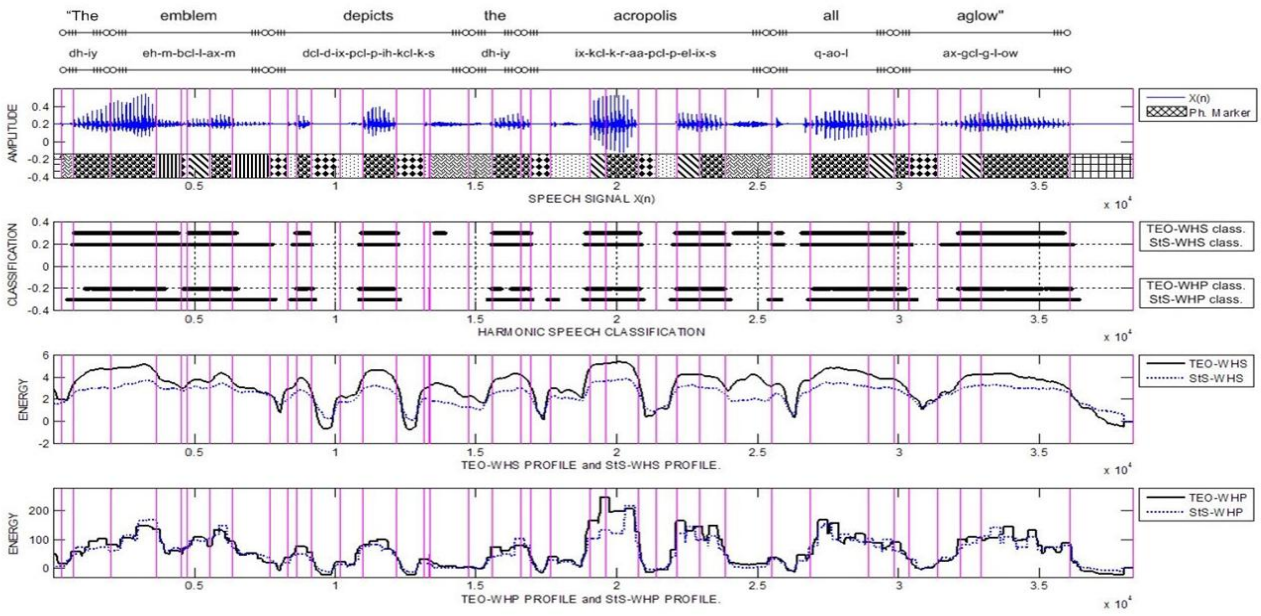

Figure 3. Harmonic speech classification and extracted energy features for a sample speech utterance. For a frame duration of 20-ms; $\beta = 10$; using a Kaiser window with $\alpha_w = 4$; the threshold coefficients $c_1 = 3$ and $c_2 = 2$.

6. CONCLUSION AND FUTURE WORK

In this paper we have introduced a novel method to perform harmonic speech classification by proposing two new TEO-based non-linear features; the TEO-WHS* and the TEO-WHP*. These features combine the pitch and formants simultaneously, to effectively extract the short-term harmonic energy present in the speech signal. The method has been applied to samples of speech from the TIMIT corpus to verify consistency for various utterances of speech. We conclude these to be valuable features for the identification of phones of speech that are characterized by strong stable voicing. The temporal resolution of the features is high, which is gained via a sliding window technique. The TEO-WHP feature was found to be more robust for working on speech consisting of all different phonemes. Future work will investigate the use of these new features as a pre-processor to enhance the ability of a modern Speech Recognition System.

7. APPENDIX

Table 2. Speech manner of articulation identifiers.

Speech manner	Texture marker	Phonetic Symbols (for TIMIT)
Stops		b, d, g, p, t, k, dx, q
Closure		bcl, dcl, gcl, kcl, pcl, tcl
Affricates		jh, ch
Fricatives		s, sh, z, zh, f, th, v, dh
Nasals		m, n, ng, em, en, eng, nx
Semivowels and Glides		l, r, w, y, hh, hv, el
Vowels		iy, ih, eh, ey, ae, aa, aw, ay, ah, ao, oy, ow, uh, uw, ux, er, ax, ix, axr, ax-h
Others		pau, epi, h#

8. ACKNOWLEDGEMENT

The authors would like to thank Dr. Naomi Harte for some fruitful discussions. D. F. Kavanagh would like to kindly acknowledge a research scholarship from the Irish Research Council for Science Engineering and Technology: funded by the National Development Plan.

9. REFERENCES

[1] C. Kotropoulos, "Speech Segmentation/Classification," *State of the Art in Speech Processing*, Khalid Daoudi, ed., September 20, 2004.

[2] Laing Gu, Kenneth Rose, "Perceptual Harmonic Cepstral Coefficients as the Front-end for Speech Recognition," Proc. ICSLP2000, October 2000.

[3] James F. Kaiser, "On a simple algorithm to calculate the 'energy' of a signal," *Proc. IEEE ICASSP-90*, Albuquerque, NM, pp. 381-384, Apr, 1990.

[4] John H. L. Hansen, Liliana Gavidia-Ceballos, James F. Kaiser, "A Nonlinear Operator-Based Speech Feature Analysis Method with Application to Vocal Fold Pathology Assessment," *IEEE Transactions on Biomedical Engineering*, Vol. 45, No. 3, March 1998.

[5] Ronald W. Schafer, Lawrence R. Rabiner, "Digital Representations of Speech Signals" *Proceedings of the IEEE*, Vol. 63, No. 4, April 1975.

[6] Herbert M. Teager, "Some observations on oral air flow during phonation," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, pp. 599-601, Oct. 1980.

[7] Herbert M. Teager, Shushan M. Teager, "A Phenomenological Model for Vowel Production in the Vocal Tract," *Speech Sciences: Recent Advances*, R. G. Daniloff, ed., College-Hill Press, San Diego, Calif.: pp. 73-109, 1983.

[8] James F. Kaiser, "On Teager's Energy Algorithm and Its Generalization to Continuous Signals," Proc. 4th IEEE Digital Signal Processing Workshop, Mohonk (New Paltz), NY, Sep. 1990.

[9] John R. Deller, John H. L. Hansen, John G. Proakis, *Discrete-Time Processing of Speech Signals*. IEEE Press, 1993, 445 Hoes Lane, P.O. Box 1331, Piscataway, NJ 08855-1331.