A MEASURE OF APERIODICITY AND PERIODICITY IN SPEECH

Om Deshmukh and Carol Espy Wilson
Speech Communication Lab, ECE Dept.
University of Maryland, College Park
{omdesh, espy}@Glue.umd.edu

ABSTRACT

In this paper, we discuss a direct measure for aperiodic energy and periodic energy in speech signals. Most measures for aperiodicity have been indirect, such as zero crossing rate, high-frequency energy and the ratio of high-frequency energy to low-frequency energy. Such indirect measurements will usually fail in situations where there is both strong periodic and aperiodic energy in the speech signal, as in the case of some voiced fricatives or when there is a need to distinguish between high frequency periodic versus high frequency aperiodic energy. We propose an AMDF based temporal method to estimate directly the amount of periodic and aperiodic energy in the speech signal. The algorithm also gives an estimate of the pitch period in periodic regions.

1. INTRODUCTION

Periodic sounds are produced by quasi-periodic lateral movements of the vocal folds. Aperiodic sounds are mainly produced by creating turbulence in the flow of air through the vocal tract. Turbulence is created when there is rapid airflow through a narrow constriction. The amount of turbulence and hence the amount of aperiodicity is increased if there is an obstacle in the main pathway that is downstream from the constriction. As the constriction narrows, the transglottal pressure decreases and the vocal fold vibration ceases. But vocal fold vibration can be maintained by active expansion of the vocal tract volume, which inhibits the buildup of intraoral pressure. This latter situation gives rise to what are called voiced consonants, which may exhibit both periodic and aperiodic energy. [1]

In previous work, algorithms have been developed to directly measure if the speech signal is periodic and a binary decision is made about whether the speech signal is voiced or unvoiced [2]. If the signal is judged unvoiced, then indirect measures such as zero crossing rate and the ratio of low- to high-frequency energy are used to determine if the signal contains noise. In [3], estimates of simultaneous voiced and turbulence-noise components in the speech signal are obtained, but the performance of the system relies on accurate estimates of the pitch period. However, pitch estimation is a difficult task that is prone to errors (pitch doubling and pitch halving).

In this paper, we discuss a system that first detects periodic and aperiodic components, and then gives an estimate of pitch period in the periodic region. The performance of the system is evaluated using a speech-like synthetic database and two speech databases that have EGG data recorded simultaneously.

This system can be used in a task such as segmentation into voiced and unvoiced regions; the recognition of regions where both excitation components exist – e.g. in a breathy vowel [1] or a voiced fricative. It will also replace the ESPS [8] tools and the indirect aperiodicity measures that were used in [7].

2. SYSTEM DISCRIPTION

Fig. 1 depicts the various stages of the signal processing involved in the analysis. The analysis starts with the speech signal being passed through a 60-channel auditory gamma-tone filterbank with characteristic frequencies (CFs) based on physiological data [4]. The CFs of the model are roughly linearly spaced at low frequencies and logarithmically spaced at higher frequencies, with bandwidths that are approximately constant-Q. The temporal envelopes $e_i(t)$ of the individual channels are obtained by the function:

$$e_i(t) = | x_i(t) + j \cdot H\{x_i(t)\} |$$

where $x_i(t)$ is the input signal, and $H\{x_i(t)\}$ is the Hilbert transform of the input signal

If the channel output is not silent, the temporal envelope is analyzed for periodicity and aperiodicity. To do so, we use the short-time average magnitude difference function (AMDF), which is defined as:

$$\gamma_n(k) = \sum_{m=-\infty}^{\infty} | x(n+m)w(m) - x(n+m-k)w(m-k) |$$

Where $x(n)$ is the input signal, k is the lag and $w(m)$ is the window. In our case, $w(m)$ is a 20ms rectangular window. For periodic sounds, the AMDF function usually attains local minima (referred as dips hereafter) at lags roughly equivalent to the pitch period and its integer multiples (see Fig. 2). In the case of aperiodic sounds, the dip locations are random. This is used as a basis for further analysis. If the windowed signal over which the AMDF is computed is such that there is a considerable monotonous energy change within the window, then the temporal envelopes may be tilted before the AMDF is computed. The AMDFs of the non-silent channel are computed at a rate of 10ms. The location of valid dips and their respective strengths are estimated by computing the convex hull of the AMDF. Any decision about the periodicity or aperiodicity is

