# 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 & nick@itl.atr.co.jp

## ABSTRACT

In the area of speech synthesis it is already possible to generate understandable speech with citation form 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 variation 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 even some elements of the intonation system are derived from naturally spoken dialogues.

## 1. INTRODUCTION

This paper presents a framework for generating intonation parameters based on existing natural speech dialogues marked with high level discourse features.

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.

Already there are a number of intonation systems which allow a specification of intonation at a higher level of abstraction than directly representing the fundamental frequency contour ($F_0$), e.g. ToBI [5], RFC & Tilt [7], [8] and the Fujisaki model [3]. Different intonation systems may represent conceptually different aspects of intonation: ToBI offers a discrete symbolic representation of linguistic intonation patterns; while Tilt offers a representation of physical pitch patterns, that difference is not significant in this work. All these intonation systems offer a method of representation from which varied $F_0$ contours may be generated.

In this paper we are primarily concerned with a system that will predict intonation parameters (for whatever system of intonation representation being used) from higher level discourse information such as speech act, discourse function, syntactic structure and part of speech information.

Different intonation systems offer different parameters which can be modified, the following is a non-exhaustive list of the sort of parameters we wish to predict.

- pitch accent type.
- boundary tone type (both start and end).
- start and end points for phrases.
- pause duration (at least in simple cases).
- reset and declination rates.
- pitch ranges

Although the above parameters are to some extent speaker dependent they may be normalized with respect to a speaker's mean $F_0$, without specifying absolute hertz values for $F_0$.

Predicting the *position* and *type* of appropriate pitch accents and boundaries tones that give the desired intentions is a non-trivial problem, and determining what information is necessary in order to predict such parameters is still an on-going research topic.

In order to be able to predict appropriate dialogue intonation we need the input utterance to be labelled so that distinctions which are not implicit in the words alone may be realised appropriately. With more appropriate information included, a greater diversity in the realisation of the intonation is possible.

The sort of information suggested as affecting intonation is

- Focusing information (global and local)
- new and old information
- speech act (including discourse function)
- contrastive and emphatic markings

In addition, specific words such as "*only*" are known to have specific effects on prosodic patterns. Also varying intonation can be used to mark discourse function, such as change of topic and end of turn.

Thus our discourse dependent intonation system takes explicit discourse features as input and generates explicit intonation parameters. This involves the more basic tasks of predicting prosodic phrasing and accent positioning which we will not discuss directly in this paper where we will concentrate on the more interesting issues of choice of accent type and boundary tone type.

An initial simple hand-crafted set of rules were written which predict intonation parameters (prosodic boundaries, pitch accents and phrase accents) from part of speech, syntactic constituent structure and speech act labels [1]. This system is adequate for simple high level control of prosody but the rules are developed by personal intuition rather than derived from actual data. A more data-driven approach is required to make this system more general.

## 2. MODELLING DISCOURSE INTONATION

In order to build models predicting intonation parameters from discourse features, our data must be labelled with both the parameters we wish to predict and the discourse features we wish to predict from. Finding large quantities of prosodically labelled data is non-trivial and the further constraint that it is labelled with discourse features makes it harder.

The ATR EMMI (English Multi-Modal Interaction) spontaneous database consists of a total of seventeen spontaneous dialogues ranging from two to eight minutes between an agent and clients asking directions and information about a conference. We transcribed the dialogues and labelled them with phonemes using an automatic aligner. They were then manually labelled with speech act classes based on those described in [6], and with prosodic labels using the ToBI system.

Two different systems were used to investigate the relationship between the discourse labels and the observed intonation patterns: one using the hand labelled ToBI system; and another using a purely automatic pitch event classification system.

### 2.1 Analysis with ToBI Labels

When the EMMI database was collected, two types of interaction were recorded: a) multi-modal, where the agent and client could see each other via video, speak through an audio channel, and a display allowed maps to be mutually seen; and b) by telephone alone. For this analysis only the agent side of the 9 multi-modal dialogues were used. As different clients were used in each dialogue, their utterances were felt to be too varied for this analysis.

Each dialogue has been labelled with phonemes, words and ToBI intonation labels (pitch accents, phrase accents, boundaries tones and break indices). Dialogues are further labelled with IFT (broad class speech acts) and discourse acts (fine detailed discourse acts). There are 22 IFT classes and 58 discourse acts [4]. This speech act labelling was done for research in discourse structure but we will show that they are also relevant in predicting prosody. The agent side of the dialogue was **chunked** into discourse act sized sections giving a total of 630 chunks, consisting of a total of 5101 words.

Initially the distribution of pitch accents was investigated. By pitch accents in ToBI we include any label containing a `*`. 1770 (35%) words were labelled with one or more pitch accents. Of these 1676 (95% of accented words) were labelled with `H*` alone. The next most frequent accent type was `L+H*` which appeared only 39 (2%) times. The next was `L*+H` at 9 times.

