Use of Temporal Information: Detection of Periodicity, Aperiodicity, and Pitch in Speech

Om Deshmukh, Carol Y. Espy-Wilson, Ariel Salomon, and Jawahar Singh

***Abstract*—In this paper, we present a time domain aperiodicity, periodicity, and pitch (APP) detector that estimates 1) the proportion of periodic and aperiodic energy in a speech signal and 2) the pitch period of the periodic component. The APP system 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 the APP system was evaluated on synthetic speech-like signals corrupted with noise at various levels of signal-to-noise ratio (SNR) and on three different natural speech databases that consist of simultaneously recorded electroglottograph (EGG) and acoustic data. When compared on a frame basis (at a frame rate of 2.5 ms) the results show excellent agreement between the periodic/aperiodic decisions made by the APP system and the estimates obtained from the EGG data (94.43% for periodicity and 96.32% for aperiodicity). The results also support previous studies that show that voiced obstruents are frequently manifested with either little or no aperiodic energy, or with strong periodic and aperiodic components. The EGG data were used as a reference for evaluating the pitch detection algorithm. The ground truth was not manually checked to rectify or exclude incorrect estimates. The overall gross error rate in pitch prediction across the three speech databases was 5.67%. In the case of synthetic speech-like data, the estimated SNR was found to be in close proportion to the actual SNR, and the pitch was always accurately found regardless of the presence of any shimmer or jitter.**

***Index Terms*—Aperiodic and periodic energy, average magnitude difference function (AMDF), pitch detection, speech preprocessing, voiced obstruents, voice quality.**

I. Introduction

In the production of speech, there are a number of sources that generate acoustic energy in the vocal tract. Aperiodic sources include aspiration, generated at the glottis; frication, generated further forward in the vocal tract; and transient bursts produced by the rapid release of complete constrictions. The periodic source in speech is created by vibration of the vocal folds creating periodic energy at the glottis. These sources are filtered by the vocal tract to generate an output signal, which will also be periodic or aperiodic depending on the source(s). Identifying and quantifying these various sources has several applications in speech coding, speech recognition and speaker recognition.

Most of the algorithms used to detect aperiodicity are passive, i.e., aperiodicity is considered the inverse of periodicity in nonsilent regions. The amount of aperiodicity is estimated using indirect measures like zero crossing rate, high-frequency energy and ratios of high-frequency energy to low-frequency energy. These measures are prone to making errors in situations where the signal has simultaneous strong periodic and aperiodic components, as is the case with some of the voiced fricatives. Such methods will also be only marginally useful in distinguishing high-frequency periodic energy from high-frequency aperiodic energy. In this paper, we present a time domain aperiodicity, periodicity, and pitch (APP) detector that estimates 1) the proportion of periodic and aperiodic energy in a speech signal and 2) the pitch period of the periodic component. The APP system uses a time domain method and is based on the distribution of the minima of the average magnitude difference function (AMDF) of the speech signal. In the previous versions [1], [2], the APP system made a binary decision about periodicity and aperiodicity of each frequency channel. In contrast, the present APP system estimates the proportion of periodic and aperiodic components in each channel. The APP system also gives an estimate of the pitch period of the periodic component.

The structure of the periodicity/aperiodicity detection part of the APP system is very similar to a pitch detection algorithm and, hence, includes a block for the estimation of the pitch of the periodic component of the signal. For the present work, pitch is defined as the frequency of the vocal fold vibration. Many methods have been proposed for reliable estimation of the pitch frequency. A comprehensive review of such methods can be found in [3]. From an auditory perspective, there are a number of different types of pitch sensations as a function of the type of sound source. For instance, a pure tone or a complex harmonic stimulus with primarily low harmonics will generate a much clearer pitch sensation than a complex harmonic stimulus with primarily high-frequency harmonics, or a click train. A variety of different models of pitch perception have been developed to attempt to explain these perceptual phenomena. There are two traditional types of models of human pitch perception: either spectral pattern recognition models based on resolved frequencies [4] or temporal models based on time-domain analysis [5]; see [6] for a review. Hybrid models based on peripheral auditory processing [7], [8], that take into account both auditory filtering and temporal processing may be able to account for most of the abilities of human listeners.

Finally, note that the decomposition of signals into periodic and aperiodic components is not a completely new idea. In the

Manuscript received December 20, 2003; revised September 4, 2004. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Bayya Yegnanarayana. This work was supported by the National Science Foundation under Grant SBR-9729688 and the National Institute of Health under Grant 1K02 DC00149-01A1.

O. Deshmukh, C. Y. Espy-Wilson, and J. Singh are with the Department of Electrical and Computer Engineering and Institute for Systems Research, University of Maryland, College Park, 20742 USA (e-mail: omdesh@Glue.umd.edu; espy@Glue.umd.edu; singhj@Glue.umd.edu).

A. Salomon is with the Research Lab for Electronics, Speech Communication Group, Massachusetts Institute of Technology, Cambridge, MA 02142 (e-mail: ariel@speech.mit.edu; asalomon@mit.edu).

Digital Object Identifier 10.1109/TSA.2005.851910

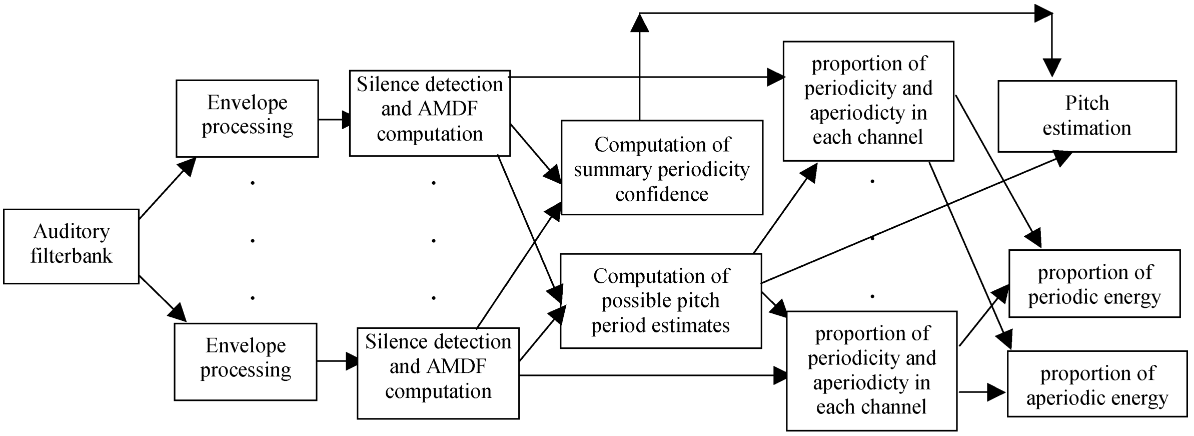

Auditory filterbank
Envelope processing
Envelope processing
Silence detection and AMDF computation
Silence detection and AMDF computation
Computation of summary periodicity confidence
Computation of possible pitch period estimates
proportion of periodicity and aperiodicty in each channel
proportion of periodicity and aperiodicty in each channel
Pitch estimation
proportion of periodic energy
proportion of aperiodic energy

Fig. 1. Block diagram of APP system.

fields of auditory scene analysis [9], [10], some work has been done on recognition of wide-band periodic ("weft," as per [9]) and aperiodic ("click" and "noise cloud") components in an audio signal. In the context of speech processing and coding, there has been considerable work that considers the problem of splitting a speech signal into periodic and aperiodic components for the purpose of manipulating the voice quality of synthetic speech [11]–[13], studying turbulent sources in natural speech [14], and sound hybridization on musical acoustics [15].

Our particular interest in developing the APP system is to study variability in the acoustic manifestations of speech sounds, particularly the voiced obstruents, and to use the measures developed and the knowledge gained to improve the performance of our speech recognition system [16]. With this goal in mind, the APP system focuses on estimating the relative spectro-temporal proportion of periodicity and aperiodicity in speech signals.

