



Heriot-Watt University  
Research Gateway

## Information density and overlap in spoken dialogue

### Citation for published version:

Dethlefs, N, Hastie, H, Cuayahuitl, H, Yu, Y, Rieser, V & Lemon, O 2016, 'Information density and overlap in spoken dialogue', *Computer Speech and Language*, vol. 37, pp. 82-97.  
<https://doi.org/10.1016/j.csl.2015.11.001>

### Digital Object Identifier (DOI):

[10.1016/j.csl.2015.11.001](https://doi.org/10.1016/j.csl.2015.11.001)

### Link:

[Link to publication record in Heriot-Watt Research Portal](#)

### Document Version:

Peer reviewed version

### Published In:

Computer Speech and Language

### General rights

Copyright for the publications made accessible via Heriot-Watt Research Portal is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

### Take down policy

Heriot-Watt University has made every reasonable effort to ensure that the content in Heriot-Watt Research Portal complies with UK legislation. If you believe that the public display of this file breaches copyright please contact [open.access@hw.ac.uk](mailto:open.access@hw.ac.uk) providing details, and we will remove access to the work immediately and investigate your claim.

Manuscript Number: CSL14-160R2

Title: Information Density and Overlap in Spoken Dialogue

Article Type: Research Article

Keywords: overlap, turn-taking, information density, incremental processing, spoken dialogue systems

Corresponding Author: Dr. Nina Dethlefs,

Corresponding Author's Institution: Heriot-Watt University

First Author: Nina Dethlefs

Order of Authors: Nina Dethlefs; Helen Hastie; Heriberto Cuayáhuitl; Yanchao Yu; Verena Rieser; Oliver Lemon

**Abstract:** Incremental dialogue systems are often perceived as more responsive and natural because they are able to address phenomena of turn-taking and overlapping speech, such as backchannels or barge-ins. Previous work in this area has often identified distinctive prosodic features, or features relating to syntactic or semantic completeness, as marking appropriate places of turn-taking. In a separate strand of work, psycholinguistic studies have established a connection between information density and prominence in language---the less expected a linguistic unit is in a particular context, the more likely it is to be linguistically marked. This has been observed across linguistic levels, including the prosodic, which plays an important role in predicting overlapping speech. In this article, we explore the hypothesis that information density (ID) also plays a role in turn-taking. Specifically, we aim to show that humans are sensitive to the peaks and troughs of information density in speech, and that overlapping speech at ID troughs is perceived as more acceptable than overlaps at ID peaks. To test our hypothesis, we collect human ratings for three models of generating overlapping speech based on features of: (1) prosody and semantic or syntactic completeness, (2) information density, and (3) both types of information. Results show that over 50\% of users preferred the version using both types of features, followed by a preference for information density features alone. This indicates a clear human sensitivity to the effects of information density in spoken language and provides a strong motivation to adopt this metric for the design, development and evaluation of turn-taking modules in spoken and incremental dialogue systems.

Hull, October 29 2015

Dr. Nina Dethlefs  
University of Hull  
Department of Modern Languages  
Hull HU6 7RX  
England, UK  
Email: n.s.dethlefs@gmail.com

The Editor of *Computer Speech and Language*

**Second Revision of Manuscript CSL14-160**

Dear Prof. Moore,

please find enclosed our revised manuscript “Information Density and Overlap in Spoken Dialogue”. The manuscript has received a further set of thorough revisions taking all the comments of the latest review round into account. We have enclosed a document that lists all points raised in the review and the way we have addressed them. We made the following main revisions:

- We have provided further clarification on the exact computation of information density from user utterances, both with regard to the term *uniform information density* as well as its incremental calculation.
- We have removed the regression experiments from the article and instead focus on analysing the results of a decision tree classifier.
- We have separated the discussions on previous work on human-human overlap and results we have obtained from human-system overlaps, so as not to cause confusion by mixing the two.
- We have provided an argument to exclude the *no interruption* option from our experiment, even though we agree that a spoken dialogue system should in general always include the option to not overlap with the user.

We thank the reviewers for their very helpful comments and believe that our revisions have substantially improved the article. Thank you for taking our submission into consideration. We look forward to hearing from you.

Kind regards,

Nina Dethlefs, on behalf of all authors.

### Highlights (for review)

- information density, related to entropy, is related to overlaps in spoken language
- humans prefer overlaps based on information density and suprasegmental features
- this is confirmed in a speech-based rating study ( $p < 0.0001$ )
- our results are relevant for spoken dialogue systems, especially incremental ones

## Author Responses CSL14-160

This paper augments prior work by Gravano and Hirschberg with the inclusion of a measure of information density to identify appropriate backchannel locations. This is a fairly narrow contribution, but one that is certainly worthy of publication. However, the presentation of what was done in this study is not clear enough to advance understanding of the investigated turn-taking phenomena nor the technological implications of the findings. Without clarification of the definition and use of information density, and the removal of the poorly motivated regression analyses it is difficult to recommend this manuscript for publication.

**Q1:** Much of the literature review comes from analyses of human-human communication, and seems to imply that this is valid for human-machine interaction. As described on page 8, this is a very common way to learn about and model human-human behavior. However on page 9 in describing Cuyáhuítl et al. 2013 the presentation seems to suggest that human-machine and human-human dialogs may be incompatible. How should a reader reconcile this?

**Response:** It is correct that we have followed related work in assuming that the turn taking principles of human-human interaction can be transferred to human-machine interaction. In order to not state a contradiction early on in the text, we have removed the paragraph in question from Page 9. Instead, we have added it as an item to be explored in more detail in future work. We have at present no detailed answer as to whether results from human-human communication are indeed one-to-one transferable to spoken dialogue systems. Most of previous work has made this assumption, and our results seem to suggest that at least to an extent they are transferable. Nonetheless, our earlier results in Cuayahuitl et al. (2013) suggest that some differences may exist for user-initiated overlap, so that future work may need to investigate more closely.

**Q2:** Section 3.2. Uniform information density definition: Clearly  $p(w|history)$  is not equal for every word  $w$  in a linguistic unit (both intuitively and as presented in Figure 3). Thus  $\log 1/p(w)$  is not equal. In what respect is the information density considered to be "uniform"? This should be made clearer in this section. As an addendum to this comment, the description of Figure 3 at the end of page 14 in section 3.3 states, "We can observe [in Figure 3] that information is distributed relatively uniformly...". This is not at all what I would consider a uniform distribution. There are clear peaks and valleys from 2-3 bit-like units up to 6-7 bit-like units. (It's not clear if these are actually bits, or a sum of the bits in the utterance so far, but I think the former.)

**Response:** "Uniform information density" is a term used in the psycholinguistic literature we refer to in the article, but we recognise that it might be slightly out of context in our article. "Uniform" is not meant to say that all linguistic units transmit exactly the same bits of information. It rather means that information

bits are kept within a certain range, practically this range seems to be between 2 and 6 bits with an average of about 4 (according to both our own experiments and previous work that has investigated information density). One could speculate that the communicative channel is roughly defined by this 2-6 bit range. To make this clearer in the article, we have removed all references to “uniform” information density and have rewritten Section 3.2 to be clearer on how information is distributed across utterances. The same has been done in Section 3.3.

**Q3:** Equation 1 isn't labeled. I assume that it is referring to the  $-\log \frac{1}{P(u_k)}$  definition. However, it is unclear in Figure 3 how the information in the utterance is plotted per word.

**Response:** Yes, thank you, we have provided a label now. We have also added a second equation that exemplifies how information density is computed per word for the first example given in Figure 3.

**Q4:** Why does is the difference in mean information density between manual transcription and ASR hypotheses a meaningful measure? It is not at all clear what we can glean from a pearson correlation coefficient of 0.71. Does this suggest that when errors are made that they are between words of approximately equal probability? Even if this were so, is that a good thing?

**Response:** Our aim was to provide a measure of how transferable the information density results for transcribed utterances would be for ASR-hypotheses, which might contain errors. However, since the Pearson correlation coefficient is admittedly not an ideal measure here and also because the experimental results in Section 4.3 are a much stronger indicator of the transferability to ASR hypotheses, we decided to omit Section 3.4 from the manuscript. It didn't seem to add anything crucial to the contents that are otherwise presented.

**Q5:** In section 3.4, "This is illustrated in Figure 6" I believe this is referring to Figure 4. Regardless, this single example does not to a particularly effective job in indicating that information density calculated from ASR hypotheses is as useful as manual transcription. The hypothesis in question is fairly high quality.

**Response:** Yes, we agree with this and removed Figure 4 from the manuscript alongside Section 3.4

**Q6:** Minor comment: In section 4.2 Figure 5 is referred to as Figure 7.

**Response:** Thank you, we have corrected this.

**Q7:** I am still confused over the use of a regression model to perform this analysis. Similar analysis can be performed on linear classifiers (like logistic regression) where informative (normalized) features have higher weights or the J48 decision tree where more informative features appear higher in the tree. Treating the target as a regression is more like asking "how correct is this placement" rather than "is this a correct placement or not". It's a rather strange way of posing this problem. As the only justification for using regression over classification is the ability to determine the relative contribution of different features, and this can be done with a classifier, the analysis of Section 4.2 (and 4.3) should be done in the context of classification rather than regression.

**Response:** We have removed the regression experiments from the article and now focus instead on classification experiments using a J48 decision tree classifier. The graphics have been replaced accordingly. We now draw our conclusions from those features that seem most predictive of user preference from the classification experiments alone.

**Q8:** Calculation of information density. In a number of instances (examples at the end of section 4.3, Figure 4) the information density of the first token is 0. Why is this? the calculation would suggest that it should be  $\log 1/p(w)$ . Certainly  $p(w) \neq 1$  for these cases. I think I must be missing something about this calculation. This should be clarified in the paper.

**Response:** This is correct and we have corrected it in the article as well.

**Q9:** I have a related question about the formulation in Equation 1 with respect to the plots presented throughout the paper. The equation suggests that information density is monotonically increasing with the length of the utterance  $k$ . Is this intentional? If the calculation of information density is pointwise (i.e.  $\log 1/p(w|history)$ ) for each word separately. {This by the way would be more like 'surprisal' than 'information density' but that's ok.} then the plots make more sense, but this is not made clear from the equation. On the other hand if the plot is the incremental information density, then it's not at all clear how there would ever be a valley in this value since  $I(w_1...w_n) \leq I(w_1...w_{n+1})$ .

