Incremental Prosody Modelling for Interaction Management in Spoken Dialogue Systems

Timo Baumann <timo@ling.uni-potsdam.de>

Background

Prosody, as manifested in accents and phrasal boundaries, is a recognized and important aspect of human language understanding. Accents are supported by syllables, whereas phrasal boundaries occur between words. The influence of both accents and boundaries on the other hand, extends to surrounding syllables and, in the case of boundaries, preceding words (even as far back as the last accent in the phrase). The manifestation and distribution of accents and boundaries strongly depends on the underlying phoneme and syllable structure and their inherent influences on acoustic prosodic parameters such as pitch, loudness and duration.

Current spoken dialogue systems use only simple acoustic feature extraction and classification, which are mostly used for end-of-turn detection and hesitation classification, ignoring the influences of the underlying structure. Simple systems use silence thresholds of fixed or contextually determined length (Ferrer et al 2002). Others use pitch and power features from a sliding window with fixed duration, but do not incorporate linguistically justified information from the underlying syllables (Nishimura et al 2007, Baumann 2008). Even when ASR results are available, only word information is used, and (implicitly available) timing information is left out.

On the other hand, Schlangen (2006) shows, that linguistic and phonologically formulated prosodic information has a great influence for the detection (and even prediction) of the end-of-turn in natural dialogue. Also, in previous work I examined the automatic detection of accents and phrasal boundaries for speech synthesis corpora (Baumann 2007) and found syllable structure, duration and lengthening to be one of the main correlates – a feature that is unavailable in current systems.

Both Schlangen and I used precomputed and partly hand-annotated information from various sources, which are currently not available in a spoken dialogue system. Furthermore, many of these features were neither calculated incrementally (that is, from left to right through the utterance, only using information available at that point in time) nor under real-time constraints. This would be necessary for the application in an incremental spoken dialogue system.

In terms of incrementality, most prosody models lack the factor of time. That is, they provide a static view on the final prosodic representation of the utterance, once it has completed. They do not take into account what intermediate interpretations or representations could have been built or would have been reasonable and the conclusions that might have been drawn from these intermediate results.

Proposal

My project concerns the design of an incremental prosody model (and the implementation of this model in a system) that takes into account the phonological structure as represented on the phonemic and syllabic level. This includes the necessity of a dynamic view of speech as it happens, including the need to form hypotheses, build alternative representations, to commit to hypotheses or to reinterprete past data.

This task requires the design of a prosodically labelled corpus, which will be used to investigate the properties specific to incremental prosody models, to train components using machine learning techniques as well as to evaluate the final system. As my work is geared towards spoken dialogue systems, I want to acquire a corpus of human-machine dialogue through Wizard-of-Oz simulations and if certain modules' performance allows, I may be able to enlarge the corpus without having to perform further WOZ experiments, but using limited versions of the system.

In a preliminary study (Baumann 2008) I have developed a toy spoken dialogue system that only deals with the prediction of the end-of-turn using simple acoustic-prosodic features and machine-learned classifiers. My next steps will be to incorporate syllable recognition into ASR for this system, to design a more phonologically grounded representation of accents and phrasal boundaries, and to setup and perform WOZ experiments.

My main focus lies on interaction management, for which boundary detection and boundary interpretation (with regard to end-of-turn, hesitation management and possibly back-channel generation) are the main topics. At the same time, better prosody modelling may also be helpful in other tasks, for example ASR, lexical disambiguation, parsing, focus, etc.

Bibliography

Baumann, Timo (2007). *Automatische Erkennung von Akzentuierungen und Phrasierungen in Sprachsynthesekorpora*. University of Hamburg. Hamburg, Germany.

Baumann, Timo (2008). “Simulating Spoken Dialogue With a Focus on Realistic Turn-Taking”. To appear in: *Proceedings of the 13th ESSLLI Summerschool*. Hamburg, Germany.

Ferrer, Luciana, Elizabeth Shriberg, and Andreas Stolcke (2002). “Is the Speaker Done Yet? Faster and More Accurate End-Of-Utterance Detection Using Prosody”. In: *Proceedings of the International Conference on Spoken Language Processing*. Denver, USA.

Nishimura, R., N. Kitaoka, and S. Nakagawa (2007). “A Spoken Dialog System for Chat-Like Conversations Considering Response Timing”. In: *Proceedings of Text, Speech and Dialogue*. Pilsen, Czech Republic.

Schlangen, David (2006). “From Reaction to Prediction: Experiments with Computational Models of Turn-Taking”. In: *Proceedings of Interspeech*. Pittsburgh, USA.