A data-driven approach to detection of interruptions in human–human conversations

Raveesh Meena, Saeed Dabbaghchian, Kalin Stefanov
Department of Speech, Music and Hearing (TMH)
KTH, Stockholm
raveesh@csc.kth.se, saeedd@kth.se, kalins@kth.se

Abstract

We report the results of our initial efforts towards automatic detection of user's interruptions in a spoken human–machine dialogue. In a first step, we explored the use of automatically extractable acoustic features, *frequency* and *intensity*, in discriminating listener's interruptions in human–human conversations. A preliminary analysis of interaction snippets from the HCRC Map Task corpus suggests that for the task at hand, intensity is a stronger feature than frequency, and using intensity in combination with feature loudness offers the best results for a k-means clustering algorithm.

Introduction

Interruptions are important elements of conversations. They contribute in mediation of the content and redirection of a conversational exchange. Human-like conversational dialogue systems should not only be able to 1) use interruptions as a means to regulate the direction of a conversation, but also 2) discriminate user's interruptions from backchannels, and turn-taking attempts, and select an appropriate response. A system's insensitivity to user's interruptions could possibly render a dialogue inefficient and have adverse effect on user experience. In this work, we aim at building a computation model for automatic detection of interruptions and in human–human conversations.

Background

Various works have analyzed the acoustic and prosodic characteristics of conversational elements such as interruptions, backchannels and turn-changes. Yang (2001) analyzed the maximum pitch and intensity in speaker turns, and described the function of interruptions in managing local and global coherence in conversation that is brought about through the systematic phrase-to-phrase prosodic patterns of discourse. For example, a speaker attempts at taking the conversational floor while the main speaker is speaking (referred to as *competitive interruptions*) are characterized by high pitch and amplitude. In contrast, speaker statements supporting the main speaker's contentions, with no intention to take the conversational floor (referred to as *cooperative interruptions*) often occur at low or medium pitch levels.

In a related work, Gravano & Hirschberg (2012) examine interruptions in a corpus of spontaneous task-oriented dialogue and report a number of significant differences between interrupting and non-interrupting turns, based on features such as speaking rate, mean intensity, mean pitch, and duration of speaker speech. Lee & Narayanan (2010) analyzed the differences between competitive and cooperative interruption with features, change and activeness, employing audio, visual, and dis-fluency data. They have shown that the using these features in combination offers better results in discriminating between the two types than using any single feature modality.

Our work is motivated from the observation made in this literature regarding the distinct acoustic characteristic of backchannels, interruptions, and turn-changes. However, unlike the supervised methods for classification used in Lee et al. (2008) and Lee & Narayanan (2010) we 1) take an unsupervised approach to automatically cluster speaker utterances into interruptions, backchannels, and turn-change categories; and 2) use a fully automatic scheme for extraction of frequency and intensity features for training a model for online use.

Method

Corpus

To get a feeling for the task at hand, we started with a relatively small subset of the HCRC Map Task corpus (Anderson et al., 1991). In the Map Task interaction one of the dialogue participants (*giver*) provide instructions to the other human participant (*follower*) about finding her way to a destination on a map. In one setting, participants have no visual contact with each other, and as the respective maps are not completely identical (absence or presence of landmarks or difference in landmark names) the conversations inevitably involves: clarifications, acknowledgements, backchannels, interruptions, turn-changes etc. Three Map Task interactions (average duration 15 min) were randomly picked and only the first 3 min of the interaction was analyzed for this initial work. The resulting dataset contains 5 unique participants (3 male and 2 female).

