Proceedings of DiSS'05, Disfluency in Spontaneous Speech Workshop.
10–12 September 2005, Aix-en-Provence, France, pp. 17-20.

Extracting the acoustic features of interruption points using non-lexical prosodic analysis

Matthew P. Aylett

ICSI, UC Berkeley, USA and
CSTR, University of Edinburgh, UK

Abstract

Non-lexical prosodic analysis is our term for the process of extracting prosodic structure from a speech waveform without reference to the lexical contents of the speech. It has been shown that human subjects are able to perceive prosodic structure within speech without lexical cues. There is some evidence that this extends to the perception of disfluency, for example, the detection interruption points (IPs) in low pass filtered speech samples. In this paper, we apply non-lexical prosodic analysis to a corpus of data collected for a speaker in a multi-person meeting environment. We show how non-lexical prosodic analysis can help structure corpus data of this kind, and reinforce previous findings that non-lexical acoustic cues can help detect IPs. These cues can be described by changes in amplitude and f0 after the IP and they can be related to the acoustic characteristics of hyper-articulated speech.

1. Introduction

Human subjects respond to prosodic structure without necessarily understanding the lexical items which make up the utterance. For example event-related brain potential (ERP) studies have shown a reliable correlation with phrase boundaries when utterances are made lexical nonsensical, either by humming the words, or by replacing them with nonsense words [9]. The use of prosodically rich pseudo speech for artistic purposes (such as R2D2 in star wars, and The Teletubbies amongst others) reinforce these findings. This effect, of apparently understanding prosodic structure without lexical cues, extends to the human perception of disfluency. Lickley [7] showed that human subjects could recognise interruption points, the boundary between disfluent and fluent speech, in low pass filtered speech where no lexical cues were present.

Non-lexical prosodic analysis (NLPA) attempts to mimic this human ability of non-lexical prosodic recognition. Initially, interest in NLPA was motivated largely by the objective of improving automatic speech recognition (ASR) technology, for example, by pre-processing the speech to find syllables [5] or prosodic prominence [3]. However, improvements in statistical modelling in ASR meant that, often, the speech recogniser itself was best left to model prosodic effects internally. Recently, there has been a renewed interest in NLPA techniques in order to address the problem of recognising, segmenting, and characterising very large spontaneous speech databases. Tamburini and Caini [10] point out that identifying prosodic phenomena is useful, not only for ASR and speech synthesis modelling, but also for disambiguating natural language and for the construction of large annotated resources. In these cases, the ability to recognise prosodic structure without lexical cues has two main advantages:

1. It does not require the resource intensive, and language dependent, engineering required for full speech recognition systems.
2. It can offer a means of modelling the human recognition of prosodic structure which in turn could lead to an improved understanding of human speech perception and production.

The ability of human subjects to recognise interruption points (IPs) without lexical information raises the question of whether NLPA can do as good a job. Although previous work has looked at this problem in some depth (e.g. [4], [7]), NLPA offers the prospect of a structured analysis that could be carried out automatically over very large speech databases. In addition, the presence of previous detailed studies allows us to validate the overall approach.

The non-lexical detection of IPs is also of interest from the perspective of determining dialogue structure. Recent work suggests that disfluency patterns could be used to signal the speakers' cognitive load [1] and thus might be used to determine areas in dialogue involving complex concepts, ideas or planning.

We will first describe in more detail the corpus of speech we analysed and the IP phenomena. Next, we will present the details of the NLPA we applied to this corpus followed by results for a set of acoustic features which may cue the non-lexical perception of IPs. Finally, we will discuss limitations with the approach and possible future work.

2. Corpus and disfluency coding

Our data was selected from the ICSI meeting corpus [6]. This consists of 75 dialogues collected from the regular weekly meetings of various ICSI research teams. Meetings in general run for under an hour and have on average 6.5 participants each recorded on a separate acoustic channel. The speech is segmented into spurts, defined as periods of speech which have no pauses greater than 0.5 seconds.

The data we present here is taken from 4 speakers taken from 10 dialogues. Disfluencies are coded as part of the dialogue act coding [2], where interruption points are shown as a hyphen in the speech transcription. In order to avoid complexity caused by multi-speaker interaction and multiple disfluencies, we looked only at phrase boundaries and IPs where:

- The same speaker continued speaking after the interruption point or phrase break
- No other speakers were speaking within 0.5 seconds of the break
- There was at least 0.5 seconds between any breaks.

