# Spotting "Hot Spots" in Meetings:
# Human Judgments and Prosodic Cues

*Britta Wrede[1,2], Elizabeth Shriberg[1,3]*

[1]International Computer Science Institute, Berkeley, USA
[2]Applied Computer Science Group, Technical Faculty, Bielefeld University, Germany
[3]Speech Technology and Research Laboratory, SRI International, Menlo Park, USA

`bwrede@techfak.uni-bielefeld.de ees@speech.sri.com`

## Abstract

Recent interest in the automatic processing of meetings is motivated by a desire to summarize, browse, and retrieve important information from lengthy archives of spoken data. One of the most useful capabilities such a technology could provide is a way for users to locate "hot spots" or regions in which participants are highly involved in the discussion (e.g. heated arguments, points of excitement, etc.). We ask two questions about hot spots in meetings in the ICSI Meeting Recorder corpus. First, we ask whether involvement can be judged reliably by human listeners. Results show that despite the subjective nature of the task, raters show significant agreement in distinguishing involved from non-involved utterances. Second, we ask whether there is a relationship between human judgments of involvement and automatically extracted prosodic features of the associated regions. Results show that there are significant differences in both F0 and energy between involved and non-involved utterances. These findings suggest that humans do agree to some extent on the judgment of hot spots, and that acoustic-only cues could be used for automatic detection of hot spots in natural meetings.

## 1. Introduction

Recent interest in the automatic processing of meetings is motivated by a desire to summarize, browse, and retrieve important information from lengthy archives of spoken data. One of the most useful capabilities such a technology could provide is a way for users to locate "hot spots", or regions in which participants are highly involved in the discussion (e.g., heated arguments, points of excitement, and so on). Such regions are likely to contain important information for users who are browsing a meeting of for applications of information retrieval.

The precise definition of a hot spot is admittedly and necessarily open-ended, because hot spots can have very different characteristics—from excited brainstorming to tense disagreement. Nevertheless, a common characteristic across hot spots is that they display high "involvement". As will be shown in the experiments involvement seems to be closely related to "activation" which is one of two basic dimensions that are useful to describe emtions [1] [2]. In this work we define hot spots as regions in a meeting in which there is high involvement on the part of two or more participants.

We ask two questions about hot spots in meetings in the ICSI corpus of naturally-occurring meetings:

- Can human listeners agree on utterance-level judgments of speaker involvement?

- Do judgments of involvement correlate with automatically extractable prosodic cues?

The questions of human judgments and acoustic correlates of involvement are addressed in two separate studies. First, a rating experiment was set up where subjects were asked to rate utterances with respect to involvement (Section 2). Based on the results of this experiment (Section 3) acoustic features based on F0 and energy were analysed with respect to their correlation with perceived involvement (Section 4).

## 2. Method

The data for the experiment was derived from a subset of meetings of the ICSI Meeting Recorder Corpus [3]. This corpus consists of recordings of naturally-occurring meetings on scientific topics. Speakers were recorded by both far field and individual close-talking microphones; we use the latter for this work.

A subset of 13 meetings was selected and analysed with respect to involvement. The subset consisted of different meetings of the same group of people, with about 4 to 8 speakers from whom 6 (2 female, 4 male) attended regularly and contributed most of the data. In a first step utterances from the meetings were labeled by one rater with respect to perceived involvement while listening to the whole meeting. Two special cases of involvement were identified in the data as having strongly differing characteristics: *amusement* and *disagreement*. Most of the remaining involved utterances were perceived as conveying interest, surprise or excitement. Therefore, the class of involved utterances was subdivided into *amused*, *disagreeing* and *other*. The fourth category was *not particularly involved*. During this pass, periods of about half a minute to one minute in the meeting where more than one participant had a high level of involvement were labeled as hot spots. It was observed that hot spots, too, can be categorized into types, based on the types of involvement in their component utterances. Thus hot spots were also labeled as either *amused*, *disagreeing*, or *other*.