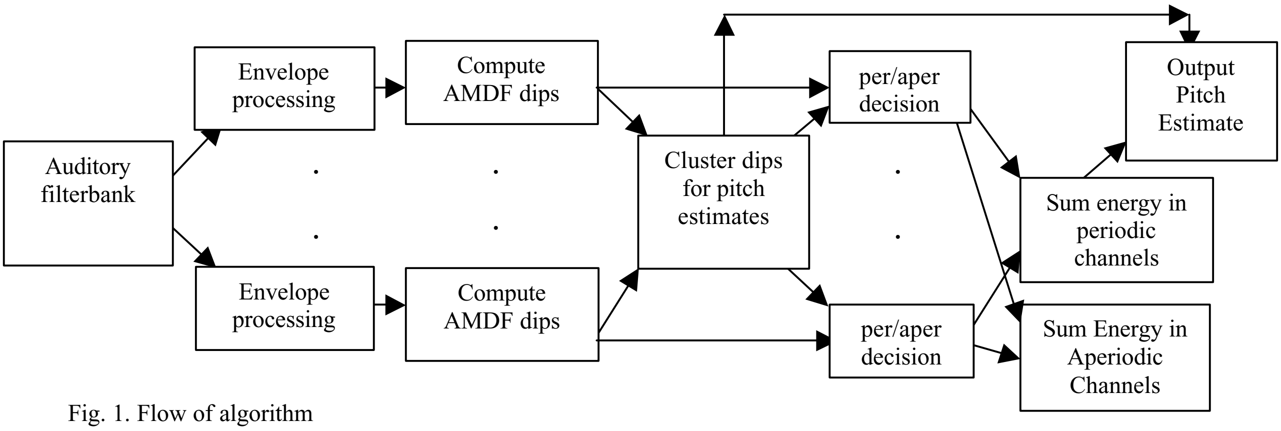

Fig. 1. Flow of algorithm

deferred until the next stage. Once the AMDF dips are computed on all the channels over the entire signal, a summary periodicity confidence is computed at a frame rate of 2.5ms. To compute this, all the dips in all the channels that are within 10ms of the particular frame are added point-by-point. In a typical periodic frame, the dips will cluster at multiples of the pitch period. On the other hand, in a typical aperiodic frame, the dips will be randomly scattered over the entire range (see fig 3). Notice that for the aperiodic frame, the depths of the dip are considerably smaller than the depths of the dips in the periodic frame.

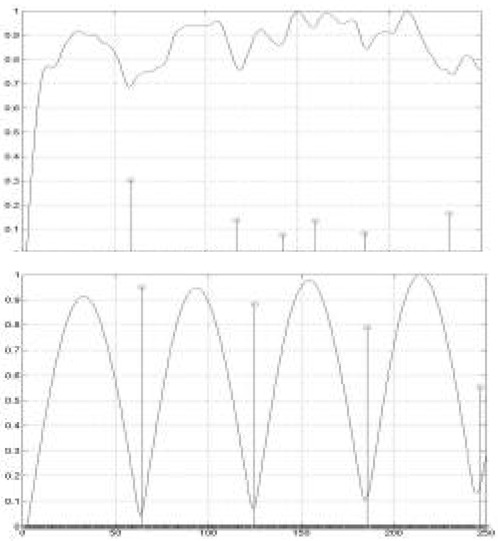

Fig 2. *Top*. AMDF of an aperiodic signal
Bottom AMDF of a periodic signal

As can be seen in Fig. 3, for periodic signals the dips cluster at T_o (fundamental period) and its multiples. A weighted sum of the strengths of all the dips that lie in the neighborhood of these locations is computed and the maximum value is the summary periodicity confidence. The dips are classified as either "*within cluster*" or "*outside cluster*" based on an exponential curve that is fitted on each side of the cluster centroid. For an aperiodic frame, the exponential curve will not fit and then the "*within cluster*" boundary is fixed to a nominal small value. Note that this confidence is just an estimate of whether the frame is periodic or aperiodic and the actual decision is deferred until the next step. This confidence value is used as a guide for further analysis.

