Detection of Periodicity and Aperiodicity in Speech Signal Based on Temporal Information

Om Deshmukh and Carol Espy-Wilson
University of Maryland College Park, College Park, USA
E-mail: omdesh@Glue.umd.edu, espy@Glue.umd.edu

ABSTRACT

In this paper, we discuss a direct measure for the proportion of periodic and aperiodic components in speech signals. Further, in the periodic regions, we estimate the pitch period. This method is particularly useful in situations where the speech signal contains simultaneous periodic and aperiodic energy, as in the case of breathy vowels and some voiced obstruents. The performance of this algorithm was evaluated on three different natural speech databases that have simultaneously recorded EGG data. The results show excellent agreement between the periodic/aperiodic decisions made by the algorithm presented here and the estimates obtained from the EGG data. To evaluate the efficiency of this algorithm in predicting pitch, reference pitch values were obtained from the EGG data using a simple peak-picking based algorithm. The gross error rate in pitch prediction was 6.1% for male subjects and 12.5% for female subjects.

This work was supported by NIH grant #1 K02 DC00149-01A1 and #BCS-9729688.

1. INTRODUCTION

In this paper, we discuss a direct measure for aperiodic energy and periodic energy in speech signals. The purpose of this analysis is to determine if the excitation signal was periodic, if it consisted of turbulent noise, or if there were simultaneous strong periodic and turbulent sources. Further, in periodic regions, we want to estimate the pitch period. 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. The system presented here extracts the proportions of periodic and aperiodic energy in the speech signal based on temporal information. We define the temporal information as the envelope of the output of a 60-channel gamma tone auditory filter bank. The structure of the system is very similar to a pitch detection algorithm, and includes estimation of the pitch of the primary periodic component of the signal.

This system can be used in tasks such as segmentation of speech signals into voiced and unvoiced regions; the detection of regions where both voiced and unvoiced components co-exist – e.g. in breathy vowels or voiced fricatives. The system is also capable of distinguishing strident fricatives from non-strident fricatives and voiced ones from their unvoiced counterparts based on the strength of aperiodicity.

2. METHOD

Fig. 1 depicts the various stages of the signal processing involved in the analysis. The analysis filterbank was a 60-channel auditory gamma-tone filter bank [1] with channel Characteristic Frequencies (CFs) based on the ERB scale (Equivalent Rectangular Bandwidth, as defined by [2]). The temporal envelopes $e_i(t)$ of the individual channels above 250 Hz 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[3]. For channels with CFs below 250Hz the channel output is used directly for further analysis.

The temporal envelope in each channel was analyzed for periodicity, aperiodicity and pitch. This system classifies the signal in every channel as silence, periodic, or aperiodic. The raw pitch estimates in each band were produced using 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, it is a 20ms rectangular window [4]. For periodic sounds, the AMDF function usually attains local minima (referred to as dips hereafter) at lags roughly equivalent to the pitch period and its integer multiples. If the signal is aperiodic, the AMDF waveform will not show such evenly spaced dips. The AMDF is computed for each non-silent channel over a 20ms window and at a rate of 5ms.

The beginning of the utterance is established by starting analysis in the frame whose total energy is within 1.5% of the maximum total energy computed across all of the frames in the utterance. For any given non-silent frame

ISBN 1-876346-nn-n © 2003 UAB

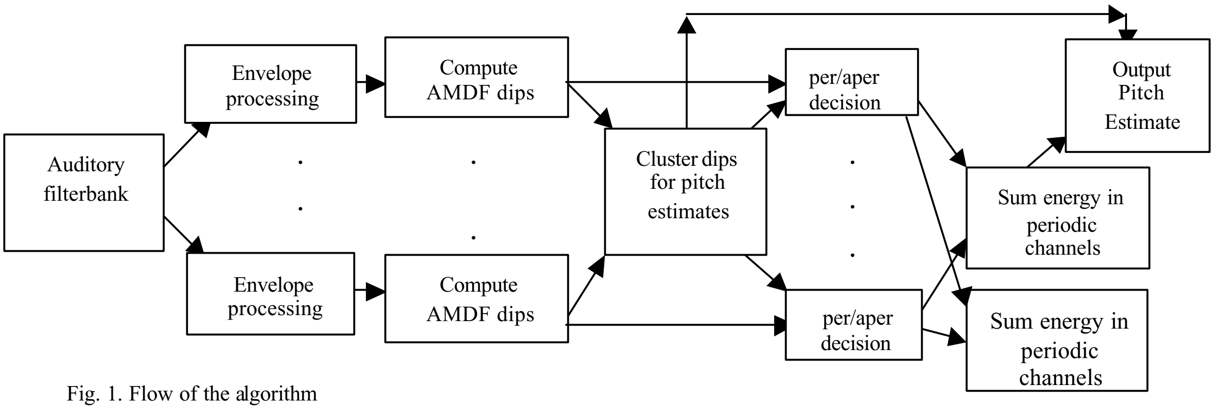


Fig. 1. Flow of the algorithm

within the utterance, a channel within that frame is considered silent if its energy is at least 45 dB down from the maximum channel energy that has been computed up to that point, including the channel energies in the present frame. If the channel is classified as silent, then no AMDF waveform is computed.