Pause duration is the clearest acoustic cue of a prosodic break and can be used to disambiguate between IPs and phrase boundaries with some success. In general, the longer the pause, the more likely the break is a phrase boundary. However there are plenty of examples of phrase boundaries followed by a short pause. An interesting question is whether we can disambiguate between these phrase boundaries and IPs followed by a similar short pause. In order to concentrate on this problem, we limited the analysis to IPs and phrase

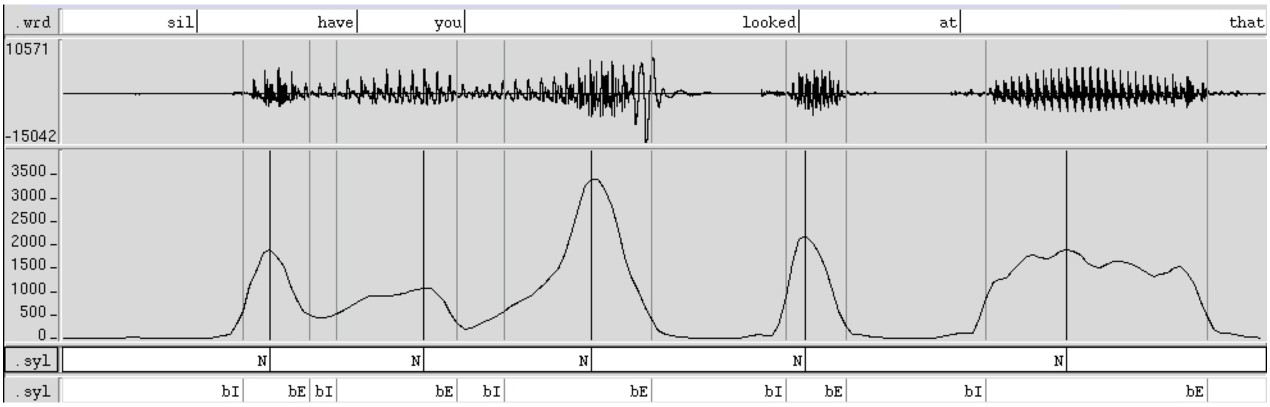

Figure 1: Automatic syllable detection. The lexical contents are shown at the top followed by the waveform, band pass (300-900Hz) energy of the speech and at the bottom labels assigned for syllable nuclei (N) and their initial boundary (bI) and end boundary (bE).

boundaries where the automatic aligner did not insert a following pause.

In this study disfluencies are categorised as repetitions:

"right to my... my right"

substitutions:

"I don't suppose you've got the balloons... the baboons?"

insertions:

"parallel with the ravine... the word ravine"

and deletions:

"oh no what... the line stops at the flagship"

The three dots in the above examples mark the interruption point (IP) (which may or may not be followed by a pause).

3. Non-lexical prosodic analysis

Any acoustic feature can be used to characterise prosody without lexical input. However, a good starting point is features which are reasonably ubiquitous, cross linguistic and have been shown to be sufficient for much human interpretation of prosodic structure. On this basis, amplitude and fundamental frequency are clear starting points. The syllable is a typical means of structuring this acoustic information. Within prosodic theory prominence is associated with syllables, in particular syllable nuclei. Therefore, a first step in any NLPA is syllable extraction. Howitt [5] reviews many of the current algorithms for segmenting speech into syllables. If we evaluate these algorithms in terms of how well they predict the syllable boundaries compared to those produced by human segmentation (or even by autosegmentation), they typically perform rather poorly. However, for NLPA we are not attempting to segment speech, our intention is rather to characterise the prosodic structure. Given that much of the perceived amplitude and pitch change occurs across the syllable nucleus, finding the extent of the nuclei is more important than determining the syllable boundaries. In fact, most simple syllable detection algorithms will find 80% of the syllable nuclei and the syllables they typically miss are unstressed, short syllables, which tend to carry much less prosodic information. In addition, Tamburini and Caini [10] found that the duration of nuclei correlates closely to the overall syllable duration and therefore the syllable nuclei duration can be used to measure the rate of speech as well as assessing prominence.

On this basis, we extracted syllable nuclei as suggested by Howitt [5]. This involves band pass filtering speech between 300-900 Hz and then using peak picking algorithms to determine the location and extent of nuclei. For these experiments we used a simpler peak picking algorithm than the modified convex-hull algorithm [8] described by Howitt [5] and used by Tamburini and Caini [10].

