DETECTING HIGH LEVEL DIALOG STRUCTURE WITHOUT LEXICAL INFORMATION

Matthew P. Aylett

ICSI UC Berkeley
CSTR Edinburgh

ABSTRACT

The potentially enormous audio resources now available to both organizations, and on the Internet, present a serious challenge to audio browsing technology. In this paper we outline a set of techniques that can be used to determine high level dialog structure without the requirement of resource intensive automatic speech recognition (ASR). Using syllable finding algorithms based on band pass energy together with prosodic feature extraction, we show that a sub-lexical approach to prosodic analysis can out-perform results based on ASR and even those based on a word alignment which requires a complete transcription. We consider how these techniques could be integrated into ASR technology and suggest a framework for extending this type of sub-lexical prosodic analysis.

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 [1]. 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 emotional content, and disfluency. Sub-lexical prosodic analysis (SLPA) attempts to mimic this human ability.

Initially, interest in SLPA was motivated largely by the objective of improving ASR technology, for example; by preprocessing the speech to find syllables or prosodic prominence. However, improvements in statistical modeling in ASR meant that, often, the speech recognizer itself was best left to model prosodic effects internally. Recently, there has been a renewed interest in SLPA techniques in order to address the problem of recognizing, segmenting, and characterizing very large spontaneous speech databases. Tamburini and Caini [2] point out that identifying prosodic phenomena is useful, not only for ASR and speech synthesis modeling, but also for disambiguating natural language and for the construction of large annotated resources. In these cases, the ability to recognize 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 characterizing and segmenting very large audio corpora.

Prosodic structure can offer extensive priors on language identity, cross talk, speaker identity, high level dialog structure (see [3] for a review), and speaker emotion [4]. In this paper we review some current work in SLPA, describe a version of SLPA in detail and present the results of applying the technique to three high level dialog structure categorization problems. We conclude by discussing the strengths and weaknesses of our version of this technique and how such an approach might be extended.

1.1. Extracting prosodic structure

The idea of extracting prosodic structure before recognition has been explored in some detail over the last 15 years. Most techniques involve an amplitude based peak picking algorithm that is used for determining syllable location. Band pass filtering and smoothing is used to improve results (see [5] for a review).

Recent work by [2, 6, 7] has applied the technique to the more abstract problems of determining fluency, disambiguation, and language identity. In this paper we evaluate SLPA directly in terms of its utility for categorizing and segmenting 3 examples of high level dialog structure:

Dialog Acts (DAs): DAs are used to determine dialog function. We focus only on the most common monosyllabic dialog acts "yeah" and "right" [3].

Involvement: Previous work [4] has shown that the emotional involvement of speakers, sometimes termed *hot spots* can be reliable coded by human subjects. Determining such involvement automatically is potentially useful for finding "interesting" areas of long dialogs.

Prosodic Boundary: Segmenting dialog acts and determining fluency is fundamental for determining dialog structure [3].

This work was supported by the European Union 6th FWP IST Integrated Project AMI.

We evaluate the performance of SLPA on these classification tasks by comparing it to results gained using syllabification based on both a complete word alignment, and on the output of a state of the art recognizer.

1.2. Corpus and dialog coding

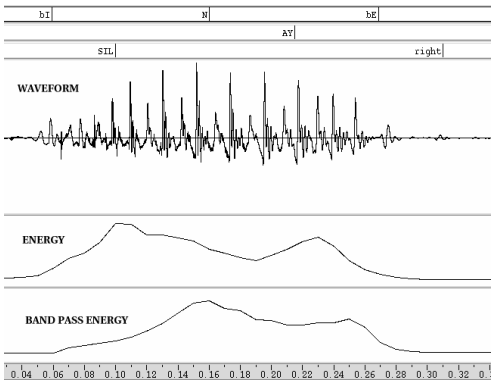
Our data was selected from the ICSI meeting corpus [8]. This consists of 75 dialogues collected from the regular weekly meetings of various ICSI research teams. Meetings in general run for under an hour and have on average 6.5 participants each recorded on a separate acoustic channel. Dialog act coding was carried out as per [9], although in this analysis of "yeah" and "right" we only considered 3 categories of dialog act: statement, back channel and question. The resulting data set contained 3044 data points for "right" and 8355 for "yeah".

