Available online at www.sciencedirect.com

ScienceDirect

Computer Speech and Language 37 (2016) 82–97

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

# Information density and overlap in spoken dialogue☆

Nina Dethlefs[a,*], Helen Hastie[b], Heriberto Cuayáhuitl[b], Yanchao Yu[b], Verena Rieser[b], Oliver Lemon[b]

[a] *University of Hull, Department of Modern Languages, Hull HU6 7RX, United Kingdom*
[b] *Heriot-Watt University, School of Mathematical and Computer Sciences, Edinburgh EH14 4AS, United Kingdom*

Received 4 October 2014; received in revised form 1 November 2015; accepted 2 November 2015
Available online 10 November 2015

**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.


*Keywords:* Overlap; Turn-taking; Information density; Incremental processing; Spoken dialogue systems

## 1. Introduction

Traditionally, the smallest unit of processing in spoken dialogue systems has been a full utterance with strict, rigid turn-taking. More recently, however, work on incremental systems has shown that processing smaller 'chunks' of user input can improve the user experience by providing faster responses and allow more flexibility in turn-taking

☆ This paper has been recommended for acceptance by R.K. Moore.
* Corresponding author. Tel.: +44 01482 462091.
*E-mail address:* n.dethlefs@hull.ac.uk (N. Dethlefs).

http://dx.doi.org/10.1016/j.csl.2015.11.001
0885-2308/

(Skantze and Schlangen, 2009; Purver and Otsuka, 2003; Skantze and Hjalmarsson, 2010; Baumann et al., 2011; Raux and Eskenazi, 2009; Dethlefs et al., 2012b). Incrementality in spoken dialogue systems enables the system designer to model several dialogue phenomena that play a vital role in human conversation (Levelt, 1989), but have so far been absent from most systems. These include more natural turn-taking and grounding through the generation of backchannels and barge-ins—which we will refer to jointly as *overlaps* in this article.

Previous studies on the triggers of backchannels and barge-ins in human–human conversation have revealed the importance of prosodic features, such as pitch, duration, and energy, and features relating to syntactic and semantic completeness (Koiso et al., 1998; Ward and Tsukahara, 2000; Cathcart et al., 2003; Morency et al., 2008; Gravano and Hirschberg, 2009; Oertel et al., 2012). The latter can refer to the grammatical completeness of constituents, e.g., such as a full NP versus just the determiner. We will refer to such features jointly as *suprasegmental*. Most previous studies have relied on manually annotated corpora for their analyses and reported results from held-out datasets, and few findings have been implemented in real spoken dialogue systems.

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

In a study that explored the relationship between information density and overlaps (Dethlefs et al., 2012a), we trained a hierarchical reinforcement learner that could generate backchannels and barge-ins in conversations with human users. The model compared a reward function that was sensitive to information density against a reward function that was not. Results showed that significantly higher human ratings were obtained for the version that took information density into account. While these results are promising, they were drawn from an exclusively text-based rating study, which potentially does not account for the peculiarities of spoken language. In this article, we therefore replicate our earlier experiments in a speech-based rating study, involving word-based as well as suprasegmental features, in order to see whether the earlier results hold in a realistic dialogue setting. Results show a clear human preference for a model that generates overlapping speech based on both suprasegmental and information density features. This is followed by overlaps based on information density features alone and then suprasegmental features alone. The results indicate a strong human sensitivity to the peaks and troughs in evolving information density in spoken language. These results hold even in the face of ASR errors.

We will start Section 2 by discussing related work on overlap in spoken dialogue systems, mainly from the perspective of incremental processing architectures. We will then describe the types of features that previous work has identified as predicting different types of overlaps, and finally the information density effects that have been observed across linguistic units in human language. Section 3 will introduce the notion of information density and exemplify some of its effects on a spoken corpus from the information-seeking dialogue domain. The relation between information density and suprasegmental features in spoken language is also discussed. In Section 4, we describe our experimental setting, data and methodology, and present results on the effect of information density on spoken overlap in dialogue. Section 5 finally draws conclusions and lays out the directions for future research.

## 2. Related work

The production of backchannels and barge-ins has long been recognised to facilitate grounding, feedback and clarifications in human spoken dialogue (e.g., Yankelovich et al., 1995). With the rise of incremental processing architectures (Schlangen and Skantze, 2009; Dethlefs et al., 2012b; Selfridge et al., 2011; DeVault et al., 2009), we now have the opportunity to integrate these phenomena into spoken dialogue systems. This section reviews the state of the art in incremental processing and the identification of triggers for backchannels and barge-ins in human dialogue. Finally, we discuss findings from information density applied to spoken language and draw conclusions on how all aspects can be brought together into an effective model.