**Response:** Yes, it is correct that we compute the information density for each word in a point-wise fashion. We have clarified this in Section 3.2 in relation to Equation 1, the formulation of information density. The peaks and troughs in information density occur because we follow previous work in computing information density based on trigrams. This means that whenever the sum of information density of the last words is lower than the sum of previous words, we'd see a trough. We have clarified this in Section 3.3 as well.

**Q10:** In response to R3Q10 "(For Fig. 6, how about no interruption as an option? You may arrive at a different conclusion") the authors write:



"Giving users the option to tick "no interruption" would very likely lead to different results, yes. However, since we're trying to answer the question of when to interrupt if we really had to / wanted to, it wouldn't greatly help our cause."

The manuscript needs to be more clear in what instances you "really have to" or "really want to" interrupt, and then guarantee that the examples fall under these criteria. It seems as though a system response of "no interruption" should always be valid. The experimental question the authors pose is "[assuming the system must interrupt the user], where the best/least intrusive/most natural place for a system to interrupt a user". It is a somewhat substantial assumption to say that the system must interrupt the user. By ignoring the possibility of no-overlap the experiment may be asking raters to evaluate the best of two bad options by eliminating the best choice. This experimental design choice needs to be strongly motivated.

**Response:** We have clarified our research question in Section 4.2 making it clear that it is to establish the best point for an overlap in a situation when the system really wanted to produce one. Possible situations for this would be the production of a backchannel, e.g., to signal continued attention to a longer user utterance or indeed the production of a barge-in to clarify an ASR error before it leads to more errors later in the interaction. In the former case, the system would not actually try to take the turn but just offer feedback at the least intrusive point. The second case of overlap is presumably more risky than the former in terms of its effect on user satisfaction, so will need to be investigated in follow-up research involving a task-based evaluation with a spoken dialogue system. We do recognise that in general a spoken dialogue system will always need to have the option to not overlap with the user, but for this study and our particular research question, we felt we would gain more by not offering a "no interrupt" option. In fact, we were concerned that users might always tick the "no interrupt" option by default because they feel it would be socially inappropriate for a system to interrupt a human user. This however would not have helped us to shed light on suitable points for overlaps in a human-system interaction. In this way, we hope to motivate future research into the options surrounding system-led overlap.

1  
2  
3  
4  
5  
6  
7  
8  
9 Information Density and Overlap in Spoken Dialogue

10  
11 Nina Dethlefs<sup>1</sup>, Helen Hastie<sup>2</sup>, Heriberto Cuayáhuitl<sup>2</sup>,  
12 Yanchao Yu<sup>2</sup>, Verena Rieser<sup>2</sup> and Oliver Lemon<sup>2</sup>

13  
14 <sup>1</sup>*University of Hull*  
15 *Department of Modern Languages*

16 *Hull HU6 7RX*  
17 *United Kingdom*

18 *Email: n.dethlefs@hull.ac.uk*

19  
20 <sup>2</sup>*Heriot-Watt University*  
21 *School of Mathematical and Computer Sciences*

22 *Edinburgh EH14 4AS*  
23 *United Kingdom*

---

24  
25  
26  
27  
28 **Abstract**

29  
30 Incremental dialogue systems are often perceived as more responsive and natu-  
31 ral because they are able to address phenomena of turn-taking and overlapping  
32 speech, such as backchannels or barge-ins. Previous work in this area has of-  
33 ten identified distinctive prosodic features, or features relating to syntactic or  
34 semantic completeness, as marking appropriate places of turn-taking. In a sep-  
35 arate strand of work, psycholinguistic studies have established a connection  
36 between information density and prominence in language—the less expected a  
37 linguistic unit is in a particular context, the more likely it is to be linguistically  
38 marked. This has been observed across linguistic levels, including the prosodic,  
39 which plays an important role in predicting overlapping speech.

40  
41 In this article, we explore the hypothesis that information density (ID) also  
42 plays a role in turn-taking. Specifically, we aim to show that humans are sensi-  
43 tive to the peaks and troughs of information density in speech, and that over-  
44 lapping speech at ID troughs is perceived as more acceptable than overlaps at  
45 ID peaks. To test our hypothesis, we collect human ratings for three models  
46 of generating overlapping speech based on features of: (1) prosody and seman-  
47 tic or syntactic completeness, (2) information density, and (3) both types of

1  
2  
3  
4  
5  
6  
7  
8  
9 information. Results show that over 50% of users preferred the version using  
10 both types of features, followed by a preference for information density features  
11 alone. This indicates a clear human sensitivity to the effects of information den-  
12 sity in spoken language and provides a strong motivation to adopt this metric  
13 for the design, development and evaluation of turn-taking modules in spoken  
14 and incremental dialogue systems.  
15  
16

17  
18 *Keywords:* overlap, turn-taking, information density, incremental processing,  
19 spoken dialogue systems  
20  
21

---

## 22 23 **1. Introduction**

24  
25  
26 Traditionally, the smallest unit of processing in spoken dialogue systems has  
27 been a full utterance with strict, rigid turn-taking. More recently, however,  
28 work on incremental systems has shown that processing smaller ‘chunks’ of user  
29 input can improve the user experience by providing faster responses and allow  
30 more flexibility in turn-taking (Skantze and Schlangen, 2009; Purver and Otsuka,  
31 2003; Skantze and Hjalmarsson, 2010; Baumann et al., 2011; Raux and Eskenazi,  
32 2009; Dethlefs et al., 2012b). Incrementality in spoken dialogue systems enables  
33 the system designer to model several dialogue phenomena that play a vital  
34 role in human conversation (Levelt, 1989), but have so far been absent from  
35 most systems. These include more natural turn-taking and grounding through  
36 the generation of backchannels and barge-ins—which we will refer to jointly as  
37 *overlaps* in this article.  
38  
39

40  
41  
42 Previous studies on the triggers of backchannels and barge-ins in human-  
43 human conversation have revealed the importance of prosodic features, such  
44 as pitch, duration, and energy, and features relating to syntactic and semantic  
45 completeness (Koiso et al., 1998; Ward and Tsukahara, 2000; Cathcart et al.,  
46 2003; Morency et al., 2008; Gravano and Hirschberg, 2009; Oertel et al., 2012).  
47 The latter can refer to the grammatical completeness of constituents, e.g., such  
48 as a full NP versus just the determiner. We will refer to such features jointly  
49 as *suprasegmental*. Most previous studies have relied on manually annotated  
50  
51  
52  
53  
54  
55  
56  
57  
58

1  
2  
3  
4  
5  
6  
7  
8  
9 corpora for their analyses and reported results from held-out datasets, and few  
10 findings have been implemented in real spoken dialogue systems.  
11

12 In a separate strand of research, psycholinguistic studies have shown that  
13 humans distribute information across linguistic units in a way so that more  
14 prominence is given to units that are less expected in a given context (Genzel and  
15 Charniak, 2002; Bell et al., 2003; Aylett and Turk, 2004; Levy and Jaeger, 2007).  
16 This evidence led us to hypothesise that there is a relation between information  
17 density and suitable places for backchannels or barge-ins in spoken conversation.  
18 Information density can be seen as a measure of entropy in human language and  
19 is computed from a language model of the domain at hand (Shannon, 1948). One  
20 advantage is therefore that it can easily be obtained incrementally for incoming  
21 strings of user speech. A further advantage of information density over other  
22 features, relating e.g. to syntactic completeness, is that it can be seen as an  
23 ‘abstract’ type of information. Information is estimated solely based on  $n$ -grams  
24 and we do not need to understand *what* is being said on a semantic level.  
25  
26  
27  
28  
29  
30  
31  
32

33 In a study that explored the relationship between information density and  
34 overlaps (Dethlefs et al., 2012a), we trained a hierarchical reinforcement learner  
35 that could generate backchannels and barge-ins in conversations with human  
36 users. The model compared a reward function that was sensitive to information  
37 density against a reward function that was not. Results showed that significantly  
38 higher human ratings were obtained for the version that took information den-  
39 sity into account. While these results are promising, they were drawn from  
40 an exclusively text-based rating study, which potentially does not account for  
41 the peculiarities of spoken language. In this article, we therefore replicate our  
42 earlier experiments in a speech-based rating study, involving word-based as well  
43 as suprasegmental features, in order to see whether the earlier results hold in  
44 a realistic dialogue setting. Results show a clear human preference for a model  
45 that generates overlapping speech based on both suprasegmental and informa-  
46 tion density features. This is followed by overlaps based on information density  
47 features alone and then suprasegmental features alone. The results indicate  
48 a strong human sensitivity to the peaks and troughs in evolving information  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58

1  
2  
3  
4  
5  
6  
7  
8  
9 density in spoken language. These results hold even in the face of ASR errors.

