A NOVEL METHOD FOR COMPUTATION OF PERIODICITY, APERIODICITY AND PITCH OF SPEECH SIGNALS

Om Deshmukh, Jawahar Singh, Carol Espy-Wilson

Speech Communications Lab. ECE Department,
University of Maryland, College Park

ABSTRACT

This paper presents improvements to our previously proposed algorithm to compute the proportion of periodic and aperiodic energies in speech signals and to estimate the pitch period. Although previously the periodic and aperiodic energies were estimated independent of each other at each frame, a binary decision was made at each of the non-silent channels. In this paper, we present an extension that replaces the binary decision with a measure of degree of periodicity and aperiodicity in each channel. Evaluation on synthetic speech-like data shows a better agreement in the estimated SNR and the actual SNR using with this improvement. Moreover, in the task of estimating the SNRs, this method significantly outperforms a method based on cepstral coefficients. When evaluated on a speech database, the periodicity and aperiodicity accuracy increased significantly. The previous pitch detector was prone to commiting pitch doubling and pitch halving errors and was unable to reliably detect pitch in weakly periodic regions. Significant changes have reduced the error rate by 28.7%. The pitch detector is also able to accurately detect the pitch of the synthetic speech-like signals and to capture the jitter present in the signals.

1. INTRODUCTION

Most of the algorithms used to detect aperiodicity are passive, i.e. non-silent regions with little or no voicing are labelled as aperiodic and the amount of aperiodicity is estimated using indirect measures like zero crossing rate, high-frequency energy and ratio of high-frequency energy to low-frequency energy. These measures are prone to making errors in situations where the signal has simultaneous strong components of both periodic and aperiodic energies, 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. A system that can reliably detect and quantify the amount of periodic vs. aperiodic energy in the speech signal has many applications including speech coding, speech recognition and speaker recognition. The inherent problem in developing such a system is to define aperiodicity in such a way that maximum and minimum possible aperiodicity roughly correspond to minimum and maximum possible periodicity respectively in a given signal but at the same time aperiodicity should not merely be a complement of periodicity. The system presented in [1] defined periodicity as the degree of regularity in the minima of the Average Magnitude Difference Function (AMDF) computed from the envelope of the filter channels. Aperiodicity was related to the degree of randomness in the minima of the AMDF waveform. For a given time instance, the system was able to grade the amount of periodicity and aperiodicity across the frequency channels, but it made a binary decision in each of the channels. The result is that the output of the system had a fine temporal resolution but a crude binary spectral resolution of the amount of periodicity/aperiodicity. The modifications presented here are able to successfully grade the amount of periodicity/aperiodicity at each of the frequency channels.

The structure of the periodicity/aperiodicity system is very similar to a pitch detection algorithm and includes estimation of the pitch of the periodic component of the signal. Pitch (the fundamental frequency of voiced speech) is defined as the frequency of vocal fold vibration. The pitch algorithm is relatively simply and is independent of the voicing decisions made by the system. The pitch detection algorithm presented at [2] was very prone to committing pitch halving and doubling errors and was unable to reliably detect pitch in weakly periodic regions and near the voiced-unvoiced boundaries. In this paper we present improvements to our Pitch Detection Algorithm (PDA) and also compare our pitch detector with the other existing PDAs. These improvements were able to reduce the error rate by 28.7%.

2. SYSTEM REVIEW

This section provides a brief review of the original algorithm. The signal analysis begins by passing the signal through a 60-channel auditory gammatone filterbank with Characteristic Frequencies (CFs) based on physiological data[3]. The temporal envelope of each channel is computed using the Hilbert transform. The channel envelope of every non-silent channel is analyzed for periodicity and aperiodicity using the short-time Average Magnitude Difference Function (AMDF). The AMDF 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 envelope signal, k is the lag value in samples and w(m) is the window. For the work presented here, w(m) is a rectangular window of 20ms.