The APP system has several applications in speech coding, speech recognition and speaker recognition. It can be used in tasks such as segmentation of speech into voiced and unvoiced regions; the recognition of regions where both periodic and aperiodic components exist, e.g., in a breathy vowel, or a voice fricative; or as a component of a system for phonological segmentation and recognition. The strength of aperiodicity can also be an important cue in distinguishing the place and voicing of fricatives.

In Section II, the various stages of signal analysis are explained. The different databases used to train and evaluate the APP system are also discussed in this section. Section III describes the different experiments performed to evaluate the performance of the APP system in pitch detection and in periodicity/aperiodicity detection. The performance results and their implications are discussed in Section IV. Some conclusions are drawn in Section V.

II. Method

A. Database

Several databases that consist of simultaneously collected acoustic and electroglottograph (EGG, also referred to as laryngograph) data were used to test the APP system. Different databases were used to demonstrate the robustness of the system to various recording conditions and to facilitate comparisons with other similar works [17], [18].

The MOCHA [19] database consists of 460 utterances, each recorded by two speakers (one male and one female) in clean environment at a sampling rate of 16 kHz. The MOCHA database is hand transcribed. A subset of 50 randomly selected sentences (25 from each speaker) was used in the development of the APP algorithm. The Bagshaw (DB2) database [20] consists of 50 utterances recorded each by one male and one female in clean environment at a sampling rate of 20 kHz. The Keele (DB5) database [21] consists of one long utterance recorded by five males and five females in clean environment at a sampling rate of 20 kHz.

A final database used to evaluate the APP system consists of synthetic speech-like signals [17]. The signals are the outputs from a 50-pole linear predictive coding (LPC) synthesis filter when it is excited by a pulse train that is corrupted by Gaussian white noise (GWN). The signal-to-noise ratio (SNR) varied from ∞ to -5 dB. The pitch period and amplitude of this pulse were perturbed by specified degrees of jitter (fluctuation in the pitch period) and shimmer (fluctuation in the amplitude of the signal).

B. Signal Analysis

Signal analysis consists of a series of stages per channel followed by across-channel processing, as detailed in Fig. 1. The signal processing performed by each of the blocks is explained in the following subsections.

1) Auditory Filterbank: The analysis begins by splitting the signal into a set of bandpass frequency channels. The analysis filterbank was a 60-channel auditory gamma-tone filter bank [22] with channel characteristic frequencies (CFs) based on the equivalent rectangular bandwidth (ERB) (as defined in [23]) scale, and ranging from 100 Hz to just below half the sampling rate. Notice that the upper limit on the channel CF is sampling rate dependent and, thus, no assumption about the sampling rate of the signal is made. An auditory filter bank was preferred for spectral analysis in order to provide an accurate weighting of the frequency components, most importantly in terms of the relative perceptual strength of the periodic and aperiodic components in speech.

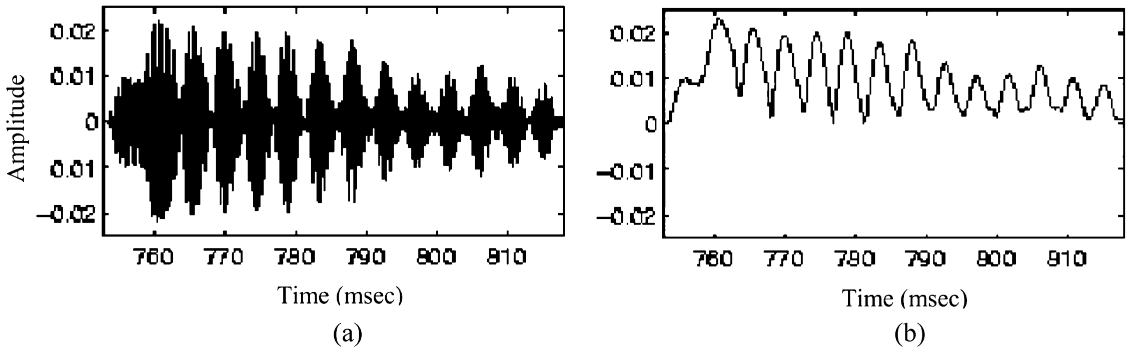


Fig. 2. (a) Bandlimited output of the channel with CF = 1935 Hz. (b) Corresponding envelope obtained using the Hilbert transform.

2) Envelope Extraction: To remove fine structure while avoiding excessive smoothing in the time domain, the Hilbert envelope was used [24]. The Hilbert transform was approximated using a Kaiser window of order 512. The 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. Given a real narrow-band signal as input, the Hilbert transform produces a version of its input signal that is precisely 90° out of phase, such that the amplitude of the complex sum of these two signals is an estimate of the low-frequency amplitude modulation applied to the signal. Fig. 2 shows that this transform preserves the abrupt changes at the maximum rate that can be captured by a particular channel given its CF.

3) Silence Detection and AMDF Computation: A frame is judged to be nonsilent if its total energy is no more than 35 dB below the maximum total energy computed across all of the frames in the utterance. All the other frames and their channels are judged to be silent. For any given nonsilent frame of the utterance, a channel within that frame is considered nonsilent if its energy is no more than 45 dB below the maximum channel energy that has been computed up to that point, including the channel energies in the present frame. All the other channels are judged as silent. If the channel is classified as silent, then no further processing is done. These very liberal thresholds for silence detection were empirically found using the training data and their use was driven by the computational complexity of the APP system. For real-time processing, the silence detector can be adjusted so that the look ahead is considerably less (around 300 ms or so to include a vowel), or an inverse strategy can be implemented where the silence threshold is based on minimum energy as opposed to maximum energy as the reference.

The temporal envelope in each nonsilent channel was analyzed for periodicity, aperiodicity and pitch. If the temporal envelope signals are either monotonically increasing or decreasing due to the amplitude variations at the boundaries of adjacent sounds, then they are flattened prior to analysis. This flattening is done by subtracting a linear function (whose slope is equal to the rate of rise/fall of the envelope) from the envelope signal. The raw pitch estimates in each channel were produced using the short-time AMDF. The AMDF was chosen over the more common autocorrelation operator due to the ease of computing a confidence metric. The autocorrelation operation consists of multiplication followed by addition, which is computationally more expensive than the AMDF operation that consists of subtraction followed by addition. The AMDF [25] 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, $w(m)$ in this case is a 20-ms rectangular window and $k$is the lag value, which varies from 0 to the sample value equivalent of 20 ms (e.g., if the sampling rate is 16 kHz, k will take values over the range {0,320}). This function looks roughly like an inverted autocorrelation function. 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. The value of these dips is referred to as the strength of the dips. If the signal is aperiodic, the AMDF waveform will not show such evenly spaced dips. The decision regarding periodicity and aperiodicity is based on the location and the strength of the dips occurring in the AMDF waveform. The dip locations and their strengths are found by computing the convex hull [26] of the AMDF. The strength of a dip can at the most be 1. The strength of the dip is the confidence of that dip location being the pitch period at that instance. The AMDF is computed for each nonsilent channel over a 20-ms window and at a rate of 5 ms. Fig. 3(a)–(3d) shows the AMDF and the dips for two typical periodic and two typical aperiodic channels, respectively.

4) Computation of Possible Pitch Period Estimates: This stage computes the possible pitch period estimates for each frame based on the distribution of the AMDF dips summed across the channels which is referred to as the summary measure. For a strongly periodic frame, the summary measure of the dip strengths will result in clusters at the pitch value and its integer multiples. For a strongly aperiodic frame, the summary measure will result in dips that are randomly scattered over the range of the possible lag values with no prominent clusters. Fig. 3(e) shows the summary measure for a periodic frame and Fig. 3(f) shows the summary measure for an aperiodic frame. Notice that the coherence of the dips for a periodic frame results in a maximum summary value that is an order of magnitude higher than that obtained for an aperiodic frame.