10 We will start Section 2 by discussing related work on overlap in spoken  
11 dialogue systems, mainly from the perspective of incremental processing archi-  
12 tectures. We will then describe the types of features that previous work has  
13 identified as predicting different types of overlaps, and finally the information  
14 density effects that have been observed across linguistic units in human lan-  
15 guage. Section 3 will introduce the notion of information density and exemplify  
16 some of its effects on a spoken corpus from the information-seeking dialogue  
17 domain. The relation between information density and suprasegmental features  
18 in spoken language is also discussed. In Section 4, we describe our experimental  
19 setting, data and methodology, and present results on the effect of information  
20 density on spoken overlap in dialogue. Section 5 finally draws conclusions and  
21 lays out the directions for future research.  
22  
23  
24  
25  
26  
27  
28  
29

## 30 **2. Related work**

31  
32  
33 The production of backchannels and barge-ins has long been recognised to  
34 facilitate grounding, feedback and clarifications in human spoken dialogue (e.g.,  
35 Yankelovich et al. (1995)). With the rise of incremental processing architectures  
36 (Schlangen and Skantze, 2009; Dethlefs et al., 2012b; Selfridge et al., 2011; De-  
37 Vault et al., 2009), we now have the opportunity to integrate these phenom-  
38 ena into spoken dialogue systems. This section reviews the state of the art in  
39 incremental processing and the identification of triggers for backchannels and  
40 barge-ins in human dialogue. Finally, we discuss findings from information den-  
41 sity applied to spoken language and draw conclusions on how all aspects can be  
42 brought together into an effective model.  
43  
44  
45  
46  
47  
48  
49

### 50 *2.1. Incremental processing*

51  
52 Traditionally, the smallest processing unit in a dialogue system has been a  
53 full user utterance with correspondingly rigid turn-taking. With the rise of incre-  
54 mental architectures in recent years, however, it has become possible to model  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

1  
2  
3  
4  
5  
6  
7  
8  
9 several discourse phenomena that have previously been exclusive to human-  
10 human conversation. These phenomena include faster turn-taking, grounding  
11 through the generation of backchannels and feedback, and facilitated clarification  
12 through barge-ins. Recent work has shown that including such phenomena  
13 into human-computer interaction can significantly improve the user’s experience  
14 in terms of automatic speech recognition (Baumann et al., 2011), dialogue man-  
15 agement (Buss et al., 2010), dialogue act recognition (Cuayáhuitl et al., 2013)  
16 and speech generation (Skantze and Hjalmarsson, 2010).  
17

18  
19 The smallest unit of processing in incremental systems is called an *incre-*  
20 *mental unit* (IU) (Schlangen and Skantze, 2009). Its instantiation depends on  
21 the particular processing module. In speech recognition, IUs can correspond  
22 to phoneme sequences that are mapped onto words (Baumann and Schlangen,  
23 2011). In dialogue management, IUs can correspond to dialogue acts (Buss  
24 et al., 2010). In natural language and speech generation, IUs can correspond  
25 to single words, phrases or full dialogue acts (Skantze and Hjalmarsson, 2010;  
26 Dethlefs et al., 2012b). Finally, in speech synthesis, IUs can correspond to  
27 speech unit sequences which are mapped to segments and speech plans (Skantze  
28 and Hjalmarsson, 2010).  
29

30  
31 Figure 1 illustrates the contrast between traditional processing units and in-  
32 cremental units, where an advantage of the latter is that they allow more flexible  
33 turn-taking. While the non-incremental case in the Figure would process a full  
34 dialogue act, e.g., *inform(restaurant, venueName=Beluga, priceRange=moderate,*  
35 *foodType=Italian, area=city centre)* without giving a user the opportunity to  
36 barge-in, the incremental case is able to process smaller unit dialogue acts, such  
37 as *inform(restaurant, venueName=Beluga)*, *inform(restaurant, priceRange=mo-*  
38 *derate)*, etc. The advantage of the latter model is that a user barge-in over an  
39 incremental dialogue act would not lead to the entire system utterance being  
40 re-prompted at the next turn.  
41

42  
43 Phenomena of turn-taking have been the focus of several studies in incremen-  
44 tal processing. For example, Raux and Eskenazi (2009) optimise turn-taking in  
45 a dialogue system based on a cost matrix and decision theoretic principles, as-  
46  
47  
48  
49  
50  
51  
52

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

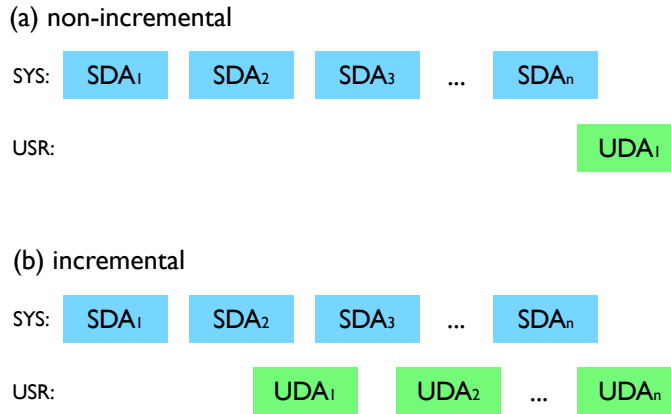


Figure 1: Contrast of traditional non-incremental processing (top) against incremental processing (bottom). The latter allows more flexible turn-taking and gives the user the opportunity to backchannel or barge-in leading to more efficient interactions. SDA here stands for ‘system dialogue act’ and UDA stands for ‘user dialogue act’.

suming that users prefer no gap and no overlap at turn boundaries. DeVault et al. (2009) allow a small “responsive” overlap between user and system utterances by predicting the completion and thus the end of a user utterance. Selfridge et al. (2011) incrementally predict the stability and accuracy of speech recognition hypotheses so as to enhance system performance without causing delays at turn boundaries.

Initial evidence therefore suggests that incremental architectures are able to offer the turn-taking flexibility required to model more of the discourse phenomena found in human language. Backchannels, barge-ins and some studies that aim to predict them in human conversation will be discussed in the following.

## 2.2. Backchannels and barge-ins

An important advantage of incremental architectures is that they are able to generate and process backchannels and barge-ins—often adding to the system’s reactivity. Figure 2 shows examples of both phenomena. *Backchannels* can

1  
2  
3  
4  
5  
6  
7  
8  
9 **Backchannel (the user backchannels)**

10 **USR** I want Italian food in the centre of town ...

11 **SYS** OK. I found 35 central Italian restaurants ...

12 **USR** OK.

13 **SYS** The restaurant *Verona* has great food but is also a bit expensive.

14 The *Roma* is cheaper, but not as central as *Verona* ...

15  
16  
17  
18 **Barge-in (the user barges in on system)**

19 **USR** I want Italian food in the centre of town ...

20 **SYS** I found 35 Indian ...

21 **USR** Not Indian, I want Italian.

22 **SYS** OK, Italian ...

23 **SYS** I have 24 Italian restaurants ...

24  
25  
26  
27 Figure 2: Examples of backchannels and barge-ins from human-computer dialogues in the  
28 restaurant domain. The first example represents a signal of grounding whereas the latter  
29 represents a correction to an initial system hypothesis. Utterances are aligned with the place  
30 at which they occur in the preceding utterance.  
31  
32

33  
34 often be interpreted as signals of grounding. Produced by the user, the system  
35 may infer that the user is following the presentation of information or is con-  
36 firming a piece of information without trying to take the turn. Similarly, we  
37 could allow a system to generate backchannels to the user to confirm that it  
38 understands the user's preferences. An important decision for a dialogue sys-  
39 tem then would be *when* to generate a backchannel. *Barge-ins* typically occur  
40 in different situations. The user may barge-in on the system to correct an ASR  
41 error (such as 'Italian' instead of 'Indian' in Figure 2). A system may want to  
42 barge-in on a user in order to confirm a low-confidence ASR hypothesis imme-  
43 diately so as to start its database look-up. In the latter case, the system will  
44 need to decide *if* and *when* to generate a barge-in. Both overlap phenomena  
45 are presumably particularly relevant to hands-free, eyes-free scenarios. Previous  
46 work has confirmed that users of spoken dialogue systems do require feedback  
47 so as to know whether the system is still listening to them or processing their  
48 request (Yankelovich et al., 1995). However, it has also been shown that feed-  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58



1  
2  
3  
4  
5  
6  
7  
8  
9 back needs to occur at the right moment in order not to confuse the user rather  
10 than helping (Hirasawa et al., 1999). Several studies have therefore investigated  
11 the linguistic cues that signal suitable points for backchannels or barge-ins in  
12 human-human dialogue. Common findings have been a final falling or rising  
13 pitch or final low/high pitch levels as distinctive prosodic features indicating  
14 the end of a turn (Ward and Tsukahara, 2000; Koiso et al., 1998). Duration  
15 and energy can sometimes play a role, as well as features relating to semantic  
16 or syntactic completeness (Ward and Tsukahara, 2000; Cathcart et al., 2003;  
17 Morency et al., 2008). The latter tend to denote the completion of a gram-  
18 matical clause or constituents as indicated through its Part-of-Speech (POS)  
19 sequence. Several authors have observed that a combination of features leads  
20 to improved performance Koiso et al. (1998); Gravano and Hirschberg (2009).

21  
22  
23  
24  
25  
26  
27  
28 A common approach to investigating turn-taking signals has been to anno-  
29 tate data sets of human-human spoken conversation and then train a statistical  
30 prediction model from them. Focusing on predicting locations of overlapping  
31 speech, for example, Oertel et al. (2012) analyse a corpus of spoken human multi-  
32 party conversations. They demonstrate in a classification study that locations of  
33 overlapping speech are prosodically different from locations of non-overlapping  
34 speech. The prosodic features of a 5 seconds window surrounding an overlap are  
35 characterised by significantly higher intensity and F0 frequency and significantly  
36 smaller F0 range than in windows of non-overlapping speech. The authors in-  
37 terpret this as representing a potentially higher level of involvement of one of  
38 the speakers which leads to the observed prosodic patterns at overlaps.

39  
40  
41  
42  
43  
44  
45  
46 Another study (Gravano and Hirschberg, 2011) looks into the prosodic, syn-  
47 tactic and lexical cues that precede turn switches and backchannels in human-  
48 human conversation. Based on a classification study from human-labelled data,  
49 the authors identify seven cues that precede smooth turn switches (i.e., with-  
50 out overlap). Their unit of analysis is the inter-pausal unit (IPU), a sequence  
51 of words preceded and followed by a silence period of more than 50 ms. The  
52 following cues signalled suitable places for turn transitions with little variation  
53 in the authors' dataset:  
54  
55  
56  
57  
58

- 1
  - 2
  - 3
  - 4
  - 5
  - 6
  - 7
  - 8
  - 9
  - 10
  - 11
  - 12
  - 13
  - 14
  - 15
  - 16
  - 17
  - 18
  - 19
  - 20
  - 21
  - 22
  - 23
  - 24
  - 25
  - 26
  - 27
  - 28
  - 29
  - 30
  - 31
  - 32
  - 33
  - 34
  - 35
  - 36
  - 37
  - 38
  - 39
  - 40
  - 41
  - 42
  - 43
  - 44
  - 45
  - 46
  - 47
  - 48
  - 49
  - 50
  - 51
  - 52
  - 53
  - 54
  - 55
  - 56
  - 57
  - 58
  - 59
  - 60
  - 61
  - 62
  - 63
  - 64
  - 65
1. a falling or high-rising intonation at the end of an IPU,
  2. a reduced lengthening of IPU-final words,
  3. a reduced intensity level,
  4. a reduced pitch level,
  5. a point of textual completion,
  6. a higher variability in frequency, amplitude of vocal-fold vibration or energy ratio of noise to harmonic components in the voiced speech signal,
  7. longer duration of the IPU.

