NONPARAMETRIC PEAK FEATURE EXTRACTION AND ITS APPLICATIONS TO SPEECH SIGNALS

HyunSoo Kim and W. Harvey Holmes
School of Electrical Engineering, The University of NSW,
UNSW SYDNEY NSW 2052, Australia

ABSTRACT: We have developed a novel peak feature algorithm that can be used for endpoint detection, transient extraction and segmentation of speech. It uses binomial probabilities to measure peak characteristics and estimate sets of endpoints to characterize an utterance. In first tests our algorithm appears to outperform existing methods. It can also be used to improve the segmentation capabilities of existing segmentation methods, and can probably benefit most speech recognition approaches. In addition to the utility of the first order peaks, the statistics of the higher order peaks are also effective for feature extraction and segmentation. The second order peaks can also be used to devise an intelligent update procedure for the feature data windows, where the window update rate changes based on the type of speech signal present. The nonparametric peak feature algorithm is flexible, efficient and very robust in noise.

INTRODUCTION

Peaks are among the most reliable components of any time series. They stand tall above the noise, unlike zero crossings, and our eyes are drawn to them when trying to obtain an impression about temporal behaviour. Peaks are also easy to measure, only requiring simple comparisons to locate. It will be demonstrated here that peak statistics can be used to give reliable and robust estimates of speech endpoints and can also be useful in segmentation.

We will propose here some novel features based primarily on peak analysis. The improved endpoint and segmentation capabilities provided by the peak feature detector benefit most speech recognition approaches. The peak feature detector can be used to investigate the effects on speech features of fixed versus variable feature window length and update rates. We will discuss a new method of varying the data windows by a few peaks to improve the stability of certain features, and also discuss window alignment based on the peak information.

In addition to the usual first order peaks we also consider second order peaks, which are defined here to be the peaks of the time series formed by the first order peaks – that is, they are the "peaks of the peaks". Third and higher order peaks can be defined in a similar way. The statistics of the higher order peaks will also be shown to be effective for feature extraction. The nature of the higher order peaks is to have higher levels, on average, than lower order peaks. Higher order peaks also occur less often – there are fewer second and third order peaks than first order peaks. Peak rate of occurrence is also an interesting feature of higher order peaks, and it is the second and third order peaks that contain more reliable pitch period information. Another interesting peak feature that can be made use of is the time, or number of points, between second or third order peaks.

NONPARAMETRIC PEAK FEATURE EXTRACTION ALGORITHM FOR SPEECH DETECTION

The peak feature extraction algorithm proposed here is motivated from sonar signal processing, where Zakarauskas (1991) proposed transient energy detectors based on peak analysis using the binomial distribution. In our case the null hypothesis H_0 is that only background noise is present, the alternative hypothesis H_1 is that speech is also present.

The peaks are extracted and their sound pressure levels are incorporated into a voltage histogram that represents an estimate of the probability density function of the peak amplitudes. When background noise only is present, the first task is to find the voltage level L above which a given proportion r of the peak amplitudes lie. This threshold L is then used to convert a sequence of peaks in the received signal into a binomial sequence of ones or zeros. If the peak is greater than L then it becomes a 1, if the peak

is less than L it becomes a 0. Following this logic, the binomial peak sequence relative to the threshold may look something like 1100011110001111, which closely resembles a "coin toss" experiment. Speech detection is based on the fact that, for background noise alone (i.e. in the absence of speech), the probability that there are N or more peaks above level L in a group of W peaks is given by the following sum of binomial coefficients:

$$P(r,N,W)=\sum_{i=N}^{W}\binom{W}{i}r^{i}(1-r)^{W-i}.$$

The parameter W controls the length of the integrator window, which is moved one peak at a time forward in time. If N is close to W, this probability is very small, so that speech may be inferred if large values of N are observed. Effectively, this method looks for unlikely combinations of high amplitude peaks to predict whether or not the set of peaks corresponds to background noise or to speech. It is a nonparametric test (i.e. independent of any assumptions about the underlying statistics of the signal or noise). If an integration window W is selected then the number of ones in the window can be counted to determine if a signal is present. The functional diagram for the complete algorithm is shown in Figure 1.