Computation of the summary measure for a particular frame depends on all of the channel estimates that were computed within 10 ms of that frame. Recall that channel estimates are

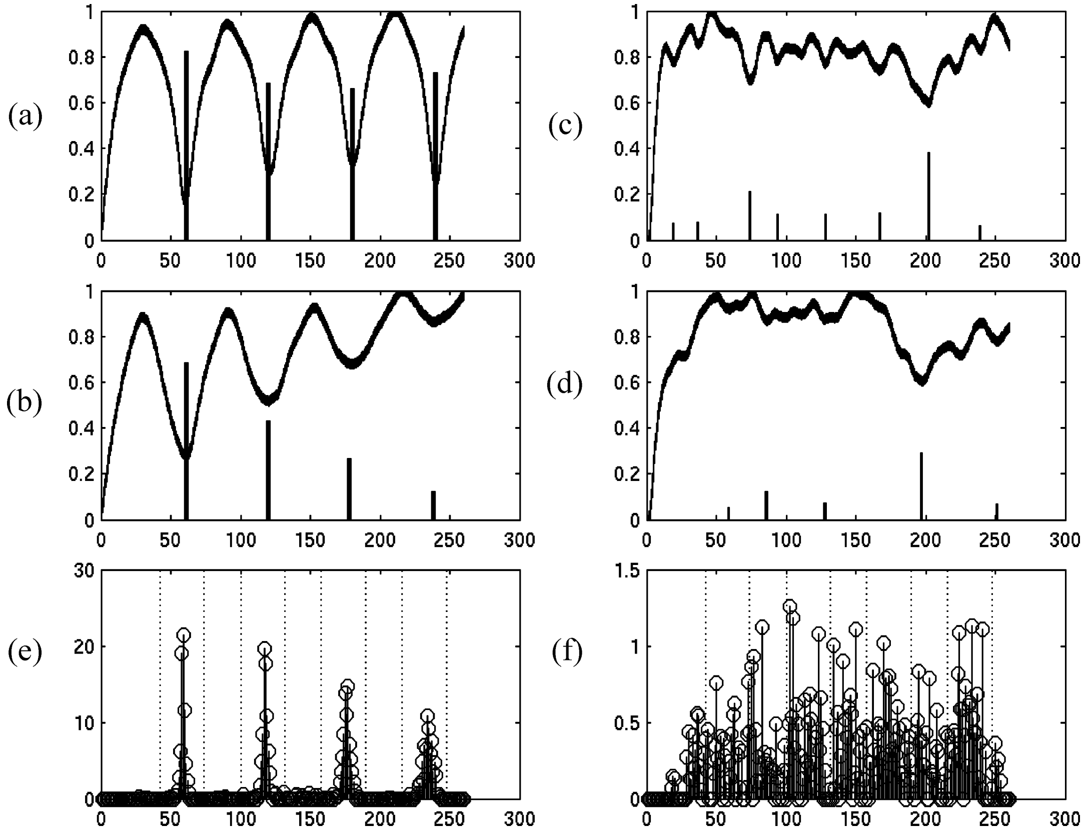


Fig. 3. (a)–(b) AMDF and its dips for typical periodic channels (x axis shows the lag values). (c)–(d) AMDF and its dips for typical aperiodic channels. (e) AMDF dips clustered across all the channels in a typical periodic frame (the vertical lines on each side of the clusters indicate the tolerance boundaries). (f) AMDF dips clustered across all the channels in a typical aperiodic frame (the vertical lines show the tolerance boundaries propogated from the previous periodic frames). Notice that the maximum value of the dip strength over the range of dip locations is 1.4 in the aperiodic frame whereas the maximum value is 22 in the case of the periodic frame.

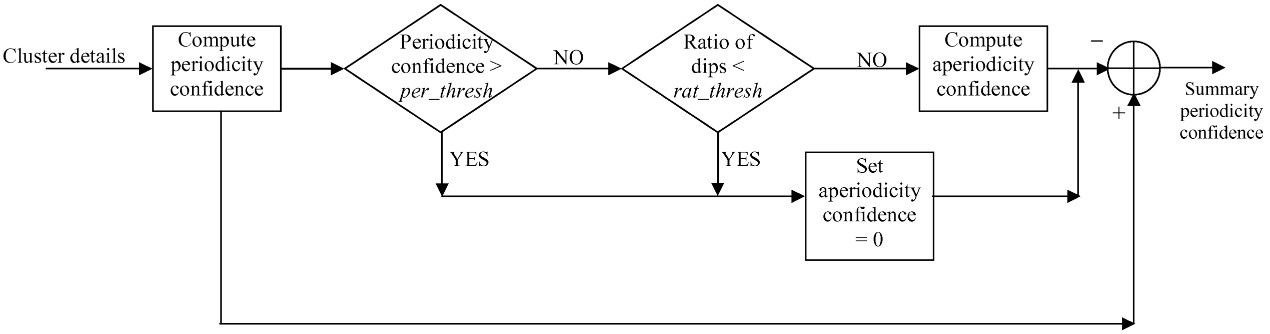

Cluster details
Compute periodicity confidence
Periodicity confidence > *per_thresh*
NO
YES
Ratio of dips < *rat_thresh*
NO
YES
Compute aperiodicity confidence
Set aperiodicity confidence = 0
−
+
Summary periodicity confidence

Fig. 4. Algorithm for the computation of summary confidence of a frame.

computed every 5 ms. A frame rate of 2.5 ms is used for analysis so that gradual changes in the relative amounts of periodicity and aperiodicity can be tracked.

At the beginning of the utterance, the lag location corresponding to the maximum of the summary measure is chosen as the peak location of a cluster. Other clusters are then formed by finding maxima near integer multiples of this location. As the analysis progresses, cluster locations from the previous frames are used to find clusters in the current frame. The locations of the peaks of the clusters of a frame are the pitch period estimates for that particular frame. Note that, at this stage, a frame can have more than one possible pitch period estimates.

5) Computation of the Summary Periodic Confidence: The algorithm to compute the *summary periodic confidence* is illustrated in Fig. 4. For each cluster of a frame, first a *periodicity confidence* measure is computed by summing the strengths of all the dips that lie within a certain tolerance region of the cluster peak. Dips outside this tolerance region are considered spurious. The tolerance region consists of samples within a 1-ms window on each side of the cluster peak location. The locations of the cluster peaks are the possible pitch period estimates of that frame and the corresponding *periodicity confidences* are the confidences of those pitch period estimates being the actual pitch period. Thus, a frame can have more than one *periodicity confidence*.