Figure 1 shows an example of the results of the syllable extraction algorithm we applied. The top shows the lexical contents of the speech, followed by a waveform. Below the waveform is the energy of the band pass filtered speech. The labels below the band pass filtered speech show the syllable nuclei (black line) and the extent of the nuclei (grey lines). The process for determining these nuclei is as follows:

1. Remove large portions of silence from the data and divide the speech into spurts - continuous speech with less than 0.5 seconds gap. Allow 0.1 seconds of silence before and after each spurt.
2. Band pass filter the speech between 300-900 Hz.
3. Examine the distribution of the energy for the speaker across the data and set a threshold for syllable energy at the 65th percentile.
4. Find the maximum points in the region. A maximum point has a previous and subsequent lower value with a number of equal values in between. Order the points by amplitude and go through the list picking syllable nuclei providing a previous nuclei has not already been picked within a range of 0.1 seconds.
5. Set the boundaries as equidistant between nuclei in the same voiced region otherwise to the threshold edge of the region.
6. Extract f0 values, using the entropics get_f0 program, for the start, centre and end of the syllable nuclei.
7. Smooth the resulting f0 contour and interpolate values across unvoiced regions.

We can assess the prominence of each syllable either based on amplitude and duration (sometimes described as stress prominence [10]), or the f0 variation over the syllable nucleus (sometimes described as accent prominence [10]). Phrase boundaries are assessed both on the basis of pauses, determined by a simple threshold silence detector, and boundary f0, taken as the f0 at the edge(s) of the surrounding syllable nuclei.

4. Analysing IP boundaries with NLPA

Previous work, Lickley [7] and Hirschberg et al [4], has shown a number of interesting acoustic features which can be associated with IPs. All the features occur after the IP with no discernable acoustic cues before the IP. Both [7][4] found a tendency for increased amplitude after the IP, higher f0 and longer duration. These are all correlates of stressed syllables and also of hyper-articulated speech. Hirschberg et al [4] describe these acoustic features as cues for *corrected speech* and describe a machine learning approach for classifying corrected speech on the basis of these features. This was then applied to reduce recognition error from 25% down to 16%. However, the extent to which these utterances contained IPs was not reported. In Lickley [7], human judgments of low pass filtered speech utterances show a significant, although far from consistent, effect across materials. Human subjects tended to

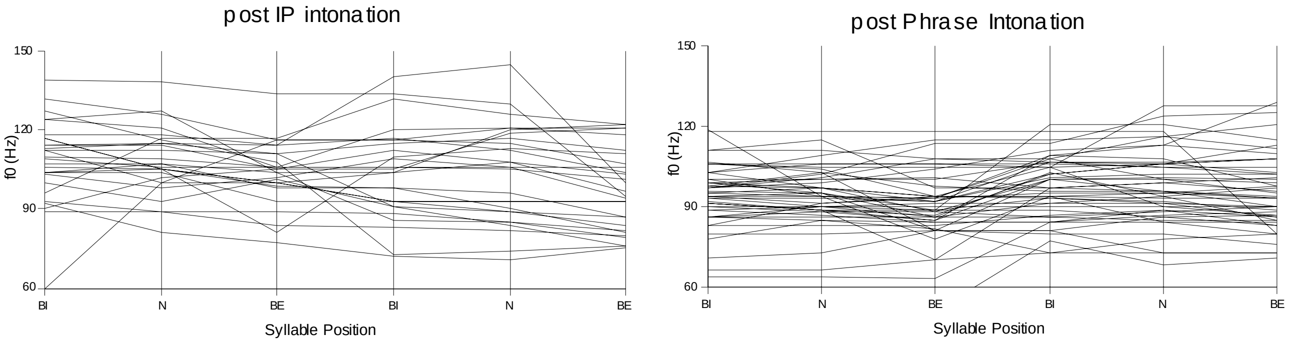

Figure 2: F0 across the two syllable nuclei following both IPs and phrase breaks where no or minimal pause cues are present. (BI - initial boundary of syllable nucleus, N - centre of syllable nucleus, BE - end boundary of syllable nucleus).

misclassify disfluent utterances as fluent utterances more then visa-versa with the best group of human subjects correctly classifying 34% of disfluent utterances as disfluent. In addition, the significant effect in this study appeared to be dominated by the presence and differences of pause durations rather than other acoustic cues.

In this data, as stated earlier, only boundaries with pauses not discernable to the aligner where examined. We compared the results for IPs and normal phrase breaks. As in [7] [4], we looked at acoustic cues in the form of f0 variation, syllabic nucleus amplitude and syllabic nucleus duration after the boundary point.

5. Results

We began by looking at the f0 change across the two syllables to the right of the boundaries. Shown in figure 2 are six f0 points. These values are taken from the first two syllable nuclei found with NLPA subsequent to the phrase or IP boundary for a single male speaker. It is interesting to note a lack of a homogeneous f0 structure in either IP or for PH (Phrase conditions). However, differences are clearly present between both groups. F0 in the IP case tends to be higher and varies more throughout the two syllables.