The decision regarding periodicity is based on the location and strength of the dips occurring in the AMDF waveform. These dips are found by computing the convex hull of the AMDF and accepting only those dip locations that have strength greater than a pre-determined threshold. The dip strength is the confidence of that dip location being the pitch period at that instance. Figure 2(a) shows the AMDF and the dips for a typical periodic and a typical aperiodic channel. Any decision of periodicity or aperiodicity of the channel is deferred until the next stage. The summary measure of periodicity across all the channels is computed at a frame rate of 2.5ms. All the channel estimates that were computed within 10ms of the frame contribute towards the decision of periodicity and aperiodicity for a particular frame. A modified histogram of all these pitch estimates *across all the channels* is computed. Part (b) of Fig. 2 shows that frames corresponding to periodic regions exhibit tight clusters at the pitch period and its integer multiples whereas the frames corresponding to aperiodic regions are more likely to show a uniform distribution of the dips. Also notice that the range of values for the periodic region is very high (0-28) as compared to that for the aperiodic region (0-1.5).

When tight clusters are formed, exponential curves are fitted on each side of the cluster to classify the dip locations as *within-cluster* dip locations or *spurious* dip locations. A weighted sum of the strengths of all the dips within a small neighborhood of the centroids of the clusters is computed and the maximum value is the summary periodicity confidence. The corresponding centroid is the pitch period estimate for that frame. If two or more clusters are comparable, then the one that yields a pitch period that is closer to the previous pitch period estimates is chosen.

In the case of aperiodic regions where there are no tight clusters, the summary periodicity confidence is low. If previous frames have been judged to be periodic, then the centroids from previous frames are used to form cluster regions. If we are at the beginning of an utterance so that there are no previous frames, then the dip locations with the maximum strength are taken as the centroids. Since the clusters are not prominent, there is no curve fitting and default values of 0.5ms around the centroids are used to define the cluster regions.

These centroid locations are used to analyze the channels for periodicity and aperiodicity. If all the dips in a channel fall in the within-cluster range then that channel is classified as periodic. Otherwise, it is called aperiodic. The proportion of periodic energy is obtained by taking the ratio of the sum of the energies in the periodic channels and the total energy in the frame. The proportion of aperiodic energy is obtained in a similar way.

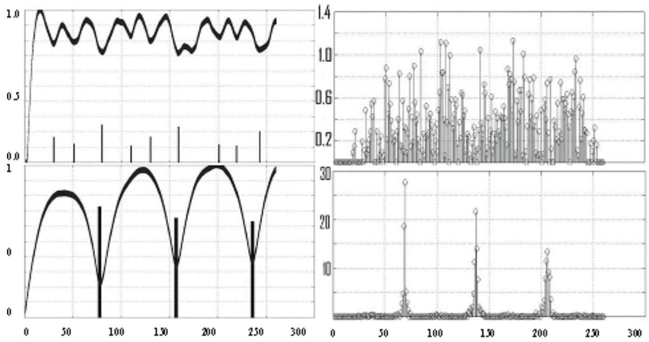
3. DATABASE

Three different corpora that consist of simultaneously recorded acoustic and electroglottograph (EGG) data were used to test the algorithms. The MOCHA [5] database consists of 460 utterances, each spoken by two speakers (one male and one female). The MOCHA database has been hand transcribed. A subset of 20 sentences (10 from each speaker) was used in the development of the algorithm. The second speech database, DB2, consists of 50 utterances spoken each by one male and one female [6]. The third database, DB5, consists of one utterance spoken by 5 males and 5 females [7].

The EGG data was used to demarcate the periodic and aperiodic regions and to compute the pitch values in periodic regions. The EGG data exhibits strong periodic fluctuations during vocalized sounds with the period equal to the pitch period of the speaker. A peak-picking algorithm was implemented on the band-pass filtered EGG data to find the locations of these peaks. The average value of the gaps between consecutive peaks over a period of 10ms is the pitch estimate at that location. The pitch estimates were computed every 2.5ms. The aperiodic regions are marked by the absence of any such regularly spaced peaks.