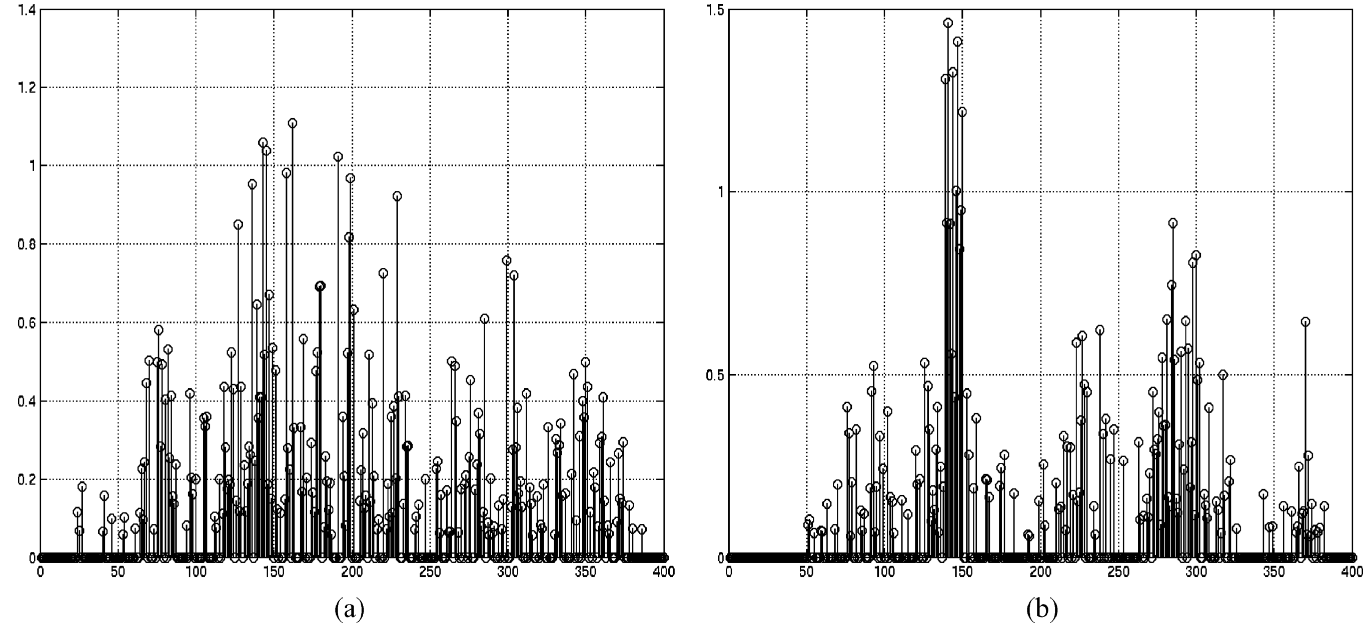


Fig. 5. Comparison of AMDF dips clustered across all the channels in (a) an aperiodic frame and (b) a weakly periodic frame. (x axis shows the lag value).

Frames belonging to strong periodic regions (e.g., the middle of a vowel) have very high *periodicity confidence* values. Frames corresponding to weakly periodic regions (e.g., voiced/unvoiced boundary regions or low-amplitude periodic sounds such as a /w/) tend to have low *periodicity confidence* values either because the dips near the pitch estimates are not strong or because many of the channels are considered silent. This leads to a considerable amount of overlap in the *periodicity confidences* of weakly periodic and aperiodic frames. On the training data, a threshold value of $0.75 * (\text{Sampling Rate}/1000)$[1] was found to give maximum separation between strongly periodic frames and weakly periodic or aperiodic frames. The phonetic transcription provided with the training data was used to label the frames as strongly periodic, weakly periodic and aperiodic. It now remains to distinguish weakly periodic frames from aperiodic frames.

Fig. 5 compares the summary measure for a frame in an aperiodic region [Fig. 5(a)] with that for a frame in a weakly periodic region [Fig. 5(b)]; both frames have comparable *periodicity confidences*. Notice that the aperiodic frame looks noisier than the weakly periodic frame, i.e., the number of lags with nonzero dips is more in the aperiodic frame than it is in the weakly periodic frame. Comparison of many such frames in the training data showed that this difference is generally true. Hence, the ratio of the number of lags with nonzero dips to the total number of lags was used as a parameter to distinguish weakly periodic frames from aperiodic frames. A value of 0.55 for this parameter was found to give maximum separation between weakly periodic frames and aperiodic frames on the training data.

The performance of the algorithm on the training data showed no considerable variation as these two thresholds were varied over a certain range near the previously mentioned optimal values. This implies that these thresholds do not need fine-tuning.

6) Pitch Estimation: For a given frame, all the clusters with negative *summary periodicity confidence* are dissolved. If no cluster survives this test, the frame is likely to be aperiodic and the frame is not assigned a pitch value. For all of the other frames, the peak of the first cluster is the pitch estimate of the frame and the corresponding *summary periodicity confidence* is the pitch confidence of the frame.

Note that the *summary periodicity confidence* is used only to decide the frames to which pitch value is to be assigned and plays no role in the computation of proportion of periodicity and aperiodicity in channels.

The APP algorithm contains special checks to rectify pitch halving and pitch doubling errors. To allow the flexibility to change a pitch value, a cluster is formed near the half multiple of the mean of previous pitch estimates if the *summary periodicity confidence* of the cluster closest to the half multiple is greater than that of the cluster near the previous pitch estimates. The pitch value is then set to the peak of the cluster closest to the half multiple. This allows the system to rectify its pitch halving errors. A similar test, where the half multiple is replaced by an integer multiple, is incorporated to rectify the pitch doubling errors.

At the same time, these criteria were chosen in such a manner that the algorithm could track the pitch correctly even when it is actually halved. Fig. 6 shows one such example where the APP system was able to detect the pitch halving. Notice that APP system starts tracking the pitch halving after a delay of about five frames since it was designed to need pitch estimates from at least five previous frames (about 12.5 ms) to capture the pitch halving before it will reflect pitch halving.

7) Proportion of Periodicity/Aperiodicity in Each Channel: The distribution and the strengths of the dips in the channels relative to the locations of the cluster peaks are used to compute the proportion of periodicity and aperiodicity

[1] Note that the threshold is sampling rate dependent, making it robust for databases with different sampling rates.

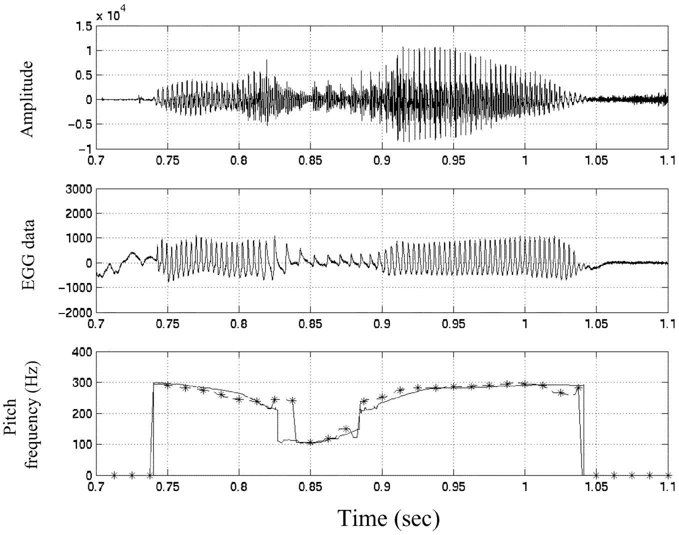


Fig. 6. (a) Time waveform for most of the word "this." (b) EGG data. (c) Comparison of F0 computed from the EGG data and the APP system (dashed line marked with "∗"). Notice that the pitch detector of the APP system is able to track the pitch halving.

in each channel. The aim is to distinguish 1) strongly periodic channels [e.g., Fig. 3(a)] from weakly periodic channels [e.g., Fig. 3(b)] and 2) strongly aperiodic channels [e.g., Fig. 3(c)] from weakly aperiodic channels [e.g., Fig. 3(d)]. For each channel, the strength of the AMDF dips closer to the cluster peaks is used to estimate the periodicity strength. The randomness in the distribution of the dips is quantified to capture the degree of aperiodicity. If the sum of the periodicity and aperiodicity measure is greater than one, they are scaled down proportionally so that the sum equals one.

Periodicity measurement: Since speech is quasi-periodic, dips with strengths closer to one should contribute one toward periodicity, whereas dips with moderate strengths should contribute their original value. To achieve this, the dip strengths are normalized using a logarithmic function that is roughly linear for smaller values and flattens for higher values. These normalized dips are then weighted such that dips closer to the cluster peaks contribute more toward periodicity. This contribution decreases rapidly with increasing distance from the cluster peaks. Consequently, we found that exponentially decaying weights perform better than linearly decaying weights.