In the next step, each channel is analyzed for periodicity and aperiodicity. If all the dips in a given channel fall in the "*within cluster*" range, then that channel is called periodic. Otherwise, it is called aperiodic. As the analysis advances, the cluster locations from the previous frames are used to decide where to form the clusters for the next frames. Thus, the system adapts to the speaker characteristics very rapidly.

The proportion of the periodic energy in the frame is obtained by taking the ratio of the sum of energies in all the periodic channels and the total energy in that frame. The proportion of the aperiodic energy is obtained in a similar way.

A pitch estimate is given in those frames where the periodic energy proportion is above a certain threshold. The centroids of the clusters obtained in the above-mentioned way are the pitch estimates for the given frame and their tightness measures (the weighted sum of all the dips that lie within a certain neighborhood of the centroid) are the confidences of those estimates being the pitch periods. We have incorporated memory into the pitch tracker to obtain smooth pitch contours

3. DATABASE

The system was evaluated on two speech databases that had the electroglottograph (EGG) data recorded simultaneously and on a database of synthetic signals. MOCHA database [5] consists of 460 utterances each spoken by two speakers. The second speech database consists of 50 utterances spoken each by one male and one female [6]. In the present work, EGG output was used to demarcate periodic and aperiodic regions.
The synthetic speech-like signals database is the same as the one used in [3]. The signals are the outputs from a 50-pole LPC (linear Predictive Coding) synthesis filter when it is excited by a pulse train that is corrupted by Gaussian white noise (GWN). Pulses of frequency 120Hz, 131Hz and 200Hz were used. The SNR varied from ∞ to –5dB. The pitch period and amplitude of this pulse was perturbed by specified degrees of jitter (0 to 5%

fluctuation in the pitch period) and shimmer (0 to 1.5dB fluctuation in the amplitude of the signal).

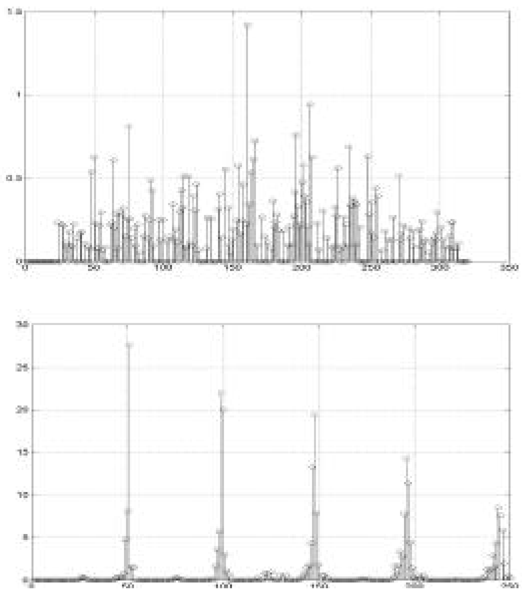

Fig. 3. Clustering of dips across channels.
Top: Aperiodic frame. Bottom: periodic frame

4. RESULTS

4.1 Evaluation on the Synthetic data

To evaluate the performance of our system we compared the output SNR with the SNR of the input signal. We define output SNR as:

$$SNR = 10 * \log_{10}(v/u)$$

where ν=periodic energy and *u* is the aperiodic energy calculated by our detector. Fig. 4. shows the actual SNR versus the output SNR for the pulse with frequency 131Hz and no jitter or shimmer. Notice that the difference between the actual SNR and the output SNR increases as the actual SNR is increased. This can be attributed to the fact that our algorithm makes a binary decision between periodicity and aperiodicity for each non-silent channel. As a result, if a particular channel that had both periodic and aperiodic energy was called periodic, then the aperiodic energy from that channel will contribute towards periodic energy. As the SNR increases the periodic component increases and the above mentioned effect dominates.