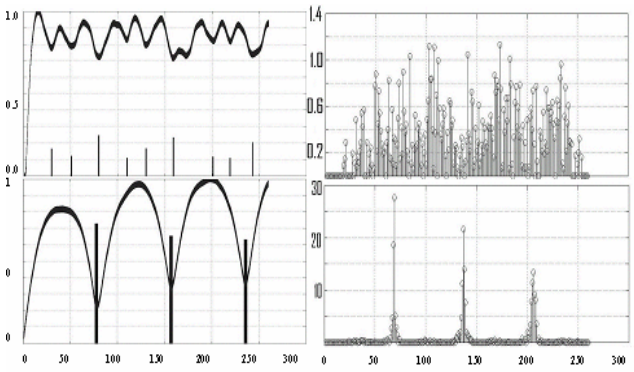
2.1. Periodicity & Aperiodicity Calculation

For a strictly periodic signal, the AMDF will attain minima (referred as dips) equal to one at lags equivalent to the pitch period and its integer multiples. For aperiodic signals these dips are at random locations (Figure 1(a)). When these strengths are added across all the channels for each lag, the output will have clusters at the pitch value and its integer multiples for strongly periodic frames. For aperiodic frams the dips will be randomly

scattered over the range of possible lag values and will have no prominent clusters (Figure 1(b)).

For each cluster a *periodicity confidence* measure is computed by summing the strengths of all the dips that lie within a certain neighborhood of the peak of the cluster. Sum of the strenghts of the dips outside this neighborhood is the *aperiodicty confidence* of that cluster and such dips are labeled as spurious dips. The difference of *periodicity confidence* and *aperiodicity_confidence* is the *summary confidence* of the cluster. All the clusters with negative *summary confidence* are dissolved. If no cluster survives this test, the frame is likely to be aperiodic. Final decision about the periodicity and aperiodicity of the frame is postponed till all the channels in the frame are analyzed. Periodic channels are defined as channel with less than two spurious dips. Aperiodic channels are defined as the channels with more than one spurious dip.

Figure 2 shows the decisions made for four channels from the voiced fricative /z/. Notice that channels (a) and (b) have no dips outside the cluster tolerances, i.e. no spurious dips, and so they are classified as periodic. Channels (c) and (d), on the other hand, have a number of spurious dips and, therefore, they are classified as aperiodic.



(a) (b)

Fig. 1. 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 periodicity/aperiodicity decision is binary at the channel level. In particular, there is no contribution of the strength of dips towards the channel periodicity or of the number of spurious dips towards channel aperiodicity. As a result, channels (a) and (b) are both labelled periodic and there is no distinction that specifies that channel (a) is strongly periodic compared to the weakly periodic channel (b). Similarly, channel (c) is strongly aperiodic compared to the weakly aperiodic channel (d) although this gradation is again not specified.
Section 3 describes the modifications made to change this binary decision into a graded decision.

2.2. Pitch Detection

The system outputs a pitch value only when at least one of the clusters has a positive *summary confidence*. The peak of the cluster with the maximum *summary confidence* is the pitch estimate for that frame and the corresponding *summary confidence* is the pitch confidence. When two or more clusters have comparable *summary confidences*, the cluster closest to the previous pitch values is chosen. The details of this pitch detector can be found in [2].

There are two types of pitch errors. First, no pitch is detected in very low amplitude regions (e.g., a weak /w/) and in some transition regions between voiced and unvoiced sounds. In both cases, the cluster strengths are significantly lower than they are in high amplitude periodic regions (e.g., the middle of a vowel). As a result, the aperiodicity confidences for clusters in these frames are equal to or more than the periodicity confidences and the clusters have non-positive summary confidences. Thus no pitch is outputted and this leads to errors.

The second type of error is the fluctuations in the estimated pitch values. It was shown in [2] that the PDA is prone to making a significant amount of halving errors. The modifications proposed in section 3 refine the continuity constraints but are also able to track the valid fluctuations in pitch contour. The modifications also address the issue of having no pitch outputs in some of the weak amplitude signals.

3. PROPOSED IMPROVEMENTS

3.1. Graded Periodicity/Aperiodicity Channel Decision

Since the strength of the AMDF dips found at multiples of the pitch period gives an estimate of the signal periodicity, a natural improvement would be to include this parameter in our measurements. At the same time, quantifying the randomness in the distribution of the dips can capture the degree of aperiodicity.

In our new periodicity measurements, these objectives are accomplished by weighting the normalized strength of each dip such that dips closer to the pitch period and its multiples contribute more towards periodicity. This contribution decreases rapidly with increasing distance from these locations. Consequently, we found that exponentially decaying weights perform better than linearly decaying weights. Because speech is only expected to be quasi-periodic, the weights in the immediate vicinity of a pitch multiple are set to unity so that dips closer to pitch multiples are not unduly penalized.

