Predicting the intonation of discourse segments from examples in dialogue speech

Alan W Black and Nick Campbell
ATR Interpreting Telecommunications Laboratories
2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-02, JAPAN
awb@itl.atr.co.jp or nick@itl.atr.co.jp

1 Abstract

In the area of speech synthesis it is already possible to generate understandable speech with neutral prosody for simple written texts. However at ATR we are researching into speech synthesis techniques for use in a speech translation environment. Dialogues in such conversations involve much richer forms of prosodic specification than are required for the reading of texts. In order for our translations to sound natural it is necessary for our synthesis system to offer a wide range of prosodic variability, which can be described at an appropriate level of abstraction.

This paper describes a multi-level intonation system which generates a fundamental frequency (F_0) contour based on input labelled with high level discourse information, including speech act type and focusing information, as well as part of speech and syntactic constituent structure. The system is rule driven but the rules and to some extent the levels themselves are derived from naturally spoken dialogues.

Keywords: speech synthesis, intonation, speech act.

2 Introduction

The goal of this study is to predict the intonation of discourse segments in spoken dialogue for synthesis in a speech-translation system. Spontaneous spoken dialogue involves more use of intonational variety than does reading of written prose, so the intonation specification component of our speech synthesizer has to take into account the prosody of different speech-act types, and must allow for the generation of utterances with the same variability as found in natural dialogue.

For example the simple English word *"okay"* is heard often in conversation but performs different functions. Sometimes it has the meaning "I understand.", sometimes "do you understand?", other times it is used as a discourse marker indicating a change of topic, or as an end-of-turn marker signalling for the other partner to speak. Different uses of the word have different intonational tunes.

To determine the degree of relationship between different uses of a word or phrase, and different intonational contours, we analysed a number of spontaneous conversations between a client and agent discussing queries about travel to a conference site. To facilitate analysis of the F_0, we used the 'RFC & tilt' representational system [?] [?].

For evaluation, we implemented an improved version of the intonation prediction rules in [?] to show how realisation of an appropriate type of intonation may be derived by rule for speech synthesis from analysis of a speech database of spontaneous dialogues.

3 Method

A total of sixteen dialogues from the ATR English Multi-modal Interaction (EMMI) database were analysed. These range from two to eight minutes of speech. To reduce speaker variability, only the agent side of each conversation was considered. The dialogues were transcribed, labelled with phonemes using an automatic aligner, and then manually labelled with speech act classes based on those described in [?].

The F_0 was extracted from the speech waveform using a pitch tracker, and then median smoothed. The smoothed contour was RFC labelled ([?]) describing the contour in terms of rise, fall and connection elements, each with a duration and amplitude specification. The phonetic labels were used for syllabification, and aligned with the RFC elements. The elements were then converted to a series of *tilt* events separated by *connections*. The canonical form of a tilt event is a simple "hat" shape, with equal degrees of rise and fall, which can be modified by four continuous parameters: amplitude, duration, accent peak position with respect to the vowel, and slope, which describes the relative height of the rise and fall of the event. -1 denotes a fall with no rise while 1 denotes a rise (with no fall). 0 denotes equal rise and fall while other values state that the rise and fall are of different heights (cf. upstep and downstep).

Tilt labelling is automatically derivable from the waveforms, and provides a higher level representation of the F_0 contour that can be used directly for resynthesis. We assume that the prosodic events labeled in this way are correlated with the linguistic events such as pitch accents and boundary tones. It is this correlation that we are trying to find.

In full synthesis, speech act, part of speech and broad syntactic constituent structure is given. Algorithms further predict prosodic phrase breaks, and accent "position" (see [?] for details). We must then map from this information to tilt events. That is, we have specified *where* pitch accents and boundary tones may go but we need to specify *which* type of tilt event to predict and what parameters should it have.

4 Mapping from speech-act to intonation

To show how such a mapping can be derived from actual data let us look at a specific example from the EMMI conversational database. In the agent side of the 12 dialogues there are 140 occurrences of the word *"okay"*. 112 of which appear alone in their own prosodic phrase. These examples fall into 12 speech act classes, only four of which occured more than twice. These four are: **frame** (37 occurrences), **ack** (31), **d-yu-q** (22) and **accept** (10). **Frame** marks the end of a discourse segment, **ack** is a general acknowledgment, **d-yu-q** is a do-you-understand question, and **accept** as in an immediate reply to a question. It should be stressed that the speech act types were not defined in terms of intonational classes but from the text with respect to discourse function, so it is not necessarily the case that the classes are distinguished by different intonational tunes.

The following table shows the mean start and end F_0 values for these examples for each speech act type. The values are normalised and given in number of standard deviations from the mean (note that the means for the start and end values are calculated separately, and thus cannot be directly compared).

Speech act	accept	d-yu-q	ack	frame
No. of occurs	10	22	31	37
start	0.0	-0.23	0.05	-0.37
end	0.29	1.32	0.02	-0.43

Note the contrast between the *d-yu-q* (a question) examples and the *frame* examples. Although both start below the mean, the question end F_0 is significantly above the mean while the *frame* examples are low.

Of more interest is the tilt event description. In most cases there is just one tilt event (i.e. one accent) in the prosodic phrase. The following table shows the mean tilt parameter for each speech act class.

Speech act	accept	d-yu-q	ack	frame
tilt	0.45	0.73	0.13	-0.27

The tilt parameter indicates whether the F_0 contour is falling (negative value) as it reaches the end of the element or is rising (positive value). Thus we can see that d-yu-q examples have rising events while **frame** examples fall. **Accept** examples rise slightly. For the more neutral **ack**, the end it almost level, having a mean tilt closer to zero.

These parameters can be used directly in the intonation specification of our synthesis system. For example, a **d-yu-q** labelled "okay" can be assigned a start value -0.23 standard deviations from the mean F_0 and event's tilt parameter a value of 0.73.

It should be noted that this example consists of only a single word, with (typically) one accent, in a single prosodic phrase, whereas the number of distinct tilt descriptions that appear more than twice is 16 (covering 86 of the 112 examples). While these do distinguish the speech-act classes well, longer phrases with more accents are harder to abstract. ToBI labelling [?] offers a small range of linguistically defined labels, but it cannot yet be achieved automatically. However, since the mapping from ToBI labels direct to F_0 is known, we are in parallel attempting to predict ToBI-like prosodic labels from the tilt specification, and to compare the contours derived from a ToBI specification with those generated from learnt the tilt parameters.

Similar findings (e.g. certain classes of questions having rising intonation, discourse markers having particular accent types, intonation of discourse cues etc.) have already been presented in the literature ([?], [?], etc.), but the focus of this work is to bring them together into one system where parameters can be automatically extracted to provide values for use in an intonation module for the synthesis of dialogue speech.

5 Summary

This abstract describes a framework for analysing and synthesizing natural dialogue utterances using an intonation coding system in which rules may be derived automatically for spontaneous speech. We are currently extending the analysis to cover the full set of speech acts and will present findings for their intonational correlates in the full paper.

In the oral presentation (or poster) we will also present a set of synthesized utterances to show the difference that such information makes, and compare it with the default intonation from hand coded rules which do not take into account speech act information.