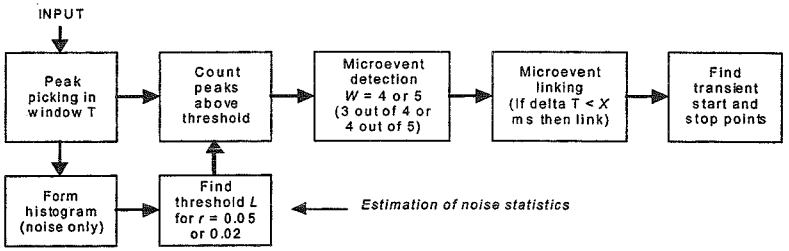

Figure 1. Peak feature extraction algorithm.

In this work we call sequences of peaks with small window lengths (such as "3 out of 4" or "4 out of 5") microevents. They appear to be the smallest, consistent packages of peaks that can be reliably detected.

When the microevents pass the temporal relationship thresholds, they can be linked together. A long chain of linked microevents leads to a valid detection of part of an utterance. However because each utterance is comprised of many microevents, there is uncertainty about the actual start and stop points when there are short gaps between microevents. Hence link criteria must be used to associate the microevents. In earlier studies of the nature and temporal consistency of speech such as that by Reaves (1991), it was suggested that two speech sounds could be related even if they were separated by up to 150 ms, but in this research 40 ms was found to work well for linking energy across pauses in the middle of spoken words. Because of this discrepancy, the microevent-linking threshold is shown in Figure 1 as X ms. Note that the linking threshold can also change somewhat depending on the value of L or even the rate of speech, with 25 ms working well in some cases. This algorithm provides a best set of endpoints that characterize an utterance. Second and third best endpoints can also be estimated, depending upon the link criteria.

The same process can be used on the higher order peaks, since they are subsets drawn from the first order peaks. One of the key advantages of this algorithm over previous energy detection algorithms is that it is better able to identify the starts and stops of the transient events within a few data samples, because it uses peaks instead of broad energy pulses or frames or blocks.

OPERATIONAL PARAMETER SELECTION FOR THE PEAK FEATURE DETECTOR (PFD)

We now briefly discuss the choice of detector parameters to optimise this algorithm for speech detection. To enable the reliable detection of voiced speech, the count ratio r (CR) must be chosen. Since the PFD is a nonparametric detector, it does not require any *a priori* information about the background because the adaptive noise threshold is determined automatically and adaptively. It is possible, however, to examine the simple case of a white Gaussian signal in a white Gaussian noise background and analytically extract the operating ranges for the PFD parameters. This example allows us to treat the nonparametric detector as a parametric detector. Note that although this example is a special case, it closely corresponds to the situations where plosives are to be detected in speech data in a white Gaussian noise background. The PFD has two groups of interrelated parameters to be chosen. First, the amplitude consistency threshold setting must be defined. This threshold drives the amplitude boundary between the noise and any incoming signal. Secondly, the peak-frequency or rate-of-occurrence thresholds must be determined. These thresholds set watch points to detect signals.