4. RESULTS

A. *Periodicity and Aperiodicity detection:*



a) b)

Fig. 2. Part (a) shows the AMDF and the prominent dips for a typical aperiodic channel (top) and for a typical periodic channel (bottom). Part (b) shows the AMDF dips clustered across all the channels in a typical aperiodic frame (top) and a typical periodic frame (bottom). Notice that the maximum value of the dip strength over the range of dip locations is 1.4 in the aperiodic frame whereas it is 28 in the case of the periodic frame.

The periodic and aperiodic measures were evaluated using the three natural speech databases. All the comparisons were made on a frame basis at a frame rate of 2.5ms. We define the periodicity accuracy as the ratio of the number of non-silent frames that have both the proportion of periodic energy no less than 0.3 (i.e., at least 30% of the energy in the frame must be periodic) and the corresponding EGG output is non-zero, to the total number of frames that have a non-zero EGG output. Similarly, the aperiodicity accuracy is defined as the ratio of the number of non-silent frames that have the proportion of aperiodic energy no less than 0.3 and the corresponding EGG output is zero, to the total number of non-silent frame that have zero EGG output. The results are tabulated in Table 1. An example of the outputs from these measures is shown in Fig. 3.

One cause of the less than perfect periodicity and aperiodicity accuracy is the boundary problem. In transition regions between adjacent sounds that differ in their manner of articulation, the frame where the switch between periodicity and aperiodicity occurs based on our algorithm may be offset from the frame where the switch occurs based on the EGG output. These scenarios are manifested in Fig. 3 around 400ms where our periodicity detector remains on for a little longer than the EGG does, and around 120ms where the EGG is on for about 5 frames longer than our periodicity estimate.

Table 2 shows results for the percentage of frames in the different broad classes that showed only strong periodicity, strong aperiodicity, or both strong aperiodicity and periodicity. For these results, the EGG data was not used as a reference. As expected, a much larger percentage of the sounds exhibiting both strong periodic and aperiodic components are voiced obstruents. Further, a large percentage of the voiced obstruents show only strong periodicity. Altogether, over half of the voiced obstruents

Table 1: Performance of Periodicity and Aperiodicity Measures

		Mocha	
	Male	Female	Overall
Per. Accuracy	95.6	92.1	93.7
Aper. Accuracy	94.0	89.0	90.0
		DB2	
Per. Accuracy	90.9	86.6	88.8
Aper. Accuracy	94.1	91.8	92.7
		DB5	
Per. Accuracy	90.9	91.4	91.2
Aper. Accuracy	84.6	85.5	85.0

Table 2: Percentage of frames in different broad classes where only strong periodicity was detected, strong aperiodicity was detected and both strong periodicity and aperiodicity were detected. Numbers in parenthesis show the total no. of frames in each category.

	only strong Periodic energy	only strong Aperiodic energy	strong Periodic and Aperiodic energy
Sonorants (540501)	83.66	9.75	6.88
Voiced obstruents (116117)	33.29	44.62	22.09
Unvoiced obstruents (199403)	3.80	96.09	0.11

exhibit strong periodicity. This finding supports previous studies that show that voiced obstruents can be lenited so that they are realized as sonorant consonants [8]. The small percentage of aperiodic sounds that show strong periodic energy and the small percentage of periodic sounds showing strong aperiodicity are probably due to boundary placement between sonorants and obstruents.

B. Pitch Detection:

The estimates from the pitch detect or were compared with the EGG-derived pitch values on a frame basis. A simple peak-picking algorithm was used to estimate the pitch from the EGG data. A temporal tolerance was incorporated in the pitch matching procedure. The EGG-derived pitch value at each frame was compared with the pitch estimates of our algorithm obtained over a small range of frames in the vicinity of the particular frame and the minimum deviation was chosen as the offset value. Using the standard established in previous studies [6], the pitch values were said to be in agreement if the offset value was less than 20% of the EGG-derived pitch value.