Using a CART technique [2], decision trees were built to predict pitch accent type for each word. It was assumed pitch accented position was known but type was not. Various trees were built but the result was always same simple tree, predicting `H*` for an accented word. Better results were hoped for but there does not seem to be enough differentiation in the input to reliably predict accents other than `H*` and there are only a few examples of non-`H*` accents. It has been suggested than accents `L+H*` and `L*+H` are used to evoke a semantic scale or a choice of value along some scale [9], but without such marking in our input no learning system will detect that.

A second investigation was to predict boundary tones at the end of discourse act sized chunks. 389 (62%) chunks were terminated with a boundary tone, the other chunks were not terminated with a prosodic phrase break. There are four sequences of phrase accent and boundary tones found at the end of chunks: `L-L%`, `H-L%`, `L-H%` and `H-H%`. The distribution of these four tones is

| Tone | Occur | Percentage |
|---|---|---|
| `L-L%` | 173 | 44% |
| `H-L%` | 110 | 28% |
| `L-H%` | 76 | 20% |
| `H-H%` | 30 | 8% |

This distribution changes for different discourse acts. For example the **instruct** discourse act and **do-you-understand** discourse act have the following distributions.

| **Instruct** | | | **Do-you-understand** | | |
|---|---|---|---|---|---|
| Tone | Occur | Percent | Tone | Occur | Percent |
| L-L% | 13 | 46% | L-H% | 6 | 46% |
| H-L% | 11 | 39% | L-L% | 3 | 23% |
| L-H% | 2 | 7% | H-H% | 2 | 15% |
| H-H% | 1 | 3% | H-L% | 2 | 15% |

The discourse acts were sorted into four groups according to which tone they most often end in.

A CART decision tree was then built to predict ending tone. Various features were used but the best results were achieved from the following factors:

- Most frequent ending tone for this discourse act
- Break index preceding final word
- Break preceding the word preceding the final word
- preceding IFT
- current IFT
- current discourse act

This produces a decision tree of depth 16 that can correctly predict the ending tone of discourse act sized prosodic phrase given the above features 60% of the time. If we simply select the most frequent ending tone the accuracy drops to 49%.

This decision tree was used in our synthesizer to predict more suitable ending tones for different discourse acts. Even the simple pitch accent prediction of only H* produces more varied dialogue speech than a more naive text to speech prediction system that ignores speech act information.

### 2.2 Analysis with Tilt Labels

In this test we used the RFC and Tilt intonation system [7], [8]. RFC and Tilt encode salient pitch events found in the speech without explicitly identifying linguistic intonation events as ToBI does. Tilt makes no distinction between boundary tones, phrase accents and pitch accents. Its main advantage is that it can automatically label data. The process of tilt labelling is achieved by the following process. The $F_0$ is extracted from the speech waveform using a pitch tracker, and then median smoothed. The smoothed contour is RFC labelled [7] segmenting the contour into a sequence of rise, fall and connection elements, each with a duration and amplitude specification. The phonetic labels are used for syllabification, and aligned with the RFC elements. The elements are 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 tilt, 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). Although no formal tests were done on the accuracy of labelling this data, measurements on other data have been carried out [7] [8].

In this test we looked at how the word "*okay*" is realized in the EMMI dialogue data. In all the dialogues (both multi-modal and telephone only modes) there are 140 occurrences of the word *"okay"* spoken by the agent, 112 of which appear alone in their own prosodic phrase. These examples fall into 12 discourse 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 noted that these discourse act types were not defined for differentiating intonational classes, they were defined to represent discourse function, so it is not necessarily the case that all classes are distinguished by different intonational tunes.

The following table shows the mean start and end $F_0$ values (standard deviations are shown in parenthesis) for these examples for each discourse 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).

| | accept | d-yu-q | ack | frame |
|---|---|---|---|---|
| occurs | 10 | 22 | 31 | 37 |
| start | -0.10 (0.6) | -0.23 (1.2) | -0.73 (1.3) | -0.13 (1.4) |
| end | 0.10 (0.8) | 0.92 (0.9) | -0.11 (0.8) | -0.47 (0.9) |

All the start values are below the mean start value, this is probably because longer phrases in general start higher and all these phrases are short. Student t-tests confirm that end values for **frame** examples are significantly lower than end values of other examples ($t = 3.9$, $df = 98$, $p < 0.001$). Also the end values of **d-yu-q** discourse acts are significantly higher than those of other discourse acts ($t = 5.55$, $df = 98$, $p < 0.001$), as would be expected for a question.