If a signal is periodic, it is expected that equally spaced dips of similar strengths will be present in the AMDF. To account for this, we consider regions around each pitch multiple separately. That is, if the detected pitch of the frame is such that it can accommodate N pitch multiples in the lags, then each of the regions from $[jF0 - F0/2 : jF0 + F0/2]\, for\ j = 1, 2 \ldots N$ is analyzed separately for periodicity. Each region is called a channel cluster and its corresponding periodicity the cluster periodicity. The following equation shows the calculation of the cluster periodicity for the jth cluster

$$p_j = s_j + (1 - s_j) \sum_{i=-\frac{f_0}{2}}^{+\frac{f_0}{2}} d_i \times w_i$$

where s_j is the strength of the dip closest to the peak of the cluster, d_i is the strength of dip i locations away from the peak and w_i is the value of the exponential weighted function at location i.

The cluster periodicity can at most equal one; if multiple dips are present in the cluster, the most significant dip closest to the pitch period location contributes its normalized and weighted strength and the other dips contribute at most one minus this value. The average across the periodic clusters is taken as the periodicity measure of the channel.

Aperiodicity measurement: The AMDF dips in channels that are predominantly aperiodic are 1) located far from the pitch period and its multiples, 2) are small in amplitude, and 3) are generally numerous. The measurement of aperiodicity also utilizes weighted strengths of AMDF dips with two important considerations. First, dips far from the cluster peaks should contribute close to their full value toward aperiodicity and this contribution should gradually decrease for dips closer to the cluster peaks. Thus, logarithmically increasing weights, with the maximum corresponding to the lag farthest from the cluster peaks, are used. Second, the strength of aperiodicity should be directly related to the number of spurious dips. Hence, the aperiodicity measure is defined as the *sum* of these weighted dips instead of the *mean* across the clusters.

The periodicity and aperiodicity measures for the channels shown in Fig. 3 are, respectively, (a) 0.74, 0 (b) 0.49, 0 (c) 0.03, 0.79 (d) 0.09, 0.26. These values are in line with our expectations.

8) Proportion of Periodic and Aperiodic Energies: The periodicity and aperiodicity measures discussed previously are multiplied by the corresponding channel energies and summed across the channels to get the overall periodic and aperiodic energies for the frame. The ratio of overall periodic energy to the total energy in the frame is the proportion of periodic energy. The proportion of aperiodic energy is calculated similarly.

It is worth mentioning that the thresholds used by this algorithm do not need to be retrained for different databases with different sampling rates. This claim is substantiated in the results section where the APP system is tested on different databases with different sampling rates.

C. Computation of Pitch From EGG Data

The EGG data captures the laryngeal behavior by measuring the change in electrical impedance across the throat during speaking [27]. The EGG waveform exhibits strong periodic fluctuations during vocalized sounds with the period equal to the pitch period of the speaker. To estimate the pitch value from the EGG data, the EGG waveforms were bandpass filtered with cutoff frequencies of 50 and 750 Hz. To highlight rapid fluctuations in the signal and to remove peaks due to extraneous noise, a first-order difference operation was performed on the filtered signal. This new signal exhibits prominent positive peaks at regular intervals in periodic regions. The gap between two consecutive peaks is the instantaneous pitch period. A peak-picking algorithm is then implemented to find the locations of these peaks. The average value of the gaps between consecutive peaks over a period of 10 ms is the pitch estimate at that location. The pitch estimates were computed every 2.5

TABLE I
Gross Errors in Pitch Detection. Reference Pitch Values for Mocha and DB2 Were Calculated With Algorithm in Section II-B-7. Reference Pitch Values for DB5 Were Taken From [21]

	Mocha			DB2			DB5		
	Half	Double	overall	Half	Double	overall	Half	Double	overall
Male	2.72	1.86	4.58	3.16	2.93	6.09	1.89	7.57	9.46
Female	4.29	2.35	6.64	2.05	1.06	3.11	3.48	3.42	6.90
Overall	3.57	2.12	**5.69**	2.61	2.02	**4.63**	2.69	5.47	**8.16**

TABLE II
Gross Errors in Pitch Prediction on DB2 by YIN and APP System When Reference Pitch Values Were Taken From [18]

	YIN	APP
Overall	1.40	2.72

ms. The aperiodic regions are marked by the absence of any such regularly spaced peaks.

III. Results

A. *Pitch Detection*

The pitch estimates from the APP system were compared with the EGG-derived pitch values on a frame basis. Using the standard established in previous studies [18], [20], the pitch value was said to be in agreement if the difference between the estimated pitch and the reference pitch was less than 20% of the reference pitch value. Otherwise, the pitch estimates were treated as a gross error.

The gross errors were split into two different categories. The *halving* errors are defined as the instances where the estimated pitch was more than 20% below the pitch value given by the EGG data. The *doubling* errors are the instances where the estimated pitch was more than 20% above the pitch value derived from the EGG data. Table I gives the details of the gross errors for the three databases. The results are given separately for males and females.

Table II compares the performance of the pitch prediction module of the APP system with the YIN pitch estimator proposed in [18]. In this case, the EGG-derived pitch values in [18] were used for both pitch detectors. There are several differences in the derivation of the results for DB2 in Table I and the results in Table II. First, the reference pitch values were derived using different algorithms. The results for DB2 in Table I used reference pitch values that were derived according to the algorithm described in Section II-B.7. The reference pitch values used for the results in Table II were derived using the YIN pitch detector and the reference values were hand corrected. Second, the results reported for the YIN pitch detector used a temporal tolerance where the reference estimates were time shifted over a certain range to give the minimum error rate. No such shifting was used to score the output of the APP pitch detector. Note that the performance of the APP pitch detector as reported in Table II is better than most of the pitch detectors evaluated on DB2 in [18]. Further, note that the error for the APP pitch detector in Table II is considerably lower than the results in Table I. We conclude that the results for the other databases in Table I would benefit favorably from hand correction of the reference values.

TABLE III
Gross Errors in Pitch Prediction on DB2 Database

	Normal	Missing fundamental
Male	6.09	6.37
Female	3.11	3.08
Overall	4.63	4.76

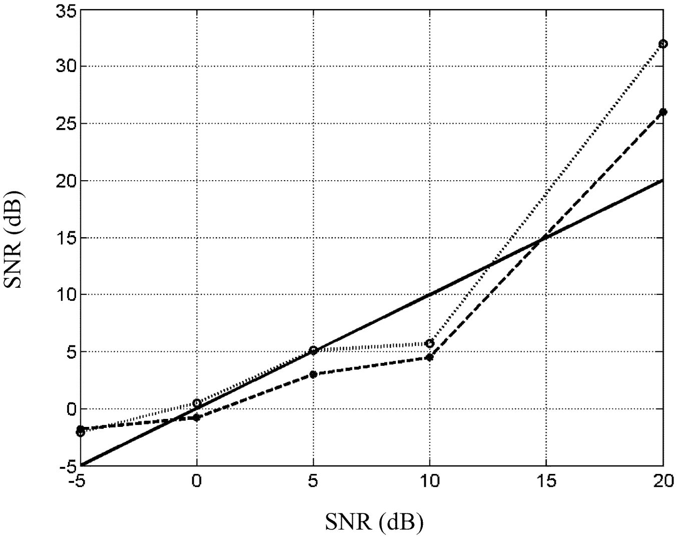


Fig. 7. Comparison of actual SNR (*solid line*). SNR computed by the APP system for 131 Hz pulse (dashed line with filled circles). SNR computed by the APP system for 120 Hz pulse (dotted line with o). Notice that the SNR predicted by the APP system closely follows the actual SNR.

A final evaluation of our system is the detection of pitch in speech where the fundamental harmonic is missing. The signal content below 300 Hz was set to zero in the DB2 database. Table III compares the performance of the pitch tracker on this database before and after high-pass filtering. The consistency in performance of the pitch detector even when the low-frequency data was removed makes it a promising candidate for pitch detection in telephone speech scenarios or for pitch detection in missing harmonic data.

