DETECTION OF FOCAL POINTS IN SPEECH PROSODY

*Lucia Valbonesi** *Rashid Ansari** *Susan Duncan*** *Karl-Eric McCullough*** *David McNeill***
*Francis Quek****

* University of Illinois at Chicago, Department of Electrical and Computer Engineering, Chicago, IL, USA
** University of Chicago, Department of Psychology, Chicago, IL, USA
*** Wright State University, Department of Computer Science and Engineering, Dayton, OH, USA

ABSTRACT

Signal processing tools are developed to help automatically detect significant events or focal points in the speech and gesture traces of audio-visual data and investigate their temporal correlation as part of a multi-disciplinary effort to assemble the computational resources for facilitating research in gesture and speech interaction. One key task in this research is to determine focal points in the speech signal by analyzing its acoustic-prosodic features. This task requires processing and analysis of an immense amount of data. In order to make the task efficient and less time consuming, algorithms for speech processing are developed to perform automatic feature extraction, parametric representation, and detection of focal points that are usually painstakingly marked by hand. In this paper a method for detecting the focal points in speech prosody is presented and the results are compared with focal points hand-marked by experts. The method produces a high rate of successful detection.

1. INTRODUCTION

Study of spontaneous gestures and accompanying speech provides a window into the nature of speakers' mental representations and processes and it is of interest in disciplines such as psychology, psycholinguistics, linguistics, anthropology and neurology. Such a study is usually carried out by perceptual analysis, i.e. analysis by unaided ear and eye, performed by experienced analysts. However perceptual analysis of large amounts of data is extremely time consuming for analysts as all the analysis and detection is performed manually [1, 2]. Analysts need to view the video data and listen to the audio data several times before identifying the events of interest in the audio-video stream.

The need for an automated analysis of gesture and speech data led to a collaboration among researchers in the fields of signal processing, computer vision, and psychology. Signal processing tools have been developed as part of this collaborative effort to help automatically detect significant events or focal points in speech and gesture traces of audio-visual data and then to investigate their temporal correlation.

This paper describes an algorithm for automatically detecting focal points in speech through the identification and extraction of F0 and amplitude cues.

Two sets of audio-visual data obtained from two natural discourse elicitations are analyzed. One data set is used as a "training" set in order to establish rules for identifying points of emphasis. The remaining data set is then used as a "test" set to check and confirm the results. The two sets of data are monologues in English spontaneous speech that is "cleaned" in order to avoid detection errors and thereby improve the performance of the algorithm.

A "focal point" is defined as an instant in speech that locally corresponds to the greatest amount of emphasis. In order to identify the position of a focal point traces of two prosodic features of speech, the fundamental frequency F0 and the amplitude, are segmented and analyzed separately. Five parameters are extracted from these traces and they are then used to detect focal point positions, using rules determined from the training set. The larger the number of cues that are located at the same instant of time the higher is the probability that such a position is a focal point. Different types of combination of cues are possible based on the rules established after analyzing the training set.

Comparing the results of this algorithm with focal points hand marked by experienced analysts, the percentage of successful detection of focal points was greater than 85%, where success is defined by agreement with focal points hand-marked by analysts.

The approach to investigating focal points based on the fundamental frequency trace is suggested by the research carried out in the Phonetics department at the University of Bonn [3, 4]

In our work, the algorithm of the University of Bonn group is modified by introducing a different method of segmentation and using a set of thresholds adjusted to the speaker's speech rate. A larger number of cues are used in our method.

The results of the focal point detection show a clear improvement when compared with those obtained at the University of Bonn.

2. EXPERIMENTAL DATA

The sets of data processed using the proposed algorithm are obtained from two different natural discourse elicitations.

While the algorithms are developed for analyzing a monologue, the actual experiments involve a subject speaking to an interlocutor. In the first experiment two subjects are recruited to serve as speaker-interlocutor pair. A female speaker is asked to describe a strategy to surround and evacuate intelligent wombats from a village. This dialogue is of a long duration. The disadvantage in using this set of

data is that the two subjects take turns at speaking. Without a sophisticated automated technique for separation of speakers the extracted features cannot be accurately associated with the person who is actually speaking. In order to "transform" the first set of data into a monologue, the audio segments corresponding to the interlocutor are eliminated manually.