Regarding relevant places for backchannels, six cues were identified:

1. a rising intonation at the end of an IPU,
2. an increased intensity level,
3. an increased pitch level,
4. a final POS bigram out of ‘DT NN’, ‘JJ NN’, ‘NN NN’,
5. a reduced noise-to harmonics ratio (NHR), and
6. an increased duration of the IPU.

While the authors reliably found some of these cues present when backchannels occurred in the data, the reverse is not true—there need not be a backchannel whenever the cues occur. This can likely be related to the optional nature of backchannels and varies between individual speakers.

Good progress has been made in identifying the triggers of backchannels and barge-ins in human-human conversation. Unfortunately, not many of these results have been implemented within real spoken dialogue systems—even incremental ones—and tested with human users. This may be to some extent because several of the suprasegmental features used in classification can be computationally intensive to obtain online (e.g., the noise-to-harmonics ratio) so that authors have preferred manual annotation. Another reason could be that backchannels, and even barge-ins, are often optional in dialogue so that human production can not be seen as much of a gold standard.

1  
2  
3  
4  
5  
6  
7  
8  
9     *2.3. Information density in spoken language*

10         Humans have a tendency to distribute information across linguistic units in  
11 a way that all information is transmitted within the bounds of a communica-  
12 tive channel, where information-dense segments are marked with an increased  
13 prominence (Bell et al., 2003; Aylett and Turk, 2004; Levy and Jaeger, 2007;  
14 Rajkumar and White, 2011). This finding has been reported at different lin-  
15 guistic levels, including the prosodic, syntactic, syllable and word levels. As an  
16 example, Bell et al. (2003) study the origins of variability in the pronunciation  
17 of function words, such as *the*, *that*, *and* and *of*. Based on the observation that  
18 these words receive a fuller or reduced pronunciation in different linguistic con-  
19 texts, the authors investigate the variation in the length of words, the form of  
20 their vowel (basic, full or reduced) and the presence of final obstruents. They  
21 find that the entropy of words (i.e., how expected they are in their given con-  
22 text) is one of three factors determining the pronunciation of function words.  
23 The other two factors are neighbouring disfluencies and the word’s position in  
24 an utterance. These findings have been confirmed for prosody. Aylett and Turk  
25 (2004) show that prosodic prominence in spoken language is strongly related  
26 to entropy—the less expected a section of speech is in an utterance, the more  
27 likely it is to be prosodically prominent.

28         Other studies have shown evidence for a role of information density—or  
29 entropy—in syntactic reduction (Levy and Jaeger, 2007; Jaeger, 2010). It is  
30 shown that speakers are more likely to produce an optional syntactic comple-  
31 mentiser (e.g., *that*) when entropy is high rather than when it is low. Given  
32 that complementisers such as *that* often have low entropy, they can be used to  
33 reduce the cognitive load on the listener in high-entropy sections. These findings  
34 have also been applied to surface realisation. Using features from information  
35 density, Rajkumar and White (2011) show that the prediction accuracy of a  
36 realisation ranking model is substantially improved for the use of optional *that*  
37 complementisers. Results by Genzel and Charniak (2002) are in line with this  
38 result, where the authors study the entropy of words in English text and find  
39 that all words in a text have roughly the same entropy.

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

Several studies thus seem to suggest a strong relationship between information density and linguistic prominence, including prosodic prominence. From the previous section we know that there is a relation between suprasegmental features and overlapping speech among humans. It is therefore worth asking whether a relationship can be established between information density and overlapping speech. In an earlier study (Dethlefs et al., 2012a), we trained a hierarchical reinforcement learner to predict the best point for a barge-in or backchannel in human-computer interaction. Results showed that a reward function that draws on information density helps to obtain significantly higher user ratings than baselines that are not sensitive to information density. While these findings seemed to point in a positive direction, they were drawn exclusively from a text-based rating study. In this article, we aim to extend them to spoken language and observe the relationship between information density and overlaps in speech.

### 3. Information Density in Spoken Utterances

This section will introduce the concepts behind information density and present some examples from actual interactions with a spoken dialogue system in the restaurant domain. We will also compare the information density in spoken utterances to suprasegmental features used in previous studies. Finally, we show that information density can be obtained from ASR analyses so that use within spoken dialogue systems is feasible.

#### 3.1. Information Theory

Information Theory (Shannon, 1948) is based on two main concepts: a *communicative channel* through which information is transferred in bits and the *information gain*, i.e., the information load carried by each bit. For natural language, the assumption is that humans aim to communicate as closely as possible to the channel’s capacity. If they exceed it, the cognitive load of the listener gets too high. If they stay too far below, too little information is transferred per bit

and the utterance is uninformative. The information gain of each word, which is indicative of how close we are to the channel’s capacity, can be computed using measures of entropy. A related measure to entropy is *information density* which measures the distribution of information across an utterance.

### 3.2. Information Density

Psycholinguistic research as discussed in Section 2.3 has shown that humans have a tendency to distribute information across the linguistic units in an utterance, e.g. words, syllables or phonetic units, in a way that keeps the overall information density within the bounds of the communicative channel. While the exact bits transmitted per unit can vary, practically they seem to lie between 2 and 6 bits with an average of about 4 bits per linguistic unit. This is shown by our own experiments in this article and in Dethlefs et al. (2012a) but also in examples shown in Jaeger (2010).

Relating information density to likelihood of words, we can say that the less frequent a word is, the more information it is likely to carry (Jaeger, 2010). In other words, the lower the probability of a word or n-gram, the higher its information density will be. Compare, for example, the word ‘*the*’ in a corpus of restaurant recommendations against the word ‘*Nepalese*’. Similarly, Jaeger (2010) has shown that the notion of information density can be used to predict the occurrence of ‘*that*’ in relative clauses where it is optional.

Information density is defined as the log-probability of an event (i.e., a word, a phrase or a whole utterance) (Shannon, 1948; Levy and Jaeger, 2007), so that for a utterance formed by  $N$  words  $\{w_1 \dots w_N\}$ , we can compute the incremental point-wise information density (for each word  $w_i$ ) as:

$$ID(w_i) = \begin{cases} \log \frac{1}{P(w_1)} & \text{for } i = 1 \\ \log \frac{1}{P(w_2)} + \log \frac{1}{P(w_2|w_1)} & \text{for } i = 2 \\ \log \frac{1}{P(w_3)} + \log \frac{1}{P(w_3|w_2)} + \log \frac{1}{P(w_3|w_1, w_2)} & \text{for } i = 3 \\ \dots & \\ \log \frac{1}{P(w_i)} + \log \frac{1}{P(w_i|w_{i-1})} + \log \frac{1}{P(w_i|w_{i-1}, w_{i-2})} & \text{for } i > 3 \end{cases} \quad (1)$$

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

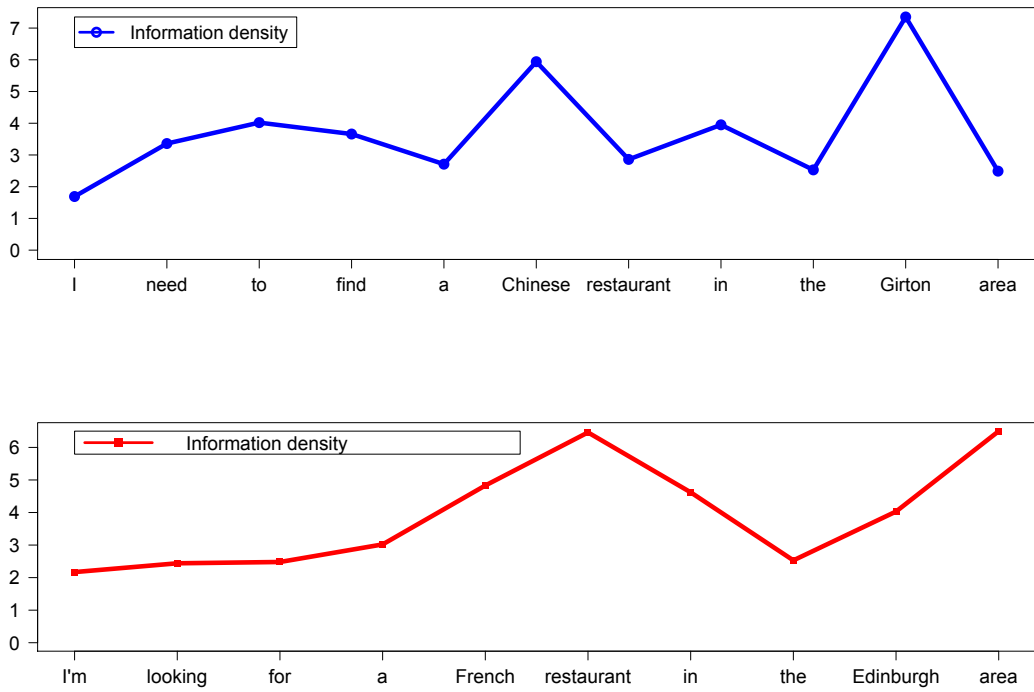


Figure 3: Evolving point-wise information density for two utterances, where information peaks occur at keywords and troughs at function words. Words in an utterance are shown on the x-axis and information density (as computed from Equation 1) is shown on the y-axis.

Note that while typically the context of a word is given by all preceding words of the utterance, several authors have restricted themselves to trigrams for practical reasons (Genzel and Charniak, 2002; Jaeger, 2010).