The same trend was observed for pulses with other frequencies. For a given SNR there was minimal deviation in the output SNR at different degrees of jitter and shimmer proving that the algorithm is robust to deviations in the pitch period and the pulse amplitude. Fig 5. shows the output of our system for a pulse at 131 Hz at ∞ SNR with 3% jitter and 1dB shimmer.

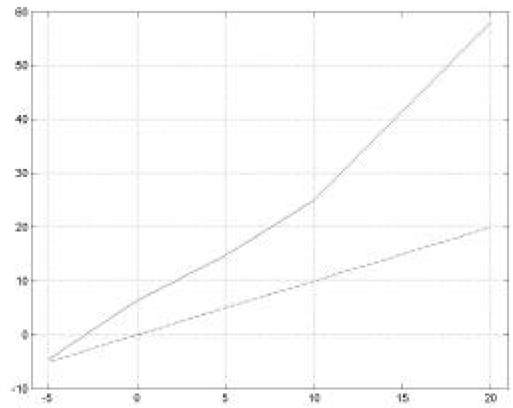

Fig.4. (*solid line*) output SNR by our system
(*dashed line*) ideal output SNR

4.2. Evaluation on the natural speech databases

All the comparisons were made on a frame basis at a frame rate of 2.5ms. Table 1 shows the performance of our system on the MOCHA database. *'per_accur'* is the periodicity accuracy and is defined as the ratio of the number of non-silent frames that have the periodic energy above a predetermined threshold and the corresponding EGG output is non-zero to the total number of frames that have a non-zero EGG output. 'aper_accur' is the aperiodicity accuracy and is defined as the ratio of the number

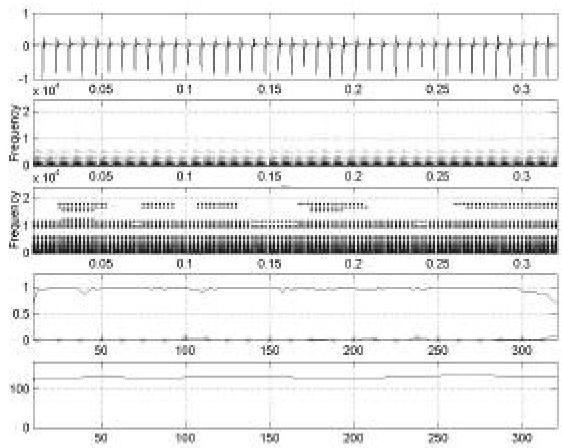

Fig. 5. output of the our system on a pulse of 131Hz at ∞ SNR and with 3% jitter and 1 dB shimmer.*Top panel;* the pulse signal. *Second:* spectogram of the signal. *Third*: the channels that were detected periodic *Fourth: solid line:* periodic energy ,*x line :* aperiodic energy. *Bottom panel:* pitch estimate, notice that the pitch estimate is not a constant indicating that, the detector is able to track the pitch changed caused by the jiter.

of non-silent frames that have the aperiodic energy above a predetermined threshold and the corresponding EGG output is zero to the total number of non-silent frame that have zero EGG output.