B. *Periodicity and Aperiodicity Detection*

1) Evaluation on Synthetic Data: To evaluate the performance of our periodic and aperiodic measures, we compared the SNR based on these measures with the known SNR of the synthetic signal. We define the SNR based on our measures as

$$\mathrm{SNR} = 20 * \log\left(\frac{\sum_i P_i}{\sum_i AP_i}\right)$$

where P_i is the periodic energy and AP_i is the aperiodic energy calculated by the APP system in the ith frame. Fig. 7 shows the actual SNR versus the computed SNR for pulses with frequencies 131 and 120 Hz and no jitter or shimmer. Notice that the difference between the actual SNR and the computed SNR is small. The harmonics-to-noise ratio (HNR) measured by the pitch-scaled harmonic filter (PSHF) system in [17] is about 3 dB higher than the actual SNR whereas the SNR measured by the APP system presented in this paper remains within 5 dB of the actual SNR. The effect of different degrees of jitter and shimmer

TABLE IV
PERFORMANCE OF THE APP SYSTEM FOR VARIOUS DEGREES OF JITTER AND SHIMMER AND AT VARIOUS SNRs

Jt %	sh dB	∞	20	10	5
0	0	46.23	25.70	4.33	2.68
	1	23.01	21.33	3.93	1.89
0.5	0	44.82	24.71	5.16	2.69
	1	20.13	18.10	2.86	0.91
3	0	22.45	16.97	2.96	3.16
	1	15.75	12.68	2.02	2.38

TABLE V
PERFORMANCE OF PERIODICITY AND APERIODICITY MEASURES

	MOCHA	DB2	DB5	Overall
Periodicity accuracy	94.69	92.56	91.13	94.43
Aperiodicity accuracy	96.59	96.22	89.08	96.32

at various SNRs on the performance of the APP system is tabulated in Table IV. In the absence of any noise (i.e., at ∞ dB SNR) increasing jitter or shimmer reduces the estimated SNR. But as the SNR reduces the effect of jitter and shimmer is less pronounced. These results are qualitatively very similar to those reported in [17].

2) Evaluation on the Natural Speech Databases: The periodic and aperiodic measures were also evaluated using the three natural speech databases. All the comparisons were made on a frame basis at a frame rate of 2.5 ms. We define the periodicity accuracy as the ratio of the number of nonsilent frames that have both the proportion of periodic energy no less than 0.25 and the corresponding EGG output is nonzero, to the total number of frames that have a nonzero EGG output. The aperiodicity accuracy is defined as the ratio of the number of nonsilent frames that have the proportion of aperiodic energy no less than 0.35 and the corresponding EGG output is zero, to the total number of nonsilent frame that have zero EGG output. These thresholds were derived from statistical analysis of the training data. The results for the periodicity and aperiodicity accuracies are shown in Table V. An example of the outputs from these measures is shown in Fig. 8. Note from part (d) of Fig. 8 that the proportion of periodic and aperiodic energies do not always sum to one since some nonsilent channels are not judged to be periodic or aperiodic (this situation occurs when there are no dips in the AMDF waveform).

One cause of the less than perfect periodicity and aperiodicity accuracy is due to the boundary problem. It was found that many of the frames that are considered to be in error according to the EGG are in the transition region between adjacent sounds that differ in their voicing. In such regions, 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. More often, the EGG turns off before our algorithm stops detecting periodicity, or starts a few frames after our algorithm starts detecting periodicity. This is probably due to weak voicing (i.e., a large drop in the peaks calculated from the EGG data). This situation will result in lower aperiodicity accuracy for our algorithm since we will detect periodicity when the EGG is off. This scenario is manifested in Fig. 8 around 480 ms where our periodicity detector remains on (i.e., the proportion of periodic energy is greater than 0.25) for 7.5 ms longer than the offset of the EGG waveform [shown with a vertical line around 480 ms in Fig. 8(d)]. The time waveform does indicate the trailing, weak periodic signal in this region. We have observed situations where the reverse is true, i.e., our algorithm stops detecting periodicity before the EGG turns off. Generally, periodicity is detected in these transition regions. However, the proportion of periodic energy is less than 0.25. An example of this situation can be seen in Fig. 8 around 650 ms, where the EGG is on for about 5 frames longer than our periodicity estimate that is just below 0.25. Such situations lower the periodicity accuracy. Table VI shows the periodicity and aperiodicity accuracies when the misclassifications at the boundaries were excluded. Notice that the overall periodicity accuracy across all the three databases increases from 94.43% to 98.60% and the aperiodicity accuracy increases from 96.32% to 99.14% when the misclassifications at the boundaries are excluded. It has been noted in [27] that several EGG pulses near voicing onset or

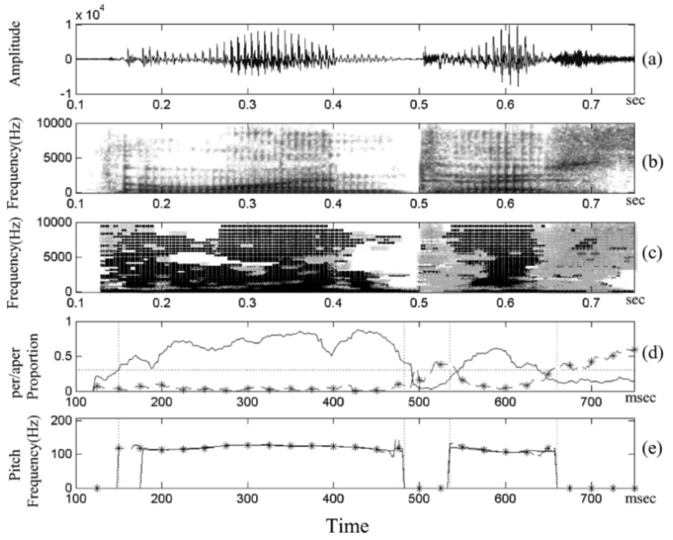


Fig. 8. Plots for the utterance "The wine tastes." (a) Time waveform. (b) Spectrogram. (c) Periodicity (dark) and aperiodicity (light) profile. (d) Proportion of the periodic energy and aperiodic (dashed with $*$) energy (horizontal dotted line indicates the periodic/aperiodic threshold). (e) Pitch estimate based on the EGG data (dashed with $*$) and pitch detected by the APP system (periodic and aperiodic regions based on the EGG data are demarcated by the vertical lines).

TABLE VI
PERFORMANCE OF THE PERIODICITY AND APERIODICITY MEASURES WHEN THE ERRORS IN THE BOUNDARY REGIONS WERE EXCLUDED

	MOCHA	DB2	DB5	Overall
Periodicity Accuracy	98.56	98.93	99.07	98.06
Aperiodicity Accuracy	99.10	99.33	99.73	99.14

TABLE VII
PERCENTAGE OF FRAMES IN DIFFERENT BROAD CLASSES (ACCORDING TO HAND TRANSCRIPTION) DETECTED AS PERIODIC AND APERIODIC AND USING EGG-DATA AS THE GROUND TRUTH. NUMBERS IN PARENTHESIS SHOW THE PERCENTAGE OF THE TOTAL FRAMES THAT FALL IN THE RESPECTIVE CATEGORY

	Mocha	
	periodic	aperiodic
Sonorants	95.26 (57.72)	92.26 (4.66)
Voiced obstruents	92.46 (5.60)	91.35 (8.26)
Unvoiced obstruents	85.90 (2.39)	99.51 (21.34)

TABLE VIII
PERCENTAGE OF FRAMES IN DIFFERENT BROAD CLASSES OF THE MOCHA DATABASE WHERE ONLY STRONG PERIODICITY WAS DETECTED, STRONG APERIODICITY WAS DETECTED AND BOTH STRONG PERIODICITY AND APERIODICITY WERE DETECTED. NUMBERS IN PARENTHESIS SHOW THE TOTAL NUMBER OF FRAMES IN EACH CATEGORY

	only strong Periodic energy	only strong Aperiodic energy	strong Periodic and Aperiodic energy
Sonorants (541758)	80.02	8.76	11.22
Voiced obstruents (120533)	29.44	36.38	34.17
Unvoiced obstruents (206038)	4.83	88.85	6.32

offset can be distorted and the pitch estimates from the EGG data in these regions can be unreliable.