3.3. Information Density in a Corpus of Spoken User Utterances

To utilise information density in our own study, we first need to estimate an  $n$ -gram model for the domain of interest, in our case information-seeking dialogues in the restaurant domain. To compute the point-wise information density of user utterances (in the form of human transcriptions) at each word

that is spoken, we estimated an  $n$ -gram language model based on the CLASSiC corpus, a corpus of spoken human-system dialogues in the restaurant domain (Lemon et al., 2012). The corpus consists of 1500 dialogues and is freely available.<sup>1</sup> We base our analysis on 1-grams, 2-grams and 3-grams, which has yielded good results in previous work. The language model was trained with the Kylm Language Modelling Toolkit<sup>2</sup> and applied Good-Turing smoothing.

Figure 3 shows examples of the evolving information density of two spoken utterances from different speakers from CLASSiC. In accordance with Equation 1, for the utterance “I need to find a Chinese restaurant in the Girton area”, we can compute the information density of each word  $w_i$  as:

$$ID(w_i) = \begin{cases} \log \frac{1}{P(I)} & \text{for } i = 1 \\ \log \frac{1}{P(need)} + \log \frac{1}{P(need|I)} & \text{for } i = 2 \\ \log \frac{1}{P(to)} + \log \frac{1}{P(to|need)} + \log \frac{1}{P(to|I,need)} & \text{for } i = 3 \\ \dots & \\ \log \frac{1}{P(area)} + \log \frac{1}{P(area|Girton)} + \log \frac{1}{P(area|Girton,the)} & \text{for } i = 11 \end{cases} \quad (2)$$

We can observe that information is distributed across linguistic units that all transmit information between 2 and 6 bits. Peaks or rising information density occur at keywords, such as *Chinese*, *Girton*, *French restaurant* or *Edinburgh* and troughs at function words such as *a* and *the*. Note that while Equation 1 might suggest that information density is ever increasing throughout an utterance (by computing the sums of previous information density scores), one cause of the peaks and troughs we can observe in information density is the fact that they are computed from trigrams of words. This means that whenever the sum of point-wise information density scores for a trigram is lower than the sum of the previous trigram, we would see a trough.

<sup>1</sup>Corpus available from <http://www.macs.hw.ac.uk/ilabarchive/classicproject/data/login.php>.

<sup>2</sup><http://www.phontron.com/kylm/>

OVERLAPS	FEATURES
<b>Barge-in</b>	Falling or high-rising intonation at the end of an IPU ( <b>F0</b> ) Reduced lengthening of IPU-final words ( <b>duration</b> ) Reduced intensity level ( <b>intensity</b> ) Reduced pitch level ( <b>F0</b> ) Point of textual completion ( <b>POS</b> ) Longer duration of the IPU ( <b>duration</b> )
<b>Backchannel</b>	Rising intonation at the end of an IPU ( <b>F0</b> ) Increased intensity level ( <b>intensity</b> ) Increased pitch level ( <b>F0</b> ) Final POS bigram out of ‘DT NN’, ‘JJ NN’, ‘NN NN’ ( <b>POS</b> ) Increased duration of the IPU ( <b>duration</b> )

Table 1: Features used to mark appropriate points of barge-ins and backchannels in spoken user utterances. The feature sets are subsets identified in previous work by Gravano and Hirschberg (2011), described in more detail in Section 2.2. Bold-face features in parentheses denote the annotations that were made to the original sound files.

From the CLASSiC corpus, we compute our language model based on 1200 dialogues (which correspond to 11,000 user utterances) and hold the remainder out for testing. This is to ensure that the information density of user utterances is not computed for the same word string as occurring in the training data.

## 4. Experiments

### 4.1. Data

Our experiments are based on spoken excerpts of interactions between a human and a spoken dialogue system in the restaurant domain. For each excerpt, we compare three alternative points of the system generating overlapping speech over a user utterance. Each triplet contains each of the following:

1. An overlap generated based on features identified by previous work (Gravano and Hirschberg, 2011; Oertel et al., 2012). This includes prosodic



1  
2  
3  
4  
5  
6  
7  
8  
9 information, but also features on bigrams of POS tags indicating com-  
10 pleteness of a constituent. We will refer to this set as **suprasegmental**  
11 **features**. Section 2.2 gave an overview of the general findings of related  
12 work and Table 1 shows the features used to mark appropriate locations  
13 of barge-ins and backchannels in a spoken user utterance. An appropriate  
14 location was marked when at least two of the features were observed at  
15 the same location in a user utterance.  
16  
17  
18  
19

- 20 2. An overlap generated based on **information density features** alone. To  
21 this end, we used the language models trained in Section 3.3 to compute  
22 the information density after each word in a user utterance. Noticeable  
23 troughs in information density (by at least a measure of 2) were marked  
24 as appropriate locations for an overlap, either backchannel or barge-in.  
25
- 26 3. An overlap generated based on **both types of features** described in  
27 points 1 and 2 above. An appropriate location was marked whenever the  
28 conditions for both 1 and 2 above were met at the same location. This  
29 was the case at least once in all user utterances.  
30  
31  
32  
33  
34

35 All tokens for the evaluation study were prepared based on automatically ex-  
36 tractable features to avoid subjective judgement. We used the Praat software for  
37 the extraction of prosodic features and the Stanford POS tagger<sup>3</sup> for POS tags.  
38 Based on these features, we extracted 60 interactions between users and sys-  
39 tems (i.e. pairs of user utterances and overlapping system utterances) from the  
40 CLASSiC corpus and prepared three versions of overlap for each extract. Half  
41 of them involved a backchannel and the other half involved a barge-in so that  
42 potential differences could be observed. Specifically, barge-ins were produced  
43 according to the rules in the top half of Table 1 and backchannels were produced  
44 according to the rules in the bottom half of the table. Table 2 shows an example  
45 excerpt with the three overlap options, here using the backchannel “*Okay*” as  
46 overlap. The overlap is shown in the three different places predicted by the  
47  
48  
49  
50  
51  
52  
53  
54

---

55 <sup>3</sup><http://nlp.stanford.edu/software/tagger.shtml>  
56  
57  
58

1  
2  
3  
4  
5  
6  
7  
8  
9 USER: I'm looking for a moderately priced restaurant in the central area.  
10  
11 SYS-SUP: Okay.  
12  
13 SYS-ID: Okay.  
14  
15 SYS-BOTH: Okay.

16 Table 2: Excerpt of an interaction between a human and three versions of a spoken dia-  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

logue system. The spoken dialogue system produces an overlap (here a backchannel) at three different points during the user’s speech: `SYS-SUP` shows the backchannel location predicted by suprasegmental features, `SYS-ID` shows the backchannel location predicted by information density features, and `SYS-BOTH` shows the backchannel location predicted by both.

different models. As can be seen, `SYS-BOTH` produces the latest backchannel in the utterance but still overlaps with the word “area” uttered by the user. While in our particular system, the *area* slot is the most likely follow-up to “central”, a system with wider coverage, e.g. for slots *park* or *town*, could have missed part of the user’s intention in this case.

We deliberately chose to investigate overlap with a spoken dialogue system, rather than in human-human dialogue, because our ultimate research objective is to integrate our results into spoken dialogue systems, such as the PARLANCE system (Hastie et al., 2013) for the restaurant domain. In the following, we will present experiments in two conditions: (a) overlaps based on transcriptions of user utterances, and (b) overlaps based on the ASR 1-best results obtained during interactions.

#### 4.2. Experiments based on transcribed utterances

*Methodology.* 200 users took part in our rating study and rated altogether 529 triplets of speech overlaps. All users were recruited via the CrowdFlower crowdsourcing platform<sup>4</sup> and were all self-rated native or fluent speakers of English. From CrowdFlower, participants were provided with a link to an external webpage, where the actual rating study was hosted. The webpage is shown in Figure 4 along with instructions on how to use it. Participants were presented with

---

<sup>4</sup><http://www.crowdfunder.com/>

Each of the three recordings below contains an interaction between a human and a dialogue system. The human is trying to ask the system for a restaurant recommendation when the system interrupts at three different points.

Please indicate (by clicking the "yes" button) which of the three points of interruption you think is the most preferable so that the human could still get their message across. Please tick one box only.

After choosing one option, please click the "Generate code" button to obtain a code to feed back into CrowdFlower to obtain payment.

You can then go to the next task by clicking the "Next" button.

Thanks!

## Most appropriate interruption point

Recording 1  yes

Recording 2  yes

Recording 3  yes

When you're done, please click the button to obtain a 4-digit code to type into crowdflower (you cannot be paid otherwise).

[Next](#)

Figure 4: Illustration of the webpage that participants used to rate utterances along with instructions presented. Participants were re-directed to the webpage via CrowdFlower and asked to listen to all three recordings carefully before choosing their preferred option.

three short excerpts of interactions between a human user and a spoken dialogue system. While the human was trying to obtain information from the system, the system would produce overlapping speech at three alternative points during the user's speech. These alternatives corresponded to our three models of producing overlap. Our research question is thus that assuming the system wants to produce an overlap, where is the best or most natural place to do so. While we would in general assume that a spoken dialogue system would always have the option to overlap or not overlap, in this experimental setup we deliberately did

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

Features	Preferences	
	ALL	<b>40 (7.56%)</b>
SUPRASEGMENTAL	BACKCHANNEL	24 (4.54%)
	BARGE-IN	16 (3.02%)
INFO DENSITY (ID)		<b>148 (28%)</b>
	ALL	<b>341 (64.46%)</b>
BOTH	BACKCHANNEL + ID	185 (35%)
	BARGE-IN + ID	156 (29.49%)

Table 3: Results comparing user preference ratings in overlapping speech based on (a) suprasegmental features alone, (b) info density features alone, and (c) both types of features. Preferences along with the percentage out of all 592 ratings are shown for all models. In addition, suprasegmental features are split into preferences of overlaps corresponding to backchannels and barge-ins. The differences between all three models are significant at  $p < 0.001$  according to a Chi-Squared test ranking one out of three options. Results based on transcribed utterances.

not offer users the option to click “no interruption”. Previous work has shown that generating system overlaps can be advantageous under certain conditions, e.g. producing a backchannel to signal continued attention or clarifying a known ASR error early on to avoid follow-up errors (Yankelovich et al., 1995; Hirasawa et al., 1999), so that in this article we are particularly interested in the question so as when would be the best point to produce such overlap. Participants in our study were asked to listen to all three versions carefully and then choose one option as their preferred one. Table 2 showed an example of a triplet.