On the basis of this plot we chose three f0 features to examine statistically: the f0 before the boundary, the f0 following the boundary and the variance of the f0 across the two syllables following the boundary. F0 values were normalised with on the basis of the mean and standard deviation of each speakers voiced f0 values. In addition, we combined the log of the raw amplitude of the first following syllable with the log of the duration of its nucleus by multiplying the factors together to give an overall prominence factor. Thus short, high energy syllable nuclei where regarded as having similar prominence to long, lower energy syllable nuclei.

An independent t-test grouped by IP and phrase boundary (PH) is shown in Table 1. All factors except post boundary f0 variance are significant with an appropriate Bonferroni correction. If we examine the cell means in Figures 3 and 4 the results are in line with previous published results. We see higher initial f0 values for after IPs and more prominence caused by amplitude and duration.

Table 1: Independent t-test for acoustic cues following IPs and Phrase Boundaries.

Acoustic Feature	t	df	Sig. (2-tailed)	Bonferroni correction
f0 pre boundary	*4.466*	*1,173*	*0.000*	*0.000*
f0 post boundary	*7.091*	*1,173*	*0.000*	*0.000*
f0 variance post boundary	*-0.403*	*1,198*	*NS*	*NS*
prominence post boundary	*3.181*	*1,154*	*0.002*	*0.008*

If we use these factors in a discriminant analysis, we find we can categorise 58.7% of the data (58% with cross validation), see Table 2. Even given the absence of pause data these results are poor and suggest significant by-speaker variation. If we analyze the speakers individually we see a varying but similar pattern for pre/post f0 and prominence as reflected in Figure 3 and 4. However the pattern of post boundary f0 variance is very different across speakers. Figure 5 shows the normalized post boundary f0 variance for all four speakers. If we look at subject *me011* (the subject whose f0 is plotted in Figure 2) we see that f0 varies more after the IP than the phrase break ($p < 0.05$ - NS with Bonferroni

Table 2: Results of discrimant analysis using acoustic cues

Discriminant Analysis		**Classification**	
		PH	IP
Original	PH	*355*	*301*
	IP	*203*	*341*

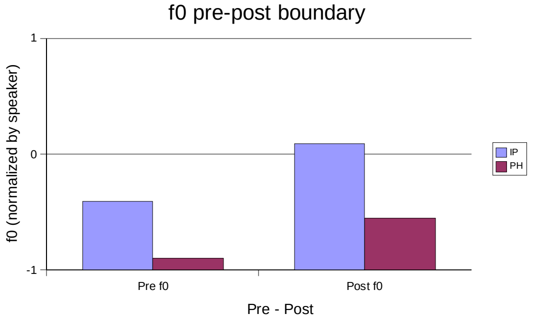

Figure 3: F0 across IP boundary and phrase boundary (PH).

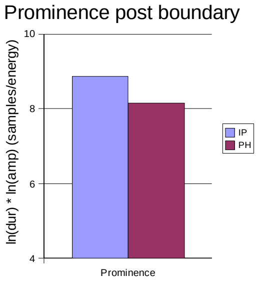

Figure 4: Prominence -nucleus ln(amplitude)ln(duration)) - after IP and phrase boundary (PH).

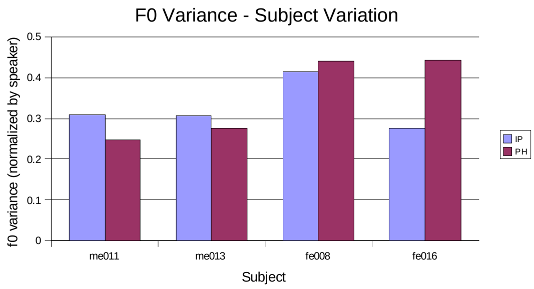

Figure 5: Post boundary f0 variance by speaker

correction). In complete contrast, subject *fe016* shows the reverse ($p < 0.005$ – $p < 0.05$ with Bonferroni correction). The success of discriminant analysis using these factors also varies from 58% to 69% across speakers. The *fe* denotes a female speaker whereas *me* a male speaker. It is possible that some of this variation is associated with the sex of the speaker although a much larger sample of speakers would be required to explore this possibility.

These results are for the IPs and phrase breaks where no pause had been inserted by the aligner. Pause duration is still, by far, the best indicator of phrasing. This raises the question of to what extent pause determined non-lexically might be useful for classifying IPs and phrase breaks.