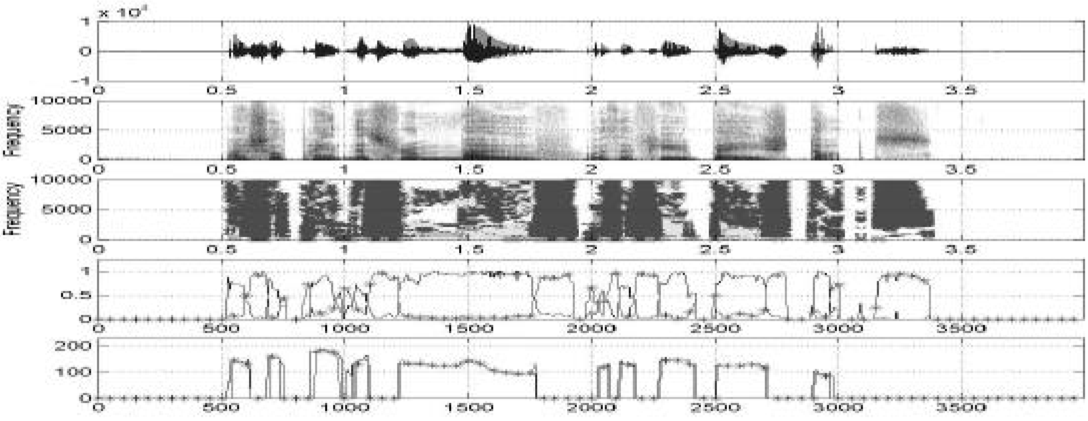

Fig 6. output of our system on “I shall paint this room mauve with a few beige dots.”
Top panel: the waveform. *Second*: spectogram. *Third*: dark spots are the locations that were judged as aperiodic and the light ones are the ones that were judged as periodic.*Fourth*: solid line: periodic energy. *x line*: aperiodic energy. *bottom panel: solid line*: pitch detected by our system. *X line*: pitch estimated from the EGG data.

	Male	Female	Overall
per_accur	0.97	0.93	0.95
aper_accur	0.94	0.89	0.90

Table 1. Performance on MOCHA database

	Male	Female	Overall
per_accur	0.93	0.87	0.90
aper_accur	0.91	0.92	0.92

Table 2. Performance on DB2 database

The phonetic transcription of the MOCHA database was used to evaluate the performance of our system on the sounds that can exhibit strong periodicity and strong aperiodicity. (voiced fricatives and voiced stops). 21.6% of the total frames of these sounds had both high periodic and high aperiodic energy whereas only 0.1% of the frames of the highly aperiodic sounds (voiceless fricatives and voiceless stops) show high periodic and high aperiodic energy. And only 6% of the frames of the highly periodic sounds (vowels, sonorant consonants) show both high periodic and high aperiodic energy. Possible reasons for the high aperiodic energy in the 6% of the periodic frames are breathiness and the transitions between periodic and aperiodic sounds.
Fig 6. shows the output of our system on an utterance from DB2.

5. DISCUSSION

We have presented an algorithm that gives direct estimates of periodic and aperiodic energy in the speech signals and have demonstrated its efficiency by evaluating it on speech-like synthetic data and on natural speech. Our future work is to extend this algorithm to give accurate estimates of pitch period and to use the periodic & aperioidic energy parameters in our Event Based Speech (EBS) recognizer.

REFERENCES

1.] K.N. Stevens, *Acoustic Phonetics*, MIT Press, Cambridge, Mass., 1998.
2. Zhang et. al. ‘Pitch estimation based on circular AMDF’ ICASSP-02, pp I-341-344
3. Jackson P.J. et. al. ‘Pitch-Scaled Estimation of Simultaneous Voiced and Turbulence-Noise Components in Speech’ IEEE Trans. Speech and Audio Processing, vol. 9, No. 7, October 2001
4. L.H. Carney, “A model for the responses of low-frequency auditory-nerve fibers in cat,” *J. Acoust. Soc. Am.*, Vol. 93, pp. 401-417, 1993
5. Wrench, A.A., "A multichannel/multispeaker ine articulatory database for continuous speech recognition ere research", In Phonus, Research Report No. 4, Institute e of Phonetics, University of Saarland, In press, 2000.
6. Bagshaw, P. C. et.al. ‘Enhanced pitch tracking and the processing of F0 contours for computer and intonation teaching’ Proc. European Conf. on Speech Comm. (Eurospeech), pp. 1003-1006
7. Deshmukh o. et. al. ‘Acoustic –phonetic speech parameters for speaker-independent speech recognition’ ICASSP-02 pp. I-593-596
8. ESPS 5.3.1, Entropic Research Laboratory, Inc., 1999.