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, i.e. If the detected pitch of the frame is such that it can accommodate N pitch multiples in the lags, then each of the *[nF0-F0/2: nF+nF0/2] for n=1,2..N* regions will be 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 j^{th} cluster.

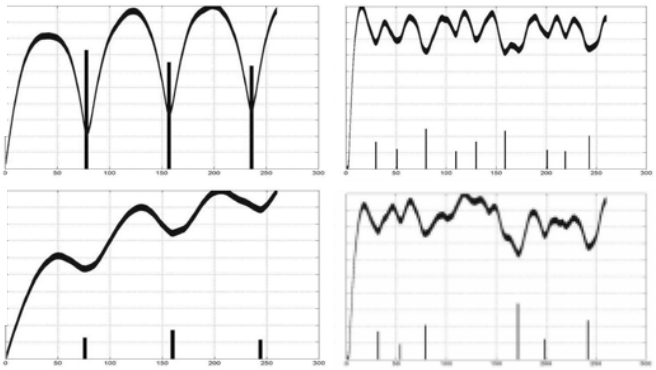


Fig. 2. The left two frames show periodic channels whereas the frames on the right show aperiodic channels.

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

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 preliminary value of periodicity.

The AMDF dips in channels that are predominantly aperiodic are located far from the pitch period and its multiples, are small in amplitude, and are generally numerous. The preliminary measurement of aperiodicity also utilizes weighted strengths of AMDF dips with two important considerations. First, dips far from the pitch period and its multiples should contribute close to their full value towards aperiodicity. Thus, logarithmically increasing weights are used. Second, the strength of aperiodicity should be directly related to the number of spurious dips. Thus, we take the sum of the dips instead of the mean across the clusters.

It should be noted that the sum of the preliminary periodicity and aperiodicity measures may be greater than one. If this is the case, they are scaled down proportionally so that the sum equals one. This approach makes it possible to obtain overall values of aperiodicity close to unity. These periodicity and aperiodicity measures are then multiplied by the corresponding channel energies and summed across the channels to get the proportion of periodic and aperiodic energies for the frame.

3.2. Pitch Improvements

For a typical low amplitude signal the strength of the dips in is much smaller compared to that for a high amplitude periodic. In fact, the strength of the dips in the low amplitude signal are comparable to those in computed from an aperiodic frame. As a result the aperiodicity confidence will be equal to or greater than the periodicity confidence resulting in no clusters and hence no pitch output. It has been noticed that the ratio of the number of lags that have non-zero values to the total number of lags is much smaller in the case of the weakly periodic frame than it is for the aperiodic frame. This ratio was trained using a small amount of pilot data and it was found that a value of 0.55 gives significant separation between weakly periodic frames and aperiodic frames. When the periodicity measure of a cluster falls below a pre-determined threshold called *per_thresh*, the ratio of number of lags with non-zero values to the total number of lags is computed and if it below 0.55 the aperiodicity confidence is set to zero. Thus some clusters will have a positive summary confidence and the corresponding frame will have a pitch value. When the ratio is above 0.55 the apriodicity confidence is computed and subtracted from the periodicity confidence. This aviods outputting pitch values for aperiodic frames.

The second avenue for improvement was to reduce the errors caused by pitch halving and doubling. This is accomplished as follows: At the initial frames of the utterance the peak corresponding to the cluster with maximum summary confidence is chosen as the pitch value. As the analysis progresses, clusters are forced at the median of the pitch values from the previous frames and at its integer multiples. By default, the peak corresponding to the first cluster is chosen as the pitch value. To allow the flexibility to change the pitch value, a cluster is formed at half the pitch value and if its summary confidence is greater than *per_thresh* the pitch value is set to the peak of this cluster. This allows the system to rectify its pitch halving errors. If the summary confidence of a cluster at the integer multiples of pitch value is greater than *2*per_thresh* and the summary confidence of the first cluster (i.e. cluster at pitch value) is less than the *per_thresh* the pitch value is changed to the peak of this new cluster. This allows the system to rectify the pitch doubling errors. At the same time, these criteria allow the algorithm to track the pitch correctly even when it is actually halved.

