# Integrating Prosodic Modelling with Incremental Speech Recognition

**Timo Baumann**

Department for Linguistics
University of Potsdam
Germany
`timo@ling.uni-potsdam.de`

## Abstract

We describe ongoing and proposed work concerning incremental prosody extraction and classification for a spoken dialogue system. The system described will be tightly integrated with the SDS's speech recogntion which also works incrementally. The proposed architecture should allow for more control over the user interaction experience, for example allowing more precise and timely end-of-utterance vs. hesitation distinction, and auditive or visual back-channel generation.

## 1 Introduction

Incremental Spoken Dialogue Systems start processing input immediately, while the user is still speaking. Thus they can respond more quickly after the user has finished, and can even back-channel to signal understanding. In order for this to work, all components of the SDS have to be incremental and interchange their partial results. While both incremental ASR (Baumann et al., 2009) and incremental prosody extraction (Edlund and Heldner, 2006) exist, we here describe work to join both for better processing results.

## 2 Related Work

Skantze and Schlangen (2009) present an incremental spoken dialogue system for a micro-domain, which uses prosody extraction for better end-of-utterance detection, reducing response time for affirmatives to 200 ms (Skantze and Schlangen, 2009). Their prosody extraction is rather crude though, and relies on the words in their number-domain being of equal length and type. We extend their work by implementing a theory-based prosody model, which should be applicable for a variety of purposes.

## 3 Prosody Modelling

The main prosodic features are pitch, loudness and duration. A combination of their contours over time determine whether syllables are *stressed* or not and whether there are intonational boundaries between adjacent words (Pierrehumbert, 1980). Stress and boundary information can then be used to further determine syntactic and semantic status of words and phrases.

Phonemes and their durations are directly available from ASR and syllables can either be reconstructed from a dictionary or computed on the fly.[1] Fundamental frequency and RMSE are calculated on the incoming audio stream. Prosodic features must be normalized by speaker (mostly pitch) and channel (mostly loudness), and phoneme identity from ASR may help with this. Also, we look into FFV (Laskowski et al., 2008) and advanced loudness metering (ITU-R, 2006) for robust pitch and loudness estimation, respectively.

In order to derive features per syllable, contours have to be parameterized. Both TILT (Taylor, 1998) and PaIntE (Möhler, 1998) require right context, which is unavailable in incremental processing, so their methods must be adapted.

Finally, the feature vectors for syllables and word boundaries should be reduced in dimensionality in order to be more useful for higher-level processing. It might also be possible to train classifiers for specific upcoming events. (like end-of-utterance prediction (Baumann, 2008)).

The dataflow through the module is shown in Figure 1. Output is generated for both prosody and word events. The frequency of these events can be different (e. g. several juncture measures could follow each other, indicating juncture growing as time proceeds) and filtering techniques similar to those by Baumann et al. (2009) will be used.

---

[1] The first approach allows predictions into the future, while the second is more flexible.
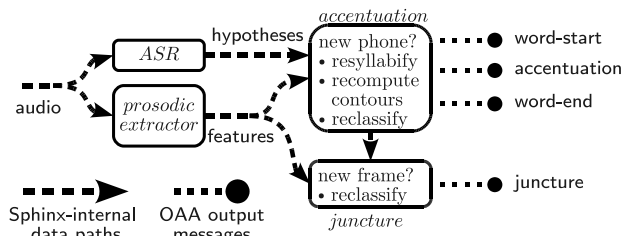
Figure 1: Dataflow diagram for the combined ASR and prosody processing.

## 4 A Prototype System

We construct a micro-domain (Edlund et al., 2008) exposing select problems we try to resolve with our system, and simplifying other problems that are outside of our focus.

The user's task is to order a robot hand to move (glowing) waste above a recycle-bin and to drop it there. In other words, the user controls a 1-dimensional motion and a final stop signal.

A data collection on user behaviour in this domain has been caried out in a Wizard-of-Oz setting with 12 subjects, comprising 40 minutes of audio and 1500 transcribed words.

The data shows the expected phenomena: sequences of directions ("left, left, left, ok; drop"), or use of lengthening ("leeeeft") to express distance. Marking of corrections (of purposeful misunderstandings by the wizard) using prosody, and stress on content words.

Another property of the domain are the consequences for different system actions: going right can easily be undone by going left, but dropping cannot be corrected. Thus, there are different levels of certainty that must be reached for the system to take different actions. Prosody should help in identifying confidences and finality of utterances.

## 5 Possible Extensions, Future Work

The model presented in Section 3 probably exceeds what would be strictly necessary for implementing the system proposed in Section 4. This is by purpose, as it allows for a basis for future extensions:

- Juncture could be calculated for all frames considered word-boundaries by the ASR and this information could be used in addition to the language model's transition probability.
- The syllable stress measure could be used in ASR rescoring to favor likely stress patterns.

- The juncture measure could be easily used in a stochastic parser.
- An obvious extension is a more complex positioning task in a 2D or 3D environment with multiple named entities in them. This would show whether the proposed system scales and introduces reference resolution problems in which prosody might be help.

## Acknowledgments

## References

Timo Baumann, Michaela Atterer, and David Schlangen. 2009. Assessing and Improving the Performance of Speech Recognition for Incremental Systems. In *Proceedings of NAACL-HLT 2009*, Boulder, USA.

Timo Baumann. 2008. Simulating Spoken Dialogue With a Focus on Realistic Turn-Taking. In *Proceedings of the 13th ESSLLI Student Session*, Hamburg, Germany.

Jens Edlund and Mattias Heldner. 2006. `/nailon/` - Software for Online Analysis of Prosody. In *Ninth International Conference on Spoken Language Processing*. ISCA.

Jens Edlund, Joakim Gustafson, Mattias Heldner, and Anna Hjalmarsson. 2008. Towards human-like spoken dialogue systems. *Speech Communication*, 50:630–645.

ITU-R. 2006. *ITU-R BS. 1770-1. Algorithm to measure audio programme loudness and true-peak audio level*. International Telecommunication Union.

Kornel Laskowski, Mattias Heldner, and Jens Edlund. 2008. The fundamental frequency variation spectrum. In *Proceedings of FONETIK 2008*.

Gregor Möhler. 1998. *Theoriebasierte Modellierung der deutschen Intonation für die Sprachsynthese*. Ph.D. thesis, Universität Stuttgart.

Janet B. Pierrehumbert. 1980. *The Phonology and Phonetics of English Intonation*. Ph.D. thesis, MIT.

Gabriel Skantze and David Schlangen. 2009. Incremental dialogue processing in a micro-domain. In *Proceedings of EACL 2009*.

Paul Taylor. 1998. The TILT Intonation Model. In *Proceedings of the ICSLP 1998*, pages 1383–1386.