Hot spot or involvement coding looks at the perceived involvement of participants in the discussion. *Involvement* is categorized as *amusement*, *disagreement*, and *other (interest, surprise or excitement)* [4]. Inter-rater agreement was relatively high (*Kappa of K=0.59(p<0.01)*). Hot spots occurred relatively infrequently and comprised of 2.6% of dialog acts (approximately 4000 instances). For each hot spot we took a random non hot dialog act with closely matching prosodic structure (number of syllables/phrasing) to act as a control.

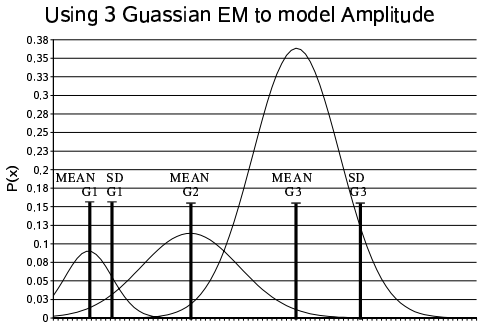
Prosodic boundaries were determined for each syllable and could be either inter-word, word boundary, dialog act or interruption. In general, dialog act boundaries coincide with prosodic phrase boundaries although not in all cases. Resource limitations prevented us from carrying out a hand analysis of prosodic phrasing. Disruptions were decided on the basis of the corpus transcription. If the speech was regarded as not fluently completed by a transcriber a hyphen was used to show a disruption point. A full word alignment was carried out using the ICSI speech recognizer [10] and used to align dialog acts and word boundaries with the speech. These time points were then used as ground truth for both the SLPA analysis an the analysis based on ASR. We took a balanced set of data points for boundaries, by randomly selecting 10% of word boundaries and 25% of inter-word boundaries. The boundary data set had approximately 160,000 data points.

2. SUB-LEXICAL PROSODIC ANALYSIS

The syllable is a typical means of structuring prosodic information. Within prosodic theory, prominence is associated with syllables, in particular syllable nuclei. Therefore, a first step in any SLPA is syllable extraction. 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 auto segmentation), they typically perform rather poorly. However, for SLPA we are not attempting to segment speech; our intention is rather to characterize the prosodic



a.



b.

Fig. 1. a) Example of SLPA extraction of the word "right". b) A 3 Gaussian 1 dimension mixture model used to model log energy distribution.

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 [2] 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 used by Howitt [5] and by Tamburini and Caini [2].

Figure 1a shows an example of the results of the syllable extraction algorithm we applied to an instance of the word "right". The top line shows the start, mid point and end point of the syllable as determined by SLPA, based on taking a threshold for silence over the energy and finding peaks in the band pass energy (both shown below the waveform). In the second line, 'AY' shows the syllable mid point as determined

by word alignment. The third line shows the start and end point of the word (and syllable in this case) as determined by the aligner. Note there are differences of up to 40ms between the two segmentations.

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. We used a 199 tap FIR filter designed using Matlab fir1 function. Use TkSnack to compute the energy over 10ms frames. Smooth the result with a low pass 50Hz filter.

3. Use TkSnack to compute the full band energy over 10ms frames and smooth the output with a 50Hz low pass filter. The log distribution of this data is then used to create a channel dependent model for both normalization and for determining thresholds. We use expectation maximization to fit a 1 dimension 3 Gaussian mixture model to the data (See Figure 1b). The Gaussians were initialized with means spread equally across the data. We used the working assumption that the Gaussian with the lowest mean would describe the non-speech silent areas while the Gaussian with the highest mean, the voiced regions. We used the mean of the middle Gaussian as a threshold for these quasi voice/non-voiced regions. A threshold for silence was set to the lowest Gaussian mean plus 4 times the standard deviation. This threshold was applied to the energy data to determine location and extent of pauses. We term these pauses *acoustic pauses* to avoid confusion with paused determined using the word alignment or the recognition output.