In the discussion of Zakaruskas' (1991) transient energy detector, the binomial cumulant of the noise was used to determine the probability of a false alarm as

$$P(FA) = \sum_{k=l}^{m} \binom{m}{k} p_n^k \, q_n^{W-k}$$

and we can write the likelihood ratio test for the PFD as a comparison between two different sums of binomial coefficients:

$$\sum_{k=l}^{n} \binom{n}{k} p_s^k \, q_s^{W-k} \underset{H_0}{\overset{H_1}{>}} \sum_{k=l}^{n} \binom{n}{k} p_n^k \, q_n^{W-k} \qquad (1)$$

where $H_0 : r_i = n_i$ and $H_1 : r_i = s_i + n_i$ for i = 1, 2, ...,N. Considering the simple binary hypothesis problem where the observations consist of a set of N statistically independent values, under H_0 only noise is present and under H_1 both signal (speech) and noise are present. The test compares the binomial cumulant of the signal distribution to the binomial cumulant of the noise. In this likelihood ratio test the binomial cumulants are also sufficient statistics. In the practical implementation of the PFD the expressions above are not computed directly. Instead, look-up tables are used to determine the threshold settings in different noise-peak distributions. These threshold settings are based on the peak histograms and ultimately determine the peak amplitude consistency settings.

The terms in equation (1) can be used to generate a "receiver operating characteristic" (ROC) curve because they represent sufficient statistics, and they define the probability of detection and failure. The term on the right hand side of (1) is the probability of false alarm $P(FA)$ for the peak feature detector, since it describes the area under the upper tail of the noise density function. Similarly, the expression on the left hand side is the probability of detection $P(D)$ which describes the area under the upper tail of the signal+noise density function.

The PFD has several interrelated parameters. First, the noise threshold p_n must be set, which in turn sets p_s, depending upon what level and type of signal appears. Then the parameters N and k need to be set according to the context of what is being detected. Properly setting N and k is the key to the success of the PFD. Note that N and k determine $P(D)$. For example, if the signal+noise peak density function has a peak far to the right, which implies a very loud signal, and little overlap occurs with the noise distribution, then $P(D) \approx 1$. However, $P(FA)$ would still be non-zero because it depends only on the noise density function tail above the threshold. We can generate a ROC curve for various settings of N and k. Each "k out of N" scenario can be thought of as a separate detector. Figure 2 contains example ROC curves of the many PFD parameter settings with the count ratios r = 0.1, 0.05 and 0.02, for N =10

and 5, and with *k* ranging from 1 to 10 and 1 to 5, respectively. The information from this analysis can be used to select acceptable operating ranges for the PFD.

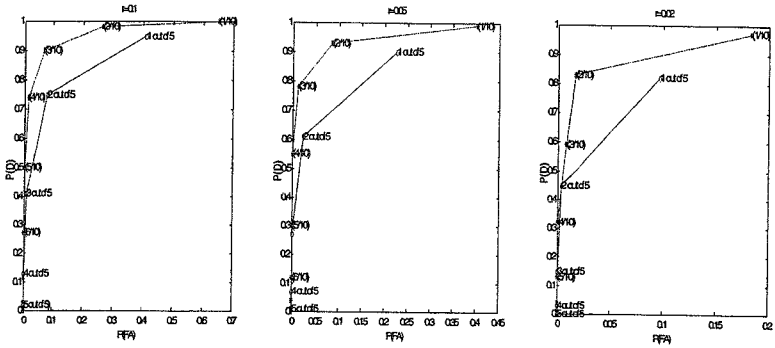

Figure 2. ROC curves for "*k* out of 10" (dashed) and "*k* out of 5" (solid)

Reaves (1991) lists several time-based limits on processing speech through energy detectors and his precision was only 40 ms for starting or stopping endpoints. To ensure that the PFD does much better than this, the maximum false alarm (FA) rate for the PFD should be set to control what microevents are linked together. A useful FA boundary can be drawn from Reaves' work, in which pulses of speech energy separated by more than 150 ms were almost always associated with separate utterances. Thus, if false alarms are required to be more than 150 ms apart they will not be linked together incorrectly. One FA every 150 ms converts to 6.67 FAs per second. This seems high, but at this setting the peak detector will clearly be able to resolve endpoints precisely. To compare this FA limit to the settings, the P(FA) for three count ratios r = 0.1, 0.05 *and* 0.02, and for "k out of 5" PFD parameter settings, are converted to FAs per second in Table 1. The cells with asterisks are satisfactory operating points for the PFD, assuming FAs must be less than one per 150 ms at the 8 kHz sampling rate.