It has been argued that the perception of emotion is strongly influenced by context [4]. Thus, context information might be crucial for the detection of involvement. However, for automatic detection it would be desirable to rely as much as possible on the acoustic features of an utterance alone. If human listeners are able to detect involvement reliably from utterances in isolation, this strongly indicates the presence of acoustic cues. Therefore, we chose to present isolated utterances. Furthermore, for the judgments of utterances we decided to rely on people who were familiar with the speakers in the meetings. It is reported in the literature (e.g. [5] p.20) that people find it difficult to determine the emotion of a person they do not know.

|       | s1  | s2  | s3  | s4  | s5  | s6  | s7▲ | s8▲ | s9▲ |
|-------|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| s1    |     | .80 | .52 | .83 | .89 | .59 | .44 | .66 | .85 |
| s2    | .72 |     | .60 | .70 | .83 | .48 | .46 | .55 | .80 |
| s3    | .49 | .51 |     | .48 | .46 | .49 | .26 | .46 | .58 |
| s4    | .81 | .62 | .35 |     | .89 | .68 | .28 | .54 | .83 |
| s5    | .79 | .70 | .41 | .78 |     | .60 | .47 | .65 | .86 |
| s6    | .48 | .40 | .45 | .53 | .48 |     | .15 | .53 | .71 |
| s7▲   | .40 | .41 | .34 | .22 | .32 | .20 |     | .32 | .43 |
| s8▲   | .52 | .50 | .46 | .38 | .60 | .33 | .30 |     | .78 |
| s9▲   | .62 | .58 | .41 | .58 | .62 | .51 | .17 | .55 |     |

Table 1: Pairwise $\kappa$ for inter-rater agreement for two categories in the upper right and for four categories in the lower left (in italics). Black triangles indicate nonnative raters.

| Group     | # Raters | # Ratings | $\kappa$ 2 cat. | 4 cat. |
|-----------|----------|-----------|-----------------|--------|
| All       | 9        | 8         | .59             | .48    |
| Nonnative | 3        | 3         | .52             | .33    |
| Native    | 6        | 5         | .63             | .55    |

Table 2: $\kappa$ for inter-rater agreement for two and four categories for the group of native speakers and the group of nonnative speakers.

This is because they do not know the relevant baseline of what level is "neutral" for that person. An automatic detector of involvement should be able to normalise for a speaker specific baseline.

The experiment was carried out via a web interface. Since it was not clear if the concept of involvement would be easy to understand we decided to mix verbal explanations and examples in the instructions for the raters. Thus, involvement was introduced with respect to hot spots, which were described as "places in conversation where multiple participants get especially involved". In order to give the raters an idea about what was meant by hot spots and involvement, they could listen to one example of each type of hot spot as labeled earlier by one annotator. Thus, the subjects could determine themselves which of the utterances in the hot spot they considered as particularly involved. This gave them an impression of involvement within the context of a hot spot, but left the decision over the involvement baseline for a specific speaker to the rater.

During the experiment the raters could listen to each utterance as often as they wanted. They had then the choice to rate an item as one of three "involved" categories (*disagreeing, amused, other*) or as *Not especially involved*, or as *Don't Know*. The raters were asked to base their judgment as much as possible on the acoustics of the utterances. This was done in order to minimise effects of the propositional content of an utterance on the judgment because we wanted to capture the pragmatic and emotional information that is conveyed by an utterance. For example, people might tend to rate the utterance "I disagree" as *disagreeing* even though it is non involved at all or said in a rather amused way. We wanted to avoid such ratings.

## 3. Inter-rater agreement

In order to assess how consistently listeners perceive involvement, inter-rater agreement was measured by Kappa for both pairwise comparisons of raters and overall agreement. Kappa computes agreement after taking chance agreement into account. In [6], pairwise Kappa is modified to capture agreement among more than two raters. This measure assumes an identical number of ratings for all utterances. Because not all listeners rated all available utterances, this reduced the number of utterances used in the all-way Kappa computation.

Nine listeners, all of whom were familiar with the speakers (as discussed above), provided ratings for at least 45 utterances. Several other raters provided additional ratings (to a maximum of 150 utterances). In the nine-way comparison, 13 of the 45 utterances were rated as *Don't Know* by at least one rater. Therefore, only 8 ratings were considered for each of these 45 utterances, by randomly omitting one rating for utterances with 9 ratings.

Inter-rater agreement for the high-level distinction between involved and non involved yielded a Kappa of $\kappa = .59$ ($p < .01$), a value considered quite reasonable for subjective categorical tasks. When Kappa was computed over all four categories, it was reduced to $\kappa = .48$ ($p < .01$). indicating that (after adjusting chance for the higher number of classes) there is more difficulty in making distinctions among the types of involvement (*amused, disagreeing* and *other*) than in making the high-level judgment of the presence of involvement. This could be due to the heterogeneous class of *other* which covers all the remaining data of involved utterances and gives the rater no canonical idea of what it refers to. More investigations are needed to determine if it is possible to establish further classes that are easier for raters to distinguish.