*Results.* The results will be analysed from two perspectives, (a) overall user preferences for our three different models, and (b) the predictive power of different features with respect to user preferences based on a regression study.

*User Preferences.* Table 3 shows the user preference results organised into three groups: (a) preferences for overlaps based on suprasegmental features only, (b) preferences based on information density features only, and (c) preferences based on both types of features. We can see that users showed a clear preference for

1  
2  
3  
4  
5  
6  
7  
8  
9 the third model, which combines different types of features. For this model, the  
10 overall preference lies at 64.46%, corresponding to 341 out of 529 ratings, and  
11 is significant at  $p < 0.001$  using a Chi-Squared test with 2 degrees of freedom. In  
12 comparison, the model based on information density features alone is preferred  
13 148 times, corresponding to 28%, which is preferred significantly more often than  
14 the version based on suprasegmental features alone (40 ratings, 7.56%). The  
15 differences between these two is again significant with  $p < 0.0001$ . In addition,  
16 we can analyse the effect that particular types of overlaps had, i.e., whether  
17 overlapping backchannels were perceived differently from overlapping barge-ins.  
18 Results are shown in rows 2-3 and 5-6 in the third column Table 3. None of the  
19 differences found are significant, though. Since users rated overlaps from the  
20 same set of samples, the variance between user preferences can be analysed. We  
21 found an average variance of 0.25 across utterances with a maximum of 0.63 for  
22 one overlap and a minimum of 0.03 for another overlap. There is no difference  
23 in the variance between preferences for barge-ins and overlaps.  
24  
25

26 Our results largely confirm the findings of previous work in highlighting the  
27 importance of suprasegmental—i.e., prosodic and grammatical completeness—  
28 features to predict overlaps in spoken language. Moreover, the results provide  
29 evidence that information density has a strong influence on the perception of  
30 appropriate points for overlapping speech, which has so far been overlooked.  
31 The effect of information density appears drastic when comparing human pref-  
32 erences for the suprasegmental system of only 7.56% (which can be said to rep-  
33 resent the current state of the art in overlap prediction) and the system which  
34 takes information density into account in addition (64.46%). The overall human  
35 preference for a model that takes both types of features into account seems to  
36 suggest that information density adds further information over previously used  
37 features, which is possibly particularly advantageous in human-computer inter-  
38 action. Since spoken dialogue systems tend to lack the sophisticated turn-taking  
39 strategies observable in human-human conversation, information density might  
40 provide valuable cues in where overlap is acceptable to humans and where it is  
41 not.  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

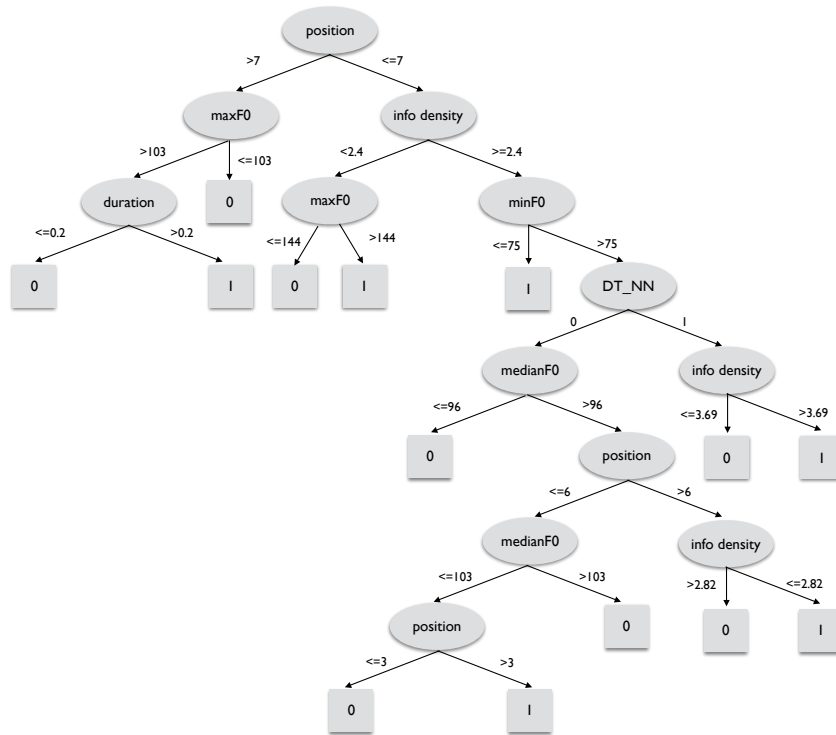


Figure 5: J48 decision tree classifier trained for predicting preferred points of overlap in the experiments with *transcribed utterances*. The features were taken from the annotations described in Section 4.1 and listed in Table 1.

*Variability in User Preferences.* Given the fact that users preferred a model that takes a mixture of features into account to produce overlaps, we were interested in the different ways that each of the features contributes to the overall user preference found. We therefore used pairs of feature vectors characterising each overlap point in our data set and their assigned user preference (1 for preferred, 0 for not-preferred) in a classification experiment. Since we found no difference between backchannel and barge-in overlaps in Section 4.2, both are treated in the same way in this experiment. Feature vectors contained the same features as shown in Table 1 (shown as bold-face in parentheses) at the point that an overlap

1  
2  
3  
4  
5  
6  
7  
8  
9 occurred. In addition, we used a “position” feature indicating the position of the  
10 word at which the overlap occurs. Using the Weka toolkit (Witten and Frank,  
11 2005), we trained a J48 decision tree classifier. In a 10-fold cross-validation, the  
12 classifier reached an accuracy of 86%. In comparison, a simple majority baseline  
13 on the same data only achieved an accuracy of 78%.  
14

15  
16  
17 Figure 5 shows an illustration of our learnt tree, where more important  
18 features can be seen as occurring higher in the tree. Again, we can observe that  
19 while the features identified in previous work play a critical role, information  
20 density is an important factor in determining the overall user preference for  
21 system overlaps. The tree also provides some insights into the cases where our  
22 combined system was not the preferred user option. This occurred most often  
23 when the system would (a) overlap over a keyword e.g. in order to clarify a  
24 previous misrecognised keyword, which is shown by the information density tree  
25 nodes and to a lesser extent by the POS-tag sequences (users did not like the  
26 system to overlap over noun phrases); and (b) when the system would overlap  
27 too early in the utterance because of a longer user silence. In addition to this,  
28 there appears to be subjectivity in the preferences of different versions of overlap  
29 as indicated in the analysis of variance.  
30  
31  
32  
33  
34  
35  
36  
37

### 38 *4.3. Experiments based on ASR analyses*

39  
40 *Results.* To demonstrate that our earlier results hold even in the face of potential  
41 ASR errors, we repeated the experiments described above with a separate set of  
42 60 tokens, i.e. pairs of user utterances with overlapping system utterances. In  
43 the new tokens, overlaps were estimated based on information density in ASR 1-  
44 best hypotheses, which could contain errors in recognition. The language model  
45 used was the same as previously, i.e. trained on transcribed utterances, so as  
46 to make sure that the system would not be trained on ASR errors and thus  
47 “expect” them. 200 users took part in an online rating study that was identical  
48 in its setup to the earlier study. 452 utterances were rated all together. The  
49 results are presented in Table 4.  
50  
51  
52  
53  
54  
55

56 We can see that the results largely confirm the earlier results obtained for  
57  
58

Features	Preferences	
	ALL	<b>42 (9.29%)</b>
SUPRASEGMENTAL	BACKCHANNEL	28 (6.19%)
	BARGE-IN	14 (3.1%)
INFO DENSITY (ID)		<b>166 (36.72%)</b>
	ALL	<b>244 (54%)</b>
BOTH	BACKCHANNEL + ID	114 (25.22%)
	BARGE-IN + ID	130 (28.76%)

Table 4: Results comparing user preference ratings in overlapping speech based on (a) suprasegmental features alone, (b) info density features alone, and (c) both types of features. Preferences along with the percentage out of all 452 ratings are shown for all models. In addition, suprasegmental features are split into preferences of overlaps corresponding to backchannels and barge-ins. The differences between all three models are significant at  $p < 0.001$  according to a Chi-Square test ranking one out of three options. Results based on ASR 1-best hypotheses.

transcribed utterances. An overall preference is revealed in favour of the model that compares suprasegmental and information density features (54%). All differences are significant at  $p < 0.001$  based on a Chi-Square test with 2 degrees of freedom. The variance between user ratings lies at 0.3 on average with a maximum variance of 0.58 for one utterance and a minimum variance of 0 for three utterances.