Table 1. Number of FAs per second for "k out of 5" peaks exceeding the threshold

	r = 0.1	r = 0.05	r = 0.02
k=1	218	121	51
k=2	43	12	2*
k=3	5*	0.6*	0.04*
k=4	0.3*	0.02*	0.004*
k=5	0.005*	N/A	0.00002*

We have discussed the optimal selection of operational ranges for the peak feature detector. In our analysis, the threshold was set as low as possible to stay within the operational boundaries discussed earlier. The ideal amplitude consistency setting, or noise threshold, which is the probability of finding a high level peak above a threshold, was found to be between 0.05 and 0.02. The rate-of-occurrence threshold setting was "3 out of 4" or "4 out of 5" to provide good detection performance and consistent utterance endpoint identification.

PEAK FEATURE PERFORMANCE IN ENDPOINT DETECTION AND OTHER APPLICATIONS

The peak feature algorithm has many speech processing applications, but the easiest to demonstrate is endpoint detection. We have compared the peak feature detector to three detectors, based respectively on energy, zero crossings, and a combination of energy and zero crossings, as used by Rabiner and

Sambur (1975), which has been the baseline for a number of later comparisons. In Table 2, typical results are summarized for two utterances of the word "eight" in both low and high noise, where A, B, A' and B' are the estimated first utterance start and stop points, and C, D, C and D' are the corresponding points for the second utterance.(the errors are given in brackets). Here, A', B', C' and D' are from words in high noise (peak SNR = 15 dB, with much lower average SNR), while A, B, C and D are from relatively clean words with peak SNR = 30 dB. The peak feature detector showed much more accurate results than the other detectors, especially in the high noise case. This is largely due to the fact that the detector exploits the highest amplitude peaks of the data stream, which tend to ride on top of the signals and noise regardless of the characteristics of the data. The performance of the new peak algorithm in noise also degrades in a predictable and graceful fashion.

Table 2. Comparison of endpoint detection estimates for five detection methods: 1=Manual (ideal), 2=Energy, 3=Zero Crossing, 4=Energy+Zero, and 5=New peak algorithm. The errors are given in brackets. See text for explanation of the letters A...D.

	A	B	C	D	A'	B'	C'	D'
1	13900	17500	28635	32400	13900	17500	28635	32400
2	13996 (+96)	17748 (+248)	28773 (+138)	32611 (+211)	10002 (-3898)	N/A(-)	N/A(-)	37427 (+5027)
3	14657 (+757)	17755 (+255)	28929 (+294)	32772 (+372)	14890 (+990)	14008 (-3492)	29896 (+1261)	30125 (-2275)
4	13996 (+96)	17735 (+235)	28773 (+138)	32772 (+372)	10002 (-3898)	N/A(-)	N/A(-)	37427 (+5027)
5	13903 (+3)	17529 (+29)	28633 (-2)	32388 (-12)	13926 (+26)	17652 (+152)	28574 (-61)	32265 (-135)

The peak feature algorithm has more applications than just endpoint detection. One potential application is to improve the segmentation capabilities of the existing approaches. By combining the peak frequency consistency information with the peak amplitude consistency information, a very useful segmentation statistic has been developed, which in first tests satisfactorily segmented the word into its underlying phonemic structure (this is the first segmentation stage, in which the phonemes are not actually identified). Higher order peak statistics have also been found effective for feature extraction.