Figure 1. Frame of Sue data

In the second experiment, a female speaker describes her house to an interlocutor. This experiment is mainly a monologue, as the interlocutor rarely interrupts the speaker.

Figure 1 shows a frame extracted from the recording of the second experiment.

For each data set, the speech is sampled at 8 KHz and processed to obtain the fundamental frequency F0 and the amplitude measured as root mean square (RMS) value over a moving window.

The first set of data, referred to as "Wombat" data, is used as "training" set for the algorithm. The results are then checked using the second set of data that is referred to as "Sue" data.

3. SEGMENTATION AND PRE-PROCESSING

The raw speech files are first processed by Entropic's xwaves software to extract F0 and amplitude data.

The F0 information extracted from the recorded audio data is inaccurate, mainly due to the presence of noise. The F0 data is filtered and then segmented into speech units that are suitable for analysis.

This is done by first identifying voiced sub-segments that are constitutive elements of a speech unit. A voiced sub-segment consists of a purely voiced stretch of data. It is separated on either side from the rest of the speech by unvoiced sound and/or silence, referred to as "pauses" or "unvoiced breaks" between voiced sub-segments.

Segments consist of one or more contiguous voiced sub-segments together with the unvoiced breaks separating the constituent sub-segments. A segment is bounded on either sides in its interior with voiced sub-segments and it is separated from other segments by unvoiced breaks whose duration exceeds a threshold. In the initial analysis a time interval of 300 ms is used as threshold because it is considered long enough to be a typical lower bound on a break between two adjacent phrases in speech. This criterion yielded a very good match with phrases marked by expert linguists.

However not all the speakers have the same speech rate. In order to account for differences in speaking rate, the segmentation procedure is modified to include an adaptive threshold that assumes different values depending on the ratio of speech and pause in the preceding and following sections. The threshold changes not only from speaker to speaker but also during the discourse of the same speaker.

The purpose of this segmentation is to isolate units of speech data within which the characteristic parameters of the audio information can be estimated for further analysis.

The next step following segmentation is to correct errors made in estimating F0 that usually occur in a burst in the F0 data generated by xwaves. The errors are corrected by eliminating abrupt changes [5] in the F0 data and by maintaining a consistency in the mean according to the average behavior of the adjacent frequency samples.

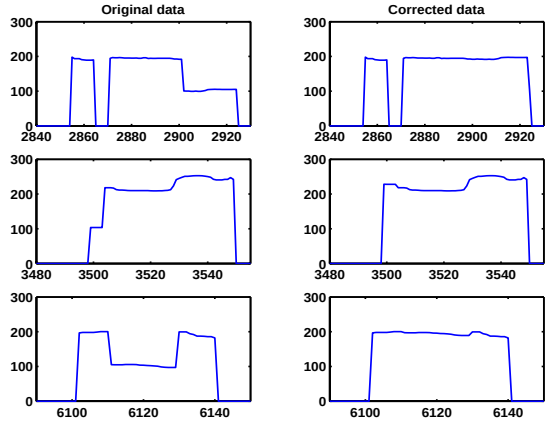

Figure 2. Examples of corrected segments

Figure 2 shows three examples of speech segments of the "Wombat" data before and after the correction of the burst error.

Pre-processing is not as critical for amplitude data as for F0 data. This is not because the noise is smaller, but because in our analysis we are more interested in the position and value of the peaks of the amplitude, and these are less sensitive to the presence of noise and they can be retrieved reliably even in a noisy environment.

4. EXTRACTION OF PARAMETERS

Once the F0 traces are pre-processed and F0 errors are corrected, significant cues to emphasis in the data are identified and extracted from both the F0 and the amplitude data.

An algorithm is developed for the purpose of extracting significant cues that will be combined together in order to detect the focal accent positions, according to "rules" that are identified through a training process and later checked with the second set of data or test data. Three of the cues are determined from the F0 information, while the other two are determined from the amplitude information.

Each segment is examined for the following cues:

- The F0 peak in the segment
- The first and second largest negative peaks of the F0 gradient
- The first and the second largest peaks of the amplitude

These cues are considered together in determining points of emphasis in speech.

4.1. FUNDAMENTAL FREQUENCY F0

Accordingly to the thesis and the results obtained at the University of Bonn on focal accent detection [3, 4], the focus of an utterance does not necessarily coincide with the highest peak of F0. The fact that the fundamental frequency falls toward the end of an utterance has to be factored in.