4. Find the maximum points in the quasi voiced regions. A maximum point was defined as having a greater value than the two points 40ms previous and subsequent. We order the maxima by amplitude and go through the list from the highest maxima downwards. We pick a maxima as a 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 when no acoustic pause is present in between, else to the edge of subsequent or previous acoustic pause.

6. Calculate f0 values, using the Entropics get_f0 program. The f0 output is then smoothed and undergoes linear fitting to produce an abstract contour consisting of falls and rises [11].

2.1. RESULTS

In order to evaluate the sub-lexical approach we compare models built using feature extraction based on syllable regions determined using SLPA against syllable regions determined using a word alignment or a recognition output. We used the ICSI recognizer which achieves a word error rate in the 20s, not unusual for this type of spontaneous material [10]. In order to determine syllables from ASR output we used maximal onset to determine syllable boundaries. Unrecognized data, or data which could not be aligned were removed.

We extracted the following features for each syllable:

Energy at the mid point of the syllable normalized using the mean and standard deviation of the G3 Gaussian (see Figure 1b);

F0 at the edges of the voiced region and F0 at a mid point in the syllable. We normalized F0 using the 5th and 95th percentile of the speakers first 10000 F0 values;

Log duration of the syllable.

Over regions (DAs and Hot spots) we computed the minimum, maximum, mean and variance for all three features. In addition, for prosodic boundaries classification, we included the length of the acoustic pause when present.

2.2. Decision tree and disciminant analysis results

A decision tree model was used for DA and Hot spot categorization. Trees were built using Weka C4.5 implementation with bagging, trained on a random 75% of the data. Results are shown from applying the trees to a held out test set of 25% of the data. Due to the larger size of the boundary data set we used SPSS to carry out linear discriminant analysis. A model was built on all data and results are shown for cross-validation on a leave one out basis.

"Yeah" and "right" account for 55% of single word dialog acts in the ICSI corpus. They can be either statements (such as acknowledgment), back channels (in order to encourage another speaker to continue), questions or floor grabbers (an attempt to enter the discussion). Due to the small number of floor grabber we only considered the first three categories. "Right" is spread evenly across all three whereas "yeah" was not normally used as a question. Figure 2a shows the percentage correctly categorized by SLPA, prosodic analysis based on a word alignment and based on recognition. For both SLPA out performs the analysis based on the recognition. Although the word alignment does do better for "right".

Figure 2b shows the results for determining whether a dialog act is a hot spot. Chance result would be 50%. SLPA does well on this material out-performing prosodic analysis based on alignment and recognition.

Figure 2c shows the results for categorizing a 4 way boundary condition. SLPA does not do quite as well as either alignment or recognition.

Results show that SLPA can perform as well as prosodic analysis from recognizer or word alignment output. In fact in several cases it can outperform recognition based output. There are a number of possible explanations for this. Firstly recognizers are not designed to extract prosody. They are ruthlessly tuned to improve word error rate. In some cases

this can harm the prosodic data extraction. For example, the acoustic models tend to gobble up short pauses.

It is interesting to note SLPA does better at categorizing regions compared to boundary types. This may be because SLPA was tuned to find mostly stressed strong syllables and miss weak ones rather than generating false alarms (50% of syllables missed by SLPA were schwa syllables). For determining hyper articulation, or prosodic contour, this may be a good thing. However for boundaries where weak syllables may play a important role in distinguishing phrase ending, this could be a problem. In future work we intend to experiment more closely with the threshold levels to see how it affects the performance on detecting different dialog structure.

3. CONCLUSION

It would be possible to amalgamate this approach with current recognition approaches. For example using the recognition of high level dialog as a secondary task for the recognizer. The use of other acoustic features such as harmonicity and spectral tilt could also improve the results. Another possibility is to augment the SLPA output with a bespoke "filled pause" recognizer. Such a light-weight recognizer could add significant power for categorizing floor-grabbing and other specific dialog events. Finally it is possible that the structured output from SLPA could aid in the detection of cross-talk and for speaker identification, both tasks which are often required before a standard recognition process can be carried out.