Feature extraction

We explored the use of two acoustic features: frequency and intensity, for the task at hand. We used inter-pausal units (IPUs): speech units separated by 200 milliseconds of silence as the basic unit of processing, i.e. deciding whether a speaker IPU is an interruption or a backchannel or a turn-change. Towards this, we first used a voice activity detector to automatically segment speakers' speech into IPUs. The automatic segmentation method is not perfect and doesn't always produce segmentation around turns with simultaneous speech and speech regions with low energy. This results in some user speech units getting lost in the next processing stage. This indicates the issues and limitations of a fully automatic system for the task at hand. Next, for each speaker IPU we extracted the maximum f0 and intensity values using Wavesurfer toolkit (Sjölander & Beskow, 2000). The feature values were z-normalized $(z = (x - \mu) /\sigma)$) for building a speaker invariant model. In addition to this, we used the perception level features: maximum pitch and loudness (the log semitone equivalents of frequency and intensity, respectively).

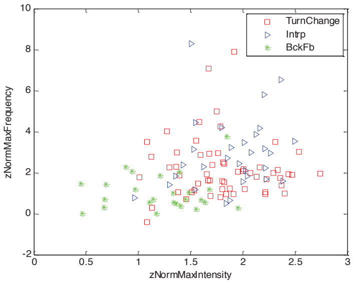

Figure 1. Spread of training instances with z-normalized max frequency and intensity

The spread of training instances in our dataset using the features zNormMaxFrequency and zNormMaxIntensity is illustrated in Figure 1. In order to obtain the ground truth of the category of speaker IPUs in our data, the authors of this paper labeled the IPUs with one of the three categories: Intrp (the IPU is an interruption), BckFb (the IPU is a backchannel), and TurnChange (the IPU marks a turn-change). Since this is still an exploratory work, the judges sat together and labeled the data unanimously (of course, in a larger study this would have been done more formally along with Kappa scores for inter-annotator agreement). In our training set we have 30 instances of interruption, 25 instances of backchannel, and 58 instances of turn-change.

A cursory look at Figure 1 suggests that interruptions (Intrp) indeed tend to have higher maximum intensity and frequency. Backchannels (BckFb), in contrast are in the lower end of the spectrum. The instances of turn-change lie somewhere in the middle, but are spread largely over interruptions, suggesting that it would be hard to discriminate interruptions and turn-changes. A univariate analysis of variances of the means of zNormMaxFrequency and zNormMaxIntensity suggest that only the backchannel category differ significantly from interruption and turn-change categories. This suggested that using these two features only we would not have much success in discriminating between the three categories. Therefore, for the remaining part of this paper we focus only on the task of discriminating interruptions from backchannels. Figure 2 illustrates the distribution of interruptions and backchannels in our training dataset.

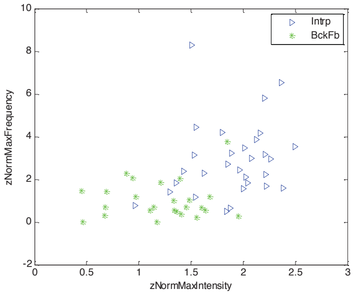

Figure 2. Spread of interruptions and backchannels in the training set (55 instances).

Result

Clustering and classification

We used the centroid based k-means clustering algorithm (with k=2) to automatically cluster our training dataset consisting of only backchannels and interruptions. In Figure 3, the two black circles indicate the two cluster centroids (1.42, 1.19) and (1.99, 4.14), with 39 and 16 instances in the respective clusters. Based on the observations made in the literature that interruptions are characterized by higher frequency and intensity, we label the cluster with centroid (1.99, 4.14) as the cluster representing interruptions, and the other cluster as representing backchannels. If we treat the cluster label of instances in these clusters as their learned category, we would correctly label 78.1% of the training dataset. The training instances with erroneous classification are indicated with red color in Figure 3.

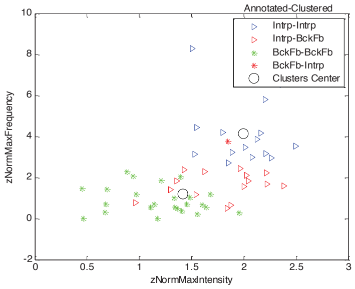

Figure 3. K-means clustering and classification results. Misclassified instances are indicated in red color.