The gross errors were split into three different categories. The *halving* errors are defined as the instances where the pitch was detected to be within 20% of half of the pitch value given by the EGG data. The *doubling* errors are the instances where the pitch was detected to be within 20% of twice the actual pitch. Finally, the error instances that did not fit either of the above criteria ar e grouped as *others.*

 ISBN 1-876346-nn-n © 2003 UAB

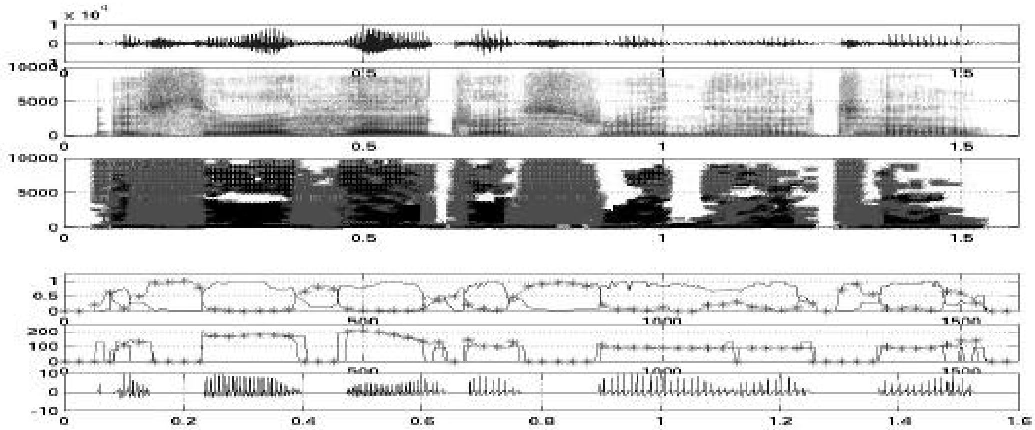


Fig. 3. (a): time waveform of the utterance “Is there a hairdresser in the hotel?”. (b): Spectrogram of the utterance. (c): periodic regions (as detected by our algorithm) are dark and the aperiodic regions are light. (d): The proportion of the periodic energy and aperiodic energy (marked with ‘x’). (e): pitch detected by our system and the pitch estimate according to the EGG data (marked with ‘x’). (f): The peaks found in the EGG data.

Table 3: Gross errors in pitch prediction

	Half	Double	Other
	Mocha		
Male	3.54	0.35	2.08
Female	12.32	0.02	0.26
Overall	9.10	0.05	0.37
	DB2		
Male	3.08	0.73	3.25
Female	11.12	0.02	0.15
Overall	8.34	0.13	0.64
	DB5		
Male	4.62	0.32	1.99
Female	12.08	0.02	0.24
Overall	8.81	0.09	0.65

Table 3 gives the details of the gross errors for the three databases. The results are given separately for males and females since the performance was consistently higher for male speakers.

5. DISCUSSION

We have presented a novel and efficient method to calculate direct measures of periodic and aperiodic energies in a speech signal that can distinguish high frequency periodic energy from high frequency aperiodic energy. The system also outputs a pitch estimate in regions that are judged to be periodic. One application of the periodicity/aperiodicity measures and pitch will be in our speech recognition algorithms [9,10]. These parameters also form a part of a landmark detection system [11] where the main emphasis is broad classification of speech signals using strictly temporal cues.

REFERENCES

[1] R. D. Patterson, “A pulse ribbon model of peripheral auditory processing,” in *Auditory Processing of Complex Sounds*, W. A. Yost and C. S. Watson, Erlbaum, New Jersey, 1987.

[2] B. R. Glasberg, and B. C. J. Moore, “Derivation of auditory filter shapes from notched-noise data,” *Hear. Res.*, vol. 47, pp. 103-138, 1990.

[3] Oppenheim, and R. Schafer, “Discrete-time Signal Processing,” Prentice Hall, Englewood Cliffs, NJ 1978.

[4] L. Rabiner, and R. Schafer, “Digital Processing of Speech Signals,” Prentice-Hall, Englewood Cliffs, NJ, 1978.

[5] http://www.cstr.ed.ac.uk/artic/

[6] P. Bagshaw, “Automatic prosody analysis,” Ph. D. thesis. University of Edinburgh. Scotland, 1994. http://www.cstr.ed.ac.uk/~pcb/fsa_eval.tar.gz

[7] ftp://ftp.cs.keele.ac.uk/pub/pitch/Speech

[8] C. Espy-Wilson, “Acoustic measures for linguistic features distinguishing the semi-vowels /w y r l/ in American English,” *J. Acoust. Soc. Am.*, vol. 92, pp. 401-417, 1993.

[9] O. Deshmukh, C. Espy-Wilson, and A. Juneja, “Acoustic-phonetic speech parameters for speaker-independent speech recognition,” in Proc. IEEE-ICASSP, 2002, 593-596

[10] A. Juneja, C. Espy-Wilson, “Segmentation of continuous speech using acoustic-phonetic parameters and statistical learning,” in Proc. ICONIP, 726-730, 2002,

[11] A. Salomon, C. Espy-Wilson, O. Deshmukh, “Detection of speech landmarks from temporal information,” *J. Acoust. Soc. Am.*, in revision