*User Preferences.* Interestingly, we can observe a slight increase in preference for the model that relies on information density features only in comparison to the combined model. A closer qualitative analysis reveals that even in the case of misrecognitions, it is often still possible to identify keywords. For example, in one case the utterance “Hi (0.52) I’m (0.07) looking (0.07) for (0.39) a (2.28) Thai (1.046) restaurant (0.05)” was misrecognised as “Hi (0.52) I’m (0.07) looking (0.07) for (0.39) a (2.28) five (5.55) restaurant (4.90)”. Despite the error, the information density distribution is similar across both utterances as indicated in parentheses after the respective words. A difference however



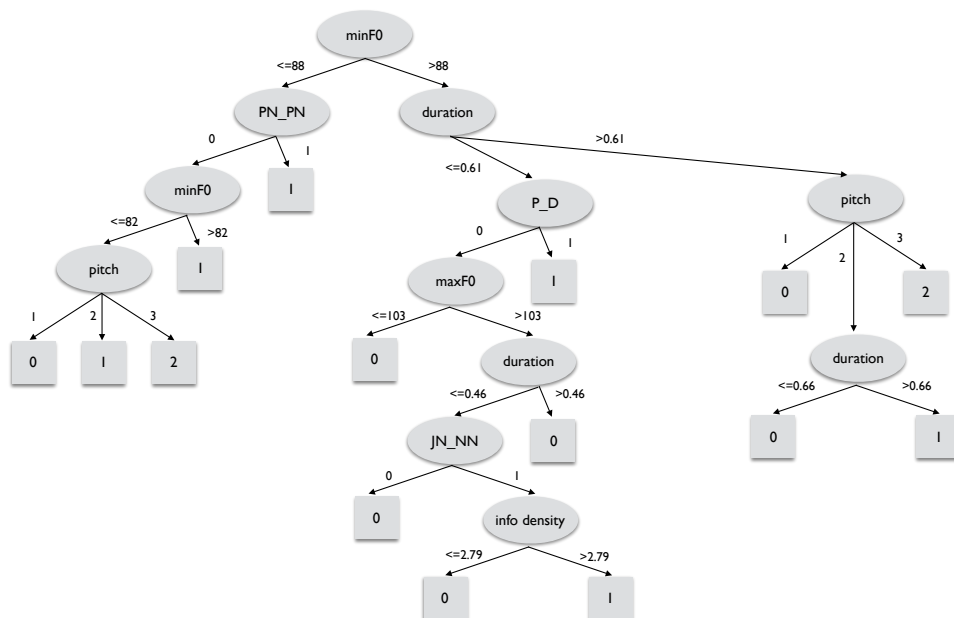


Figure 6: J48 decision tree classifier trained for predicting preferred points of overlap in the experiments with *ASR 1-best hypotheses*. The features were taken from the annotations described in Section 4.1 and listed in Table 1.

occurs around the misrecognised “five”, which results in a sharp increase in information density. This would lead to an overlap occurring following the misrecognised keyword because information density falls again after the word. In this particular example, this leads to an overlap occurring later in the utterance than in for its transcribed counterpart: while in the transcribed utterance, the overlap would occur just after “Thai”, in the ASR utterance, it would get delayed until just after “restaurant”. It is likely that such differences could have led to the overall increase in user preferences for the information density model. Incidentally, when analysing to what extent taking account of the full ASR N-best list would help to improve results, we find that the correct utterance is only part of the N-best list in 41.7% when it is not ranked as 1-best. The automatic speech recogniser that was used in the CLASSiC corpus, on which our results

1  
2  
3  
4  
5  
6  
7  
8  
9 are based had a WER of 53.6, see Rieser et al. (2011).

10  
11 *Variability in User Preferences.* Similarly to the experiments based on tran-  
12 scribed utterances, we can train a J48 decision tree classifier to reveal the most  
13 important determining factors in user preference. The resulting model is shown  
14 in Figure 6. Our decision tree classifier achieved an accuracy of 85% in a 10-  
15 fold cross-validation, while a simple majority baseline only achieved in accuracy  
16 of 79% in the same experimental setup. In contrast to the tree trained from  
17 transcribed utterances, we can see here that user preferences were sensitive to  
18 certain POS-sequences as well the duration and pitch of overlapped segments.  
19 Information density plays a role in connection with POS-sequences as was also  
20 observed in the decision tree for transcribed utterances.  
21  
22  
23  
24  
25  
26  
27

## 28 **5. Conclusions and Future Work**

29  
30

31 Previous work investigating spoken overlap in human-human conversation  
32 has often highlighted the importance of suprasegmental features in marking  
33 suitable points for backchannels or barge-ins. A separate strand of investi-  
34 gations has established that there is a strong relationship between entropy and  
35 prominence in language, manifesting itself in less expected linguistic units being  
36 realised in a marked form. The notion of entropy or expectation in language has  
37 often been related to information density and has been shown to be operational  
38 at the phonetic and prosodic levels of language, among others.  
39  
40  
41  
42  
43

44 In this article, we bring these two strands together. We explore the hypothe-  
45 sis that information density can also be related to the occurrence of overlapping  
46 speech, such as backchannels and barge-ins, in spoken dialogue. In an exper-  
47 iment with human judges, we collected ratings of three models that generate  
48 overlaps based on features relating to (1) prosody and semantic and syntactic  
49 completeness, (2) information density, and (3) both types of information. Hu-  
50 man raters showed a significant preference for the third model ( $p < 0.001$ ). This  
51 demonstrates that besides prosodic and completeness features, humans are in-  
52 deed sensitive to the evolving information density in spoken language. They  
53  
54  
55  
56  
57  
58

1  
2  
3  
4  
5  
6  
7  
8  
9 significantly preferred overlaps at points of low information density over such  
10 occurring at points of high information density. These results are relevant to  
11 research on spoken dialogue systems, especially those that make use of incre-  
12 mental processing, and are therefore able to address communicative features  
13 such as backchannels or barge-ins. Previous work has provided evidence that  
14 interactive feedback mechanisms such as backchannels and barge-ins have im-  
15 portant positive effects on spoken dialogue systems in terms of indicating the  
16 system’s status to the user (Yankelovich et al., 1995; Hirasawa et al., 1999). Our  
17 work represents a further step in the direction of equipping spoken dialogue sys-  
18 tems with interactive feedback mechanisms, which are easy to implement based  
19 on a measure of entropy that requires access to only the words spoken and a  
20 language model of the domain. In particular, we have made the following novel  
21 contributions:  
22

- 23 • We have presented the first study that investigates experimentally the  
24 effects of different suprasegmental features on human turn-taking pref-  
25 erences in spoken dialogue (previous studies have relied on classification  
26 experiments only).
- 27 • Our experiment shows that humans are (at least partially) sensitive to the  
28 peaks and troughs of information density in spoken language.
- 29 • We have demonstrated that the effects found hold for transcribed utter-  
30 ances and for (potentially erroneous) outputs of an ASR module equally.

31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44 Our results show that humans are sensitive to the troughs and peaks in informa-  
45 tion density in spoken language and that this sensitivity does at least partially  
46 guide their preferences on system-initiated overlap. Equipping spoken dialogue  
47 systems therefore with more interactive feedback mechanisms, such as spoken  
48 overlaps in the form of backchannels or barge-ins, could help to make them more  
49 responsive to user overlaps, more communicative of their own status and more  
50 natural and human-like to interact with.

51  
52  
53  
54  
55  
56 Future research will explore the following directions:  
57  
58

- 1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65
1. In this article, we have been able to confirm the positive effect of information density on overlap in spoken language reported earlier in Dethlefs et al. (2012a). While the present speech-based scenario was more natural than the earlier text-based rating study, we plan further experiments using a full spoken dialogue system. This step is essential to show that users perceive and approve of the points of overlap predicted by our method even in a task-based scenario when they are not particularly focused on the differences. This will also be important to show whether system-initiated overlaps are preferable to users over repetition or clarification requests.
2. Information density has been shown to be operational across linguistic levels in human language. However, its effects in natural language processing have not been explored, with a few exceptions (Rajkumar and White, 2011; Dethlefs et al., 2012a). It would be interesting and important to further explore the role that information density can play within spoken dialogue systems or natural language generators. This could address the rankings of a set of competing dialogue acts or surface realisations in the face of the communicative channel and its capacity at different times during an interaction.
3. While this paper has focused on system-initiated overlaps, future work could also explore phenomena of user-initiated overlap. For example, information density might help the system to decide whether the user is offering a backchannel or is barging-in on the system. In the latter case, the system could weigh up the importance of what the user is saying in order to decide whether to yield or try to keep the current turn.
4. Our experiments were based on conventional n-gram language models in accordance with earlier work on information density in language. Since the quality of overlap prediction crucially depends on the robustness of the underlying language model, future work could explore alternative methods. A candidate for investigation are recurrent neural networks, which have been shown to superior to conventional n-gram models in a number of respects (Mikolov et al., 2010).

- 1  
2  
3  
4  
5  
6  
7  
8  
9 5. Information density has so far only been explored at the linguistic level.  
10 However, it has been shown that phenomena of salience and information  
11 structure are also at work in visual and multimodal language scenarios  
12 (Elsner et al., 2014). Future work could attempt to establish a relationship  
13 between information density and multimodal dialogue or natural language  
14 generation scenarios or investigate the effect of information density on  
15 multimodal turn-taking scenarios (Chao and Thomaz, 2013; Nalin et al.,  
16 2012).  
17  
18 6. Throughout this article, we have followed previous work in assuming that  
19 the principles of human-human turn taking are readily transferable to  
20 turn taking in human-computer interaction. In a recent study on the  
21 occurrence of barge-ins of users over a spoken dialogue system, however,  
22 Cuayáhuitl et al. (2013) find that the most likely place for a user to barge-  
23 in on a system is after the first or second dialogue act. This substantially  
24 differs from human-human communication, where barge-ins tend to occur  
25 much later, supposedly because very early barge-ins may be less socially  
26 acceptable among humans. As a conclusion, it may be that not all findings  
27 from human-human data are as readily transferable to spoken dialogue  
28 systems as is often assumed. Future work could provide a systematic  
29 comparison of the differences between the two modes of interaction.  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41

## 42 **Acknowledgements**

43  
44 The research leading to this work was supported by the European Commis-  
45 sion’s FP7 programme under grant agreement number 287615 (PARLANCE)  
46 and the EPSRC grant no. EP/L026775/1 “Generation for Uncertain Informa-  
47 tion (GUI)”. Thanks to Michael White for making the initial suggestion to  
48 explore information density in dialogue and to Chika Ugwuanyi for her help in  
49 setting up the website for experimentation.  
50  
51  
52  
53  
54  
55  
56  
57  
58