Since the majority class in our dataset is interruptions (54%), the accuracy of 78.1% is a huge improvement over the majority class baseline. Table 1 summarizes the performances of the various (additive) features explored in this work. Both z-normalized max intensity and loudness are stronger features in comparison to frequency and pitch. Using the combination of z-normalized maximum intensity and loudness we obtained the best clustering performance of 80.0%. Table 2 presents the recall and performance corresponding to this feature combination. The model achieves a high F-measure for interruptions.

Table 1. Feature performances (where + indicates additive feature combinations)

Feature (s)	*Accuracy*
zNormMaxFrequency	70.9%
+ zNormMaxIntensity	78.1%
zNormMaxLoudness	76.3%
+ zNormMaxPitch	61.8%
zNormMaxIntensity +zNormMaxLoudness	**80.0%**

Table 2. Precision, Recall and F-measure of clustering using z-normalized max intensity and loudness

Precision		Recall		F-measure	
Intrp	BckFb	Intrp	BckFb	Intrp	BckFb
0.77	0.85	0.90	0.68	0.83	0.76

Discussion

We have presented the preliminary results from our efforts towards automatic detection of user interruptions in a spoken human–human conversation. We formulated the task as that of clustering speaker utterances in three categories: interruption, backchannel, and turn-change. We explored the two acoustic features: z-normalized maximum frequency (and pitch) and intensity (as well as loudness) in speaker utterances. A preliminary analysis of interaction snippets from the HCRC Map Task corpus suggested that the task of discriminating between backchannel and interruption is instead more feasible on the dataset at hand. Using our fully automated approach to extract feature values, we have observed that intensity is a stronger feature in comparison to frequency, and using intensity in combination with loudness offers the best performance results for discriminating interruptions and backchannels.

The results obtained in this work are encouraging. In a next step it would be interesting to see whether scaling up the training set would provide similar or better results. More data should help us return to our original task of automatic discrimination between the three categories.

It would be interesting to see if the performance of models presented here could be improved with using additional features, such as duration as suggested in Gravano & Hirschberg (2012), or measures of activity (how the values fluctuate in the overlapping speech regions) and change (shift in peak values) in Lee & Narayanan (2010). The dataset contains both overlap and non-overlap speech segments. A similar analysis on the data with a clear separation of the two cases would be an interesting investigation.

A major limitation of this work is that we have excluded turn-changes from the current dataset. As observed, with regard to their acoustic property (max frequency and intensity) turn-change overlap largely with interruptions. This suggests that one may want to explore contextual (dialogue act) and lexico-syntactic features for telling them apart from interruptions and backchannels.

References

Anderson, A., Bader, M., Bard, E., Boyle, E., Doherty, G., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., Sotillo, C., Thompson, H., & Weinert, R. (1991). The HCRC Map Task corpus. *Language and Speech, 34*(4), 351-366.

Gravano, A., & Hirschberg, J. (2012). A Corpus-Based Study of Interruptions in Spoken Dialogue. In *INTERSPEECH*. ISCA.

Lee, C-C., Lee, S., & Narayanan, S. S. (2008). An analysis of multimodal cues of interruption in dyadic spoken interactions. In *INTERSPEECH* (pp. 1678-1681). ISCA.

Lee, C-C., & Narayanan, S. (2010). Predicting interruptions in dyadic spoken interactions. In *ICASSP* (pp. 5250-5253). IEEE.

Sjölander, K., & Beskow, J. (2000). WaveSurfer - an open source speech tool. In Yuan, B., Huang, T., & Tang, X. (Eds.), *Proceedings of ICSLP 2000, 6th Intl Conf on Spoken Language Processing* (pp. 464-467). Beijing.

Yang, L-C. (2001). Visualizing Spoken Discourse: Prosodic Form and Discourse Functions of Interruptions. In *Proceedings of the Second SIGdial Workshop on Discourse and Dialogue - Volume 16* (pp. 1-10). Stroudsburg, PA, USA: Association for Computational Linguistics.