4. REFERENCES

[1] A. Pannekamp et al, "An automatic system for detecting prosodic prominence in american english continuous speech," *Journal of Cognitive Neuroscience*, vol. 17, pp. 407–21, 2005.

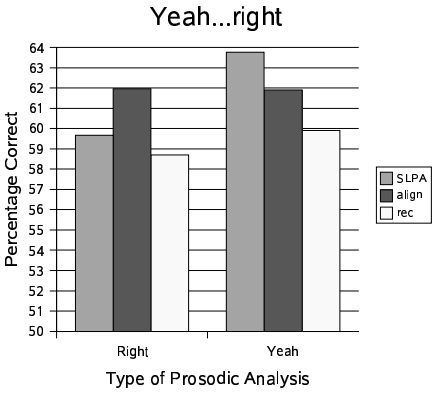
[2] F. Tamburini and C. Caini, "An automatic system for detecting prosodic prominence in american english continuous speech," *International Journal of Speech Technology*, vol. 8, pp. 33–44, 2005.

[3] E. Shriberg and A. Stolcke, "Prosody modeling for automatic speech recognition and understanding," in *Mathematical Foundations of Speech and Language Processing*, M. Johnson, S. Khudanpur, M. Ostendorf, and R. Rosenfeld, Eds., pp. 229–254. Springer-Verlag, 2004.

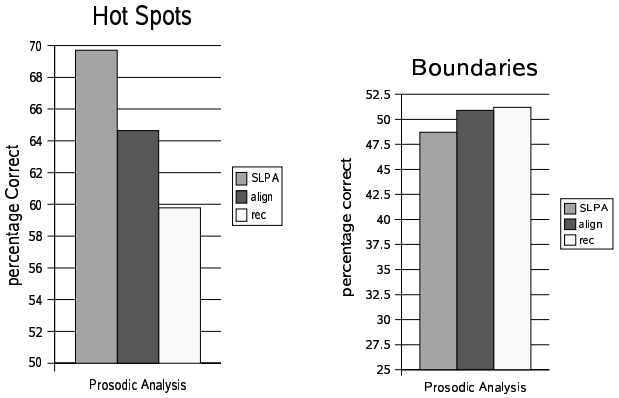
[4] B. Wrede and E. Shriberg, "Spotting 'hot spots' in meetings: Human judgments and prosodic cues," in *Eurospeech*, 2003, pp. 2805–8.

[5] A.W. Howitt, *Automatic Syllable Detection of Vowel Landmarks*, Ph.D. thesis, MIT, 2000.

[6] M.P. Aylett, "Extracting the acoustic features of interruption points using non-lexical prosodic analysis," in *DISS'05: ISCA Workshop*, 2005.



a.



b. c.

Fig. 2. Categorization results compared across SLPA, word alignment (align) and ASR recognition (rec). Results using decision trees, a) Dialog act identity for "yeah" and "right", b) Hot spots. Results using discriminant analysis, c) between syllable boundary types.

[7] J-L. Rouas, "Modeling long and short-term prosody for language identification," in *Interspeech*, 2005, pp. 2257–60.

[8] A. Janin et al, "The ICSI meeting corpus," in *ICASSP*, 2003.

[9] R. Dhillon, H. Bhagat, H. Carvey, and E. Shriberg, "Meeting recorder project: Dialog act labelling guide," in *Technical Report TR-04-002*, 2001.

[10] A. Stolcke et al, "Progress in meeting recognition: The ICSI-SRI-UW spring 2004 evaluation system," in *NIST ICASSP Meeting Recognition Workshop*, 2004.

[11] K. Sönmez, E. Shriberg, L. Heck, and M. Weintraub, "Modeling dynamic prosodic variation for speaker verification," in *ICSLP*, 1998.