### 3.1. Pairwise agreement

To obtain more fine-grained results, $\kappa$ was also computed for each pair of raters (it thus also included substantially more rated utterances), as shown in table 1. Both the two-way and four-way results are shown in the Table. As can be seen, agreement values differ depending on the raters. It seems that some raters are simply better than others at the task, since high and low agreement values tend to correlate with raters.

### 3.2. Native vs. nonnative raters

One possible explanation for the agreement differences by raters is that ratings differ for natives and nonnatives, when judging native utterances. All utterances used were spoken by native (or in one case a perceptually-native) speakers of American English. Nonnative raters are marked in Table 1 by a black triangle. In order to determine the effect of nativeness on ratings we computed the Kappas for these two groups separately. Results are shown in Table 2. As shown, Kappas for the nonnative group are lower than those for the native speakers. In particular, agreement with respect to the more detailed 4 categories is noticeably lower in the nonnative group ($\kappa = .33$). Although from this small sample it is not possible to draw conclusions on the nature of the effect (it could be linguistic or cultural) it is interesting to note that nativeness plays a role in such judgments.

## 4. Acoustic cues to involvement

Since our long term goal is automatically detecting hot spots, an important question is: What cues could help us to detect involvement? Here we focus on prosody for two reasons. First, at least for now, there is not enough data in the corpus to allow robust language modeling. Second, prosody does not require the results of an automatic speech recogniser, which might not be available for certain audio browsing applications or have a poor performance on the meeting data.

Involvement seems to be related to emotion although not in a very distinctive way. Involvement comprises a wide range of emotion-related states such as amusement, surprise, excitement, curiosity etc. From research on speech and emotion it is known that prosodic features, especially F0, show good correlations with certain types of emotions. It has proven useful to describe emotions in terms of two dimensions: *evaluation* and *activation* (e.g. [1]). Evaluation describes the positive or negative valence that is associated with a feeling. Activation is defined as "the strength of the person's disposition to take some action" ([2] p. 39). This is in fact closely related to the notion of involvement which means that a speaker displays a high interest or concern in a current topic — be it in terms of positive approval or negative disagreement.

In [2], results of several investigations are reported that indicate that acoustic features tend to be more dependent on such dimensions than on emotions. According to this study, an increase in mean and range of the fundamental frequency (F0) can be observed in more activated speech as well as tense voice quality. In general, pitch related measures, energy and duration can be useful indicators of emotion.

## 4.1. Acoustic features

Given the hypothesized relationship between involvement and activation, several F0 features were computed and compared between utterances rated as involved and those rated as non involved. Additionally, energy based features of voiced segments were computed. For each word either the average (**av-**), minimum (**mi-**), or maximum (**ma-**) value was considered[1]. In order to obtain a single value for an utterance, the average (**av-**), minimum (**mi-**) or maximum (**ma-**) over all words was computed. Only the most meaningful features were considered. For example, instead of computing the minimum of the maximum values the range (**rg-**) was computed as the difference between the **ma-** and **mi-** values of an utterance. For F0 and energy, either absolute (**a-**) or normalised (**b- z- bz-**) values were used.

The most simple normalisation scheme that was used was the z-score. It is computed by removing the mean obtained over all values of a speaker in a meeting and dividing by the corresponding standard deviation. This normalisation is denoted by **z-**. For F0 a more sophisticated normalisation was performed with respect to the baseline $BL_{spk}$, i.e. the lowest assumed F0 value, of a speaker[2]. The normalised F0 value was achieved by

$$\textbf{b-F0} = \log \frac{\text{F0}}{BL_{spk}} \tag{1}$$

These baseline normalised F0 values can also be mean and variance normalised over all values of a speaker in a meeting. This is indicated by **bz-**. Note that the **b-** and **bz-** normalisations can only be applied to F0. In summary, the name of a feature is generally composed by four parts:

| normal-isation | utterance level | word level | basic feature |
|---|---|---|---|
| $\begin{bmatrix} a- \\ b- \\ z- \\ bz- \end{bmatrix}$ | $\begin{bmatrix} ma- \\ mi- \\ av- \\ rg- \end{bmatrix}$ | $\begin{bmatrix} ma- \\ mi- \\ av- \end{bmatrix}$ | $\begin{bmatrix} F0 \\ En \end{bmatrix}$ |

[1]We used word based features out of convenience because they are provided in a prosody database for the meeting corpus. We expect similar performance for frame based features.