4. RESULTS

The system was evaluated on a speech database that had Electroglottograh (EGG) data recorded simultaneously and on a database of synthetic speech like signals. The speech database consists of 50 utterances spoken by one male and one female subject in clean environment [4]. The synthetic database is the same as the one used in [1]. This database consists of signals that are outputs of a 50-pole LPC synthesis filter when it is excited by a pulse train corrupted with Gaussian white noise. Pulses at frequencies 131 Hz, 120 Hz and 200 Hz at SNRs varying from Inf to –5dB. To evaluate the performance of our periodic and aperiodic detector, we compared the SNR based on these measures with the known SNR of the synthetic signal and with the SNRs obtained by our old method. We define the SNR based on our measures as:

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

where v is the periodic energy and *u* is the aperiodic energy calculated by our detector. Fig. 3 shows the actual SNR versus the computed SNR for the pulse with frequency 131 Hz for the old system and for the refined system after incorportating the above mentioned modifications.

As can be inferred from the figure, the improved system is able to track the SNR more closely compared to the old system. To

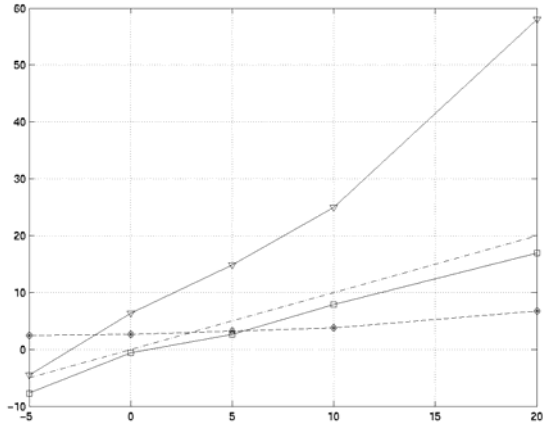


Fig. 3 The dashed line shows the actual SNR. The line with triangles represents the old results. The line with squares represents the new results. The dashed line with circles shows the results obtained from cepstral coefficients.

compare the performance of our parameters with that of other standard set of parameters we also evaluated the SNR using cepstral coefficients. The high pass liftered version of cepstrum was used for this purpose. The maximum value of cepstrum in a small neighborhood of the expected pitch value was taken as the periodicity confidence and the average of the rest of the cepstrum was taken as the aperiodicity confidence. The SNR values obtained using the cepstral coefficients are shown in Figure 3. Its evident from the figure that our periodicity and aperiodicity measures outperform the cepstral coefficients.

The performace of the system was also evaluated on the natural speech database. The evaluation was made on a frame basis at a frame rate of 2.5ms. The periodicity accuracy is defined as the ratio of the number of non-silent frames where the periodicity measure was above 0.3 and the EGG output was non-zero. The aperiodicity accuracy is defined as the ratio of the number of non-silent frames where the aperiodicity measure was above 0.3 and the EGG data was zero. The previous system had a periodicity accuracy of 88.8% whereas the refined system give periodicity accuracy of 95.1%. This is a significant improvement. The aperiodicity accuracy for the old system was 92.7% whereas it is 87.8% for the new system. Although there is a slight drop in the aperiodicity accuracy the overall performance is improved.

The pitch detection algorithm was also evaluated on this database. The refined pitch detector is able to reduce the pitch error from 9.8% to 6.5% thus giving a 27.8 % reduction in the relative error.

5. CONCLUSION

The refinements proposed here have significantly improved the performance of the system. One application of the periodicity/aperiodicity measures and pitch will be in our speech recognition algorithms. These parameters also form a part of a landmark detection system where the main emphasis is broad classification of speech signals using strictly temporal cues.

REFERENCES

[1] Om Deshmukh, Carol Espy-Wilson, "*A measure of Periodicity and Aperiodicity in Speech*", in Proc. IEEE ICASSP 2003, Hong Kong, pp. 448-451.

[2] Om Deshmukh, Carol Espy-Wilson, "*Detection of Periodicity and Aperiodicity in Speech Signal Based on Temporal Information* ", The 15th International Congress of Phonetic Sciences, Barcelona, Spain, 2003.

[3] 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.

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