1  
2  
3  
4  
5  
6  
7  
8  
9 **References**

- 10  
11 Aylett, M., Turk, A., 2004. The smooth signal redundancy hypothesis: A func-  
12 tional explanation for the relationships between redundancy, prosodic promi-  
13 nence, and duration in spontaneous speech. *Language and Speech* 47(1), 31–  
14 56.  
15  
16  
17  
18 Baumann, T., Buss, O., Schlangen, D., 2011. Evaluation and Optimisation of  
19 Incremental Processors. *Dialogue and Discourse* 2(1), 113–141.  
20  
21 Baumann, T., Schlangen, D., 2011. Predicting the Micro-Timing of User Input  
22 for an Incremental Spoken Dialogue System that Completes a User’s Ongoing  
23 Turn. In: *Proceedings of 12th Annual SIGdial Meeting on Discourse and*  
24 *Dialogue*. Portland, OR.  
25  
26  
27  
28 Bell, A., Jurafsky, D., Fossler-Lussier, E., Girand, C., Gregory, M., Gildea, D.,  
29 2003. Effects of disfluencies, predictability, and utterance position on word  
30 form variation in English conversation. *Journal of the Acoustic Society of*  
31 *America* 113(2), 1001–1024.  
32  
33  
34  
35 Buss, O., Baumann, T., Schlangen, D., 2010. Collaborating on Utterances with  
36 a Spoken Dialogue System Using an ISU-based Approach to Incremental Di-  
37 alogue Management. In: *Proceedings of 11th Annual SIGdial Meeting on*  
38 *Discourse and Dialogue*.  
39  
40  
41  
42 Cathcart, N., Carletta, J., Klein, E., 2003. A Shallow Model of Backchannel  
43 Continuers in Spoken Dialogue. In: *Proceedings of the 11th Conference of the*  
44 *European Chapter of the Association for Computational Linguistics (EACL)*.  
45 *Budapest, Hungary*.  
46  
47  
48  
49 Chao, C., Thomaz, A., 2013. Controlling social dynamics with a parametrized  
50 model of floor regulation. *Journal of Human-Robot Interaction* 2(1), 4–29.  
51  
52  
53 Cuayáhuitl, H., Dethlefs, N., Hastie, H., Lemon, O., 2013. Barge-in Effects in  
54 Bayesian Dialogue Act Recognition and Simulation. In: *Proceedings of the*  
55 *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*.  
56  
57  
58

- 1  
2  
3  
4  
5  
6  
7  
8  
9 Dethlefs, N., Hastie, H., Rieser, V., Lemon, O., 2012a. Optimising Incremental  
10 Dialogue Decisions Using Information Density for Interactive Systems. In:  
11 Proceedings of the Conference on Empirical Methods in Natural Language  
12 Processing (EMNLP-CoNLL). Jeju, South Korea.  
13  
14  
15  
16 Dethlefs, N., Hastie, H., Rieser, V., Lemon, O., 2012b. Optimising Incremental  
17 Generation for Spoken Dialogue Systems: Reducing the Need for Fillers. In:  
18 Proceedings of the International Conference on Natural Language Generation  
19 (INLG). Chicago, Illinois, USA.  
20  
21  
22  
23 DeVault, D., Sagae, K., Traum, D., 2009. Can I finish? Learning when to  
24 respond to incremental interpretation result in interactive dialogue. In: Pro-  
25 ceedings of the 10th Annual SigDial Meeting on Discourse and Dialogue.  
26 Queen Mary University, UK.  
27  
28  
29  
30 Elsner, M., Rohde, H., Clarke, A., 2014. Information Structure Prediction for  
31 Visual-world Referring Expressions. In: Proceedings of the European Associ-  
32 ation for Computational Linguistics (EACL).  
33  
34  
35  
36 Genzel, D., Charniak, E., 2002. Entropy Rate Constancy in Text. In: Proceed-  
37 ings of the 40th Annual Meeting on Association for Computational Linguis-  
38 tics. pp. 199–206.  
39  
40  
41  
42 Gravano, A., Hirschberg, J., 2009. Backchannel-Inviting Cues in Task-Oriented  
43 Dialogue. In: Proceedings of INTERSPEECH. Brighton, UK.  
44  
45  
46 Gravano, A., Hirschberg, J., 2011. Turn-taking cues in task-oriented dialogue.  
47 Computer Speech and Language 25, 601–634.  
48  
49  
50  
51 Hastie, H., Afaure, M.-A., Alexopoulos, P., Cuayhuitl, H., Dethlefs, N., Mil-  
52 ica Gasic, J. H., Lemon, O., Liu, X., Mika, P., Mustapha, N. B., Rieser, V.,  
53 Thomson, B., Tsiakoulis, P., Vanrompay, Y., Villazon-Terrazas, B., Young,  
54 S., 2013. Demonstration of the PARLANCE System: A Data-Driven, Incre-  
55 mental, Spoken Dialogue System for Interactive Search. In: Proceedings of  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

1  
2  
3  
4  
5  
6  
7  
8  
9 the 14th Annual Meeting of the Special Interest Group on Discourse and  
10 Dialogue (SIGdial).

11  
12 Hirasawa, J., Nakano, M., Kawabata, T., Aikawa, K., 1999. Effects of system  
13 barge-in responses on user impressions. In: Proceedings of the Sixth European  
14 Conference on Speech Communication and Technology.

15  
16  
17  
18 Jaeger, T. F., 2010. Redundancy and reduction: Speakers manage syntactic  
19 information density. *Cognitive Psychology* 61, 23–62.

20  
21  
22 Koiso, H., Horiuchi, Y., Tutiya, S., Ichikawa, A., Den, Y., 1998. An analysis  
23 of turn-taking and backchannels based on prosodic and syntactic features.  
24 *Language and Speech* 41, 295–321.

25  
26  
27  
28 Lemon, O., Pietquin, O., (editors), 2012. *Data-Driven Methods for Adaptive*  
29 *Spoken Dialogue Systems*. Springer.

30  
31  
32 Levelt, W., 1989. *Speaking: From Intention to Articulation*. MIT Press.

33  
34  
35  
36  
37  
38 Levy, R., Jaeger, T. F., 2007. Speakers optimize information density through  
39 syntactic reduction. In: Schölkopf, B., Platt, J., Hoffman, T. (Eds.), *Advances*  
40 *in Neural Information Processing Systems 19*. MIT Press, Cambridge, MA.

41  
42  
43  
44 Mikolov, T., Karafiát, M., Burget, L., Černocký, J., Khudanpur, S., 2010. Re-  
45 current neural network based language models. In: Proceedings of INTER-  
46 SPEECH. Makuhari, Chiba, Japan.

47  
48  
49  
50 Morency, L.-P., de Kok, I., Gratch, J., 2008. Predicting listener backchannels:  
51 A probabilistic multimodal approach. In: Proceedings of the International  
52 Conference on Intelligent Virtual Agents. Tokyo, Japan.

53  
54  
55  
56  
57  
58 Nalin, M., Baroni, I., Kruijff-Korbayová, I., Cañamero, L., Lewis, M., Beck,  
59 A., Cuayáhuitl, H., Sanna, A., 2012. Children’s adaptation in multi-session  
60 interaction with a humanoid robot. In: The 21st IEEE International Sym-  
61 posium on Robot and Human Interactive Communication, IEEE RO-MAN  
62 2012, Paris, France, September 9-13, 2012. pp. 351–357.



- 1  
2  
3  
4  
5  
6  
7  
8  
9 Oertel, C., Włodarczak, M., Tarasov, A., Campbell, N., Wagner, P., 2012. Con-  
10 text cues for classification of competitive and collaborative overlaps. In: Pro-  
11 ceedings of the Speech Prosody Conference. Shanghai, China.  
12  
13  
14 Purver, M., Otsuka, M., 2003. Incremental Generation by Incremental Parsing.  
15 In: Proceedings of the 6th UK Special-Interesting Group for Computational  
16 Linguistics (CLUK) Colloquium.  
17  
18  
19  
20 Rajkumar, R., White, M., 2011. Linguistically Motivated Complementizer  
21 Choice in Surface Realization. In: Proceedings of the EMNLP-11 Workshop  
22 on Using Corpora in NLG. Edinburgh, Scotland.  
23  
24  
25 Raux, A., Eskenazi, M., 2009. A Finite-State Turn-Taking Model for Spoken Di-  
26 alog Systems. In: Proceedings of the 10th Conference of the North American  
27 Chapter of the Association for Computational Linguistics—Human Language  
28 Technologies (NAACL-HLT). Boulder, Colorado.  
29  
30  
31  
32 Rieser, V., Keizer, S., Liu, X., Lemon, O., 2011. Adaptive Information Presenta-  
33 tion for Spoken Dialogue Systems: Evaluation with Human Subjects. In:  
34 Proceedings of the 13th European Workshop on Natural Language Generation  
35 (ENLG). Nancy, France.  
36  
37  
38  
39 Schlangen, D., Skantze, G., 2009. A General, Abstract Model of Incremental  
40 Dialogue Processing. In: Proceedings of the 12th Conference of the European  
41 Chapter of the Association for Computational Linguistics (EACL). Athens,  
42 Greece.  
43  
44  
45  
46 Selfridge, E., Arizmendi, I., Heeman, P., Williams, J., 2011. Stability and Accu-  
47 racy in Incremental Speech Recognition. In: Proceedings of the 12th Annual  
48 SigDial Meeting on Discourse and Dialogue. Portland, Oregon.  
49  
50  
51  
52 Shannon, C., 1948. A Mathematical Theory of Communications. Bell Systems  
53 Technical Journal 27(4), 623–656.  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

- 1  
2  
3  
4  
5  
6  
7  
8  
9 Skantze, G., Hjalmarsson, A., 2010. Towards Incremental Speech Generation in  
10 Dialogue Systems. In: Proceedings of the 11th Annual SigDial Meeting on  
11 Discourse and Dialogue (SIGDIAL). Tokyo, Japan.  
12  
13  
14 Skantze, G., Schlangen, D., 2009. Incremental Dialogue Processing in a Micro-  
15 Domain. In: Proceedings of the 12th Conference of the European Chapter of  
16 the Association for Computational Linguistics (EACL). Athens, Greece.  
17  
18  
19  
20 Ward, N., Tsukahara, W., 2000. Prosodic features which cue back-channel re-  
21 sponses in English and Japanese. *Journal of Pragmatics* 32(8), 1117–1207.  
22  
23  
24 Witten, I. H., Frank, E., 2005. *Data Mining: Practical Machine Learning Tools*  
25 *and Techniques*. San Francisco: Morgan Kaufman.  
26  
27  
28 Yankelevich, N., Levow, G.-A., Marx, M., 1995. Designing SpeechActs: Issues  
29 in speech user interfaces. *Language and Speech* 41(3/4), 256–294.  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

**LaTeX Source Files**

[Click here to download LaTeX Source Files: R2-edit.zip](#)