[2]For more details on how to determine the baseline frequency cf. [7]. For a short overview of the computation of the prosodic features in general cf. [8]

For example, the measure **bz-ma-av-F0** is the maximum value in an utterance of the average F0 values of its words with the F0 values being baseline and z-score normalised. Features where the same operations are performed on word and on utterance level only receive one index. Thus, **a-av-F0** is the mean absolute F0 value of each word averaged over the utterance. In total, 48 measures based on F0 and energy were computed in this manner.

## 4.2. Correlations with perceived involvement

In order to determine which features are useful for a classifier of hot spots the rated utterances have to be related to the prosodic features. For the analysis of the acoustic features 88 utterances were taken into account for which at least 3 ratings were available. Before correlations can be computed the utterances have to be classified. But how can a class be assigned to an utterance when not all raters agree? Since there is no ground truth in determining the class label of an utterance, the class assigned to each utterance was determined as a weighted version of the ratings. Thus, if an utterance had ratings for both classes its features contributed accordingly with different weights to each class. For example, for an utterance with 5 ratings as involved and the same number of ratings as non involved the value of the feature was weighted by 0.5 for each class. This is in fact a soft decision and accounts for the different ratings in an adequate way. The results of these weighted means are displayed in Figure 1.

The Figure shows the means and standard deviations for the 16 most distinguishing features for involved (stars) and non involved (crosses) utterances (cf. Table 2 for the identity of the features). For purposes of comparison across features with different ranges the values are mean and variance normalised with respect to the values of each feature over all rated utterances. On the left side are the features with a large difference between the means of the two classes; on the right side are those with more overlap. The differences between the two classes are significant (t-test, $p < .01$) for all features except the last three. Note that the most affected features given in Table 3 are all F0 based — the first energy feature appears only at place 14. Also, the first features are all either baseline or z-score normalised or both. In general, baseline and variance normalised features (bz-) show a clearer separation indicating that F0 needs speaker normalisation with respect to both, variance and baseline. Not surprisingly, absolute measures can distinguish less well between the two classes. It is interesting to note that the features which are affected most by involvement tend to be combinations of mean and maximum values of F0. This suggests that F0 is increased in general whereas the range is not affected to the same extent. Indeed, the range of F0 **b-rg-F0** occurs only at rank 19 in the list of the most affected features.

### 4.2.1. Within-speaker comparison

Results so far are for all speakers. In order to show that the features do not behave differently for individual speakers, Figure 2 shows the values of the 16 features from Table 3 for one particular speaker. It turns out that the pattern remains similar and the most distinguishing features are roughly the same. Further investigations are necessary to analyse how much of the variance in the features of the involvement classes in figure 1 are due to differences between speakers. But this example suggests that the normalisation removes a significant part of the variability due to specific speakes.
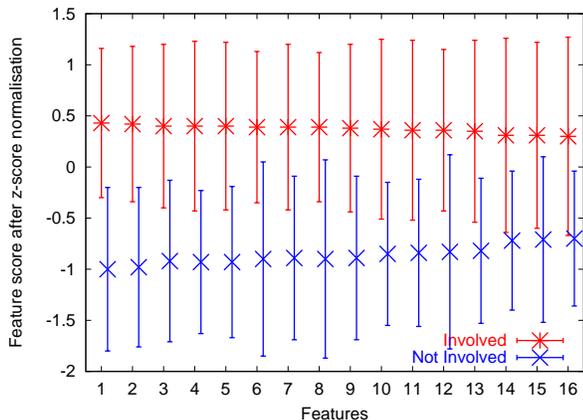
Figure 1: Means and standard deviations of the normalised prosodic features of utterances of all speakers rated as involved and not involved. The top 16 features are shown in the order of the differences between the means of involved versus non involved utterances.
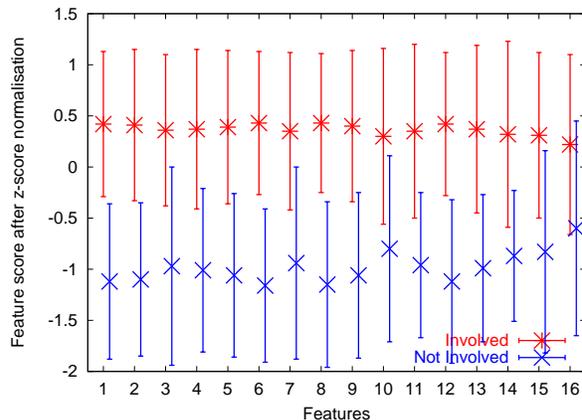


Figure 2: Means and standard deviations of the normalised prosodic features of utterances of one speaker rated as involved and not involved. The top 16 features are shown in the order of the differences between the means of involved versus non involved utterances for all speakers.