In order to address this we examined the boundaries and determined pauses based on the band pass energy being below the 25th percentile. Results suggest the aligner has a tendency to absorb short pauses. The mean pause calculated in this way for boundaries with no pause discernable to the aligner was 45 milliseconds. If we add this acoustic pause factor to our discriminant analysis we see a jump from 58% to 68% success of classification.

6. Conclusion

Results show that NLPA can be used for characterising disfluency. Furthermore, that it would seem to perform as well, or better, than human subjects given the same task. Perhaps the most interesting feature of the work is that NLPA offers a non-lexical structure for dealing with timing. Using the syllable nucleus we can implicitly scale f0 contours which might allow a more structured approach to characterizing intonation non-lexically. Although the prominence feature presented in this work is perhaps an over simplification of the perceptual effect of duration and amplitude, it does allow a starting point for an improved system. Similarly it would be an interesting idea to replace the f0 variance with a more perceptually based model of accentedness.

Results for four speakers suggest that a great deal of inter-subject variation makes it hard to produce a general model of these factors. This is further complicated by treating all IPs as the same when previous work (i.e. [7]) has shown that there are differences between the typical acoustic effects of, for example, repetitions as opposed to deletions.

The results for pauses determined acoustically and non-lexically offer a salutary lesson against depending blindly on automatically aligned results for corpora analysis. A simple non lexical acoustic analysis can be taken quite far: it requires less resources, is less language dependent and arguably less prescriptive.

However, the success of NLPA depends largely on the autosyllabification process. Overgeneration of syllables and overestimation of syllable nuclei, for example, caused by liquids or nasals, can present a significant problem in terms of aligning f0 contours with the output. In future work we will evaluate the syllabification algorithm quantitatively against state-of-the-art automatic alignment. Preliminary results suggest the current NLPA matches 65% of syllables from the alignment. However over 50% of the syllables missed are schwa nuclei which reinforces the idea that NLPA might do a better job at finding prosodically *pertinent* syllables than automatic alignment. The system appears to have around a 10% false alarm rate and, as expected, these are very much associated with nasals and roticization.

The IP analysis reinforces findings from previously published work. The results for automatic disambiguation (especially given the lack of pause information) are promising. However, in order to really test how useful these factors are for discrimination, we must also see to what extent they can tell any boundary (syllable/word) from an IP. In addition, as pointed out by Hirschberg et al [4], different speakers have different characteristics in terms of hyper-articulation. On this basis further work requires the analysis of many more subjects.

7. Acknowledgements

This work was partly supported by the European Union 6th FWP IST Integrated Project AMI (Augmented Multi-party Interaction, FP6-506811 publication).

8. References

[1] Bard, E., Lickley, R.J. & Aylett, M.P. 2001. Is Disfluency Just Difficult? *Proceedings of DISS 01, ISCA Tutorial and Research Workshop*. Edinburgh.

[2] Dhillon, R., Bhagat, H., Carvey, H. & Shriberg, E. 2004. Meeting Recorder Project: Dialog Act Labelling Guide. *Technical Report TR-04-002*. ICSI.

[3] Hironymous, J.L., McKelvie, D. & McInnes, F.R. 1992. Use of Acoustic Sentence Level and Lexical Stress in HSMM Speech Recognition. *ICASSP '92 Proceedings.* San Francisco. California, pp225-227.

[4] Hirschberg, J., Litman, D. & Swerts M. 1999. Prosodic Cues to Recognition Errors. *ASRU-99.*

[5] Howitt A.W. 2000. *Automatic Syllable Detection of Vowel Landmarks.* PhD Thesis, MIT.

[6] Janin, A., Baron, D., Edwards, J., Ellis, D., Gelbart, D., Morgan, N., Peskin, B., Pfau, T., Shriberg, E., Stolcke, A. & Wooters, C. 2003. The ICSI Meeting Corpus. *Proceedings ICASSP-03.* Hong Kong.

[7] Lickely, R.J. 1994. *Detecting Disfluency in Spontaneous Speech.* PhD Thesis, University of Edinburgh.

[8] Mermelstein, P. 1975. Automatic segmentation of speech into syllabic units. *JASA.* 58(4) pp880-883.

[9] Pannekamp A., Toepel, U., Alter, K., Hahne, A. & Friederici, A.D. 2005. Prosody-driven Sentence Processing: An Event-related Brain Potential Study. *Journal of Cognitive Neuroscience.* 17(3) pp407-421.

[10] Tamburini, F. & Caini, C. 2005. An Automatic System for Detecting Prosodic Prominence in American English Continuous Speech. *International Journal of Speech Technology.* 8 pp33-44.