Table VII provides more details about the performance of the aperiodic and periodic measures for different types of sounds. More specifically, the hand transcriptions provided with the MOCHA database were used to pull out the results for sonorant sounds (vowels, semivowels and nasals) that should be primarily periodic, voiced obstruent sounds that may have been produced with both strong periodic and aperiodic sources and unvoiced obstruent sounds that should be strictly aperiodic. Using this division of the speech signal, the results in Table VII for periodic sounds are only for those frames where the EGG was also on. Similarly, the results for the aperiodic sounds are based only on those frames where the EGG was off.

Table VIII shows the results for the MOCHA database for the percentage of frames in the different broad classes that showed only strong periodicity, strong aperiodicity, or both strong periodicity and aperiodicity. For these results, the EGG signal is 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, about 30% of the voiced obstruents show only strong periodicity. This finding is in agreement with previous studies that show that voiced obstruents can be lenited so that they are realized as sonorant consonants [28]. 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. It is often difficult to know where to place the boundary between a sonorant and a strident fricative. The transition region may show several periods resembling those in the vowel region, but with a simultaneous aperiodic component riding on top of the lower frequency waveform. Thus, it is not clear if the coarticulated region should be included in the vowel region or in the obstruent region. The fact that there is 94.69% agreement between the EGG data and our periodicity detector suggests that many of these coarticulated regions were included within labels of the sonorant regions, leading to 8.76% of the frames showing strong aperiodicity.

TABLE IX
APERIODICITY STRENGTH FOR THE FRICATIVES

	sh/zh	s/z	f/v	th/dh	ch/jh
aper strength	111.1/85.5	101.5/83.1	95.2/50.5	90.7/62.8	74.0/68.1

The strength of aperiodicity defined as the average number of spurious dips across all the channels, can be used to distinguish strident fricatives from nonstrident and voiced ones from their unvoiced counterparts. This strength of aperiodicity is tabulated in Table IX.

IV. DISCUSSION

The APP system presented here incorporates a relatively simple algorithm to extract the proportions and frequency range of periodic and aperiodic energies in speech signals. The system is robust in the sense that its efficiency does not depend on the accuracy of some inherently difficult tasks, like the detection of F0 [17], [29]. The performance of the algorithm implemented in the APP system does not depend on the sampling rate of the database. The performance across different databases with different sampling rates is comparable (Tables I and V). Informal evaluation of the APP system on telephone speech (sampling rate 8 kHz) was also encouraging. The thresholds trained on MOCHA database were not retrained to test the other databases. This shows that the system parameters do not need database specific fine-tuning.

Unlike many popular pitch detection algorithms, the algorithm presented in this paper does not involve any low-pass filtering nor is there an upper limit set by the algorithm on the pitch frequency at which it can distinguish periodic sounds from the aperiodic sounds. The lower limit depends on the analysis window size. However, as the pitch frequency increases, the consecutive clusters (mentioned in Section II-B.4) will get closer so that it will be progressively more difficult to demarcate the cluster boundaries, thus, affecting the periodicity/aperiodicity decision. (Making it more biased toward periodic decisions). Fig. 8(e) shows the pitch contour estimated by the APP system overlaid with the reference pitch values obtained from the EGG data. Notice that the APP system does not start detecting pitch until about 175 ms (when the *periodicity confidence* rises above *per_thresh*) whereas the EGG data starts showing pitch estimates from about 150 ms.

Table IX shows that, for a particular place, the degree of randomness in the distribution of the AMDF dips differs considerably for the voiced versus unvoiced obstruents. Further, as expected, the degree of randomness is more for strident fricatives and considerably less for the nonstrident fricatives. Finally we have also observed that the frequency range of the aperiodic

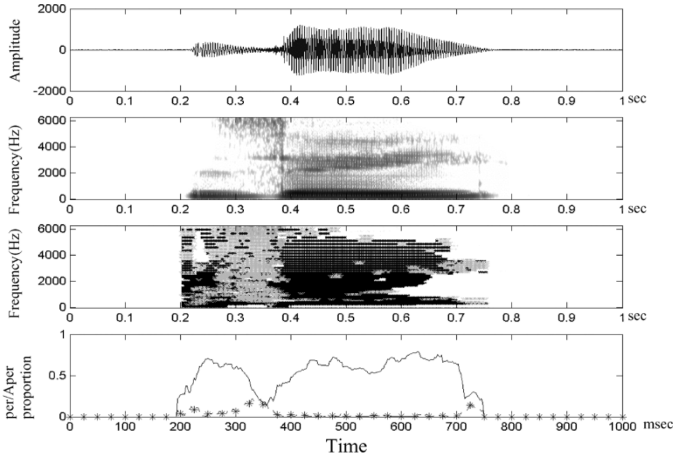


Fig. 9. Plots for the utterance "z." Top panel: time waveform. Second panel: spectrogram. Third panel: periodicity (dark) and aperiodicity (light) profile. Bottom panel: the proportion of the periodic energy and aperiodic energy (dashed with $*$).

noise may help to distinguish frication from aspiration. Consider the aspirated /t/ in the word 'tastes' shown in Fig. 8 (between 500 ms and 550 ms). The /t/ burst is followed by frication noise and then aspiration noise as the alveolar constriction is released further and the articulators move toward the position needed for the following vowel. Thus, the frication noise generated by the narrow constriction at the alveolar ridge is manifest in the high-frequency channels whereas the aspiration noise generated at the glottis is manifest in lower frequency channels.

One of the goals of the present work was to detect periodicity and aperiodicity in voiced obstruents that were manifest with strong periodicity like the /z/ in the alphabet 'z' shown in Fig. 9. Whereas most simple measures such as zero crossing rate would be able to detect the turbulence seen at high frequencies for a canonical /z/, when there is strong voicing, the aperiodicity "rides" on top of the low-frequency signal so that this measure fails for the /z/ shown in Fig. 9. High-pass filtering the signal and then computing the zero crossing rate is one possible solution to this problem. However, now the very strong high-frequency energy that sometimes occurs in vowels like /iy/ (where F2, F3, and F4 may be close together in a high-frequency region) will also exhibit a high zero crossing rate. Thus, the issue is to distinguish that in one case, the strong high-frequency energy is periodic, whereas in the other case, the high-frequency energy is aperiodic. The algorithm presented in this paper is able to make that distinction.

One of the applications of the periodicity/aperiodicity measures and pitch will be in our speech recognition algorithms [16], [30]. These parameters also form a part of a landmark detection system [31] where the main emphasis is broad classification of speech signals using primarily temporal cues. Finally, the present algorithm can also be used to detect breathy voice quality. The efficiency of the APP system in different speech-in-noise conditions will be evaluated in future.

The current implementation of the APP system is mostly in MATLAB and is several times real-time. We have identified several avenues where the amount of computation can be reduced and this will be addressed in near future. On a Pentium M processor with 1.7-MHz clock speed and 1-GB RAM the APP system is about 110 times real-time.

V. Conclusion

We have presented a novel, simple yet 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. The system was tested on three natural speech databases, which also had EGG data recorded simultaneously, and on synthetic speech like data. The robustness of the system to predict pitch in missing harmonic cases was also exhibited. The system is also successful in detecting simultaneous high periodicity and high aperiodicity when they are both present in voiced obstruent regions. The amount of aperiodicity in predominantly voiced regions can potentially be used to evaluate the voice quality. The amount of aperiodicity in obstruent regions can be a useful cue to distinguish the voicing and place of the fricatives. This system has vast applications since it improves on the typical binary voiced/unvoiced decision made in most speech recognition and speech coding systems.