Not only is the location of the peak important, but also the gradient of F0, because the focus of the utterance is found to be in the area of the steepest fall in the F0 trace. Therefore the points with the highest negative gradient are the ones that serve as cues to the focus. Consequently the algorithm looks not just for the highest peak within a segment, but also for the peak closest to the negative maximum of the gradient.

In the procedure for determining the focal points, first the F0 data are segmented and within each segment the peak is found. Then the gradient of the fundamental frequency is computed, the location of its most negative value is found, and the peak of F0 closest to this extremum is identified.

In order to make the algorithm less sensitive to errors, the second most negative peak is also found within each segment and the corresponding closest peak of F0 identified. This second value is taken into consideration only if the difference between the F0 values of the first and the second peaks does not exceed a fixed threshold.

4.2. AMPLITUDE

For the amplitude data a segmentation procedure similar to that developed for frequency is used. Within each segment the first and the second highest values of the amplitude are identified, and the peaks of F0 closest to these peaks of amplitude are found.

The second amplitude peak is taken into consideration only if the difference between the values of the amplitude in the two positions does not exceed a fixed threshold.

5. COMBINATION OF THE PARAMETER

Once the five cues are identified within each segment, they have to be combined according to suitable "rules" for focal point detection. In order to define these rules, the co-occurrence of cues and their match with focal points hand-marked by experienced analysts is examined for one data set. The Wombat data set is used as a "training" set, i.e. the positions of the parameters are compared to the "correct" position of the foci, where "correct" positions are focal points hand-marked by analysts.

Obviously hand-marking is subjective and it is not possible to have a set of universally "correct" focal points. Here we employ hand-marking done by English speakers who are linguists experienced in identifying focal points in speech.

The rules are chosen to be simple and intuitively reasonable. The intent is not to cover all points in the training set.

The following criteria are used to combine the information of the frequency and the amplitude, and they are used in finding one or two focal points for each segment.

A peak of the fundamental frequency is considered to be a focus of the speech if:

- it is the highest peak within a segment and its value is significantly greater than the values of other peaks;
- it coincides with the position of the first peak of the frequency gradient and of the first peak of the amplitude trace
- it coincides with the position of the second peak of the frequency gradient and of the second peak of the amplitude trace
- it coincides with the position of the first peak of the frequency gradient and of the second peak of the amplitude
- it coincides with the position of the second peak of the frequency gradient and of the first peak of the amplitude

Figure 3 shows an example of the detection of two focal points.

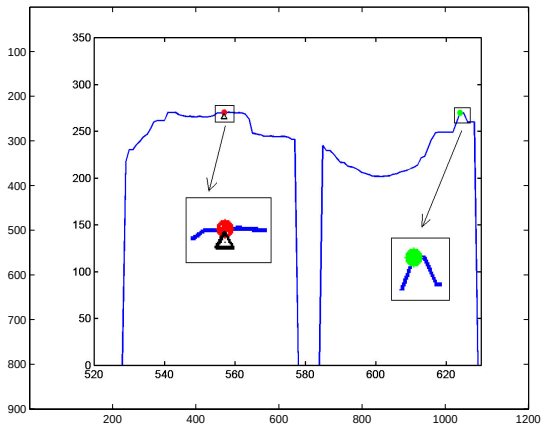

Figure 3. Example of focal detection analyzing cues position

In the picture the symbols have the following meaning:

- triangle: F0 peak
- stars: first and second F0 gradient
- circles: first and second amplitude peak

6. RESULTS AND COMPARISON WITH UNIVERSITY OF BONN METHOD

Once all the focal points are identified in the F0 data, their validity has to be checked with focal points hand-marked by experienced linguists.

Table 1 shows the results for both the Wombat and Sue data. The last columns summarize the performance of the algorithm, summing up the results of the two sets of data.

The algorithm provides an automated way to detect the position of foci with an accuracy greater than 88% and thus

	WOMBAT		SUE		TOTAL	
	#	%	#	%	#	%
TOTAL	63		27		90	
CORRECT	59	93.65	21	77.77	80	88.89
INSERTIONS	8	12.69	4	14.81	12	13.33
DELETIONS	4	6.35	6	22.22	10	11.11

Table 1. Results and percentages of success

it can provide an immense saving in time with respect to the hand-marking procedure.