| | | | |
|---|---|---|---|
| 1 bz-ma-av-F0 | 13 b-av-F0 | 25 bz-rg-F0 | 37 a-ma-F0 |
| 2 z-ma-av-F0 | 14 z-av-En | 26 a-av-En | 38 a-av-ma-F0 |
| 3 z-av-ma-F0 | 15 z-av-mi-F0 | 27 a-ma-En | 39 a-av-F0 |
| 4 bz-av-F0 | 16 z-av-ma-En | 28 a-rg-En | 40 a-mi-En |
| 5 bz-av-ma-F0 | 17 b-av-mi-F0 | 29 a-ma-av-En | 41 b-mi-F0 |
| 6 z-ma-F0 | 18 z-ma-av-En | 30 a-av-mi-En | 42 z-mi-En |
| 7 z-av-F0 | 19 b-rg-F0 | 31 b-mi-av-F0 | 43 a-rg-F0 |
| 8 bz-ma-F0 | 20 z-rg-En | 32 z-mi-av-En | 44 a-av-mi-F0 |
| 9 b-ma-av-F0 | 21 z-ma-En | 33 a-mi-av-En | 45 a-mi-av-F0 |
| 10 bz-av-mi-F0 | 22 z-av-mi-En | 34 z-rg-F0 | 46 z-mi-av-F0 |
| 11 b-av-ma-F0 | 23 a-av-ma-En | 35 bz-mi-F0 | 47 a-mi-F0 |
| 12 b-ma-F0 | 24 bz-mi-av-F0 | 36 a-ma-av-F0 | 48 z-mi-F0 |

Table 3: Features sorted according to the differences between the means of involved versus non involved utterances.

## 5. Conclusion

Despite the subjective nature of the task, raters show significant agreement in distinguishing involved from non-involved utterances. Since the utterances of the perception task were given in isolation to the human raters, it is likely that the judgments are mainly based on acoustic cues. However, differences in performance between native and nonnative raters indicate that judgments on involvement are also influenced by the native language of the listener.

Furthermore, reliable acoustic cues for involvement have been found. The prosodic features of the rated utterances indicate that involvement can be characterised by deviations in F0 and energy. It is likely that this is a general effect over all speakers as it was shown for a least one speaker that the most affected features of an individual speaker were similar to the most affected features that were computed over all speakers. If this holds true for all speakers this is an indication that the applied mean and variance as well as baseline normalisations are able to remove most of the variability between speakers.

These results are promising towards our longer term goal of automatically detecting hot spots in multi-party conversations.

## 7. References

[1] R. Cowie, E. Douglas-Cowie, S. Savvidou, E. MkMahon, M. Sawey, and M. Schröder, "FEELTRACE: An instrument for recording perceived emotion in real time," in *Proc. ISCA ITRW on Speech and Emotion: Developing a Conceptual Framework*, R. Cowie, Douglas-Cowie, and E. Schröder, Eds., Belfast, 2000, pp. 19–24.

[2] R. Cowie, E. Couglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. Taylor, "Emotion recognition in human-computer interaction," *IEEE Signal Processing Magazine*, vol. 18, no. 1, pp. 32–80, 2001.

[3] A. Janin, D. Baron, J. Edwards, E. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters, "The ICSI Meeting Corpus," in *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing*, 2003.

[4] R. T Cauldwell, "Where did the anger go? The role of context in interpreting emotion in speech," in *Proc. ISCA ITRW on Speech and Emotion: Developing a Conceptual Framework*, R. Cowie, Douglas-Cowie, and E. Schröder, Eds., Belfast, 2000, pp. 127–131.

[5] R. Cowie and R. R. Cornelius, "Describing the emotional states that are expressed in speech," *Speech Communication*, vol. 40, no. 1-2, pp. 5–32, 2003.

[6] J. L. Fleiss, "Measuring nominal scale agreement among many raters," *Psychological Bulletin*, vol. 76, no. 5, pp. 378–382, 1971.

[7] E. Shriberg, A. Stolcke, D. Hakkani-Tür, and G. Tür, "Prosody-based automatic segmentation of speech into sentences and topics," *Speech Communication*, vol. 32, no. 1-2, pp. 127–154, Sept. 2000.

[8] D. Baron, E. Shriberg, and A. Stolcke, "Automatic punctuation and disfluency detection in multi-party meetings using prosodic and lexical cues," in *Proc. Int. Conf. on Spoken Language Processing*, Denver, USA, 2002, pp. 949–952.