Acknowledgment

The authors would like to thank A. Wrench for providing us with the Mocha database for free, P. Bagshaw for distributing the DB2 database, C. Shadle and P. Jackson for the synthetic data, and G. Meyer for DB5, A. de Chaveigne for providing the hand corrected reference pitch values used in [18].

References

[1] O. Deshmukh and C. Espy-Wilson, "Detection of periodicity and aperiodicity in speech signal based on temporal information," in *Proc. 15th Int. Congr. Phonetic Sciences*, Barcelona, Spain, 2003, pp. 1365–1368.

[2] ——, "A measure of periodicity and aperiodicity in speech," in *Proc. IEEE ICASSP*, Hong Kong, 2003, pp. 448–451.

[3] W. J. Hess, *Pitch Determination of Speech Signals*. New York: Springer-Verlag, 1983.

[4] J. L. Goldstein, "An optimum processor for the central formation of pitch of complex tones," *J. Acoust. Soc. Amer.*, vol. 54, pp. 1496–1516, 1973.

[5] J. F. Schouten, R. J. Ritsma, and D. L. Cardozo, "Pitch of the residue," *J. Acoust. Soc. Amer.*, vol. 34, pp. 1418–1424, 1962.

[6] B. C. J. Moore, *An Introduction to the Psychology of Hearing*, 4th ed. New York: Academic, 1997.

[7] R. D. Patterson, M. Allerhand, and C. Giguere, "Time-domain modeling of peripheral auditory processing: a modular architecture and software platform," *J. Acoust. Soc. Amer.*, vol. 98, pp. 1890–1894, 1995.

[8] R. Meddis and M. Hewitt, "Virtual pitch and phase sensitivity of a compute model of the auditory periphery. I: pitch identification," *J. Acoust. Soc. Amer.*, vol. 6, pp. 2866–2882, 1991.

[9] G. Brown and M. Cooke, "Computational auditory scene analysis," *Comp. Speech Lang.*, vol. 8, pp. 297–336, 1994.

[10] D. P. W. Ellis, "Using knowledge to organize sound: the prediction-driven approach to computational auditory scene analysis, and its application to speech/nonspeech mixtures," *Speech Commun.*, vol. 27, pp. 281–298, 1999.

[11] B. Yegnanarayana, C. d'Alessandro, and V. Darsinos, "An iterative algorithm for decomposition of speech signals into periodic and aperiodic components," *IEEE Trans. Speech Audio Process.*, vol. 6, no. 1, pp. 1–11, Jan. 1998.

[12] C. d'Alessandro *et al.*, "Effectiveness of a periodic and pperiodic decomposition method for analysis of voice sources," *IEEE Trans. Speech Audio Process.*, vol. 6, no. 1, pp. 12–23, Jan. 1998.

[13] O. Fujimura, "Approximation to voice aperiodicity," *IEEE Trans. Audio Electroacoust.*, vol. AU-16, no. 1, pp. 68–73, 1968.

[14] P. Jackson and C. Shadle, "Frication noise modulated by voicing, as revealed by pitch-scaled decomposition," *J. Acoust. Soc. Amer.*, vol. 108, pp. 1421–1434, 2000.

[15] X. Serra and J. Smith, "Spectral modeling synthesis: a sound analysis/synthesis system based on a deterministic plus stochastic decomposition," *Comput. Music J.*, vol. 14, no. 4, 1990.

[16] A. Juneja and C. Espy-Wilson, "Segmentation of continuous speech using acoustic-phonetic parameters and statistical learning," in *Proc. ICONIP*, 2002, pp. 726–730.

[17] P. Jackson and C. Shadle, "Pitch-scaled estimation of simultaneous voiced and turbulence-noise components in speech," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 7, pp. 713–726, Oct. 2001.

[18] A. de Cheveigne and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *J. Acoust. Soc. Amer.*, vol. 111, pp. 1917–1929, 2002.

[19] A. Wrench, "A Multichannel/Multispeaker Articulatory Database for Continuous Speech Recognition Research," Inst. Phonetics, Univ. Saarland, Res. Rep. 4, 2000.

[20] P. Bagshaw, "Automatic prosody analysis," Ph.D. dissertation, Univ. Edinburgh, Edinburgh, U.K., 1994.

[21] F. Plante *et al.*, "A pitch extraction reference database," in *Eurospeech*, vol. 1, Madrid, Spain, 1995, pp. 837–840.

[22] R. D. Patterson, "A pulse ribbon model of peripheral auditory processing," in *Auditory Processing of Complex Sounds*, W. A. Yost and C. S. Watson, Eds, NJ: Erlbaum, 1987.

[23] 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.

[24] A. Oppenheim and R. Schafer, *Discrete-time Signal Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1978.

[25] L. Rabiner and R. Schafer, *Digital Processing of Speech Signals*. Englewood Cliffs, NJ: Prentice-Hall, 1978.

[26] P. Mermelstein, "Automatic segmentation of speech into syllabic units," *J. Acoust. Soc. Amer.*, vol. 58, pp. 880–883, 1975.

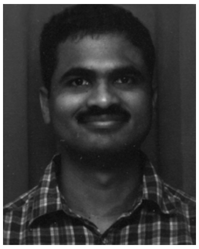
[27] D. G. Childers and A. K. Krishnamurthy, "A critical review of electroglottography," *Crit. Rev. Biomed. Eng.*, vol. 12, pp. 131–161, 1985.

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

[29] J. Lim, A. Oppenheim, and L. Braida, "Evaluation of an adaptive comb filtering method for enhancing speech degraded by white noise," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP–28, pp. 354–358, 1978.

[30] O. Deshmukh, C. Espy-Wilson, and A. Juneja, "Acoustic-phonetic speech parameters for speaker-independent speech recognition," in *Proc. IEEE-ICASSP*, 2002, pp. 593–596.

[31] A. Salomon, C. Espy-Wilson, and O. Deshmukh, "Detection of speech landmarks: use of temporal information," *J. Acoust. Soc. Amer.*, vol. 115, 2004.

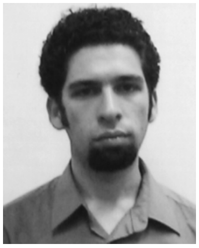


Om Deshmukh received the M.S. degree in electrical engineering from Boston University, Boston, MA, in 2001 and the B.E. degree in electrical and electronics engineering from Birla Institute of Technology and Science, Pilani, India, in 1999. He is currently working toward the Ph.D. degree in the Electrical and Computer Engineering Department, University of Maryland, College Park.

His current research interests include developing robust speech recognition techniques based on knowledge of acoustic phonetics and auditory processing.

Carol Espy-Wilson received the B.S. degree in electrical engineering from Stanford University, Stanford, CA, in 1979, and the M.S. and Ph.D. degrees in electrical engineering from the Massachusetts Institute of Technology, Cambridge, MA in 1981 and 1987, respectively.

She is currently an Associate Professor in the Department of Electrical and Computer Engineering, University of Maryland, College Park. Previously, she has been on the faculty at Boston University in the Electrical and Computer Engineering Department and a Research Scientist in the Research Laboratory of Electronics at the Massachusetts Institute of Technology. Her current research interests are in speech acoustics, speech recognition, and speaker recognition.



Ariel Salomon received the M.S. degree in electrical engineering from Boston University, Boston, MA, in 2000 and the S.B. degree in cognitive science with minor in linguistics from the Massachusetts Institute of Technology (MIT), Cambridge, MA, in 1996. He is currently working toward the Ph.D. degree in electrical engineering at MIT.

Jawahar Singh received the B.S. degree in computer engineering from the University of Maryland, College Park in 2004.

In 2003, he worked at the University of Maryland Speech Communication Lab as a research intern. He is currently working as a Consultant pursuing his other interests in computer networks and computer security.