The idea of considering the minima of the F0 gradient as cues to focal points was suggested by an analysis carried out at the University of Bonn on automatic speech recognition systems [3, 4]. In the investigation of speech data that consists of German spontaneous speech from several speakers it was noticed that in a pre-focal position there is no down-stepping, but after a focal accent down-stepping is significant and characteristic. The physiological reason for this phenomenon is that the physical effort in producing an utterance seems to be unequally distributed. The effort remains high until the focus is reached, while after the focus the effort sinks to a significantly lower level.

The algorithm developed by researchers at the University of Bonn for identifying the points of steepest gradient is slightly different form the one presented in the previous sections, but leads to similar results in identifying F0 gradient cues. The method proposed by the Bonn group consists of first to identifying significant local maxima and minima, and then computing the average values between the maximum and the minimum, and connecting these points to get the global reference line. Using the new frequency data line, the gradient analysis described earlier can be applied. The gradient of F0 is computed and within each segment its minima are identified in view of the fact that the gradient must be in the area of the steepest fall in the F0 trace. Again, the position of the focus is the nearest local F0 maximum in the vicinity of the steepest fall.

The only difference in the method proposed by the University of Bonn researchers and the one proposed in the paper is that the University of Bonn procedure relies on the computation of the reference line that substitutes the frequency data in the analysis. Despite this difference, the results produced are identical for our data sets, due to the fact that the frequency data were already cleaned and preprocessed and the additional processing is therefore not necessary.

Even if the analysis of the F0 gradient information leads to the same results using the two methods, the proposed algorithm has a recognition rate significantly higher than that of the method proposed by the University of Bonn, for which mean recognition rate is equal to 66.6%, versus 88.89% for the proposed algorithm. This difference in performance is due to the fact that the University of Bonn method takes into consideration only the information of the F0 gradient, while the proposed algorithm also considers factors such as F0 peaks and amplitude information and is found to yield a better success rate. As an example, we applied the University of Bonn algorithm to the Wombat data and the results are summarized in table 2.

One observes that the recognition rate is around 65%,

WOMBAT DATA		
	#	%
TOTAL	63	
CORRECT	41	65.07 %
INSERTIONS	18	28.57 %
DELETIONS	22	34.92 %

Table 2. Results applying the University of Bonn algorithm

which is similar to the result obtained by the University of Bonn group.

7. CONCLUSION

In this paper an algorithm is proposed for detecting focal accents in prosody through the identification and extraction of F0 and amplitude cues.

The results are compared to focal points hand-marked by experienced analysts. The percentage of correct detection for the training set, i.e. the set of data used to identify the "rules" of the algorithm, is around 93%, while in the case of the test set, i.e. the set used to check the validity of the algorithm, the success rate is 77%, with an average success rate of 88.89%. This average success rate as well as the success rate in the test case is an improvement over the performance of the algorithm proposed by researchers at the University of Bonn in which only the F0 gradient is used as cue and where the percentage of correct detection is 65%. The improvement is attributed to the introduction of additional F0 and amplitude cues and to the adaptive segmentation method.

In a companion paper, these results will be used to investigate the correlation between speech and gestures and to determine the nature of their temporal relationship.

REFERENCES

[1] D. McNeill, *Hand and mind: what gestures reveal about thought*, University of Chicago Press, 1992

[2] F. Quek, R. Bryll, D. McNeill, C. Kirbas, H. Arslan, K.E. McCullough, N. Furuyama and R. Ansari *Gesture, Speech, and Gaze Cues for Discourse Segmentation* Proc. IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2000, Hilton Head Island, South Carolina, Vol. 2, pp. 247-254, June 13-15, 2000.

[3] A. Elsner, *Prediction and Perception of Focal Accents*, Proc. International Congress of Phonetic Sciences ICPhS'99, pp. 1549-1552, San Francisco, August 14-18 1999

[4] A. Petzold, *Strategies for focal accent detection in spontaneous speech*, Proc. International Congress of Phonetic Sciences ICPhS 95, pp. 672-675, Stockholm, Sweden, August 13-19 1995

[5] R. Ansari, Y. Day, J. Lou, D. McNeill, F. Quek, *Representation of prosodic structure in speech using nonlinear methods*, Proceedings Universiof IEEE-EURASIP Workshop on Nonlinear Signal and Image Processing (NSIP'99), Antalya, Turkey, 20-23 June 1999