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 (and standard deviation) for each discourse act class.

| | accept | d-yu-q | ack | frame |
|---|---|---|---|---|
| tilt | 0.45 (0.89) | 0.74 (0.55) | 0.19 (0.93) | -0.28 (0.79) |

The tilt parameter indicates the amount of rise and fall at that point in the $F_0$ contour. Values near zero represent events with equal rise and fall, values closer to 1.0 represent rise only while values closer to -1.0 represent a fall with no preceding rise. Thus we can see **frame** examples have significantly more downward tilt than the other discourse acts ($t = 4.13$, $df = 98$, $p < 0.001$), while **d-yu-q** examples are predominantly rising events ($t = 3.68$, $df = 98$, $p < 0.001$)

These three results show a significant difference

between different renderings of "*okay*". **Frame** examples start higher and tilt more downward to end lower than **ack** examples which tend to start lower and not tilt as much ending higher. **D-yu-q** start relatively neutral but rise up to significantly higher values than other examples.

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 $F_0$ be given an event whose tilt parameter has a value of 0.74.

## 3. DISCUSSION

It is important to realise that although it may be possible to predict so-called "default intonation" for plain text any variation from default emphasis, focus, discourse function etc. would have to be derived from the text. The additional discourse features are not intonation features in themselves but describe function and are necessary to predict more appropriately varying intonation. At ATR within a framework of telephone translation a much richer input is available as part of the translation process, so IFT, focus etc. are available directly as input information, with no special processing required to predict them.

Because the number of non-`H*` pitch accents in the EMMI database is so small, it seems unlikely that a more complex pitch accent prediction model than simply predicting `H*` (or its Tilt equivalent) for accented words, can be found based on this current data and its labelling. Even with a larger database with more variation in pitch accents, in order to differentiate between pitch accent types we would most probably need richer labelling of the input data identifying focus, new and old information, contrastive marking, emphasis etc.

We do not yet wish to choose between the two methods of labelling intonation system presented here, in fact we are likely to add to them. Lack of prosodically labelled data is probably our greatest hurdle. Any reasonable form of labelling cannot be ignored. Tilt labelling has the advantage of being automatically derivable from waveforms, though does not explicitly distinguish between pitch accents, phrase accents and boundary tones. Automatic ToBI labelling is under consideration [10] and would aid us greatly in labelling of more databases. Although hand labelling is resource intensive it is becoming easier with appropriate tools. Also as it is becoming a standard it is likely that more suitable data will soon become widely available.

The framework presented here has been designed to be language independent and to some extent intonation theory independent. A Japanese version of the same speech dialogue database has been recorded and is currently being labelled with J-ToBI labels, and we will apply similar analysis techniques to that data.

## 4. SUMMARY

This paper discusses the synthesis of intonation for dialogue speech. It presents a framework which allows prediction of intonation parameters from input labelled with factors describing discourse function. If factors such as speech act, syntactic constituent structure, focus, emphasis, part of speech, etc. are labelled in the input then more varied intonation patterns can be predicted.

Rather than writing translation rules directly, techniques for building such rules from prosodically labelled natural dialogue speech are presented. Two analyses of aspects of the ATR EMMI dialogue database are presented showing how speech act information can be used to distinguish different intonational tunes. The main conclusion we can draw from these analyses is that discourse act plays a significant role in predicting intonational tune.

The initial results look promising and we will continue to expand the system for English and also for Japanese. There are still questions as to which modeling techniques to use but at present the greatest problems lie in labelling, both in the task of actually labelling data, and in deciding on what level to label the data.

## 5. REFERENCES


[1] A. W. Black and P. Taylor. Assigning intonation elements and prosodic phrasing for English speech synthesis from high level linguistic input. In *Proceedings of ICSLP 94*, volume 2, pages 715–718, Yokohama, Japan, 1994.

[2] L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification and Regression Trees.* Wadsworth & Brooks, Pacific Grove, CA., 1984.

[3] H. Fujisaki. Dynamic characteristics of voice fundamental frequency in speech and singing. In P MacNeilage, editor, *The Production of Speech*, pages 39–55. Springer-verlag, 1983.

[4] M. Seligman, L. Fais, and M. Tomokiyo. A bilingual set of communicative act labels for spontaneous dialogues. Technical Report TR-IT-0081, ATR Interpreting Telecommunications Laboratories, Kyoto, Japan, 1994.

[5] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg. ToBI: a standard for labelling English prosody. In *Proceedings of ICSLP92*, volume 2, pages 867–870, 1992.

[6] A. Stenström. *An Introduction to Spoken Interaction.* Longman, London., 1994.

[7] P. Taylor. The Rise/Fall/Connection model of intonation. Speech Communications, forthcoming, 1994.

[8] P. Taylor and A. W. Black. Synthesizing conversational intonation from a linguistically rich input. In *Proc. ESCA Workshop on Speech Synthesis*, pages 175–178, Mohonk, NY., 1994.

[9] G. Ward and J. Hirschberg. Implicating uncertainty: the pragmatics of fall-rise intonation. *Language*, 61:747–776, 1985.

[10] C. Wightman and N. Campbell. Automatic labeling of prosodic structure. Technical Report TR-IT-0061, ATR Interpreting Telecommunications Laboratories, Kyoto, Japan, 1994.