These techniques offer the potential to be much more accurate than block-oriented methods, since the boundaries are not limited to block boundaries. They can also be used to improve speech recognition processing. For acoustic-phonetic speech recognizers, peak feature information could be used as "data" features for input into an early segmentation stage, as just described. The peak features could also be used as "signal" features in the late segmentation stage to help determine which phoneme is present. In other speech recognizers such as Vector Quantizers (VQ), the peak features may also prove useful as part of the codebook. Other interesting peak feature that can be made use of are the times, or number of points, between the first, second or third order peaks. These characterize the frequency components of the time series without the need for computationally expensive formant extraction, and can be used as the basis for simple and effective pitch estimation algorithms.

INTELLIGENT WINDOW ALIGNMENT USING HIGHER ORDER PEAKS

Speech processing systems have traditionally employed fixed length data windows for computing features. When fixed window update rates are used the method could be called an "unintelligent" update procedure (UUP), since no logic is used to determine the update rate or the start and stop points of the data windows, and the features extracted are generated without regard for the data in the window. The second order peaks can be used to devise an "intelligent" update procedure (IUP) for the feature data windows, where the window update rate changes based upon the type of speech signal present.

Thus, consider an IUP that uses the second order peaks to align the start of the data windows. The goal of the IUP is to improve feature repeatability when presented with a consistent speech signal (e.g. voiced speech). One way to do this is to shift the data window so that the next data window is more correlated

with the previous window. Second or third order peaks provide such a shifting mechanism, based on the fact that they allow the data window to align to the peak of the glottal waveform. The result is a higher degree of correlation between adjacent data windows, especially for voiced sounds, which should translate into improved feature repeatability and stability (i.e. reduced frame-to-frame variability).

To demonstrate the effectiveness of this IUP, we employed cepstral coefficients and used voiced sounds. By splitting the "consistent" region of data into 80 separate 128 -point data windows, the variability of the cepstral coefficients can be evaluated. The solid top curve in Figure 3 is a plot of the standard deviation of the cepstral coefficients as a function of time for the UUP using a fixed update window of 128 points and no overlap. The middle dashed curve is based on the IUP with an update of 128 samples plus the number of samples required to get to the largest second order peak within 0 to 30 points. This had the effect of *positioning* a dominant second order peak at the start of the data window, but not necessarily the maximum peak in the glottal pulse. The bottom dotted curve was computed using the IUP with an update of 128 points plus or minus the number of points required to get to the maximum second order peak within ±30 points, which almost guaranteed that the data window would start on the pitch peak. The IUPs reduced the standard deviation significantly.

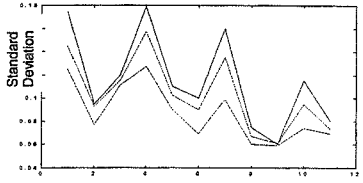

Figure 3. Variance reduction using the IUP (for the /EY/ sound). See text for description.

CONCLUSION

A new speech feature based on the analysis of peaks of digitised acoustic time series has been introduced. This gives a simple, flexible and efficient approach to speech processing that can be used for endpoint detection, glottal closure detection, and landmark detection or segmentation, depending on the type of features employed. The algorithm is based on a nonparametric test using binomial probabilities of certain peak characteristics. The statistics of the higher order peaks have also been found to be effective for feature extraction, including an intelligent window update and shift mechanism for improving feature extraction. The new method has several advantages over existing methods for endpoint determination and other applications, especially in noisy environments. It can also be combined with existing features to obtain potentially even better performance. The new algorithm is data-flexible and application-flexible.

REFERENCES

Rabiner, L. and Sambur, M.R., "An algorithm for determining the endpoints of isolated utterances", *Bell System Technical Journal*, pp. 297-315, February 1975.

Reaves, B., "Comments on: An improved endpoint detector for isolated word recognition", *IEEE Trans. on Signal Processing*, vol. 39, no. 2, February 1991, pp. 526-527.

Zakarauskas, P., Parfitt, C.J and Thorleifson, J.M., "Automatic extraction of spring-time Arctic ambient noise transients", *J. Acoust. Soc. Am.*, vol. 90, July 1991, pp. 470-474.